DECLARATION OF
VALERI VASIOUKHIN

Exhibit 1008 (H) Part 2

Detection of Residual Lymphoma Cells by Polymerase Chain Reaction in Peripheral Blood Is Significantly Less Predictive for Relapse Than Detection in Bone Marrow

By John G. Gribben, Donna Neuberg, Michael Barber, Jennifer Moore, Kelly W. Pesek, Arnold S. Freedman, and Lee M. Nadler

Polymerase chain reaction (PCR) amplification of the t(14;18) has been shown to be a highly sensitive method to detect minimal residual disease in patients with non-Hodgkin's lymphoma (NHL) whose tumors bear this translocation. The ideal tissue source to detect residual lymphoma would be from a previously involved lymph node. However, lymphoid tissue is rarely available once patients achieve complete remission. Although PCR amplification has been used to detect residual lymphoma cells in both bone marrow (BM) and peripheral blood (PB) of patients in complete remission, it is presently unknown whether BM and PB are equivalent tissue sources to detect residual disease. In the present study, we compared the clinical utility of the detection of residual lymphoma in both the BM and the PB of patients with advanced-stage non-Hodgkin's lymphoma before, at the time of, and after high-dose therapy and autologous BM transplantation (ABMT). The detection of residual lymphoma in either the BM or PB was associated with decreased disease-free survival. However, in the present study, 44% of patients who relapsed had no evidence of circulating lymphoma cells in their PB. At the time of BM harvest, PCR-detectable residual lymphoma cells were detected in 211 of 212 patients; although, in a subset of these patients analyzed, lymphoma cells were detected in the peripheral blood of only 49% of patients. When residual lymphoma cells within the autologous BM are infused into the patient these cells are rapidly detectable circulating in the PB in the patient. These cells continue to circulate during the immediate posttransplant period and be detectable in the PB in the majority of patients who are infused with marrow containing residual lymphoma. We conclude that BM is a more informative tissue source than PB in detecting minimal residual disease at the time of and after ABMT, and that contamination of PB early after ABMT appears to be the consequence of reinfusion of lymphoma cells within autologous marrow.
© 1994 by The American Society of Hematology.

DESPITE ADVANCES in the treatment of advanced stage lymphomas, the majority of patients ultimately relapse. Relapse from clinical complete remission (CR) is believed to be from minimal numbers of residual lymphoma cells that are below the limit of detection by clinical staging procedures. A number of studies using more sensitive techniques have now clearly shown that patients in clinical CR may harbor minimal numbers of residual malignant cells in the bone marrow (BM) and peripheral blood (PB). This has been shown by detection of monoclonal B cells,[1] by DNA restriction-fragment analysis of Ig gene rearrangements,[2,3] by clonogenic assays[4,5] and most recently, by polymerase chain reaction (PCR) amplification.[6-8] Specifically, cloning of the chromosomal breakpoints in the t(14;18)[9-11] has made it possible to use PCR amplification to identify lymphoma cells containing the t(14;18). This translocation is present in up to 85% of patients with follicular lymphomas and 30% of patients with diffuse lymphoma.[12-14] In this translocation the *BCL2* proto-oncogene on chromosome 18 is translocated to the Ig heavy chain (IgH) region on chromosome 14, creating a hybrid *BCL2*/IgH gene product. Using PCR amplification, one lymphoma cell with this translocation can be detected in up to 10^6 normal cells.[15]

Clearly, the ideal tissue source to determine whether a patient has residual lymphoma would be a previously involved lymph node. However, lymphoid tissue is rarely available from patients once they achieve CR. Because of their accessibility of sampling, BM and PB samples have been studied most extensively by PCR analysis, both at diagnosis,[16-18] and after treatment.[19,20] Although these studies have illustrated that PCR could be used for the sensitive detection of lymphoma, the clinical significance of the detection of such minimal residual lymphoma in the PB or BM has never been conclusively established. For example, PCR amplification detected minimal residual lymphoma cells in the circulation of patients presenting with advanced stage follicular lymphoma who had very long-term clinical CR.[21] Recently it has also been shown that residual lymphoma cells can be detected by PCR after radiotherapy for localized lymphoma.[22] In patients with advanced stage lymphoma, PCR amplification invariably detected cells bearing the *BCL2*/IgH translocation in the BM before and after conventional dose chemotherapy.[15,23] However, after high-dose myeloablative therapy and autologous BM transplantation (ABMT), the persistence or reappearance of PCR-detectable residual lymphoma in the BM was associated with an increased risk of subsequent relapse of disease.[23]

In the present report we assess the clinical utility of the assessment of minimal residual lymphoma by PCR analysis of PB samples after ABMT in patients with advanced stage lymphoma. The detection of residual lymphoma cells both in the BM and PB was associated with decreased disease-free survival. However, whereas lymphoma cells were detectable in the BM before relapse in 96% of patients, 44% of patients had no detectable cells in the PB before relapse. In light of these findings, we compared PB and BM samples at the time of BM harvest. We show that, whereas the BM is almost invariably infiltrated after conventional chemotherapy, circulating lymphoma cells are detected in only half of

From the Dana-Farber Cancer Institute, Boston, MA.
Submitted December 16, 1993; accepted February 14, 1994.

Supported by Grants No. CA-34183 (L.M.N) and CA-55207 (A.S.F) from the National Cancer Institute, FO5-TWO44 (JGG) from the Fogarty International Center, and CA 06516-28 (D.N.) from the National Institutes of Health.

Address reprint requests to John G. Gribben, MD, Division of Hematologic Malignancies, Dana-Farber Cancer Institute, 44 Binney St, Boston, MA 02215.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. section 1734 solely to indicate this fact.

© 1994 by The American Society of Hematology.
0006-4971/94/8312-0022$3.00/0

Exhibit M

the patients studied. In addition, PCR amplification is capable of detecting circulating lymphoma cells within 2 hours of infusion of residual lymphoma in the autologous harvested BM. We conclude that the clinical significance of the assessment of minimal residual lymphoma is clearly influenced by the tissue source used to detect such disease and that BM is a more informative source of tissue to assess minimal residual lymphomatous infiltration than PB.

MATERIALS AND METHODS

Patients. From 1982 to December 1992, 386 patients with advanced stage B-cell NHL have undergone ABMT at a single institution. In all cases, immunophenotyping confirmed that the lymphoma cells expressed CD20. All patients had chemosensitive disease at the time of ABMT as assessed by the achievement of protocol-eligible minimal disease after induction or salvage therapy. Minimal disease criteria were defined as either a clinical CR or a partial remission (PR) to maximum tumor masses of 2 cm or less, and BM infiltration of less than 20% of the intratrabecular space. A PCR-amplifiable *BCL2*/IgH translocation was identified in BM or lymph node samples obtained before ABMT from 212 of these patients. Samples analyzed from these 212 patients are included for subsequent analysis. All patients were treated with an identical high-dose therapy regimen consisting of high-dose cyclophosphamide and total body irradiation (TBI) before ABMT. Immunologic purging was performed using a cocktail of three monoclonal antibodies (MoAbs) as previously described.[15] All patients were monitored for palpable disease, radiologic progression or evidence of morphologic infiltration in the BM at six monthly intervals for the first 2 years after ABMT and at yearly intervals thereafter.

BM and PB samples. BM aspirates and biopsies were obtained from all 212 patients at the time of initial evaluation and at the time of BM harvest before ABMT. These samples were used to document that the lymphoma expressed the *BCL2*/IgH translocation. PB and BM samples were obtained at the time of follow up clinic visits at 6, 12, 18, and 24 months after ABMT, and yearly thereafter until the time of documented clinical relapse of disease. All samples obtained before January 1990 were cryopreserved. Previous analysis has shown that identical results are obtained by PCR analysis of fresh and cryopreserved BM samples.[15] Normal BM and PB were obtained from normal volunteers. In addition samples were collected from all patients who had NHL that did not contain a PCR detectable *BCL2*/IgH translocation and these samples were used throughout the study to ensure that false positive results did not occur. Samples of the harvested BM after immunologic purging were available for analysis in 202 of these 212 patients. Although the investigative protocol included the collection and cryopreservation of BM samples from these patients, PB samples were not routinely collected. Therefore, at the time of BM harvest, only 153 PB samples, collected from 45 patients, were available for analysis. To address particular questions regarding the utility of PCR to detect infused lymphoma cells after ABMT, PB samples were collected prospectively from 60 consecutive patients undergoing ABMT who had a documented *BCL2*/IgH translocation. Samples from these patients were collected before and at 2 hours after infusion of autologous BM, on days 1 and 7 after ABMT, and weekly thereafter until discharge from hospital. All samples were obtained with the patients' informed consent and with the approval of the human protection committee of our institution. Genomic DNA was isolated by cell lysis using nonionic detergents and digestion with proteinase K (Sigma Chemical Co, St Louis MO).[24] DNA was extracted from the cell lines DHL-6 and RL (which contain a PCR-amplifiable translocation at the major breakpoint region [MBR]) and from DHL-16 (which contains a PCR-amplifiable translocation at the minor cluster region [mcr]) and were used as positive controls. RL was a gift from Dr W. Urba, National Institutes of Health Biologic Response Modifiers Branch, Frederick, MD. DHL-6 and DHL-16 were gifts from Dr A. Epstein, University of Southern California, Los Angeles, CA.

PCR amplification. PCR amplification of the major[9-11] and minor[25] breakpoint regions of the *BCL2*/IgH translocation were performed using nested oligonucleotides in a modification of the method previously described.[19] Samples containing 1 μg of genomic DNA were amplified for 25 cycles in a thermal cycler (Cetus, Emoryville, CA) in a 50-μL final volume containing 50 mmol/L potassium chloride, 10 mmol/L TRIS hydrochloride (pH 8.3), 2.25 mmol/L magnesium chloride, 0.001% gelatin, 200 μmol/L each of deoxyadenosine triphosphate, deoxycytidine triphosphate, deoxythymidine triphosphate and deoxyguanosine triphosphate, 1.5 U Taq polymerase (Cetus), and 20 nmol/L oligonucleotide primers, 5′ ACC TGA GGA GAC GGT GAC C 3′ for the consensus J_H region, with 5′ CAG CCT TGA AAC ATT GAT GG 3′ for the MBR or 5′ CGT GCT GGT ACC ACT CCT G 3′ for the mcr. Amplification was performed by denaturation for 1 minute at 94°C, annealing for 1 minute at 55°C for the MBR and 58°C for the mcr and extension at 72°C for 1 minute with a final extension period of 10 minutes. Reamplification for 30 cycles of a 5-μL aliquot of the initially amplified mixture was reamplified in a final volume of 50 μL using oligonucleotides internal to the original primers, 5′ ACC AGG GTC CCT TGG CCC CA 3′; for the consensus J_H region, with 5′ TAT GGT GGT TTG ACC TTT AG 3′ for the MBR or 5′ CCT GGC TTC CTT CCC TCT G 3′ for the mcr, using the identical cycle parameters used in the initial amplification step. Aliquots of the final reaction product were analyzed by electrophoresis in 4% agarose gels containing ethidium bromide and visualized under ultraviolet light. Standard precautions against cross-contamination of amplified material were taken.[26] With each amplification a weak-positive control consisting of DNA from a 10^{-5} dilution of the appropriate cell line in normal BM cells and a negative control consisting of DNA from a cell line that did not contain the appropriate *BCL2*/IgH translocation were performed. In samples with no detectable PCR product, PCR amplification was repeated using oligonucleotides for the V region of the human B-cell activation antigen B7 and confirmed that PCR-amplifiable DNA was isolated from each sample that was included in the present analysis.

Statistical analysis. Disease-free survival was estimated by the method of Kaplan and Meier.[27] Differences in disease-free survival were assessed using the logrank test.[28] Associations between the results of PCR analysis of BM samples after immunologic purging and of PB samples, both at the time of marrow harvest and within 1 month after ABMT, were assessed using the Fisher exact test.[29] The results of PCR analysis of PB samples obtained before and after infusion of autologous BM were examined using McNemar's test.[30]

RESULTS

Identification of patients with a BCL2 translocation. At the time of analysis, 386 patients with B-cell NHL had been treated at our institution with an identical protocol that included treatment to a protocol-eligible minimal disease state followed by TBI, high-dose cyclophosphamide and immunologic purging of autologous marrow before transplantation. A PCR-amplifiable *BCL2*/IgH translocation was identified in diagnostic tissue obtained from 212 patients. The patient characteristics are shown in Table 1.

Presence of residual lymphoma cells in BM and in PB is predictive for outcome after ABMT. Serial BM and PB samples were obtained after ABMT and PCR analysis performed to determine whether the presence of residual lymphoma cells in the BM and PB predicted for subsequent relapse. PCR analysis at the MBR and mcr were performed

Table 1. Patient Characteristics

BCL2/IgH translocation	
MBR	167
mcr	45
Histology	
Low grade	149
Transformed from low grade	17
Intermediate grade	44
High grade	2
At ABMT	
CR	72
PR	140
BM infiltration	
In history	147
At ABMT	96

and investigators were blinded to clinical outcome. For inclusion in this analysis, patients had a documented history of lymphoma containing a *BCL2*/IgH translocation that could be detected by PCR, and samples were available for analysis that were obtained while the patient was documented to be in clinical CR after ABMT. Remission samples were not available from all of the 212 patients with a documented PCR-detectable *BCL2*/IgH translocation. A total of 863 remission BM samples were available after ABMT from 146 patients, 47 of whom have relapsed to date. A total of 1478 remission PB samples were available after ABMT from 168 patients, 36 of whom have relapsed to date. A further 108 PB and BM samples were obtained at the time of documented relapse and were excluded from subsequent analysis because it was felt that these samples had no predictive value.

No detectable lymphoma cells were found at any time after ABMT in any of the BM samples analyzed from 49 patients (34%). Two of these patients (4%) relapsed at 10 and 16 months after ABMT. BM samples from these two patients showed no evidence of lymphoma on PCR analysis at 3 or 6 months after ABMT and no further BM samples were available for analysis before their subsequent relapse. PCR analysis showed lymphoma in the BM at some time after ABMT in 97 patients, 43 of whom (44%) have relapsed. The median disease-free survival for this group was 27 months after ABMT. All 43 of these patients had lymphoma detected by PCR analysis in the BM samples obtained at the previous clinic visit before their subsequent relapse. Therefore, PCR analysis detected lymphoma cells in the BM of 43 of 45 patients (96%) who have relapsed to date. The estimated disease-free survival after ABMT of these patient groups are shown in Fig 1A. The detection of lymphoma cells by PCR analysis at any time after ABMT was associated with a decreased disease-free survival ($P < .00001$).

No PCR-detectable lymphoma cells were detected by PCR analysis of PB samples analyzed from 95 of these patients (57%), and 16 of these patients (17%) have relapsed to date. Residual lymphoma cells were detected at some time after ABMT in PB samples obtained from 73 patients (43%), and 20 of these patients (27%) have relapsed to date (median disease-free survival 43.5 months). Therefore, lymphoma cells in the PB were detectable by PCR preceding relapse in 20 of 36 patients (55%) who have relapsed to date. The estimated disease-free survival of these patient groups is shown in Fig 1B. The presence of PCR-detectable lymphoma cells in the PB after ABMT was associated with an increased likelihood of subsequent relapse ($P = .01$ by log rank test). Although the detection of lymphoma cells in both the BM and PB was associated with decreased disease-free survival after ABMT, a detection of lymphoma in PB was less predictive than the detection of lymphoma in BM (logrank $P = .001$). As a predictive assay for subsequent relapse, the detection of PCR-detectable lymphoma cells in the BM at any time after ABMT was associated with a 96% sensitivity and a 44% specificity. Further follow-up will be required to determine whether this low specificity is a result of insufficient follow-up of the patients who currently have minimal detectable disease or if indeed the detection of minimal disease is not always a harbinger of subsequent relapse in these patients. In contrast, the detection of residual lymphoma in PB was associated with only a 55% sensitivity and a 67% specificity. Therefore, the BM is a more informative tissue source in which to detect residual disease.

Detection of residual lymphoma cells in the BM and PB at the time of marrow harvest. Because we found that BM was a more informative tissue source to detect minimal residual disease after ABMT than PB, we sought to determine whether this was also the case at the time of BM harvest. We previously reported that PCR analysis of BM samples at the time of BM harvest detected residual disease in all patients analyzed.[23] Although 96 patients (45%) had morphologic evidence of BM infiltration at the time of marrow harvest, PCR analysis of the BM obtained showed residual BM infiltration in 211 of 212 patients (99.5%) in the present series. Only one patient had no evidence of BM infiltration by PCR analysis in any of four BM samples obtained at the time of harvest.

Unfortunately, PB samples at the time of their marrow harvest had not been cryopreserved because collection of PB samples were not included in the investigative protocol. Therefore, we obtained multiple PB samples at the time of marrow harvest from 45 patients. The patient characteristics of these 45 patients are shown in Table 2. This subgroup of patients are representative of the whole group of 212 patients in all aspects except the results obtained after purging. After purging, residual lymphoma cells were detected after immunologic purging in 67% of patients in this group, but in 52% of the remaining 157 patients for whom post purging samples were available ($P = .09$).

A total of 153 PB samples and 144 BM samples were obtained from these 45 patients at the time of marrow harvest. Lymphoma cells were detected in the BM by PCR analysis in 44 of these patients (98%). In contrast, PCR analysis detected circulating lymphoma cells in the PB in only 22 of these patients (49%). Of note, the single patient who had no residual lymphoma cells detected in any of the four BM samples obtained at the time of marrow harvest had lymphoma cells detectable in 1 of 6 PB samples obtained at that time. This suggests that this patient had residual lymphoma present only at the very limit of detection by PCR analysis.

Association between lymphoma in PB and detection of

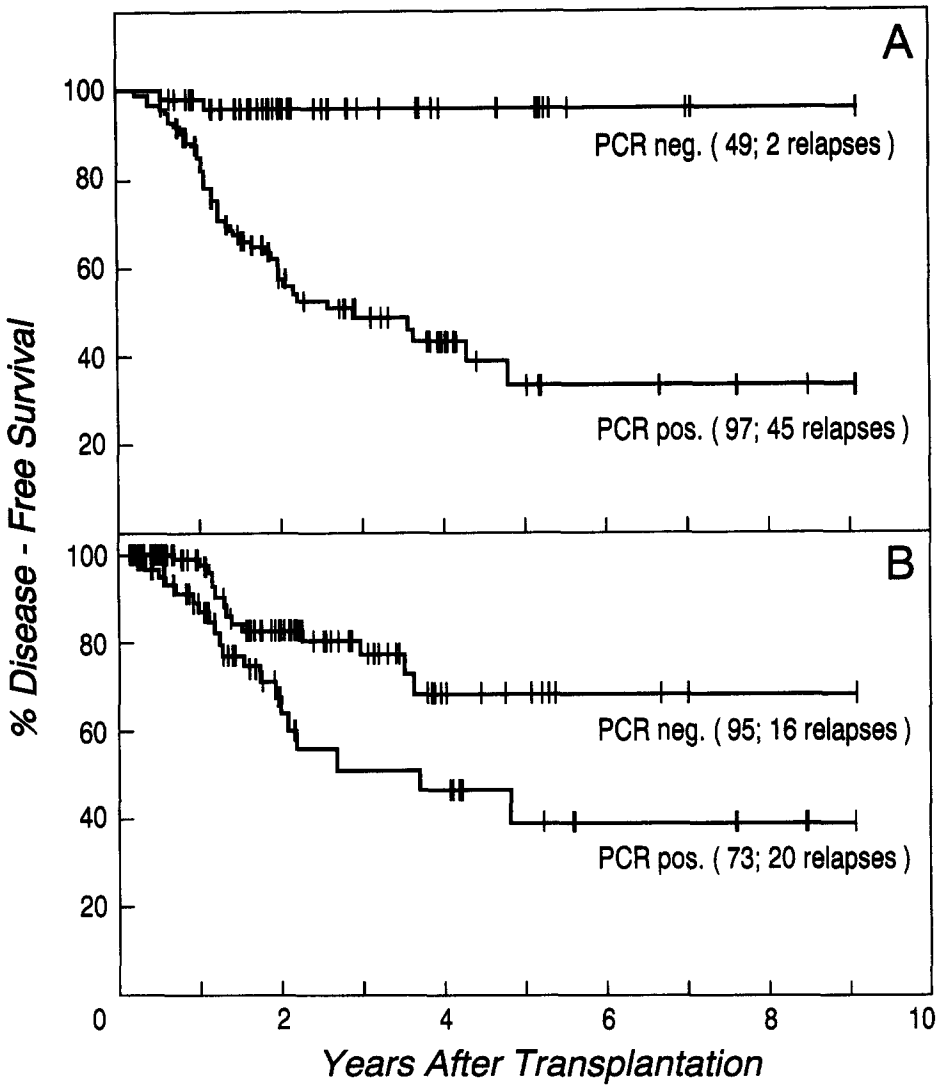


Fig 1. Actuarial probability of disease-free survival after ABMT of patients with a PCR-amplifiable *BCL2*/IgH translocation. PCR neg refers to those patients in whom PCR did not detect any residual lymphoma cells after transplantation. PCR pos refers to those patients in whom PCR-detectable residual lymphoma cells were found at some time after ABMT. (A) Assessment of BM samples from 146 patients; (B) assessment of PB samples from 168 patients.

lymphoma after immunologic purging. After immunologic purging, samples were available for analysis from 202 of the 212 patients. After purging, no lymphoma cells were detectable by PCR in the BM of 91 patients (45%) including the single patient who had no detectable lymphoma cells in the harvested BM before purging. Residual lymphoma cells were detected in the BM after purging in 111 patients (55%). Of the 45 patients in whom PB and BM samples were available for analysis at the time of marrow harvest, those patients who had residual lymphoma cells detected in both the BM and PB at the time of ABMT were more likely to contain residual lymphoma cells after immunologic purging than those patients who had residual lymphoma in the BM but not in the PB (P = .01). As can be seen in Table 2, of 30 patients in whom PCR-detectable lymphoma cells were found after purging, 19 patients (63%) had detectable lymphoma cells in the PB at the time of BM harvest. In contrast, of 15 patients who had no detectable lymphoma cells after purging, 3 patients (20%) had detectable lymphoma cells in PB at the time of BM harvest. Results obtained from two patient samples are shown in Fig 2. Patient A has detectable lymphoma cells in BM but not in PB and has no detectable lymphoma cells after purging. Patient B has detectable lymphoma cells in BM and PB, and has detectable lymphoma cells after purging.

Detection of lymphoma cells in the circulation after infusion of autologous BM. Because we had previously shown by PCR analysis that we were able to eradicate residual lymphoma cells by immunologic purging in only 50% of patients,[15] we were interested to determine whether lymphoma cells could be detected in the patient after the infusion of autologous BM that contained PCR detectable lymphoma cells. For this reason we collected prospectively multiple PB samples from patients 2 hours before and at 2 hours after infusion of purged autologous BM. Samples were analyzed from 60 consecutive patients undergoing ABMT in whom it was documented that the lymphoma contained a PCR-detectable *BCL2*/IgH translocation. A total of 388 PB samples were obtained from these 60 patients, 195 samples obtained before infusion and 193 samples obtained 2 hours

Table 2. PCR Analysis of PB at the Time of BM Harvest

	PCR⁺	PCR⁻
Histology		
Low grade	19	16
Intermediate grade	3	7
Status at ABMT		
CR	6	9
PR	16	14
Morphologic BM involvement in history		
BM involved	14	16
BM not involved	8	7
Morphologic BM involvement at harvest		
BM involved	10	5
BM not involved	12	18
PCR analysis of BM at harvest		
PCR⁺	21	23
PCR⁻	1	0
PCR analysis after immunologic purging		
PCR⁺	19	11
PCR⁻	3	12

after infusion of autologous marrow. These samples were not obtained from the lumen of the Hickman line that was used to infuse the BM.

The results obtained by PCR analysis of PB samples obtained before and after infusion of marrow are shown in Fig 3. There is an obvious difference in the detection of lymphoma cells in the circulation after infusion of autologous BM ($P < .0001$ by McNemar's test). After high-dose therapy, but before infusion of autologous BM, lymphoma cells were detected by PCR analysis in the PB samples of only 8 patients (13%). This suggests that high-dose therapy is relatively effective in clearing detectable lymphoma cells from the majority of patients. In contrast, 2 hours after infusion of autologous marrow, PCR-detectable lymphoma cells were detected in the PB samples from 35 patients (58%). In 29 of 35 patients (83%), there was no evidence of circulating lymphoma cells before infusion of the autologous BM. Six patients had detectable lymphoma cells before and after infusion of BM. The reason 2 patients had detectable lymphoma cells in the PB before infusion of BM that were not detectable later that day is not clear, but these may represent patients in whom the lymphoma cells were in the process of being cleared from the circulation after high-dose therapy.

Lymphoma cells infused with autologous marrow are detectable in the PB. In these 60 patients in whom PB samples were collected after infusion of autologous marrow, samples before and after immunologic purging were available for PCR analysis from 56 (93%) patients. All 56 patients had PCR-detectable lymphoma cells in their harvested BM. These 60 patients appeared representative of the entire patient group except for the results obtained after immunologic purging. Despite the fact that immunologic purging is capable of depleting PCR-detectable lymphoma cells from almost 50% of the total population studied, no residual lymphoma cells were detectable in the BM after purging in only 18 (32%) of these patients ($P = .03$). The results of PCR analysis of PB samples obtained before and after infusion of marrow that could be purged of all PCR-detectable lymphoma cells are shown in Fig 4. Before infusion of marrow, lymphoma cells were detected in the PB of 8 patients, 7 of whom had PCR-detectable lymphoma cells in their BM after purging. After infusion of marrow that contained no detectable lymphoma cells, only 1 of 18 patients had detectable lymphoma cells in the PB, and this patient had lymphoma cells detected in the circulation before infusion. In contrast, 2 hours after infusion of marrow that contained residual PCR-detectable lymphoma cells after purging, 32 of the 38 patients (84%) had circulating lymphoma cells detected in the PB after infusion. There was a significant association between the detection of residual lymphoma cells in the BM after purging and the subsequent detection of lymphoma cells in the PB after infusion of marrow ($P = .0001$). The appearance of detectable lymphoma cells in the PB of these patients by 2 hours after infusion of autologous BM that contains PCR-detectable lymphoma cells suggests strongly that these lymphoma cells are infused with the autologous BM. In addition, because DNA can be extracted from lymphoma cells in the circulation, this also suggests that PCR analysis is detecting residual lymphoma cells after immunologic purging and not simply detecting residual DNA containing the translocation that has been released from lysed cells. However, definitive proof that the lymphoma cells appearing in the circulation represent those same lymphoma cells that were infused within the autologous BM will require gene-marking studies. These studies are currently being undertaken in a number of centers.

Lymphoma cells continue to circulate during the early posttransplant period. During the immediate posttransplant period, while the patients remained hospitalized, a total

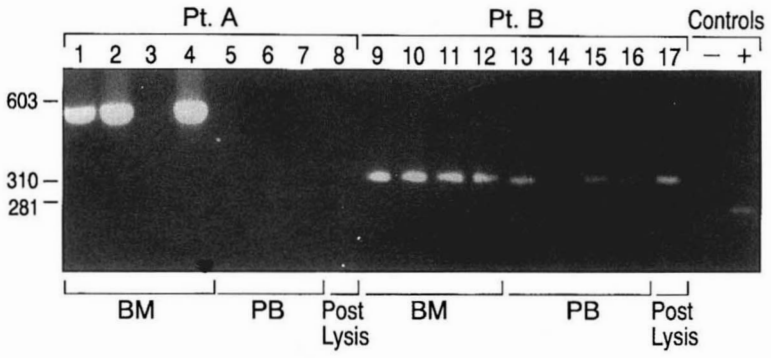


Fig 2. Ethidium bromide-stained agarose gel electrophoresis of PCR-amplified samples obtained at the time of BM harvest from two representative patients. Prelysis and postlysis samples obtained before and after complement mediated purging. Controls: ⁻PCR amplification of DNA from a cell line that does not contain a PCR-amplifiable *BCL2*/IgH translocation; ⁺PCR amplification of DNA from a 10 to 5 dilution of the cell line RL, that contains a *BCL2*/IgH translocation in normal BM mononuclear cells.

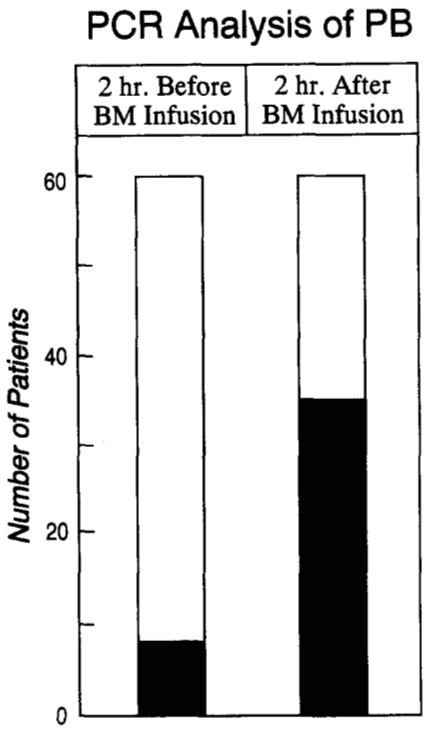


Fig 3. Results of PCR amplification at either the MBR or mcr of the *BCL2*/IgH translocation of PB samples obtained immediately before and at 2 hours after infusion of harvested autologous BM. (□) PCR-negative; (■) PCR-positive.

of 637 samples were obtained from 60 patients from day 1 to day 26 after infusion of autologous BM. PCR-amplifiable DNA could be obtained from only 38 patients' samples during days 1 to 6 because of the paucity of cells obtained from the PB during this period, from 57 patients during days 7 to 13, and from 53 patients during days 14 to 20. Samples were available for analysis from only 27 patients during days 21 to 27 as the remaining patients had already been discharged from hospital admission by this time. During each of these time periods, there was a statistically significant association between the detection of residual lymphoma cells after immunologic purging and the detection of circulating lymphoma cells in the PB ($P < .0001$ for each of the periods from days 1 to 20 and $P = .0002$ for days 21 to 27). Results obtained on representative samples are shown in Fig 5. Patients A and B have PCR-detectable lymphoma cells remaining after immunologic purging. Lymphoma cells were also detected in the PB at 2 hours after infusion of autologous BM. These lymphoma cells remained detectable in PB throughout the entire hospital admission for ABMT. In contrast, patient C has no PCR-detectable lymphoma after immunologic purging and no lymphoma cells were detected by PCR analysis of PB samples obtained during admission. PB samples contained either no PCR-detectable lymphoma cells in any sample or contained PCR-detectable lymphoma cells at all time periods in all but 12 of the patients. In these 12 patients, PCR analysis detected lymphoma cells at some, but not all time periods. No discernible pattern was apparent on detailed analysis of the samples analyzed from this small group of patients. These findings suggest that once lymphoma cells in the autologous marrow are infused into the patient, they appear in the circulation and continue to circulate for extended periods. This is in keeping with our previous observation that patients who were infused with marrow that contained residual lymphoma after purging had detectable lymphoma cells in the circulation at 6 months after ABMT.[15]

DISCUSSION

Although residual lymphoma cells have been detected by PCR in both BM and PB of patients who are in complete clinical remission, it is presently unknown whether these BM and PB are equivalent tissue sources to detect residual lymphoma. The present study compared the sensitivity and clinical utility of the detection of residual lymphoma cells by PCR amplification of the *BCL2*/IgH translocation in both BM and PB of patients with advanced stage B-cell non-Hodgkin's lymphoma (NHL) before, at the time of, and after ABMT. Those patients who had no residual lymphoma cells detectable in BM samples obtained after ABMT had excellent prognosis. The presence of residual lymphoma cells in the PB in the present study was also associated with a statistically significant decrease in disease-free survival after ABMT. However, 44% of patients who subsequently relapsed had no detectable lymphoma cells in their PB. We conclude that analysis of BM is a far more informative and

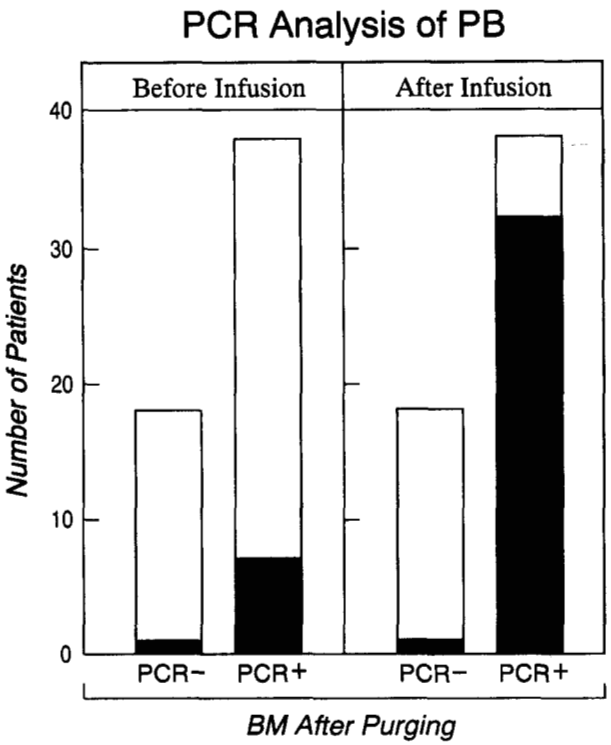


Fig 4. Results of PCR amplification at either the MBR or mcr of the *BCL2*/IgH translocation of PB samples obtained immediately before and at 2 hours after infusion of harvested autologous BM stratified by the result of PCR performed after immunologic purging of the harvested BM. (□) PCR-negative; (■) PCR-positive.

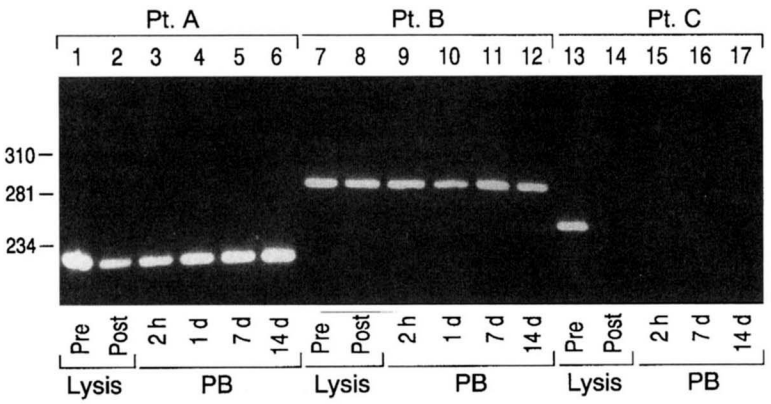


Fig 5. Ethidium bromide-stained agarose gel electrophoresis of PCR-amplified samples of three representative patients. Prelysis and postlysis samples refer to the harvested BM before and after immunologic purging. PB samples were obtained at 2 hours, 1 day, 7 days, and 14 days after infusion of BM.

sensitive detector of minimal residual disease than analysis of PB, and, more importantly, is significantly more predictive of which patients will relapse after ABMT.

A major concern with the use of ABMT has always been the fear that malignant cells harbored within the BM will be infused to the patient and contribute to subsequent relapse. Recently, three independent lines of evidence have suggested that the infusion of malignant cells within the autologous BM may contribute to subsequent relapse. First, Sharp and his colleagues[5] at the University of Nebraska have shown that the detection of clonogenic lymphoma cells within autologous BM was associated with an increased rate of relapse. Second, the infusion of PCR-detectable lymphoma cells after immunologic purging was also associated with an increased probability of relapse after ABMT.[15] Most recently, gene-marker studies have shown the presence of a marker gene at the time of relapse, suggesting that these infused marrow cells contributed to subsequent relapse.[31] In the present report, we show that when residual lymphoma cells are infused into the patient, these cells are rapidly detectable circulating in the PB in the patient. These cells continue to circulate during the immediate posttransplant period and are detectable in the PB in the majority of patients who are infused with marrow containing residual lymphoma. If lymphoma cells infused within the autologous marrow do indeed contribute to subsequent relapse, then it would seem likely that patients would relapse in a disseminated manner. However, after ABMT, patients relapse at the site of prior disease. A potential reason why patients subsequently relapse at the sites of prior disease might be that the infused lymphoma cells home back to previously involved lymph nodes.[32]

PCR analysis of BM at the time of harvest showed residual lymphoma cells in 211 of 212 patients studied (99%). However, analysis of PB samples obtained at the time of BM harvest from 45 of these patients showed circulating lymphoma cells in only 22 patients (49%). The finding that the PB is less likely than BM to have residual detectable lymphoma cells at the time of BM harvest might well suggest that PB is a less contaminated source of autologous hematopoietic stem cells than BM. However, because 49% of patients in the present study had residual lymphoma cells in PB samples, this suggests that PB stem cell (PBSC) collections from a considerable number of patients with NHL with the *BCL2*/IgH translocation are likely to be contaminated with lymphoma. These results are in keeping with a previously published study that showed that PB stem cell harvests were contaminated with lymphoma in up to 50% of the patients studied.[33] This number is comparable with the results obtained using immunologic purging of autologous BM.[15] In addition, it is presently unclear whether the use of chemotherapy and growth factors that are currently used to mobilize PBSC might also mobilize tumor cells. If the present results are confirmed in larger studies, then it will not be clear whether a patient would benefit more from the use of PBSC collection or purged autologous BM. To answer this question definitively, clinical trials comparing PBSC versus purged autologous BM would be required. Ongoing clinical studies in patients undergoing high-dose therapy with PBSC support can address whether the detection of residual lymphoma cells in PBSC collections will have the same adverse effect on subsequent outcome than the infusion of residual lymphoma cells in autologous BM. The results of such studies might help to determine the optimal tissue source as stem cell support for individual patients undergoing ABMT.

In conclusion, the results of concomitant assessment of BM and PB after ABMT show that serial BM samples are necessary to predict best for subsequent relapse. Moreover, these data are consistent with the notion that post-ABMT contamination of PB appears to be the consequence of reinfusion of lymphoma cells within autologous marrow and that to improve disease-free survival after ABMT, attention should be given to remove contaminating PCR-detectable lymphoma cells from either harvested autologous BM or PBSC collections. Furthermore, future studies will be required to address the issue of whether treatment attempting to deplete PCR-detectable lymphoma cells after ABMT will also result in improved survival after ABMT.

REFERENCES

1. Ault KA: Detection of small numbers of monoclonal B lymphocytes in the blood of patients with B cell lymphoma. N Engl J Med 300:1401, 1979

2. Hu E, Trela M, Thompson J, Lowder J, Horning S, Levy R, Sklar J: Detection of B cell lymphoma in peripheral blood by DNA hybridization. Lancet 2:1092, 1985

3. Horning SJ, Galila N, Cleary M, Sklar J: Detection of non-Hodgkin's lymphoma in the peripheral blood by analysis of the antigen receptor gene rearrangements: Results of a prospective trial. Blood 75:1139, 1990

4. Benjamin D, Magrath iT, Douglass EC, Corash LM: Derivation of lymphoma cell lines from microscopically normal bone marrow in patients with undifferentiated lymphoma: Evidence of occult bone marrow involvement. Blood 61:1017, 1983

5. Sharp JG, Joshi SS, Armitage JO, Bierman P, Coccia PF, Harrington DS, Kessinger A, Crouse DA, Mann SL, Weisenberger DD: Significance of detection of occult non-Hodgkin's lymphoma in histologically uninvolved bone marrow by culture technique. Blood 79:1074, 1992

6. Crescenzi M, Seto M, Herzig GP, Weiss PD, Griffith RC, Korsmeyer SJ: Thermostable DNA polymerase chain amplification of t(14;18) chromosome breakpoints and detection of minimal residual disease. Proc Natl Acad Sci USA 85:4869, 1988

7. Lee MS, Chang KS, Cabanillas F, Freireich EJ, Trujillo JM, Stass SA: Detection of minimal residual disease carrying the t(14;18) by DNA sequence amplification. Science 237:175, 1987

8. Yamada M, Hudson S, Tourney O, Birttenbender S, Shane SS, Lange B, Tsujimoto Y, Caton AJ, Rovera G: Detection of minimal disease in hematopoietic malignancies of the B-cell lineage by using third complementarity-determining region probes. Proc Natl Acad Sci USA 86:5123, 1989

9. Tsujimoto Y, Finger LR, Yunis J, Norwell PC, Croce CM: Cloning of the chromosome breakpoint of neoplastic B cells with the t(14;18) chromosome translocation. Science 226:1097, 1984

10. Bakshi A, Jensen JP, Goldman P, Wright JJ, McBride OW, Epstein AL, Korsmeyer SJ: Cloning the chromosomal breakpoint of t(14;18) human lymphomas: Clustering around J_H on chromosome 14 and near a transcriptional unit on 18. Cell 41:899, 1985

11. Cleary ML, Sklar J: Nucleotide sequence of a t(14;18) chromosomal breakpoint in follicular lymphoma and demonstration of a breakpoint cluster region near a transcriptionally active locus on chromosome 18. Proc Natl Acad Sci USA 81:593, 1985

12. Yunis JJ, Oken MM, Kaplan ME, Theologides RR, Howe A: Distinctive chromosomal abnormalities in histological subtypes of non-Hodgkin's lymphoma. N Engl J Med 307:1231, 1982

13. Weiss LM, Warnke RA, Sklar J, Cleary ML: Molecular analysis of the t(14;18) chromosomal translocation in malignant lymphomas. N Engl J Med 317:1185, 1987

14. Lee MS, Blick MB, Pathak S, Trujillo JM, Butler JJ, Katz RL, McLaughlin P, Hagemeister FB, Velasquez WS, Goodacre A, Cork A, Gutterman JU, Cabinillas F: The gene located at chromosome 18 band q21 is rearranged in uncultured diffuse lymphomas as well as follicular lymphomas. Blood 70:90, 1987

15. Gribben JG, Freedman AS, Neuberg D, Roy DC, Blake KW, Woo SD, Grossbard ML, Rabinowe SN, Coral F, Freeman GJ, et al: Immunologic purging of marrow assessed by PCR before autologous bone marrow transplantation for B-cell lymphoma. N Engl J Med 325:1525, 1991

16. Lambrechts AC, de Ruiter PE, Dorssers LC, van't Veer MB: Detection of residual disease in translocation (14;18) positive non-Hodgkin's lymphoma, using the polymerase chain reaction: A comparison with conventional staging methods. Leukemia 6:29, 1992

17. Yuan R, Dowling P, Zucca E, Diggelmann H, Cavalli F: Detection of bcl-2/JH rearrangement in follicular and diffuse lymphoma: Concordant results of peripheral blood and bone marrow analysis at diagnosis. Br J Cancer 67:922, 1993

18. Berinstein NL, Reis MD, Ngan BY, Sawka CA, Jamal HH, Kuzniar B: Detection of occult lymphoma in the peripheral blood and bone marrow of patients with untreated early stage and advanced stage follicular lymphoma. J Clin Oncol 11:1344, 1993

19. Gribben JG, Freedman AS, Woo SD, Blake K, Shu RS, Freeman G, Longtine JA, Pinkus GS, Nadler LM: All advanced stage non-Hodgkin's lymphomas with a polymerase chain reaction amplifiable breakpoint of bcl-2 have residual cells containing the bcl-2 rearrangement at evaluation and after treatment. Blood 78:3275, 1991

20. Hickish TF, Purvies H, Mansi J, Soukop M, Cunningham D: Molecular monitoring of low grade non-Hodgkin's lymphoma by gene amplification. Br J Cancer 64:1161, 1991

21. Price CGA, Meerabux J, Murtagh S, Cotter FE, Rohatiner AZS, Young BD, Lister TA: The significance of circulating cells carrying t(14;18) in long remission from follicular lymphoma. J Clin Oncol 9:1527, 1991

22. Finke J, Slanina J, Lange W, Dolkenm G: Persistence of circulating t(14;18)-positive cells in long-term remission after radiation therapy for localized-stage follicular lymphoma. J Clin Oncol 11:1668, 1993

23. Gribben JG, Neuberg D, Freedman AS, Gimmi CD, Pesek KW, Barber M, Saporito L, Woo SD, Coral F, Spector N, Rabinowe SN, Grossbard M, Ritz J, Nadler LM: Detection by polymerase chain reaction of residual cells with the bcl-2 translocation is associated with increased risk of relapse after autologous bone marrow transplantation for B-cell lymphoma. Blood 81:3449, 1993

24. Higuchi R: Simple and rapid preparation of samples for PCR, in Erlich HA (ed): PCR Technology. New York, NY, Stockton, 1989 p 31

25. Cleary ML, Galili N, Sklar J: Detection of a second t(14;18) breakpoint cluster region in human follicular lymphomas. J Exp Med 164:315, 1986

26. Kwok S, Higuchi R: Avoiding false positives with PCR. Nature 339:237, 1989

27. Kaplan EL, Meier P: Nonparametric estimation from incomplete observations. J Am Stat Assoc 53:457, 1958

28. Cox DR, Oakes D: Analysis of Survival Data. New York, NY, Chapman and Hall, 1984, p 124

29. Rosner B: Fundamentals of Biostatistics. Boston, MA, PWS-Kent, 1990, p 336

30. Rosner B: Fundamentals of Biostatistics. Boston, MA, PWS-Kent, 1990, p 342

31. Brenner MK, Rill DR, Moen RC, Krance RA, Mirro J, Anderson WF, Ihle JN: Gene-marking to trace origin of relapse after autologous bone-marrow transplantation. Lancet 341:85, 1993

32. Freedman AS, Saporito L, Rhynhart K, Morimoto C, Nadler LM: Adhesion of follicular lymphoma cells to lymphoid germinal centers—A potential mechanism of tumor cell homing following autologous transplantation. Leuk Lymph (in press)

33. Hardingham JE, Kotasek D, Sage RE, Dobrovic A, Gooley T, Dale BM: Molecular detection of residual lymphoma cells in peripheral blood stem cell harvests and following autologous transplantation. Bone Marrow Transplant 11:15, 1993

Detection of Circulating Tumor Cells in Colorectal Cancer by Immunobead-PCR Is a Sensitive Prognostic Marker for Relapse of Disease

Jennifer E. Hardingham,*† Dusan Kotasek,* Robert E. Sage,*‖ Michael C. Eaton,‡ Vivienne H. Pascoe,§ and Alexander Dobrovic*‖

Departments of *Hematology-Oncology, ‡Surgery, and §Histopathology, The Queen Elizabeth Hospital, Woodville, South Australia
†Departments of Physiology and ‖Medicine, University of Adelaide, Adelaide, South Australia

ABSTRACT

Background: Recurrent and metastatic carcinoma of the colorectum remains a major problem, with survival at 5 years post curative resection still only about 50%. Moreover, up to 30% of patients who present with early stage disease also relapse and die within 5 years, suggesting the presence of micrometastatic disease at diagnosis. One route of metastatic spread is via the blood stream, hence the detection of tumor cells in blood is likely to provide an important predictive tool with respect to relapse of disease. We have developed a sensitive molecular technique to identify tumor cells in blood using mutations in codon 12 of the K-*ras* gene as a marker.
Materials and Methods: Twenty-seven patients whose tumor carried a mutation in codon 12 of K-*ras* were studied for the presence of tumor cells in perioperative peripheral blood samples. Immunomagnetic beads, labeled with an epithelial-specific antibody, were used to harvest epithelial cells from blood. K-*ras* mutations were identified in this selected population using a polymerase chain reaction (PCR)–based analysis (immunobead-PCR).
Results: Circulating K-*ras* mutant cells were detected in 9 of 27 patients; seven of these nine patients have since died due to recurrent or metastatic disease. Mutant cells were not detected in 18 patients, and 16 of 18 have remained disease free (median follow-up: 16 months; range: 7–42 months). Kaplan-Meier analysis showed that detection of K-*ras* mutant cells in blood was associated with significantly reduced disease-free survival ($p = 0.0001$).
Conclusion: This study indicates that detection of circulating tumor cells perioperatively by immunobead-PCR provides a sensitive prognostic marker for recurrent and metastatic colorectal cancer.

INTRODUCTION

Despite recent advances in surgical and diagnostic techniques, a significant number of patients with early stage colorectal cancer who undergo "curative" resection, still suffer recurrent and metastatic disease and die within 5 years (1). The inference is that occult metastatic disease was already present at diagnosis, or that tumor cells with metastatic potential were released into the blood at the time of surgery. Detection of circulating tumor cells preoperatively in these "good prognosis" patients would imply that the disease is more advanced than indicated by conventional

Address correspondence and reprint requests to: A. Dobrovic, Hematology-Oncology Department, The Queen Elizabeth Hospital, 28 Woodville Road, Woodville, South Australia 5011.

Copyright © 1995, *Molecular Medicine*, 1076-1551/95/$10.50/0

Exhibit N

staging. These patients might then also be offered adjuvant chemotherapy or radiotherapy post surgery.

Mutations in codon 12 of the K-*ras* gene have been identified as an early event in the development of up to 50% of colorectal tumors (2). Identification of this mutation in a tumor thus provides a marker for tumor cells that may be present in the blood. We have developed a sensitive method, immunobead–polymerase chain reaction (PCR), for the detection of tumor cells bearing this mutation in blood (3) and have used this technique to address the question of the tumorigenic potential of K-*ras*–positive cells in the peripheral circulation and tumor cell release during surgical resection.

The process of metastasis involves sequential steps of tumor cell detachment from the primary tumor, adherence to and proteolysis of endothelial basement membranes, intravasation into the circulation, and initiation of secondary tumors in distal organs. Most of the tumor cells that enter the circulation die, and only a very small proportion succeed in establishing metastases (4,5). The aims of this study were to detect the presence of circulating tumor cells perioperatively and to determine the relationship of tumor cells in blood to disease-free survival following curative resection for colorectal cancer.

MATERIALS AND METHODS

Patients and Samples

Twenty-seven consecutive patients whose colorectal tumor was positive for a K-*ras* codon 12 mutation were included in the study. The disease was staged according to the Dukes' classification (Astler Coller modified) (1). The median age was 74 years (range: 56–91) with 13 males and 14 females. The study was approved by the institutional Ethics of Research Committee and informed consent was obtained from each patient. Tumor specimens were obtained either fresh, or from formalin fixed, paraffin-embedded sections. Peripheral venous blood samples (20 ml) were collected before surgery from 24 cases, and post surgery from 26 cases (immediately after mobilization of the tumor in 13 cases and the following morning in 24 cases). Normal controls consisted of 20-ml blood samples from 11 nonsurgical patients without cancer and from two patients undergoing colonic surgery for benign disease.

Immunobead-PCR Analysis

Tumor samples were analyzed for K-*ras* codon 12 mutations by PCR as previously described (3). Following digestion of PCR products with BstN1 (New England Biolabs, Beverly, MA, U.S.A.), the restriction enzyme fragments were run on a 10% polyacrylamide (19:1 acrylamide:*bis*acrylamide) gel and visualized by ethidium bromide or silver staining (6).

To detect tumor cells in blood, the immunobead-PCR technique was performed (3). Firstly, epithelial cells were isolated from blood using immunomagnetic beads (Dynal, Oslo, Norway) labeled with the epithelial-specific antibody Ber-EP4 (Dakopatts, Gestrop, Denmark). This antibody was found to react strongly with all 27 tumor samples, either by flow cytometry or immunohistochemistry. Secondly, the PCR analysis was performed as described above, on the isolated epithelial cell/bead aggregates to determine the presence of a K-*ras* mutation. The sensitivity of detection of K-*ras* mutations using immunobead-PCR and silver staining was in the order of 1 tumor cell/10^6 white blood cells (i.e., 10 tumor cells/ml of whole blood [data not shown]). The study was performed in a blinded fashion such that the patients' clinical details and histopathology were unknown to the investigator performing immunobead-PCR on the blood samples. A univariate Kaplan-Meier analysis (BMDP Statistical Software Inc., Los Angeles, CA, U.S.A.) was used to assess the significance of circulating K-*ras* mutant cells with respect to disease-free survival in our patient cohort.

RESULTS

K-*ras* mutant cells were detected in perioperative peripheral blood samples in 9 of 27 (33%) patients (Table 1). Results from blood taken post mobilization of the bowel and day 1 post surgery were combined since no additional information was obtained from those 11 cases (three positive, eight negative) in which we tested both samples. In six of nine patients, mutant cells were detected in blood samples before surgery indicating that tumor dissemination had already occurred. Five of the six patients (Table 1, Nos. 17, 18, 20, 21, and 25) have relapsed and died from recurrent or metastatic disease at a median of 14 months (range: 7–23 months) post surgery. The other patient (No. 5) had been staged as B1 (tumor contained within the muscularis propria

TABLE 1. Patient characteristics and detection of tumor cells in blood

Patient No.	Patient ID	Age	Sex	Tumor Cells		Stage	Relapse (months)	Survival (months)	Status
				Pre-Op	Post-Op				
1	535	73	M	Neg	Neg	A	Nil	23.5	Alive
2	425	56	F	Neg	Neg	A	Nil	23	Alive
3	169	64	M	Neg	Pos	A	Nil	17.5	Alive
4	173	79	M	Neg	Neg	B1	Nil	38	Alive
5	304	68	F	Pos	Pos	B1	Nil	15	Alive
6	954	82	F	Neg	Neg	B1	Nil	28	Alive
7	711	77	M	Neg	Neg	B1	Nil	15	Alive
8	647	65	M	Neg	Neg	B1	Nil	15	Alive
9	177	88	F	Neg	Neg	B1	Nil	16	Dec[a]
10	966	80	M	NA	Neg	B2	Nil	42	Alive
11	988	69	M	Neg	Neg	B2	Nil	14	Alive
12	703	72	M	Neg	Neg	B2	Nil	16	Alive
13	356	60	F	NA	Pos	B2	7 m	10	Dec
14	712	91	F	Neg	Neg	B2	Nil	37.5	Alive
15	831	76	M	Neg	Neg	B2	Nil	29	Alive
16	433	76	F	Neg	Neg	B2	Nil	30	Alive
17	840	80	F	Pos	Pos	B2	4 m	7	Dec
18	352	82	M	Pos	NA	B2	22 r	23	Dec
19	925	79	F	Neg	Neg	C1	Nil	12.5	Alive
20	606	59	M	Pos	Pos	C1	14.5 r,m	18	Dec
21	477	68	F	Pos	Pos	C1	4 m	10	Dec
22	748	75	F	Neg	Pos	C1	5 r	9.5	Dec
23	43	73	M	Neg	Neg	C1	Nil	32	Dec[a]
24	895	71	F	Neg	Neg	C2	Nil	33	Alive
25	412	63	F	Pos	Pos	D	1 m	7	Dec
26	195	83	F	Neg	Neg	D	6 r,m	8	Dec
27	625	76	M	NA	Neg	D	7 r	7	Dec

r, local recurrence; m, metastasis; NA, not available; Dec, deceased.
[a]Died of unrelated causes.

and no nodal metastases) and is presently disease free at 15 months. Tumor cells were also present in postoperative blood in all but one of these six patients (in this patient we did not receive a postsurgery specimen). K-*ras* positive cells were not detected in any of the normal control samples.

There were two cases in which tumor cells were detected in the postoperative but not the preoperative blood specimen, suggesting that tumor cells were shed into the circulation during surgical resection. One of these patients (No. 3) is alive at 17.5 months post surgery; the other (No. 22) suffered a local recurrence and died 9.5 months post surgery. In the remaining case (No. 13), tumor cells were found in postoperative blood, but we did not receive a presurgery blood sample; this patient, who was staged as B2, developed a metastatic deposit in the bladder and died 10 months post surgery.

Overall, 9 of 27 patients subsequently relapsed with either local recurrence or metastatic disease or both, and seven of the nine had circulating tumor cells detected in perioperative blood

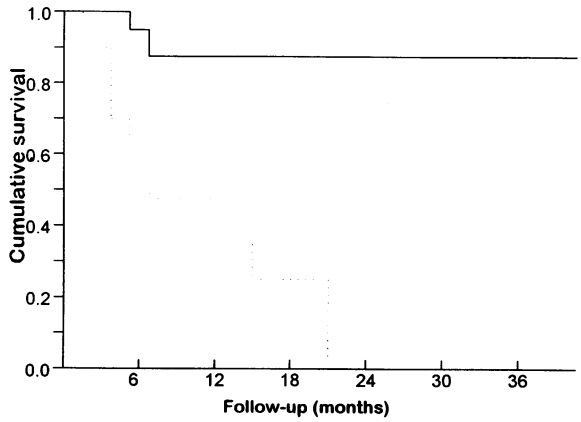


FIG. 1. Kaplan-Meier disease-free survival analysis

Solid line, patients in whom tumor cells were not detected; dashed line, patients in whom tumor cells were detected, $p = 0.0001$.

specimens. Of 18 patients who have remained disease free post surgery, 16 of 18 were negative for circulating tumor cells at the time of surgery (median follow-up: 24 months; range: 12.5–42 months). Kaplan-Meier analysis (Fig. 1) showed that disease-free survival was significantly reduced ($p = 0.0001$) in patients in whom circulating tumor cells were detected (mean survival time: 11.2 versus 30.4 months; SE 2.88 months).

DISCUSSION

Tumor cells were detected in perioperative peripheral blood samples in a third of the patients, and this finding showed a strong association with shortened disease-free survival (Fig. 1). Due to the small number of patients in this study, we were not able to show whether or not detection of tumor cells in blood was independent of stage as a prognostic indicator.

Intravasation of viable tumorigenic cells into the blood stream is one step in the metastatic cascade leading to dissemination of the primary tumor. The presence of tumor cells in the circulation before surgery suggests that an active metastatic process is involved, and five out of six patients succumbed to recurrent or metastatic disease within 2 years of surgery. The presence of tumor cells in the circulation in two patients appeared to be due to direct release of tumor cells during surgical resection, as they were negative for tumor cells in blood before surgery but positive post surgery. One of these patients (Table 1, No. 22) suffered a local recurrence which may have arisen from a preexisting subclinical adenoma, or may have occurred as a result of "reseeding" of the mucosa by blood-borne tumorigenic cells. The other patient (Table 1, no. 3) has remained disease-free at 17.5 months post surgery, so perhaps tumor cells that gain direct access into the circulation may not necessarily have metastatic potential; Tarin et al. (7) demonstrated the absence of metastatic disease in some patients after months of continuous infusion of malignant ascites fluid into the bloodstream.

Previously, conventional cytology and immunocytochemistry have been used to detect tumor cells in blood. An immunofluorescence method, using the anti-cytokeratin antibody AE3, detected carcinoma cells in blood from the renal vein in 8 of 10 patients undergoing nephrectomy for renal carcinoma; three of eight patients had clinically detectable metastases prior to the assay, while five of eight have remained disease free for up to 35 months (8), suggesting that most of the tumor cells released into the renal vein were arrested and destroyed in the pulmonary microvasculature. Another study, using both cytology and immunocytochemistry, detected circulating colorectal cancer cells in 4 of 42 patients, including 2 of 4 with Dukes' B disease (9). The low detection rate was attributed to the fact that strict morphological criteria was applied in determination of malignant cells. Immunocytochemistry for an epithelial marker, cytokeratin 18, was used by Lindemann et al. (10) to determine the prognostic significance of tumor cells in bone marrow taken preoperatively from patients with colorectal cancer, 32% of the patients were positive and these patients had a significantly reduced disease-free survival.

More recently, the molecular techniques of reverse transcription and PCR have been used to detect very small numbers of tumor cells in blood in a range of neoplasms (11–15); however, results from these studies were not correlated with clinical outcome after long-term follow-up. Brown et al. (15) detected breast carcinoma cells in preoperative blood in of nine patients, in intraoperative blood in four of nine, but in zero of nine in blood collected 24 hr post surgery, suggesting a proportion of tumors shed cells as a result of surgery and that they were removed from the circulation within 24 hr. This is in contrast to our findings, where eight patients were

positive for tumor cells in peripheral blood taken 16–24 hr post surgery.

The biological significance of very small numbers of tumor cells in the circulation remains to be elucidated. A study by Mayhem and Glaves (16) using a murine model of metastasis for Lewis lung carcinoma determined that 1000 Lewis carcinoma cells were needed to generate at least one pulmonary metastasis. At the limit of detection of tumor cells using our assay (i.e., 10 tumor cells/ml of blood), some 50,000 tumor cells may be circulating (based on an average blood volume of 5 l) at a given time point. If similar tumor kinetics can be applied to human carcinoma, then this implies that 50,000 circulating tumor cells would provide sufficient tumorigenic cells to enable metastases to form. Taking into account the inefficiency of the metastatic process such that less than 0.01% of these cells will survive (17,18), there is still the potential for some five metastatic deposits to develop. The observed frequency is even less in our patient group, with only two patients developing secondary deposits at more than one site, although multiple seeding at the same site cannot be excluded.

In two patients with stage D (metastatic) disease we did not find circulating tumor cells prior to surgery and this may reflect the sporadic release of tumor cells from the primary (or secondary) tumor(s) into the blood (19). Hence, samples taken at any one time will necessarily contain variable numbers of cancer cells perhaps below the threshold of detection, depending on the rate of shedding at the time.

Our study has identified circulating tumor cells at the time of surgery in five patients whose tumors did not show any evidence of metastasis by conventional staging. These patients would not have been offered adjuvant therapy since this is currently restricted to patients with more advanced disease. Perhaps patients with circulating micrometastatic disease would benefit most from adjuvant chemotherapy or radiotherapy, or systemic treatment with monoclonal antibodies such as 17-1A, which in Dukes' C patients has been shown to reduce the overall death rate by 30% and the recurrence rate by 27% (20).

In conclusion, this preliminary study shows that K-*ras* immunobead-PCR is a sensitive and specific technique for identifying tumor cells in blood. The correlation between the presence of tumor cells in blood and shortened disease-free survival shows that the technique is useful for identifying colorectal patients at diagnosis who are at increased risk of relapse. This should result in a more appropriate selection of patients for further treatment and so lead to a reduction in morbidity and mortality in these patients.

ACKNOWLEDGMENTS

This work was supported by a grant from the Anticancer Foundation of the Universities of South Australia. The cooperation of the Department of Surgery, in particular Assoc. Professor W. E. W. Roediger, was greatly appreciated. We thank Mrs. Jing-Xian Mi for excellent technical assistance.

REFERENCES

1. Cohen AM, Shank B, Friedman MA. (1989) Colorectal cancer. In: Devita VT, Hellman S, Rosenberg SA (eds). *Cancer: Principles and Practice of Oncology. 3rd ed.* Lippincott, Philadelphia, pp. 895–952.
2. Bos JL. (1989) *ras* oncogenes in human cancer: A review. *Cancer Res.* **49:** 4682–4689.
3. Hardingham JE, Kotasek D, Farmer B, et al. (1993) Immunobead-PCR: A technique for the detection of circulating tumor cells using immunomagnetic beads and the polymerase chain reaction. *Cancer Res.* **53:** 3455–3458.
4. Fidler IJ. (1991) Cancer metastasis. *Br. Med. Bull.* **47:** 157–177.
5. Liotta LA, Stetler-Stevenson WG. (1991) Tumor invasion and metastasis: An imbalance of positive and negative regulation. *Cancer Res.* **51**(Suppl): 5054s–5059s.
6. Bassam BJ, Caetano-Anolles G. (1993) Silver staining of DNA in polyacrylamide gels. *Appl. Biochem. Biotechnol.* **42:** 181–188.
7. Tarin D, Price JE, Kettlewell MGW, Souter RG, Vass ACR, Grossley B. (1984) Mechanisms of human tumor metastases studied in patients with peritoneovenous shunts. *Cancer Res.* **44:** 3584–3592.
8. Glaves D, Huben RP, Weiss L. (1988) Haematogenous dissemination of cells from human renal adenocarcinomas. *Br. J. Cancer* **57:** 32–35.
9. Leather AJM, Gallegos NC, Kocjan G, et al. (1993) Detection and enumeration of circulating tumour cells in colorectal cancer. *Br. J. Surg.* **80:** 777–780.
10. Lindemann F, Schlimok G, Dirschedl P, Witte J, Riethmuller G. (1992) Prognostic

significance of micrometastatic tumour cells in bone marrow of colorectal cancer patients. *Lancet* **340:** 685–689.
11. Smith B, Selby P, Southgate J, Pittman K, Bradley C, Blair GE. (1991) Detection of melanoma cells in peripheral blood by means of reverse transcriptase and polymerase chain reaction. *Lancet* **338:** 1227–1229.
12. Mattano LA, Moss TJ, Emerson SG. (1992) Sensitive detection of rare circulating neuroblastoma cells by the reverse transcriptase-polymerase chain reaction. *Cancer Res.* **52:** 4701–4705.
13. Seiden MV, Kantoff PW, Krithivas K, et al. (1994) Detection of circulating tumor cells in men with localized prostate cancer. *J. Clin. Oncol.* **12:** 2634–2639.
14. Datta YH, Adams PT, Drobyski WR, Ethier SP, Terry VH, Roth MS. (1994) Sensitive detection of occult breast cancer by the reverse-transcriptase polymerase chain reaction. *J. Clin. Oncol.* **12:** 475–482.
15. Brown DC, Purushotham AD, Birnie GD, George WD. (1995) Detection of intraoperative tumor cell dissemination in patients with breast cancer by use of reverse transcription and polymerase chain reaction. *Surgery* **117:** 96–101.
16. Mayhew E, Glaves D. (1984) Quantitation of tumorigenic disseminating and arrested cancer cells. *Br. J. Cancer* **50:** 159–166.
17. Weiss L. (1986) Metastatic inefficiency: Causes and consequences. *Cancer Rev.* **3:** 1–24.
18. Fidler IJ. (1970) Metastasis: Quantitative analysis of distribution and fate of tumor emboli labeled with ^{125}I-5-iodo-2′-deoxyuridine. *J. Natl. Cancer Inst.* **45:** 773–778.
19. Glaves D. (1983) Correlation between circulating cancer cells and incidence of metastases. *Br. J. Cancer* **48:** 665–673.
20. Riethmuller G, Schneider-Gadicke E, Schlimok G, et al. (1994) Randomised trial of monoclonal antibody for adjuvant therapy of resected Dukes' C colorectal carcinoma. *Lancet* **343:** 1177–1183.

Contributed by G. J. V. Nossal on July 21, 1995.

This material may be protected by Copyright law (Title 17 U.S. Code)

Detection of Occult Melanoma Cells in Blood With a Multiple-Marker Polymerase Chain Reaction Assay

By Dave S.B. Hoon, Yuan Wang, Paul S. Dale, Andrew J. Conrad, Peter Schmid, Darryl Garrison, Christine Kuo, Leland J. Foshag, Anne J. Nizze, and Donald L. Morton

Purpose: **The objective of the study was to develop a sensitive multimarker polymerase chain reaction (PCR) assay to detect circulating melanoma cells in patient blood. The rationale was that malignant melanoma is heterogeneous in regards to antigen expression.**

Patients and Methods: **A PCR assay that uses four melanoma-associated gene markers (tyrosinase, p97, MUC18, and MAGE-3) was developed. Sensitivity and specificity of the PCR assay for individual markers were assessed using 10 melanoma cell lines and peripheral-blood lymphocytes (PBL) from 39 normal volunteers as controls. The assay's sensitivity and specificity were improved using nested primers and Southern blot analysis. Patients (N = 119) with American Joint Committee on Cancer (AJCC) stages I to IV disease were evaluated for circulating melanoma cells using the four gene markers under optimal conditions.**

Results: **All melanoma-associated gene markers were expressed in at least 80% of the melanoma lines, whereas 37 of 39 normal PBL tested negative for all markers; the remaining two PBL were positive for MUC18. Using four markers in the PCR assay was significantly better than using tyrosinase alone. There was a significant correlation between the number of positive PCR markers, AJCC stage of disease, and progression of disease. In all AJCC stages, there were more PCR-positive patients with disease than without disease.**

Conclusion: **A multimarker PCR assay is more reliable and sensitive than a single-marker assay for detection of melanoma cells in blood of patients. This assay can provide important insight into tumor progression kinetics without major surgical or conventional radiologic diagnostic procedures.**

J Clin Oncol 13:2109-2116. © 1995 by American Society of Clinical Oncology.

MELANOMA METASTATIC to distant sites often leads to a poor prognosis, with reported survival times less than 1 year.[1-3] In general, melanoma metastasizes to regional lymph nodes and organ sites via the lymphatics and blood vessels.[3] Considering the small size of most primary melanoma lesions and their frequency of metastasis, the metastatic potential of melanoma is considerably greater than that of most other human solid tumors. Animal experiments indicate that approximately 0.01% of circulating cancer cells from solid tumors establish successful metastatic colonies.[4]

The detection of metastatic melanoma cells in circulation may be beneficial to the clinician in assessing tumor progression and potential for metastasis. Previous tumor detection studies have focused on detection of tumor-associated antigens or distinct molecules released from tumor cells.[5-7] At best, these assays only detect tumor, with no indication of metastatic potential or tumor progression. In general, most assays measure a single antigen, the release of which is often proportional to the size of the tumor. They do not consider the heterogeneity of individual markers in tumor lesions within individual patients or among patients. The recent development of the polymerase chain reaction (PCR) assay[8,9] for detection of occult metastatic tumor cells in blood using specific markers has provided a new approach to the assessment of tumor progression.[10,11]

In a recent study, circulating melanoma cells in blood were detected by PCR analysis using the tyrosinase gene marker.[10] In this study, only seven melanoma patients with metastatic disease were analyzed: four were positive. Studies on the detection tumor cells in circulation by PCR in melanoma, prostate, breast, and neuroblastoma patients have used only one specific marker and demonstrated the potential feasibility of this approach.[10-13] These studies were limited by the number of patients, use of only one marker, and presence of the marker gene in normal cells. Tumors such as melanoma are heterogeneous, particularly in advanced stages of tumor progression.[3,14-17] Although tumor cells within a primary tumor or metastasis may all express the same marker gene, the level of specific mRNA expression can vary considerably. This difference must be carefully considered in developing a sensitive, specific, and reliable PCR detection assay. The use of multiple markers is one approach to improve the sensitivity of the PCR assay system.

In this study, we used tyrosinase, p97, MAGE-3, and MUC18 as markers in a PCR assay to detect circulating

From the John Wayne Cancer Institute, Saint John's Hospital and Health Center, Santa Monica; and National Genetics Institute, Los Angeles, CA.

Submitted August 11, 1994; accepted April 10, 1995.

Supported in part by grant no. CA 29605 from the National Cancer Institute, Bethesda, MD.

Address reprint requests to Dave S.B. Hoon, MSc, PhD, Division of Molecular and Cellular Immunology, John Wayne Cancer Institute, 2200 Santa Monica Blvd, Santa Monica, CA 90404.

© 1995 by American Society of Clinical Oncology.

0732-183X/95/1308-0036$3.00/0

Exhibit O

tumor cells in the blood of melanoma patients. All four markers are expressed in melanoma.[18-22] Tyrosinase is a key enzyme in melanin synthesis, which is characteristic of melanocytes and melanoma cells.[18,23] The markers p97, MAGE-3, and MUC18 are tumor-associated antigens.[20-22] MAGE-3 gene is transcribed in approximately 69% of melanomas and has been detected only in tumor cells and not in normal tissue, except testes.[21] MUC18 gene encodes a 113-kd glycoprotein, and mRNA expression has been detected in approximately 80% of melanomas and correlated with metastasis.[22,24] P97 gene encodes a glycoprotein that is a tumor-associated antigen on melanoma, referred to as melanotransferrin.[19,20] This study demonstrates the efficacy of using these individual and combined gene markers transcribed to detect occult melanoma cells in the blood.

PATIENTS AND METHODS

Melanoma Cell Lines

Melanoma cell lines M10, M12, M24, M101, Mke, Mst, Mmu, Mka, and Mkn were established and characterized at the John Wayne Cancer Institute (JWCI). Cells were grown in RPMI 1640 plus 10% fetal calf serum (heat-inactivated) (Gemini, Calabasas, CA) plus penicillin and streptomycin (GIBCO, Long Island, NY) in 75-cm^2 flasks. Adherent cell lines were routinely passaged by trypsinization every 3 to 4 days.[25] Cell lines were used for PCR studies when 75% to 85% confluent.

Patients

All melanoma patients who had complete documented physical and medical histories were accrued from JWCI. Their disease stage was recorded as I, II, III, or IV according to American Joint Committee on Cancer (AJCC) guidelines. Clinical disease status was documented when blood was drawn for the PCR assay and again at 8 to 15 months of follow-up evaluation. Disease status was categorized as no evidence of clinical disease (NED), alive with clinical disease (AWD), or died during follow-up evaluation (DFU), disease progression was determined by a change in disease status.

The study was conducted in a double-blind fashion: the patient's disease status was not known to the individual who ran the PCR assay or analyzed the PCR data, and PCR results were not known to the individual who recorded disease status.

Blood Preparation and RNA Extraction

Fifteen milliliters of blood were obtained from patients and collected in sodium citrate tubes. All blood was drawn at JWCI using the same procedure after obtaining written consent from the patient. The protocol for the study was approved by the Saint John's Hospital and JWCI Human Subjects Committee. Samples were centrifuged for 20 minutes at 2,000 *g* in the presence of Ficoll-Hypaque (Sigma, St Louis, MO) and the buffy coat carefully removed. The cells were washed by centrifugation for 5 minutes.

Total cellular RNA was extracted using the UltraSpec isolation system (Biotecx, Houston, TX) as described by the manufacturer. Briefly, the cells were lysed in 2 mL of UltraSpec RNA reagent by repetitive pipetting and placed on ice for 5 minutes. Four hundred microliters of chloroform was added; after rigorous mixing for 15 seconds, the solution was placed on ice for 5 minutes and then centrifuged at 12,000 *g* at 4°C for 15 minutes. The upper phase was transferred into a RNAse-free Eppendorf tube, 1 vol of isopropanol was added, and the solution was precipitated at 4°C for 10 minutes. The tube was centrifuged at 12,000 *g* at 4°C for 20 minutes. The sample was washed with 70% ethanol, dried, and resuspended in 50 μL of diethyl pyrocarbonate (DEPC)-treated Tris-ethylenediaminetetraacetic acid (TE) buffer. All RNA extraction procedures for each specimen were performed in a designated laminar-flow hood under sterile conditions. PCR reagent set-up and gel electrophoresis were performed in separate rooms to avoid potential RNA cross-contamination.

Oligonucleotide Primers and Probes

Oligonucleotide primers were synthesized and purified at the Molecular Biology Institute Core Facility, University of California, Los Angeles. Oligonucleotide 5′ and 3′ primers for individual genes were as follows: MAGE-3—5′-GAAGCCGGCCCAGGCTCG3′and 5′-GGAGTCCTCATAGGATTGGCTCC-3′; MUC18—5′-CCAAGGCAACCTCAGCCATGTC-3′ and 5′-CTCGACTCCACAGTCTGGGACGACT-3′; MUC18 nested—5′-GTCATCTTCCGTGTGCGCCA-3′ and 5′-GTAGCGACCTCCTCAGGCTCCTTAC-3′; tyrosinase—5′-TTGGCAGATTGTCTGTAGCC-3′ and 5′-AGGCATTGTGCATGCTGCTT-3′; tyrosinase nested—5′-GTCTTTATGCAATGGAACGC-3′ and 5′-GCTATCCCAGTAAGTGGACT-3′; and p97—5′-TACCTGGTGGAGAGCGGCCGCCTC-3′ and 5′-AGCGTCTTCCCATCCGTGT-3′. The PCR cDNA products of MAGE-3, MUC18, MUC18 nested, tyrosinase, tyrosinase nested, and p97 were 423, 437, 262, 284, 207 and 286 bp, respectively.

PCR

The PCR assay was performed as previously described with some modifications.[26] The reverse-transcription (RT) mixture consisted of 4 μL of 25-mmol/L magnesium chloride, 2 μL of 10× RT buffer, 4 μL of 10-mmol/L dinucleotide triphosphate (dNTPs) mixture, 0.5 μL of RNAsin (40 U/μL), 1 μL of avian myeloblastosis virus reverse transcriptase (9 U/μL), and 1 μL of oligo(dT)$_{15}$ primer (1.5 μg/μL). Three micrograms of sample RNA was added to the RT mixture and water was added to increase the volume to 20 μL. All reagents were obtained from Promega (Madison, WI). The reaction was incubated at 42°C for 2 hours, at 99°C for 5 minutes, and on ice for 5 minutes.

The PCR mixture consisted of 10 μL of 10× PCR buffer (Perkin Elmer Cetus, Norwalk, CT), 8 μL of 10-mmol/L dNTPs mixture, 1 μL of 5′ oligonucleotide primer (100 pmol/μL), 1 μL of 5′ and 3′ oligonucleotide primer, 0.5 μL of AmpliTaq 5 U/μL (Perkin Elmer Cetus), and 20 μL of RT mixture. Sterile double-distilled water was added to the mixture to increase its volume to 100 μL. The mixture was overlaid with mineral oil. The PCR conditions were set up as follows: 95°C for 5 minutes, followed by 95°C for 70 seconds, 52°C for 70 seconds, 72°C for 70 seconds for 40 cycles, and 72°C for 10 minutes extension time and soaked at 4°C. The PCR reaction was performed in a OmmniGene temperature cycler (Hybaid, Middlesex, England). The preparation of PCR mixture for the temperature cycler was performed in a designated PCR room in a specified laminar-flow hood.

To assess nested primers to a particular gene, 10 μL of the cycled PCR mixture was added to 10 μL of 10× PCR buffer, 8 μL of 10-

mmol/L dNTPs, 1 μL of 5′ oligonucleotide nested primer (100 pmol/μL), 1 μL of 3′ oligonucleotide nested primer (100 pmol/μL), and 0.5 μL of AmpliTaq polymerase (5 U/μL). The volume of the mixture was increased to 100 μL. PCR cycling was repeated using an annealing temperature of 55°C.

The PCR cDNA product was detected by electrophoresis on a 2% agarose gel (GIBCO BRL, Grand Island, NY) and visualized by ethidium bromide staining under UV light. An 0X174RF DNA/*Hae*III fragment DNA ladder (BRL) was used as a size reference marker for all assays.

Southern Blot Analysis

After electrophoresis of PCR cDNA products, agarose gels were transferred overnight onto a nitrocellulose membrane (Schleicher & Schull, Keene, NH) with 20× SSC buffer (1× SSC = 0.15 mol/L sodium chloride + 0.015 mol/L sodium citrate) as previously described.[27] Tyrosinase cDNA probe was a gift of Dr. B. Kwon. Other cDNA probes were prepared from PCR cDNA products of tyrosinase and p97 sequences inside the outermost PCR primers. Probes were prepared and labeled; specific binding was detected using the Genius Kit (Boehringer Mannheim, Indianapolis, IN) as previously described.[28]

Statistical Analysis

The exact coefficient in level for small-sample proportion analysis was performed to determine whether there was any significant difference between using tyrosinase alone and using tyrosinase with MUC18, p97, and MAGE-3. Pearson's χ^2 analysis was used to assess the significance between disease stage and PCR markers.

RESULTS

Detection of Specific Genes in Melanoma Cell Lines

The objective of the study was to develop a PCR melanoma assay that included several prominent gene markers transcribed in melanoma cells but not in normal peripheral-blood lymphocytes (PBL). Tyrosinase and p97, already well established as markers for melanoma, were first assessed. We then screened two other markers, MAGE-3 and MUC18, recently demonstrated to be present at a high frequency in melanoma. The screening process involved examination of 10 established melanoma cell lines (10^6 cells per line) and 39 normal PBL (10^7 cells per blood draw) as controls. PBL were obtained from the buffy coat of 15 mL of blood from healthy normal donors. Table 1 lists the expression of these markers for melanoma cells and PBL. A positive reaction was considered to be a visible specific PCR cDNA product identified by gel electrophoresis with ethidium bromide staining.

Table 1. PCR Analysis of Melanoma Marker Genes

Marker Gene	Melanoma Cell Lines	PBL
Tyrosinase	9/10	0/39
p97	10/10	0/39
MAGE-3	8/10	0/39
MUC18	10/10	2/39

NOTE. RNA was extracted from melanoma and PBL and assessed for expression of individual markers by PCR. Data are presented as positive cell lines or PBL over total number of specimens assessed. Positive PCR refers to PCR cDNA product assessed by gel electrophoresis.

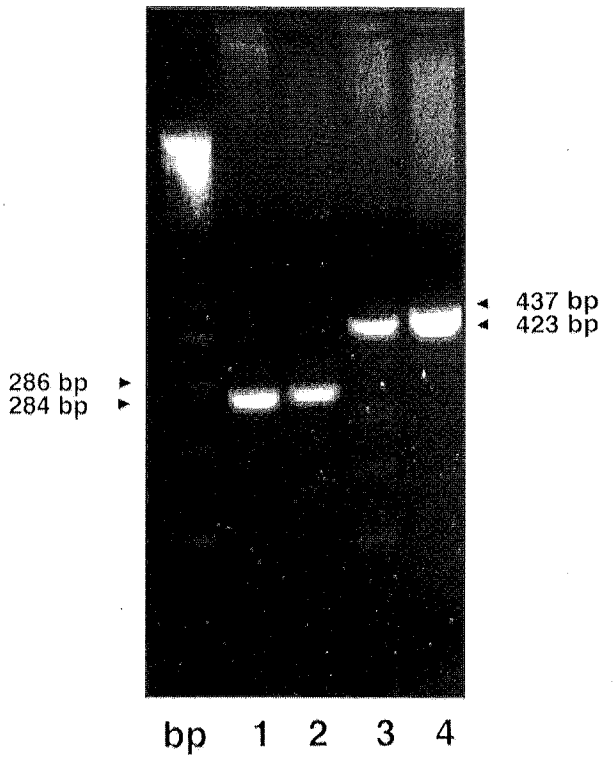


Fig 1. Detection of specific gene markers in melanoma cell lines. Representative example of a melanoma cell line (M12) expression of marker tyrosinase (284 bp; lane 1), p97 (286 bp; lane 2), MAGE-3 (423 bp; lane 3), and MUC18 (437 bp; lane 4). Markers assessed by PCR and gel electrophoresis with ethidium bromide staining, bp, base pair ladder.

Transcription of all four markers in the 10 melanoma cell lines was evaluated. An example of a melanoma cell line that expresses all four markers is shown in Fig 1. As listed in Table 1, one melanoma cell line, tyrosinase expression by PCR, was negative; however, when nested PCR for tyrosinase was performed, tyrosinase gene expression was detected. There was no detection of PCR markers in PBL from 39 normal donors, except two donors who were positive for MUC18 gene transcription. These individuals were tested multiple times from separate blood draws and always remained positive for MUC18. This indicated they were not false-positive results due to PCR contamination or contamination from normal tissue during blood drawing. In all assays, MUC18 nested-primer PCR was performed; this procedure increased the sensitivity to allow verification and

amplification of weak bands produced by PCR with only MUC18 primers. Melanoma cell lines and PBL were tested at least twice to verify specificity. Respective controls in each assay included samples that contained marker gene RNA, PCR reagents and primers without RNA, human tumor cell lines that were negative for individual gene expression, and β-actin gene expression.

MAGE-3 gene expression was found in eight of 10 melanoma cell lines. MAGE-1, a member of the MAGE gene family, was also tested (data not shown) and found to be transcribed in less than 50% of the melanoma cell lines. We therefore decided not to use this marker, since MAGE-3 is more often expressed in melanomas and MAGE-3 is usually found when MAGE-1 is expressed.[21] Neither gene was expressed in PBL of normal volunteer donors.

PCR Sensitivity for Marker Genes

After establishing potential marker genes for melanoma, we assessed each marker's sensitivity. RNA was isolated and quantitated from melanoma cell lines positive for individual markers. Specific-marker PCR analysis was then performed on serial diluted RNA. In general, all markers could be detected at picogram levels of RNA by visual examination of gel-electrophoresed PCR cDNA products stained with ethidium bromide. Figure 2A through D shows representative examples of RNA diluted from melanoma cell lines and assessed for tyrosinase, p97, MUC18, and MAGE-3. Sensitivity varied for individual lines with different levels of gene expression. In general, mRNA for p97, MUC18, and MAGE-3 was detected at approximately 10 to 100 pg by PCR. Tryosinase mRNA could be detected at 10 to 100 fg by PCR. Specificity of the cDNA products was demonstrated by Southern blotting with respective specific cDNA probes for tyrosinase and p97. Sensitivity of the assay could be enhanced 10- to 100-fold using PCR followed by Southern blotting with specific cDNA probe. Nested-primer PCR for tyrosinase enhanced detection levels 10- to 100-fold above PCR for tyrosinase, whereas nested-primer PCR for MUC18 enhanced detection levels no more than 10-fold above PCR for MUC18.

Detection of Melanoma Cells Mixed With PBL In Vitro

An in vitro model system that mimics circulating melanoma cells in blood was developed to detect melanoma cells mixed with PBL. In this assay system, 10^7 normal PBL were mixed with serial dilutions of melanoma cells (10^6 to 1 cell) and assessed by PCR for individual marker genes. The cDNA PCR products were then assessed by gel electrophoresis followed by Southern blot analysis (Fig 3). Southern blot analysis for tyrosinase verified the specificity of the PCR cDNA product and demonstrated enhanced sensitivity. Using PCR and Southern blot analysis or nested-primer PCR, about one melanoma cell in 10^7 PBL could be detected using tyrosinase, p97, or MUC18 as markers. PBL controls were negative for individual markers in gel electrophoresis, nested-primer, and Southern blot analysis. We also set up specific dilutions of melanoma cells in 50 million PBL and demonstrated that approximately one to five melanoma cells could be detected in 50 million PBL with nested-primer tyrosinase PCR followed by Southern blotting.

To demonstrate the sensitivity and reproducibility of detecting one melanoma cell in 10 million PBL, a Poisson distribution analysis was performed. In eight of 11 samples tested by tyrosinase gene, PCR showed a cDNA PCR product by gel electrophoresis analysis. The level of detection was enhanced by greater than 90% using tyrosinase-nested PCR primers or Southern blot analysis with tyrosinase cDNA probe.

Assessment of Circulating Melanoma Cells

PBL from melanoma patients were examined using our optimal PCR assay detection system. The protocol was as follows: PCR assays were performed to detect transcripts of tyrosinase, p97, MAGE-3, and MUC18 in PBL from 119 melanoma patients. If the sample was negative for tyrosinase or MUC18, nested PCR was performed with respective primers. If the test was negative for tyrosinase nested primer and p97, then Southern blot analysis was performed with respective cDNA probes. PBL negative for all the markers and tests were considered truly negative.

Initially, PBL isolated by Ficoll-Hypaque gradient centrifugation were compared with those isolated from buffy-coat separation after centrifugation. Our analysis demonstrated buffy-coat–isolated cells were better for detecting melanoma cells in PBL.

Results of PCR analysis of PBL from melanoma patients are listed in Table 2. MUC18 was expressed most often (72%), followed by tyrosinase (59%), p97 (56%), and MAGE-3 (10%). The use of nested primers of tyrosinase in the tyrosinase PCR versus tyrosinase primers alone significantly increased the number of positive patients from two to 56. Using PCR followed by Southern blotting with p97 and tyrosinase cDNA probes also significantly increased the number of positive patients. The most significant increase was observed by Southern blotting with the p97 cDNA probe. Seven patients were positive for all four markers. All seven patients had stage IV disease.

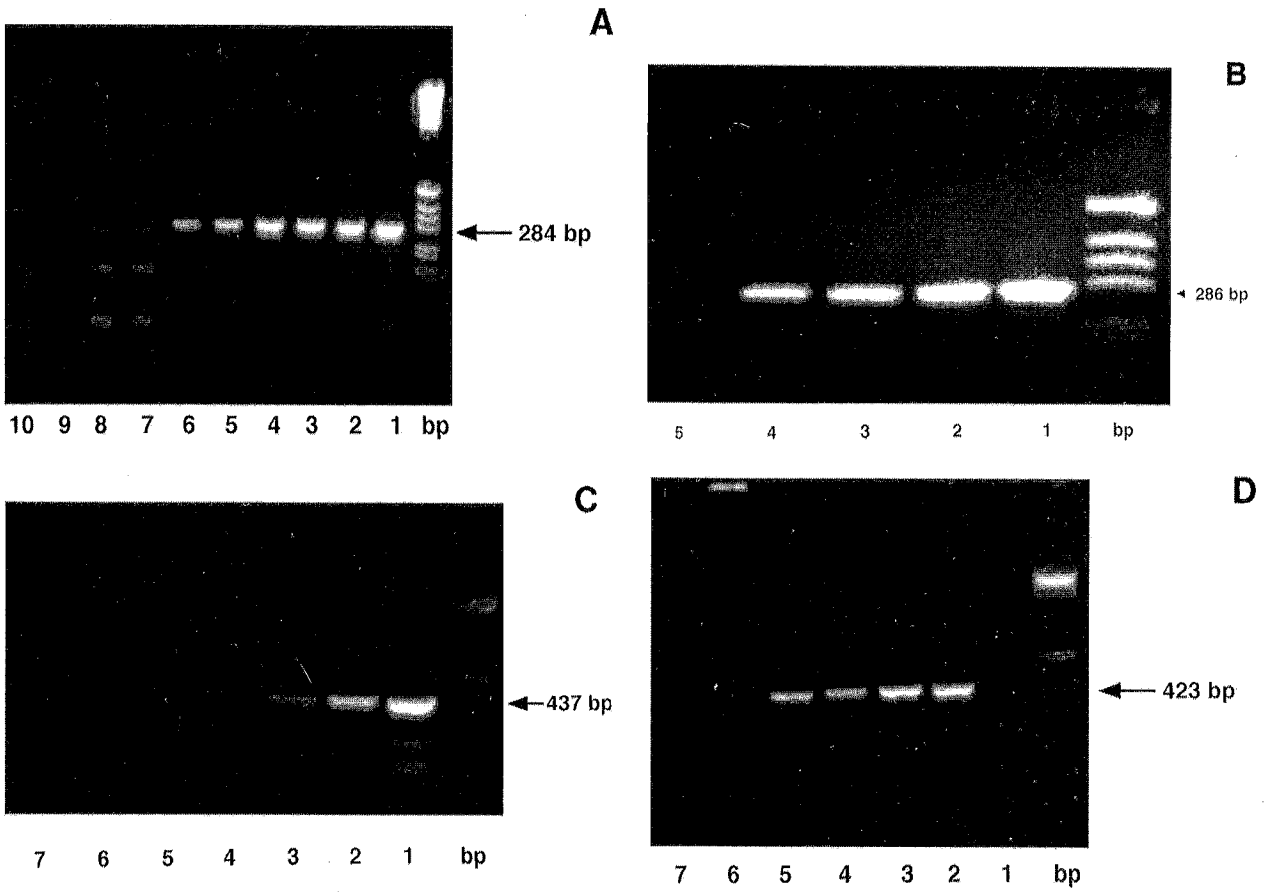


Fig 2. RNA dilutions from melanoma cell lines assessment by PCR markers. PCR markers assessed were (A) tyrosinase, (B) p97, (C) MUC18, and (D) MAGE-3. Columns 1 to 10 represent RNA serial dilutions of 1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}, and 10^{-9} μg, respectively. RNA was extracted from cell lines and diluted, and PCR was performed. Positive and negative controls were run.

PCR analyses were correlated with disease stage and status. The follow-up time for clinical status after blood drawing for PCR analysis was 8 to 15 months. There were six stage I, 11 stage II, 36 stage III, and 66 stage IV patients (Table 3). The majority of the stage II, III, and IV patients were PCR-positive (Table 3). Of 119 patients assessed, 49 were negative for tyrosinase expression. Combining all four markers was significantly ($P <$ .01) better than using the tyrosinase marker alone. There was a positive correlation between the number of positive markers to both stage ($P = .0029$; Table 4) and the progression of disease ($P < .05$). Tyrosinase was the only single marker to show significant correlation ($P < .01$) with disease stage.

Breslow thickness and Clark's level of the primary melanoma play a major role in determining patients' prognosis.[29,30] Patients were divided into subgroups based on Breslow thickness of $\leq$ 0.75 mm, greater than 0.75 mm to 1.49 mm, $\geq$ 1.5 mm to 3.0 mm, and greater than 3.0 mm. There was no significant correlation of Breslow thickness and detection of PCR markers (data not shown). The majority of patients were either Clark's level 3 or 4, and no significant pattern was observed for Clark's level and detection of PCR markers (data not shown). The number of tumor-positive regional lymph nodes or the sites of distal metastases did not significantly correlate with the detection of PCR markers (data not shown).

DISCUSSION

In this study, we have developed a sensitive and reproducible multimarker PCR assay to detect occult melanoma cells in the blood of patients with or without clinical evidence of disease. The in vitro models provided an excellent approach to assess the efficacy of the assay. The assay was able to detect about one melanoma cell in 10 to 50 million normal PBL. Sensitivity varied with the number of mRNA copies of the expressed gene. Although tyrosinase has been used as the single marker in a previ-

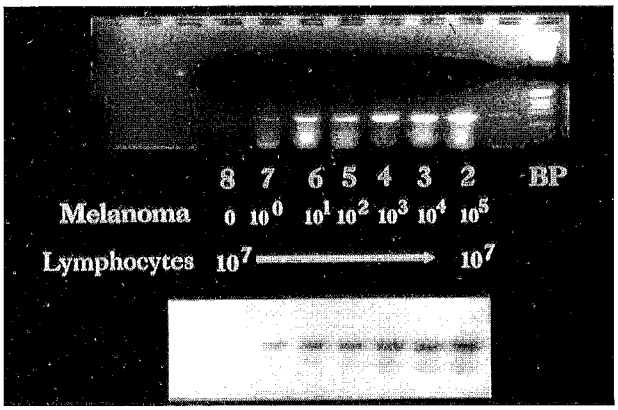


Fig 3. Sensitivity of tyrosinase PCR assay to detect melanoma cells in PBL. Melanoma cells were serially diluted (10^0 to 10^5) in 10^7 PBL (columns 2 to 7). Controls are 10^7 PBL and 10^1 melanoma cells (columns 8 and 1, respectively). PCR was performed on RNA from cell mixtures, melanoma cells, and PBL. Southern blot analysis was performed with tyrosinase cDNA probe (columns 1 to 8). Melanoma cells were detected only in columns 1 to 7. Column 8 was negative.

ous report,[10] our studies indicate that it is not always sensitive alone in detecting circulating melanoma cells. The use of more than one marker can verify the presence of occult melanoma cells and significantly increase the sensitivity to detect melanoma cells that express few or no copies of tyrosinase mRNA. The study demonstrated that using four markers was significantly better than tyrosinase alone. This was our initial hypothesis. The number of markers detected in individual patients was significantly correlated with disease stage and progression. This suggests that there could be higher expression of individual marker genes, greater heterogeneity of tumor cells, or more tumor cells in the circulation at advanced stages of disease. Interestingly, PCR results showed no correlation with either Breslow thickness or Clark's level of the primary melanoma, both of which are important prognostic factors.[29,30] This finding indicates that tumor progression is no longer dependent on the initial pathologic parameters of the primary tumor after it has been removed.

Overall, the level of detection was highest for MUC18, similar for tyrosinase and p97, and lowest for MAGE-3. Although MAGE-3 is expressed with high frequency in cell lines and biopsies, the number of mRNA copies in single cells is likely to be low. This may be related to the state of the cell during circulation in the blood or clonal phenotype.

Previous PCR studies, have not analyzed large numbers of patients with different clinical stages of melanoma. These parameters are important to evaluate the sensitivity and clinical significance of the assay. At this point, we do not fully know the short- or long-term clinical significance of circulating tumor cells in the blood, other than the potential for metastasis to other organs. However, this information may be useful to identify patients who may need more intensive therapeutic intervention. Immediate therapeutic intervention may be an efficacious means to control tumor progression in NED patients with circulating tumor cells. The detection of circulating melanoma cells would also be useful to monitor a patient's response to operative and adjuvant therapies. In our analysis of stage III and IV patients, there was no major difference

Table 2. Analysis of Melanoma Patients Using Multiple-Marker PCR Assay

Assays	No. of Patients Positive			
	p97	tyrosinase	MAGE-3	MUC18
PCR	16	2	12	79
Nested PCR	—	56	—	7
Southern blot	48	12	—	—
Total patients	64	70	12	86

NOTE. One hundred twenty melanoma patients were evaluated. PCR and Southern blot refer to assays positive for individual marker genes. Nested PCR was performed for tyrosinase and MUC18 when negative by PCR. Nested PCR was also performed when MUC18 PCR was positive. Southern blotting with specific cDNA probes was performed after p97 and tyrosinase PCR when PCR was negative by gel electrophoresis and ethidium bromide staining.

Table 3. PCR-Positive Patients: Correlation to Stage and Disease Status

AJCC Stage	AWD & DFU	NED
I	1/1	2/6
II	NP	10/11
III	5/6	25/29
IV	46/48	16/18
Total positive		
No.	52/55	58/64
%	95	90

NOTE. Values represent patients who were PCR-positive (≥ 1 markers) over total patients evaluated.
Abbreviation: NP, no patients.

Table 4. Number of PCR Markers Correlated to Disease Stage

AJCC Stage	No. of Positive Markers (no. of patients)					Total Patients
	0	1	2	3	4	
I	3	0	1	2	0	6
II	1	6	2	2	0	11
III	5	6	18	7	0	36
IV	3	15	21	20	7	66
Total	12	27	42	31	7	119

NOTE. Positive markers refer to detection of tyrosinase, p97, MUC18, and MAGE-3 by PCR, nested PCR, or Southern blotting.

in disease progression and PCR results. Future studies on serial blood analysis and longer-term follow-up evaluation of patients will allow a more detail assessment of PCR results to disease progression and prognosis.

The presence of circulating tumor cells in the blood is only a segment of the tumor progression cascade toward establishment of distant metastases. The time course from the detection of melanoma cells in correlation to detection of clinical detectable metastasis may vary considerably among patients. Multiple parameters, which include the hosts' antitumor immune status, the tumor's invasion capacity, survival during blood circulation, growth potential, and colonization potential, all play a role in the establishment of a successful metastasis.[4,31-33] Metastasis to a distal organ may or may not further metastasize to other organs. Since subclinical metastasis can remain dormant for many years, monitoring patients blood for circulating tumor cells may be advantageous for detecting tumor progression before clinically evident metastases to other organs are detected.

Applying multiple melanoma markers to the evaluation of circulating melanoma cells also provides information about the tumor's phenotype. Identification of specific tumor-associated antigen(s) expressed by tumor cells will aid in the use of antigen-specific immunotherapy such as monoclonal antibodies.[26] The PCR assay would also provide a rapid monitoring system to determine if this specific therapy is effective. These protocols are currently being developed.

ACKNOWLEDGMENT

We thank the JWCI nursing staff, David Elashoff, PhD, and Dorit Natflin for technical assistance. We also thank Janet Sinsheimer, PhD, and Stacy Stern, MS, of the JWCI Biostatistical Division.

REFERENCES

1. Morton DL, Davtyan DG, Wanek LA, et al: Multivariate analysis of the relationship between survival and the microstage of primary melanoma by Clark level and Breslow thickness. Cancer 71:3737-3743, 1993

2. Balch CM, Soong SW, Shaw HM, et al: An analysis of prognostic factors in 4000 patients with cutaneous melanoma, in Balch CM, Milton GW (eds): Cutaneous Melanoma: Clinical Management and Treatment Results Worldwide. Philadelphia, PA, Lippincott 1985, pp 321-352

3. Morton DL, Wong JH, Kirkwood JM, et al: Malignant melanoma, in Holland JF, Frei E III, Bast CC Jr, et al (eds): Cancer Medicine (ed 3). Philadelphia, PA, Lea & Febiger, 1993, pp 1793-1824

4. Fidler IJ: Critical Factors in the biology of human cancer metastasis: Twenty-Eighth G.H.A. Clowes Memorial Award Lecture. Cancer Res 50:6130-6138, 1990

5. Smart CR: Screening and early cancer detection. Semin Oncol 17:456-462, 1990

6. Moertel CG, Fleming TR, MacDonald JS, et al: An evaluation of the carcinoembryonic antigen (CEA) test for monitoring patients with resected patients with resected colon cancer. JAMA 270:943-947, 1993

7. Stamey TA, Kabalin JN, McNeal JE, et al: Prostate-specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. J Urol 141:1076-1083, 1989

8. Mullis KB, Faloona F: Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. Methods Enzymol 155:335-350, 1987

9. Erlich HA (ed): PCR Technology: Principles and Applications for DNA Amplification. New York, NY, Stockton, 1989

10. Smith B, Selby P, Southgate J, et al: Detection of melanoma cells in peripheral blood by means of reverse transcriptase and polymerase chain reaction. Lancet 338:1227-1229, 1991

11. Naito H, Kuzumaki N, Uchino J-I, et al: Detection of tyrosine hydroxylase mRNA and minimal neuroblastoma cells by the reverse transcription-polymerase chain reaction. Eur J Cancer 27:762-765, 1991

12. Moreno JG, Croce CM, Fisher R, et al: Detection of hematogenous micrometastasis in patients with prostate cancer. Cancer Res 52:6110-6112, 1992

13. Datta YH, Adams PT, Drobyski WR, et al: Sensitive detection of occult breast cancer by the reverse-transcriptase polymerase chain reaction. J Clin Oncol 12:475-482, 1994

14. Fidler IJ, Hart IR: Biologic diversity in metastatic neoplasms origins and implications. Science 217:998-1001, 1982

15. Nowell PC: Genetic instability in cancer cells: relationship to tumor cell heterogeneity, in Owens AH, Coffey DS, Baylin SB (eds): Tumor Cell Heterogeneity. New York, NY, Academic, 1982, pp 351-365

16. Elder DE, Rodeck U, Thurin J, et al: Pigmented lesion-associated antigens distinguish between benign and malignant melanocytic lesions. Cancer Res 49:5091-5096, 1989

17. Bystryn J-C, Bernstein P, Liu P, et al: Immunophenotype of human melanoma cells in different metastases. Cancer Res 45:5603-5607, 1985

18. Nordlund JJ, Abdel-Malek ZA, Boissy RE, et al: Pigment cell biology: An historical review. J Invest Dermatol 92:53S-60S, 1989

19. Brown JP, Nishiyama K, Hellstrom I, et al: Structural characterization of human melanoma-associated antigen p97 using monoclonal antibodies. J Immunol 127:539-546, 1981

20. Rose TM, Plowman GD, Teplow DB, et al: Primary structure of the human melanoma-antigen p97 (melanotransferrin) deduced from the mRNA sequence. Proc Natl Acad Sci USA 83:1261-1265, 1986

21. Gaugler B, Van den Eynde B, van der Bruggen P, et al: Human gene MAGE-3 codes for an antigen recognized on melanoma by autologous cytolytic T lymphocytes. J Exp Med 179:921-930, 1994

22. Lehmann JM, Riethmuller G, Johnson JP: MUC18, a marker of tumor progression in human melanoma, shows a sequence similarity to the neural cell adhesion molecules of the immunoglobulin superfamily. Proc Natl Acad Sci USA 86:9891-9895, 1989

23. Kwon BS, Haq AK, Pomerantz SH, et al: Isolation and sequence of a cDNA clone for human tyrosinase that maps at the mouse c-albino locus. Proc Natl Acad Sci USA 84:7473-7477, 1987

24. Lehmann JM, Holzmann B, Breitbart EW, et al: Discrimination between benign and malignant cells of melanocytic lineage by two novel antigens, a glycoprotein with a molecular weight of 113,000 and a protein with a molecular weight of 76,000. Cancer Res 47:841-845, 1987

25. Hoon DSB, Hayashi Y, Morisaki T, et al: Interleukin-4 plus tumor necrosis factor α augments the antigenicity of melanoma cells. Cancer Immunol Immunother 37:378-384, 1993

26. Hoon DSB, Wang Y, Sze L, et al: Molecular cloning of a human monoclonal antibody reactive to ganglioside GM3 antigen on human cancers. Cancer Res 53:5244-5250, 1993

27. Sambrook J, Fritsch EF, Maniatis T (eds): Molecular Cloning. Cold Spring Harbor, NY, Cold Spring Harbor Laboratory, 1989

28. Morisaki T, Yuzuki DH, Lin RT, et al: Interleukin 4 receptor expression and growth inhibition of gastric carcinoma cells by interleukin 4. Cancer Res 52:6059-6065, 1992

29. Breslow A: Thickness, cross-sectional area and depth of invasion in the prognosis of cutaneous melanoma. Ann Surg 172:902-908, 1970

30. Morton DL, Davtyan DG, Wanek LA, et al: Multivariate analysis of the relationship between survival and the microstage of primary melanoma by Clark's level and Breslow thickness. Cancer 71:3737-3743, 1993

31. Kaiser HE: Characteristics and pattern of direct tumor spreading, in Brunson KW (ed): Local Invasion and Spread of Cancer. Dordrecht, the Netherlands, Kluwer Academic, 1989, pp 1-16

32. Morton DL, Foshag LJ, Hoon DSB, et al: Prolongation of survival in metastatic melanoma after active specific immunotherapy with a new polyvalent melanoma vaccine. Ann Surg 216:463-482, 1992

33. Clark WH Jr, Elder DE, Guerry D IV, et al: A study of tumor progression: The precursor lesions of superficial spreading and nodular melanoma. Hum Pathol 15:1147-1165, 1984

Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium*

* *A partial list of authors appears on the opposite page. Affiliations are listed at the end of the paper.*

The human genome holds an extraordinary trove of information about human development, physiology, medicine and evolution. Here we report the results of an international collaboration to produce and make freely available a draft sequence of the human genome. We also present an initial analysis of the data, describing some of the insights that can be gleaned from the sequence.

The rediscovery of Mendel's laws of heredity in the opening weeks of the 20th century[1–3] sparked a scientific quest to understand the nature and content of genetic information that has propelled biology for the last hundred years. The scientific progress made falls naturally into four main phases, corresponding roughly to the four quarters of the century. The first established the cellular basis of heredity: the chromosomes. The second defined the molecular basis of heredity: the DNA double helix. The third unlocked the informational basis of heredity, with the discovery of the biological mechanism by which cells read the information contained in genes and with the invention of the recombinant DNA technologies of cloning and sequencing by which scientists can do the same.

The last quarter of a century has been marked by a relentless drive to decipher first genes and then entire genomes, spawning the field of genomics. The fruits of this work already include the genome sequences of 599 viruses and viroids, 205 naturally occurring plasmids, 185 organelles, 31 eubacteria, seven archaea, one fungus, two animals and one plant.

Here we report the results of a collaboration involving 20 groups from the United States, the United Kingdom, Japan, France, Germany and China to produce a draft sequence of the human genome. The draft genome sequence was generated from a physical map covering more than 96% of the euchromatic part of the human genome and, together with additional sequence in public databases, it covers about 94% of the human genome. The sequence was produced over a relatively short period, with coverage rising from about 10% to more than 90% over roughly fifteen months. The sequence data have been made available without restriction and updated daily throughout the project. The task ahead is to produce a finished sequence, by closing all gaps and resolving all ambiguities. Already about one billion bases are in final form and the task of bringing the vast majority of the sequence to this standard is now straightforward and should proceed rapidly.

The sequence of the human genome is of interest in several respects. It is the largest genome to be extensively sequenced so far, being 25 times as large as any previously sequenced genome and eight times as large as the sum of all such genomes. It is the first vertebrate genome to be extensively sequenced. And, uniquely, it is the genome of our own species.

Much work remains to be done to produce a complete finished sequence, but the vast trove of information that has become available through this collaborative effort allows a global perspective on the human genome. Although the details will change as the sequence is finished, many points are already clear.

- The genomic landscape shows marked variation in the distribution of a number of features, including genes, transposable elements, GC content, CpG islands and recombination rate. This gives us important clues about function. For example, the developmentally important HOX gene clusters are the most repeat-poor regions of the human genome, probably reflecting the very complex coordinate regulation of the genes in the clusters.
- There appear to be about 30,000–40,000 protein-coding genes in the human genome—only about twice as many as in worm or fly. However, the genes are more complex, with more alternative splicing generating a larger number of protein products.
- The full set of proteins (the 'proteome') encoded by the human genome is more complex than those of invertebrates. This is due in part to the presence of vertebrate-specific protein domains and motifs (an estimated 7% of the total), but more to the fact that vertebrates appear to have arranged pre-existing components into a richer collection of domain architectures.
- Hundreds of human genes appear likely to have resulted from horizontal transfer from bacteria at some point in the vertebrate lineage. Dozens of genes appear to have been derived from transposable elements.
- Although about half of the human genome derives from transposable elements, there has been a marked decline in the overall activity of such elements in the hominid lineage. DNA transposons appear to have become completely inactive and long-terminal repeat (LTR) retroposons may also have done so.
- The pericentromeric and subtelomeric regions of chromosomes are filled with large recent segmental duplications of sequence from elsewhere in the genome. Segmental duplication is much more frequent in humans than in yeast, fly or worm.
- Analysis of the organization of Alu elements explains the long-standing mystery of their surprising genomic distribution, and suggests that there may be strong selection in favour of preferential retention of Alu elements in GC-rich regions and that these 'selfish' elements may benefit their human hosts.
- The mutation rate is about twice as high in male as in female meiosis, showing that most mutation occurs in males.
- Cytogenetic analysis of the sequenced clones confirms suggestions that large GC-poor regions are strongly correlated with 'dark G-bands' in karyotypes.
- Recombination rates tend to be much higher in distal regions (around 20 megabases (Mb)) of chromosomes and on shorter chromosome arms in general, in a pattern that promotes the occurrence of at least one crossover per chromosome arm in each meiosis.
- More than 1.4 million single nucleotide polymorphisms (SNPs) in the human genome have been identified. This collection should allow the initiation of genome-wide linkage disequilibrium mapping of the genes in the human population.

In this paper, we start by presenting background information on the project and describing the generation, assembly and evaluation of the draft genome sequence. We then focus on an initial analysis of the sequence itself: the broad chromosomal landscape; the repeat elements and the rich palaeontological record of evolutionary and biological processes that they provide; the human genes and proteins and their differences and similarities with those of other

Exhibit P

© 2001 Macmillan Magazines Ltd

Genome Sequencing Centres (Listed in order of total genomic sequence contributed, with a partial list of personnel. A full list of contributors at each centre is available as Supplementary Information.)

Whitehead Institute for Biomedical Research, Center for Genome Research: Eric S. Lander[1]*, Lauren M. Linton[1], Bruce Birren[1]*, Chad Nusbaum[1]*, Michael C. Zody[1]*, Jennifer Baldwin[1], Keri Devon[1], Ken Dewar[1], Michael Doyle[1], William FitzHugh[1]*, Roel Funke[1], Diane Gage[1], Katrina Harris[1], Andrew Heaford[1], John Howland[1], Lisa Kann[1], Jessica Lehoczky[1], Rosie LeVine[1], Paul McEwan[1], Kevin McKernan[1], James Meldrim[1], Jill P. Mesirov[1]*, Cher Miranda[1], William Morris[1], Jerome Naylor[1], Christina Raymond[1], Mark Rosetti[1], Ralph Santos[1], Andrew Sheridan[1], Carrie Sougnez[1], Nicole Stange-Thomann[1], Nikola Stojanovic[1], Aravind Subramanian[1] & Dudley Wyman[1]

The Sanger Centre: Jane Rogers[2], John Sulston[2]*, Rachael Ainscough[2], Stephan Beck[2], David Bentley[2], John Burton[2], Christopher Clee[2], Nigel Carter[2], Alan Coulson[2], Rebecca Deadman[2], Panos Deloukas[2], Andrew Dunham[2], Ian Dunham[2], Richard Durbin[2]*, Lisa French[2], Darren Grafham[2], Simon Gregory[2], Tim Hubbard[2]*, Sean Humphray[2], Adrienne Hunt[2], Matthew Jones[2], Christine Lloyd[2], Amanda McMurray[2], Lucy Matthews[2], Simon Mercer[2], Sarah Milne[2], James C. Mullikin[2]*, Andrew Mungall[2], Robert Plumb[2], Mark Ross[2], Ratna Shownkeen[2] & Sarah Sims[2]

Washington University Genome Sequencing Center: Robert H. Waterston[3]*, Richard K. Wilson[3], LaDeana W. Hillier[3]*, John D. McPherson[3], Marco A. Marra[3], Elaine R. Mardis[3], Lucinda A. Fulton[3], Asif T. Chinwalla[3]*, Kymberlie H. Pepin[3], Warren R. Gish[3], Stephanie L. Chissoe[3], Michael C. Wendl[3], Kim D. Delehaunty[3], Tracie L. Miner[3], Andrew Delehaunty[3], Jason B. Kramer[3], Lisa L. Cook[3], Robert S. Fulton[3], Douglas L. Johnson[3], Patrick J. Minx[3] & Sandra W. Clifton[3]

US DOE Joint Genome Institute: Trevor Hawkins[4], Elbert Branscomb[4], Paul Predki[4], Paul Richardson[4], Sarah Wenning[4], Tom Slezak[4], Norman Doggett[4], Jan-Fang Cheng[4], Anne Olsen[4], Susan Lucas[4], Christopher Elkin[4], Edward Uberbacher[4] & Marvin Frazier[4]

Baylor College of Medicine Human Genome Sequencing Center: Richard A. Gibbs[5]*, Donna M. Muzny[5], Steven E. Scherer[5], John B. Bouck[5]*, Erica J. Sodergren[5], Kim C. Worley[5]*, Catherine M. Rives[5], James H. Gorrell[5], Michael L. Metzker[5], Susan L. Naylor[6], Raju S. Kucherlapati[7], David L. Nelson, & George M. Weinstock[8]

RIKEN Genomic Sciences Center: Yoshiyuki Sakaki[9], Asao Fujiyama[9], Masahira Hattori[9], Tetsushi Yada[9], Atsushi Toyoda[9], Takehiko Itoh[9], Chiharu Kawagoe[9], Hidemi Watanabe[9], Yasushi Totoki[9] & Todd Taylor[9]

Genoscope and CNRS UMR-8030: Jean Weissenbach[10], Roland Heilig[10], William Saurin[10], Francois Artiguenave[10], Philippe Brottier[10], Thomas Bruls[10], Eric Pelletier[10], Catherine Robert[10] & Patrick Wincker[10]

GTC Sequencing Center: Douglas R. Smith[11], Lynn Doucette-Stamm[11], Marc Rubenfield[11], Keith Weinstock[11], Hong Mei Lee[11] & JoAnn Dubois[11]

Department of Genome Analysis, Institute of Molecular Biotechnology: André Rosenthal[12], Matthias Platzer[12], Gerald Nyakatura[12], Stefan Taudien[12] & Andreas Rump[12]

Beijing Genomics Institute/Human Genome Center: Huanming Yang[13], Jun Yu[13], Jian Wang[13], Guyang Huang[14] & Jun Gu[15]

Multimegabase Sequencing Center, The Institute for Systems Biology: Leroy Hood[16], Lee Rowen[16], Anup Madan[16] & Shizen Qin[16]

Stanford Genome Technology Center: Ronald W. Davis[17], Nancy A. Federspiel[17], A. Pia Abola[17] & Michael J. Proctor[17]

Stanford Human Genome Center: Richard M. Myers[18], Jeremy Schmutz[18], Mark Dickson[18], Jane Grimwood[18] & David R. Cox[18]

University of Washington Genome Center: Maynard V. Olson[19], Rajinder Kaul[19] & Christopher Raymond[19]

Department of Molecular Biology, Keio University School of Medicine: Nobuyoshi Shimizu[20], Kazuhiko Kawasaki[20] & Shinsei Minoshima[20]

University of Texas Southwestern Medical Center at Dallas: Glen A. Evans[21]†, Maria Athanasiou[21] & Roger Schultz[21]

University of Oklahoma's Advanced Center for Genome Technology: Bruce A. Roe[22], Feng Chen[22] & Huaqin Pan[22]

Max Planck Institute for Molecular Genetics: Juliane Ramser[23], Hans Lehrach[23] & Richard Reinhardt[23]

Cold Spring Harbor Laboratory, Lita Annenberg Hazen Genome Center: W. Richard McCombie[24], Melissa de la Bastide[24] & Neilay Dedhia[24]

GBF—German Research Centre for Biotechnology: Helmut Blöcker[25], Klaus Hornischer[25] & Gabriele Nordsiek[25]

* **Genome Analysis Group (listed in alphabetical order, also includes individuals listed under other headings):** Richa Agarwala[26], L. Aravind[26], Jeffrey A. Bailey[27], Alex Bateman[2], Serafim Batzoglou[1], Ewan Birney[28], Peer Bork[29,30], Daniel G. Brown[1], Christopher B. Burge[31], Lorenzo Cerutti[28], Hsiu-Chuan Chen[26], Deanna Church[26], Michele Clamp[2], Richard R. Copley[30], Tobias Doerks[29,30], Sean R. Eddy[32], Evan E. Eichler[27], Terrence S. Furey[33], James Galagan[1], James G. R. Gilbert[2], Cyrus Harmon[34], Yoshihide Hayashizaki[35], David Haussler[36], Henning Hermjakob[28], Karsten Hokamp[37], Wonhee Jang[26], L. Steven Johnson[32], Thomas A. Jones[32], Simon Kasif[38], Arek Kaspryzk[28], Scot Kennedy[39], W. James Kent[40], Paul Kitts[26], Eugene V. Koonin[26], Ian Korf[3], David Kulp[34], Doron Lancet[41], Todd M. Lowe[42], Aoife McLysaght[37], Tarjei Mikkelsen[38], John V. Moran[43], Nicola Mulder[28], Victor J. Pollara[1], Chris P. Ponting[44], Greg Schuler[26], Jörg Schultz[30], Guy Slater[28], Arian F. A. Smit[45], Elia Stupka[28], Joseph Szustakowki[38], Danielle Thierry-Mieg[26], Jean Thierry-Mieg[26], Lukas Wagner[26], John Wallis[3], Raymond Wheeler[34], Alan Williams[34], Yuri I. Wolf[26], Kenneth H. Wolfe[37], Shiaw-Pyng Yang[3] & Ru-Fang Yeh[31]

Scientific management: National Human Genome Research Institute, US National Institutes of Health: Francis Collins[46]*, Mark S. Guyer[46], Jane Peterson[46], Adam Felsenfeld[46]* & Kris A. Wetterstrand[46]; **Office of Science, US Department of Energy:** Aristides Patrinos[47]; **The Wellcome Trust:** Michael J. Morgan[48]

© 2001 Macmillan Magazines Ltd

organisms; and the history of genomic segments. (Comparisons are drawn throughout with the genomes of the budding yeast *Saccharomyces cerevisiae*, the nematode worm *Caenorhabditis elegans*, the fruitfly *Drosophila melanogaster* and the mustard weed *Arabidopsis thaliana*; we refer to these for convenience simply as yeast, worm, fly and mustard weed.) Finally, we discuss applications of the sequence to biology and medicine and describe next steps in the project. A full description of the methods is provided as Supplementary Information on *Nature*'s web site (http://www.nature.com).

We recognize that it is impossible to provide a comprehensive analysis of this vast dataset, and thus our goal is to illustrate the range of insights that can be gleaned from the human genome and thereby to sketch a research agenda for the future.

Background to the Human Genome Project

The Human Genome Project arose from two key insights that emerged in the early 1980s: that the ability to take global views of genomes could greatly accelerate biomedical research, by allowing researchers to attack problems in a comprehensive and unbiased fashion; and that the creation of such global views would require a communal effort in infrastructure building, unlike anything previously attempted in biomedical research. Several key projects helped to crystallize these insights, including:

(1) The sequencing of the bacterial viruses ΦX174[4,5] and lambda[6], the animal virus SV40[7] and the human mitochondrion[8] between 1977 and 1982. These projects proved the feasibility of assembling small sequence fragments into complete genomes, and showed the value of complete catalogues of genes and other functional elements.

(2) The programme to create a human genetic map to make it possible to locate disease genes of unknown function based solely on their inheritance patterns, launched by Botstein and colleagues in 1980 (ref. 9).

(3) The programmes to create physical maps of clones covering the yeast[10] and worm[11] genomes to allow isolation of genes and regions based solely on their chromosomal position, launched by Olson and Sulston in the mid-1980s.

(4) The development of random shotgun sequencing of complementary DNA fragments for high-throughput gene discovery by Schimmel[12] and Schimmel and Sutcliffe[13], later dubbed expressed sequence tags (ESTs) and pursued with automated sequencing by Venter and others[14–20].

The idea of sequencing the entire human genome was first proposed in discussions at scientific meetings organized by the US Department of Energy and others from 1984 to 1986 (refs 21, 22). A committee appointed by the US National Research Council endorsed the concept in its 1988 report[23], but recommended a broader programme, to include: the creation of genetic, physical and sequence maps of the human genome; parallel efforts in key model organisms such as bacteria, yeast, worms, flies and mice; the development of technology in support of these objectives; and research into the ethical, legal and social issues raised by human genome research. The programme was launched in the US as a joint effort of the Department of Energy and the National Institutes of Health. In other countries, the UK Medical Research Council and the Wellcome Trust supported genomic research in Britain; the Centre d'Etude du Polymorphisme Humain and the French Muscular Dystrophy Association launched mapping efforts in France; government agencies, including the Science and Technology Agency and the Ministry of Education, Science, Sports and Culture supported genomic research efforts in Japan; and the European Community helped to launch several international efforts, notably the programme to sequence the yeast genome. By late 1990, the Human Genome Project had been launched, with the creation of genome centres in these countries. Additional participants subsequently joined the effort, notably in Germany and China. In addition, the Human Genome Organization (HUGO) was founded to provide a forum for international coordination of genomic research. Several books[24–26] provide a more comprehensive discussion of the genesis of the Human Genome Project.

Through 1995, work progressed rapidly on two fronts (Fig. 1). The first was construction of genetic and physical maps of the human and mouse genomes[27–31], providing key tools for identification of disease genes and anchoring points for genomic sequence. The second was sequencing of the yeast[32] and worm[33] genomes, as

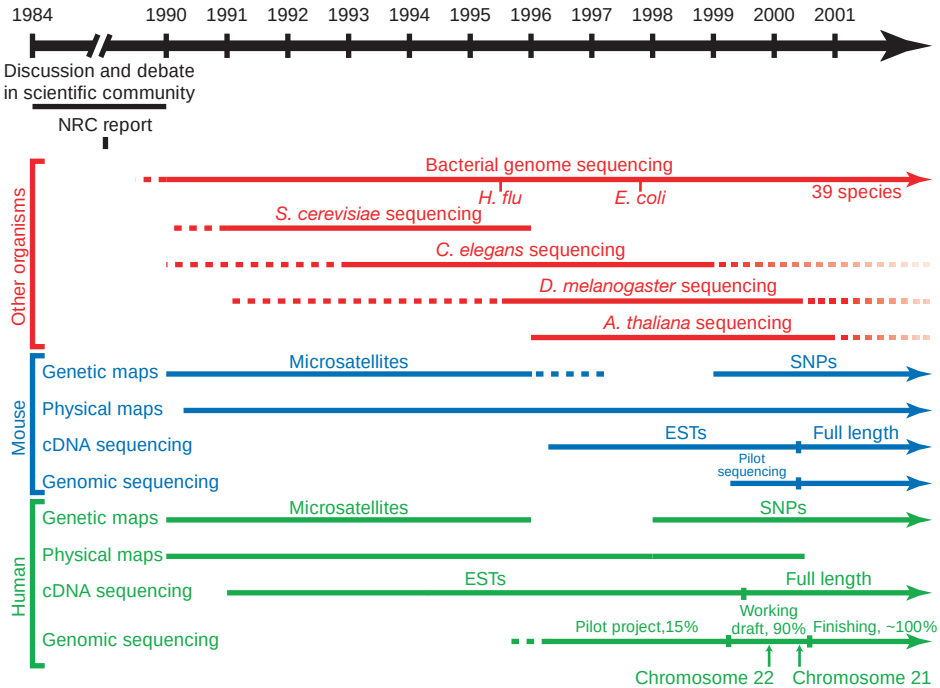


Figure 1 Timeline of large-scale genomic analyses. Shown are selected components of work on several non-vertebrate model organisms (red), the mouse (blue) and the human (green) from 1990; earlier projects are described in the text. SNPs, single nucleotide polymorphisms; ESTs, expressed sequence tags.

© 2001 Macmillan Magazines Ltd

well as targeted regions of mammalian genomes[34–37]. These projects showed that large-scale sequencing was feasible and developed the two-phase paradigm for genome sequencing. In the first, 'shotgun', phase, the genome is divided into appropriately sized segments and each segment is covered to a high degree of redundancy (typically, eight- to tenfold) through the sequencing of randomly selected subfragments. The second is a 'finishing' phase, in which sequence gaps are closed and remaining ambiguities are resolved through directed analysis. The results also showed that complete genomic sequence provided information about genes, regulatory regions and chromosome structure that was not readily obtainable from cDNA studies alone.

In 1995, genome scientists considered a proposal[38] that would have involved producing a draft genome sequence of the human genome in a first phase and then returning to finish the sequence in a second phase. After vigorous debate, it was decided that such a plan was premature for several reasons. These included the need first to prove that high-quality, long-range finished sequence could be produced from most parts of the complex, repeat-rich human genome; the sense that many aspects of the sequencing process were still rapidly evolving; and the desirability of further decreasing costs.

Instead, pilot projects were launched to demonstrate the feasibility of cost-effective, large-scale sequencing, with a target completion date of March 1999. The projects successfully produced finished sequence with 99.99% accuracy and no gaps[39]. They also introduced bacterial artificial chromosomes (BACs)[40], a new large-insert cloning system that proved to be more stable than the cosmids and yeast artificial chromosomes (YACs)[41] that had been used previously. The pilot projects drove the maturation and convergence of sequencing strategies, while producing 15% of the human genome sequence. With successful completion of this phase, the human genome sequencing effort moved into full-scale production in March 1999.

The idea of first producing a draft genome sequence was revived at this time, both because the ability to finish such a sequence was no longer in doubt and because there was great hunger in the scientific community for human sequence data. In addition, some scientists favoured prioritizing the production of a draft genome sequence over regional finished sequence because of concerns about commercial plans to generate proprietary databases of human sequence that might be subject to undesirable restrictions on use[42–44].

The consortium focused on an initial goal of producing, in a first production phase lasting until June 2000, a draft genome sequence covering most of the genome. Such a draft genome sequence, although not completely finished, would rapidly allow investigators to begin to extract most of the information in the human sequence. Experiments showed that sequencing clones covering about 90% of the human genome to a redundancy of about four- to fivefold ('half-shotgun' coverage; see Box 1) would accomplish this[45,46]. The draft genome sequence goal has been achieved, as described below.

The second sequence production phase is now under way. Its aims are to achieve full-shotgun coverage of the existing clones during 2001, to obtain clones to fill the remaining gaps in the physical map, and to produce a finished sequence (apart from regions that cannot be cloned or sequenced with currently available techniques) no later than 2003.

Strategic issues

Hierarchical shotgun sequencing

Soon after the invention of DNA sequencing methods[47,48], the shotgun sequencing strategy was introduced[49–51]; it has remained the fundamental method for large-scale genome sequencing[52–54] for the past 20 years. The approach has been refined and extended to make it more efficient. For example, improved protocols for fragmenting and cloning DNA allowed construction of shotgun libraries with more uniform representation. The practice of sequencing from both ends of double-stranded clones ('double-barrelled' shotgun sequencing) was introduced by Ansorge and others[37] in 1990, allowing the use of 'linking information' between sequence fragments.

The application of shotgun sequencing was also extended by applying it to larger and larger DNA molecules—from plasmids (~4 kilobases (kb)) to cosmid clones[37] (40 kb), to artificial chromosomes cloned in bacteria and yeast[55] (100–500 kb) and bacterial genomes[56] (1–2 megabases (Mb)). In principle, a genome of arbitrary size may be directly sequenced by the shotgun method, provided that it contains no repeated sequence and can be uniformly sampled at random. The genome can then be assembled using the simple computer science technique of 'hashing' (in which one detects overlaps by consulting an alphabetized look-up table of all *k*-letter words in the data). Mathematical analysis of the expected number of gaps as a function of coverage is similarly straightforward[57].

Practical difficulties arise because of repeated sequences and cloning bias. Small amounts of repeated sequence pose little problem for shotgun sequencing. For example, one can readily assemble typical bacterial genomes (about 1.5% repeat) or the euchromatic portion of the fly genome (about 3% repeat). By contrast, the human genome is filled (>50%) with repeated sequences, including interspersed repeats derived from transposable elements, and long genomic regions that have been duplicated in tandem, palindromic or dispersed fashion (see below). These include large duplicated segments (50–500 kb) with high sequence identity (98–99.9%), at which mispairing during recombination creates deletions responsible for genetic syndromes. Such features complicate the assembly of a correct and finished genome sequence.

There are two approaches for sequencing large repeat-rich genomes. The first is a whole-genome shotgun sequencing approach, as has been used for the repeat-poor genomes of viruses, bacteria and flies, using linking information and computational

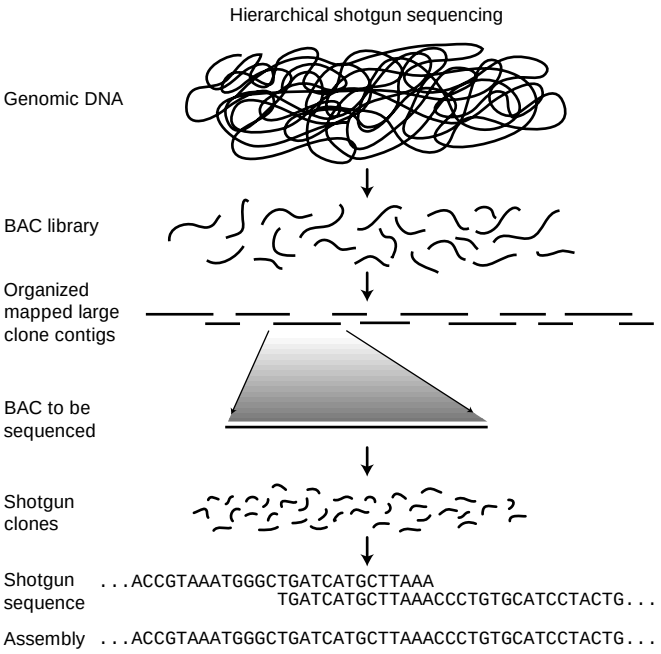


Figure 2 Idealized representation of the hierarchical shotgun sequencing strategy. A library is constructed by fragmenting the target genome and cloning it into a large-fragment cloning vector; here, BAC vectors are shown. The genomic DNA fragments represented in the library are then organized into a physical map and individual BAC clones are selected and sequenced by the random shotgun strategy. Finally, the clone sequences are assembled to reconstruct the sequence of the genome.

© 2001 Macmillan Magazines Ltd

analysis to attempt to avoid misassemblies. The second is the 'hierarchical shotgun sequencing' approach (Fig. 2), also referred to as 'map-based', 'BAC-based' or 'clone-by-clone'. This approach involves generating and organizing a set of large-insert clones (typically 100–200 kb each) covering the genome and separately performing shotgun sequencing on appropriately chosen clones. Because the sequence information is local, the issue of long-range misassembly is eliminated and the risk of short-range misassembly is reduced. One caveat is that some large-insert clones may suffer rearrangement, although this risk can be reduced by appropriate quality-control measures involving clone fingerprints (see below).

The two methods are likely to entail similar costs for producing finished sequence of a mammalian genome. The hierarchical approach has a higher initial cost than the whole-genome approach, owing to the need to create a map of clones (about 1% of the total cost of sequencing) and to sequence overlaps between clones. On the other hand, the whole-genome approach is likely to require much greater work and expense in the final stage of producing a finished sequence, because of the challenge of resolving misassemblies. Both methods must also deal with cloning biases, resulting in under-representation of some regions in either large-insert or small-insert clone libraries.

There was lively scientific debate over whether the human genome sequencing effort should employ whole-genome or hierarchical shotgun sequencing. Weber and Myers[58] stimulated these discussions with a specific proposal for a whole-genome shotgun approach, together with an analysis suggesting that the method could work and be more efficient. Green[59] challenged these conclusions and argued that the potential benefits did not outweigh the likely risks.

In the end, we concluded that the human genome sequencing effort should employ the hierarchical approach for several reasons. First, it was prudent to use the approach for the first project to sequence a repeat-rich genome. With the hierarchical approach, the ultimate frequency of misassembly in the finished product would probably be lower than with the whole-genome approach, in which it would be more difficult to identify regions in which the assembly was incorrect.

Second, it was prudent to use the approach in dealing with an outbred organism, such as the human. In the whole-genome shotgun method, sequence would necessarily come from two different copies of the human genome. Accurate sequence assembly could be complicated by sequence variation between these two copies—both SNPs (which occur at a rate of 1 per 1,300 bases) and larger-scale structural heterozygosity (which has been documented in human chromosomes). In the hierarchical shotgun method, each large-insert clone is derived from a single haplotype.

Third, the hierarchical method would be better able to deal with inevitable cloning biases, because it would more readily allow targeting of additional sequencing to under-represented regions. And fourth, it was better suited to a project shared among members of a diverse international consortium, because it allowed work and responsibility to be easily distributed. As the ultimate goal has always been to create a high-quality, finished sequence to serve as a foundation for biomedical research, we reasoned that the advantages of this more conservative approach outweighed the additional cost, if any.

A biotechnology company, Celera Genomics, has chosen to incorporate the whole-genome shotgun approach into its own efforts to sequence the human genome. Their plan[60,61] uses a mixed strategy, involving combining some coverage with whole-genome shotgun data generated by the company together with the publicly available hierarchical shotgun data generated by the International Human Genome Sequencing Consortium. If the raw sequence reads from the whole-genome shotgun component are made available, it may be possible to evaluate the extent to which the sequence of the human genome can be assembled without the need for clone-based information. Such analysis may help to refine sequencing strategies for other large genomes.

Technology for large-scale sequencing

Sequencing the human genome depended on many technological improvements in the production and analysis of sequence data. Key innovations were developed both within and outside the Human Genome Project. Laboratory innovations included four-colour fluorescence-based sequence detection[62], improved fluorescent dyes[63–66], dye-labelled terminators[67], polymerases specifically designed for sequencing[68–70], cycle sequencing[71] and capillary gel electrophoresis[72–74]. These studies contributed to substantial improvements in the automation, quality and throughput of collecting raw DNA sequence[75,76]. There were also important advances in the development of software packages for the analysis of sequence data. The PHRED software package[77,78] introduced the concept of assigning a 'base-quality score' to each base, on the basis of the probability of an erroneous call. These quality scores make it possible to monitor raw data quality and also assist in determining whether two similar sequences truly overlap. The PHRAP computer package (http://bozeman.mbt.washington.edu/phrap.docs/phrap.html) then systematically assembles the sequence data using the base-quality scores. The program assigns 'assembly-quality scores' to each base in the assembled sequence, providing an objective criterion to guide sequence finishing. The quality scores were based on and validated by extensive experimental data.

Another key innovation for scaling up sequencing was the development by several centres of automated methods for sample preparation. This typically involved creating new biochemical protocols suitable for automation, followed by construction of appropriate robotic systems.

Coordination and public data sharing

The Human Genome Project adopted two important principles with regard to human sequencing. The first was that the collaboration would be open to centres from any nation. Although potentially less efficient, in a narrow economic sense, than a centralized approach involving a few large factories, the inclusive approach was strongly favoured because we felt that the human genome sequence is the common heritage of all humanity and the work should transcend national boundaries, and we believed that scientific progress was best assured by a diversity of approaches. The collaboration was coordinated through periodic international meetings (referred to as 'Bermuda meetings' after the venue of the first three gatherings) and regular telephone conferences. Work was shared flexibly among the centres, with some groups focusing on particular chromosomes and others contributing in a genome-wide fashion.

The second principle was rapid and unrestricted data release. The centres adopted a policy that all genomic sequence data should be made publicly available without restriction within 24 hours of assembly[79,80]. Pre-publication data releases had been pioneered in mapping projects in the worm[11] and mouse genomes[30,81] and were prominently adopted in the sequencing of the worm, providing a direct model for the human sequencing efforts. We believed that scientific progress would be most rapidly advanced by immediate and free availability of the human genome sequence. The explosion of scientific work based on the publicly available sequence data in both academia and industry has confirmed this judgement.

Generating the draft genome sequence

Generating a draft sequence of the human genome involved three steps: selecting the BAC clones to be sequenced, sequencing them and assembling the individual sequenced clones into an overall draft genome sequence. A glossary of terms related to genome sequencing and assembly is provided in Box 1.

The draft genome sequence is a dynamic product, which is regularly updated as additional data accumulate en route to the

© 2001 Macmillan Magazines Ltd

ultimate goal of a completely finished sequence. The results below are based on the map and sequence data available on 7 October 2000, except as otherwise noted. At the end of this section, we provide a brief update of key data.

Clone selection

The hierarchical shotgun method involves the sequencing of overlapping large-insert clones spanning the genome. For the Human Genome Project, clones were largely chosen from eight large-insert libraries containing BAC or P1-derived artificial chromosome (PAC) clones (Table 1; refs 82–88). The libraries were made by partial digestion of genomic DNA with restriction enzymes. Together, they represent around 65-fold coverage (redundant sampling) of the genome. Libraries based on other vectors, such as cosmids, were also used in early stages of the project.

The libraries (Table 1) were prepared from DNA obtained from anonymous human donors in accordance with US Federal Regulations for the Protection of Human Subjects in Research (45CFR46) and following full review by an Institutional Review Board. Briefly, the opportunity to donate DNA for this purpose was broadly advertised near the two laboratories engaged in library

Box 1

Genome glossary

Sequence

Raw sequence Individual unassembled sequence reads, produced by sequencing of clones containing DNA inserts.

Paired-end sequence Raw sequence obtained from both ends of a cloned insert in any vector, such as a plasmid or bacterial artificial chromosome.

Finished sequence Complete sequence of a clone or genome, with an accuracy of at least 99.99% and no gaps.

Coverage (or depth) The average number of times a nucleotide is represented by a high-quality base in a collection of random raw sequence. Operationally, a 'high-quality base' is defined as one with an accuracy of at least 99% (corresponding to a PHRED score of at least 20).

Full shotgun coverage The coverage in random raw sequence needed from a large-insert clone to ensure that it is ready for finishing; this varies among centres but is typically 8–10-fold. Clones with full shotgun coverage can usually be assembled with only a handful of gaps per 100 kb.

Half shotgun coverage Half the amount of full shotgun coverage (typically, 4–5-fold random coverage).

Clones

BAC clone Bacterial artificial chromosome vector carrying a genomic DNA insert, typically 100–200 kb. Most of the large-insert clones sequenced in the project were BAC clones.

Finished clone A large-insert clone that is entirely represented by finished sequence.

Full shotgun clone A large-insert clone for which full shotgun sequence has been produced.

Draft clone A large-insert clone for which roughly half-shotgun sequence has been produced. Operationally, the collection of draft clones produced by each centre was required to have an average coverage of fourfold for the entire set and a minimum coverage of threefold for each clone.

Predraft clone A large-insert clone for which some shotgun sequence is available, but which does not meet the standards for inclusion in the collection of draft clones.

Contigs and scaffolds

Contig The result of joining an overlapping collection of sequences or clones.

Scaffold The result of connecting contigs by linking information from paired-end reads from plasmids, paired-end reads from BACs, known messenger RNAs or other sources. The contigs in a scaffold are ordered and oriented with respect to one another.

Fingerprint clone contigs Contigs produced by joining clones inferred to overlap on the basis of their restriction digest fingerprints.

Sequenced-clone layout Assignment of sequenced clones to the physical map of fingerprint clone contigs.

Initial sequence contigs Contigs produced by merging overlapping sequence reads obtained from a single clone, in a process called sequence assembly.

Merged sequence contigs Contigs produced by taking the initial sequence contigs contained in overlapping clones and merging those found to overlap. These are also referred to simply as 'sequence contigs' where no confusion will result.

Sequence-contig scaffolds Scaffolds produced by connecting sequence contigs on the basis of linking information.

Sequenced-clone contigs Contigs produced by merging overlapping sequenced clones.

Sequenced-clone-contig scaffolds Scaffolds produced by joining sequenced-clone contigs on the basis of linking information.

Draft genome sequence The sequence produced by combining the information from the individual sequenced clones (by creating merged sequence contigs and then employing linking information to create scaffolds) and positioning the sequence along the physical map of the chromosomes.

N50 length A measure of the contig length (or scaffold length) containing a 'typical' nucleotide. Specifically, it is the maximum length L such that 50% of all nucleotides lie in contigs (or scaffolds) of size at least L.

Computer programs and databases

PHRED A widely used computer program that analyses raw sequence to produce a 'base call' with an associated 'quality score' for each position in the sequence. A PHRED quality score of X corresponds to an error probability of approximately $10^{-X/10}$. Thus, a PHRED quality score of 30 corresponds to 99.9% accuracy for the base call in the raw read.

PHRAP A widely used computer program that assembles raw sequence into sequence contigs and assigns to each position in the sequence an associated 'quality score', on the basis of the PHRED scores of the raw sequence reads. A PHRAP quality score of X corresponds to an error probability of approximately $10^{-X/10}$. Thus, a PHRAP quality score of 30 corresponds to 99.9% accuracy for a base in the assembled sequence.

GigAssembler A computer program developed during this project for merging the information from individual sequenced clones into a draft genome sequence.

Public sequence databases The three coordinated international sequence databases: GenBank, the EMBL data library and DDBJ.

Map features

STS Sequence tagged site, corresponding to a short (typically less than 500 bp) unique genomic locus for which a polymerase chain reaction assay has been developed.

EST Expressed sequence tag, obtained by performing a single raw sequence read from a random complementary DNA clone.

SSR Simple sequence repeat, a sequence consisting largely of a tandem repeat of a specific k-mer (such as $(CA)_{15}$). Many SSRs are polymorphic and have been widely used in genetic mapping.

SNP Single nucleotide polymorphism, or a single nucleotide position in the genome sequence for which two or more alternative alleles are present at appreciable frequency (traditionally, at least 1%) in the human population.

Genetic map A genome map in which polymorphic loci are positioned relative to one another on the basis of the frequency with which they recombine during meiosis. The unit of distance is centimorgans (cM), denoting a 1% chance of recombination.

Radiation hybrid (RH) map A genome map in which STSs are positioned relative to one another on the basis of the frequency with which they are separated by radiation-induced breaks. The frequency is assayed by analysing a panel of human–hamster hybrid cell lines, each produced by lethally irradiating human cells and fusing them with recipient hamster cells such that each carries a collection of human chromosomal fragments. The unit of distance is centirays (cR), denoting a 1% chance of a break occuring between two loci.

© 2001 Macmillan Magazines Ltd

construction. Volunteers of diverse backgrounds were accepted on a first-come, first-taken basis. Samples were obtained after discussion with a genetic counsellor and written informed consent. The samples were made anonymous as follows: the sampling laboratory stripped all identifiers from the samples, applied random numeric labels, and transferred them to the processing laboratory, which then removed all labels and relabelled the samples. All records of the labelling were destroyed. The processing laboratory chose samples at random from which to prepare DNA and immortalized cell lines. Around 5–10 samples were collected for every one that was eventually used. Because no link was retained between donor and DNA sample, the identity of the donors for the libraries is not known, even by the donors themselves. A more complete description can be found at http://www.nhgri.nih.gov/Grant_info/Funding/Statements/RFA/human_subjects.html.

During the pilot phase, centres showed that sequence-tagged sites (STSs) from previously constructed genetic and physical maps could be used to recover BACs from specific regions. As sequencing expanded, some centres continued this approach, augmented with additional probes from flow sorting of chromosomes to obtain long-range coverage of specific chromosomes or chromosomal regions[89–94].

For the large-scale sequence production phase, a genome-wide physical map of overlapping clones was also constructed by systematic analysis of BAC clones representing 20-fold coverage of the human genome[86]. Most clones came from the first three sections of the RPCI-11 library, supplemented with clones from sections of the RPCI-13 and CalTech D libraries (Table 1). DNA from each BAC clone was digested with the restriction enzyme *Hin*dIII, and the sizes of the resulting fragments were measured by agarose gel electrophoresis. The pattern of restriction fragments provides a 'fingerprint' for each BAC, which allows different BACs to be distinguished and the degree of overlaps to be assessed. We used these restriction-fragment fingerprints to determine clone overlaps, and thereby assembled the BACs into fingerprint clone contigs.

The fingerprint clone contigs were positioned along the chromosomes by anchoring them with STS markers from existing genetic and physical maps. Fingerprint clone contigs were tied to specific STSs initially by probe hybridization and later by direct search of the sequenced clones. To localize fingerprint clone contigs that did not contain known markers, new STSs were generated and placed onto chromosomes[95]. Representative clones were also positioned by fluorescence *in situ* hybridization (FISH) (ref. 86 and C. McPherson, unpublished).

We selected clones from the fingerprint clone contigs for sequencing according to various criteria. Fingerprint data were reviewed[86,90] to evaluate overlaps and to assess clone fidelity (to bias against rearranged clones[83,96]). STS content information and BAC end sequence information were also used[91,92]. Where possible, we tried to select a minimally overlapping set spanning a region. However, because the genome-wide physical map was constructed concurrently with the sequencing, continuity in many regions was low in early stages. These small fingerprint clone contigs were nonetheless useful in identifying validated, nonredundant clones

Table 1 Key large-insert genome-wide libraries

Library name*	GenBank abbreviation	Vector type	Source DNA	Library segment or plate numbers	Enzyme digest	Average insert size (kb)	Total number of clones in library	Number of fingerprinted clones†	BAC-end sequence (ends/clones/clones with both ends sequenced)‡	Number of clones in genome layout§	Sequenced clones used in construction of the draft genome sequence		
											Number‖	Total bases (Mb)¶	Fraction of total from library
Caltech B	CTB	BAC	987SK cells	All	*Hin*dIII	120	74,496	16	2/1/1	528	518	66.7	0.016
Caltech C	CTC	BAC	Human sperm	All	*Hin*dIII	125	263,040	144	21,956/14,445/7,255	621	606	88.4	0.021
Caltech D1 (CITB-H1)	CTD	BAC	Human sperm	All	*Hin*dIII	129	162,432	49,833	403,589/226,068/156,631	1,381	1,367	185.6	0.043
Caltech D2 (CITB-E1)		BAC	Human sperm	All									
				2,501–2,565	*Eco*RI	202	24,960						
				2,566–2,671	*Eco*RI	182	46,326						
				3,000–3,253	*Eco*RI	142	97,536						
RPCI-1	RP1	PAC	Male, blood	All	*Mbo*I	110	115,200	3,388		1,070	1,053	117.7	0.028
RPCI-3	RP3	PAC	Male, blood	All	*Mbo*I	115	75,513			644	638	68.5	0.016
RPCI-4	RP4	PAC	Male, blood	All	*Mbo*I	116	105,251			889	881	95.5	0.022
RPCI-5	RP5	PAC	Male, blood	All	*Mbo*I	115	142,773			1,042	1,033	116.5	0.027
RPCI-11	RP11	BAC	Male, blood	All		178	543,797	267,931	379,773/243,764/134,110	19,405	19,145	3,165.0	0.743
				1	*Eco*RI	164	108,499						
				2	*Eco*RI	168	109,496						
				3	*Eco*RI	181	109,657						
				4	*Eco*RI	183	109,382						
				5	*Mbo*I	196	106,763						
Total of top eight libraries							1,482,502	321,312	805,320/484,278/297,997	25,580	25,241	3,903.9	0.916
Total all libraries								354,510	812,594/488,017/100,775	30,445	29,298	4,260.5	1

* For the CalTech libraries[82], see http://www.tree.caltech.edu/lib_status.html; for RPCI libraries[83], see http://www.chori.org/bacpac/home.htm.
† For the FPC map and fingerprinting[84–86], see http://genome.wustl.edu/gsc/human/human_database.shtml.
‡ The number of raw BAC end sequences (clones/ends/clones with both ends sequenced) available for use in human genome sequencing. Typically, for clones in which sequence was obtained from both ends, more than 95% of both end sequences contained at least 100 bp of nonrepetitive sequence. BAC-end sequencing of RPCI-11 and of the CalTech libraries was done at The Institute for Genomic Research, the California Institute of Technology and the University of Washington High Throughput Sequencing Center. The sources for the Table were http://www.ncbi.nlm.nih.gov/genome/clone/BESstat.shtml and refs 87, 88.
§ These are the clones in the sequenced-clone layout map (http://genome.wustl.edu/gsc/human/Mapping/index.shtml) that were pre-draft, draft or finished.
‖ The number of sequenced clones used in the assembly. This number is less than that in the previous column owing to removal of a small number of obviously contaminated, combined or duplicated projects; in addition, not all of the clones from completed chromosomes 21 and 22 were included here because only the available finished sequence from those chromosomes was used in the assembly.
¶ The number reported is the total sequence from the clones indicated in the previous column. Potential overlap between clones was not removed here, but Ns were excluded.

© 2001 Macmillan Magazines Ltd

that were used to 'seed' the sequencing of new regions. The small fingerprint clone contigs were extended or merged with others as the map matured.

The clones that make up the draft genome sequence therefore do not constitute a minimally overlapping set—there is overlap and redundancy in places. The cost of using suboptimal overlaps was justified by the benefit of earlier availability of the draft genome sequence data. Minimizing the overlap between adjacent clones would have required completing the physical map before undertaking large-scale sequencing. In addition, the overlaps between BAC clones provide a rich collection of SNPs. More than 1.4 million SNPs have already been identified from clone overlaps and other sequence comparisons[97].

Because the sequencing project was shared among twenty centres in six countries, it was important to coordinate selection of clones across the centres. Most centres focused on particular chromosomes or, in some cases, larger regions of the genome. We also maintained a clone registry to track selected clones and their progress. In later phases, the global map provided an integrated view of the data from all centres, facilitating the distribution of effort to maximize coverage of the genome. Before performing extensive sequencing on a clone, several centres routinely examined an initial sample of 96 raw sequence reads from each subclone library to evaluate possible overlap with previously sequenced clones.

Figure 3 The automated production line for sample preparation at the Whitehead Institute, Center for Genome Research. The system consists of custom-designed factory-style conveyor belt robots that perform all functions from purifying DNA from bacterial cultures through setting up and purifying sequencing reactions.

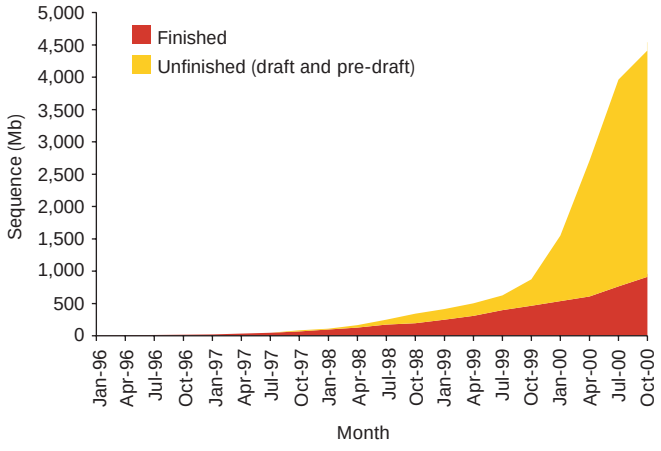


Figure 4 Total amount of human sequence in the High Throughput Genome Sequence (HTGS) division of GenBank. The total is the sum of finished sequence (red) and unfinished (draft plus predraft) sequence (yellow).

Sequencing

The selected clones were subjected to shotgun sequencing. Although the basic approach of shotgun sequencing is well established, the details of implementation varied among the centres. For example, there were differences in the average insert size of the shotgun libraries, in the use of single-stranded or double-stranded cloning vectors, and in sequencing from one end or both ends of each insert. Centres differed in the fluorescent labels employed and in the degree to which they used dye-primers or dye-terminators. The sequence detectors included both slab gel- and capillary-based devices. Detailed protocols are available on the web sites of many of the individual centres (URLs can be found at www.nhgri.nih.gov/genome_hub). The extent of automation also varied greatly among the centres, with the most aggressive automation efforts resulting in factory-style systems able to process more than 100,000 sequencing reactions in 12 hours (Fig. 3). In addition, centres differed in the amount of raw sequence data typically obtained for each clone (so-called half-shotgun, full shotgun and finished sequence). Sequence information from the different centres could be directly integrated despite this diversity, because the data were analysed by a common computational procedure. Raw sequence traces were processed and assembled with the PHRED and PHRAP software packages[77,78] (P. Green, unpublished). All assembled contigs of more than 2 kb were deposited in public databases within 24 hours of assembly.

The overall sequencing output rose sharply during production (Fig. 4). Following installation of new sequence detectors beginning in June 1999, sequencing capacity and output rose approximately eightfold in eight months to nearly 7 million samples processed per month, with little or no drop in success rate (ratio of useable reads to attempted reads). By June 2000, the centres were producing raw sequence at a rate equivalent to onefold coverage of the entire human genome in less than six weeks. This corresponded to a continuous throughput exceeding 1,000 nucleotides per second, 24 hours per day, seven days per week. This scale-up resulted in a concomitant increase in the sequence available in the public databases (Fig. 4).

A version of the draft genome sequence was prepared on the basis of the map and sequence data available on 7 October 2000. For this version, the mapping effort had assembled the fingerprinted BACs into 1,246 fingerprint clone contigs. The sequencing effort had sequenced and assembled 29,298 overlapping BACs and other large-insert clones (Table 2), comprising a total length of 4.26 gigabases (Gb). This resulted from around 23 Gb of underlying raw shotgun sequence data, or about 7.5-fold coverage averaged across the genome (including both draft and finished sequence). The various contributions to the total amount of sequence deposited in the HTGS division of GenBank are given in Table 3.

Table 2 Total genome sequence from the collection of sequenced clones, by sequence status

Sequence status	Number of clones	Total clone length (Mb)	Average number of sequence reads per kb*	Average sequence depth†	Total amount of raw sequence (Mb)
Finished	8,277	897	20–25	8–12	9,085
Draft	18,969	3,097	12	4.5	13,395
Predraft	2,052	267	6	2.5	667
Total					23,147

*The average number of reads per kb was estimated based on information provided by each sequencing centre. This number differed among sequencing centres, based on the actual protocols used.

†The average depth in high quality bases (≥99% accuracy) was estimated from information provided by each sequencing centre. The average varies among the centres, and the number may vary considerably for clones with the same sequencing status. For draft clones in the public databases (keyword: HTGS_draft), the number can be computed from the quality scores listed in the database entry.

© 2001 Macmillan Magazines Ltd

By agreement among the centres, the collection of draft clones produced by each centre was required to have fourfold average sequence coverage, with no clone below threefold. (For this purpose, sequence coverage was defined as the average number of times that each base was independently read with a base-quality score corresponding to at least 99% accuracy.) We attained an overall average of 4.5-fold coverage across the genome for draft clones. A few of the sequenced clones fell below the minimum of threefold sequence coverage or have not been formally designated by centres as meeting draft standards; these are referred to as predraft (Table 2). Some of these are clones that span remaining gaps in the draft genome sequence and were in the process of being sequenced on 7 October 2000; a few are old submissions from centres that are no longer active.

The lengths of the initial sequence contigs in the draft clones vary as a function of coverage, but half of all nucleotides reside in initial sequence contigs of at least 21.7 kb (see below). Various properties of the draft clones can be assessed from instances in which there was substantial overlap between a draft clone and a finished (or nearly finished) clone. By examining the sequence alignments in the overlap regions, we estimated that the initial sequence contigs in a draft sequence clone cover an average of about 96% of the clone and are separated by gaps with an average size of about 500 bp.

Although the main emphasis was on producing a draft genome sequence, the centres also maintained sequence finishing activities during this period, leading to a twofold increase in finished sequence from June 1999 to June 2000 (Fig. 4). The total amount of human sequence in this final form stood at more than 835 Mb on 7 October 2000, or more than 25% of the human genome. This includes the finished sequences of chromosomes 21 and 22 (refs 93, 94). As centres have begun to shift from draft to finished sequencing in the last quarter of 2000, the production of finished sequence has increased to an annualized rate of 1 Gb per year and is continuing to rise.

Table 3 Total human sequence deposited in the HTGS division of GenBank

Sequencing centre	Total human sequence (kb)	Finished human sequence (kb)
Whitehead Institute, Center for Genome Research*	1,196,888	46,560
The Sanger Centre*	970,789	284,353
Washington University Genome Sequencing Center*	765,898	175,279
US DOE Joint Genome Institute	377,998	78,486
Baylor College of Medicine Human Genome Sequencing Center	345,125	53,418
RIKEN Genomic Sciences Center	203,166	16,971
Genoscope	85,995	48,808
GTC Sequencing Center	71,357	7,014
Department of Genome Analysis, Institute of Molecular Biotechnology	49,865	17,788
Beijing Genomics Institute/Human Genome Center	42,865	6,297
Multimegabase Sequencing Center; Institute for Systems Biology	31,241	9,676
Stanford Genome Technology Center	29,728	3,530
The Stanford Human Genome Center and Department of Genetics	28,162	9,121
University of Washington Genome Center	24,115	14,692
Keio University	17,364	13,058
University of Texas Southwestern Medical Center at Dallas	11,670	7,028
University of Oklahoma Advanced Center for Genome Technology	10,071	9,155
Max Planck Institute for Molecular Genetics	7,650	2,940
GBF – German Research Centre for Biotechnology	4,639	2,338
Cold Spring Harbor Laboratory Lita Annenberg Hazen Genome Center	4,338	2,104
Other	59,574	35,911
Total	4,338,224	842,027

Total human sequence deposited in GenBank by members of the International Human Genome Sequencing Consortium, as of 8 October 2000.The amount of total sequence (finished plus draft plus predraft) is shown in the second column and the amount of finished sequence is shown in the third column. Total sequence differs from totals in Tables 1 and 2 because of inclusion of padding characters and of some clones not used in assembly. HTGS, high throughput genome sequence.
*These three centres produced an additional 2.4 Gb of raw plasmid paired-end reads (see Table 4), consisting of 0.99 Gb from Whitehead Institute, 0.66 Gb from The Sanger Centre and 0.75 Gb from Washington University.

In addition to sequencing large-insert clones, three centres generated a large collection of random raw sequence reads from whole-genome shotgun libraries (Table 4; ref. 98). These 5.77 million successful sequences contained 2.4 Gb of high-quality bases; this corresponds to about 0.75-fold coverage and would be statistically expected to include about 50% of the nucleotides in the human genome (data available at http://snp.cshl.org/data). The primary objective of this work was to discover SNPs, by comparing these random raw sequences (which came from different individuals) with the draft genome sequence. However, many of these raw sequences were obtained from both ends of plasmid clones and thereby also provided valuable 'linking' information that was used in sequence assembly. In addition, the random raw sequences provide sequence coverage of about half of the nucleotides not yet represented in the sequenced large-insert clones; these can be used as probes for portions of the genome not yet recovered.

Assembly of the draft genome sequence

We then set out to assemble the sequences from the individual large-insert clones into an integrated draft sequence of the human genome. The assembly process had to resolve problems arising from the draft nature of much of the sequence, from the variety of clone sources, and from the high fraction of repeated sequences in the human genome. This process involved three steps: filtering, layout and merging.

The entire data set was filtered uniformly to eliminate contamination from nonhuman sequences and other artefacts that had not already been removed by the individual centres. (Information about contamination was also sent back to the centres, which are updating the individual entries in the public databases.) We also identified instances in which the sequence data from one BAC clone was substantially contaminated with sequence data from another (human or nonhuman) clone. The problems were resolved in most instances; 231 clones remained unresolved, and these were eliminated from the assembly reported here. Instances of lower levels of cross-contamination (for example, a single 96-well microplate misassigned to the wrong BAC) are more difficult to detect; some undoubtedly remain and may give rise to small spurious sequence contigs in the draft genome sequence. Such issues are readily resolved as the clones progress towards finished sequence, but they necessitate some caution in certain applications of the current data.

The sequenced clones were then associated with specific clones on the physical map to produce a 'layout'. In principle, sequenced clones that correspond to fingerprinted BACs could be directly assigned by name to fingerprint clone contigs on the fingerprint-based physical map. In practice, however, laboratory mixups occasionally resulted in incorrect assignments. To eliminate such problems, sequenced clones were associated with the fingerprint clone contigs in the physical map by using the sequence data to calculate a

Table 4 Plasmid paired-end reads

	Total reads deposited*	Read pairs†	Size range of inserts (kb)
Random-sheared	3,227,685	1,155,284	1.8–6
Enzyme digest	2,539,222	761,010	0.8–4.7
Total	5,766,907	1,916,294	

The plasmid paired-end reads used a mixture of DNA from a set of 24 samples from the DNA Polymorphism Discovery Resource (http://locus.umdnj.edu/nigms/pdr.html). This set of 24 anonymous US residents contains samples from European-Americans, African-Americans, Mexican-Americans, Native Americans and Asian-Americans, although the ethnicities of the individual samples are not identified. Informed consent to contribute samples to the DNA Polymorphism Discovery Resource was obtained from all 450 individuals who contributed samples. Samples from the European-American, African-American and Mexican-American individuals came from NHANES (http://www.cdc.gov/nchs/nhanes.htm); individuals were recontacted to obtain their consent for the Resource project. New samples were obtained from Asian-Americans whose ancestry was from a variety of East and South Asian countries. New samples were also obtained for the Native Americans; tribal permission was obtained first, and then individual consents. See http://www.nhgri.nih.gov/Grant_info/Funding/RFA/discover_polymorphisms.html and ref. 98.
*Reflects data deposited with and released by The SNP Consortium (see http://snp.cshl.org/data).
† Read pairs represents the number of cases in which sequence from both ends of a genomic cloned fragment was determined and used in this study as linking information.

© 2001 Macmillan Magazines Ltd

partial list of restriction fragments *in silico* and comparing that list with the experimental database of BAC fingerprints. The comparison was feasible because the experimental sizing of restriction fragments was highly accurate (to within 0.5–1.5% of the true size, for 95% of fragments from 600 to 12,000 base pairs (bp))[84,85]. Reliable matching scores could be obtained for 16,193 of the clones. The remaining sequenced clones could not be placed on the map by this method because they were too short, or they contained too many small initial sequence contigs to yield enough restriction fragments, or possibly because their sequences were not represented in the fingerprint database.

An independent approach to placing sequenced clones on the physical map used the database of end sequences from fingerprinted BACs (Table 1). Sequenced clones could typically be reliably mapped if they contained multiple matches to BAC ends, with all corresponding to clones from a single genomic region (multiple matches were required as a safeguard against errors known to exist in the BAC end database and against repeated sequences). This approach provided useful placement information for 22,566 sequenced clones.

Altogether, we could assign 25,403 sequenced clones to fingerprint clone contigs by combining *in silico* digestion and BAC end sequence match data. To place most of the remaining sequenced clones, we exploited information about sequence overlap or BAC-end paired links of these clones with already positioned clones. This left only a few, mostly small, sequenced clones that could not be placed (152 sequenced clones containing 5.5 Mb of sequence out of 29,298 sequenced clones containing more than 4,260 Mb of sequence); these are being localized by radiation hybrid mapping of STSs derived from their sequences.

The fingerprint clone contigs were then mapped to chromosomal locations, using sequence matches to mapped STSs from four human radiation hybrid maps[95,99,100], one YAC and radiation hybrid map[29], and two genetic maps[101,102], together with data from FISH[86,90,103]. The mapping was iteratively refined by comparing the order and orientation of the STSs in the fingerprint clone contigs and the various STS-based maps, to identify and refine discrepancies (Fig. 5). Small fingerprint clone contigs (< 1 Mb) were difficult to orient and, sometimes, to order using these methods. In all, 942 fingerprint clone contigs contained sequenced clones. (An additional 304 of the 1,246 fingerprint clone contigs did not contain sequenced clones, but these tended to be extremely small and together contain less than 1% of the mapped clones. About one-third have been targeted for sequencing. A few derive from the Y chromosome, for which the map was constructed separately[89]. Most of the remainder are fragments of other larger contigs or represent other artefacts. These are being eliminated in subsequent versions of the database.) Of these 942 contigs with sequenced clones, 852 (90%, containing 99.2% of the total sequence) were localized to specific chromosome locations in this way. An additional 51 fingerprint clone contigs, containing 0.5% of the sequence, could be assigned to a specific chromosome but not to a precise position.

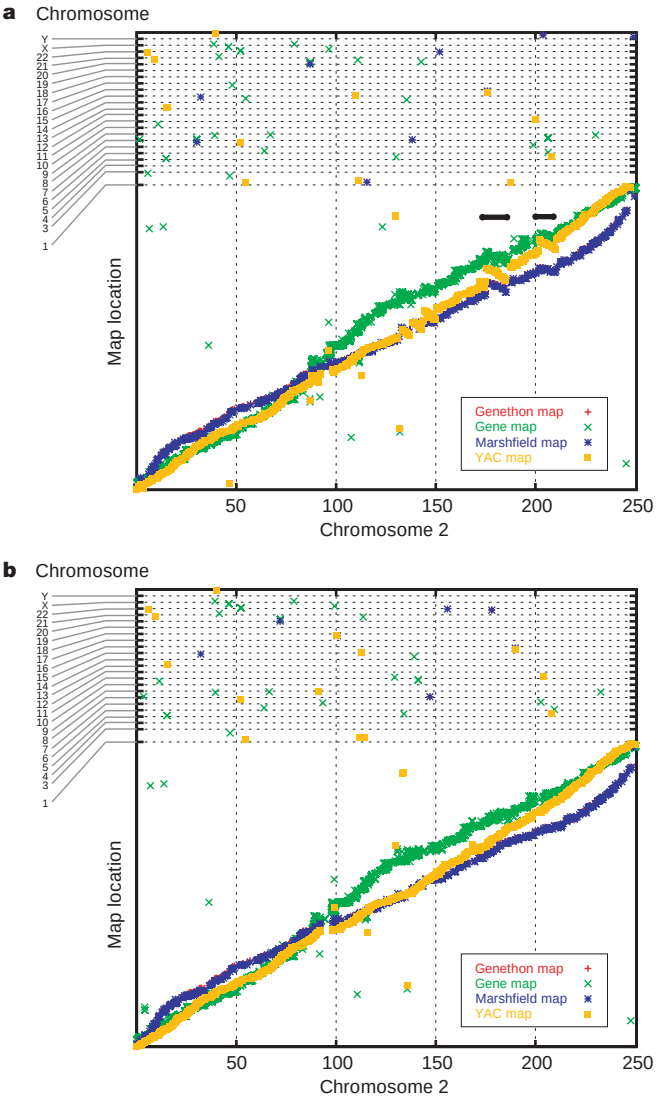


Figure 5 Positions of markers on previous maps of the genome (the Genethon[101] genetic map and Marshfield genetic map (http://research.marshfieldclinic.org/genetics/genotyping_service/mgsver2.htm), the GeneMap99 radiation hybrid map[100], and the Whitehead YAC and radiation hybrid map[29]) plotted against their derived position on the draft sequence for chromosome 2. The horizontal units are Mb but the vertical units of each map vary (cM, cR and so on) and thus all were scaled so that the entire map spans the full vertical range. Markers that map to other chromosomes are shown in the chromosome lines at the top.The data sets generally follow the diagonal, indicating that order and orientation of the marker sets on the different maps largely agree (note that the two genetic maps are completely superimposed). In **a**, there are two segments (bars) that are inverted in an earlier version draft sequence relative to all the other maps. **b,** The same chromosome after the information was used to reorient those two segments.

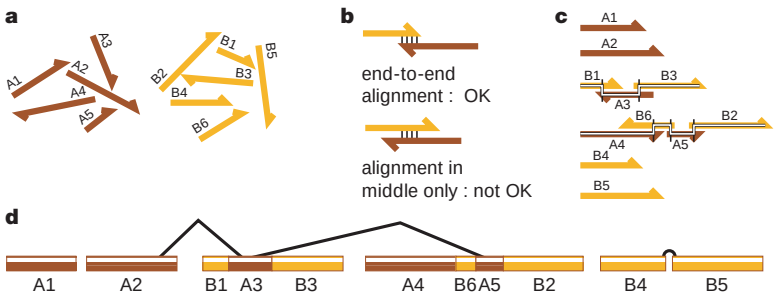


Figure 6 The key steps (**a–d**) in assembling individual sequenced clones into the draft genome sequence. A1–A5 represent initial sequence contigs derived from shotgun sequencing of clone A, and B1–B6 are from clone B.

© 2001 Macmillan Magazines Ltd

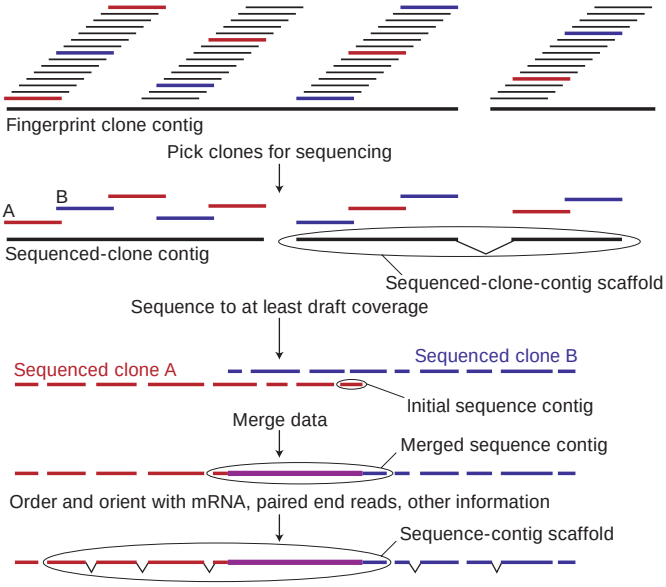


Figure 7 Levels of clone and sequence coverage. A 'fingerprint clone contig' is assembled by using the computer program FPC[84,451] to analyse the restriction enzyme digestion patterns of many large-insert clones. Clones are then selected for sequencing to minimize overlap between adjacent clones. For a clone to be selected, all of its restriction enzyme fragments (except the two vector-insert junction fragments) must be shared with at least one of its neighbours on each side in the contig. Once these overlapping clones have been sequenced, the set is a 'sequenced-clone contig'. When all selected clones from a fingerprint clone contig have been sequenced, the sequenced-clone contig will be the same as the fingerprint clone contig. Until then, a fingerprint clone contig may contain several sequenced-clone contigs. After individual clones (for example, A and B) have been sequenced to draft coverage and the clones have been mapped, the data are analysed by GigAssembler (Fig. 6), producing merged sequence contigs from initial sequence contigs, and linking these to form sequence-contig scaffolds (see Box 1).

The remaining 39 contigs containing 0.3% of the sequence were not positioned at all.

We then merged the sequences from overlapping sequenced clones (Fig. 6), using the computer program GigAssembler[104]. The program considers nearby sequenced clones, detects overlaps between the initial sequence contigs in these clones, merges the overlapping sequences and attempts to order and orient the sequence contigs. It begins by aligning the initial sequence contigs from one clone with those from other clones in the same fingerprint clone contig on the basis of length of alignment, per cent identity of the alignment, position in the sequenced clone layout and other factors. Alignments are limited to one end of each initial sequence contig for partially overlapping contigs or to both ends of an initial sequence contig contained entirely within another; this eliminates internal alignments that may reflect repeated sequence or possible misassembly (Fig. 6b). Beginning with the highest scoring pairs, initial sequence contigs are then integrated to produce 'merged sequence contigs' (usually referred to simply as 'sequence contigs'). The program refines the arrangement of the clones within the fingerprint clone contig on the basis of the extent of sequence overlap between them and then rebuilds the sequence contigs. Next, the program selects a sequence path through the sequence contigs (Fig. 6c). It tries to use the highest quality data by preferring longer initial sequence contigs and avoiding the first and last 250 bases of initial sequence contigs where possible. Finally, it attempts to order and orient the sequence contigs by using additional information, including sequence data from paired-end plasmid and BAC reads, known messenger RNAs and ESTs, as well as additional linking information provided by centres. The sequence contigs are thereby linked together to create 'sequence-contig scaffolds' (Fig. 6d). The process also joins overlapping sequenced clones into sequenced-clone contigs and links sequenced-clone contigs to form sequenced-clone-contig scaffolds. A fingerprint clone contig may contain several sequenced-clone contigs, because bridging clones remain to be sequenced. The assembly contained 4,884 sequenced-clone

Table 5 The draft genome sequence

Chromosome	Sequence from clones (kb)			Sequence from contigs (kb)		
	Finished clones	Draft clones	Pre-draft clones	Contigs containing finished clones	Deep coverage sequence contigs	Draft/predraft sequence contigs
All	826,441	1,734,995	131,476	958,922	840,815	893,175
1	50,851	149,027	12,356	61,001	78,773	72,461
2	46,909	167,439	7,210	53,775	81,569	86,214
3	22,350	152,840	11,057	26,959	79,649	79,638
4	15,914	134,973	17,261	19,096	66,165	82,887
5	37,973	129,581	2,160	48,895	61,387	59,431
6	75,312	76,082	6,696	93,458	28,204	36,428
7	94,845	47,328	4,047	103,188	14,434	28,597
8	14,538	102,484	7,236	16,659	47,198	60,400
9	18,401	77,648	10,864	24,030	42,653	40,230
10	16,889	99,181	11,066	21,421	54,054	51,662
11	13,162	111,092	4,352	16,145	65,147	47,314
12	32,156	84,653	7,651	37,519	43,995	42,946
13	16,818	68,983	7,136	22,191	38,319	32,429
14	58,989	27,370	565	78,302	3,267	5,355
15	2,739	67,453	3,211	3,112	34,758	35,533
16	22,987	48,997	1,143	27,751	20,892	24,484
17	29,881	36,349	6,600	33,531	14,671	24,628
18	5,128	65,284	2,352	6,656	40,947	25,160
19	28,481	26,568	369	32,228	7,188	16,003
20	54,217	5,302	976	56,534	1,065	2,896
21	33,824	0	0	33,824	0	0
22	33,786	0	0	33,786	0	0
X	77,630	45,100	4,941	83,796	14,056	29,820
Y	18,169	3,221	363	20,222	333	1,198
NA	2,434	1,858	844	2,446	122	2,568
UL	2,056	6,182	1,020	2,395	1,969	4,894

The table presents summary statistics for the draft genome sequence over the entire genome and by individual chromosome. NA, clones that could not be placed into the sequenced clone layout. UL, clones that could be placed in the layout, but that could not reliably be placed on a chromosome. First three columns, data from finished clones, draft clones and predraft clones. The last three columns break the data down according to the type of sequence contig. Contigs containing finished clones represent sequence contigs that consist of finished sequence plus any (small) extensions from merged sequence contigs that arise from overlap with flanking draft clones. Deep coverage sequence contigs include sequence from two or more overlapping unfinished clones; they consist of roughly full shotgun coverage and thus are longer than the average unfinished sequence contig. Draft/predraft sequence contigs are all of the other sequence contigs in unfinished clones. Thus, the draft genome sequence consists of approximately one-third finished sequence, one-third deep coverage sequence and one-third draft/pre-draft coverage sequence. In all of the statistics, we count only nonoverlapping bases in the draft genome sequence.

© 2001 Macmillan Magazines Ltd

contigs in 942 fingerprint clone contigs.

The hierarchy of contigs is summarized in Fig. 7. Initial sequence contigs are integrated to create merged sequence contigs, which are then linked to form sequence-contig scaffolds. These scaffolds reside within sequenced-clone contigs, which in turn reside within fingerprint clone contigs.

The draft genome sequence

The result of the assembly process is an integrated draft sequence of the human genome. Several features of the draft genome sequence are reported in Tables 5–7, including the proportion represented by finished, draft and predraft categories. The Tables also show the numbers and lengths of different types of contig, for each chromosome and for the genome as a whole.

The contiguity of the draft genome sequence at each level is an important feature. Two commonly used statistics have significant drawbacks for describing contiguity. The 'average length' of a contig is deflated by the presence of many small contigs comprising only a small proportion of the genome, whereas the 'length-weighted average length' is inflated by the presence of large segments of finished sequence. Instead, we chose to describe the contiguity as a property of the 'typical' nucleotide. We used a statistic called the 'N50 length', defined as the largest length L such that 50% of all nucleotides are contained in contigs of size at least L.

The continuity of the draft genome sequence reported here and the effectiveness of assembly can be readily seen from the following: half of all nucleotides reside within an initial sequence contig of at least 21.7 kb, a sequence contig of at least 82 kb, a sequence-contig scaffold of at least 274 kb, a sequenced-clone contig of at least 826 kb and a fingerprint clone contig of at least 8.4 Mb (Tables 6, 7). The cumulative distributions for each of these measures of contiguity are shown in Fig. 8, in which the N50 values for each measure can be seen as the value at which the cumulative distributions cross 50%. We have also estimated the size of each chromosome, by estimating the gap sizes (see below) and the extent of missing heterochromatic sequence[93,94,105–108] (Table 8). This is undoubtedly an oversimplification and does not adequately take into account the sequence status of each chromosome. Nonetheless, it provides a useful way to relate the draft sequence to the chromosomes.

Quality assessment

The draft genome sequence already covers the vast majority of the genome, but it remains an incomplete, intermediate product that is regularly updated as we work towards a complete finished sequence. The current version contains many gaps and errors. We therefore sought to evaluate the quality of various aspects of the current draft genome sequence, including the sequenced clones themselves, their assignment to a position in the fingerprint clone contigs, and the assembly of initial sequence contigs from the individual clones into sequence-contig scaffolds.

Nucleotide accuracy is reflected in a PHRAP score assigned to each base in the draft genome sequence and available to users through the Genome Browsers (see below) and public database entries. A summary of these scores for the unfinished portion of the genome is shown in Table 9. About 91% of the unfinished draft genome sequence has an error rate of less than 1 per 10,000 bases (PHRAP score >40), and about 96% has an error rate of less than 1 in 1,000 bases (PHRAP >30). These values are based only on the quality scores for the bases in the sequenced clones; they do not reflect additional confidence in the sequences that are represented in overlapping clones. The finished portion of the draft genome sequence has an error rate of less than 1 per 10,000 bases.

Individual sequenced clones. We assessed the frequency of misassemblies, which can occur when the assembly program PHRAP joins two nonadjacent regions in the clone into a single initial sequence contig. The frequency of misassemblies depends heavily on the depth and quality of coverage of each clone and the nature of the underlying sequence; thus it may vary among genomic regions and among individual centres. Most clone misassemblies are readily corrected as coverage is added during finishing, but they may have been propagated into the current version of the draft genome sequence and they justify caution for certain applications.

We estimated the frequency of misassembly by examining instances in which there was substantial overlap between a draft clone and a finished clone. We studied 83 Mb of such overlaps, involving about 9,000 initial sequence contigs. We found 5.3 instances per Mb in which the alignment of an initial sequence contig to the finished sequence failed to extend to within 200 bases

Table 6 Clone level contiguity of the draft genome sequence

Chromosome	Sequenced-clone contigs		Sequenced-clone-contig scaffolds		Fingerprint clone contigs with sequence	
	Number	N50 length (kb)	Number	N50 length (kb)	Number	N50 length (kb)
All	4,884	826	2,191	2,279	942	8,398
1	453	650	197	1,915	106	3,537
2	348	1,028	127	3,140	52	10,628
3	409	672	201	1,550	73	5,077
4	384	606	163	1,659	41	6,918
5	385	623	164	1,642	48	5,747
6	292	814	98	3,292	17	24,680
7	224	1.074	86	3,527	29	20,401
8	292	542	115	1,742	43	6,236
9	143	1,242	78	2,411	21	29,108
10	179	1,097	105	1,952	16	30,284
11	224	887	89	3,024	31	9,414
12	196	1,138	76	2,717	28	9,546
13	128	1,151	56	3,257	13	25,256
14	54	3,079	27	8,489	14	22,128
15	123	797	56	2,095	19	8,274
16	159	620	92	1,317	57	2,716
17	138	831	58	2,138	43	2,816
18	137	709	47	2,572	24	4,887
19	159	569	79	1,200	51	1,534
20	42	2,318	20	6,862	9	23,489
21	5	28,515	5	28,515	5	28,515
22	11	23,048	11	23,048	11	23,048
X	325	572	181	1,082	143	1,436
Y	27	1,539	20	3,290	8	5,135
UL	47	227	40	281	40	281

Number and size of sequenced-clone contigs, sequenced-clone-contig scaffolds and those fingerprint clone contigs (see Box 1) that contain sequenced clones; some small fingerprint clone contigs do not as yet have associated sequence. UL, fingerprint clone contigs that could not reliably be placed on a chromosome. These length estimates are from the draft genome sequence, in which gaps between sequence contigs are arbitrarily represented with 100 Ns and gaps between sequence clone contigs with 50,000 Ns for 'bridged gaps' and 100,000 Ns for 'unbridged gaps'. These arbitrary values differ minimally from empirical estimates of gap size (see text), and using the empirically derived estimates would change the N50 lengths presented here only slightly. For unfinished chromosomes, the N50 length ranges from 1.5 to 3 times the arithmetic mean for sequenced-clone contigs, 1.5 to 3 times for sequenced-clone-contig scaffolds, and 1.5 to 6 times for fingerprint clone contigs with sequence.

© 2001 Macmillan Magazines Ltd

of the end of the contig, suggesting a possible false join in the assembly of the initial sequence contig. In about half of these cases, the potential misassembly involved fewer than 400 bases, suggesting that a single raw sequence read may have been incorrectly joined. We found 1.9 instances per Mb in which the alignment showed an internal gap, again suggesting a possible misassembly; and 0.5 instances per Mb in which the alignment indicated that two initial sequence contigs that overlapped by at least 150 bp had not been merged by PHRAP. Finally, there were another 0.9 instances per Mb with various other problems. This gives a total of 8.6 instances per Mb of possible misassembly, with about half being relatively small issues involving a few hundred bases.

Some of the potential problems might not result from misassembly, but might reflect sequence polymorphism in the population, small rearrangements during growth of the large-insert clones, regions of low-quality sequence or matches between segmental duplications. Thus, the frequency of misassemblies may be overstated. On the other hand, the criteria for recognizing overlap between draft and finished clones may have eliminated some misassemblies.

Layout of the sequenced clones. We assessed the accuracy of the layout of sequenced clones onto the fingerprinted clone contigs by calculating the concordance between the positions assigned to a sequenced clone on the basis of *in silico* digestion and the position assigned on the basis of BAC end sequence data. The positions agreed in 98% of cases in which independent assignments could be made by both methods. The results were also compared with well studied regions containing both finished and draft genome sequence. These results indicated that sequenced clone order in the fingerprint map was reliable to within about half of one clone length (~100 kb).

A direct test of the layout is also provided by the draft genome sequence assembly itself. With extensive coverage of the genome, a correctly placed clone should usually (although not always) show sequence overlap with its neighbours in the map. We found only 421 instances of 'singleton' clones that failed to overlap a neighbouring clone. Close examination of the data suggests that most of these are correctly placed, but simply do not yet overlap an adjacent sequenced clone. About 150 clones appeared to be candidates for being incorrectly placed.

Alignment of the fingerprint clone contigs. The alignment of the fingerprint clone contigs with the chromosomes was based on the radiation hybrid, YAC and genetic maps of STSs. The positions of most of the STSs in the draft genome sequence were consistent with these previous maps, but the positions of about 1.7% differed from one or more of them. Some of these disagreements may be due to errors in the layout of the sequenced clones or in the underlying

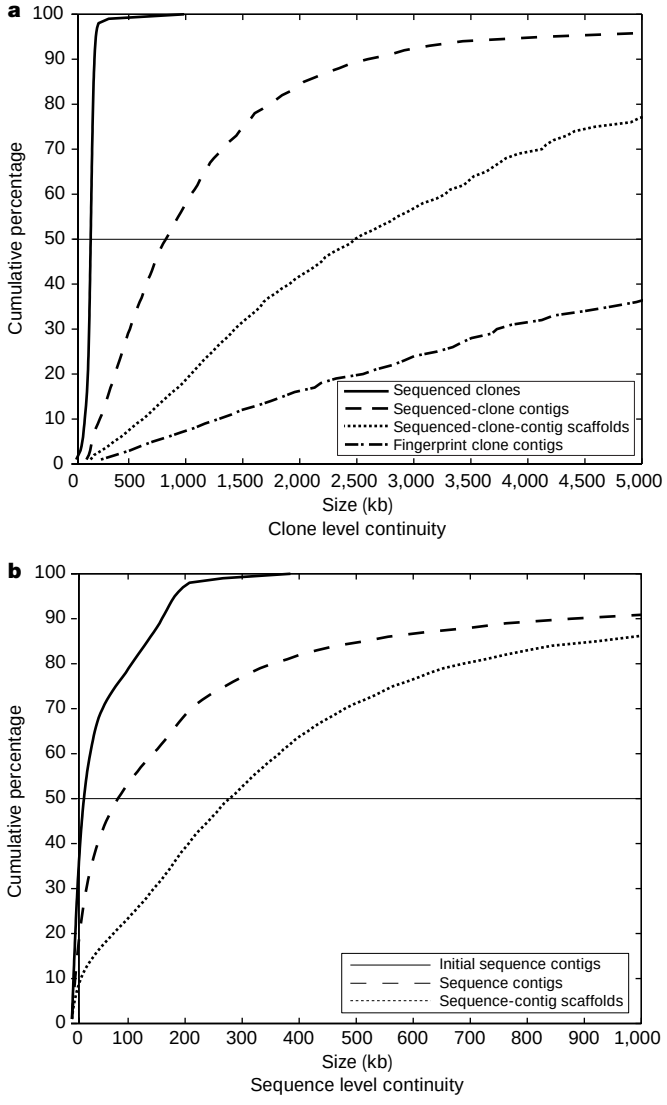


Figure 8 Cumulative distributions of several measures of clone level contiguity and sequence contiguity. The figures represent the proportion of the draft genome sequence contained in contigs of at most the indicated size. **a**, Clone level contiguity. The clones have a tight size distribution with an N50 of ~160 kb (corresponding to 50% on the cumulative distribution). Sequenced-clone contigs represent the next level of continuity, and are linked by mRNA sequences or pairs of BAC end sequences to yield the sequenced-clone-contig scaffolds. The underlying contiguity of the layout of sequenced clones against the fingerprinted clone contigs is only partially shown at this scale. **b**, Sequence contiguity. The input fragments have low continuity (N50 = 21.7 kb). After merging, the sequence contigs grow to an N50 length of about 82 kb. After linking, sequence-contig scaffolds with an N50 length of about 274 kb are created.

Figure 9 Overview of features of draft human genome. The Figure shows the occurrences of twelve important types of feature across the human genome. Large grey blocks represent centromeres and centromeric heterochromatin (size not precisely to scale). Each of the feature types is depicted in a track, from top to bottom as follows. (1) Chromosome position in Mb. (2) The approximate positions of Giemsa-stained chromosome bands at the 800 band resolution. (3) Level of coverage in the draft genome sequence. Red, areas covered by finished clones; yellow, areas covered by predraft sequence. Regions covered by draft sequenced clones are in orange, with darker shades reflecting increasing shotgun sequence coverage. (4) GC content. Percentage of bases in a 20,000 base window that are C or G. (5) Repeat density. Red line, density of SINE class repeats in a 100,000-base window; blue line, density of LINE class repeats in a 100,000-base window. (6) Density of SNPs in a 50,000-base window. The SNPs were detected by sequencing and alignments of random genomic reads. Some of the heterogeneity in SNP density reflects the methods used for SNP discovery. Rigorous analysis of SNP density requires comparing the number of SNPs identified to the precise number of bases surveyed. (7) Non-coding RNA genes. Brown, functional RNA genes such as tRNAs, snoRNAs and rRNAs; light orange, RNA pseudogenes. (8) CpG islands. Green ticks represent regions of ~200 bases with CpG levels significantly higher than in the genome as a whole, and GC ratios of at least 50%. (9) Exofish ecores. Regions of homology with the pufferfish *T. nigroviridis*[292] are blue. (10) ESTs with at least one intron when aligned against genomic DNA are shown as black tick marks. (11) The starts of genes predicted by Genie or Ensembl are shown as red ticks. The starts of known genes from the RefSeq database[110] are shown in blue. (12) The names of genes that have been uniquely located in the draft genome sequence, characterized and named by the HGM Nomenclature Committee. Known disease genes from the OMIM database are red, other genes blue. This Figure is based on an earlier version of the draft genome sequence than analysed in the text, owing to production constraints. We are aware of various errors in the Figure, including omissions of some known genes and misplacements of others. Some genes are mapped to more than one location, owing to errors in assembly, close paralogues or pseudogenes. Manual review was performed to select the most likely location in these cases and to correct other regions. For updated information, see http://genome.ucsc.edu/ and http://www.ensembl.org/.

© 2001 Macmillan Magazines Ltd

Table 7 Sequence level contiguity of the draft genome sequence

Chromosome	Initial sequence contigs		Sequence contigs		Sequence-contig scaffolds	
	Number	N50 length (kb)	Number	N50 length (kb)	Number	N50 length (kb)
All	396,913	21.7	149,821	81.9	87,757	274.3
1	37,656	16.5	12,256	59.1	5,457	278.4
2	32,280	19.9	13,228	57.3	6,959	248.5
3	38,848	15.6	15,098	37.7	8,964	167.4
4	28,600	16.0	13,152	33.0	7,402	158.9
5	30,096	20.4	10,689	72.9	6,378	241.2
6	17,472	43.6	5,547	180.3	2,554	485.0
7	12,733	86.4	4,562	335.7	2,726	591.3
8	19,042	18.1	8,984	38.2	4,631	198.9
9	15,955	20.1	6,226	55.6	3,766	216.2
10	21,762	18.7	9,126	47.9	6,886	133.0
11	29,723	14.3	8,503	40.0	4,684	193.2
12	22,050	19.1	8,422	63.4	5,526	217.0
13	13,737	21.7	5,193	70.5	2,659	300.1
14	4,470	161.4	829	1,371.0	541	2,009.5
15	13,134	15.3	5,840	30.3	3,229	149.7
16	10,297	34.4	4,916	119.5	3,337	356.3
17	10,369	22.9	4,339	90.6	2,616	248.9
18	16,266	15.3	4,461	51.4	2,540	216.1
19	6,009	38.4	2,503	134.4	1,551	375.5
20	2,884	108.6	511	1,346.7	312	813.8
21	103	340.0	5	28,515.3	5	28,515.3
22	526	113.9	11	23,048.1	11	23,048.1
X	11,062	58.8	4,607	218.6	2,610	450.7
Y	557	154.3	140	1,388.6	106	1,439.7
UL	1,282	21.4	613	46.0	297	166.4

This Table is similar to Table 6 but shows the number and N50 length for various types of sequence contig (see Box 1). See legend to Table 6 concerning treatment of gaps. For sequence contigs in the draft genome sequence, the N50 length ranges from 1.7 to 5.5 times the arithmetic mean for initial sequence contigs, 2.5 to 8.2 times for merged sequence contigs, and 6.1 to 10 times for sequence-contig scaffolds.

Table 8 Chromosome size estimates

Chromosome*	Sequenced bases† (Mb)	FCC gaps‡		SCC gaps‖		Sequence gaps#		Heterochromatin and short arm adjustments**(Mb)	Total estimated chromosome size (including artefactual duplication in draft genome sequence)†† (Mb)	Previously estimated chromosome size‡‡ (Mb)
		Number	Total bases in gaps§ (Mb)	Number	Total bases in gaps¶ (Mb)	Number	Total bases in gaps☆ (Mb)			
All	2,692.9	897	152.0	4,076	142.7	145,514	80.6	212	3,289	3,286
1	212.2	104	17.7	347	12.1	11,803	6.5	30	279	263
2	221.6	50	8.5	296	10.4	12,880	7.1	3	251	255
3	186.2	71	12.1	336	11.8	14,689	8.1	3	221	214
4	168.1	39	6.6	343	12.0	12,768	7.1	3	197	203
5	169.7	46	7.8	337	11.8	10,304	5.7	3	198	194
6	158.1	15	2.6	275	9.6	5,225	2.9	3	176	183
7	146.2	27	4.6	195	6.8	4,338	2.4	3	163	171
8	124.3	41	7.0	249	8.7	8,692	4.8	3	148	155
9	106.9	19	3.2	122	4.3	6,083	3.4	22	140	145
10	127.1	14	2.4	163	5.7	8,947	5.0	3	143	144
11	128.6	29	4.9	193	6.8	8,279	4.6	3	148	144
12	124.5	26	4.4	168	5.9	8,226	4.6	3	142	143
13	92.9	12	2.0	115	4.0	5,065	2.8	16	118	114
14	86.9	13	2.2	40	1.4	775	0.4	16	107	109
15	73.4	18	3.1	104	3.6	5,717	3.2	17	100	106
16	73.1	55	9.4	102	3.6	4,757	2.6	15	104	98
17	72.8	41	7.0	95	3.3	4,261	2.4	3	88	92
18	72.9	22	3.7	113	4.0	4,324	2.4	3	86	85
19	55.4	49	8.3	108	3.8	2,344	1.3	3	72	67
20	60.5	7	1.2	33	1.2	469	0.3	3	66	72
21	33.8	4	0.1	0	0.0	0	0.0	11	45	50
22	33.8	10	1.0	0	0.0	0	0.0	13	48	56
X	127.7	141	24.0	182	6.4	4,282	2.4	3	163	164
Y	21.8	6	1.0	19	0.7	113	0.1	27	51	59
NA	5.1	0	0	134	0.0	577	0.3	0	0	0
UL	9.3	38	0	7	0.0	566	0.3	0	0	0

* NA, sequenced clones that could not be associated with fingerprint clone contigs. UL, clone contigs that could not be reliably placed on a chromosome.
† Total number of bases in the draft genome sequence, excluding gaps. Total length of scaffold (including gaps contained within clones) is 2.916 Gb.
‡ Gaps between those fingerprint clone contigs that contain sequenced clones excluding gaps for centromeres.
§ For unfinished chromosomes, we estimate an average size of 0.17 Mb per FCC gap, based on retrospective estimates of the clone coverage of chromosomes 21 and 22. Gap estimates for chromosomes 21 and 22 are taken from refs 93, 94.
‖ Gaps between sequenced-clone contigs within a fingerprint clone contig.
¶ For unfinished chromosomes, we estimate sequenced clone gaps at 0.035 Mb each, based on evaluation of a sample of these gaps.
# Gaps between two sequence contigs within a sequenced-clone contig.
☆ We estimate the average number of bases in sequence gaps from alignments of the initial sequence contigs of unfinished clones (see text) and extrapolation to the whole chromosome.
** Including adjustments for estimates of the sizes of the short arms of the acrocentric chromosomes 13, 14, 15, 21 and 22 (ref. 105), estimates for the centromere and heterochromatic regions of chromosomes 1, 9 and 16 (refs 106, 107) and estimates of 3 Mb for the centromere and 24 Mb for telomeric heterochromatin for the Y chromosome[108].
†† The sum of the five lengths in the preceding columns. This is an overestimate, because the draft genome sequence contains some artefactual sequence owing to inability to correctly to merge all underlying sequence contigs. The total amount of artefactual duplication varies among chromosomes; the overall amount is estimated by computational analysis to be about 100 Mb, or about 3% of the total length given, yielding a total estimated size of about 3,200 Mb for the human genome.
‡‡ Including heterochromatic regions and acrocentric short arm(s)[105].

© 2001 Macmillan Magazines Ltd

fingerprint map. However, many involve STSs that have been localized on only one or two of the previous maps or that occur as isolated discrepancies in conflict with several flanking STSs. Many of these cases are probably due to errors in the previous maps (with error rates for individual maps estimated at 1–2%[100]). Others may be due to incorrect assignment of the STSs to the draft genome sequence (by the electronic polymerase chain reaction (e-PCR) computer program) or to database entries that contain sequence data from more than one clone (owing to cross-contamination).

Graphical views of the independent data sets were particularly useful in detecting problems with order or orientation (Fig. 5). Areas of conflict were reviewed and corrected if supported by the underlying data. In the version discussed here, there were 41 sequenced clones falling in 14 sequenced-clone contigs with STS content information from multiple maps that disagreed with the flanking clones or sequenced-clone contigs; the placement of these clones thus remains suspect. Four of these instances suggest errors in the fingerprint map, whereas the others suggest errors in the layout of sequenced clones. These cases are being investigated and will be corrected in future versions.

Assembly of the sequenced clones. We assessed the accuracy of the assembly by using a set of 148 draft clones comprising 22.4 Mb for which finished sequence subsequently became available[104]. The initial sequence contigs lack information about order and orientation, and GigAssembler attempts to use linking data to infer such information as far as possible[104]. Starting with initial sequence contigs that were unordered and unoriented, the program placed 90% of the initial sequence contigs in the correct orientation and 85% in the correct order with respect to one another. In a separate test, GigAssembler was tested on simulated draft data produced from finished sequence on chromosome 22 and similar results were obtained.

Some problems remain at all levels. First, errors in the initial sequence contigs persist in the merged sequence contigs built from them and can cause difficulties in the assembly of the draft genome sequence. Second, GigAssembler may fail to merge some overlapping sequences because of poor data quality, allelic differences or misassemblies of the initial sequence contigs; this may result in apparent local duplication of a sequence. We have estimated by various methods the amount of such artefactual duplication in the assembly from these and other sources to be about 100 Mb. On the other hand, nearby duplicated sequences may occasionally be incorrectly merged. Some sequenced clones remain incorrectly placed on the layout, as discussed above, and others ($< 0.5\%$) remain unplaced. The fingerprint map has undoubtedly failed to resolve some closely related duplicated regions, such as the Williams region and several highly repetitive subtelomeric and pericentric regions (see below). Detailed examination and sequence finishing may be required to sort out these regions precisely, as has been done with chromosome Y[89]. Finally, small sequenced-clone contigs with limited or no STS landmark content remain difficult to place. Full utilization of the higher resolution radiation hybrid map (the TNG map) may help in this[95]. Future targeted FISH experiments and increased map continuity will also facilitate positioning of these sequences.

Table 9 Distribution of PHRAP scores in the draft genome sequence

PHRAP score	Percentage of bases in the draft genome sequence
0–9	0.6
10–19	1.3
20–29	2.2
30–39	4.8
40–49	8.1
50–59	8.7
60–69	9.0
70–79	12.1
80–89	17.3
>90	35.9

PHRAP scores are a logarithmically based representation of the error probability. A PHRAP score of X corresponds to an error probability of $10^{-X/10}$. Thus, PHRAP scores of 20, 30 and 40 correspond to accuracy of 99%, 99.9% and 99.99%, respectively. PHRAP scores are derived from quality scores of the underlying sequence reads used in sequence assembly. See http://www.genome.washington.edu/UWGC/analysistools/phrap.htm.

Genome coverage

We next assessed the nature of the gaps within the draft genome sequence, and attempted to estimate the fraction of the human genome not represented within the current version.

Gaps in draft genome sequence coverage. There are three types of gap in the draft genome sequence: gaps within unfinished sequenced clones; gaps between sequenced-clone contigs, but within fingerprint clone contigs; and gaps between fingerprint clone contigs. The first two types are relatively straightforward to close simply by performing additional sequencing and finishing on already identified clones. Closing the third type may require screening of additional large-insert clone libraries and possibly new technologies for the most recalcitrant regions. We consider these three cases in turn.

We estimated the size of gaps within draft clones by studying instances in which there was substantial overlap between a draft clone and a finished clone, as described above. The average gap size in these draft sequenced clones was 554 bp, although the precise estimate was sensitive to certain assumptions in the analysis. Assuming that the sequence gaps in the draft genome sequence are fairly represented by this sample, about 80 Mb or about 3% (likely range 2–4%) of sequence may lie in the 145,514 gaps within draft sequenced clones.

The gaps between sequenced-clone contigs but within fingerprint clone contigs are more difficult to evaluate directly, because the draft genome sequence flanking many of the gaps is often not precisely aligned with the fingerprinted clones. However, most are much smaller than a single BAC. In fact, nearly three-quarters of these gaps are bridged by one or more individual BACs, as indicated by linking information from BAC end sequences. We measured the sizes of a subset of gaps directly by examining restriction fragment fingerprints of overlapping clones. A study of 157 'bridged' gaps and 55 'unbridged' gaps gave an average gap size of 25 kb. Allowing for the possibility that these gaps may not be fully representative and that some restriction fragments are not included in the calculation, a more conservative estimate of gap size would be 35 kb. This would indicate that about 150 Mb or 5% of the human genome may reside in the 4,076 gaps between sequenced-clone contigs. This sequence should be readily obtained as the clones spanning them are sequenced.

The size of the gaps between fingerprint clone contigs was estimated by comparing the fingerprint maps to the essentially completed chromosomes 21 and 22. The analysis shows that the fingerprinted BAC clones in the global database cover 97–98% of the sequenced portions of those chromosomes[86]. The published sequences of these chromosomes also contain a few small gaps (5 and 11, respectively) amounting to some 1.6% of the euchromatic sequence, and do not include the heterochromatic portion. This suggests that the gaps between contigs in the fingerprint map contain about 4% of the euchromatic genome. Experience with closure of such gaps on chromosomes 20 and 7 suggests that many of these gaps are less than one clone in length and will be closed by clones from other libraries. However, recovery of sequence from these gaps represents the most challenging aspect of producing a complete finished sequence of the human genome.

As another measure of the representation of the BAC libraries, Riethman[109] has found BAC or cosmid clones that link to telomeric half-YACs or to the telomeric sequence itself for 40 of the 41 non-satellite telomeres. Thus, the fingerprint map appears to have no substantial gaps in these regions. Many of the pericentric regions are also represented, but analysis is less complete here (see below).

Representation of random raw sequences. In another approach to measuring coverage, we compared a collection of random raw sequence reads to the existing draft genome sequence. In principle,

© 2001 Macmillan Magazines Ltd

the fraction of reads matching the draft genome sequence should provide an estimate of genome coverage. In practice, the comparison is complicated by the need to allow for repeat sequences, the imperfect sequence quality of both the raw sequence and the draft genome sequence, and the possibility of polymorphism. Nonetheless, the analysis provides a reasonable view of the extent to which the genome is represented in the draft genome sequence and the public databases.

We compared the raw sequence reads against both the sequences used in the construction of the draft genome sequence and all of GenBank using the BLAST computer program. Of the 5,615 raw sequence reads analysed (each containing at least 100 bp of contiguous non-repetitive sequence), 4,924 had a match of $\geq 97\%$ identity with a sequenced clone, indicating that $88 \pm 1.5\%$ of the genome was represented in sequenced clones. The estimate is subject to various uncertainties. Most serious is the proportion of repeat sequence in the remainder of the genome. If the unsequenced portion of the genome is unusually rich in repeated sequence, we would underestimate its size (although the excess would be comprised of repeated sequence).

We examined those raw sequences that failed to match by comparing them to the other publicly available sequence resources. Fifty (0.9%) had matches in public databases containing cDNA sequences, STSs and similar data. An additional 276 (or 43% of the remaining raw sequence) had matches to the whole-genome shotgun reads discussed above (consistent with the idea that these reads cover about half of the genome).

We also examined the extent of genome coverage by aligning the cDNA sequences for genes in the RefSeq dataset[110] to the draft genome sequence. We found that 88% of the bases of these cDNAs could be aligned to the draft genome sequence at high stringency (at least 98% identity). (A few of the alignments with either the random raw sequence reads or the cDNAs may be to a highly similar region in the genome, but such matches should affect the estimate of genome coverage by considerably less than 1%, based on the estimated extent of duplication within the genome (see below).)

These results indicate that about 88% of the human genome is represented in the draft genome sequence and about 94% in the combined publicly available sequence databases. The figure of 88% agrees well with our independent estimates above that about 3%, 5% and 4% of the genome reside in the three types of gap in the draft genome sequence.

Finally, a small experimental check was performed by screening a large-insert clone library with probes corresponding to 16 of the whole genome shotgun reads that failed to match the draft genome sequence. Five hybridized to many clones from different fingerprint clone contigs and were discarded as being repetitive. Of the remaining eleven, two fell within sequenced clones (presumably within sequence gaps of the first type), eight fell in fingerprint clone contigs but between sequenced clones (gaps of the second type) and one failed to identify clones in the fingerprint map (gaps of the third type) but did identify clones in another large-insert library. Although these numbers are small, they are consistent with the view that the much of the remaining genome sequence lies within already identified clones in the current map.

Estimates of genome and chromosome sizes. Informed by this analysis of genome coverage, we proceeded to estimate the sizes of the genome and each of the chromosomes (Table 8). Beginning with the current assigned sequence for each chromosome, we corrected for the known gaps on the basis of their estimated sizes (see above). We attempted to account for the sizes of centromeres and heterochromatin, neither of which are well represented in the draft sequence. Finally, we corrected for around 100 Mb of artefactual duplication in the assembly. We arrived at a total human genome size estimate of around 3,200 Mb, which compares favourably with previous estimates based on DNA content.

We also independently estimated the size of the euchromatic portion of the genome by determining the fraction of the 5,615 random raw sequences that matched the finished portion of the human genome (whose total length is known with greater precision). Twenty-nine per cent of these raw sequences found a match among 835 Mb of nonredundant finished sequence. This leads to an estimate of the euchromatic genome size of 2.9 Gb. This agrees reasonably with the prediction above based on the length of the draft genome sequence (Table 8).

Update. The results above reflect the data on 7 October 2000. New data are continually being added, with improvements being made to the physical map, new clones being sequenced to close gaps and draft clones progressing to full shotgun coverage and finishing. The draft genome sequence will be regularly reassembled and publicly released.

Currently, the physical map has been refined such that the number of fingerprint clone contigs has fallen from 1,246 to 965; this reflects the elimination of some artefactual contigs and the closure of some gaps. The sequence coverage has risen such that 90% of the human genome is now represented in the sequenced clones and more than 94% is represented in the combined publicly available sequence databases. The total amount of finished sequence is now around 1 Gb.

Broad genomic landscape

What biological insights can be gleaned from the draft sequence? In this section, we consider very large-scale features of the draft genome sequence: the distribution of GC content, CpG islands and recombination rates, and the repeat content and gene content of the human genome. The draft genome sequence makes it possible to integrate these features and others at scales ranging from individual

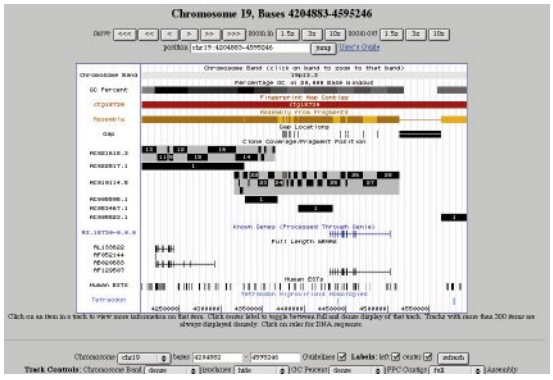


Figure 10 Screen shot from UCSC Draft Human Genome Browser. See http://genome.ucsc.edu/.

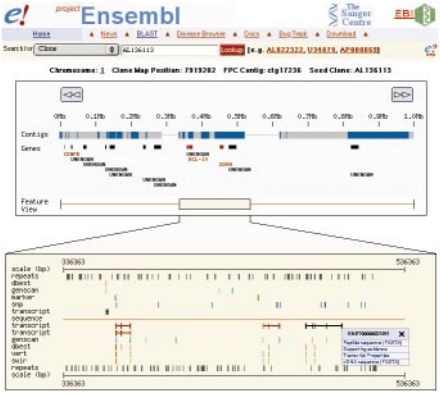


Figure 11 Screen shot from the Genome Browser of Project Ensembl. See http://www.ensembl.org.

© 2001 Macmillan Magazines Ltd

nucleotides to collections of chromosomes. Unless noted, all analyses were conducted on the assembled draft genome sequence described above.

Figure 9 provides a high-level view of the contents of the draft genome sequence, at a scale of about 3.8 Mb per centimetre. Of course, navigating information spanning nearly ten orders of magnitude requires computational tools to extract the full value. We have created and made freely available various 'Genome Browsers'. Browsers were developed and are maintained by the University of California at Santa Cruz (Fig. 10) and the EnsEMBL project of the European Bioinformatics Institute and the Sanger Centre (Fig. 11). Additional browsers have been created; URLs are listed at www.nhgri.nih.gov/genome_hub. These web-based computer tools allow users to view an annotated display of the draft genome sequence, with the ability to scroll along the chromosomes and zoom in or out to different scales. They include: the nucleotide sequence, sequence contigs, clone contigs, sequence coverage and finishing status, local GC content, CpG islands, known STS markers from previous genetic and physical maps, families of repeat sequences, known genes, ESTs and mRNAs, predicted genes, SNPs and sequence similarities with other organisms (currently the pufferfish *Tetraodon nigroviridis*). These browsers will be updated as the draft genome sequence is refined and corrected as additional annotations are developed.

In addition to using the Genome Browsers, one can download from these sites the entire draft genome sequence together with the annotations in a computer-readable format. The sequences of the underlying sequenced clones are all available through the public sequence databases. URLs for these and other genome websites are listed in Box 2. A larger list of useful URLs can be found at www.nhgri.nih.gov/genome_hub. An introduction to using the draft genome sequence, as well as associated databases and analytical tools, is provided in an accompanying paper[111].

In addition, the human cytogenetic map has been integrated with the draft genome sequence as part of a related project. The BAC Resource Consortium[103] established dense connections between the maps using more than 7,500 sequenced large-insert clones that had been cytogenetically mapped by FISH; the average density of the map is 2.3 clones per Mb. Although the precision of the integration is limited by the resolution of FISH, the links provide a powerful tool for the analysis of cytogenetic aberrations in inherited diseases and cancer. These cytogenetic links can also be accessed through the Genome Browsers.

Long-range variation in GC content

The existence of GC-rich and GC-poor regions in the human genome was first revealed by experimental studies involving density gradient separation, which indicated substantial variation in average GC content among large fragments. Subsequent studies have indicated that these GC-rich and GC-poor regions may have different biological properties, such as gene density, composition of repeat sequences, correspondence with cytogenetic bands and recombination rate[112–117]. Many of these studies were indirect, owing to the lack of sufficient sequence data.

The draft genome sequence makes it possible to explore the variation in GC content in a direct and global manner. Visual inspection (Fig. 9) confirms that local GC content undergoes substantial long-range excursions from its genome-wide average of 41%. If the genome were drawn from a uniform distribution of GC content, the local GC content in a window of size n bp should be $41 \pm \sqrt{((41)(59)/n)}$%. Fluctuations would be modest, with the standard deviation being halved as the window size is quadrupled—for example, 0.70%, 0.35%, 0.17% and 0.09% for windows of size 5, 20, 80 and 320 kb.

The draft genome sequence, however, contains many regions with much more extreme variation. There are huge regions (> 10 Mb) with GC content far from the average. For example, the most distal 48 Mb of chromosome 1p (from the telomere to about STS marker D1S3279) has an average GC content of 47.1%, and chromosome 13 has a 40-Mb region (roughly between STS marker A005X38 and

Box 2

Sources of publicly available sequence data and other relevant genomic information

http://genome.ucsc.edu/
University of California at Santa Cruz
Contains the assembly of the draft genome sequence used in this paper and updates

http://genome.wustl.edu/gsc/ human/Mapping/
Washington University
Contains links to clone and accession maps of the human genome

http://www.ensembl.org
EBI/Sanger Centre
Allows access to DNA and protein sequences with automatic baseline annotation

http://www.ncbi.nlm.nih.gov/ genome/guide/
NCBI
Views of chromosomes and maps and loci with links to other NCBI resources

http://www.ncbi.nlm.nih.gov/ genemap99/
Gene map 99: contains data and viewers for radiation hybrid maps of EST-based STSs

http://compbio.ornl.gov/channel/index.html
Oak Ridge National Laboratory
Java viewers for human genome data

http://hgrep.ims.u-tokyo.ac.jp/
RIKEN and the University of Tokyo
Gives an overview of the entire human genome structure

http://snp.cshl.org/
The SNP Consortium
Includes a variety of ways to query for SNPs in the human genome

http://www.ncbi.nlm.nih.gov/Omim/
Online *Mendelian Inheritance in Man*
Contain information about human genes and disease

http://www.nhgri.nih.gov/ELSI/ and http://www.ornl.gov/hgmis/elsi/elsi.html
NHGRI and DOE
Contains information, links and articles on a wide range of social, ethical and legal issues

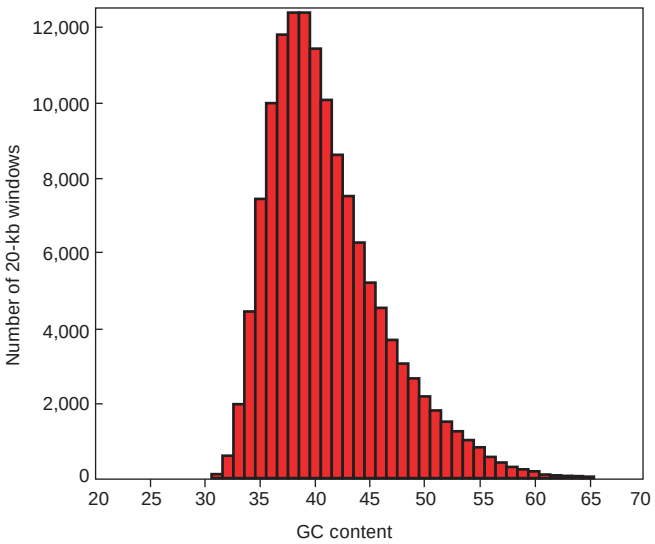


Figure 12 Histogram of GC content of 20-kb windows in the draft genome sequence.

© 2001 Macmillan Magazines Ltd

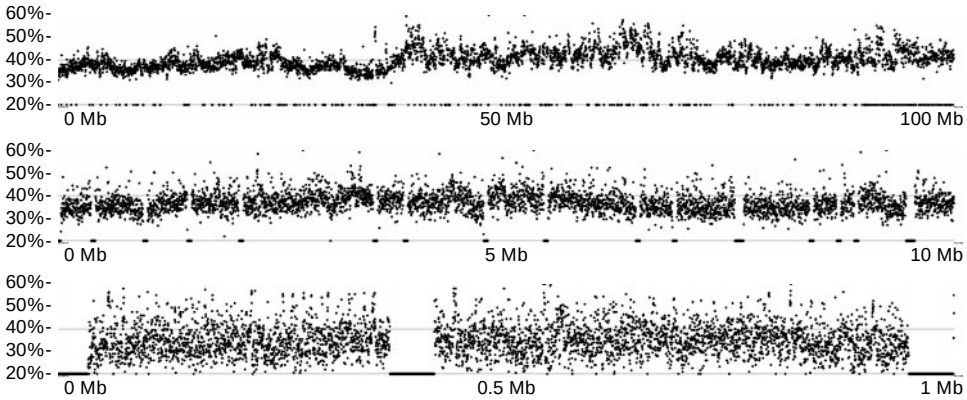


Figure 13 Variation in GC content at various scales. The GC content in subregions of a 100-Mb region of chromosome 1 is plotted, starting at about 83 Mb from the beginning of the draft genome sequence. This region is AT-rich overall. Top, the GC content of the entire 100-Mb region analysed in non-overlapping 20-kb windows. Middle, GC content of the first 10 Mb, analysed in 2-kb windows. Bottom, GC content of the first 1 Mb, analysed in 200-bp windows. At this scale, gaps in the sequence can be seen.

stsG30423) with only 36% GC content. There are also examples of large shifts in GC content between adjacent multimegabase regions. For example, the average GC content on chromosome 17q is 50% for the distal 10.3 Mb but drops to 38% for the adjacent 3.9 Mb. There are regions of less than 300 kb with even wider swings in GC content, for example, from 33.1% to 59.3%.

Long-range variation in GC content is evident not just from extreme outliers, but throughout the genome. The distribution of average GC content in 20-kb windows across the draft genome sequence is shown in Fig. 12. The spread is 15-fold larger than predicted by a uniform process. Moreover, the standard deviation barely decreases as window size increases by successive factors of four—5.9%, 5.2%, 4.9% and 4.6% for windows of size 5, 20, 80 and 320 kb. The distribution is also notably skewed, with 58% below the average and 42% above the average of 41%, with a long tail of GC-rich regions.

Bernardi and colleagues[118,119] proposed that the long-range variation in GC content may reflect that the genome is composed of a mosaic of compositionally homogeneous regions that they dubbed 'isochores'. They suggested that the skewed distribution is composed of five normal distributions, corresponding to five distinct types of isochore (L1, L2, H1, H2 and H3, with GC contents of $<38\%$, 38–42%, 42–47%, 47–52% and $>52\%$, respectively).

We studied the draft genome sequence to see whether strict isochores could be identified. For example, the sequence was divided into 300-kb windows, and each window was subdivided into 20-kb subwindows. We calculated the average GC content for each window and subwindow, and investigated how much of the variance in the GC content of subwindows across the genome can be statistically 'explained' by the average GC content in each window. About three-quarters of the genome-wide variance among 20-kb windows can be statistically explained by the average GC content of 300-kb windows that contain them, but the residual variance among subwindows (standard deviation, 2.4%) is still far too large to be consistent with a homogeneous distribution. In fact, the hypothesis of homogeneity could be rejected for each 300-kb window in the draft genome sequence.

Similar results were obtained with other window and subwindow sizes. Some of the local heterogeneity in GC content is attributable to transposable element insertions (see below). Such repeat elements typically have a higher GC content than the surrounding sequence, with the effect being strongest for the most recent insertions.

These results rule out a strict notion of isochores as compositionally homogeneous. Instead, there is substantial variation at many different scales, as illustrated in Fig. 13. Although isochores do not appear to merit the prefix 'iso', the genome clearly does contain large regions of distinctive GC content and it is likely to be worth redefining the concept so that it becomes possible rigorously to partition the genome into regions. In the absence of a precise definition, we will loosely refer to such regions as 'GC content domains' in the context of the discussion below.

Fickett *et al.*[120] have explored a model in which the underlying preference for a particular GC content drifts continuously throughout the genome, an approach that bears further examination. Churchill[121] has proposed that the boundaries between GC content domains can in some cases be predicted by a hidden Markov model, with one state representing a GC-rich region and one representing an AT-rich region. We found that this approach tended to identify only very short domains of less than a kilobase (data not shown), but variants of this approach deserve further attention.

The correlation between GC content domains and various biological properties is of great interest, and this is likely to be the most fruitful route to understanding the basis of variation in GC content. As described below, we confirm the existence of strong correlations with both repeat content and gene density. Using the integration between the draft genome sequence and the cytogenetic map described above, it is possible to confirm a statistically significant correlation between GC content and Giemsa bands (G-bands). For example, 98% of large-insert clones mapping to the darkest G-bands are in 200-kb regions of low GC content (average 37%), whereas more than 80% of clones mapping to the lightest G-bands are in regions of high GC content (average 45%)[103]. Estimated band locations can be seen in Fig. 9 and viewed in the context of other genome annotation at http://genome.ucsc.edu/goldenPath/mapPlots/ and http://genome.ucsc.edu/goldenPath/hgTracks.html.

CpG islands

A related topic is the distribution of so-called CpG islands across the genome. The dinucleotide CpG is notable because it is greatly under-represented in human DNA, occurring at only about one-fifth of the roughly 4% frequency that would be expected by simply multiplying the typical fraction of Cs and Gs (0.21×0.21). The deficit occurs because most CpG dinucleotides are methylated on the cytosine base, and spontaneous deamination of methyl-C residues gives rise to T residues. (Spontaneous deamination of ordinary cytosine residues gives rise to uracil residues that are readily recognized and repaired by the cell.) As a result, methyl-CpG dinucleotides steadily mutate to TpG dinucleotides. However, the genome contains many 'CpG islands' in which CpG dinucleotides are not methylated and occur at a frequency closer to that predicted by the local GC content. CpG islands are of particular interest because many are associated with the 5′ ends of genes[122–127].

We searched the draft genome sequence for CpG islands. Ideally, they should be defined by directly testing for the absence of cytosine methylation, but that was not practical for this report. There are

© 2001 Macmillan Magazines Ltd

Table 10 Number of CpG islands by GC content

GC content of island	Number of islands	Percentage of islands	Nucleotides in islands	Percentage of nucleotides in islands
Total	28,890	100	19,818,547	100
>80%	22	0.08	5,916	0.03
70–80%	5,884	20	3,111,965	16
60–70%	18,779	65	13,110,924	66
50–60%	4,205	15	3,589,742	18

Potential CpG islands were identified by searching the draft genome sequence one base at a time, scoring each dinucleotide (+17 for GC, −1 for others) and identifying maximally scoring segments. Each segment was then evaluated to determine GC content (≥50%), length (>200) and ratio of observed proportion of GC dinucleotides to the expected proportion on the basis of the GC content of the segment (>0.60), using a modification of a program developed by G. Micklem (personal communication).

various computer programs that attempt to identify CpG islands on the basis of primary sequence alone. These programs differ in some important respects (such as how aggressively they subdivide long CpG-containing regions), and the precise correspondence with experimentally undermethylated islands has not been validated. Nevertheless, there is a good correlation, and computational analysis thus provides a reasonable picture of the distribution of CpG islands in the genome.

To identify CpG islands, we used the definition proposed by Gardiner-Garden and Frommer[128] and embodied in a computer program. We searched the draft genome sequence for CpG islands, using both the full sequence and the sequence masked to eliminate repeat sequences. The number of regions satisfying the definition of a CpG island was 50,267 in the full sequence and 28,890 in the repeat-masked sequence. The difference reflects the fact that some repeat elements (notably Alu) are GC-rich. Although some of these repeat elements may function as control regions, it seems unlikely that most of the apparent CpG islands in repeat sequences are functional. Accordingly, we focused on those in the non-repeated sequence. The count of 28,890 CpG islands is reasonably close to the previous estimate of about 35,000 (ref. 129, as modified by ref. 130). Most of the islands are short, with 60–70% GC content (Table 10). More than 95% of the islands are less than 1,800 bp long, and more than 75% are less than 850 bp. The longest CpG island (on chromosome 10) is 36,619 bp long, and 322 are longer than 3,000 bp. Some of the larger islands contain ribosomal pseudogenes, although RNA genes and pseudogenes account for only a small proportion of all islands (< 0.5%). The smaller islands are consistent with their previously hypothesized function, but the role of these larger islands is uncertain.

The density of CpG islands varies substantially among some of the chromosomes. Most chromosomes have 5–15 islands per Mb, with a mean of 10.5 islands per Mb. However, chromosome Y has an unusually low 2.9 islands per Mb, and chromosomes 16, 17 and 22 have 19–22 islands per Mb. The extreme outlier is chromosome 19, with 43 islands per Mb. Similar trends are seen when considering the percentage of bases contained in CpG islands. The relative density of CpG islands correlates reasonably well with estimates of relative gene density on these chromosomes, based both on previous mapping studies involving ESTs (Fig. 14) and on the distribution of gene predictions discussed below.

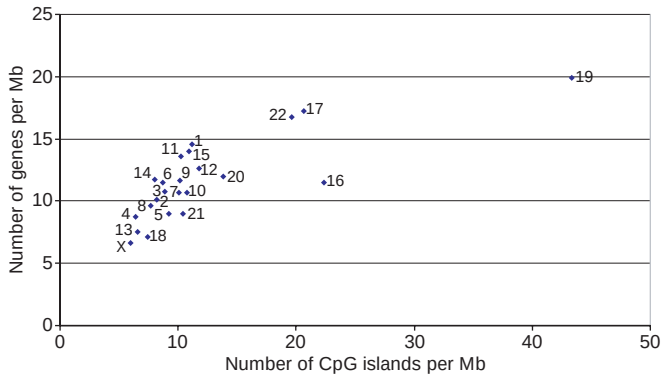


Figure 14 Number of CpG islands per Mb for each chromosome, plotted against the number of genes per Mb (the number of genes was taken from GeneMap98 (ref. 100)). Chromosomes 16, 17, 22 and particularly 19 are clear outliers, with a density of CpG islands that is even greater than would be expected from the high gene counts for these four chromosomes.

Comparison of genetic and physical distance

The draft genome sequence makes it possible to compare genetic and physical distances and thereby to explore variation in the rate of recombination across the human chromosomes. We focus here on large-scale variation. Finer variation is examined in an accompanying paper[131].

The genetic and physical maps are integrated by 5,282 polymorphic loci from the Marshfield genetic map[102], whose positions are known in terms of centimorgans (cM) and Mb along the chromosomes. Figure 15 shows the comparison of the draft genome sequence for chromosome 12 with the male, female and sex-averaged maps. One can calculate the approximate ratio of cM per Mb across a chromosome (reflected in the slopes in Fig. 15) and the average recombination rate for each chromosome arm.

Two striking features emerge from analysis of these data. First, the average recombination rate increases as the length of the chromosome arm decreases (Fig. 16). Long chromosome arms have an average recombination rate of about 1 cM per Mb, whereas the shortest arms are in the range of 2 cM per Mb. A similar trend has been seen in the yeast genome[132,133], despite the fact that the physical scale is nearly 200 times as small. Moreover, experimental studies have shown that lengthening or shortening yeast chromosomes results in a compensatory change in recombination rate[132].

The second observation is that the recombination rate tends to be suppressed near the centromeres and higher in the distal portions of most chromosomes, with the increase largely in the terminal

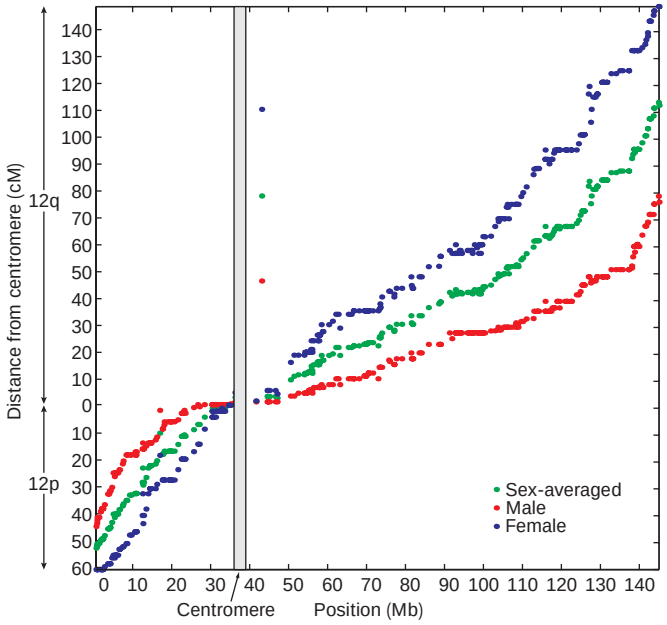


Figure 15 Distance in cM along the genetic map of chromosome 12 plotted against position in Mb in the draft genome sequence. Female, male and sex-averaged maps are shown. Female recombination rates are much higher than male recombination rates. The increased slopes at either end of the chromosome reflect the increased rates of recombination per Mb near the telomeres. Conversely, the flatter slope near the centromere shows decreased recombination there, especially in male meiosis. This is typical of the other chromosomes as well (see http://genome.ucsc.edu/goldenPath/mapPlots). Discordant markers may be map, marker placement or assembly errors.

© 2001 Macmillan Magazines Ltd

20–35 Mb. The increase is most pronounced in the male meiotic map. The effect can be seen, for example, from the higher slope at both ends of chromosome 12 (Fig. 15). Regional and sex-specific effects have been observed for chromosome 21 (refs 110, 134).

Why is recombination higher on smaller chromosome arms? A higher rate would increase the likelihood of at least one crossover during meiosis on each chromosome arm, as is generally observed in human chiasmata counts[135]. Crossovers are believed to be necessary for normal meiotic disjunction of homologous chromosome pairs in eukaryotes. An extreme example is the pseudoautosomal regions on chromosomes Xp and Yp, which pair during male meiosis; this physical region of only 2.6 Mb has a genetic length of 50 cM (corresponding to 20 cM per Mb), with the result that a crossover is virtually assured.

Mechanistically, the increased rate of recombination on shorter chromosome arms could be explained if, once an initial recombination event occurs, additional nearby events are blocked by positive crossover interference on each arm. Evidence from yeast mutants in which interference is abolished shows that interference plays a key role in distributing a limited number of crossovers among the various chromosome arms in yeast[136]. An alternative possibility is that a checkpoint mechanism scans for and enforces the presence of at least one crossover on each chromosome arm.

Variation in recombination rates along chromosomes and between the sexes is likely to reflect variation in the initiation of meiosis-induced double-strand breaks (DSBs) that initiate recombination. DSBs in yeast have been associated with open chromatin[137,138], rather than with specific DNA sequence motifs. With the availability of the draft genome sequence, it should be possible to explore in an analogous manner whether variation in human recombination rates reflects systematic differences in chromosome accessibility during meiosis.

Repeat content of the human genome

A puzzling observation in the early days of molecular biology was that genome size does not correlate well with organismal complexity. For example, *Homo sapiens* has a genome that is 200 times as large as that of the yeast *S. cerevisiae*, but 200 times as small as that of *Amoeba dubia*[139,140]. This mystery (the C-value paradox) was largely resolved with the recognition that genomes can contain a large quantity of repetitive sequence, far in excess of that devoted to protein-coding genes (reviewed in refs 140, 141).

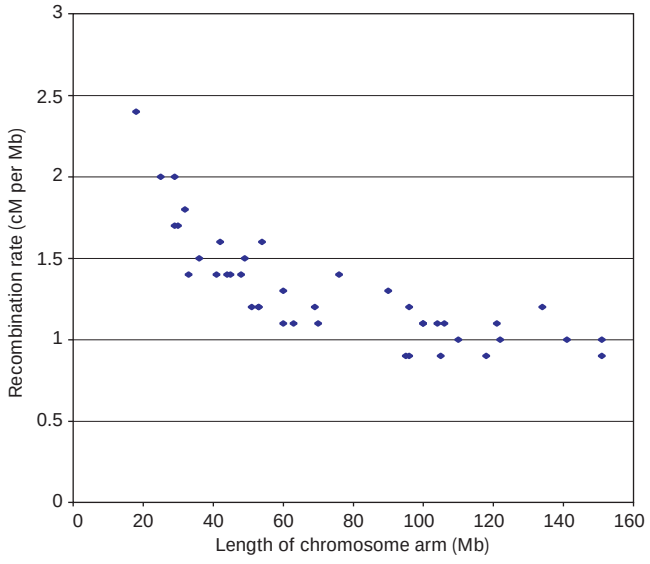


Figure 16 Rate of recombination averaged across the euchromatic portion of each chromosome arm plotted against the length of the chromosome arm in Mb. For large chromosomes, the average recombination rates are very similar, but as chromosome arm length decreases, average recombination rates rise markedly.

In the human, coding sequences comprise less than 5% of the genome (see below), whereas repeat sequences account for at least 50% and probably much more. Broadly, the repeats fall into five classes: (1) transposon-derived repeats, often referred to as interspersed repeats; (2) inactive (partially) retroposed copies of cellular genes (including protein-coding genes and small structural RNAs), usually referred to as processed pseudogenes; (3) simple sequence repeats, consisting of direct repetitions of relatively short *k*-mers such as $(A)_n$, $(CA)_n$ or $(CGG)_n$; (4) segmental duplications, consisting of blocks of around 10–300 kb that have been copied from one region of the genome into another region; and (5) blocks of tandemly repeated sequences, such as at centromeres, telomeres, the short arms of acrocentric chromosomes and ribosomal gene clusters. (These regions are intentionally under-represented in the draft genome sequence and are not discussed here.)

Repeats are often described as 'junk' and dismissed as uninteresting. However, they actually represent an extraordinary trove of information about biological processes. The repeats constitute a rich palaeontological record, holding crucial clues about evolutionary events and forces. As passive markers, they provide assays for studying processes of mutation and selection. It is possible to recognize cohorts of repeats 'born' at the same time and to follow their fates in different regions of the genome or in different species. As active agents, repeats have reshaped the genome by causing ectopic rearrangements, creating entirely new genes, modifying and reshuffling existing genes, and modulating overall GC content. They also shed light on chromosome structure and dynamics, and provide tools for medical genetic and population genetic studies.

The human is the first repeat-rich genome to be sequenced, and so we investigated what information could be gleaned from this majority component of the human genome. Although some of the general observations about repeats were suggested by previous studies, the draft genome sequence provides the first comprehensive view, allowing some questions to be resolved and new mysteries to emerge.

Transposon-derived repeats

Most human repeat sequence is derived from transposable elements[142,143]. We can currently recognize about 45% of the genome as belonging to this class. Much of the remaining 'unique' DNA must also be derived from ancient transposable element copies that have diverged too far to be recognized as such. To describe our analyses of interspersed repeats, it is necessary briefly to review the relevant features of human transposable elements.

Classes of transposable elements. In mammals, almost all transposable elements fall into one of four types (Fig. 17), of which three transpose through RNA intermediates and one transposes directly as DNA. These are long interspersed elements (LINEs), short interspersed elements (SINEs), LTR retrotransposons and DNA transposons.

LINEs are one of the most ancient and successful inventions in eukaryotic genomes. In humans, these transposons are about 6 kb long, harbour an internal polymerase II promoter and encode two open reading frames (ORFs). Upon translation, a LINE RNA assembles with its own encoded proteins and moves to the nucleus, where an endonuclease activity makes a single-stranded nick and the reverse transcriptase uses the nicked DNA to prime reverse transcription from the 3′ end of the LINE RNA. Reverse transcription frequently fails to proceed to the 5′ end, resulting in many truncated, nonfunctional insertions. Indeed, most LINE-derived repeats are short, with an average size of 900 bp for all LINE1 copies, and a median size of 1,070 bp for copies of the currently active LINE1 element (L1Hs). New insertion sites are flanked by a small

© 2001 Macmillan Magazines Ltd

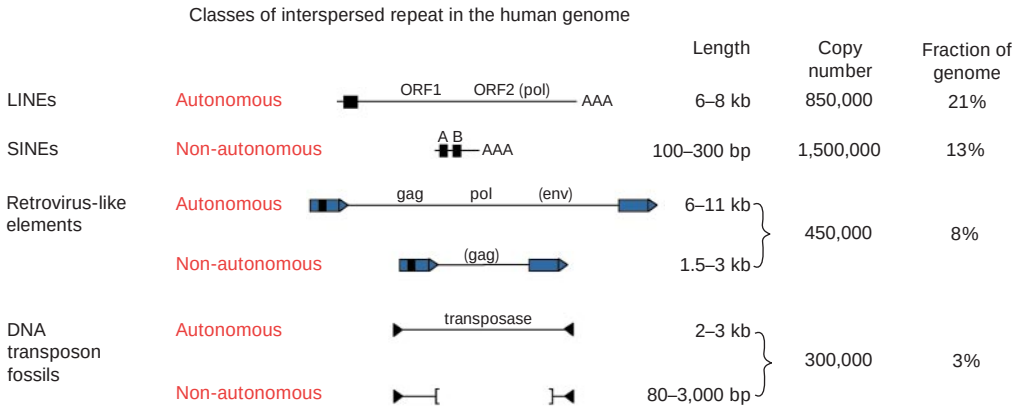


Figure 17 Almost all transposable elements in mammals fall into one of four classes. See text for details.

target site duplication of 7–20 bp. The LINE machinery is believed to be responsible for most reverse transcription in the genome, including the retrotransposition of the non-autonomous SINEs[144] and the creation of processed pseudogenes[145,146]. Three distantly related LINE families are found in the human genome: LINE1, LINE2 and LINE3. Only LINE1 is still active.

SINEs are wildly successful freeloaders on the backs of LINE elements. They are short (about 100–400 bp), harbour an internal polymerase III promoter and encode no proteins. These non-autonomous transposons are thought to use the LINE machinery for transposition. Indeed, most SINEs 'live' by sharing the 3′ end with a resident LINE element[144]. The promoter regions of all known SINEs are derived from tRNA sequences, with the exception of a single monophyletic family of SINEs derived from the signal recognition particle component 7SL. This family, which also does not share its 3′ end with a LINE, includes the only active SINE in the human genome: the Alu element. By contrast, the mouse has both tRNA-derived and 7SL-derived SINEs. The human genome contains three distinct monophyletic families of SINEs: the active Alu, and the inactive MIR and Ther2/MIR3.

LTR retroposons are flanked by long terminal direct repeats that contain all of the necessary transcriptional regulatory elements. The autonomous elements (retrotransposons) contain *gag* and *pol* genes, which encode a protease, reverse transcriptase, RNAse H and integrase. Exogenous retroviruses seem to have arisen from endogenous retrotransposons by acquisition of a cellular *envelope* gene (*env*)[147]. Transposition occurs through the retroviral mechanism with reverse transcription occurring in a cytoplasmic virus-like particle, primed by a tRNA (in contrast to the nuclear location and chromosomal priming of LINEs). Although a variety of LTR retrotransposons exist, only the vertebrate-specific endogenous retroviruses (ERVs) appear to have been active in the mammalian genome. Mammalian retroviruses fall into three classes (I–III), each comprising many families with independent origins. Most (85%) of the LTR retroposon-derived 'fossils' consist only of an isolated LTR, with the internal sequence having been lost by homologous recombination between the flanking LTRs.

DNA transposons resemble bacterial transposons, having terminal inverted repeats and encoding a transposase that binds near the inverted repeats and mediates mobility through a 'cut-and-paste' mechanism. The human genome contains at least seven major classes of DNA transposon, which can be subdivided into many families with independent origins[148] (see RepBase, http://www.girinst.org/~server/repbase.html). DNA transposons tend to have short life spans within a species. This can be explained by contrasting the modes of transposition of DNA transposons and LINE elements. LINE transposition tends to involve only functional elements, owing to the *cis*-preference by which LINE proteins assemble with the RNA from which they were translated. By contrast, DNA transposons cannot exercise a *cis*-preference: the encoded transposase is produced in the cytoplasm and, when it returns to the nucleus, it cannot distinguish active from inactive elements. As inactive copies accumulate in the genome, transposition becomes less efficient. This checks the expansion of any DNA transposon family and in due course causes it to die out. To survive, DNA transposons must eventually move by horizontal transfer to virgin genomes, and there is considerable evidence for such transfer[149–153].

Transposable elements employ different strategies to ensure their evolutionary survival. LINEs and SINEs rely almost exclusively on vertical transmission within the host genome[154] (but see refs 148, 155). DNA transposons are more promiscuous, requiring relatively frequent horizontal transfer. LTR retroposons use both strategies, with some being long-term active residents of the human genome (such as members of the ERVL family) and others having only short residence times.

Table 11 Number of copies and fraction of genome for classes of interspersed repeat

	Number of copies (× 1,000)	Total number of bases in the draft genome sequence (Mb)	Fraction of the draft genome sequence (%)	Number of families (subfamilies)
SINEs	1,558	359.6	13.14	
Alu	1,090	290.1	10.60	1 (~20)
MIR	393	60.1	2.20	1 (1)
MIR3	75	9.3	0.34	1 (1)
LINEs	868	558.8	20.42	
LINE1	516	462.1	16.89	1 (~55)
LINE2	315	88.2	3.22	1 (2)
LINE3	37	8.4	0.31	1 (2)
LTR elements	443	227.0	8.29	
ERV-class I	112	79.2	2.89	72 (132)
ERV(K)-class II	8	8.5	0.31	10 (20)
ERV (L)-class III	83	39.5	1.44	21 (42)
MaLR	240	99.8	3.65	1 (31)
DNA elements	294	77.6	2.84	
hAT group				
MER1-Charlie	182	38.1	1.39	25 (50)
Zaphod	13	4.3	0.16	4 (10)
Tc-1 group				
MER2-Tigger	57	28.0	1.02	12 (28)
Tc2	4	0.9	0.03	1 (5)
Mariner	14	2.6	0.10	4 (5)
PiggyBac-like	2	0.5	0.02	10 (20)
Unclassified	22	3.2	0.12	7 (7)
Unclassified	3	3.8	0.14	3 (4)
Total interspersed repeats		1,226.8	44.83	

The number of copies and base pair contributions of the major classes and subclasses of transposable elements in the human genome. Data extracted from a RepeatMasker analysis of the draft genome sequence (RepeatMasker version 09092000, sensitive settings, using RepBase Update 5.08). In calculating percentages, RepeatMasker excluded the runs of Ns linking the contigs in the draft genome sequence. In the last column, separate consensus sequences in the repeat databases are considered subfamilies, rather than families, when the sequences are closely related or related through intermediate subfamilies.

© 2001 Macmillan Magazines Ltd

Census of human repeats. We began by taking a census of the transposable elements in the draft genome sequence, using a recently updated version of the RepeatMasker program (version 09092000) run under sensitive settings (see http://repeatmasker.genome.washington.edu). This program scans sequences to identify full-length and partial members of all known repeat families represented in RepBase Update (version 5.08; see http://www.girinst.org/~server/repbase.html and ref. 156). Table 11 shows the number of copies and fraction of the draft genome sequence occupied by each of the four major classes and the main subclasses.

The precise count of repeats is obviously underestimated because the genome sequence is not finished, but their density and other properties can be stated with reasonable confidence. Currently recognized SINEs, LINEs, LTR retroposons and DNA transposon copies comprise 13%, 20%, 8% and 3% of the sequence, respectively. We expect these densities to grow as more repeat families are recognized, among which will be lower copy number LTR elements and DNA transposons, and possibly high copy number ancient (highly diverged) repeats.

Age distribution. The age distribution of the repeats in the human genome provides a rich 'fossil record' stretching over several hundred million years. The ancestry and approximate age of each fossil can be inferred by exploiting the fact that each copy is derived from, and therefore initially carried the sequence of, a then-active transposon and, being generally under no functional constraint, has accumulated mutations randomly and independently of other copies. We can infer the sequence of the ancestral active elements by clustering the modern derivatives into phylogenetic trees and building a consensus based on the multiple sequence alignment of a cluster of copies. Using available consensus sequences for known repeat subfamilies, we calculated the per cent divergence from the inferred ancestral active transposon for each of three million interspersed repeats in the draft genome sequence.

The percentage of sequence divergence can be converted into an approximate age in millions of years (Myr) on the basis of evolutionary information. Care is required in calibrating the clock, because the rate of sequence divergence may not be constant over time or between lineages[139]. The relative-rate test[157] can be used to calculate the sequence divergence that accumulated in a lineage after a given timepoint, on the basis of comparison with a sibling species that diverged at that time and an outgroup species. For example, the substitution rate over roughly the last 25 Myr in the human lineage can be calculated by using old world monkeys (which diverged about 25 Myr ago) as a sibling species and new world monkeys as an outgroup. We have used currently available calibrations for the human lineage, but the issue should be revisited as sequence information becomes available from different mammals.

Figure 18a shows the representation of various classes of trans-

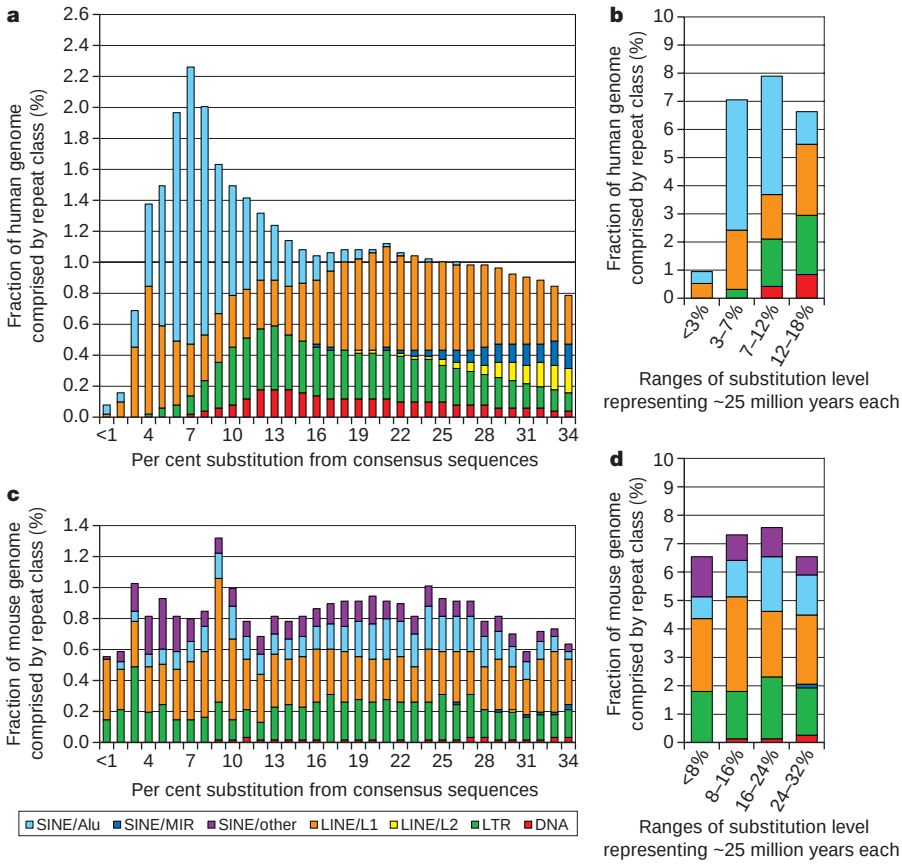


Figure 18 Age distribution of interspersed repeats in the human and mouse genomes. Bases covered by interspersed repeats were sorted by their divergence from their consensus sequence (which approximates the repeat's original sequence at the time of insertion). The average number of substitutions per 100 bp (substitution level, K) was calculated from the mismatch level p assuming equal frequency of all substitutions (the one-parameter Jukes–Cantor model, $K = -3/4\ln(1 - 4/3p)$). This model tends to underestimate higher substitution levels. CpG dinucleotides in the consensus were excluded from the substitution level calculations because the C⟶T transition rate in CpG pairs is about tenfold higher than other transitions and causes distortions in comparing transposable elements with high and low CpG content. **a**, The distribution, for the human genome, in bins corresponding to 1% increments in substitution levels. **b**, The data grouped into bins representing roughly equal time periods of 25 Myr. **c**,**d**, Equivalent data for available mouse genomic sequence. There is a different correspondence between substitution levels and time periods owing to different rates of nucleotide substitution in the two species. The correspondence between substitution levels and time periods was largely derived from three-way species comparisons (relative rate test[139,157]) with the age estimates based on fossil data. Human divergence from gibbon 20–30 Myr; old world monkey 25–35 Myr; prosimians 55–80 Myr; eutherian mammalian radiation ~100 Myr.

© 2001 Macmillan Magazines Ltd

posable elements in categories reflecting equal amounts of sequence divergence. In Fig. 18b the data are grouped into four bins corresponding to successive 25-Myr periods, on the basis of an approximate clock. Figure 19 shows the mean ages of various subfamilies of DNA transposons. Several facts are apparent from these graphs. First, most interspersed repeats in the human genome predate the eutherian radiation. This is a testament to the extremely slow rate with which nonfunctional sequences are cleared from vertebrate genomes (see below concerning comparison with the fly).

Second, LINE and SINE elements have extremely long lives. The monophyletic LINE1 and Alu lineages are at least 150 and 80 Myr old, respectively. In earlier times, the reigning transposons were LINE2 and MIR[148,158]. The SINE MIR was perfectly adapted for reverse transcription by LINE2, as it carried the same 50-base sequence at its 3′ end. When LINE2 became extinct 80–100 Myr ago, it spelled the doom of MIR.

Third, there were two major peaks of DNA transposon activity (Fig. 19). The first involved Charlie elements and occurred long before the eutherian radiation; the second involved Tigger elements and occurred after this radiation. Because DNA transposons can produce large-scale chromosome rearrangements[159–162], it is possible that widespread activity could be involved in speciation events.

Fourth, there is no evidence for DNA transposon activity in the past 50 Myr in the human genome. The youngest two DNA transposon families that we can identify in the draft genome sequence (MER75 and MER85) show 6–7% divergence from their respective consensus sequences representing the ancestral element (Fig. 19), indicating that they were active before the divergence of humans and new world monkeys. Moreover, these elements were relatively unsuccessful, together contributing just 125 kb to the draft genome sequence.

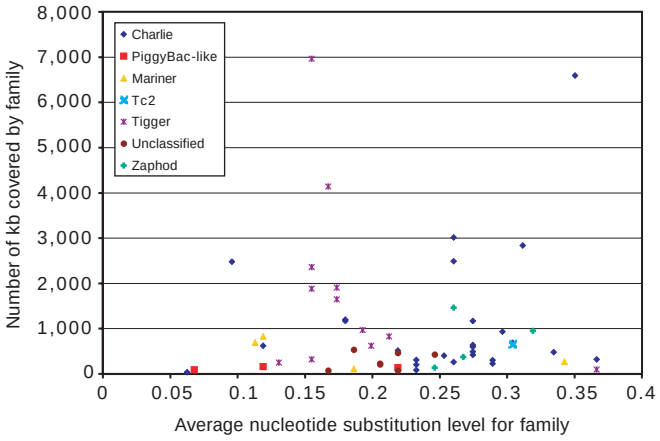


Figure 19 Median ages and per cent of the genome covered by subfamilies of DNA transposons. The Charlie and Zaphod elements were hobo-Activator-Tam3 (hAT) DNA transposons; Mariner, Tc2 and Tigger were Tc1-like elements. Unlike retroposons, DNA transposons are thought to have a short life span in a genome. Thus, the average or median divergence of copies from the consensus is a particularly accurate measure of the age of the DNA transposon copies.

Finally, LTR retroposons appear to be teetering on the brink of extinction, if they have not already succumbed. For example, the most prolific elements (ERVL and MaLRs) flourished for more than 100 Myr but appear to have died out about 40 Myr ago[163,164]. Only a single LTR retroposon family (HERVK10) is known to have transposed since our divergence from the chimpanzee 7 Myr ago, with only one known copy (in the HLA region) that is not shared between all humans[165]. In the draft genome sequence, we can identify only three full-length copies with all ORFs intact (the final total may be slightly higher owing to the imperfect state of the draft genome sequence).

More generally, the overall activity of all transposons has declined markedly over the past 35–50 Myr, with the possible exception of LINE1 (Fig. 18). Indeed, apart from an exceptional burst of activity of Alus peaking around 40 Myr ago, there would appear to have been a fairly steady decline in activity in the hominid lineage since the mammalian radiation. The extent of the decline must be even greater than it appears because old repeats are gradually removed by random deletion and because old repeat families are harder to recognize and likely to be under-represented in the repeat databases. (We confirmed that the decline in transposition is not an artefact arising from errors in the draft genome sequence, which, in principle, could increase the divergence level in recent elements. First, the sequence error rate (Table 9) is far too low to have a significant effect on the apparent age of recent transposons; and second, the same result is seen if one considers only finished sequence.)

What explains the decline in transposon activity in the lineage leading to humans? We return to this question below, in the context of the observation that there is no similar decline in the mouse genome.

Comparison with other organisms. We compared the complement of transposable elements in the human genome with those of the other sequenced eukaryotic genomes. We analysed the fly, worm and mustard weed genomes for the number and nature of repeats (Table 12) and the age distribution (Fig. 20). (For the fly, we analysed the 114 Mb of unfinished 'large' contigs produced by the whole-genome shotgun assembly[166], which are reported to represent euchromatic sequence. Similar results were obtained by analysing 30 Mb of finished euchromatic sequence.) The human genome stands in stark contrast to the genomes of the other organisms.

(1) The euchromatic portion of the human genome has a much higher density of transposable element copies than the euchromatic DNA of the other three organisms. The repeats in the other organisms may have been slightly underestimated because the repeat databases for the other organisms are less complete than for the human, especially with regard to older elements; on the other hand, recent additions to these databases appear to increase the repeat content only marginally.

(2) The human genome is filled with copies of ancient transposons, whereas the transposons in the other genomes tend to be of more recent origin. The difference is most marked with the fly, but is clear for the other genomes as well. The accumulation of old repeats is likely to be determined by the rate at which organisms engage in 'housecleaning' through genomic deletion. Studies of pseudogenes

Table 12 Number and nature of interspersed repeats in eukaryotic genomes

	Human		Fly		Worm		Mustard weed	
	Percentage of bases	Approximate number of families	Percentage of bases	Approximate number of families	Percentage of bases	Approximate number of families	Percentage of bases	Approximate number of families
LINE/SINE	33.40%	6	0.70%	20	0.40%	10	0.50%	10
LTR	8.10%	100	1.50%	50	0.00%	4	4.80%	70
DNA	2.80%	60	0.70%	20	5.30%	80	5.10%	80
Total	44.40%	170	3.10%	90	6.50%	90	10.50%	160

The complete genomes of fly, worm, and chromosomes 2 and 4 of mustard weed (as deposited at ncbi.nlm.nih.gov/genbank/genomes) were screened against the repeats in RepBase Update 5.02 (September 2000) with RepeatMasker at sensitive settings.

© 2001 Macmillan Magazines Ltd

have suggested that small deletions occur at a rate that is 75-fold higher in flies than in mammals; the half-life of such nonfunctional DNA is estimated at 12 Myr for flies and 800 Myr for mammals[167]. The rate of large deletions has not been systematically compared, but seems likely also to differ markedly.

(3) Whereas in the human two repeat families (LINE1 and Alu) account for 60% of all interspersed repeat sequence, the other organisms have no dominant families. Instead, the worm, fly and mustard weed genomes all contain many transposon families, each consisting of typically hundreds to thousands of elements. This difference may be explained by the observation that the vertically transmitted, long-term residential LINE and SINE elements represent 75% of interspersed repeats in the human genome, but only 5–25% in the other genomes. In contrast, the horizontally transmitted and shorter-lived DNA transposons represent only a small portion of all interspersed repeats in humans (6%) but a much larger fraction in fly, mustard weed and worm (25%, 49% and 87%, respectively). These features of the human genome are probably general to all mammals. The relative lack of horizontally transmitted elements may have its origin in the well developed immune system of mammals, as horizontal transfer requires infectious vectors, such as viruses, against which the immune system guards.

We also looked for differences among mammals, by comparing the transposons in the human and mouse genomes. As with the human genome, care is required in calibrating the substitution clock for the mouse genome. There is considerable evidence that the rate of substitution per Myr is higher in rodent lineages than in the hominid lineages[139,168,169]. In fact, we found clear evidence for different rates of substitution by examining families of transposable elements whose insertions predate the divergence of the human and mouse lineages. In an analysis of 22 such families, we found that the substitution level was an average of 1.7-fold higher in mouse than human (not shown). (This is likely to be an underestimate because of an ascertainment bias against the most diverged copies.) The faster clock in mouse is also evident from the fact that the ancient LINE2 and MIR elements, which transposed before the mammalian radiation and are readily detectable in the human genome, cannot be readily identified in available mouse genomic sequence (Fig. 18).

We used the best available estimates to calibrate substitution levels and time[169]. The ratio of substitution rates varied from about 1.7-fold higher over the past 100 Myr to about 2.6-fold higher over the past 25 Myr.

The analysis shows that, although the overall density of the four transposon types in human and mouse is similar, the age distribution is strikingly different (Fig. 18). Transposon activity in the mouse genome has not undergone the decline seen in humans and proceeds at a much higher rate. In contrast to their possible extinction in humans, LTR retroposons are alive and well in the mouse with such representatives as the active IAP family and putatively active members of the long-lived ERVL and MaLR families. LINE1 and a variety of SINEs are quite active. These evolutionary findings are consistent with the empirical observations that new spontaneous mutations are 30 times more likely to be caused by LINE insertions in mouse than in human (~3% versus 0.1%)[170] and 60 times more likely to be caused by transposable elements in general. It is estimated that around 1 in 600 mutations in human are due to transpositions, whereas 10% of mutations in mouse are due to transpositions (mostly IAP insertions).

The contrast between human and mouse suggests that the explanation for the decline of transposon activity in humans may lie in some fundamental difference between hominids and rodents. Population structure and dynamics would seem to be likely suspects. Rodents tend to have large populations, whereas hominid populations tend to be small and may undergo frequent bottlenecks. Evolutionary forces affected by such factors include inbreeding and genetic drift, which might affect the persistence of active transposable elements[171]. Studies in additional mammalian lineages may shed light on the forces responsible for the differences in the activity of transposable elements[172].

Variation in the distribution of repeats. We next explored variation in the distribution of repeats across the draft genome sequence, by calculating the repeat density in windows of various sizes across the genome. There is striking variation at smaller scales.

Some regions of the genome are extraordinarily dense in repeats. The prizewinner appears to be a 525-kb region on chromosome Xp11, with an overall transposable element density of 89%. This region contains a 200-kb segment with 98% density, as well as a segment of 100 kb in which LINE1 sequences alone comprise 89% of the sequence. In addition, there are regions of more than 100 kb with extremely high densities of Alu (> 56% at three loci, including one on 7q11 with a 50-kb stretch of > 61% Alu) and the ancient transposons MIR (> 15% on chromosome 1p36) and LINE2 (> 18% on chromosome 22q12).

In contrast, some genomic regions are nearly devoid of repeats. The absence of repeats may be a sign of large-scale *cis*-regulatory elements that cannot tolerate being interrupted by insertions. The four regions with the lowest density of interspersed repeats in the human genome are the four homeobox gene clusters, HOXA, HOXB, HOXC and HOXD (Fig. 21). Each locus contains regions of around 100 kb containing less than 2% interspersed repeats. Ongoing sequence analysis of the four HOX clusters in mouse, rat and baboon shows a similar absence of transposable elements, and reveals a high density of conserved noncoding elements (K. Dewar and B. Birren, manuscript in preparation). The presence of a complex collection of regulatory regions may explain why individual HOX genes carried in transgenic mice fail to show proper regulation.

It may be worth investigating other repeat-poor regions, such as a region on chromosome 8q21 (1.5% repeat over 63 kb) containing a gene encoding a homeodomain zinc-finger protein (homologous to mouse pID 9663936), a region on chromosome 1p36 (5% repeat over 100 kb) with no obvious genes and a region on chromosome 18q22 (4% over 100 kb) containing three genes of unknown function (among which is KIAA0450). It will be interesting to see whether the homologous regions in the mouse genome have

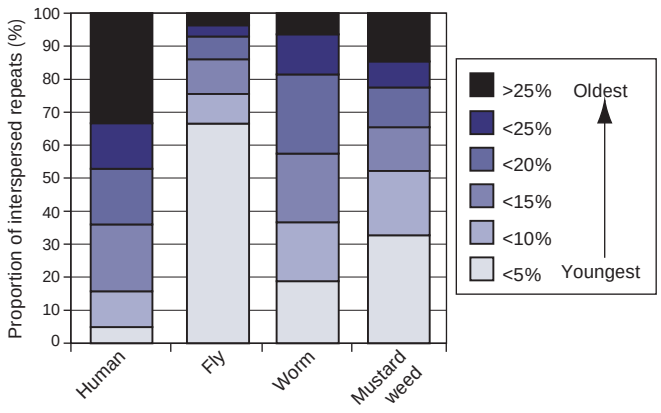


Figure 20 Comparison of the age of interspersed repeats in eukaryotic genomes. The copies of repeats were pooled by their nucleotide substitution level from the consensus.

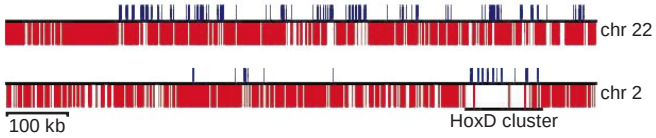


Figure 21 Two regions of about 1 Mb on chromosomes 2 and 22. Red bars, interspersed repeats; blue bars, exons of known genes. Note the deficit of repeats in the HoxD cluster, which contains a collection of genes with complex, interrelated regulation.

© 2001 Macmillan Magazines Ltd

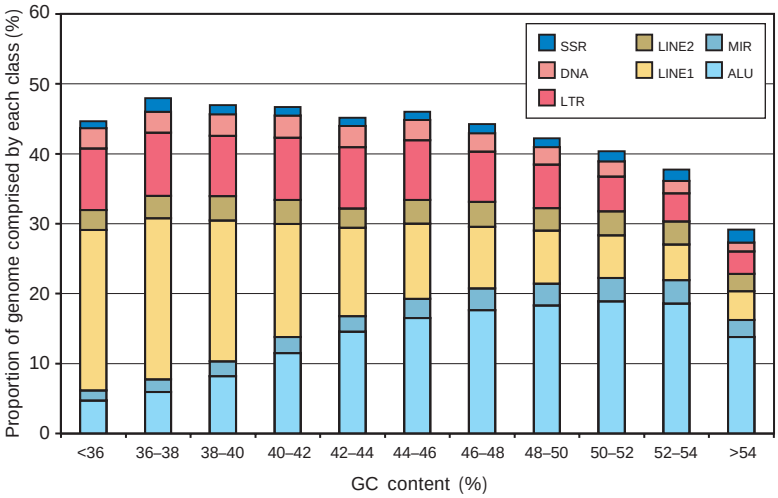


Figure 22 Density of the major repeat classes as a function of local GC content, in windows of 50 kb.

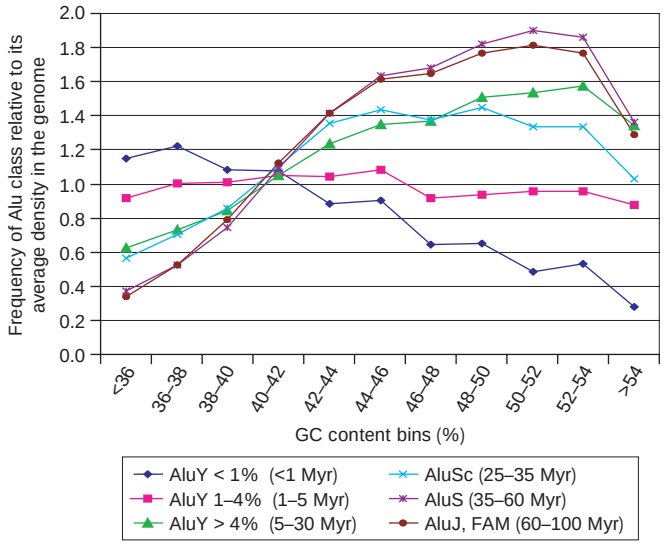


Figure 23 Alu elements target AT-rich DNA, but accumulate in GC-rich DNA. This graph shows the relative distribution of various Alu cohorts as a function of local GC content. The divergence levels (including CpG sites) and ages of the cohorts are shown in the key.

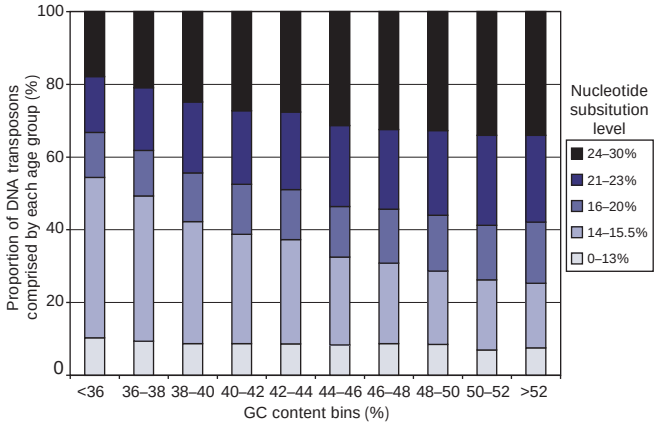


Figure 24 DNA transposon copies in AT-rich DNA tend to be younger than those in more GC-rich DNA. DNA transposon families were grouped into five age categories by their median substitution level (see Fig. 19). The proportion attributed to each age class is shown as a function of GC content. Similar patterns are seen for LINE1 and LTR elements.

similarly resisted the insertion of transposable elements during rodent evolution.

Distribution by GC content. We next focused on the correlation between the nature of the transposons in a region and its GC content. We calculated the density of each repeat type as a function of the GC content in 50-kb windows (Fig. 22). As has been reported[142,173–176], LINE sequences occur at much higher density in AT-rich regions (roughly fourfold enriched), whereas SINEs (MIR, Alu) show the opposite trend (for Alu, up to fivefold lower in AT-rich DNA). LTR retroposons and DNA transposons show a more uniform distribution, dipping only in the most GC-rich regions.

The preference of LINEs for AT-rich DNA seems like a reasonable way for a genomic parasite to accommodate its host, by targeting gene-poor AT-rich DNA and thereby imposing a lower mutational burden. Mechanistically, selective targeting is nicely explained by the fact that the preferred cleavage site of the LINE endonuclease is TTTT/A (where the slash indicates the point of cleavage), which is used to prime reverse transcription from the poly(A) tail of LINE RNA[177].

The contrary behaviour of SINEs, however, is baffling. How do SINEs accumulate in GC-rich DNA, particularly if they depend on the LINE transposition machinery[178]? Notably, the same pattern is seen for the Alu-like B1 and the tRNA-derived SINEs in mouse and for MIR in human[142]. One possibility is that SINEs somehow target GC-rich DNA for insertion. The alternative is that SINEs initially insert with the same proclivity for AT-rich DNA as LINEs, but that the distribution is subsequently reshaped by evolutionary forces[142,179].

We used the draft genome sequence to investigate this mystery by comparing the proclivities of young, adolescent, middle-aged and old Alus (Fig. 23). Strikingly, recent Alus show a preference for AT-rich DNA resembling that of LINEs, whereas progressively older Alus show a progressively stronger bias towards GC-rich DNA. These results indicate that the GC bias must result from strong pressure: Fig. 23 shows that a 13-fold enrichment of Alus in GC-rich DNA has occurred within the last 30 Myr, and possibly more recently.

These results raise a new mystery. What is the force that produces the great and rapid enrichment of Alus in GC-rich DNA? One explanation may be that deletions are more readily tolerated in gene-poor AT-rich regions than in gene-rich GC-rich regions, resulting in older elements being enriched in GC-rich regions. Such an enrichment is seen for transposable elements such as

© 2001 Macmillan Magazines Ltd

DNA transposons (Fig. 24). However, this effect seems too slow and too small to account for the observed remodelling of the Alu distribution. This can be seen by performing a similar analysis for LINE elements (Fig. 25). There is no significant change in the LINE distribution over the past 100 Myr, in contrast to the rapid change seen for Alu. There is an eventual shift after more than 100 Myr, although its magnitude is still smaller than seen for Alus.

These observations indicate that there may be some force acting particularly on Alus. This could be a higher rate of random loss of Alus in AT-rich DNA, negative selection against Alus in AT-rich DNA or positive selection in favour of Alus in GC-rich DNA. The first two possibilities seem unlikely because AT-rich DNA is gene-poor and tolerates the accumulation of other transposable elements. The third seems more feasible, in that it involves selecting in favour of the minority of Alus in GC-rich regions rather than against the majority that lie in AT-rich regions. But positive selection for Alus in GC-rich regions would imply that they benefit the organism.

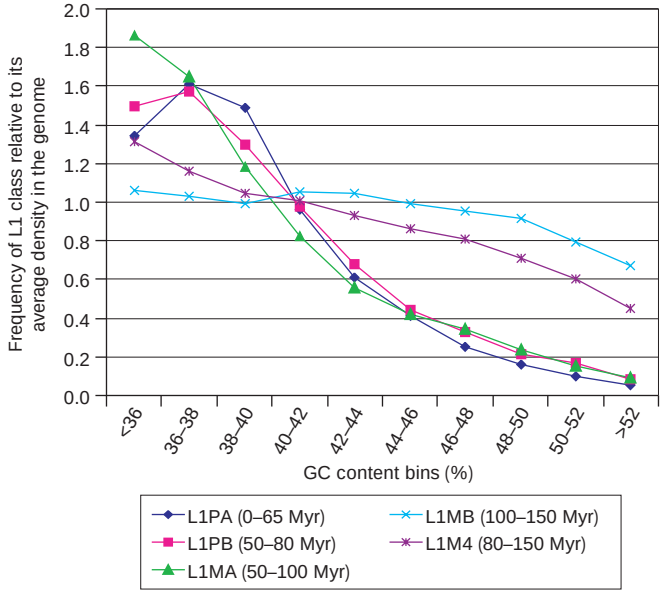


Figure 25 Distribution of various LINE cohorts as a function of local GC content. The divergence levels and ages of the cohorts are shown in the key. (The divergence levels were measured for the 3′ UTR of the LINE1 element only, which is best characterized evolutionarily. This region contains almost no CpG sites, and thus 1% divergence level corresponds to a much longer time than for CpG-rich Alu copies).

Schmid[180] has proposed such a function for SINEs. This hypothesis is based on the observation that in many species SINEs are transcribed under conditions of stress, and the resulting RNAs specifically bind a particular protein kinase (PKR) and block its ability to inhibit protein translation[181–183]. SINE RNAs would thus promote protein translation under stress. SINE RNA may be well suited to such a role in regulating protein translation, because it can be quickly transcribed in large quantities from thousands of elements and it can function without protein translation. Under this theory, there could be positive selection for SINEs in readily transcribed open chromatin such as is found near genes. This could explain the retention of Alus in gene-rich GC-rich regions. It is also consistent with the observation that SINE density in AT-rich DNA is higher near genes[142].

Further insight about Alus comes from the relationship between Alu density and GC content on individual chromosomes (Fig. 26). There are two outliers. Chromosome 19 is even richer in Alus than predicted by its (high) GC content; the chromosome comprises 2% of the genome, but contains 5% of Alus. On the other hand, chromosome Y shows the lowest density of Alus relative to its GC content, being higher than average for GC content less than 40% and lower than average for GC content over 40%. Even in AT-rich DNA, Alus are under-represented on chromosome Y compared with other young interspersed repeats (see below). These phenomena may be related to an unusually high gene density on chromosome 19 and an unusually low density of somatically active genes on chromosome Y (both relative to GC content). This would be consistent with the idea that Alu correlates not with GC content but with actively transcribed genes.

Our results may support the controversial idea that SINEs actually earn their keep in the genome. Clearly, much additional work will be needed to prove or disprove the hypothesis that SINEs are genomic symbionts.

Biases in human mutation. Indirect studies have suggested that nucleotide substitution is not uniform across mammalian

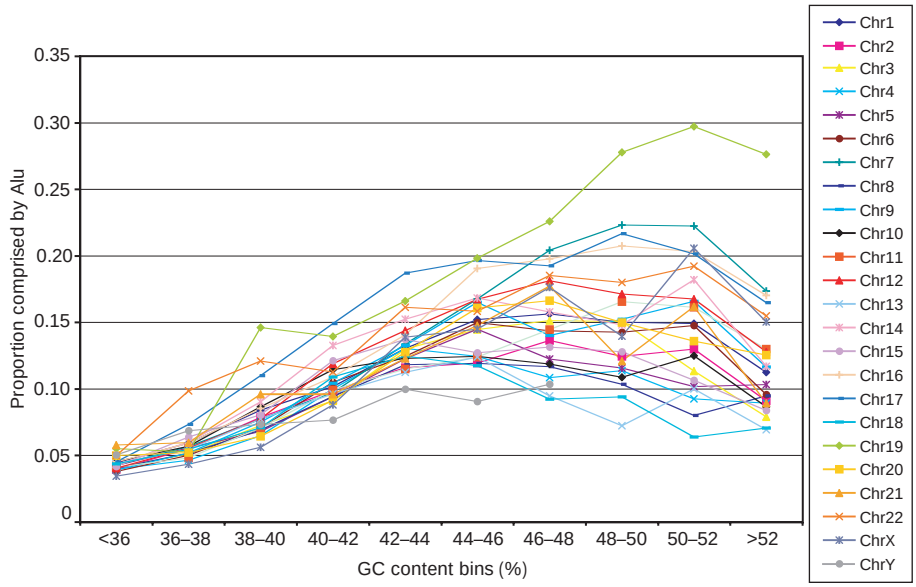


Figure 26 Comparison of the Alu density of each chromosome as a function of local GC content. At higher GC levels, the Alu density varies widely between chromosomes, with chromosome 19 being a particular outlier. In contrast, the LINE1 density pattern is quite uniform for most chromosomes, with the exception of a 1.5 to 2-fold over-representation in AT-rich regions of the X and Y chromosomes (not shown).

© 2001 Macmillan Magazines Ltd

genomes[184–187]. By studying sets of repeat elements belonging to a common cohort, one can directly measure nucleotide substitution rates in different regions of the genome. We find strong evidence that the pattern of neutral substitution differs as a function of local GC content (Fig. 27). Because the results are observed in repetitive elements throughout the genome, the variation in the pattern of nucleotide substitution seems likely to be due to differences in the underlying mutational process rather than to selection.

The effect can be seen most clearly by focusing on the substitution process $\gamma \leftrightarrow \alpha$, where γ denotes GC or CG base pairs and α denotes AT or TA base pairs. If K is the equilibrium constant in the direction of α base pairs (defined by the ratio of the forward and reverse rates), then the equilibrium GC content should be $1/(1+K)$. Two observations emerge.

First, there is a regional bias in substitution patterns. The equilibrium constant varies as a function of local GC content: γ base pairs are more likely to mutate towards α base pairs in AT-rich regions than in GC-rich regions. For the analysis in Fig. 27, the equilibrium constant K is 2.5, 1.9 and 1.2 when the draft genome sequence is partitioned into three bins with average GC content of 37, 43 and 50%, respectively. This bias could be due to a reported tendency for GC-rich regions to replicate earlier in the cell cycle than AT-rich regions and for guanine pools, which are limiting for DNA replication, to become depleted late in the cell cycle, thereby resulting in a small but significant shift in substitution towards α base pairs[186,188]. Another theory proposes that many substitutions are due to differences in DNA repair mechanisms, possibly related to transcriptional activity and thereby to gene density and GC content[185,189,190].

There is also an absolute bias in substitution patterns resulting in directional pressure towards lower GC content throughout the human genome. The genome is not at equilibrium with respect to the pattern of nucleotide substitution: the expected equilibrium GC content corresponding to the values of K above is 29, 35 and 44% for regions with average GC contents of 37, 43 and 50%, respectively. Recent observations on SNPs[190] confirm that the mutation pattern in GC-rich DNA is biased towards α base pairs; it should be possible to perform similar analyses throughout the genome with the availability of 1.4 million SNPs[97,191]. On the basis solely of nucleotide substitution patterns, the GC content would be expected to be about 7% lower throughout the genome.

What accounts for the higher GC content? One possible explanation is that in GC-rich regions, a considerable fraction of the nucleotides is likely to be under functional constraint owing to the high gene density. Selection on coding regions and regulatory CpG islands may maintain the higher-than-predicted GC content. Another is that throughout the rest of the genome, a constant influx of transposable elements tends to increase GC content (Fig. 28). Young repeat elements clearly have a higher GC content than their surrounding regions, except in extremely GC-rich regions. Moreover, repeat elements clearly shift with age towards a lower GC content, closer to that of the neighbourhood in which they reside. Much of the 'non-repeat' DNA in AT-rich regions probably consists of ancient repeats that are not detectable by current methods and that have had more time to approach the local equilibrium value.

The repeats can also be used to study how the mutation process is affected by the immediately adjacent nucleotide. Such 'context effects' will be discussed elsewhere (A. Kas and A. F. A. Smit, unpublished results).

Fast living on chromosome Y. The pattern of interspersed repeats can be used to shed light on the unusual evolutionary history of chromosome Y. Our analysis shows that the genetic material on

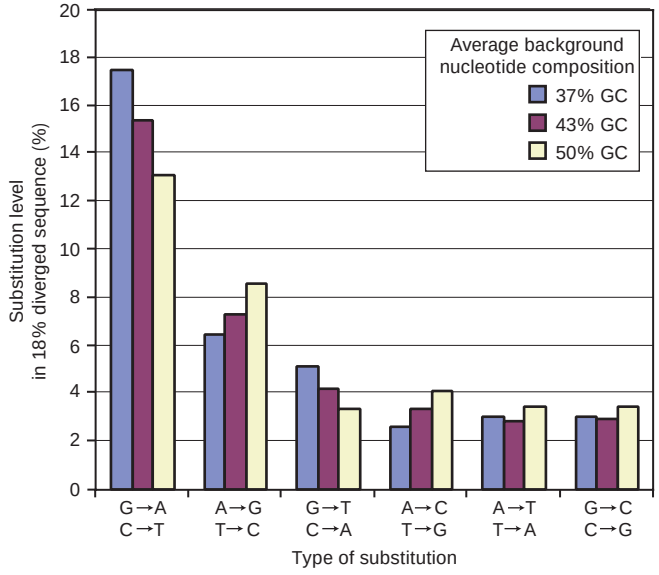


Figure 27 Substitution patterns in interspersed repeats differ as a function of GC content. We collected all copies of five DNA transposons (Tigger1, Tigger2, Charlie3, MER1 and HSMAR2), chosen for their high copy number and well defined consensus sequences. DNA transposons are optimal for the study of neutral substitutions: they do not segregate into subfamilies with diagnostic differences, presumably because they are short-lived and new active families do not evolve in a genome (see text). Duplicates and close paralogues resulting from duplication after transposition were eliminated. The copies were grouped on the basis of GC content of the flanking 1,000 bp on both sides and aligned to the consensus sequence (representing the state of the copy at integration). Recursive efforts using parameters arising from this study did not change the alignments significantly. Alignments were inspected by hand, and obvious misalignments caused by insertions and duplications were eliminated. Substitutions ($n = 80,000$) were counted for each position in the consensus, excluding those in CpG dinucleotides, and a substitution frequency matrix was defined. From the matrices for each repeat (which corresponded to different ages), a single rate matrix was calculated for these bins of GC content (< 40% GC, 40–47% GC and > 47% GC). Data are shown for a repeat with an average divergence (in non-CpG sites) of 18% in 43% GC content (the repeat has slightly higher divergence in AT-rich DNA and lower in GC-rich DNA). From the rate matrix, we calculated log-likelihood matrices with different entropies (divergence levels), which are theoretically optimal for alignments of neutrally diverged copies to their common ancestral state (A. Kas and A. F. A. Smit, unpublished). These matrices are in use by the RepeatMasker program.

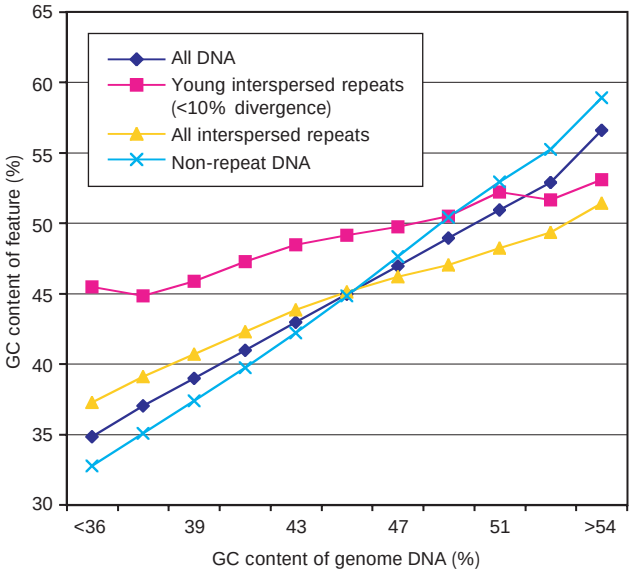


Figure 28 Interspersed repeats tend to diminish the differences between GC bins, despite the fact that GC-rich transposable elements (specifically Alu) accumulate in GC-rich DNA, and AT-rich elements (LINE1) in AT-rich DNA. The GC content of particular components of the sequence (repeats, young repeats and non-repeat sequence) was calculated as a function of overall GC content.

© 2001 Macmillan Magazines Ltd

chromosome Y is unusually young, probably owing to a high tolerance for gain of new material by insertion and loss of old material by deletion. Several lines of evidence support this picture. For example, LINE elements on chromosome Y are on average much younger than those on autosomes (not shown). Similarly, MaLR-family retroposons on chromosome Y are younger than those on autosomes, with the representation of subfamilies showing a strong inverse correlation with the age of the subfamily. Moreover, chromosome Y has a relative over-representation of the younger retroviral class II (ERVK) and a relative under-representation of the primarily older class III (ERVL) compared with other chromosomes. Overall, chromosome Y seems to maintain a youthful appearance by rapid turnover.

Interspersed repeats on chromosome Y can also be used to estimate the relative mutation rates, α_m and α_f, in the male and female germlines. Chromosome Y always resides in males, whereas chromosome X resides in females twice as often as in males. The substitution rates, μ_Y and μ_X, on these two chromosomes should thus be in the ratio $\mu_Y:\mu_X = (\alpha_m):(\alpha_m + 2\alpha_f)/3$, provided that one considers equivalent neutral sequences. Several authors have estimated the mutation rate in the male germline to be fivefold higher than in the female germline, by comparing the rates of evolution of X- and Y-linked genes in humans and primates. However, Page and colleagues[192] have challenged these estimates as too high. They studied a 39-kb region that is apparently devoid of genes and resides within a large segmental duplication from X to Y that occurred 3–4 Myr ago in the human lineage. On the basis of phylogenetic analysis of the sequence on human Y and human, chimp and gorilla X, they obtained a much lower estimate of $\mu_Y:\mu_X = 1.36$, corresponding to $\alpha_m:\alpha_f = 1.7$. They suggested that the other estimates may have been higher because they were based on much longer evolutionary periods or because the genes studied may have been under selection.

Our database of human repeats provides a powerful resource for addressing this question. We identified the repeat elements from recent subfamilies (effectively, birth cohorts dating from the past 50 Myr) and measured the substitution rates for subfamily members on chromosomes X and Y (Fig. 29). There is a clear linear relationship with a slope of $\mu_Y:\mu_X = 1.57$ corresponding to $\alpha_m:\alpha_f = 2.1$. The estimate is in reasonable agreement with that of Page *et al.*, although it is based on much more total sequence (360 kb on Y, 1.6 Mb on X) and a much longer time period. In particular, the discrepancy with earlier reports is not explained by recent changes in the human lineage. Various theories have been proposed for the higher mutation rate in the male germline, including the greater number of cell divisions in the formation of sperm than eggs and different repair mechanisms in sperm and eggs.

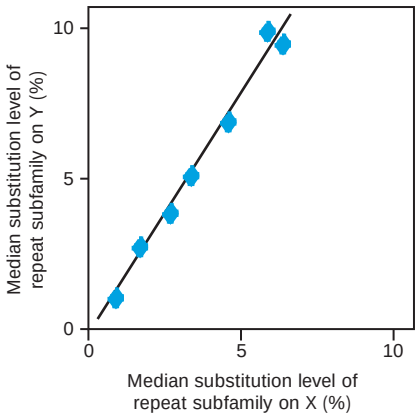


Figure 29 Higher substitution rate on chromosome Y than on chromosome X. We calculated the median substitution level (excluding CpG sites) for copies of the most recent L1 subfamilies (L1Hs–L1PA8) on the X and Y chromosomes. Only the 3′ UTR of the L1 element was considered because its consensus sequence is best established.

Active transposons. We were interested in identifying the youngest retrotransposons in the draft genome sequence. This set should contain the currently active retrotransposons, as well as the insertion sites that are still polymorphic in the human population.

The youngest branch in the phylogenetic tree of human LINE1 elements is called L1Hs (ref. 158); it differs in its 3′ untranslated region (UTR) by 12 diagnostic substitutions from the next oldest subfamily (L1PA2). Within the L1Hs family, there are two subsets referred to as Ta and pre-Ta, defined by a diagnostic trinucleotide[193,194]. All active L1 elements are thought to belong to these two subsets, because they account for all 14 known cases of human disease arising from new L1 transposition (with 13 belonging to the Ta subset and one to the pre-Ta subset)[195,196]. These subsets are also of great interest for population genetics because at least 50% are still segregating as polymorphisms in the human population[194,197]; they provide powerful markers for tracing population history because they represent unique (non-recurrent and non-revertible) genetic events that can be used (along with similarly polymorphic Alus) for reconstructing human migrations.

LINE1 elements that are retrotransposition-competent should consist of a full-length sequence and should have both ORFs intact. Eleven such elements from the Ta subset have been identified, including the likely progenitors of mutagenic insertions into the factor VIII and dystrophin genes[198–202]. A cultured cell retrotransposition assay has revealed that eight of these elements remain retrotransposition-competent[200,202,203].

We searched the draft genome sequence and identified 535 LINEs belonging to the Ta subset and 415 belonging to the pre-Ta subset. These elements provide a large collection of tools for probing human population history. We also identified those consisting of full-length elements with intact ORFs, which are candidate active LINEs. We found 39 such elements belonging to the Ta subset and 22 belonging to the pre-Ta subset; this substantially increases the number in the first category and provides the first known examples in the second category. These elements can now be tested for retrotransposition competence in the cell culture assay. Preliminary analysis resulted in the identification of two of these elements as the likely progenitors of mutagenic insertions into the β-globin and RP2 genes (R. Badge and J. V. Moran, unpublished data). Similar analyses should allow the identification of the progenitors of most, if not all, other known mutagenic L1 insertions.

L1 elements can carry extra DNA if transcription extends through the native transcriptional termination site into flanking genomic DNA. This process, termed L1-mediated transduction, provides a means for the mobilization of DNA sequences around the genome and may be a mechanism for 'exon shuffling'[204]. Twenty-one per cent of the 71 full-length L1s analysed contained non-L1-derived sequences before the 3′ target-site duplication site, in cases in which the site was unambiguously recognizable. The length of the transduced sequence was 30–970 bp, supporting the suggestion that 0.5–1.0% of the human genome may have arisen by LINE-based transduction of 3′ flanking sequences[205,206].

Our analysis also turned up two instances of 5′ transduction (145 bp and 215 bp). Although this possibility had been suggested on the basis of cell culture models[195,203], these are the first documented examples. Such events may arise from transcription initiating in a cellular promoter upstream of the L1 elements. L1 transcription is generally confined to the germline[207,208], but transcription from other promoters could explain a somatic L1 retrotransposition event that resulted in colon cancer[206].

Transposons as a creative force. The primary force for the origin and expansion of most transposons has been selection for their ability to create progeny, and not a selective advantage for the host. However, these selfish pieces of DNA have been responsible for important innovations in many genomes, for example by contributing regulatory elements and even new genes.

Twenty human genes have been recognized as probably derived

© 2001 Macmillan Magazines Ltd

from transposons[142,209]. These include the RAG1 and RAG2 recombinases and the major centromere-binding protein CENPB. We scanned the draft genome sequence and identified another 27 cases, bringing the total to 47 (Table 13; refs 142, 209). All but four are derived from DNA transposons, which give rise to only a small proportion of the interspersed repeats in the genome. Why there are so many DNA transposase-like genes, many of which still contain the critical residues for transposase activity, is a mystery.

To illustrate this concept, we describe the discovery of one of the new examples. We searched the draft genome sequence to identify the autonomous DNA transposon responsible for the distribution of the non-autonomous MER85 element, one of the most recently (40–50 Myr ago) active DNA transposons. Most non-autonomous elements are internal deletion products of a DNA transposon. We identified one instance of a large (1,782 bp) ORF flanked by the 5′ and 3′ halves of a MER85 element. The ORF encodes a novel protein (partially published as pID 6453533) whose closest homologue is the transposase of the piggyBac DNA transposon, which is found in insects and has the same characteristic TTAA target-site duplications[210] as MER85. The ORF is actively transcribed in fetal brain and in cancer cells. That it has not been lost to mutation in 40–50 Myr of evolution (whereas the flanking, noncoding, MER85-like termini show the typical divergence level of such elements) and is actively transcribed provides strong evidence that it has been adopted by the human genome as a gene. Its function is unknown.

LINE1 activity clearly has also had fringe benefits. We mentioned above the possibility of exon reshuffling by cotranscription of neighbouring DNA. The LINE1 machinery can also cause reverse transcription of genic mRNAs, which typically results in nonfunctional processed pseudogenes but can, occasionally, give rise to functional processed genes. There are at least eight human and eight mouse genes for which evidence strongly supports such an origin[211] (see http://www-ifi.uni-muenster.de/exapted-retrogenes/tables.html). Many other intronless genes may have been created in the same way.

Transposons have made other creative contributions to the genome. A few hundred genes, for example, use transcriptional terminators donated by LTR retroposons (data not shown). Other genes employ regulatory elements derived from repeat elements[211].

Simple sequence repeats

Simple sequence repeats (SSRs) are a rather different type of repetitive structure that is common in the human genome—perfect or slightly imperfect tandem repeats of a particular *k*-mer. SSRs with a short repeat unit (n = 1–13 bases) are often termed microsa-

Table 13 Human genes derived from transposable elements

GenBank ID*	Gene name	Related transposon family†	Possible fusion gene§	Newly recognized derivation‖
nID 3150436	BC200	FLAM Alu‡		
pID 2330017	Telomerase	non-LTR retrotransposon		
pID 1196425	HERV-3 env	Retroviridae/HERV-R‡		
pID 4773880	Syncytin	Retroviridae/HERV-W‡		
pID 131827	RAG1 and 2	Tc1-like		
pID 29863	CENP-B	Tc1/Pogo		
EST 2529718		Tc1/Pogo		+
PID 10047247		Tc1/Pogo/Pogo		+
EST 4524463		Tc1/Pogo/Pogo		+
pID 4504807	Jerky	Tc1/Pogo/Tigger		
pID 7513096	JRKL	Tc1/Pogo/Tigger		
EST 5112721		Tc1/Pogo/Tigger		+
EST 11097233		Tc1/Pogo/Tigger		+
EST 6986275	Sancho	Tc1/Pogo/Tigger		
EST 8616450		Tc1/Pogo/Tigger		+
EST 8750408		Tc1/Pogo/Tigger		+
EST 5177004		Tc1/Pogo/Tigger		+
PID 3413884	KIAA0461	Tc1/Pogo/Tc2	+	
PID 7959287	KIAA1513	Tc1/Pogo/Tc2		+
PID 2231380		Tc1/Mariner/Hsmar1‡	+	
EST 10219887		hAT/Hobo	+	+
PID 6581095	Buster1	hAT/Charlie	+	
PID 7243087	Buster2	hAT/Charlie	+	
PID 6581097	Buster3	hAT/Charlie		
PID 7662294	KIAA0766	hAT/Charlie	+	
PID 10439678		hAT/Charlie		+
PID 7243087	KIAA1353	hAT/Charlie		+
PID 7021900		hAT/Charlie/Charlie3‡		+
PID 4263748		hAT/Charlie/Charlie8‡	+	
EST 8161741		hAT/Charlie/Charlie9‡		+
pID 4758872	DAP4,pP52[rIPK]	hAT/Tip100/Zaphod		
EST 10990063		hAT/Tip100/Zaphod		+
EST 10101591		hAT/Tip100/Zaphod		+
pID 7513011	KIAA0543	hAT/Tip100/Tip100	+	
pID 10439744		hAT/Tip100/Tip100		+
pID 10047247	KIAA1586	hAT/Tip100/Tip100		+
pID 10439762		hAT/Tip100	+	+
EST 10459804		hAT/Tip100		+
pID 4160548	Tramp	hAT/Tam3		+
BAC 3522927		hAT/Tam3		+
pID 3327088	KIAA0637	hAT/Tam3	+	
EST 1928552		hAT/Tam3		+
pID 6453533		piggyBac/MER85‡	+	
EST 3594004		piggyBac/MER85‡		+
BAC 4309921		piggyBac/MER85‡		+
EST 4073914		piggyBac/MER75‡		+
EST 1963278		piggyBac		+

The Table lists 47 human genes, with a likely origin in up to 38 different transposon copies.
* Where available, the GenBank ID numbers are given for proteins, otherwise a representative EST or a clone name is shown. Six groups (two or three genes each) have similarity at the DNA level well beyond that observed between different DNA transposon families in the genome; they are indicated in italics, with all but the initial member of each group indented. This could be explained if the genes were paralogous (derived from a single inserted transposon and subsequently duplicated).
† Classification of the transposon.
‡ Indicates that the transposon from which the gene is derived is precisely known.
§ Proteins probably formed by fusion of a cellular and transposon gene; many have acquired zinc-finger domains.
‖ Not previously reported as being derived from transposable element genes. The remaining genes can be found in refs 142, 209.

© 2001 Macmillan Magazines Ltd

Table 14 SSR content of the human genome

Length of repeat unit	Average bases per Mb	Average number of SSR elements per Mb
1	1,660	36.7
2	5,046	43.1
3	1,013	11.8
4	3,383	32.5
5	2,686	17.6
6	1,376	15.2
7	906	8.4
8	1,139	11.1
9	900	8.6
10	1,576	8.6
11	770	8.7

SSRs were identified by using the computer program Tandem Repeat Finder with the following parameters: match score 2, mismatch score 3, indel 5, minimum alignment 50, maximum repeat length 500, minimum repeat length 1.

tellites, whereas those with longer repeat units (n = 14–500 bases) are often termed minisatellites. With the exception of poly(A) tails from reverse transcribed messages, SSRs are thought to arise by slippage during DNA replication[212,213].

We compiled a catalogue of all SSRs over a given length in the human draft genome sequence, and studied their properties (Table 14). SSRs comprise about 3% of the human genome, with the greatest single contribution coming from dinucleotide repeats (0.5%). (The precise criteria for the number of repeat units and the extent of divergence allowed in an SSR affect the exact census, but not the qualitative conclusions.)

There is approximately one SSR per 2 kb (the number of non-overlapping tandem repeats is 437 per Mb). The catalogue confirms various properties of SSRs that have been inferred from sampling approaches (Table 15). The most frequent dinucleotide repeats are AC and AT (50 and 35% of dinucleotide repeats, respectively), whereas AG repeats (15%) are less frequent and GC repeats (0.1%) are greatly under-represented. The most frequent trinucleotides are AAT and AAC (33% and 21%, respectively), whereas ACC (4.0%), AGC (2.2%), ACT (1.4%) and ACG (0.1%) are relatively rare. Overall, trinucleotide SSRs are much less frequent than dinucleotide SSRs[214].

SSRs have been extremely important in human genetic studies, because they show a high degree of length polymorphism in the human population owing to frequent slippage by DNA polymerase during replication. Genetic markers based on SSRs—particularly $(CA)_n$ repeats—have been the workhorse of most human disease-mapping studies[101,102]. The availability of a comprehensive catalogue of SSRs is thus a boon for human genetic studies.

The SSR catalogue also allowed us to resolve a mystery regarding mammalian genetic maps. Such genetic maps in rat, mouse and human have a deficit of polymorphic $(CA)_n$ repeats on chromosome X[30,101]. There are two possible explanations for this deficit. There may simply be fewer $(CA)_n$ repeats on chromosome X; or $(CA)_n$ repeats may be as dense on chromosome X but less polymorphic in the population. In fact, analysis of the draft genome sequence shows that chromosome X has the same density of $(CA)_n$ repeats per Mb as the autosomes (data not shown). Thus, the deficit of polymorphic markers relative to autosomes results from population genetic forces. Possible explanations include that chromosome X has a smaller effective population size, experiences more frequent selective sweeps reducing diversity (owing to its hemizygosity in males), or has a lower mutation rate (owing to its more frequent passage through the less mutagenic female germline). The availability of the draft genome sequence should provide ways to test these alternative explanations.

Table 15 SSRs by repeat unit

Repeat unit	Number of SSRs per Mb
AC	27.7
AT	19.4
AG	8.2
GC	0.1
AAT	4.1
AAC	2.6
AGG	1.5
AAG	1.4
ATG	0.7
CGG	0.6
ACC	0.4
AGC	0.3
ACT	0.2
ACG	0.0

SSRs were identified as in Table 14.

Segmental duplications

A remarkable feature of the human genome is the segmental duplication of portions of genomic sequence[215–217]. Such duplications involve the transfer of 1–200-kb blocks of genomic sequence to one or more locations in the genome. The locations of both donor and recipient regions of the genome are often not tandemly arranged, suggesting mechanisms other than unequal crossing-over for their origin. They are relatively recent, inasmuch as strong sequence identity is seen in both exons and introns (in contrast to regions that are considered to show evidence of ancient duplications, characterized by similarities only in coding regions). Indeed, many such duplications appear to have arisen in very recent evolutionary time, as judged by high sequence identity and by their absence in closely related species.

Segmental duplications can be divided into two categories. First, interchromosomal duplications are defined as segments that are duplicated among nonhomologous chromosomes. For example, a 9.5-kb genomic segment of the adrenoleukodystrophy locus from Xq28 has been duplicated to regions near the centromeres of chromosomes 2, 10, 16 and 22 (refs 218, 219). Anecdotal observations suggest that many interchromosomal duplications map near the centromeric and telomeric regions of human chromosomes[218–233].

The second category is intrachromosomal duplications, which occur within a particular chromosome or chromosomal arm. This category includes several duplicated segments, also known as low copy repeat sequences, that mediate recurrent chromosomal structural rearrangements associated with genetic disease[215,217]. Examples on chromosome 17 include three copies of a roughly 200-kb repeat separated by around 5 Mb and two copies of a roughly 24-kb repeat separated by 1.5 Mb. The copies are so similar (99% identity) that paralogous recombination events can occur, giving rise to contiguous gene syndromes: Smith–Magenis syndrome and Charcot–Marie–Tooth syndrome 1A, respectively[34,234]. Several other examples are known and are also suspected to be responsible for recurrent microdeletion syndromes (for example, Prader–Willi/Angelman,

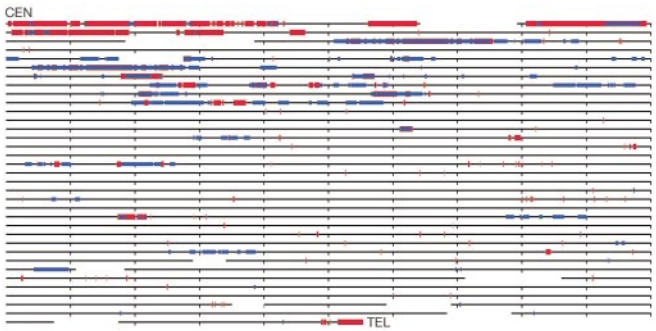


Figure 30 Duplication landscape of chromosome 22. The size and location of intrachromosomal (blue) and interchromosomal (red) duplications are depicted for chromosome 22q, using the PARASIGHT computer program (Bailey and Eichler, unpublished). Each horizontal line represents 1 Mb (ticks, 100-kb intervals). The chromosome sequence is oriented from centromere (top left) to telomere (bottom right). Pairwise alignments with > 90% nucleotide identity and > 1 kb long are shown. Gaps within the chromosomal sequence are of known size and shown as empty space.

© 2001 Macmillan Magazines Ltd

velocardiofacial/DiGeorge and Williams' syndromes[215,235–240]).

Until now, the identification and characterization of segmental duplications have been based on anecdotal reports—for example, finding that certain probes hybridize to multiple chromosomal sites or noticing duplicated sequence at certain recurrent chromosomal breakpoints. The availability of the entire genomic sequence will make it possible to explore the nature of segmental duplications more systematically. This analysis can begin with the current state of the draft genome sequence, although caution is required because some apparent duplications may arise from a failure to merge sequence contigs from overlapping clones. Alternatively, erroneous assembly of closely related sequences from nonoverlapping clones may underestimate the true frequency of such features, particularly among those segments with the highest sequence similarity. Accordingly, we adopted a conservative approach for estimating such duplication from the available draft genome sequence.

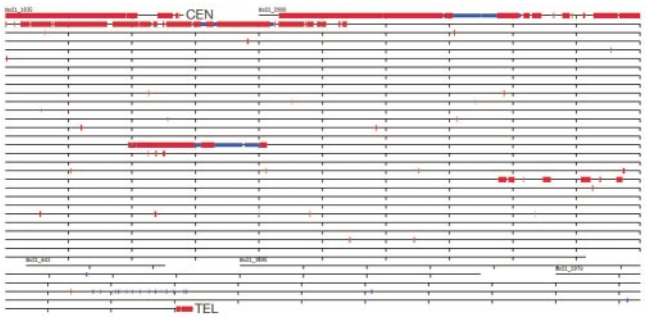


Figure 31 Duplication landscape of chromosome 21. The size and location of intrachromosomal (blue) and interchromosomal (red) duplications are depicted along the sequence of the long arm of chromosome 21. Gaps between finished sequence are denoted by empty space but do not represent actual gap size.

Pericentromeres and subtelomeres. We began by re-evaluating the finished sequences of chromosomes 21 and 22. The initial papers on these chromosomes[93,94] noted some instances of interchromosomal duplication near each centromere. With the ability now to compare these chromosomes to the vast majority of the genome, it is apparent that the regions near the centromeres consist almost entirely of interchromosomal duplicated segments, with little or no unique sequence. Smaller regions of interchromosomal duplication are also observed near the telomeres.

Chromosome 22 contains a region of 1.5 Mb adjacent to the centromere in which 90% of sequence can now be recognized to consist of interchromosomal duplication (Fig. 30). Conversely, 52% of the interchromosomal duplications on chromosome 22 were located in this region, which comprises only 5% of the chromosome. Also, the subtelomeric end consists of a 50-kb region consisting almost entirely of interchromosomal duplications.

Chromosome 21 presents a similar landscape (Fig. 31). The first 1 Mb after the centromere is composed of interchromosomal repeats, as well as the largest (> 200 kb) block of intrachromosomally duplicated material. Again, most interchromosomal duplications on the chromosome map to this region and the most subtelomeric region (30 kb) shows extensive duplication among nonhomologous chromosomes.

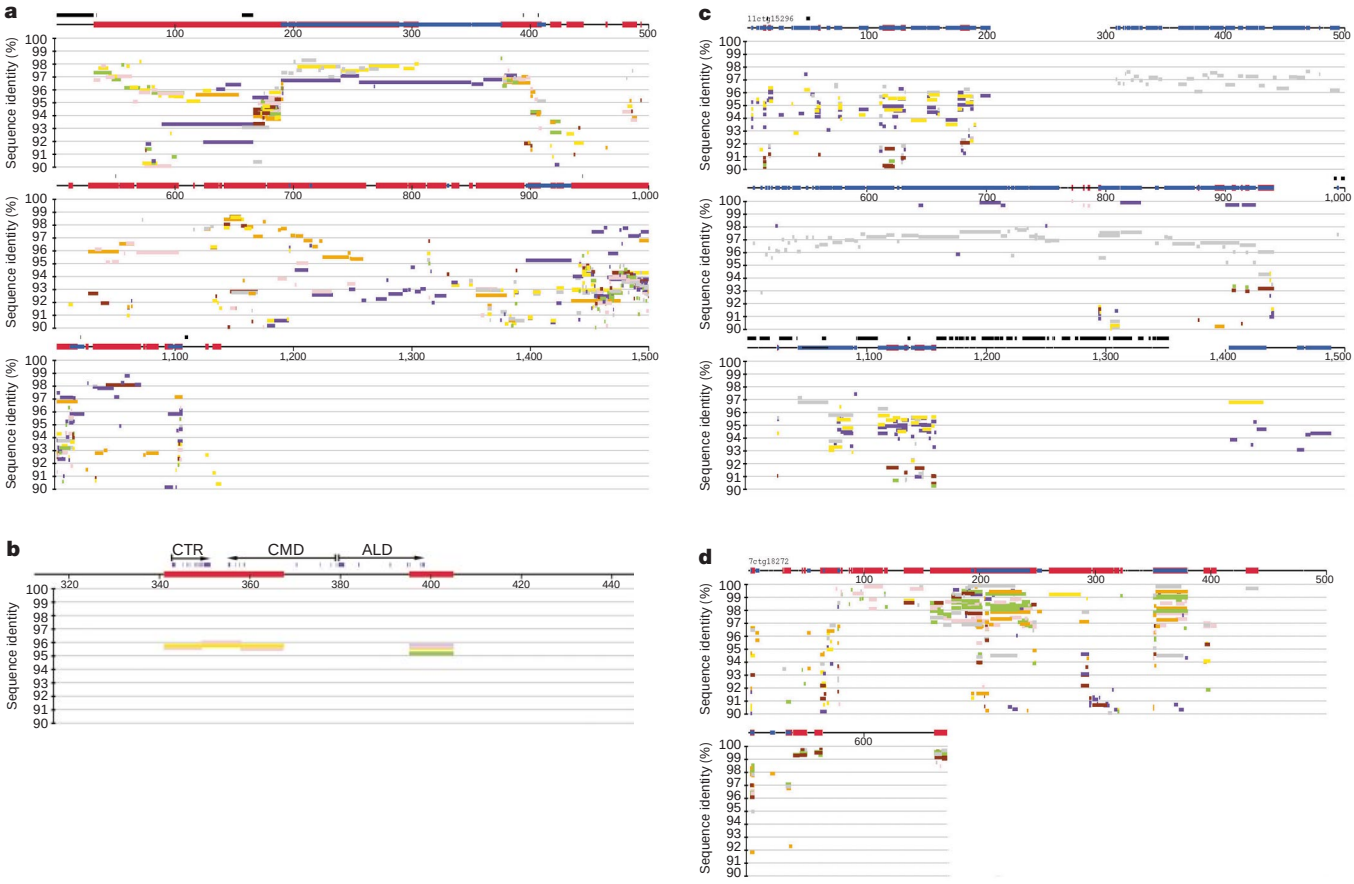


Figure 32 Mosaic patterns of duplications. Panels depict various patterns of duplication within the human genome (PARASIGHT). For each region, a segment of draft genome sequence (100–500 kb) is shown with both interchromosomal (red) and intrachromosomal (blue) duplications displayed along the horizontal line. Below the line, each separate sequence duplication is indicated (with a distinct colour) relative to per cent nucleotide identity for the duplicated segment (*y* axis). Black bars show the relative locations of large blocks of heterochromatic sequences (alpha, gamma and HSAT sequence). **a**, An active pericentromeric region on chromosome 21. **b**, An ancestral region from Xq28 that has contributed various 'genic' segments to pericentromeric regions. **c**, A pericentromeric

© 2001 Macmillan Magazines Ltd

The pericentromeric regions are structurally very complex, as illustrated for chromosome 21 in Fig. 32a. The pericentromeric regions appear to have been bombarded by successive insertions of duplications; the insertion events must be fairly recent because the degree of sequence conservation with the genomic source loci is fairly high (90–100%, with an apparent peak around 96%). Distinct insertions are typically separated by AT-rich or GC-rich minisatellite-like repeats that have been hypothesized to have a functional role in targeting duplications to these regions[233,241].

A single genomic source locus often gives rise to pericentromeric copies on multiple chromosomes, with each having essentially the same breakpoints and the same degree of divergence. An example of such a source locus on Xq28 is shown in Fig. 32b. Phylogenetic analysis has suggested a two-step mechanism for the origin and dispersal of these segments, whereby an initial segmental duplication in the pericentromeric region of one chromosome occurs and is then redistributed as part of a larger cassette to other such regions[242].

A comprehensive analysis for all chromosomes will have to await complete sequencing of the genome, but the evidence from the draft genome sequence indicates that the same picture is likely to be seen throughout the genome. Several papers have analysed finished segments within pericentromeric regions of chromosomes 2 (160 kb), 10 (400 kb) and 16 (300 kb), all of which show extensive interchromosomal segmental duplication[215,219,232,233]. An example from another pericentromeric region on chromosome 11 is shown in Fig. 32c. Interchromosomal duplications in subtelomeric regions also appear to be a fairly general phenomenon, as illustrated by a large tract (~500 kb) of complex duplication on chromosome 7 (Fig. 32d).

The explanation for the clustering of segmental duplications may be that the genome has a damage-control mechanism whereby chromosomal breakage products are preferentially inserted into pericentromeric and, to a lesser extent, subtelomeric regions. The possibility of a specific mechanism for the insertion of these sequences has been suggested on the basis of the unusual sequences found flanking the insertions. Although it is also possible that these regions simply have greater tolerance for large insertions, many large gene-poor 'deserts' have been identified[93] and there is no accumulation of duplicated segments within these regions. Along with the fact that transitions between duplicons (from different regions of the genome) occur at specific sequences, this suggests that active recruitment of duplications to such regions may occur. In any case, the duplicated regions are in general young (with many duplications showing <6% nucleotide divergence from their source loci) and in constant flux, both through additional duplications and by large-scale exchange among similar chromosomal environments. There is evidence of structural polymorphism in the human population, such as the presence or absence of olfactory receptor segments located within the telomeric regions of several human chromosomes[226,227].

Genome-wide analysis of segmental duplications. We also performed a global genome-wide analysis to characterize the amount of segmental duplication in the genome. We 'repeat-masked' the known interspersed repeats in the draft genome sequence and compared the remaining draft genomic sequence with itself in a massive all-by-all BLASTN similarity search. We excluded matches in which the sequence identity was so high that it might reflect artefactual duplications resulting from a failure to overlap sequence contigs correctly in assembling the draft genome sequence. Specifically, we considered only matches with less than 99.5% identity for finished sequence and less than 98% identity for unfinished sequence.

We took several approaches to avoid counting artefactual duplications in the sequence. In the first approach, we studied only finished sequence. We compared the finished sequence with itself, to identify segments of at least 1 kb and 90–99.5% sequence identity. This analysis will underestimate the extent of segmental duplication, because it requires that at least two copies of the segment are present in the finished sequence and because some true duplications have over 99.5% identity.

The finished sequence consists of at least 3.3% segmental duplication (Table 16). Interchromosomal duplication accounts for about 1.5% and intrachromosomal duplication for about 2%, with some overlap (0.2%) between these categories. We analysed the lengths and divergence of the segmental duplications (Fig. 33). The duplications tend to be large (10–50 kb) and highly homologous, especially for the interchromosomal segments. The sequence divergence for the interchromosomal duplications appears to peak between 96.5% and 97.5%. This may indicate that interchromosomal duplications occurred in a punctuated manner. It will be intriguing to investigate whether such genomic upheaval has a role in speciation events.

In a second approach, we compared the entire human draft genome sequence (finished and unfinished) with itself to identify duplications with 90–98% sequence identity (Table 17). The draft genome sequence contains at least 3.6% segmental duplication. The actual proportion will be significantly higher, because we excluded many true matches with more than 98% sequence identity (at least 1.1% of the finished sequence). Although exact measurement must await a finished sequence, the human genome seems likely to contain about 5% segmental duplication, with most of this sequence in large blocks (>10 kb). Such a high proportion of large duplications clearly distinguishes the human genome from other sequenced genomes, such as the fly and worm (Table 18).

The structure of large highly paralogous regions presents one of the 'serious and unanticipated challenges' to producing a finished sequence of the genome[46]. The absence of unique STS or fingerprint signatures over large genomic distances (~1 Mb) and the high degree of sequence similarity makes the distinction between paralogous sequence variation and allelic polymorphism problematic. Furthermore, the fact that such regions frequently harbour intron–exon structures of genuine unique sequence will complicate efforts to generate a genome-wide SNP map. The data indicate that a modest portion of the human genome may be relatively recalcitrant to genomic-based methods for SNP detection. Owing to their repetitive nature and their location in the genome, segmental

Table 16 Fraction of finished sequence in inter- and intrachromosomal duplications

Chromosome	Intrachromosomal (%)	Interchromosomal (%)	All (%)
1	1.4	0.5	1.9
2	0.1	0.6	0.7
3	0.3	1.1	1.1
4	0.0	1.0	1.0
5	0.6	0.3	0.9
6	0.8	0.4	1.1
7	3.4	1.3	4.1
8	0.3	0.1	0.3
9	0.8	2.9	3.7
10	2.1	0.8	2.9
11	1.2	2.1	2.3
12	1.5	0.3	1.8
13	0.0	0.5	0.5
14	0.6	0.4	1.0
15	3.0	6.9	6.9
16	4.5	2.0	5.8
17	1.6	0.3	1.8
18	0.0	0.7	0.7
19	3.6	0.3	3.8
20	0.2	0.3	0.5
21	1.4	1.6	3.0
22	6.1	2.6	7.5
X	1.8	3.2	5.0
Y	12.1	16.0	27.4
Un	0.0	0.5	0.5
Total	2.0	1.5	3.3

Excludes duplications with identities >99.5% to avoid artefactual duplication due to incomplete merger in the assembly process. Calculation was performed on the finished sequence available in September 2000 and reflects the duplications found within the total amount of finished sequence then. Note that there is some overlap between the interchromosomal and intrachromosomal sets.

© 2001 Macmillan Magazines Ltd

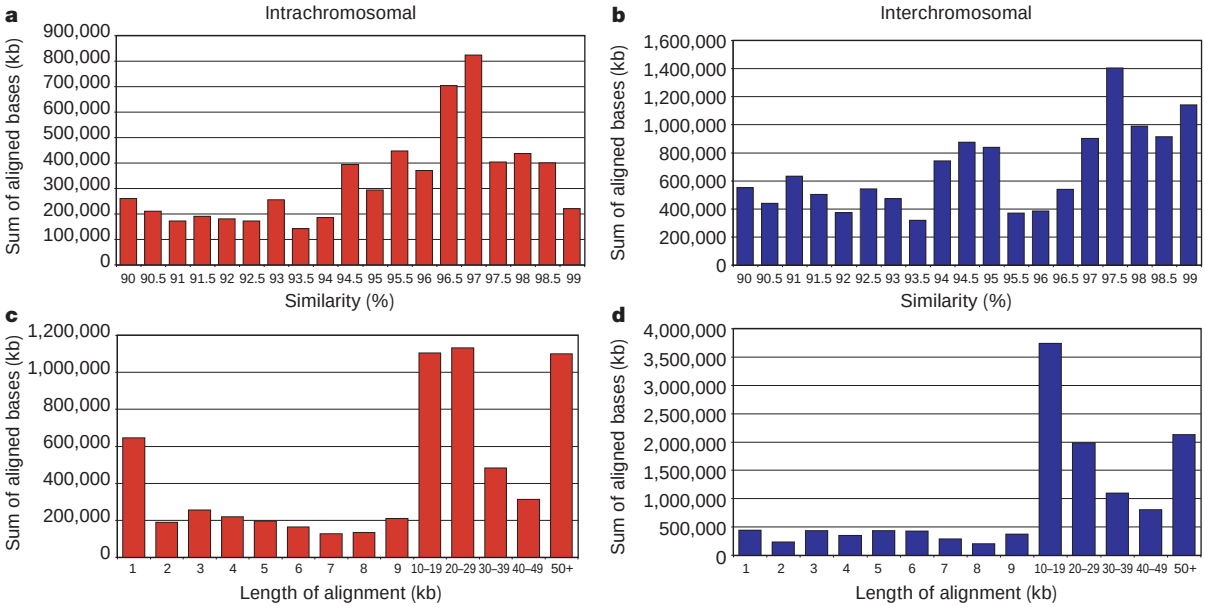


Figure 33 a–d, Sequence properties of segmental duplications. Distributions of length and per cent nucleotide identity for segmental duplications are shown as a function of the number of aligned bp, for the subset of finished genome sequence. Intrachromosomal, red; interchromosomal, blue.

duplications may well be underestimated by the current analysis. An understanding of the biology, pathology and evolution of these duplications will require specialized efforts within these exceptional regions of the human genome. The presence and distribution of such segments may provide evolutionary fodder for processes of exon shuffling and a general increase in protein diversity associated with domain accretion. It will be important to consider both genome-wide duplication events and more restricted punctuated events of genome duplication as forces in the evolution of vertebrate genomes.

Table 17 Fraction of the draft genome sequence in inter- and intrachromosomal duplications

Chromosome	Intrachromosomal (%)	Interchromosomal (%)	All (%)
1	2.1	1.7	3.4
2	1.6	1.6	2.6
3	1.8	1.4	2.7
4	1.5	2.2	3.0
5	1.0	0.9	1.8
6	1.5	1.4	2.7
7	3.6	1.8	4.5
8	1.2	1.5	2.1
9	2.1	2.3	3.8
10	3.3	2.0	4.7
11	2.7	1.4	3.7
12	2.1	1.2	2.8
13	1.7	1.6	3.0
14	0.6	0.6	1.2
15	4.1	4.4	6.7
16	3.4	3.4	5.5
17	4.4	1.7	5.7
18	0.9	1.0	1.9
19	5.4	1.6	6.3
20	0.8	1.4	2.0
21	1.9	4.0	4.8
22	6.8	7.7	11.9
X	1.2	1.1	2.2
Y	10.9	13.1	20.8
NA	2.3	7.8	8.3
UL	11.6	20.8	22.2
Total	2.3	2.0	3.6

Excludes duplications with identities >98% to avoid artefactual duplication due to incomplete merger in the assembly process. Calculation was performed on an earlier version of the draft genome sequence based on data available in July 2000 and reflects the duplications found within the total amount of finished sequence then. Note that there is some overlap between the interchromosomal and intrachromosomal sets.

Gene content of the human genome

Genes (or at least their coding regions) comprise only a tiny fraction of human DNA, but they represent the major biological function of the genome and the main focus of interest by biologists. They are also the most challenging feature to identify in the human genome sequence.

The ultimate goal is to compile a complete list of all human genes and their encoded proteins, to serve as a 'periodic table' for biomedical research[243]. But this is a difficult task. In organisms with small genomes, it is straightforward to identify most genes by the presence of long ORFs. In contrast, human genes tend to have small exons (encoding an average of only 50 codons) separated by long introns (some exceeding 10 kb). This creates a signal-to-noise problem, with the result that computer programs for direct gene prediction have only limited accuracy. Instead, computational prediction of human genes must rely largely on the availability of cDNA sequences or on sequence conservation with genes and proteins from other organisms. This approach is adequate for strongly conserved genes (such as histones or ubiquitin), but may be less sensitive to rapidly evolving genes (including many crucial to speciation, sex determination and fertilization).

Here we describe our efforts to recognize both the RNA genes and protein-coding genes in the human genome. We also study the properties of the predicted human protein set, attempting to discern how the human proteome differs from those of invertebrates such as worm and fly.

Noncoding RNAs

Although biologists often speak of a tight coupling between 'genes

Table 18 Cross-species comparison for large, highly homologous segmental duplications

	Percentage of genome (%)		
	Fly	Worm	Human (finished)*
> 1 kb	1.2	4.25	3.25
> 5 kb	0.37	1.50	2.86
> 10 kb	0.08	0.66	2.52

*This is an underestimate of the total amount of segmental duplication in the human genome because it only reflects duplication detectable with available finished sequence. The proportion of segmental duplications of > 1 kb is probably about 5% (see text).

© 2001 Macmillan Magazines Ltd

and their encoded protein products', it is important to remember that thousands of human genes produce noncoding RNAs (ncRNAs) as their ultimate product[244]. There are several major classes of ncRNA. (1) Transfer RNAs (tRNAs) are the adapters that translate the triplet nucleic acid code of RNA into the amino-acid sequence of proteins; (2) ribosomal RNAs (rRNAs) are also central to the translational machinery, and recent X-ray crystallography results strongly indicate that peptide bond formation is catalysed by rRNA, not protein[245,246]; (3) small nucleolar RNAs (snoRNAs) are required for rRNA processing and base modification in the nucleolus[247,248]; and (4) small nuclear RNAs (snRNAs) are critical components of spliceosomes, the large ribonucleoprotein (RNP) complexes that splice introns out of pre-mRNAs in the nucleus. Humans have both a major, U2 snRNA-dependent spliceosome that splices most introns, and a minor, U12 snRNA-dependent spliceosome that splices a rare class of introns that often have AT/AC dinucleotides at the splice sites instead of the canonical GT/AG splice site consensus[249].

Other ncRNAs include both RNAs of known biochemical function (such as telomerase RNA and the 7SL signal recognition particle RNA) and ncRNAs of enigmatic function (such as the large Xist transcript implicated in X dosage compensation[250], or the small vault RNAs found in the bizarre vault ribonucleoprotein complex[251], which is three times the mass of the ribosome but has unknown function).

ncRNAs do not have translated ORFs, are often small and are not polyadenylated. Accordingly, novel ncRNAs cannot readily be found by computational gene-finding techniques (which search for features such as ORFs) or experimental sequencing of cDNA or EST libraries (most of which are prepared by reverse transcription using a primer complementary to a poly(A) tail). Even if the complete finished sequence of the human genome were available, discovering novel ncRNAs would still be challenging. We can, however, identify genomic sequences that are homologous to known ncRNA genes, using BLASTN or, in some cases, more specialized methods.

It is sometimes difficult to tell whether such homologous genes are orthologues, paralogues or closely related pseudogenes (because inactivating mutations are much less obvious than for protein-coding genes). For tRNA, there is sufficiently detailed information about the cloverleaf secondary structure to allow true genes and pseudogenes to be distinguished with high sensitivity. For many other ncRNAs, there is much less structural information and so we employ an operational criterion of high sequence similarity ($>95\%$ sequence identity and $>95\%$ full length) to distinguish true genes from pseudogenes. These assignments will eventually need to be reconciled with experimental data.

Transfer RNA genes. The classical experimental estimate of the number of human tRNA genes is 1,310 (ref. 252). In the draft genome sequence, we find only 497 human tRNA genes (Tables 19, 20). How do we account for this discrepancy? We believe that the original estimate is likely to have been inflated in two respects. First, it came from a hybridization experiment that probably counted closely related pseudogenes; by analysis of the draft genome sequence, there are in fact 324 tRNA-derived putative pseudogenes (Table 20). Second, the earlier estimate assumed too high a value for the size of the human genome; repeating the calculation using the correct value yields an estimate of about 890 tRNA-related loci, which is in reasonable accord with our count of 821 tRNA genes and pseudogenes in the draft genome sequence.

Table 19 Number of tRNA genes in various organisms

Organism	Number of canonical tRNAs	SeCys tRNA
Human	497	1
Worm	584	1
Fly	284	1
Yeast	273	0
Methanococcus jannaschii	36	1
Escherichia coli	86	1

Number of tRNA genes in each of six genome sequences, according to analysis by the computer program tRNAscan-SE . Canonical tRNAs read one of the standard 61 sense codons; this category excludes pseudogenes, undetermined anticodons, putative supressors and selenocysteine tRNAs. Most organisms have a selenocysteine (SeCys) tRNA species, but some unicellular eukaryotes do not (such as the yeast *S. cerevisiae*).

The human tRNA gene set predicted from the draft genome sequence appears to include most of the known human tRNA species. The draft genome sequence contains 37 of 38 human tRNA species listed in a tRNA database[253], allowing for up to one mismatch. This includes one copy of the known gene for a specialized selenocysteine tRNA, one of several components of a baroque translational mechanism that reads UGA as a selenocysteine codon in certain rare mRNAs that carry a specific *cis*-acting RNA regulatory site (a so-called SECIS element) in their 3′ UTRs. The one tRNA gene in the database not found in the draft genome sequence is DE9990, a tRNAGlu species, which differs in two positions from the most related tRNA gene in the human genome. Possible explanations are that the database version of this tRNA contains two errors, the gene is polymorphic or this is a genuine functional tRNA that is missing from the draft genome sequence. (The database also lists one additional tRNA gene (*DS9994*), but this is apparently a contaminant, most similar to bacterial tRNAs; the parent entry (*Z13399*) was withdrawn from the DNA database, but the tRNA entry has not yet been removed from the tRNA database.) Although the human set appears substantially complete by this test, the tRNA gene numbers in Table 19 should be considered tentative and used with caution. The human and fly (but not the worm) are known to be missing significant amounts of heterochromatic DNA, and additional tRNA genes could be located there.

With this caveat, the results indicate that the human has fewer tRNA genes than the worm, but more than the fly. This may seem surprising, but tRNA gene number in metazoans is thought to be related not to organismal complexity, but more to idiosyncrasies of the demand for tRNA abundance in certain tissues or stages of embryonic development. For example, the frog *Xenopus laevis*, which must load each oocyte with a remarkable 40 ng of tRNA, has thousands of tRNA genes[254].

The degeneracy of the genetic code has allowed an inspired economy of tRNA anticodon usage. Although 61 sense codons need to be decoded, not all 61 different anticodons are present in tRNAs. Rather, tRNAs generally follow stereotyped and conserved wobble rules[255–257]. Wobble reduces the number of required anticodons substantially, and provides a connection between the genetic code and the hybridization stability of modified and unmodified RNA bases. In eukaryotes, the rules proposed by Guthrie and Abelson[256] predict that about 46 tRNA species will be sufficient to read the 61 sense codons (counting the initiator and elongator methionine tRNAs as two species). According to these rules, in the codon's third (wobble) position, U and C are generally decoded by a single tRNA species, whereas A and G are decoded by two separate tRNA species.

In 'two-codon boxes' of the genetic code (where codons ending with U/C encode a different amino acid from those ending with A/G), the U/C wobble position should be decoded by a G at position 34 in the tRNA anticodon. Thus, in the top left of Fig. 34, there is no tRNA with an AAA anticodon for Phe, but the GAA anticodon can recognize both UUU and UUC codons in the mRNA. In 'four-codon boxes' of the genetic code (where U, C, A and G in the wobble position all encode the same amino acid), the U/C wobble position is almost always decoded by I34 (inosine) in the tRNA, where the inosine is produced by post-transcriptional modification of an adenine (A). In the bottom left of Fig. 34, for example, the GUU and GUC codons of the four-codon Val box are decoded by a tRNA with an anticodon of AAC, which is no doubt modified to IAC. Presumably this pattern, which is strikingly conserved in eukaryotes, has to do with the fact that IA base pairs are also possible; thus

© 2001 Macmillan Magazines Ltd

the IAC anticodon for a Val tRNA could recognize GUU, GUC and even GUA codons. Were this same I34 to be utilized in two-codon boxes, however, misreading of the NNA codon would occur, resulting in translational havoc. Eukaryotic glycine tRNAs represent a conserved exception to this last rule; they use a GCC anticodon to decode GGU and GGC, rather than the expected ICC anticodon.

Satisfyingly, the human tRNA set follows these wobble rules almost perfectly (Fig. 34). Only three unexpected tRNA species are found: single genes for a tRNATyr-AUA, tRNAIle-GAU, and tRNAAsn-AUU. Perhaps these are pseudogenes, but they appear to be plausible tRNAs. We also checked the possibility of sequencing errors in their anticodons, but each of these three genes is in a region of high sequence accuracy, with PHRAP quality scores higher than 70 for every base in their anticodons.

As in all other organisms, human protein-coding genes show codon bias—preferential use of one synonymous codon over another[258] (Fig. 34). In less complex organisms, such as yeast or bacteria, highly expressed genes show the strongest codon bias. Cytoplasmic abundance of tRNA species is correlated with both codon bias and overall amino-acid frequency (for example, tRNAs for preferred codons and for more common amino acids are more abundant). This is presumably driven by selective pressure for efficient or accurate translation[259]. In many organisms, tRNA abundance in turn appears to be roughly correlated with tRNA gene copy number, so tRNA gene copy number has been used as a proxy for tRNA abundance[260]. In vertebrates, however, codon bias is not so obviously correlated with gene expression level. Differing codon biases between human genes is more a function of their location in regions of different GC composition[261]. In agreement with the literature, we see only a very rough correlation of human tRNA gene number with either amino-acid frequency or codon bias (Fig. 34). The most obvious outliers in these weak correlations are the strongly preferred CUG leucine codon, with a mere six tRNA-Leu-CAG genes producing a tRNA to decode it, and the relatively rare cysteine UGU and UGC codons, with 30 tRNA genes to decode them.

The tRNA genes are dispersed throughout the human genome. However, this dispersal is nonrandom. tRNA genes have sometimes been seen in clusters at small scales[262,263] but we can now see striking clustering on a genome-wide scale. More than 25% of the tRNA genes (140) are found in a region of only about 4 Mb on chromosome 6. This small region, only about 0.1% of the genome, contains an almost sufficient set of tRNA genes all by itself. The 140 tRNA genes contain a representative for 36 of the 49 anticodons found in the complete set; and of the 21 isoacceptor types, only tRNAs to decode Asn, Cys, Glu and selenocysteine are missing. Many of these tRNA genes, meanwhile, are clustered elsewhere; 18 of the 30 Cys tRNAs are found in a 0.5-Mb stretch of chromosome 7 and many of the Asn and Glu tRNA genes are loosely clustered on chromosome 1. More than half of the tRNA genes (280 out of 497) reside on either chromosome 1 or chromosome 6. Chromosomes 3, 4, 8, 9, 10, 12, 18, 20, 21 and X appear to have fewer than 10 tRNA genes each; and chromosomes 22 and Y have none at all (each has a single pseudogene).

Ribosomal RNA genes. The ribosome, the protein synthetic machine of the cell, is made up of two subunits and contains four rRNA species and many proteins. The large ribosomal subunit contains 28S and 5.8S rRNAs (collectively called 'large subunit' (LSU) rRNA) and also a 5S rRNA. The small ribosomal subunit contains 18S rRNA ('small subunit' (SSU) rRNA). The genes for LSU and SSU rRNA occur in the human genome as a 44-kb tandem repeat unit[264]. There are thought to be about 150–200 copies of this repeat unit arrayed on the short arms of acrocentric chromosomes 13, 14, 15, 21 and 22 (refs 254, 264). There are no true complete

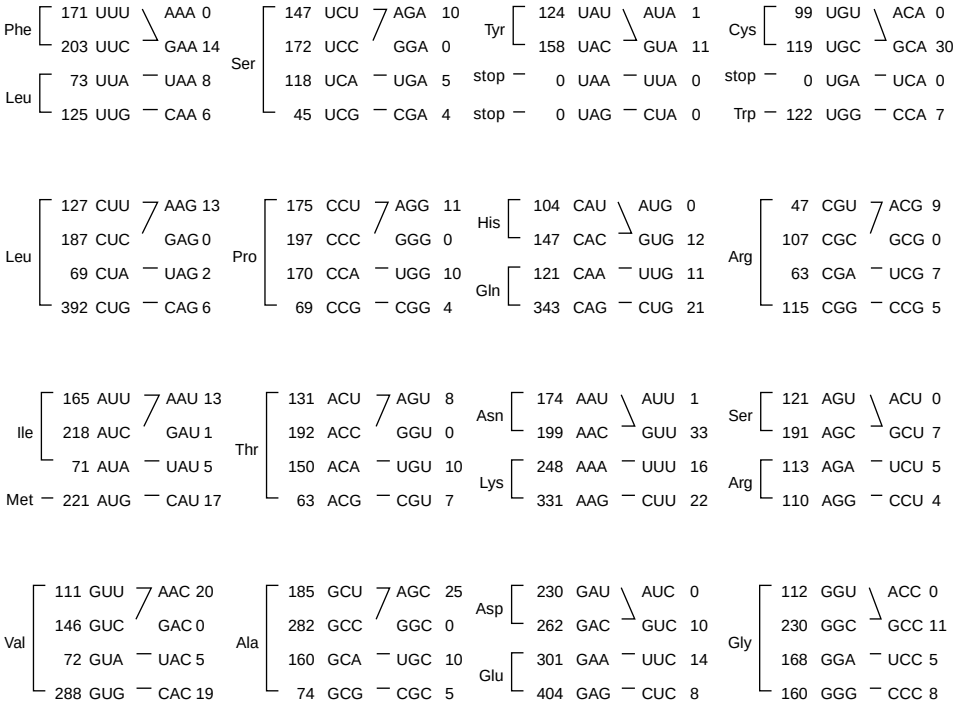


Figure 34 The human genetic code and associated tRNA genes. For each of the 64 codons, we show: the corresponding amino acid; the observed frequency of the codon per 10,000 codons; the codon; predicted wobble pairing to a tRNA anticodon (black lines); an unmodified tRNA anticodon sequence; and the number of tRNA genes found with this anticodon. For example, phenylalanine is encoded by UUU or UUC; UUC is seen more frequently, 203 to 171 occurrences per 10,000 total codons; both codons are expected to be decoded by a single tRNA anticodon type, GAA, using a G/U wobble; and there are 14 tRNA genes found with this anticodon. The modified anticodon sequence in the mature tRNA is not shown, even where post-transcriptional modifications can be confidently predicted (for example, when an A is used to decode a U/C third position, the A is almost certainly an inosine in the mature tRNA). The Figure also does not show the number of distinct tRNA species (such as distinct sequence families) for each anticodon; often there is more than one species for each anticodon.

© 2001 Macmillan Magazines Ltd

copies of the rDNA tandem repeats in the draft genome sequence, owing to the deliberate bias in the initial phase of the sequencing effort against sequencing BAC clones whose restriction fragment fingerprints showed them to contain primarily tandemly repeated sequence. Sequence similarity analysis with the BLASTN computer program does, however, detect hundreds of rDNA-derived sequence fragments dispersed throughout the complete genome, including one 'full-length' copy of an individual 5.8S rRNA gene not associated with a true tandem repeat unit (Table 20).

The 5S rDNA genes also occur in tandem arrays, the largest of which is on chromosome 1 between 1q41.11 and 1q42.13, close to the telomere[265,266]. There are 200–300 true 5S genes in these arrays[265,267]. The number of 5S-related sequences in the genome, including numerous dispersed pseudogenes, is classically cited as 2,000 (refs 252, 254). The long tandem array on chromosome 1 is not yet present in the draft genome sequence because there are no *Eco*RI or *Hin*dIII sites present, and thus it was not cloned in the most heavily utilized BAC libraries (Table 1). We expect to recover it during the finishing stage. We do detect four individual copies of 5S rDNA by our search criteria ($\geq$95% identity and $\geq$95% full length). We also find many more distantly related dispersed sequences (520 at $P \leq 0.001$), which we interpret as probable pseudogenes (Table 20).

Small nucleolar RNA genes. Eukaryotic rRNA is extensively processed and modified in the nucleolus. Much of this activity is directed by numerous snoRNAs. These come in two families: C/D box snoRNAs (mostly involved in guiding site-specific 2′-*O*-ribose methylations of other RNAs) and H/ACA snoRNAs (mostly involved in guiding site-specific pseudouridylations)[247,248]. We compiled a set of 97 known human snoRNA gene sequences; 84 of these (87%) have at least one copy in the draft genome sequence (Table 20), almost all as single-copy genes.

It is thought that all 2′-*O*-ribose methylations and pseudouridylations in eukaryotic rRNA are guided by snoRNAs. There are 105–107 methylations and around 95 pseudouridylations in human rRNA[268]. Only about half of these have been tentatively assigned to known guide snoRNAs. There are also snoRNA-directed modifications on other stable RNAs, such as U6 (ref. 269), and the extent of this is just beginning to be explored. Sequence similarity has so far proven insufficient to recognize all snoRNA genes. We therefore expect that there are many unrecognized snoRNA genes that are not detected by BLAST queries.

Spliceosomal RNAs and other ncRNA genes. We also looked for copies of other known ncRNA genes. We found at least one copy of 21 (95%) of 22 known ncRNAs, including the spliceosomal snRNAs. There were multiple copies for several ncRNAs, as expected; for example, we find 44 dispersed genes for U6 snRNA, and 16 for U1 snRNA (Table 20).

For some of these RNA genes, homogeneous multigene families that occur in tandem arrays are again under-represented owing to the restriction enzymes used in constructing the BAC libraries and, in some instances, the decision to delay the sequencing of BAC clones with low complexity fingerprints indicative of tandemly repeated DNA. The U2 RNA genes are located at the RNU2 locus, a tandem array of 10–20 copies of nearly identical 6.1-kb units at 17q21–q22 (refs 270–272). Similarly, the U3 snoRNA genes (included in the aggregate count of C/D snoRNAs in Table 20) are clustered at the RNU3 locus at 17p11.2, not in a tandem array, but in a complex inverted repeat structure of about 5–10 copies per haploid genome[273]. The U1 RNA genes are clustered with about 30 copies at the RNU1 locus at 1p36.1, but this cluster is thought to be loose and irregularly organized; no two U1 genes have been cloned on the same cosmid[271]. In the draft genome sequence, we see six copies of U2 RNA that meet our criteria for true genes, three of which appear to be in the expected position on chromosome 17. For U3, so far we see one true copy at the correct place on chromosome 17p11.2. For U1, we see 16 true genes, 6 of which are loosely clustered within 0.6 Mb at 1p36.1 and another 6 are elsewhere on chromosome 1. Again, these and other clusters will be a matter for the finishing process.

Table 20 Known non-coding RNA genes in the draft genome sequence

RNA gene*	Number expected†	Number found‡	Number of related genes§	Function
tRNA	1,310	497	324	Protein synthesis
SSU (18S) rRNA	150–200	0	40	Protein synthesis
5.8S rRNA	150–200	1	11	Protein synthesis
LSU (28S) rRNA	150–200	0	181	Protein synthesis
5S rRNA	200–300	4	520	Protein synthesis
U1	~30	16	134	Spliceosome component
U2	10–20	6	94	Spliceosome component
U4	??	4	87	Spliceosome component
U4atac	??	1	20	Component of minor (U11/U12) spliceosome
U5	??	1	31	Spliceosome component
U6	??	44	1,135	Spliceosome component
U6atac	??	4	32	Component of minor (U11/U12) spliceosome
U7	1	1	3	Histone mRNA 3′ processing
U11	1	0	6	Component of minor (U11/U12) spliceosome
U12	1	1	0	Component of minor (U11/U12) spliceosome
SRP (7SL) RNA	4	3	773	Component of signal recognition particle (protein secretion)
RNAse P	1	1	2	tRNA 5′ end processing
RNAse MRP	1	1	6	rRNA processing
Telomerase RNA	1	1	4	Template for addition of telomeres
hY1	1	1	353	Component of Ro RNP, function unknown
hY3	1	25	414	Component of Ro RNP, function unknown
hY4	1	3	115	Component of Ro RNP, function unknown
hY5 (4.5S RNA)	1	1	9	Component of Ro RNP, function unknown
Vault RNAs	3	3	1	Component of 13-MDa vault RNP, function unknown
7SK	1	1	330	Unknown
H19	1	1	2	Unknown
Xist	1	1	0	Initiation of X chromosome inactivation (dosage compensation)
Known C/D snoRNAs	81	69	558	Pre-rRNA processing or site-specific ribose methylation of rRNA
Known H/ACA snoRNAs	16	15	87	Pre-rRNA processing or site-specific pseudouridylation of rRNA

* Known ncRNA genes (or gene families, such as the C/D and H/ACA snoRNA families); reference sequences were extracted from GenBank and used to probe the draft genome sequence.
† Number of genes that were expected in the human genome, based on previous literature (note that earlier experimental techniques probably tend to overestimate copy number, by counting closely related pseudogenes).
‡ The copy number of 'true' full-length genes identified in the draft genome sequence.
§ The copy number of other significantly related copies (pseudogenes, fragments, paralogues) found. Except for the 497 true tRNA genes, all sequence similarities were identified by WashU BLASTN 2.0MP (W. Gish, unpublished; http://blast.wustl.edu), with parameters '-kap wordmask = seg B = 50000 W = 8' and the default +5/–4 DNA scoring matrix. True genes were operationally defined as BLAST hits with $\geq$95% identity over $\geq$95% of the length of the query. Related sequences were operationally defined as all other BLAST hits with *P*-values $\leq$ 0.001.

© 2001 Macmillan Magazines Ltd

Table 21 Characteristics of human genes

	Median	Mean	Sample (size)
Internal exon	122 bp	145 bp	RefSeq alignments to draft genome sequence, with confirmed intron boundaries (43,317 exons)
Exon number	7	8.8	RefSeq alignments to finished sequence (3,501 genes)
Introns	1,023 bp	3,365 bp	RefSeq alignments to finished sequence (27,238 introns)
3′ UTR	400 bp	770 bp	Confirmed by mRNA or EST on chromosome 22 (689)
5′ UTR	240 bp	300 bp	Confirmed by mRNA or EST on chromosome 22 (463)
Coding sequence	1,100 bp	1,340 bp	Selected RefSeq entries (1,804)
(CDS)	367 aa	447 aa	
Genomic extent	14 kb	27 kb	Selected RefSeq entries (1,804)

Median and mean values for a number of properties of human protein-coding genes. The 1,804 selected RefSeq entries were those that could be unambiguously aligned to finished sequence over their entire length.

Our observations also confirm the striking proliferation of ncRNA-derived pseudogenes (Table 20). There are hundreds or thousands of sequences in the draft genome sequence related to some of the ncRNA genes. The most prolific pseudogene counts generally come from RNA genes transcribed by RNA polymerase III promoters, including U6, the hY RNAs and SRP-RNA. These ncRNA pseudogenes presumably arise through reverse transcription. The frequency of such events gives insight into how ncRNA genes can evolve into SINE retroposons, such as the tRNA-derived SINEs found in many vertebrates and the SRP-RNA-derived Alu elements found in humans.

Protein-coding genes

Identifying the protein-coding genes in the human genome is one of the most important applications of the sequence data, but also one of the most difficult challenges. We describe below our efforts to create an initial human gene and protein index.

Exploring properties of known genes. Before attempting to identify new genes, we explored what could be learned by aligning the cDNA sequences of known genes to the draft genome sequence. Genomic alignments allow one to study exon–intron structure and local GC content, and are valuable for biomedical studies because they connect genes with the genetic and cytogenetic map, link them with regulatory sequences and facilitate the development of polymerase chain reaction (PCR) primers to amplify exons. Until now, genomic alignment was available for only about a quarter of known genes.

The 'known' genes studied were those in the RefSeq database[110], a manually curated collection designed to contain nonredundant representatives of most full-length human mRNA sequences in GenBank (RefSeq intentionally contains some alternative splice forms of the same genes). The version of RefSeq used contained 10,272 mRNAs.

The RefSeq genes were aligned with the draft genome sequence, using both the Spidey (S. Wheelan, personal communication) and Acembly (D. Thierry-Mieg and J. Thierry-Mieg, unpublished; http://www.acedb.org) computer programs. Because this sequence is incomplete and contains errors, not all genes could be fully aligned and some may have been incorrectly aligned. More than 92% of the RefSeq entries could be aligned at high stringency over at least part of their length, and 85% could be aligned over more than half of their length. Some genes (16%) had high stringency alignments to more than one location in the draft genome sequence owing, for example, to paralogues or pseudogenes. In such cases, we considered only the best match. In a few of these cases, the assignment may not be correct because the true matching region has not yet been sequenced. Three per cent of entries appeared to be alternative splice products of the same gene, on the basis of their alignment to the same location in the draft genome sequence. In all, we obtained at least partial genomic alignments for 9,212 distinct known genes and essentially complete alignment for 5,364 of them.

Previous efforts to study human gene structure[116,274,275] have been hampered by limited sample sizes and strong biases in favour of compact genes. Table 21 gives the mean and median values of some basic characteristics of gene structures. Some of the values may be underestimates. In particular, the UTRs given in the RefSeq database are likely to be incomplete; they are considerably shorter, for example, than those derived from careful reconstructions on chromosome 22. Intron sizes were measured only for genes in finished genomic sequence, to mitigate the bias arising from the fact that

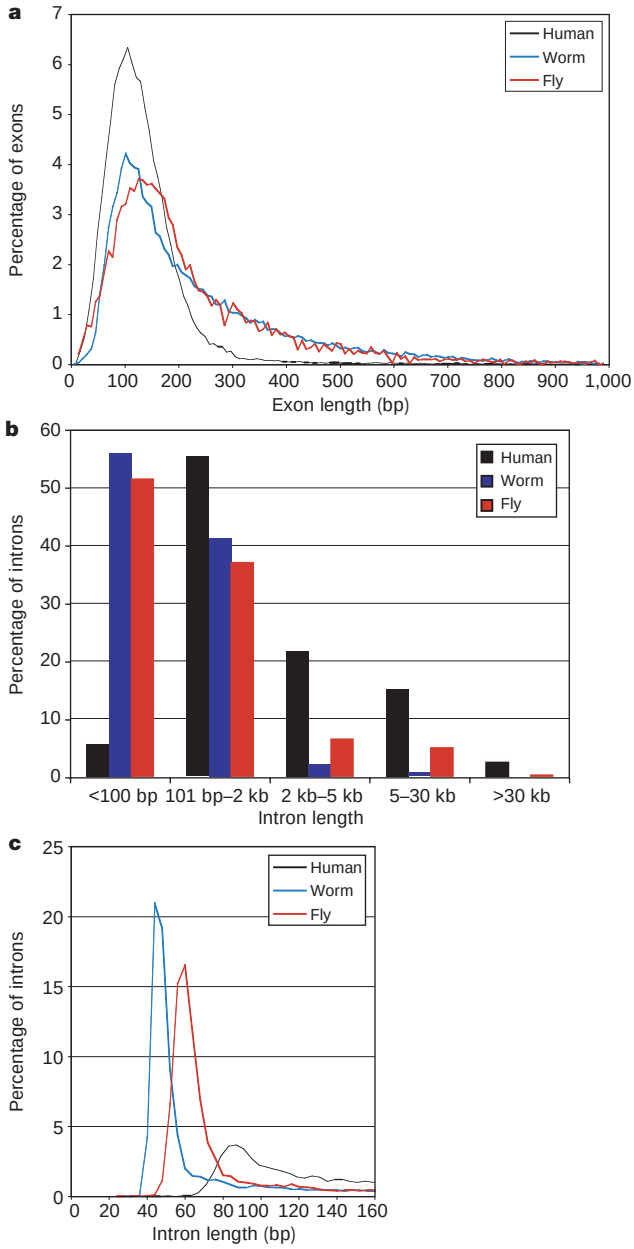


Figure 35 Size distributions of exons, introns and short introns, in sequenced genomes. **a**, Exons; **b**, introns; **c**, short introns (enlarged from **b**). Confirmed exons and introns for the human were taken from RefSeq alignments and for worm and fly from Acembly alignments of ESTs (J. and D. Thierry-Mieg and, for worm, Y. Kohara, unpublished).

© 2001 Macmillan Magazines Ltd

long introns are more likely than short introns to be interrupted by gaps in the draft genome sequence. Nonetheless, there may be some residual bias against long genes and long introns.

There is considerable variation in overall gene size and intron size, with both distributions having very long tails. Many genes are over 100 kb long, the largest known example being the dystrophin gene (DMD) at 2.4 Mb. The variation in the size distribution of coding sequences and exons is less extreme, although there are still some remarkable outliers. The titin gene[276] has the longest currently known coding sequence at 80,780 bp; it also has the largest number of exons (178) and longest single exon (17,106 bp).

It is instructive to compare the properties of human genes with those from worm and fly. For all three organisms, the typical length of a coding sequence is similar (1,311 bp for worm, 1,497 bp for fly and 1,340 bp for human), and most internal exons fall within a common peak between 50 and 200 bp (Fig. 35a). However, the worm and fly exon distributions have a fatter tail, resulting in a larger mean size for internal exons (218 bp for worm versus 145 bp for human). The conservation of preferred exon size across all three species supports suggestions of a conserved exon-based component of the splicing machinery[277]. Intriguingly, the few extremely short human exons show an unusual base composition. In 42 detected human exons of less than 19 bp, the nucleotide frequencies of A, G, T and C are 39, 33, 15 and 12%, respectively, showing a strong purine bias. Purine-rich sequences may enhance splicing[278,279], and it is possible that such sequences are required or strongly selected for to ensure correct splicing of very short exons. Previous studies have shown that short exons require intronic, but not exonic, splicing enhancers[280].

In contrast to the exons, the intron size distributions differ substantially among the three species (Fig. 35b, c). The worm and fly each have a reasonably tight distribution, with most introns near the preferred minimum intron length (47 bp for worm, 59 bp for fly) and an extended tail (overall average length of 267 bp for worm and 487 bp for fly). Intron size is much more variable in humans, with a peak at 87 bp but a very long tail resulting in a mean of more than 3,300 bp. The variation in intron size results in great variation in gene size.

The variation in gene size and intron size can partly be explained by the fact that GC-rich regions tend to be gene-dense with many compact genes, whereas AT-rich regions tend to be gene-poor with many sprawling genes containing large introns. The correlation of gene density with GC content is shown in Fig. 36a, b; the relative density increases more than tenfold as GC content increases from

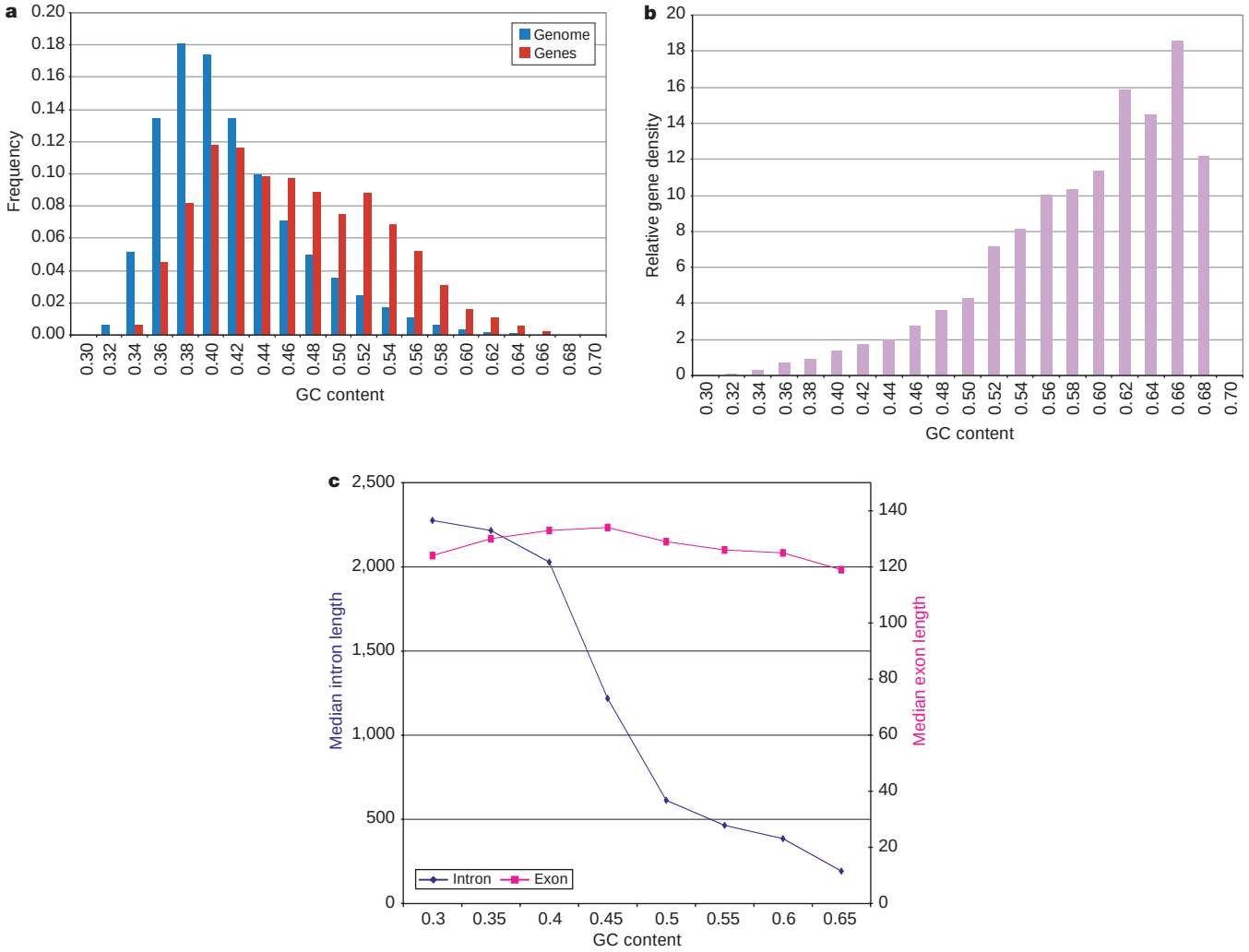


Figure 36 GC content. **a**, Distribution of GC content in genes and in the genome. For 9,315 known genes mapped to the draft genome sequence, the local GC content was calculated in a window covering either the whole alignment or 20,000 bp centred around the midpoint of the alignment, whichever was larger. Ns in the sequence were not counted. GC content for the genome was calculated for adjacent nonoverlapping 20,000-bp windows across the sequence. Both the gene and genome distributions have been normalized to sum to one. **b**, Gene density as a function of GC content, obtained by taking the ratio of the data in **a**. Values are less accurate at higher GC levels because the denominator is small. **c**, Dependence of mean exon and intron lengths on GC content. For exons and introns, the local GC content was derived from alignments to finished sequence only, and were calculated from windows covering the feature or 10,000 bp centred on the feature, whichever was larger.

© 2001 Macmillan Magazines Ltd

30% to 50%. The correlation appears to be due primarily to intron size, which drops markedly with increasing GC content (Fig. 36c). In contrast, coding properties such as exon length (Fig. 36c) or exon number (data not shown) vary little. Intergenic distance is also probably lower in high-GC areas, although this is hard to prove directly until all genes have been identified.

The large number of confirmed human introns allows us to analyse variant splice sites, confirming and extending recent reports[281]. Intron positions were confirmed by applying a stringent criterion that EST or mRNA sequence show an exact match of 8 bp in the flanking exonic sequence on each side. Of 53,295 confirmed introns, 98.12% use the canonical dinucleotides GT at the 5′ splice site and AG at the 3′ site (GT–AG pattern). Another 0.76% use the related GC–AG. About 0.10% use AT–AC, which is a rare alternative pattern primarily recognized by the variant U12 splicing machinery[282]. The remaining 1% belong to 177 types, some of which undoubtedly reflect sequencing or alignment errors.

Finally, we looked at alternative splicing of human genes. Alternative splicing can allow many proteins to be produced from a single gene and can be used for complex gene regulation. It appears to be prevalent in humans, with lower estimates of about 35% of human genes being subject to alternative splicing[283–285]. These studies may have underestimated the prevalence of alternative splicing, because they examined only EST alignments covering only a portion of a gene.

To investigate the prevalence of alternative splicing, we analysed reconstructed mRNA transcripts covering the entire coding regions of genes on chromosome 22 (omitting small genes with coding regions of less than 240 bp). Potential transcripts identified by alignments of ESTs and cDNAs to genomic sequence were verified by human inspection. We found 642 transcripts, covering 245 genes (average of 2.6 distinct transcripts per gene). Two or more alternatively spliced transcripts were found for 145 (59%) of these genes. A similar analysis for the gene-rich chromosome 19 gave 1,859 transcripts, corresponding to 544 genes (average 3.2 distinct transcripts per gene). Because we are sampling only a subset of all transcripts, the true extent of alternative splicing is likely to be greater. These figures are considerably higher than those for worm, in which analysis reveals alternative splicing for 22% of genes for which ESTs have been found, with an average of 1.34 (12,816/9,516) splice variants per gene. (The apparently higher extent of alternative splicing seen in human than in worm was not an artefact resulting from much deeper coverage of human genes by ESTs and mRNAs. Although there are many times more ESTs available for human than worm, these ESTs tend to have shorter average length (because many were the product of early sequencing efforts) and many match no human genes. We calculated the actual coverage per bp used in the analysis of the human and worm genes; the coverage is only modestly higher (about 50%) for the human, with a strong bias towards 3′ UTRs which tend to show much less alternative splicing. We also repeated the analysis using equal coverage for the two organisms and confirmed that higher levels of alternative splicing were still seen in human.)

Seventy per cent of alternative splice forms found in the genes on chromosomes 19 and 22 affect the coding sequence, rather than merely changing the 3′ or 5′ UTR. (This estimate may be affected by the incomplete representation of UTRs in the RefSeq database and in the transcripts studied.) Alternative splicing of the terminal exon was seen for 20% of 6,105 mRNAs that were aligned to the draft genome sequence and correspond to confirmed 3′ EST clusters. In addition to alternative splicing, we found evidence of the terminal exon employing alternative polyadenylation sites (separated by >100 bp) in 24% of cases.

Towards a complete index of human genes. We next focused on creating an initial index of human genes and proteins. This index is quite incomplete, owing to the difficulty of gene identification in human DNA and the imperfect state of the draft genome sequence. Nonetheless, it is valuable for experimental studies and provides important insights into the nature of human genes and proteins.

The challenge of identifying genes from genomic sequence varies greatly among organisms. Gene identification is almost trivial in bacteria and yeast, because the absence of introns in bacteria and their paucity in yeast means that most genes can be readily recognized by *ab initio* analysis as unusually long ORFs. It is not as simple, but still relatively straightforward, to identify genes in animals with small genomes and small introns, such as worm and fly. A major factor is the high signal-to-noise ratio—coding sequences comprise a large proportion of the genome and a large proportion of each gene (about 50% for worm and fly), and exons are relatively large.

Gene identification is more difficult in human DNA. The signal-to-noise ratio is lower: coding sequences comprise only a few per cent of the genome and an average of about 5% of each gene; internal exons are smaller than in worms; and genes appear to have more alternative splicing. The challenge is underscored by the work on human chromosomes 21 and 22. Even with the availability of finished sequence and intensive experimental work, the gene content remains uncertain, with upper and lower estimates differing by as much as 30%. The initial report of the finished sequence of chromosome 22 (ref. 94) identified 247 previously known genes, 298 predicted genes confirmed by sequence homology or ESTs and 325 *ab initio* predictions without additional support. Many of the confirmed predictions represented partial genes. In the past year, 440 additional exons (10%) have been added to existing gene annotations by the chromosome 22 annotation group, although the number of confirmed genes has increased by only 17 and some previously identified gene predictions have been merged[286].

Before discussing the gene predictions for the human genome, it is useful to consider background issues, including previous estimates of the number of human genes, lessons learned from worms and flies and the representativeness of currently 'known' human genes.

Previous estimates of human gene number. Although direct enumeration of human genes is only now becoming possible with the advent of the draft genome sequence, there have been many attempts in the past quarter of a century to estimate the number of genes indirectly. Early estimates based on reassociation kinetics estimated the mRNA complexity of typical vertebrate tissues to be 10,000–20,000, and were extrapolated to suggest around 40,000 for the entire genome[287]. In the mid-1980s, Gilbert suggested that there might be about 100,000 genes, based on the approximate ratio of the size of a typical gene ($\sim 3\times10^4$ bp) to the size of the genome (3×10^9 bp). Although this was intended only as a back-of-the-envelope estimate, the pleasing roundness of the figure seems to have led to it being widely quoted and adopted in many textbooks. (W. Gilbert, personal communication; ref. 288). An estimate of 70,000–80,000 genes was made by extrapolating from the number of CpG islands and the frequency of their association with known genes[129].

As human sequence information has accumulated, it has been possible to derive estimates on the basis of sampling techniques[289]. Such studies have sought to extrapolate from various types of data, including ESTs, mRNAs from known genes, cross-species genome comparisons and analysis of finished chromosomes. Estimates based on ESTs[290] have varied widely, from 35,000 (ref. 130) to 120,000 genes[291]. Some of the discrepancy lies in differing estimates of the amount of contaminating genomic sequence in the EST collection and the extent to which multiple distinct ESTs correspond to a single gene. The most rigorous analyses[130] exclude as spurious any ESTs that appear only once in the data set and carefully calibrate sensitivity and specificity. Such calculations consistently produce low estimates, in the region of 35,000.

Comparison of whole-genome shotgun sequence from the pufferfish *T. nigroviridis* with the human genome[292] can be used to estimate the density of exons (detected as conserved sequences

© 2001 Macmillan Magazines Ltd

between fish and human). These analyses also suggest around 30,000 human genes.

Extrapolations have also been made from the gene counts for chromosomes 21 and 22 (refs 93, 94), adjusted for differences in gene densities on these chromosomes, as inferred from EST mapping. These estimates are between 30,500 and 35,500, depending on the precise assumptions used[286].

Insights from invertebrates. The worm and fly genomes contain a large proportion of novel genes (around 50% of worm genes and 30% of fly genes), in the sense of showing no significant similarity to organisms outside their phylum[293–295]. Such genes may have been present in the original eukaryotic ancestor, but were subsequently lost from the lineages of the other eukaryotes for which sequence is available; they may be rapidly diverging genes, so that it is difficult to recognize homologues solely on the basis of sequence; they may represent true innovations developed within the lineage; or they may represent acquisitions by horizontal transfer. Whatever their origin, these genes tend to have different biological properties from highly conserved genes. In particular, they tend to have low expression levels as assayed both by direct studies and by a paucity of corresponding ESTs, and are less likely to produce a visible phenotype in loss-of-function genetic experiments[294,296].

Gene prediction. Current gene prediction methods employ combinations of three basic approaches: direct evidence of transcription provided by ESTs or mRNAs[297–299]; indirect evidence based on sequence similarity to previously identified genes and proteins[300,301]; and *ab initio* recognition of groups of exons on the basis of hidden Markov models (HMMs) that combine statistical information about splice sites, coding bias and exon and intron lengths (for example, Genscan[275], Genie[302,303] and FGENES[304]).

The first approach relies on direct experimental data, but is subject to artefacts arising from contaminating ESTs derived from unspliced mRNAs, genomic DNA contamination and nongenic transcription (for example, from the promoter of a transposable element). The first two problems can be mitigated by comparing transcripts with the genomic sequence and using only those that show clear evidence of splicing. This solution, however, tends to discard evidence from genes with long terminal exons or single exons. The second approach tends correctly to identify gene-derived sequences, although some of these may be pseudogenes. However, it obviously cannot identify truly novel genes that have no sequence similarity to known genes. The third approach would suffice alone if one could accurately define the features used by cells for gene recognition, but our current understanding is insufficient to do so. The sensitivity and specificity of *ab initio* predictions are greatly affected by the signal-to-noise ratio. Such methods are more accurate in the fly and worm than in human. In fly, *ab initio* methods can correctly predict around 90% of individual exons and can correctly predict all coding exons of a gene in about 40% of cases[303]. For human, the comparable figures are only about 70% and 20%, respectively[94,305]. These estimates may be optimistic, owing to the design of the tests used.

In any collection of gene predictions, we can expect to see various errors. Some gene predictions may represent partial genes, because of inability to detect some portions of a gene (incomplete sensitivity) or to connect all the components of a gene (fragmentation); some may be gene fusions; and others may be spurious predictions (incomplete specificity) resulting from chance matches or pseudogenes.

Creating an initial gene index. We set out to create an initial integrated gene index (IGI) and an associated integrated protein index (IPI) for the human genome. We describe the results obtained from a version of the draft genome sequence based on the sequence data available in July 2000, to allow time for detailed analysis of the gene and protein content. The additional sequence data that has since become available will affect the results quantitatively, but are unlikely to change the conclusions qualitatively.

We began with predictions produced by the Ensembl system[306]. Ensembl starts with *ab initio* predictions produced by Genscan[275] and then attempts to confirm them by virtue of similarity to proteins, mRNAs, ESTs and protein motifs (contained in the Pfam database[307]) from any organism. In particular, it confirms introns if they are bridged by matches and exons if they are flanked by confirmed introns. It then attempts to extend protein matches using the GeneWise computer program[308]. Because it requires confirmatory evidence to support each gene component, it frequently produces partial gene predictions. In addition, when there is evidence of alternative splicing, it reports multiple overlapping transcripts. In total, Ensembl produced 35,500 gene predictions with 44,860 transcripts.

To reduce fragmentation, we next merged Ensembl-based gene predictions with overlapping gene predictions from another program, Genie[302]. Genie starts with mRNA or EST matches and employs an HMM to extend these matches by using *ab initio* statistical approaches. To avoid fragmentation, it attempts to link information from 5′ and 3′ ESTs from the same cDNA clone and thereby to produce a complete coding sequence from an initial ATG to a stop codon. As a result, it may generate complete genes more accurately than Ensembl in cases where there is extensive EST support. (Genie also generates potential alternative transcripts, but we used only the longest transcript in each group.) We merged 15,437 Ensembl predictions into 9,526 clusters, and the longest transcript in each cluster (from either Genie or Ensembl) was taken as the representative.

Next, we merged these results with known genes contained in the RefSeq (version of 29 September 2000), SWISSPROT (release 39.6 of 30 August 2000) and TrEMBL databases (TrEMBL release 14.17 of 1 October 2000, TrEMBL_new of 1 October 2000). Incorporating these sequences gave rise to overlapping sequences because of alternative splice forms and partial sequences. To construct a nonredundant set, we selected the longest sequence from each overlapping set by using direct protein comparison and by mapping the gene predictions back onto the genome to construct the overlapping sets. This may occasionally remove some close paralogues in the event that the correct genomic location has not yet been sequenced, but this number is expected to be small.

Finally, we searched the set to eliminate any genes derived from contaminating bacterial sequences, recognized by virtue of near identity to known bacterial plasmids, transposons and chromosomal genes. Although most instances of such contamination had been removed in the assembly process, a few cases had slipped through and were removed at this stage.

The process resulted in version 1 of the IGI (IGI.1). The composition of the corresponding IPI.1 protein set, obtained by translating IGI.1, is given in Table 22. There are 31,778 protein predictions, with 14,882 from known genes, 4,057 predictions from Ensembl merged with Genie and 12,839 predictions from Ensembl alone. The average lengths are 469 amino acids for the known proteins, 443 amino acids for protein predictions from the Ensembl–Genie merge, and 187 amino acids for those from Ensembl alone. (The smaller average size for the predictions from Ensembl alone reflects its tendency to predict partial genes where there is supporting evidence for only part of the gene; the remainder of the gene will often not be predicted at all, rather than included as part of another prediction. Accordingly, the smaller size cannot be used to estimate the rate of fragmentation in such predictions.)

The set corresponds to fewer than 31,000 actual genes, because some genes are fragmented into more than one partial prediction and some predictions may be spurious or correspond to pseudogenes. As discussed below, our best estimate is that IGI.1 includes about 24,500 true genes.

Evaluation of IGI/IPI. We used several approaches to evaluate the sensitivity, specificity and fragmentation of the IGI/IPI set.

Comparison with 'new' known genes. One approach was to examine

© 2001 Macmillan Magazines Ltd

Table 22 Properties of the IGI/IPI human protein set

Source	Number	Average length (amino acids)	Matches to nonhuman proteins	Matches to RIKEN mouse cDNA set	Matches to RIKEN mouse cDNA set but not to nonhuman proteins
RefSeq/SwissProt/TrEMBL	14,882	469	12,708 (85%)	11,599 (78%)	776 (36%)
Ensembl–Genie	4,057	443	2,989 (74%)	3,016 (74%)	498 (47%)
Ensembl	12,839	187	81,126 (63%)	7,372 (57%)	1,449 (31%)
Total	31,778	352	23,813 (75%)	219,873 (69%)	2,723 (34%)

The matches to nonhuman proteins were obtained by using Smith-Waterman sequence alignment with an *E*-value threshold of 10^{-3} and the matches to the RIKEN mouse cDNAs by using TBLASTN with an *E*-value threshold of 10^{-6}. The last column shows that a significant number of the IGI members that do not have nonhuman protein matches do match sequences in the RIKEN mouse cDNA set, suggesting that both the IGI and the RIKEN sets contain a significant number of novel proteins.

newly discovered genes arising from independent work that were not used in our gene prediction effort. We identified 31 such genes: 22 recent entries to RefSeq and 9 from the Sanger Centre's gene identification program on chromosome X. Of these, 28 were contained in the draft genome sequence and 19 were represented in the IGI/IPI. This suggests that the gene prediction process has a sensitivity of about 68% (19/28) for the detection of novel genes in the draft genome sequence and that the current IGI contains about 61% (19/31) of novel genes in the human genome. On average, 79% of each gene was detected. The extent of fragmentation could also be estimated: 14 of the genes corresponded to a single prediction in the IGI/IPI, three genes corresponded to two predictions, one gene to three predictions and one gene to four predictions. This corresponds to a fragmentation rate of about 1.4 gene predictions per true gene.

Comparison with RIKEN mouse cDNAs. In a less direct but larger-scale approach, we compared the IGI gene set to a set of mouse cDNAs sequenced by the Genome Exploration Group of the RIKEN Genomic Sciences Center[309]. This set of 15,294 cDNAs, subjected to full-insert sequencing, was enriched for novel genes by selecting cDNAs with novel 3′ ends from a collection of nearly one million ESTs from diverse tissues and developmental timepoints. We determined the proportion of the RIKEN cDNAs that showed sequence similarity to the draft genome sequence and the proportion that showed sequence similarity to the IGI/IPI. Around 81% of the genes in the RIKEN mouse set showed sequence similarity to the human genome sequence, whereas 69% showed sequence similarity to the IGI/IPI. This suggests a sensitivity of 85% (69/81). This is higher than the sensitivity estimate above, perhaps because some of the matches may be due to paralogues rather than orthologues. It is consistent with the IGI/IPI representing a substantial fraction of the human proteome.

Conversely, 69% (22,013/31,898) of the IGI matches the RIKEN cDNA set. Table 22 shows the breakdown of these matches among the different components of the IGI. This is lower than the proportion of matches among known proteins, although this is expected because known proteins tend to be more highly conserved (see above) and because the predictions are on average shorter than known proteins. Table 22 also shows the numbers of matches to the RIKEN cDNAs among IGI members that do not match known proteins. The results indicate that both the IGI and the RIKEN set contain a significant number of genes that are novel in the sense of not having known protein homologues.

Comparison with genes on chromosome 22. We also compared the IGI/IPI with the gene annotations on chromosome 22, to assess the proportion of gene predictions corresponding to pseudogenes and to estimate the rate of overprediction. We compared 477 IGI gene predictions to 539 confirmed genes and 133 pseudogenes on chromosome 22 (with the immunoglobulin lambda locus excluded owing to its highly atypical gene structure). Of these, 43 hit 36 annotated pseudogenes. This suggests that 9% of the IGI predictions may correspond to pseudogenes and also suggests a fragmentation rate of 1.2 gene predictions per gene. Of the remaining hits, 63 did not overlap with any current annotations. This would suggest a rate of spurious predictions of about 13% (63/477), although the true rate is likely to be much lower because many of these may correspond to unannotated portions of existing gene predictions or to currently unannotated genes (of which there are estimated to be about 100 on this chromosome[94]).

Chromosomal distribution. Finally, we examined the chromosomal distribution of the IGI gene set. The average density of gene predictions is 11.1 per Mb across the genome, with the extremes being chromosome 19 at 26.8 per Mb and chromosome Y at 6.4 per Mb. It is likely that a significant number of the predictions on chromosome Y are pseudogenes (this chromosome is known to be rich in pseudogenes) and thus that the density for chromosome Y is an overestimate. The density of both genes and Alus on chromosome 19 is much higher than expected, even accounting for the high GC content of the chromosome; this supports the idea that Alu density is more closely correlated with gene density than with GC content itself.

Summary. We are clearly still some way from having a complete set of human genes. The current IGI contains significant numbers of partial genes, fragmented and fused genes, pseudogenes and spurious predictions, and it also lacks significant numbers of true genes. This reflects the current state of gene prediction methods in vertebrates even in finished sequence, as well as the additional challenges related to the current state of the draft genome sequence. Nonetheless, the gene predictions provide a valuable starting point for a wide range of biological studies and will be rapidly refined in the coming year.

The analysis above allows us to estimate the number of distinct genes in the IGI, as well as the number of genes in the human genome. The IGI set contains about 15,000 known genes and about 17,000 gene predictions. Assuming that the gene predictions are subject to a rate of overprediction (spurious predictions and pseudogenes) of 20% and a rate of fragmentation of 1.4, the IGI would be estimated to contain about 24,500 actual human genes. Assuming that the gene predictions contain about 60% of previously unknown human genes, the total number of genes in the human genome would be estimated to be about 31,000. This is consistent with most recent estimates based on sampling, which suggest a gene number of 30,000–35,000. If there are 30,000–35,000 genes, with an average coding length of about 1,400 bp and average genomic extent of about 30 kb, then about 1.5% of the human genome would consist of coding sequence and one-third of the genome would be transcribed in genes.

The IGI/IPI was constructed primarily on the basis of gene predictions from Ensembl. However, we also generated an expanded set (IGI+) by including additional predictions from two other gene prediction programs, Genie and GenomeScan (C. Burge, personal communication). These predictions were not included in the core IGI set, because of the concern that each additional set will provide diminishing returns in identifying true genes while contributing its own false positives (increased sensitivity at the expense of specificity). Genie produced an additional 2,837 gene predictions not overlapping the IGI, and GenomeScan produced 6,534 such gene predictions. If all of these gene predictions were included in the IGI, the number of the 31 new 'known' genes (see above) contained in the IGI would rise from 19 to 24. This would amount to an increase of about 26% in sensitivity, at the expense of increasing the number of predicted genes (excluding knowns) by 55%. Allowing a higher

© 2001 Macmillan Magazines Ltd

overprediction rate of 30% for gene predictions in this expanded set, the analysis above suggests that IGI+ set contains about 28,000 true genes and yields an estimate of about 32,000 human genes. We are investigating ways to filter the expanded set, to produce an IGI with the advantage of the increased sensitivity resulting from combining multiple gene prediction programs without the corresponding loss of specificity. Meanwhile, the IGI+ set can be used by researchers searching for genes that cannot be found in the IGI.

Some classes of genes may have been missed by all of the gene-finding methods. Genes could be missed if they are expressed at low levels or in rare tissues (being absent or very under-represented in EST and mRNA databases) and have sequences that evolve rapidly (being hard to detect by protein homology and genome comparison). Both the worm and fly gene sets contain a substantial number of such genes[293,294]. Single-exon genes encoding small proteins may also have been missed, because EST evidence that supports them cannot be distinguished from genomic contamination in the EST dataset and because homology may be hard to detect for small proteins[310].

The human thus appears to have only about twice as many genes as worm or fly. However, human genes differ in important respects from those in worm and fly. They are spread out over much larger regions of genomic DNA, and they are used to construct more alternative transcripts. This may result in perhaps five times as many primary protein products in the human as in the worm or fly.

The predicted gene and protein sets described here are clearly far from final. Nonetheless, they provide a valuable starting point for experimental and computational research. The predictions will improve progressively as the sequence is finished, as further confirmatory evidence becomes available (particularly from other vertebrate genome sequences, such as those of mouse and *T. nigroviridis*), and as computational methods improve. We intend to create and release updated versions of the IGI and IPI regularly, until they converge to a final accurate list of every human gene. The gene predictions will be linked to RefSeq, HUGO and SWISSPROT identifiers where available, and tracking identifiers between versions will be included, so that individual genes under study can be traced forwards as the human sequence is completed.

Comparative proteome analysis

Knowledge of the human proteome will provide unprecedented opportunities for studies of human gene function. Often clues will be provided by sequence similarity with proteins of known function in model organisms. Such initial observations must then be followed up by detailed studies to establish the actual function of these molecules in humans.

For example, 35 proteins are known to be involved in the vacuolar protein-sorting machinery in yeast. Human genes encoding homologues can be found in the draft human sequence for 34 of these yeast proteins, but precise relationships are not always clear. In nine cases there appears to be a single clear human orthologue (a gene that arose as a consequence of speciation); in 12 cases there are matches to a family of human paralogues (genes that arose owing to intra-genome duplication); and in 13 cases there are matches to specific protein domains[311–314]. Hundreds of similar stories emerge from the draft sequence, but each merits a detailed interpretation in context. To treat these subjects properly, there will be many following studies, the first of which appear in accompanying papers[315–323].

Here, we aim to take a more global perspective on the content of the human proteome by comparing it with the proteomes of yeast, worm, fly and mustard weed. Such comparisons shed useful light on the commonalities and differences among these eukaryotes[294,324,325]. The analysis is necessarily preliminary, because of the imperfect nature of the human sequence, uncertainties in the gene and protein sets for all of the multicellular organisms considered and our incomplete knowledge of protein structures. Nonetheless, some general patterns emerge. These include insights into fundamental mechanisms that create functional diversity, including invention of protein domains, expansion of protein and domain families, evolution of new protein architectures and horizontal transfer of genes. Other mechanisms, such as alternative splicing, post-translational modification and complex regulatory networks, are also crucial in generating diversity but are much harder to discern from the primary sequence. We will not attempt to consider the effects of alternative splicing on proteins; we will consider only a single splice form from each gene in the various organisms, even when multiple splice forms are known.

Functional and evolutionary classification. We began by classifying the human proteome on the basis of functional categories and evolutionary conservation. We used the InterPro annotation protocol to identify conserved biochemical and cellular processes. InterPro is a tool for combining sequence-pattern information from four databases. The first two databases (PRINTS[326] and Prosite[327]) primarily contain information about motifs corresponding to specific family subtypes, such as type II receptor tyrosine kinases (RTK-II) in particular or tyrosine kinases in general. The second two databases (Pfam[307] and Prosite Profile[327]) contain information (in the form of profiles or HMMs) about families of structural domains—for example, protein kinase domains. InterPro integrates the motif and domain assignments into a hierarchical classification system; so a protein might be classified at the most detailed level as being an RTK-II, at a more general level as being a kinase specific for tyrosine, and at a still more general level as being a protein kinase. The complete hierarchy of InterPro entries is described at http://www.ebi.ac.uk/interpro/. We collapsed the InterPro entries into 12 broad categories, each reflecting a set of cellular functions.

The InterPro families are partly the product of human judgement and reflect the current state of biological and evolutionary knowledge. The system is a valuable way to gain insight into large collections of proteins, but not all proteins can be classified at present. The proportions of the yeast, worm, fly and mustard weed protein sets that are assigned to at least one InterPro family is, for each organism, about 50% (Table 23; refs 307, 326, 327).

About 40% of the predicted human proteins in the IPI could be assigned to InterPro entries and functional categories. On the basis of these assignments, we could compare organisms according to the number of proteins in each category (Fig. 37). Compared with the two invertebrates, humans appear to have many proteins involved in cytoskeleton, defence and immunity, and transcription and translation. These expansions are clearly related to aspects of vertebrate physiology. Humans also have many more proteins that are classified as falling into more than one functional category (426 in human versus 80 in worm and 57 in fly, data not shown). Interestingly, 32% of these are transmembrane receptors.

We obtained further insight into the evolutionary conservation of proteins by comparing each sequence to the complete nonredundant database of protein sequences maintained at NCBI, using the BLASTP computer program[328] and then breaking down the matches according to organismal taxonomy (Fig. 38). Overall, 74% of the proteins had significant matches to known proteins.

Such classifications are based on the presence of clearly detectable homologues in existing databases. Many of these genes have surely evolved from genes that were present in common ancestors but have since diverged substantially. Indeed, one can detect more distant relationships by using sensitive computer programs that can recognize weakly conserved features. Using PSI-BLAST, we can recognize probable nonvertebrate homologues for about 45% of the 'vertebrate-specific' set. Nonetheless, the classification is useful for gaining insights into the commonalities and differences among the proteomes of different organisms.

Probable horizontal transfer. An interesting category is a set of 223 proteins that have significant similarity to proteins from bacteria, but no comparable similarity to proteins from yeast, worm, fly and

© 2001 Macmillan Magazines Ltd

Table 23 Properties of genome and proteome in essentially completed eukaryotic proteomes

	Human	Fly	Worm	Yeast	Mustard weed
Number of identified genes	~32,000*	13,338	18,266	6,144	25,706
% with InterPro matches	51	56	50	50	52
Number of annotated domain families	1,262	1,035	1,014	851	1,010
Number of InterPro entries per gene	0.53	0.84	0.63	0.6	0.62
Number of distinct domain architectures	1,695	1,036	1,018	310	–
Percentage of 1-1-1-1	1.40	4.20	3.10	9.20	–
% Signal sequences	20	20	24	11	–
% Transmembrane proteins	20	25	28	15	–
% Repeat-containing	10	11	9	5	–
% Coiled-coil	11	13	10	9	–

The numbers of distinct architectures were calculated using SMART[339] and the percentages of repeat-containing proteins were estimated using Prospero[452] and a *P*-value threshold of 10^{-5}. The protein sets used in the analysis were taken from http://www.ebi.ac.uk/proteome/ for yeast, worm and fly. The proteins from mustard weed were taken from the TAIR website (http:// www.arabidopsis.org/) on 5 September 2000. The protein set was searched against the InterPro database (http://www.ebi.ac.uk/interpro/) using the InterProscan software. Comparison of protein sequences with the InterPro database allows prediction of protein families, domain and repeat families and sequence motifs. The searches used Pfam release 5.2[307], Prints release 26.1[326], Prosite release 16[327] and Prosite preliminary profiles. InterPro analysis results are available as Supplementary Information. The fraction of 1-1-1-1 is the percentage of the genome that falls into orthologous groups composed of only one member each in human, fly, worm and yeast.
*The gene number for the human is still uncertain (see text). Table is based on 31,778 known genes and gene predictions.

mustard weed, or indeed from any other (nonvertebrate) eukaryote. These sequences should not represent bacterial contamination in the draft human sequence, because we filtered the sequence to eliminate sequences that were essentially identical to known bacterial plasmid, transposon or chromosomal DNA (such as the host strains for the large-insert clones). To investigate whether these were genuine human sequences, we designed PCR primers for 35 of these genes and confirmed that most could be readily detected directly in human genomic DNA (Table 24). Orthologues of many of these genes have also been detected in other vertebrates (Table 24).

A more detailed computational analysis indicated that at least 113 of these genes are widespread among bacteria, but, among eukaryotes, appear to be present only in vertebrates. It is possible that the genes encoding these proteins were present in both early prokaryotes and eukaryotes, but were lost in each of the lineages of yeast, worm, fly, mustard weed and, possibly, from other nonvertebrate eukaryote lineages. A more parsimonious explanation is that these genes entered the vertebrate (or prevertebrate) lineage by horizontal transfer from bacteria. Many of these genes contain introns, which presumably were acquired after the putative horizontal transfer event. Similar observations indicating probable lineage-specific horizontal gene transfers, as well as intron insertion in the acquired genes, have been made in the worm genome[329].

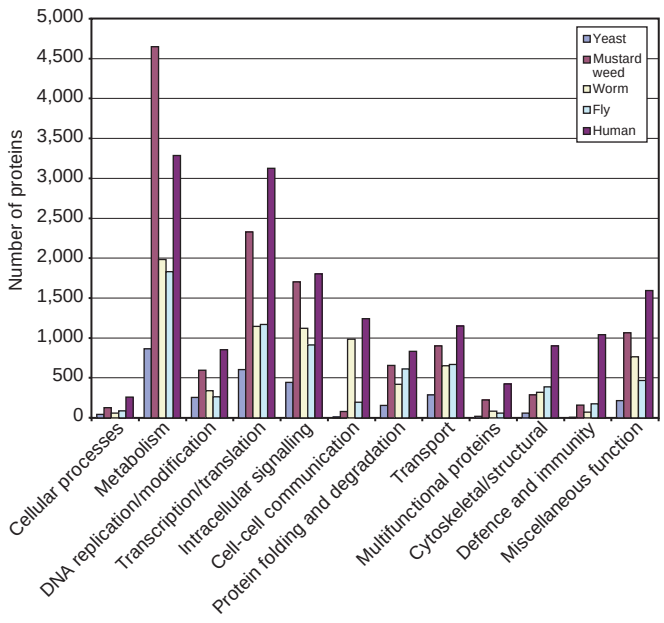


Figure 37 Functional categories in eukaryotic proteomes. The classification categories were derived from functional classification systems, including the top-level biological function category of the Gene Ontology project (GO; see http://www.geneontology.org).

We cannot formally exclude the possibility that gene transfer occurred in the opposite direction—that is, that the genes were invented in the vertebrate lineage and then transferred to bacteria. However, we consider this less likely. Under this scenario, the broad distribution of these genes among bacteria would require extensive horizontal dissemination after their initial acquisition. In addition, the functional repertoire of these genes, which largely encode intracellular enzymes (Table 24), is uncharacteristic of vertebrate-specific evolutionary innovations (which appear to be primarily extracellular proteins; see below).

We did not identify a strongly preferred bacterial source for the putative horizontally transferred genes, indicating the likelihood of multiple independent gene transfers from different bacteria (Table 24). Notably, several of the probable recent acquisitions have established (or likely) roles in metabolism of xenobiotics or stress response. These include several hydrolases of different specificities, including epoxide hydrolase, and several dehydrogenases (Table 24). Of particular interest is the presence of two paralogues of monoamine oxidase (MAO), an enzyme of the mitochondrial outer membrane that is central in the metabolism of neuromediators and is a target of important psychiatric drugs[330–333]. This example shows that at least some of the genes thought to be horizontally transferred into the vertebrate lineage appear to be involved in important physiological functions and so probably have been fixed and maintained during evolution because

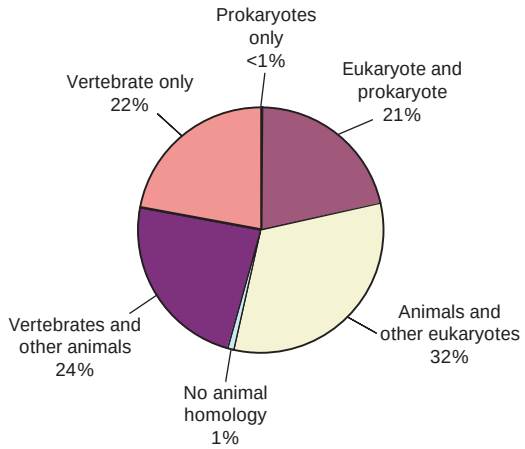


Figure 38 Distribution of the homologues of the predicted human proteins. For each protein, a homologue to a phylogenetic lineage was considered present if a search of the NCBI nonredundant protein sequence database, using the gapped BLASTP program, gave a random expectation (*E*) value of ≤ 0.001. Additional searches for probable homologues with lower sequence conservation were performed using the PSI-BLAST program, run for three iterations using the same cut-off for inclusion of sequences into the profile[328].

© 2001 Macmillan Magazines Ltd

of the increased selective advantage(s) they provide.

Genes shared with fly, worm and yeast. IPI.1 contains apparent homologues of 61% of the fly proteome, 43% of the worm proteome and 46% of the yeast proteome. We next considered the groups of proteins containing likely orthologues and paralogues (genes that arose from intragenome duplication) in human, fly, worm and yeast.

Briefly, we performed all-against-all sequence comparison[334] for the combined protein sets of human, yeast, fly and worm. Pairs of sequences that were one another's best matches in their respective genomes were considered to be potential orthologues. These were then used to identify orthologous groups across three organisms[335]. Recent species-specific paralogues were defined by using the all-against-all sequence comparison to cluster the protein set for each organism. For each sequence found in an orthologous group, the recent paralogues were defined to be the largest species-specific cluster including it. The set of paralogues may be inflated by unrecognized splice variants and by fragmentation.

We identified 1,308 groups of proteins, each containing at least one predicted orthologue in each species and many containing additional paralogues. The 1,308 groups contained 3,129 human proteins, 1,445 fly proteins, 1,503 worm proteins and 1,441 yeast proteins. These 1,308 groups represent a conserved core of proteins that are mostly responsible for the basic 'housekeeping' functions of the cell, including metabolism, DNA replication and repair, and translation.

In 564 of the 1,308 groups, one orthologue (and no additional paralogues) could be unambiguously assigned for each of human, fly, worm and yeast. These groups will be referred to as 1-1-1-1 groups. More than half (305) of these groups could be assigned to the functional categories shown in Fig. 37. Within these functional categories, the numbers of groups containing single orthologues in each of the four proteomes was: 19 for cellular processes, 66 for metabolism, 31 for DNA replication and modification, 106 for transcription/translation, 13 for intracellular signalling, 24 for protein folding and degradation, 38 for transport, 5 for

Table 24 Probable vertebrate-specific acquisitions of bacterial genes

Human protein (accession)	Predicted function	Known orthologues in other vertebrates	Bacterial homologues		Human origin confirmed by PCR
			Range	Best hit	
AAG01853.1	Formiminotransferase cyclodeaminase	Pig, rat, chicken	*Thermotoga*, *Thermoplasma*, *Methylobacter*	*Thermotoga maritima*	Yes
CAB81772.1 AAB59448.1	Na/glucose cotransporter	Rodents, ungulates	Most bacteria	*Vibrio parahaemolyticus*	Yes (CAB81772, AAC41747.1) NT* (AAB59448.1, AAA36608.1)
AAA36608.1 AAC41747.1					
BAA1143.21	Epoxide hydrolase (α/β-hydrolase)	Mouse, *Danio*, fugu fish	Most bacteria	*Pseudomonas aeruginosa*	Yes
CAB59628.1	Protein-methionine-*S*-oxide reductase	Cow	Most bacteria	*Synechocystis sp.*	Yes
BAA91273.1	Hypertension-associated protein SA/acetate-CoA ligase	Mouse, rat, cow	Most bacteria	*Bacillus halodurans*	NT*
CAA75608.1	Glucose-6-phosphate transporter/glycogen storage disease type 1b protein	Mouse, rat	Most bacteria	*Chlamydophila pneumoniae*	Yes
AAA59548.1 AAB27229.1	Monoamine oxidase	Cow, rat, salmon	Most bacteria	*Mycobacterium tuberculosis*	Yes
AAF12736.1 AAA51565.1	Acyl-CoA dehydrogenase, mitochondrial protein	Mouse, rat, pig	Most bacteria	*P. aeruginosa*	Yes
IGI_M1_ctg19153_147	Aldose-1-epimerase	Pig (also found in plants)	*Streptomyces*, *Bacillus*	*Streptomyces coelicolor*	Yes
BAA92632.1	Predicted carboxylase (C-terminal domain, N-terminal domain unique)	None	*Streptomyces, Rhizobium, Bacillus*	*S. coelicolor*	Yes
BAA34458.1	Uncharacterized protein	None	Gamma-proteobacteria	*Escherichia coli*	Yes
AAF24044.1	Uncharacterized protein	None	Most bacteria	*T. maritima*	Yes
BAA34458.1	β-Lactamase superfamily hydrolase	None	Most bacteria	*Synechocystis* sp.	Yes
BAA91839.1	Oxidoreductase (Rossmann fold) fused to a six-transmembrane protein	None (several human paralogues of both parts)	*Actinomycetes*, *Leptospira*; more distant homologues in other bacteria	*S. coelicolor*	Yes
BAA92073.1	Oxidoreductase (Rossmann fold)	None	*Synechocystis, Pseudomonas*	*Synechocystis* sp.	Yes
BAA92133.1	α/β-hydrolase	None	*Rickettsia*; more distant homologues in other bacteria	*Rickettsia prowazekii*	Yes
BAA91174.1	ADP-ribosylglycohydrolase	None	*Streptomyces*, *Aquifex*, *Archaeoglobus* (archaeon), *E. coli*	*S. coelicolor*	Yes
AAA60043.1	Thymidine phosporylase/endothelial cell growth factor	None	Most bacteria	*Bacillus stearothermophilus*	Yes
BAA86552.1	Ribosomal protein S6-glutamic acid ligase	None	Most bacteria and archaea	*Haemophilus influenzae*	Yes
IGI_M1_ctg12741_7	Ribosomal protein S6-glutamic acid ligase (paralogue of the above)	None	Most bacteria and archaea	*H. influenzae*	Yes
IGI_M1_ctg13238_61	Hydratase	None	*Synechocystis*, *Sphingomonas*	*Synechocystis* sp.	Yes
IGI_M1_ctg13305_116	Homologue of histone macro-2A C-terminal domain, predicted phosphatase	None (several human paralogues, RNA viruses)	*Thermotoga*, *Alcaligenes*, *E. coli*, more distant homologues in other bacteria	*T. maritima*	Yes
IGI_M1_ctg14420_10	Sugar transporter	None	Most bacteria	*Synechocystis* sp.	Yes
IGI_M1_ctg16010_18	Predicted metal-binding protein	None	Most bacteria	*Borrelia burgdorferi*	Yes
IGI_M1_ctg16227_58	Pseudouridine synthase	None	Most bacteria	*Zymomonas mobilis*	Yes
IGI_M1_ctg25107_24	Surfactin synthetase domain	None	Gram-positive bacteria, *Actinomycetes*, *Cyanobacteria*	*Bacillus subtilis*	Yes

*NT, not tested.

Representative genes confirmed by PCR to be present in the human genome. The similarity to a bacterial homologue was considered to be 'significantly' greater than that to eukaryotic homologues if the difference in alignment scores returned by BLASTP was greater than 30 bits (~9 orders of magnitude in terms of *E*-value). A complete, classified and annotated list of probable vertebrate-specific horizontal gene transfers detected in this analysis is available as Supplementary Information. cDNA sequences for each protein were searched, using the SSAHA algorithm, against the draft genome sequence. Primers were designed and PCR was performed using three human genomic samples and a random BAC clone. The predicted genes were considered to be present in the human genome if a band of the expected size was found in all three human samples but not in the control clone.

© 2001 Macmillan Magazines Ltd

multifunctional proteins and 3 for cytoskeletal/structural. No such groups were found for defence and immunity or cell–cell communication.

The 1-1-1-1 groups probably represent key functions that have not undergone duplication and elaboration in the various lineages. They include many anabolic enzymes responsible for such functions as respiratory chain and nucleotide biosynthesis. In contrast, there are few catabolic enzymes. As anabolic pathways branch less frequently than catabolic pathways, this indicates that alternative routes and displacements are more frequent in catabolic reactions. If proteins from the single-celled yeast are excluded from the analysis, there are 1,195 1-1-1 groups. The additional groups include many examples of more complex signalling proteins, such as receptor-type and src-like tyrosine kinases, likely to have arisen early in the metazoan lineage. The fact that this set comprises only a small proportion of the proteome of each of the animals indicates that, apart from a modest conserved core, there has been extensive elaboration and innovation within the protein complement.

Most proteins do not show simple 1-1-1 orthologous relationships across the three animals. To illustrate this, we investigated the nuclear hormone receptor family. In the human proteome, this family consists of 60 different 'classical' members, each with a zinc finger and a ligand-binding domain. In comparison, the fly proteome has 19 and the worm proteome has 220. As shown in Fig. 39, few simple orthologous relationships can be derived among these homologues. And, where potential subgroups of orthologues and paralogues could be identified, it was apparent that the functions of the subgroup members could differ significantly. For example, the fly receptor for the fly-specific hormone ecdysone and the human retinoic acid receptors cluster together on the basis of sequence similarity. Such examples underscore that the assignment of functional similarity on the basis of sequence similarities among these three organisms is not trivial in most cases.

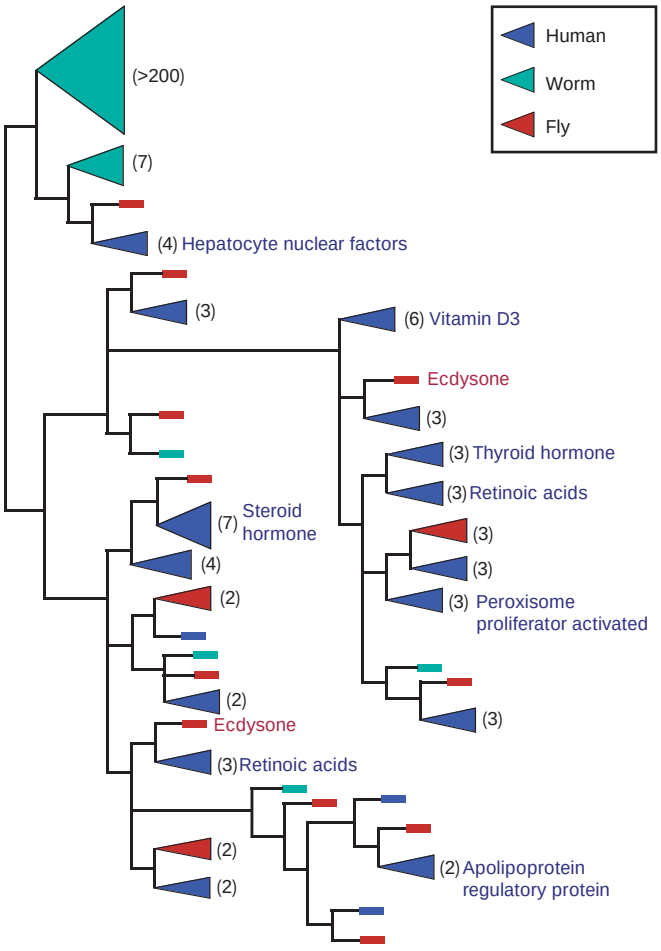


Figure 39 Simplified cladogram (relationship tree) of the 'many-to-many' relationships of classical nuclear receptors. Triangles indicate expansion within one lineage; bars represent single members. Numbers in parentheses indicate the number of paralogues in each group.

New vertebrate domains and proteins. We then explored how the proteome of vertebrates (as represented by the human) differs from those of the other species considered. The 1,262 InterPro families were scanned to identify those that contain only vertebrate proteins. Only 94 (7%) of the families were 'vertebrate-specific'. These represent 70 protein families and 24 domain families. Only one of the 94 families represents enzymes, which is consistent with the ancient origins of most enzymes[336]. The single vertebrate-specific enzyme family identified was the pancreatic or eosinophil-associated ribonucleases. These enzymes evolved rapidly, possibly to combat vertebrate pathogens[337].

The relatively small proportion of vertebrate-specific multicopy families suggests that few new protein domains have been invented in the vertebrate lineage, and that most protein domains trace at least as far back as a common animal ancestor. This conclusion must be tempered by the fact that the InterPro classification system is incomplete; additional vertebrate-specific families undoubtedly exist that have not yet been recognized in the InterPro system.

The 94 vertebrate-specific families appear to reflect important physiological differences between vertebrates and other eukaryotes. Defence and immunity proteins (23 families) and proteins that function in the nervous system (17 families) are particularly enriched in this set. These data indicate the recent emergence or rapid divergence of these proteins.

Representative human proteins were previously known for nearly all of the vertebrate-specific families. This was not surprising, given the anthropocentrism of biological research. However, the analysis did identify the first mammalian proteins belonging to two of these families. Both of these families were originally defined in fish. The first is the family of polar fish antifreeze III proteins. We found a human sialic acid synthase containing a domain homologous to polar fish antifreeze III protein (BAA91818.1). This finding suggests that fish created the antifreeze function by adaptation of this domain. We also found a human protein (CAB60269.1) homologous to the ependymin found in teleost fish. Ependymins are major glycoproteins of fish brains that have been claimed to be involved in long-term memory formation[338]. The function of the mammalian ependymin homologue will need to be elucidated.

New architectures from old domains. Whereas there appears to be only modest invention at the level of new vertebrate protein domains, there appears to be substantial innovation in the creation of new vertebrate proteins. This innovation is evident at the level of domain architecture, defined as the linear arrangement of domains within a polypeptide. New architectures can be created by shuffling, adding or deleting domains, resulting in new proteins from old parts.

We quantified the number of distinct protein architectures found in yeast, worm, fly and human by using the SMART annotation resource[339] (Fig. 40). The human proteome set contained 1.8 times as many protein architectures as worm or fly and 5.8 times as many as yeast. This difference is most prominent in the recent evolution of novel extracellular and transmembrane architectures in the human lineage. Human extracellular proteins show the greatest innovation: the human has 2.3 times as many extracellular architectures as fly and 2.0 times as many as worm. The larger number of human architectures does not simply reflect differences in the number of domains known in these organisms; the result remains qualitatively the same even if the number of architectures in each organism is normalized by dividing by the total number of domains (not shown). (We also checked that the larger number of human

© 2001 Macmillan Magazines Ltd

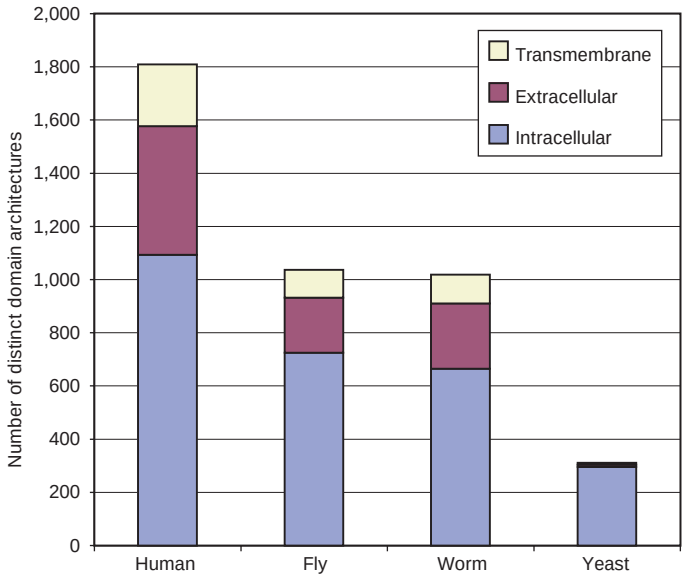


Figure 40 Number of distinct domain architectures in the four eukaryotic genomes, predicted using SMART[339]. The number of architectures is split into three cellular environments: intracellular, extracellular and membrane-associated. The increase in architectures for the human, relative to the other lineages, is seen when these numbers are normalized with respect to the numbers of domains predicted in each phylum. To avoid artefactual results from the relatively low detection rate for some repeat types, tandem occurrences of tetratricopeptide, armadillo, EF-hand, leucine-rich, WD40 or ankyrin repeats or C2H2-type zinc fingers were treated as single occurrences.

architectures could not be an artefact resulting from erroneous gene predictions. Three-quarters of the architectures can be found in known genes, which already yields an increase of about 50% over worm and fly. We expect the final number of human architectures to grow as the complete gene set is identified.)

A related measure of proteome complexity can be obtained by considering an individual domain and counting the number of different domain types with which it co-occurs. For example, the trypsin-like serine protease domain (number 12 in Fig. 41) co-occurs with 18 domain types in human (including proteins involved in the mammalian complement system, blood coagulation, and fibrinolytic and related systems). By contrast, the trypsin-like serine protease domain occurs with only eight other domains in fly, five in worm and one in yeast. Similar results for 27 common domains are shown in Fig. 41. In general, there are more different co-occurring domains in the human proteome than in the other proteomes.

One mechanism by which architectures evolve is through the fusion of additional domains, often at one or both ends of the proteins. Such 'domain accretion'[340] is seen in many human proteins when compared with proteins from other eukaryotes. The effect is illustrated by several chromatin-associated proteins (Fig. 42). In these examples, the domain architectures of human proteins differ from those found in yeast, worm and fly proteins only by the addition of domains at their termini.

Among chromatin-associated proteins and transcription factors, a significant proportion of domain architectures is shared between the vertebrate and fly, but not with worm (Fig. 43a). The trend was even more prominent in architectures of proteins involved in another key cellular process, programmed cell death (Fig. 43b). These examples might seem to bear upon the unresolved issue of the evolutionary branching order of worms, flies and humans, suggesting that worms branched off first. However, there were other cases in which worms and humans shared architectures not present in fly. A global analysis of shared architectures could not conclusively distinguish between the two models, given the possibility of lineage-specific loss of architectures. Comparison of protein architectures may help to resolve the evolutionary issue, but it will require more detailed analyses of many protein families.

New physiology from old proteins. An important aspect of

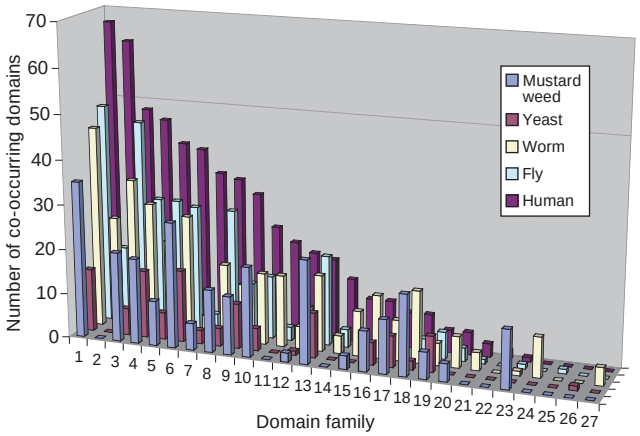


Figure 41 Number of different Pfam domain types that co-occur in the same protein, for each of the 10 most common domain families in each of the five eukaryotic proteomes. Because some common domain families are shared, there are 27 families rather than 50. The data are ranked according to decreasing numbers of human co-occurring Pfam domains. The domain families are: (1) eukaryotic protein kinase [IPR000719]; (2) immunoglobulin domain [IPR003006]; (3) ankyrin repeat [IPR002110]; (4) RING finger [IPR001841]; (5) C2H2-type zinc finger [IPR000822]; (6) ATP/GTP-binding P-loop [IPR001687]; (7) reverse transcriptase (RNA-dependent DNA polymerase) [IPR000477]; (8) leucine-rich repeat [IPR001611]; (9) G-proteinβ WD-40 repeats [IPR001680]; (10) RNA-binding region RNP-1 (RNA recognition motif) [IPR000504]; (11) C-type lectin domain [IPR001304]; (12) serine proteases, trypsin family [IPR001254]; (13) helicase C-terminal domain [IPR001650]; (14) collagen triple helix repeat [IPR000087]; (15) rhodopsin-like GPCR superfamily [IPR000276]; (16) esterase/lipase/thioesterase [IPR000379]; (17) Myb DNA-binding domain [IPR001005]; (18) F-box domain [IPR001810]; (19) ATP-binding transport protein, 2nd P-loop motif [IPR001051]; (20) homeobox domain [IPR001356]; (21) C4-type steroid receptor zinc finger [IPR001628]; (22) sugar transporter [IPR001066]; (23) PPR repeats [IPR002885]; (24) seven-helix G-protein-coupled receptor, worm (probably olfactory) family [IPR000168]; (25) cytochrome P450 enzyme [IPR001128]; (26) fungal transcriptional regulatory protein, N terminus [IPR001138]; (27) domain of unknown function DUF38 [IPR002900].

© 2001 Macmillan Magazines Ltd

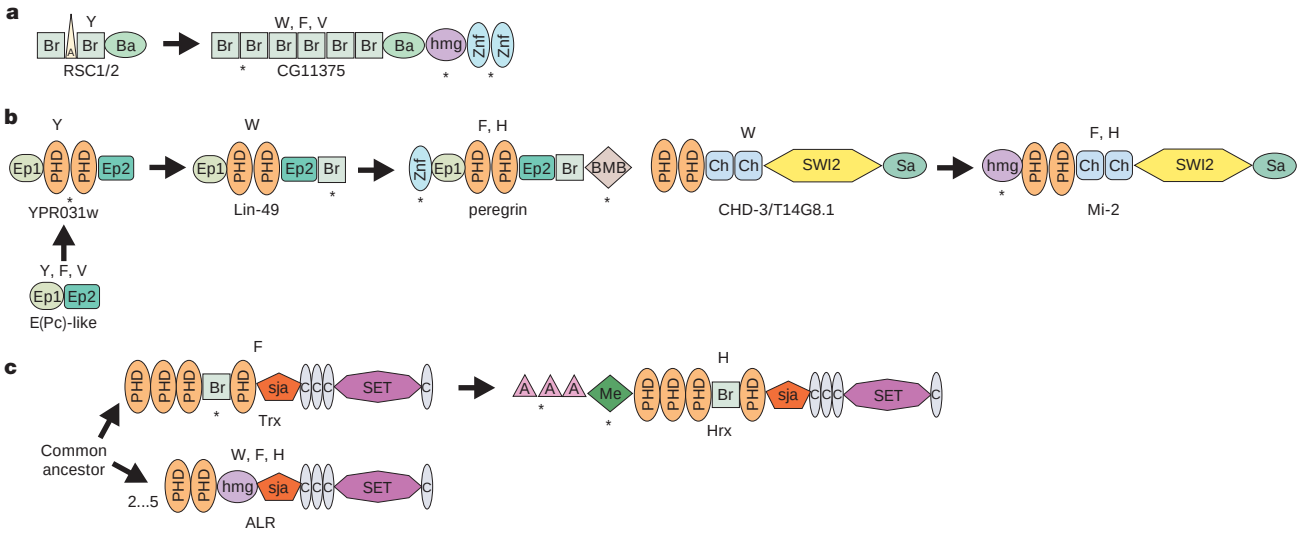


Figure 42 Examples of domain accretion in chromatin proteins. Domain accretion in various lineages before the animal divergence, in the apparent coelomate lineage and the vertebrate lineage are shown using schematic representations of domain architectures (not to scale). Asterisks, mobile domains that have participated in the accretion. Species in which a domain architecture has been identified are indicated above the diagram (Y, yeast; W, worm; F, fly; V, vertebrate). Protein names are below the diagrams. The domains are SET, a chromatin protein methyltransferase domain; SWI2, a superfamily II helicase/ATPase domain; Sa, sant domain; Br, bromo domain; Ch, chromodomain; C, a cysteine triad motif associated with the Msl-2 and SET domains; A, AT hook motif; EP1/EP2, enhancer of polycomb domains 1 and 2; Znf, zinc finger; sja, SET-JOR-associated domain (L. Aravind, unpublished); Me, DNA methylase/Hrx-associated DNA binding zinc finger; Ba, bromo-associated homology motif. **a**–**c**, Different examples of accretion.

vertebrate innovation lies in the expansion of protein families. Table 25 shows the most prevalent protein domains and protein families in humans, together with their relative ranks in the other species. About 60% of families are more numerous in the human than in any of the other four organisms. This shows that gene duplication has been a major evolutionary force during vertebrate evolution. A comparison of relative expansions in human versus fly is shown in Fig. 44.

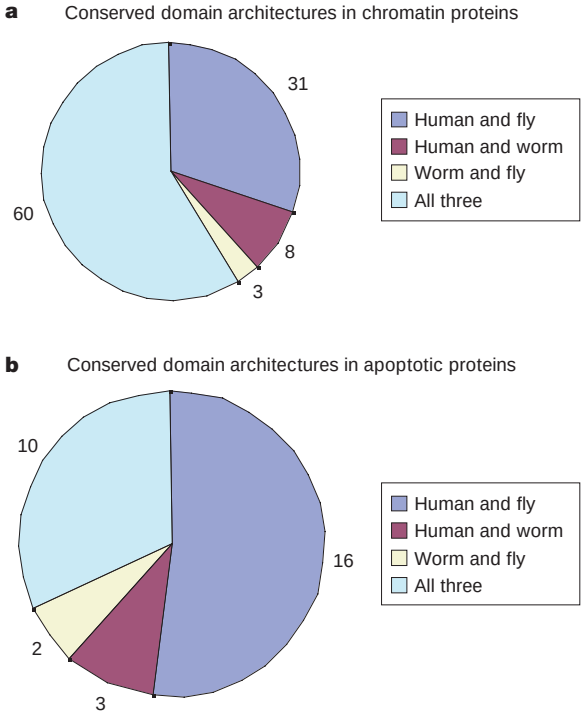


Figure 43 Conservation of architectures between animal species. The pie charts illustrate the shared domain architectures of apparent orthologues that are conserved in at least two of the three sequenced animal genomes. If an architecture was detected in fungi or plants, as well as two of the animal lineages, it was omitted as ancient and its absence in the third animal lineage attributed to gene loss. **a**, Chromatin-associated proteins. **b**, Components of the programmed cell death system.

Many of the families that are expanded in human relative to fly and worm are involved in distinctive aspects of vertebrate physiology. An example is the family of immunoglobulin (IG) domains, first identified in antibodies thirty years ago. Classic (as opposed to divergent) IG domains are completely absent from the yeast and mustard weed proteomes and, although prokaryotic homologues exist, they have probably been transferred horizontally from metazoans[341]. Most IG superfamily proteins in invertebrates are cell-surface proteins. In vertebrates, the IG repertoire includes immune functions such as those of antibodies, MHC proteins, antibody receptors and many lymphocyte cell-surface proteins. The large expansion of IG domains in vertebrates shows the versatility of a single family in evoking rapid and effective response to infection.

Two prominent families are involved in the control of development. The human genome contains 30 fibroblast growth factors (FGFs), as opposed to two FGFs each in the fly and worm. It contains 42 transforming growth factor-βs (TGFβs) compared with nine and six in the fly and worm, respectively. These growth factors are involved in organogenesis, such as that of the liver and the lung. A fly FGF protein, branchless, is involved in developing respiratory organs (tracheae) in embryos[342]. Thus, developmental triggers of morphogenesis in vertebrates have evolved from related but simpler systems in invertebrates[343].

Another example is the family of intermediate filament proteins, with 127 family members. This expansion is almost entirely due to 111 keratins, which are chordate-specific intermediate filament proteins that form filaments in epithelia. The large number of human keratins suggests multiple cellular structural support roles for the many specialized epithelia of vertebrates.

Finally, the olfactory receptor genes comprise a huge gene family of about 1,000 genes and pseudogenes[344,345]. The number of olfactory receptors testifies to the importance of the sense of smell in vertebrates. A total of 906 olfactory receptor genes and pseudogenes could be identified in the draft genome sequence, two-thirds of which were not previously annotated. About 80% are found in about two dozen clusters ranging from 6 to 138 genes and encompassing about 30 Mb (~1%) of the human genome. Despite the importance of smell among our vertebrate ancestors, hominids

© 2001 Macmillan Magazines Ltd

appear to have considerably less interest in this sense. About 60% of the olfactory receptors in the draft genome sequence have disrupted ORFs and appear to be pseudogenes, consistent with recent reports[344,346] suggesting massive functional gene loss in the last 10 Myr[347,348]. Interestingly, there appears to be a much higher proportion of intact genes among class I than class II olfactory receptors, suggesting functional importance.

Vertebrates are not unique in employing gene family expansion. For many domain types, expansions appear to have occurred independently in each of the major eukaryotic lineages. A good example is the classical C2H2 family of zinc finger domains, which have expanded independently in the yeast, worm, fly and human lineages (Fig. 45). These independent expansions have resulted in numerous C2H2 zinc finger domain-containing proteins that are specific to each lineage. In flies, the important components of the C2H2 zinc finger expansion are architectures in which it is combined with the POZ domain and the C4DM domain (a metal-binding domain found only in fly). In humans, the most prevalent expansions are combinations of the C2H2 zinc finger with POZ (independent of the one in insects) and the vertebrate-specific KRAB and SCAN domains.

The homeodomain is similarly expanded in all animals and is present in both architectures that are conserved and lineage-specific architectures (Fig. 45). This indicates that the ancestral animal probably encoded a significant number of homeodomain proteins, but subsequent evolution involved multiple, independent expansions and domain shuffling after lineages diverged. Thus, the most prevalent transcription factor families are different in worm, fly and human (Fig. 45). This has major biological implications because transcription factors are critical in animal development and differentiation. The emergence of major variations in the developmental body plans that accompanied the early radiation of the animals[349] could have been driven by lineage-specific proliferation of such transcription factors. Beyond these large expansions of protein families, protein components of particular functional systems such as the cell death signalling system show a general increase in diversity and numbers in the vertebrates relative to other animals. For example, there are greater numbers of and more novel architectures in cell death regulatory proteins such as BCL-2, TNFR and NFκB from vertebrates.

Conclusion. Five lines of evidence point to an increase in the complexity of the proteome from the single-celled yeast to the multicellular invertebrates and to vertebrates such as the human. Specifically, the human contains greater numbers of genes, domain and protein families, paralogues, multidomain proteins with multiple functions, and domain architectures. According to these measures, the relatively greater complexity of the human proteome is a consequence not simply of its larger size, but also of large-scale protein innovation.

An important question is the extent to which the greater phenotypic complexity of vertebrates can be explained simply by two- or threefold increases in proteome complexity. The real explanation may lie in combinatorial amplification of these modest differences, by mechanisms that include alternative splicing, post-translational modification and cellular regulatory networks. The potential numbers of different proteins and protein–protein interactions are vast, and their actual numbers cannot readily be discerned from the genome sequence. Elucidating such system-level properties presents one of the great challenges for modern biology.

Table 25 The most populous InterPro families in the human proteome and other species

	Human		Fly		Worm		Yeast		Mustard weed		
InterPro ID	No. of genes	Rank	No. of genes	Rank	No. of genes	Rank	No. of genes	Rank	No. of genes	Rank	
IPR003006	765	(1)	140	(9)	64	(34)	0	(na)	0	(na)	Immunoglobulin domain
PR000822	706	(2)	357	(1)	151	(10)	48	(7)	115	(20)	C2H2 zinc finger
IPR000719	575	(3)	319	(2)	437	(2)	121	(1)	1049	(1)	Eukaryotic protein kinase
IPR000276	569	(4)	97	(14)	358	(3)	0	(na)	16	(84)	Rhodopsin-like GPCR superfamily
IPR001687	433	(5)	198	(4)	183	(7)	97	(2)	331	(5)	P-loop motif
IPR000477	350	(6)	10	(65)	50	(41)	6	(36)	80	(35)	Reverse transcriptase (RNA-dependent DNA polymerase)
IPR000504	300	(7)	157	(6)	96	(21)	54	(6)	255	(8)	rrm domain
IPR001680	277	(8)	162	(5)	102	(19)	91	(3)	210	(10)	G-protein β WD-40 repeats
IPR002110	276	(9)	105	(13)	107	(17)	19	(23)	120	(18)	Ankyrin repeat
IPR001356	267	(10)	148	(7)	109	(15)	9	(33)	118	(19)	Homeobox domain
IPR001849	252	(11)	77	(22)	71	(31)	27	(17)	27	(73)	PH domain
IPR002048	242	(12)	111	(12)	81	(25)	15	(27)	167	(12)	EF-hand family
IPR000561	222	(13)	81	(20)	113	(14)	0	(na)	17	(83)	EGF-like domain
IPR001452	215	(14)	72	(23)	62	(35)	25	(18)	3	(97)	SH3 domain
IPR001841	210	(15)	114	(11)	126	(12)	35	(12)	379	(4)	RING finger
IPR001611	188	(16)	115	(10)	54	(38)	7	(35)	392	(2)	Leucine-rich repeat
IPR001909	171	(17)	0	(na)	0	(na)	0	(na)	0	(na)	KRAB box
IPR001777	165	(18)	63	(27)	51	(40)	2	(40)	4	(96)	Fibronectin type III domain
IPR001478	162	(19)	70	(24)	66	(33)	2	(40)	15	(85)	PDZ domain
IPR001650	155	(20)	87	(17)	78	(27)	79	(4)	148	(13)	Helicase C-terminal domain
IPR001440	150	(21)	86	(18)	46	(43)	36	(11)	125	(17)	TPR repeat
IPR002216	133	(22)	65	(26)	99	(20)	2	(40)	31	(69)	Ion transport protein
IPR001092	131	(23)	84	(19)	41	(46)	7	(35)	106	(24)	Helix–loop–helix DNA-binding domain
IPR000008	123	(24)	43	(34)	36	(49)	9	(33)	82	(34)	C2 domain
IPR001664	119	(25)	4	(71)	22	(63)	1	(41)	2	(98)	SH2 domain
IPR001254	118	(26)	210	(3)	12	(73)	1	(41)	15	(85)	Serine protease, trypsin family
IPR002126	114	(27)	19	(56)	16	(69)	0	(na)	0	(na)	Cadherin domain
IPR000210	113	(28)	78	(21)	117	(13)	1	(41)	54	(50)	BTB/POZ domain
IPR000387	112	(29)	35	(40)	108	(16)	12	(30)	21	(79)	Tyrosine-specific protein phosphatase and dual specificity protein phosphatase family
IPR000087	106	(30)	18	(57)	169	(9)	0	(na)	5	(95)	Collagen triple helix repeat
IPR000379	94	(31)	141	(8)	134	(11)	40	(10)	194	(11)	Esterase/lipase/thioesterase
IPR000910	89	(32)	38	(38)	18	(67)	8	(34)	18	(82)	HMG1/2 (high mobility group) box
IPR000130	87	(33)	56	(29)	92	(22)	8	(34)	12	(88)	Neutral zinc metallopeptidase
IPR001965	84	(34)	37	(39)	24	(61)	16	(26)	71	(39)	PHD-finger
IPR000636	83	(35)	32	(43)	24	(61)	1	(41)	14	(86)	Cation channels (non-ligand gated)
IPR001781	81	(36)	38	(38)	36	(49)	4	(38)	8	(92)	LIM domain
IPR002035	81	(36)	8	(67)	45	(44)	3	(39)	17	(83)	VWA domain
IPR001715	80	(37)	33	(42)	30	(55)	3	(39)	18	(82)	Calponin homology domain
IPR000198	77	(38)	20	(55)	20	(65)	10	(32)	9	(91)	RhoGAP domain

Forty most populous Interpro families found in the human proteome compared with equivalent numbers from other species. na, not applicable (used when there are no proteins in an organism in that family).

© 2001 Macmillan Magazines Ltd

Segmental history of the human genome

In bacteria, genomic segments often convey important information about function: genes located close to one another often encode proteins in a common pathway and are regulated in a common operon. In mammals, genes found close to each other only rarely have common functions, but they are still interesting because they have a common history. In fact, the study of genomic segments can shed light on biological events as long as 500 Myr ago and as recently as 20,000 years ago.

Conserved segments between human and mouse

Humans and mice shared a common ancestor about 100 Myr ago. Despite the 200 Myr of evolutionary distance between the species, a significant fraction of genes show synteny between the two, being preserved within conserved segments. Genes tightly linked in one mammalian species tend to be linked in others. In fact, conserved segments have been observed in even more distant species: humans show conserved segments with fish[350,351] and even with invertebrates such as fly and worm[352]. In general, the likelihood that a syntenic relationship will be disrupted correlates with the physical distance between the loci and the evolutionary distance between the species.

Studying conserved segments between human and mouse has several uses. First, conservation of gene order has been used to identify likely orthologues between the species, particularly when investigating disease phenotypes. Second, the study of conserved segments among genomes helps us to deduce evolutionary ancestry.

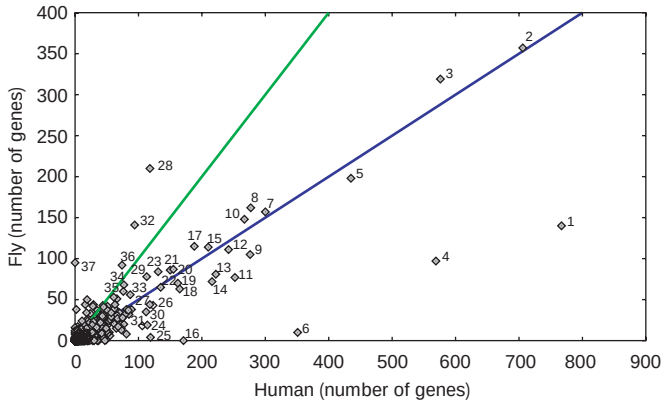


Figure 44 Relative expansions of protein families between human and fly. These data have not been normalized for proteomic size differences. Blue line, equality between normalized family sizes in the two organisms. Green line, equality between unnormalized family sizes. Numbered InterPro entries: (1) immunoglobulin domain [IPR003006]; (2) zinc finger, C2H2 type [IPR000822]; (3) eukaryotic protein kinase [IPR000719]; (4) rhodopsin-like GPCR superfamily [IPR000276]; (5) ATP/GTP-binding site motif A (P-loop) [IPR001687]; (6) reverse transcriptase (RNA-dependent DNA polymerase) [IPR000477]; (7) RNA-binding region RNP-1 (RNA recognition motif) [IPR000504]; (8) G-proteinβ WD-40 repeats [IPR001680]; (9) ankyrin repeat [IPR002110]; (10) homeobox domain [IPR001356]; (11) PH domain [IPR001849]; (12) EF-hand family [IPR002048]; (13) EGF-like domain [IPR000561]; (14) Src homology 3 (SH3) domain [IPR001452]; (15) RING finger [IPR001841]; (16) KRAB box [IPR001909]; (17) leucine-rich repeat [IPR001611]; (18) fibronectin type III domain [IPR001777]; (19) PDZ domain (also known as DHR or GLGF) [IPR001478]; (20) TPR repeat [IPR001440]; (21) helicase C-terminal domain [IPR001650]; (22) ion transport protein [IPR002216]; (23) helix–loop–helix DNA-binding domain [IPR001092]; (24) cadherin domain [IPR002126]; (25) intermediate filament proteins [IPR001664]; (26) C2 domain [IPR000008]; (27) Src homology 2 (SH2) domain [IPR000980]; (28) serine proteases, trypsin family [IPR001254]; (29) BTB/POZ domain [IPR000210]; (30) tyrosine-specific protein phosphatase and dual specificity protein phosphatase family [IPR000387]; (31) collagen triple helix repeat [IPR000087]; (32) esterase/lipase/thioesterase [IPR000379]; (33) neutral zinc metallopeptidases, zinc-binding region [IPR000130]; (34) ATP-binding transport protein, 2nd P-loop motif [IPR001051]; (35) ABC transporters family [IPR001617]; (36) cytochrome P450 enzyme [IPR001128]; (37) insect cuticle protein [IPR000618].

And third, detailed comparative maps may assist in the assembly of the mouse sequence, using the human sequence as a scaffold.

Two types of linkage conservation are commonly described[353]. 'Conserved synteny' indicates that at least two genes that reside on a common chromosome in one species are also located on a common chromosome in the other species. Syntenic loci are said to lie in a 'conserved segment' when not only the chromosomal position but the linear order of the loci has been preserved, without interruption by other chromosomal rearrangements.

An initial survey of homologous loci in human and mouse[354] suggested that the total number of conserved segments would be about 180. Subsequent estimates based on increasingly detailed comparative maps have remained close to this projection[353,355,356] (http://www.informatics.jax.org). The distribution of segment lengths has corresponded reasonably well to the truncated negative exponential curve predicted by the random breakage model[357].

The availability of a draft human genome sequence allows the first global human–mouse comparison in which human physical distances can be measured in Mb, rather than cM or orthologous gene counts. We identified likely orthologues by reciprocal comparison of the human and mouse mRNAs in the LocusLink database, using megaBLAST. For each orthologous pair, we mapped the location of the human gene in the draft genome sequence and then checked the location of the mouse gene in the Mouse Genome Informatics database (http://www.informatics.jax.org). Using a conservative threshold, we identified 3,920 orthologous pairs in which the human gene could be mapped on the draft genome sequence with high confidence. Of these, 2,998 corresponding mouse genes had a known position in the mouse genome. We then searched for definitive conserved segments, defined as human regions containing orthologues of at least two genes from the same mouse chromosome region (< 15 cM) without interruption by segments from other chromosomes.

We identified 183 definitive conserved segments (Fig. 46). The average segment length was 15.4 Mb, with the largest segment being 90.5 Mb and the smallest 24 kb. There were also 141 'singletons', segments that contained only a single locus; these are not counted in the statistics. Although some of these could be short conserved segments, they could also reflect incorrect choices of orthologues or problems with the human or mouse maps. Because of this conservative approach, the observed number of definitive segments is likely be lower than the correct total. One piece of evidence for this conclusion comes from a more detailed analysis on human chromosome 7 (ref. 358), which identified 20 conserved segments, of which three were singletons. Our analysis revealed only 13 definitive segments on this chromosome, with nine singletons.

The frequency of observing a particular gene count in a conserved segment is plotted on a logarithmic scale in Fig. 47. If chromosomal breaks occur in a random fashion (as has been proposed) and differences in gene density are ignored, a roughly straight line should result. There is a clear excess for $n = 1$, suggesting that 50% or more of the singletons are indeed artefactual. Thus, we estimate that true number of conserved segments is around 190–230, in good agreement with the original Nadeau–Taylor prediction[354].

Figure 48 shows a plot of the frequency of lengths of conserved segments, where the x-axis scale is shown in Mb. As before, there is a fair amount of scatter in the data for the larger segments (where the numbers are small), but the trend appears to be consistent with a random breakage model.

We attempted to ascertain whether the breakpoint regions have any special characteristics. This analysis was complicated by imprecision in the positioning of these breaks, which will tend to blur any relationships. With 2,998 orthologues, the average interval within which a break is known to have occurred is about 1.1 Mb. We compared the aggregate features of these breakpoint intervals with the genome as a whole. The mean gene density was lower in breakpoint regions than in the conserved segments (13.8 versus

© 2001 Macmillan Magazines Ltd

18.6 per Mb). This suggests that breakpoints may be more likely to occur or to undergo fixation in gene-poor intervals than in gene-rich intervals. The occurrence of breakpoints may be promoted by homologous recombination among repeated sequences[359]. When the sequence of the mouse genome is finished, this analysis can be revisited more precisely.

A number of examples of extended conserved segments and syntenies are apparent in Fig. 46. As has been noted, almost all human genes on chromosome 17 are found on mouse chromosome 11, with two members of the placental lactogen family from mouse 13 inserted. Apart from two singleton loci, human chromosome 20 appears to be entirely orthologous to mouse chromosome 2, apparently in a single segment. The largest apparently contiguous conserved segment in the human genome is on chromosome 4, including roughly 90.5 Mb of human DNA that is orthologous to mouse chromosome 5. This analysis also allows us to infer the likely location of thousands of mouse genes for which the human orthologue has been located in the draft genome sequence but the mouse locus has not yet been mapped.

With about 200 conserved segments between mouse and human and about 100 Myr of evolution from their common ancestor[360], we obtain an estimated rate of about 1.0 chromosomal rearrangement being fixed per Myr. However, there is good evidence that the rate of chromosomal rearrangement (like the rate of nucleotide substitutions; see above) differs between the two species. Among mammals, rodents may show unusually rapid chromosome alteration. By comparison, very few rearrangements have been observed among primates, and studies of a broader array of mammalian orders, including cats, cows, sheep and pigs, suggest an average rate of chromosome alteration of only about 0.2 rearrangements per Myr in these lineages[361]. Additional evidence that rodents are outliers comes from a recent analysis of synteny between the human and zebrafish genomes. From a study of 523 orthologues, it was possible to project 418 conserved segments[350]. Assuming 400 Myr since a common vertebrate ancestor of zebrafish and humans[362], we obtain an estimate of 0.52 rearrangements per Myr. Recent estimates of rearrangement rates in plants have suggested bimodality, with some pairs showing rates of 0.15–0.41 rearrangements per Myr, and others showing higher rates of 1.1–1.3 rearrangements per Myr[363]. With additional detailed genome maps of multiple species, it should be possible to determine whether this particular molecular clock is truly operating at a different rate in various branches of the evolutionary tree, and whether variations in that rate are bimodal or continuous. It should also be possible to reconstruct the karyotypes of common ancestors.

Ancient duplicated segments in the human genome

Another approach to genomic history is to study segmental duplications within the human genome. Earlier, we discussed examples of recent duplications of genomic segments to pericentromeric and subtelomeric regions. Most of these events appear to be evolutionary dead-ends resulting in nonfunctional pseudogenes; however, segmental duplication is also an important mode of evolutionary innovation: a duplication permits one copy of each gene to drift and potentially to acquire a new function.

Segmental duplications can occur through unequal crossing over to create gene families in specific chromosomal regions. This mechanism can create both small families, such as the five related genes of the β-globin cluster on chromosome 11, and large ones, such as the olfactory receptor gene clusters, which together contain nearly 1,000 genes and pseudogenes.

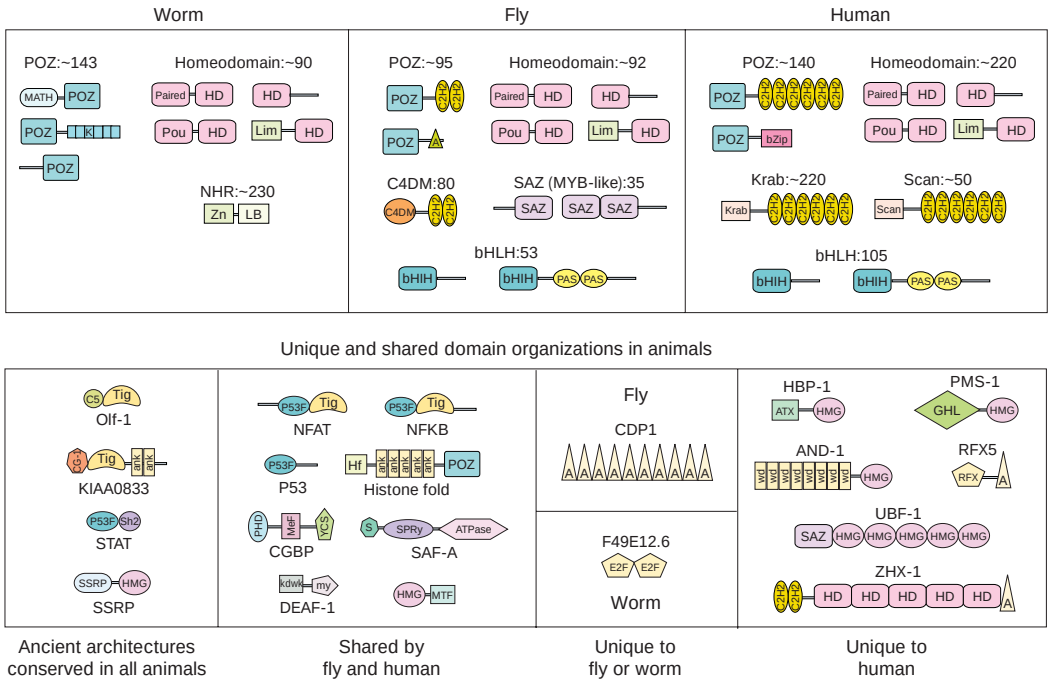


Figure 45 Lineage-specific expansions of domains and architectures of transcription factors. Top, specific families of transcription factors that have been expanded in each of the proteomes. Approximate numbers of domains identified in each of the (nearly) complete proteomes representing the lineages are shown next to the domains, and some of the most common architectures are shown. Some are shared by different animal lineages; others are lineage-specific. Bottom, samples of architectures from transcription factors that are shared by all animals (ancient architectures), shared by fly and human and unique to each lineage. Domains: K, kelch; HD, homeodomain; Zn, zinc-binding domain; LB, ligand-binding domain; C4DM, novel Zn cluster with four cysteines, probably involved in protein–protein interactions (L. Aravind, unpublished); MATH, meprin-associated TRAF domain; CG-1, novel domain in KIAA0909-like transcription factors (L. Aravind, unpublished); MTF, myelin transcription factor domain; SAZ, specialized Myb-like helix-turn-helix (HTH) domain found in Stonewall, ADF-1 and Zeste (L. Aravind, unpublished); A, AT-hook motif; E2F, winged HTH DNA-binding domain; GHL, gyraseB-histidine kinase-MutL ATPase domain; ATX, ATaXin domain; RFX, RFX winged HTH DNA binding domain; My, MYND domain; KDWK, KDWK DNA-binding domain; POZ, Pox zinc finger domain; S, SAP domain; P53F, P53 fold domain; HF, histone fold; ANK, ankyrin repeat; TIG, transcription factor Ig domain; SSRP, structure-specific recognition protein domain; C5, 5-cysteine metal binding domain; C2H2, classic zinc finger domain; WD, WD40 repeats.

© 2001 Macmillan Magazines Ltd

The most extreme mechanism is whole-genome duplication (WGD), through a polyploidization event in which a diploid organism becomes tetraploid. Such events are classified as autopolyploidy or allopolyploidy, depending on whether they involve hybridization between members of the same species or different species. Polyploidization is common in the plant kingdom, with many known examples among wild and domesticated crop species. Alfalfa (*Medicago sativa*) is a naturally occurring autotetraploid[364], and *Nicotiana tabacum*, some species of cotton (*Gossypium*) and several of the common brassicas are allotetraploids containing pairs of 'homeologous' chromosome pairs.

In principle, WGD provides the raw material for great bursts of innovation by allowing the duplication and divergence of entire pathways. Ohno[365] suggested that WGD has played a key role in evolution. There is evidence for an ancient WGD event in the ancestry of yeast and several independent such events in the ancestry of mustard weed[366–369]. Such ancient WGD events can be hard to detect because only a minority of the duplicated loci may be retained, with the result that the genes in duplicated segments cannot be aligned in a one-to-one correspondence but rather require many gaps. In addition, duplicated segments may be subsequently rearranged. For example, the ancient duplication in the yeast genome appears to have been followed by loss of more than 90% of the newly duplicated genes[366].

One of the most controversial hypotheses about vertebrate evolution is the proposal that two WGD events occurred early in the vertebrate lineage, around the time of jawed fishes some 500 Myr ago. Some authors[370–373] have seen support for this theory in the fact that many human genes occur in sets of four homologues—most notably the four extensive HOX gene clusters on chromosomes 2, 7, 12 and 17, whose duplication dates to around the correct time. However, other authors have disputed this interpretation[374], suggesting that these cases may reflect unrelated duplications of specific regions rather than successive WGD.

We analysed the draft genome sequence for evidence that might bear on this question. The analysis provides many interesting observations, but no convincing evidence of ancient WGD. We looked for evidence of pairs of chromosomal regions containing many homologous genes. Although we found many pairs containing a few homologous genes, the human genome does not appear to contain any pairs of regions where the density of duplicated genes approaches the densities seen in yeast or mustard weed[366–369].

We also examined human proteins in the IPI for which the orthologues among fly or worm proteins occur in the ratios 2:1:1, 3:1:1, 4:1:1 and so on (Fig. 49). The number of such families falls smoothly, with no peak at four and some instances of five or more homologues. Although this does not rule out two rounds of WGD followed by extensive gene loss and some unrelated gene duplication, it provides no support for the theory. More probatively, if two successive rounds of genome duplication occurred, phylogenetic analysis of the proteins having 4:1:1 ratios between human, fly and worm would be expected to show more trees with the topology (A,B)(C,D) for the human sequences than (A,(B,(C,D)))[375]. However, of 57 sets studied carefully, only 24% of the trees constructed from the 4:1:1 set have the former topology; this is not significantly different from what would be expected under the hypothesis of random sequential duplication of individual loci.

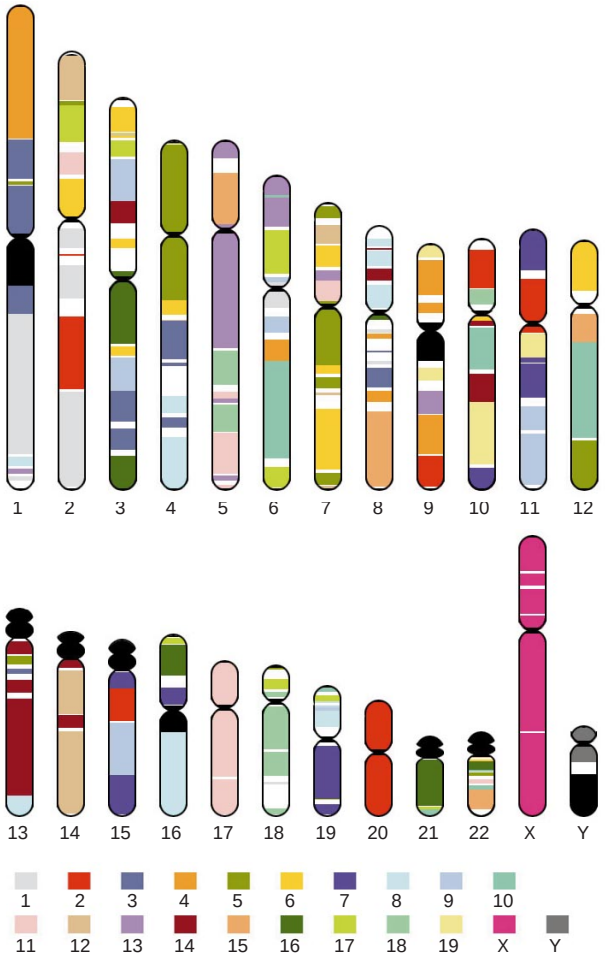


Figure 46 Conserved segments in the human and mouse genome. Human chromosomes, with segments containing at least two genes whose order is conserved in the mouse genome as colour blocks. Each colour corresponds to a particular mouse chromosome. Centromeres, subcentromeric heterochromatin of chromosomes 1, 9 and 16, and the repetitive short arms of 13, 14, 15, 21 and 22 are in black.

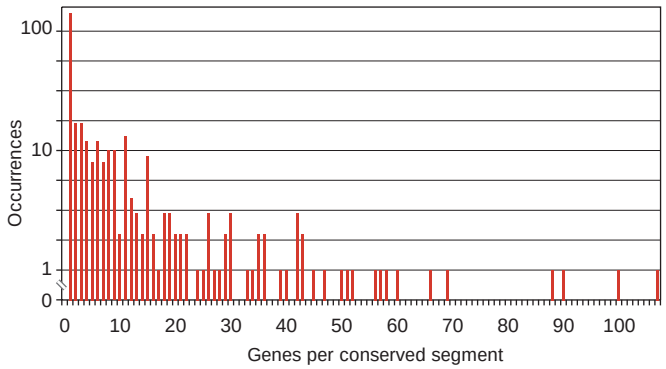


Figure 47 Distribution of number of genes per conserved segment between human and mouse genomes.

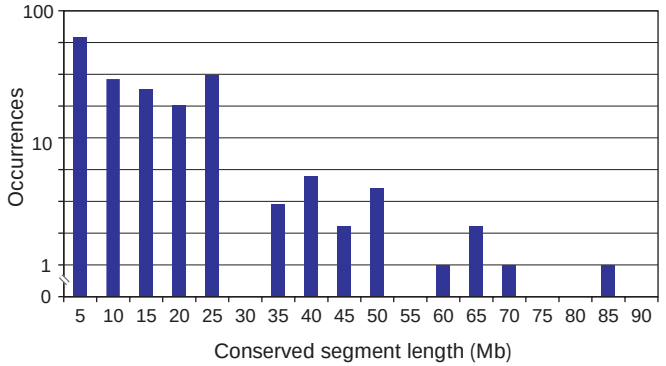


Figure 48 Distribution of lengths (in 5-Mb bins) of conserved segments between human and mouse genomes, omitting singletons.

© 2001 Macmillan Magazines Ltd

We also searched for sets of four chromosomes where there are multiple genes with homologues on each of the four. The strongest example was chromosomes 2, 7, 12 and 17, containing the HOX clusters as well as additional genes. These four chromosomes appear to have an excess of quadruplicated genes. The genes are not all clustered in a single region; this may reflect intrachromosomal rearrangement since the duplication of these genes, or it may indicate that they result from several independent events. Of the genes with homologues on chromosomes 2, 12 and 17, many of those missing on chromosome 7 are clustered on chromosome 3, suggesting a translocation. Several additional examples of groups of four chromosomes were found, although they were connected by fewer homologous genes.

Although the analyses are sensitive to the imperfect quality of the gene predictions, our results so far are insufficient to settle whether two rounds of WGD occurred around 500 Myr ago. It may be possible to resolve the issue by systematically estimating the time of each of the many gene duplication events on the basis of sequence divergence, although this is beyond the scope of this report. Another approach to determining whether a widespread duplication occurred at a particular time in vertebrate evolution would be to sequence the genomes of organisms whose lineages diverged from vertebrates at appropriate times, such as amphioxus.

Recent history from human polymorphism

The recent history of genomic segments can be probed by studying the properties of SNPs segregating in the current human population. The sequence information generated in the course of this project has yielded a huge collection of SNPs. These SNPs were extracted in two ways: by comparing overlapping large-insert clones derived from distinct haplotypes (either different individuals or different chromosomes within an individual) and by comparing random reads from whole-genome shotgun libraries derived from multiple individuals. The analysis confirms an average heterozygosity rate in the human population of about 1 in 1,300 bp (ref. 97).

More than 1.42 million SNPs have been assembled into a genome-wide map and are analysed in detail in an accompanying paper[97]. SNP density is also displayed across the genome in Fig. 9. The SNPs have an average spacing of 1.9 kb and 63% of 5-kb intervals contain a SNP. These polymorphisms are of immediate utility for medical genetic studies. Whereas investigators studying a gene previously had to expend considerable effort to discover polymorphisms across the region of interest, the current collection now provides then with about 15 SNPs for gene loci of average size.

The density of SNPs (adjusted for ascertainment—that is, polymorphisms per base screened) varies considerably across the genome[97] and sheds light on the unique properties and history of each genomic region. The average heterozygosity at a locus will tend to increase in proportion to the local mutation rate and the 'age' of the locus (which can be defined as the average number of generations since the most recent common ancestor of two randomly chosen copies in the population). For example, positive selection can cause a locus to be unusually 'young' and balancing selection can cause it to be unusually 'old'. An extreme example is the HLA region, in which a high SNP density is observed, reflecting the fact that diverse HLA haplotypes have been maintained for many millions of years by balancing selection and greatly predate the origin of the human species.

SNPs can also be used to study linkage disequilibrium in the human genome[376]. Linkage disequilibrium refers to the persistence of ancestral haplotypes—that is, genomic segments carrying particular combinations of alleles descended from a common ancestor. It can provide a powerful tool for mapping disease genes[377,378] and for probing population history[379–381]. There has been considerably controversy concerning the typical distance over which linkage disequilibrium extends in the human genome[382–387]. With the collection of SNPs now available, it should be possible to resolve this important issue.

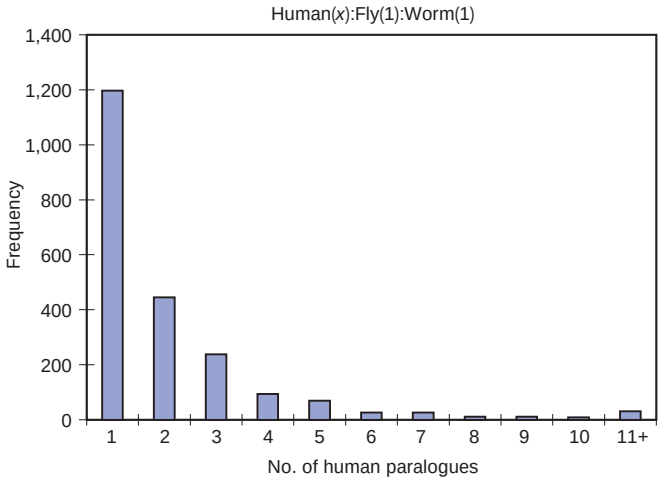


Figure 49 Number of human paralogues of genes having single orthologues in worm and fly.

Applications to medicine and biology

In most research papers, the authors can only speculate about future applications of the work. Because the genome sequence has been released on a daily basis over the past four years, however, we can already cite many direct applications. We focus on a handful of applications chosen primarily from medical research.

Disease genes

A key application of human genome research has been the ability to find disease genes of unknown biochemical function by positional cloning[388]. This method involves mapping the chromosomal region containing the gene by linkage analysis in affected families and then scouring the region to find the gene itself. Positional cloning is powerful, but it has also been extremely tedious. When the approach was first proposed in the early 1980s[9], a researcher wishing to perform positional cloning had to generate genetic markers to trace inheritance; perform chromosomal walking to obtain genomic DNA covering the region; and analyse a region of around 1 Mb by either direct sequencing or indirect gene identification methods. The first two barriers were eliminated with the development in the mid-1990s of comprehensive genetic and physical maps of the human chromosomes, under the auspices of the Human Genome Project. The remaining barrier, however, has continued to be formidable.

All that is changing with the availability of the human draft genome sequence. The human genomic sequence in public databases allows rapid identification *in silico* of candidate genes, followed by mutation screening of relevant candidates, aided by information on gene structure. For a mendelian disorder, a gene search can now often be carried out in a matter of months with only a modestly sized team.

At least 30 disease genes[55,389–422] (Table 26) have been positionally cloned in research efforts that depended directly on the publicly available genome sequence. As most of the human sequence has only arrived in the past twelve months, it is likely that many similar discoveries are not yet published. In addition, there are many cases in which the genome sequence played a supporting role, such as providing candidate microsatellite markers for finer genetic linkage analysis.

The genome sequence has also helped to reveal the mechanisms leading to some common chromosomal deletion syndromes. In several instances, recurrent deletions have been found to result from homologous recombination and unequal crossing over between large, nearly identical intrachromosomal duplications. Examples

© 2001 Macmillan Magazines Ltd

include the DiGeorge/velocardiofacial syndrome region on chromosome 22 (ref. 238) and the Williams–Beuren syndrome recurrent deletion on chromosome 7 (ref. 239).

The availability of the genome sequence also allows rapid identification of paralogues of disease genes, which is valuable for two reasons. First, mutations in a paralogous gene may give rise to a related genetic disease. A good example, discovered through use of the genome sequence, is achromatopsia (complete colour blindness). The *CNGA3* gene, encoding the α-subunit of the cone photoreceptor cyclic GMP-gated channel, had been shown to harbour mutations in some families with achromatopsia. Computational searching of the genome sequences revealed the paralogous gene encoding the corresponding β-subunit, *CNGB3* (which had not been apparent from EST databases). The *CNGB3* gene was rapidly shown to be the cause of achromatopsia in other families[407,408]. Another example is provided by the presenilin-1 and presenilin-2 genes, in which mutations can cause early-onset Alzheimer's disease[423,424]. Second, the paralogue may provide an opportunity for therapeutic intervention, as exemplified by attempts to reactivate the fetally expressed haemoglobin genes in individuals with sickle cell disease or β-thalassaemia, caused by mutations in the β-globin gene[425].

We undertook a systematic search for paralogues of 971 known human disease genes with entries in both the Online Mendelian Inheritance in Man (OMIM) database (http://www.ncbi.nlm.nih.gov/Omim/) and either the SwissProt or TrEMBL protein databases. We identified 286 potential paralogues (with the requirement of a match of at least 50 amino acids with identity greater than 70% but less than 90% if on the same chromosome, and less than 95% if on a different chromosome). Although this analysis may have identified some pseudogenes, 89% of the matches showed homology over more than one exon in the new target sequence, suggesting that many are functional. This analysis shows the potential for rapid identification of disease gene paralogues *in silico*.

Drug targets

Over the past century, the pharmaceutical industry has largely depended upon a limited set of drug targets to develop new therapies. A recent compendium[426,427] lists 483 drug targets as accounting for virtually all drugs on the market. Knowing the complete set of human genes and proteins will greatly expand the search for suitable drug targets. Although only a minority of human genes may be drug targets, it has been predicted that the number will exceed several thousand, and this prospect has led to a massive expansion of genomic research in pharmaceutical research and development. A few examples will illustrate the point.

(1) The neurotransmitter serotonin (5-HT) mediates rapid excitatory responses through ligand-gated channels. The previously identified 5-HT_{3A} receptor gene produces functional receptors, but with a much smaller conductance than observed *in vivo*. Cross-hybridization experiments and analysis of ESTs failed to reveal any other homologues of the known receptor. Recently, however, by searching the human draft genome sequence at low stringency, a putative homologue was identified within a PAC clone from the long arm of chromosome 11 (ref. 428). The homologue was shown to be expressed in the amygdala, caudate and hippocampus, and a full-length cDNA was subsequently obtained. The gene, which codes for a serotonin receptor, was named 5-HT_{3B}. When assembled in a heterodimer with 5-HT_{3A}, it was shown to account for the large-conductance neuronal serotonin channel. Given the central role of the serotonin pathway in mood disorders and schizophrenia, the discovery of a major new therapeutic target is of considerable interest.

(2) The contractile and inflammatory actions of the cysteinyl leukotrienes, formerly known as the slow reacting substance of anaphylaxis (SRS-A), are mediated through specific receptors. The second such receptor, $CysLT_2$, was identified using the combination of a rat EST and the human genome sequence. This led to the cloning of a gene with 38% amino-acid identity to the only other receptor that had previously been identified[429]. This new receptor, which shows high-affinity binding to several leukotrienes, maps to a region of chromosome 13 that is linked to atopic asthma. The gene is expressed in airway smooth muscles and in the heart. As the leukotriene pathway has been a significant target for the development of drugs against asthma, the discovery of a new receptor has obvious and important consequences.

(3) Abundant deposition of β-amyloid in senile plaques is the hallmark of Alzheimer's disease. β-Amyloid is generated by proteolytic processing of the amyloid precursor protein (APP). One of the enzymes involved is the β-site APP-cleaving enzyme (BACE), which is a transmembrane aspartyl protease. Computational searching of the public human draft genome sequence recently identified a new sequence homologous to BACE, encoding a protein now named BACE2[430,431]. BACE2, which has 52% amino-acid sequence identity to BACE, contains two active protease sites and maps to the obligatory Down's syndrome region of chromosome 21, as does APP. This raises the question of whether the extra copies of both BACE2 and APP may contribute to accelerated deposition of β-amyloid in the brains of Down's syndrome patients. The development of antagonists to BACE and BACE2 represents a promising approach to preventing Alzheimer's disease.

Given these examples, we undertook a systematic effort to identify paralogues of the classic drug target proteins in the draft genome sequence. The target list[427] was used to identify 603 entries in the SwissProt database with unique accession numbers. These were then searched against the current genome sequence database, using the requirement that a match should have 70–100% identity to at least 50 amino acids. Matches to named proteins were ignored, as we assumed that these represented known homologues.

We found 18 putative novel paralogues (Table 27), including apparent dopamine receptors, purinergic receptors and insulin-like growth factor receptors. In six cases, the novel paralogue matches at least one EST, adding confidence that this search process can identify novel functional genes. For the remaining 12 putative paralogues without an EST match, all have long ORFs and all but

Table 26 Disease genes positionally cloned using the draft genome sequence

Locus	Disorder	Reference(s)
BRCA2	Breast cancer susceptibility	55
AIRE	Autoimmune polyglandular syndrome type 1 (APS1 or APECED)	389
PEX1	Peroxisome biogenesis disorder	390, 391
PDS	Pendred syndrome	392
XLP	X-linked lymphoproliferative disease	393
DFNA5	Nonsyndromic deafness	394
ATP2A2	Darier's disease	395
SEDL	X-linked spondyloepiphyseal dysplasia tarda	396
WISP3	Progressive pseudorheumatoid dysplasia	397
CCM1	Cerebral cavernous malformations	398, 399
COL11A2/DFNA13	Nonsyndromic deafness	400
LGMD 2G	Limb-girdle muscular dystrophy	401
EVC	Ellis-Van Creveld syndrome, Weyer's acrodental dysostosis	402
ACTN4	Familial focal segmental glomerulosclerosis	403
SCN1A	Generalized epilepsy with febrile seizures plus type 2	404
AASS	Familial hyperlysinaemia	405
NDRG1	Hereditary motor and sensory neuropathy-Lom	406
CNGB3	Total colour-blindness	407, 408
MUL	Mulibrey nanism	409
USH1C	Usher type 1C	410, 411
MYH9	May-Hegglin anomaly	412, 413
PRKAR1A	Carney's complex	414
MYH9	Nonsyndromic hereditary deafness DFNA17	415
SCA10	Spinocerebellar ataxia type 10	416
OPA1	Optic atrophy	417
XLCSNB	X-linked congenital stationary night blindness	418
FGF23	Hypophosphataemic rickets	419
GAN	Giant axonal neuropathy	420
AAAS	Triple-A syndrome	421
HSPG2	Schwartz-Jampel syndrome	422

© 2001 Macmillan Magazines Ltd

one show similarity spanning multiple exons separated by introns, so these are not processed pseudogenes. They are likely to represent interesting new candidate drug targets.

Basic biology

Although the examples above reflect medical applications, there are also many similar applications to basic physiology and cell biology. To cite one satisfying example, the publicly available sequence was used to solve a mystery that had vexed investigators for several decades: the molecular basis of bitter taste[432]. Humans and other animals are polymorphic for response to certain bitter tastes. Recently, investigators mapped this trait in both humans and mice and then searched the relevant region of the human draft genome sequence for G-protein coupled receptors. These studies led, in quick succession, to the discovery of a new family of such proteins, the demonstration that they are expressed almost exclusively in taste buds, and the experimental confirmation that the receptors in cultured cells respond to specific bitter substances[433–435].

The next steps

Considerable progress has been made in human sequencing, but much remains to be done to produce a finished sequence. Even more work will be required to extract the full information contained in the sequence. Many of the key next steps are already underway.

Finishing the human sequence

The human sequence will serve as a foundation for biomedical research in the years ahead, and it is thus crucial that the remaining gaps be filled and ambiguities be resolved as quickly as possible. This will involve a three-step program.

The first stage involves producing finished sequence from clones spanning the current physical map, which covers more than 96% of the euchromatic regions of the genome. About 1 Gb of finished sequence is already completed. Almost all of the remaining clones are already sequenced to at least draft coverage, and the rest have been selected for sequencing. All clones are expected to reach 'full shotgun' coverage (8–10-fold redundancy) by about mid-2001 and finished form (99.99% accuracy) not long thereafter, using established and increasingly automated protocols.

The next stage will be to screen additional libraries to close gaps between clone contigs. Directed probing of additional large-insert clone libraries should close many of the remaining gaps. Unclosed gaps will be sized by FISH techniques or other methods. Two chromosomes, 22 and 21, have already been assembled in this 'essentially complete' form in this manner[93,94], and chromosomes 20, Y, 19, 14 and 7 are likely to reach this status in the next few months. All chromosomes should be essentially completed by 2003, if not sooner.

Finally, techniques must be developed to close recalcitrant gaps. Several hundred such gaps in the euchromatic sequence will probably remain in the genome after exhaustive screening of existing large-insert libraries. New methodologies will be needed to recover sequence from these segments, and to define biological reasons for their lack of representation in standard libraries. Ideally, it would be desirable to obtain complete sequence from all heterochromatic regions, such as centromeres and ribosomal gene clusters, although most of this sequence will consist of highly polymorphic tandem repeats containing few protein-coding genes.

Developing the IGI and IPI

The draft genome sequence has provided an initial look at the human gene content, but many ambiguities remain. A high priority will be to refine the IGI and IPI to the point where they accurately reflect every gene and every alternatively spliced form. Several steps are needed to reach this ambitious goal.

Finishing the human sequence will assist in this effort, but the experiences gained on chromosomes 21 and 22 show that sequence alone is not enough to allow complete gene identification. One powerful approach is cross-species sequence comparison with related organisms at suitable evolutionary distances. The sequence coverage from the pufferfish *T. nigroviridis* has already proven valuable in identifying potential exons[292]; this work is expected to continue from its current state of onefold coverage to reach at least fivefold coverage later this year. The genome sequence of the laboratory mouse will provide a particularly powerful tool for exon identification, as sequence similarity is expected to identify 95–97% of the exons, as well as a significant number of regulatory domains[436–438]. A public-private consortium is speeding this effort, by producing freely accessible whole-genome shotgun coverage that can be readily used for cross-species comparison[439]. More than onefold coverage from the C57BL/6J strain has already been completed and threefold is expected within the next few months. In the slightly longer term, a program is under way to produce a finished sequence of the laboratory mouse.

Another important step is to obtain a comprehensive collection of full-length human cDNAs, both as sequences and as actual clones. The Mammalian Gene Collection project has been underway for a year[18] and expects to produce 10,000–15,000 human full-length cDNAs over the coming year, which will be available without restrictions on use. The Genome Exploration Group of the RIKEN Genomic Sciences Center is similarly developing a collection of cDNA clones from mouse[309], which is a valuable complement

Table 27 New paralogues of common drug targets identified by searching the draft human genome sequence

Drug target	Drug target			Novel match IGI number	Chromosome containing paralogue	Per cent identity	dbEST GenBank accession
	HGM symbol	SwissProt accession	Chromosome				
Aquaporin 7	AQP7	O14520	9	IGI_M1_ctg15869_11	2	92.3	AW593324
Arachidonate 12-lipoxygenase	ALOX12	P18054	17	IGI_M1_ctg17216_23	17	70.1	
Calcitonin	CALCA	P01258	11	IGI_M1_ctg14138_20	12	93.6	
Calcium channel, voltage-dependent, γ-subunit	CACNG2	Q9Y698	22	IGI_M1_ctg17137_10	19	70.7	
DNA polymerase-δ, small subunit	POLD2	P49005	7	IGI_M1_ctg12903_29	5	86.8	
Dopamine receptor, D1-α	DRD1	P21728	5	IGI_M1_ctg25203_33	16	70.7	
Dopamine receptor, D1-β	DRD5	P21918	4	IGI_M1_ctg17190_14	1	88.0	AI148329
Eukaryotic translation elongation factor, 1δ	EEF1D	P29692	2	IGI_M1_ctg16401_37	17	77.6	BE719683
FKBP, tacrolimus binding protein, FK506 binding protein	FKBP1B	Q16645	2	IGI_M1_ctg14291_56	6	79.4	
Glutamic acid decarboxylase	GAD1	Q99259	2	IGI_M1_ctg12341_103	18	70.5	
Glycine receptor, α1	GLRA1	P23415	5	IGI_M1_ctg16547_14	X	85.5	
Heparan *N*-deacetylase/*N*-sulphotransferase	NDST1	P52848	5	IGI_M1_ctg13263_18	4	81.5	
Insulin-like growth factor-1 receptor	IGF1R	P08069	15	IGI_M1_ctg18444_3	19	71.8	
Na,K-ATPase, α-subunit	ATP1A1	P05023	1	IGI_M1_ctg14877_54	1	83	
Purinergic receptor 7 (P2X), ligand-gated ion channel	P2RX7	Q99572	12	IGI_M1_ctg15140_15	12	80.3	H84353
Tubulin, ε-chain	TUBE*	Q9UJT0	6	IGI_M1_ctg13826_4	5	78.5	AA970498
Tubulin, χ-chain	TUBG1	P23258	17	IGI_M1_ctg12599_5	7	84.0	
Voltage-gated potassium channel, KV3.3	KCNC3	Q14003	19	IGI_M1_ctg13492_5	12	80.1	H49142

* HGM symbol unknown.

© 2001 Macmillan Magazines Ltd

because of the availability of tissues from all developmental time points. A challenge will be to define the gene-specific patterns of alternative splicing, which may affect half of human genes. Existing collections of ESTs and cDNAs may allow identification of the most abundant of these isoforms, but systematic exploration of this problem may require exhaustive analysis of cDNA libraries from multiple tissues or perhaps high-throughput reverse transcription–PCR studies. Deep understanding of gene function will probably require knowledge of the structure, tissue distribution and abundance of these alternative forms.

Large-scale identification of regulatory regions

The one-dimensional script of the human genome, shared by essentially all cells in all tissues, contains sufficient information to provide for differentiation of hundreds of different cell types, and the ability to respond to a vast array of internal and external influences. Much of this plasticity results from the carefully orchestrated symphony of transcriptional regulation. Although much has been learned about the *cis*-acting regulatory motifs of some specific genes, the regulatory signals for most genes remain uncharacterized. Comparative genomics of multiple vertebrates offers the best hope for large-scale identification of such regulatory sites[440]. Previous studies of sequence alignment of regulatory domains of orthologous genes in multiple species has shown a remarkable correlation between sequence conservation, dubbed 'phylogenetic footprints'[441], and the presence of binding motifs for transcription factors. This approach could be particularly powerful if combined with expression array technologies that identify cohorts of genes that are coordinately regulated, implicating a common set of *cis*-acting regulatory sequences[442–445]. It will also be of considerable interest to study epigenetic modifications such as cytosine methylation on a genome-wide scale, and to determine their biological consequences[446,447]. Towards this end, a pilot Human Epigenome Project has been launched[448,449].

Sequencing of additional large genomes

More generally, comparative genomics allows biologists to peruse evolution's laboratory notebook—to identify conserved functional features and recognize new innovations in specific lineages. Determination of the genome sequence of many organisms is very desirable. Already, projects are underway to sequence the genomes of the mouse, rat, zebrafish and the pufferfishes *T. nigroviridis* and *Takifugu rubripes*. Plans are also under consideration for sequencing additional primates and other organisms that will help define key developments along the vertebrate and nonvertebrate lineages.

To realize the full promise of comparative genomics, however, it needs to become simple and inexpensive to sequence the genome of any organism. Sequencing costs have dropped 100-fold over the last 10 years, corresponding to a roughly twofold decrease every 18 months. This rate is similar to 'Moore's law' concerning improvements in semiconductor manufacture. In both sequencing and semiconductors, such improvement does not happen automatically, but requires aggressive technological innovation fuelled by major investment. Improvements are needed to move current dideoxy sequencing to smaller volumes and more rapid sequencing times, based upon advances such as microchannel technology. More revolutionary methods, such as mass spectrometry, single-molecule sequencing and nanopore approaches[76], have not yet been fully developed, but hold great promise and deserve strong encouragement.

Completing the catalogue of human variation

The human draft genome sequence has already allowed the identification of more than 1.4 million SNPs, comprising a substantial proportion of all common human variation. This program should be extended to obtain a nearly complete catalogue of common variants and to identify the common ancestral haplotypes present in the population. In principle, these genetic tools should make it possible to perform association studies and linkage disequilibrium studies[376] to identify the genes that confer even relatively modest risk for common diseases. Launching such an intense era of human molecular epidemiology will also require major advances in the cost efficiency of genotyping technology, in the collection of carefully phenotyped patient cohorts and in statistical methods for relating large-scale SNP data to disease phenotype.

From sequence to function

The scientific program outlined above focuses on how the genome sequence can be mined for biological information. In addition, the sequence will serve as a foundation for a broad range of functional genomic tools to help biologists to probe function in a more systematic manner. These will need to include improved techniques and databases for the global analysis of: RNA and protein expression, protein localization, protein–protein interactions and chemical inhibition of pathways. New computational techniques will be needed to use such information to model cellular circuitry. A full discussion of these important directions is beyond the scope of this paper.

Concluding thoughts

The Human Genome Project is but the latest increment in a remarkable scientific program whose origins stretch back a hundred years to the rediscovery of Mendel's laws and whose end is nowhere in sight. In a sense, it provides a capstone for efforts in the past century to discover genetic information and a foundation for efforts in the coming century to understand it.

We find it humbling to gaze upon the human sequence now coming into focus. In principle, the string of genetic bits holds long-sought secrets of human development, physiology and medicine. In practice, our ability to transform such information into understanding remains woefully inadequate. This paper simply records some initial observations and attempts to frame issues for future study. Fulfilling the true promise of the Human Genome Project will be the work of tens of thousands of scientists around the world, in both academia and industry. It is for this reason that our highest priority has been to ensure that genome data are available rapidly, freely and without restriction.

The scientific work will have profound long-term consequences for medicine, leading to the elucidation of the underlying molecular mechanisms of disease and thereby facilitating the design in many cases of rational diagnostics and therapeutics targeted at those mechanisms. But the science is only part of the challenge. We must also involve society at large in the work ahead. We must set realistic expectations that the most important benefits will not be reaped overnight. Moreover, understanding and wisdom will be required to ensure that these benefits are implemented broadly and equitably. To that end, serious attention must be paid to the many ethical, legal and social implications (ELSI) raised by the accelerated pace of genetic discovery. This paper has focused on the scientific achievements of the human genome sequencing efforts. This is not the place to engage in a lengthy discussion of the ELSI issues, which have also been a major research focus of the Human Genome Project, but these issues are of comparable importance and could appropriately fill a paper of equal length.

Finally, it is has not escaped our notice that the more we learn about the human genome, the more there is to explore.

"We shall not cease from exploration. And the end of all our exploring will be to arrive where we started, and know the place for the first time."—T. S. Eliot[450] □

Received 7 December 2000; accepted 9 January 2001.

1. Correns, C. Untersuchungen über die Xenien bei Zea mays. *Berichte der Deutsche Botanische Gesellschaft* **17,** 410–418 (1899).
2. De Vries, H. Sur la loie de disjonction des hybrides. *Comptes Rendue Hebdemodaires, Acad. Sci. Paris* **130,** 845–847 (1900).
3. von Tschermack, E. Uber Künstliche Kreuzung bei Pisum sativum. *Berichte der Deutsche Botanische Gesellschaft* **18,** 232–239. (1900).
4. Sanger, F. *et al.* Nucleotide sequence of bacteriophage Φ X174 DNA. *Nature* **265,** 687–695 (1977).
5. Sanger, F. *et al.* The nucleotide sequence of bacteriophage ΦX174. *J Mol Biol* **125,** 225–246 (1978).

© 2001 Macmillan Magazines Ltd

6. Sanger, F., Coulson, A. R., Hong, G. F., Hill, D. F. & Petersen, G. B. Nucleotide-sequence of bacteriophage Lambda DNA. *J. Mol. Biol.* **162,** 729–773 (1982).
7. Fiers, W. *et al.* Complete nucleotide sequence of SV40 DNA. *Nature* **273,** 113–120 (1978).
8. Anderson, S. *et al.* Sequence and organization of the human mitochondrial genome. *Nature* **290,** 457–465 (1981).
9. Botstein, D., White, R. L., Skolnick, M. & Davis, R. W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* **32,** 314–331 (1980).
10. Olson, M. V. *et al.* Random-clone strategy for genomic restriction mapping in yeast. *Proc. Natl Acad. Sci. USA* **83,** 7826–7830 (1986).
11. Coulson, A., Sulston, J., Brenner, S. & Karn, J. Toward a physical map of the genome of the nematode *Caenorhabditis elegans*. *Proc. Natl Acad. Sci. USA* **83,** 7821–7825 (1986).
12. Putney, S. D., Herlihy, W. C. & Schimmel, P. A new troponin T and cDNA clones for 13 different muscle proteins, found by shotgun sequencing. *Nature* **302,** 718–721 (1983).
13. Milner, R. J. & Sutcliffe, J. G. Gene expression in rat brain. *Nucleic Acids Res.* **11,** 5497–5520 (1983).
14. Adams, M. D. *et al.* Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* **252,** 1651–1656 (1991).
15. Adams, M. D. *et al.* Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* **377,** 3–174 (1995).
16. Okubo, K. *et al.* Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. *Nature Genet.* **2,** 173–179 (1992).
17. Hillier, L. D. *et al.* Generation and analysis of 280,000 human expressed sequence tags. *Genome Res.* **6,** 807–828 (1996).
18. Strausberg, R. L., Feingold, E. A., Klausner, R. D. & Collins, F. S. The mammalian gene collection. *Science* **286,** 455–457 (1999).
19. Berry, R. *et al.* Gene-based sequence-tagged-sites (STSs) as the basis for a human gene map. *Nature Genet.* **10,** 415–423 (1995).
20. Houlgatte, R. *et al.* The Genexpress Index: a resource for gene discovery and the genic map of the human genome. *Genome Res.* **5,** 272–304 (1995).
21. Sinsheimer, R. L. The Santa Cruz Workshop—May 1985. *Genomics* **5,** 954–956 (1989).
22. Palca, J. Human genome—Department of Energy on the map. *Nature* **321,** 371 (1986).
23. National Research Council *Mapping and Sequencing the Human Genome* (National Academy Press, Washington DC, 1988).
24. Bishop, J. E. & Waldholz, M. *Genome* (Simon and Schuster, New York, 1990).
25. Kevles, D. J. & Hood, L. (eds) *The Code of Codes: Scientific and Social Issues in the Human Genome Project* (Harvard Univ. Press, Cambridge, Massachusetts, 1992).
26. Cook-Deegan, R. *The Gene Wars: Science, Politics, and the Human Genome* (W. W. Norton & Co., New York, London, 1994).
27. Donis-Keller, H. *et al.* A genetic linkage map of the human genome. *Cell* **51,** 319–337 (1987).
28. Gyapay, G. *et al.* The 1993–94 Genethon human genetic linkage map. *Nature Genet.* **7,** 246–339 (1994).
29. Hudson, T. J. *et al.* An STS-based map of the human genome. *Science* **270,** 1945–1954 (1995).
30. Dietrich, W. F. *et al.* A comprehensive genetic map of the mouse genome. *Nature* **380,** 149–152 (1996).
31. Nusbaum, C. *et al.* A YAC-based physical map of the mouse genome. *Nature Genet.* **22,** 388–393 (1999).
32. Oliver, S. G. *et al.* The complete DNA sequence of yeast chromosome III. *Nature* **357,** 38–46 (1992).
33. Wilson, R. *et al.* 2.2 Mb of contiguous nucleotide sequence from chromosome III of *C. elegans*. *Nature* **368,** 32–38 (1994).
34. Chen, E. Y. *et al.* The human growth hormone locus: nucleotide sequence, biology, and evolution. *Genomics* **4,** 479–497 (1989).
35. McCombie, W. R. *et al.* Expressed genes, Alu repeats and polymorphisms in cosmids sequenced from chromosome 4p16.3. *Nature Genet.* **1,** 348–353 (1992).
36. Martin-Gallardo, A. *et al.* Automated DNA sequencing and analysis of 106 kilobases from human chromosome 19q13.3. *Nature Genet.* **1,** 34–39 (1992).
37. Edwards, A. *et al.* Automated DNA sequencing of the human HPRT locus. *Genomics* **6,** 593–608 (1990).
38. Marshall, E. A strategy for sequencing the genome 5 years early. *Science* **267,** 783–784 (1995).
39. Project to sequence human genome moves on to the starting blocks. *Nature* **375,** 93–94 (1995).
40. Shizuya, H. *et al.* Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. *Proc. Natl Acad. Sci. USA* **89,** 8794–8797 (1992).
41. Burke, D. T., Carle, G. F. & Olson, M. V. Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. *Science* **236,** 806–812 (1987).
42. Marshall, E. A second private genome project. *Science* **281,** 1121 (1998).
43. Marshall, E. NIH to produce a 'working draft' of the genome by 2001. *Science* **281,** 1774–1775 (1998).
44. Pennisi, E. Academic sequencers challenge Celera in a sprint to the finish. *Science* **283,** 1822–1823 (1999).
45. Bouck, J., Miller, W., Gorrell, J. H., Muzny, D. & Gibbs, R. A. Analysis of the quality and utility of random shotgun sequencing at low redundancies. *Genome Res.* **8,** 1074–1084 (1998).
46. Collins, F. S. *et al.* New goals for the U. S. Human Genome Project: 1998–2003. *Science* **282,** 682–689 (1998).
47. Sanger, F. & Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* **94,** 441–448 (1975).
48. Maxam, A. M. & Gilbert, W. A new method for sequencing DNA. *Proc. Natl Acad. Sci. USA* **74,** 560–564 (1977).
49. Anderson, S. Shotgun DNA sequencing using cloned DNase I-generated fragments. *Nucleic Acids Res.* **9,** 3015–3027 (1981).
50. Gardner, R. C. *et al.* The complete nucleotide sequence of an infectious clone of cauliflower mosaic virus by M13mp7 shotgun sequencing. *Nucleic Acids Res.* **9,** 2871–2888 (1981).
51. Deininger, P. L. Random subcloning of sonicated DNA: application to shotgun DNA sequence analysis. *Anal. Biochem.* **129,** 216–223 (1983).
52. Chissoe, S. L. *et al.* Sequence and analysis of the human ABL gene, the BCR gene, and regions involved in the Philadelphia chromosomal translocation. *Genomics* **27,** 67–82 (1995).
53. Rowen, L., Koop, B. F. & Hood, L. The complete 685-kilobase DNA sequence of the human beta T cell receptor locus. *Science* **272,** 1755–1762 (1996).
54. Koop, B. F. *et al.* Organization, structure, and function of 95 kb of DNA spanning the murine T-cell receptor C alpha/C delta region. *Genomics* **13,** 1209–1230 (1992).
55. Wooster, R. *et al.* Identification of the breast cancer susceptibility gene BRCA2. *Nature* **378,** 789–792 (1995).
56. Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269,** 496–512 (1995).
57. Lander, E. S. & Waterman, M. S. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2,** 231–239 (1988).
58. Weber, J. L. & Myers, E. W. Human whole-genome shotgun sequencing. *Genome Res.* **7,** 401–409 (1997).
59. Green, P. Against a whole-genome shotgun. *Genome Res.* **7,** 410–417 (1997).
60. Venter, J. C. *et al.* Shotgun sequencing of the human genome. *Science* **280,** 1540–1542 (1998).
61. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291,** 1304–1351 (2001).
62. Smith, L. M. *et al.* Fluorescence detection in automated DNA sequence analysis. *Nature* **321,** 674–679 (1986).
63. Ju, J. Y., Ruan, C. C., Fuller, C. W., Glazer, A. N. & Mathies, R. A. Fluorescence energy-transfer dye-labeled primers for DNA sequencing and analysis. *Proc. Natl Acad. Sci. USA* **92,** 4347–4351 (1995).
64. Lee, L. G. *et al.* New energy transfer dyes for DNA sequencing. *Nucleic Acids Res.* **25,** 2816–2822 (1997).
65. Rosenblum, B. B. *et al.* New dye-labeled terminators for improved DNA sequencing patterns. *Nucleic Acids Res.* **25,** 4500–4504 (1997).
66. Metzker, M. L., Lu, J. & Gibbs, R. A. Electrophoretically uniform fluorescent dyes for automated DNA sequencing. *Science* **271,** 1420–1422 (1996).
67. Prober, J. M. *et al.* A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science* **238,** 336–341 (1987).
68. Reeve, M. A. & Fuller, C. W. A novel thermostable polymerase for DNA sequencing. *Nature* **376,** 796–797 (1995).
69. Tabor, S. & Richardson, C. C. Selective inactivation of the exonuclease activity of bacteriophage T7 DNA polymerase by in vitro mutagenesis. *J. Biol. Chem.* **264,** 6447–6458 (1989).
70. Tabor, S. & Richardson, C. C. DNA sequence analysis with a modified bacteriophage T7 DNA polymerase—effect of pyrophosphorolysis and metal ions. *J. Biol. Chem.* **265,** 8322–8328 (1990).
71. Murray, V. Improved double-stranded DNA sequencing using the linear polymerase chain reaction. *Nucleic Acids Res.* **17,** 8889 (1989).
72. Guttman, A., Cohen, A. S., Heiger, D. N. & Karger, B. L. Analytical and micropreparative ultrahigh resolution of oligonucleotides by polyacrylamide-gel high-performance capillary electrophoresis. *Anal. Chem.* **62,** 137–141 (1990).
73. Luckey, J. A. *et al.* High-speed DNA sequencing by capillary electrophoresis. *Nucleic Acids Res.* **18,** 4417–4421 (1990).
74. Swerdlow, H., Wu, S., Harke, H. & Dovichi, N. J. Capillary gel-electrophoresis for DNA sequencing—laser-induced fluorescence detection with the sheath flow cuvette. *J. Chromatogr.* **516,** 61–67 (1990).
75. Meldrum, D. Automation for genomics, part one: preparation for sequencing. *Genome Res.* **10,** 1081–1092 (2000).
76. Meldrum, D. Automation for genomics, part two: sequencers, microarrays, and future trends. *Genome Res.* **10,** 1288–1303 (2000).
77. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8,** 186–194 (1998).
78. Ewing, B., Hillier, L., Wendl, M. C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8,** 175–185 (1998).
79. Bentley, D. R. Genomic sequence information should be released immediately and freely in the public domain. *Science* **274,** 533–534 (1996).
80. Guyer, M. Statement on the rapid release of genomic DNA sequence. *Genome Res.* **8,** 413 (1998).
81. Dietrich, W. *et al.* A genetic map of the mouse suitable for typing intraspecific crosses. *Genetics* **131,** 423–447 (1992).
82. Kim, U. J. *et al.* Construction and characterization of a human bacterial artificial chromosome library. *Genomics* **34,** 213–218 (1996).
83. Osoegawa, K. *et al.* Bacterial artificial chromosome libraries for mouse sequencing and functional analysis. *Genome Res.* **10,** 116–128 (2000).
84. Marra, M. A. *et al.* High throughput fingerprint analysis of large-insert clones. *Genome Res.* **7,** 1072–1084 (1997).
85. Marra, M. *et al.* A map for sequence analysis of the *Arabidopsis thaliana* genome. *Nature Genet.* **22,** 265–270 (1999).
86. The International Human Genome Mapping Consortium. A physical map of the human genome. *Nature* **409,** 934–941 (2001).
87. Zhao, S. *et al.* Human BAC ends quality assessment and sequence analyses. *Genomics* **63,** 321–332 (2000).
88. Mahairas, G. G. *et al.* Sequence-tagged connectors: A sequence approach to mapping and scanning the human genome. *Proc. Natl Acad. Sci. USA* **96,** 9739–9744 (1999).
89. Tilford, C. A. *et al.* A physical map of the human Y chromosome. *Nature* **409,** 943–945 (2001).
90. Bentley, D. R. *et al.* The physical maps for sequencing human chromosomes 1, 6, 9, 10, 13, 20 and X. *Nature* **409,** 942–943 (2001).
91. Montgomery, K. T. *et al.* A high-resolution map of human chromosome 12. *Nature* **409,** 945–946 (2001).
92. Brüls, T. *et al.* A physical map of human chromosome 14. *Nature* **409,** 947–948 (2001).
93. Hattori, M. *et al.* The DNA sequence of human chromosome 21. *Nature* **405,** 311–319 (2000).
94. Dunham, I. *et al.* The DNA sequence of human chromosome 22. *Nature* **402,** 489–495 (1999).
95. Cox, D. *et al.* Radiation hybrid map of the human genome. *Science* (in the press).
96. Osoegawa, K. *et al.* An improved approach for construction of bacterial artificial chromosome libraries. *Genomics* **52,** 1–8 (1998).
97. The International SNP Map Working Group. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409,** 928–933 (2001).
98. Collins, F. S., Brooks, L. D. & Chakravarti, A. A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res.* **8,** 1229–1231 (1998).
99. Stewart, E. A. *et al.* An STS-based radiation hybrid map of the human genome. *Genome Res.* **7,** 422–433 (1997).
100. Deloukas, P. *et al.* A physical map of 30,000 human genes. *Science* **282,** 744–746 (1998).
101. Dib, C. *et al.* A comprehensive genetic map of the human genome based on 5,264 microsatellites.

© 2001 Macmillan Magazines Ltd

Nature **380,** 152–154 (1996).
102. Broman, K. W., Murray, J. C., Sheffield, V. C., White, R. L. & Weber, J. L. Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am. J. Hum. Genet.* **63,** 861–869 (1998).
103. The BAC Resource Consortium. Integration of cytogenetic landmarks into the draft sequence of the human genome. *Nature* **409,** 953–958 (2001).
104. Kent, W. J. & Haussler, D. GigAssembler: an algorithm for the initial assembly of the human working draft . Technical Report UCSC-CRL-00-17 (Univ. California at Santa Cruz, Santa Cruz, California, 2001).
105. Morton, N. E. Parameters of the human genome. *Proc. Natl Acad. Sci. USA* **88,** 7474–7476 (1991).
106. Podugolnikova, O. A. & Blumina, M. G. Heterochromatic regions on chromosomes 1, 9, 16, and Y in children with some disturbances occurring during embryo development. *Hum. Genet.* **63,** 183–188 (1983).
107. Lundgren, R., Berger, R. & Kristoffersson, U. Constitutive heterochromatin C-band polymorphism in prostatic cancer. *Cancer Genet. Cytogenet.* **51,** 57–62 (1991).
108. Lee, C., Wevrick, R., Fisher, R. B., Ferguson-Smith, M. A. & Lin, C. C. Human centromeric DNAs. *Hum. Genet.* **100,** 291–304 (1997).
109. Riethman, H. C. *et al.* Integration of telomere sequences with the draft human genome sequence. *Nature* **409,** 953–958 (2001).
110. Pruit, K. D. & Maglott, D. R. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.* **29,** 137–140 (2001).
111. Wolfsberg, T. G., McEntyre, J. & Schuler, G. D. Guide to the draft human genome. *Nature* **409,** 824–826 (2001).
112. Hurst, L. D. & Eyre-Walker, A. Evolutionary genomics: reading the bands. *Bioessays* **22,** 105–107 (2000).
113. Saccone, S. *et al.* Correlations between isochores and chromosomal bands in the human genome. *Proc. Natl Acad. Sci. USA* **90,** 11929–11933 (1993).
114. Zoubak, S., Clay, O. & Bernardi, G. The gene distribution of the human genome. *Gene* **174,** 95–102 (1996).
115. Gardiner, K. Base composition and gene distribution: critical patterns in mammalian genome organization. *Trends Genet.* **12,** 519–524 (1996).
116. Duret, L., Mouchiroud, D. & Gautier, C. Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. *J. Mol. Evol.* **40,** 308–317 (1995).
117. Saccone, S., De Sario, A., Della Valle, G. & Bernardi, G. The highest gene concentrations in the human genome are in telomeric bands of metaphase chromosomes. *Proc. Natl Acad. Sci. USA* **89,** 4913–4917 (1992).
118. Bernardi, G. *et al.* The mosaic genome of warm-blooded vertebrates. *Science* **228,** 953–958 (1985).
119. Bernardi, G. Isochores and the evolutionary genomics of vertebrates. *Gene* **241,** 3–17 (2000).
120. Fickett, J. W., Torney, D. C. & Wolf, D. R. Base compositional structure of genomes. *Genomics* **13,** 1056–1064 (1992).
121. Churchill, G. A. Stochastic models for heterogeneous DNA sequences. *Bull. Math. Biol.* **51,** 79–94 (1989).
122. Bird, A., Taggart, M., Frommer, M., Miller, O. J. & Macleod, D. A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. *Cell* **40,** 91–99 (1985).
123. Bird, A. P. CpG islands as gene markers in the vertebrate nucleus. *Trends Genet.* **3,** 342–347 (1987).
124. Chan, M. F., Liang, G. & Jones, P. A. Relationship between transcription and DNA methylation. *Curr. Top. Microbiol. Immunol.* **249,** 75–86 (2000).
125. Holliday, R. & Pugh, J. E. DNA modification mechanisms and gene activity during development. *Science* **187,** 226–232 (1975).
126. Larsen, F., Gundersen, G., Lopez, R. & Prydz, H. CpG islands as gene markers in the human genome. *Genomics* **13,** 1095–1107 (1992).
127. Tazi, J. & Bird, A. Alternative chromatin structure at CpG islands. *Cell* **60,** 909–920 (1990).
128. Gardiner-Garden, M. & Frommer, M. CpG islands in vertebrate genomes. *J. Mol. Biol.* **196,** 261–282 (1987).
129. Antequera, F. & Bird, A. Number of CpG islands and genes in human and mouse. *Proc. Natl Acad. Sci. USA* **90,** 11995–11999 (1993).
130. Ewing, B. & Green, P. Analysis of expressed sequence tags indicates 35,000 human genes. *Nature Genet.* **25,** 232–234 (2000).
131. Yu, A. Comparison of human genetic and sequence-based physical maps. *Nature* **409,** 951–953 (2001).
132. Kaback, D. B., Guacci, V., Barber, D. & Mahon, J. W. Chromosome size-dependent control of meiotic recombination. *Science* **256,** 228–232 (1992).
133. Riles, L. *et al.* Physical maps of the 6 smallest chromosomes of *Saccharomyces cerevisiae* at a resolution of 2.6-kilobase pairs. *Genetics* **134,** 81–150 (1993).
134. Lynn, A. *et al.* Patterns of meiotic recombination on the long arm of human chromosome 21. *Genome Res.* **10,** 1319–1332 (2000).
135. Laurie, D. A. & Hulten, M. A. Further studies on bivalent chiasma frequency in human males with normal karyotypes. *Ann. Hum. Genet.* **49,** 189–201 (1985).
136. Roeder, G. S. Meiotic chromosomes: it takes two to tango. *Genes Dev.* **11,** 2600–2621 (1997).
137. Wu, T.-C. & Lichten, M. Meiosis-induced double-strand break sites determined by yeast chromatin structure. *Science* **263,** 515–518 (1994).
138. Gerton, J. L. *et al.* Global mapping of meiotic recombination hotspots and coldspots in the yeast *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA* **97,** 11383–11390 (2000).
139. Li, W. -H. *Molecular Evolution* (Sinauer, Sunderland, Massachusetts, 1997).
140. Gregory, T. R. & Hebert, P. D. The modulation of DNA content: proximate causes and ultimate consequences. *Genome Res.* **9,** 317–324 (1999).
141. Hartl, D. L. Molecular melodies in high and low C. *Nature Rev. Genet.* **1,** 145–149 (2000).
142. Smit, A. F. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.* **9,** 657–663 (1999).
143. Prak, E. L. & Haig, H. K. Jr Mobile elements and the human genome. *Nature Rev. Genet.* **1,** 134–144 (2000).
144. Okada, N., Hamada, M., Ogiwara, I. & Ohshima, K. SINEs and LINEs share common 3′ sequences: a review. *Gene* **205,** 229–243 (1997).
145. Esnault, C., Maestre, J. & Heidmann, T. Human LINE retrotransposons generate processed pseudogenes. *Nature Genet.* **24,** 363–367 (2000).
146. Wei, W. *et al.* Human L1 retrotransposition: cis-preference vs. trans-complementation. *Mol. Cell. Biol.* **21,** 1429–1439 (2001)
147. Malik, H. S., Henikoff, S. & Eickbush, T. H. Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. *Genome Res.* **10,** 1307–1318 (2000).
148. Smit, A. F. The origin of interspersed repeats in the human genome. *Curr. Opin. Genet. Dev.* **6,** 743–748 (1996).
149. Clark, J. B. & Tidwell, M. G. A phylogenetic perspective on P transposable element evolution in Drosophila. *Proc. Natl Acad. Sci. USA* **94,** 11428–11433 (1997).
150. Haring, E., Hagemann, S. & Pinsker, W. Ancient and recent horizontal invasions of Drosophilids by P elements. *J. Mol. Evol.* **51,** 577–586 (2000).
151. Koga, A. *et al.* Evidence for recent invasion of the medaka fish genome by the Tol2 transposable element. *Genetics* **155,** 273–281 (2000).
152. Robertson, H. M. & Lampe, D. J. Recent horizontal transfer of a mariner transposable element among and between Diptera and Neuroptera. *Mol. Biol. Evol.* **12,** 850–862 (1995).
153. Simmons, G. M. Horizontal transfer of hobo transposable elements within the *Drosophila melanogaster* species complex: evidence from DNA sequencing. *Mol. Biol. Evol.* **9,** 1050–1060 (1992).
154. Malik, H. S., Burke, W. D. & Eickbush, T. H. The age and evolution of non-LTR retrotransposable elements. *Mol. Biol. Evol.* **16,** 793–805 (1999).
155. Kordis, D. & Gubensek, F. Bov-B long interspersed repeated DNA (LINE) sequences are present in *Vipera ammodytes* phospholipase A2 genes and in genomes of Viperidae snakes. *Eur. J. Biochem.* **246,** 772–779 (1997).
156. Jurka, J. Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet.* **16,** 418–420 (2000).
157. Sarich, V. M. & Wilson, A. C. Generation time and genome evolution in primates. *Science* **179,** 1144–1147 (1973).
158. Smit, A. F., Toth, G., Riggs, A. D., & Jurka, J. Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. *J. Mol. Biol.* **246,** 401–417 (1995).
159. Lim, J. K. & Simmons, M. J. Gross chromosome rearrangements mediated by transposable elements in *Drosophila melanogaster*. *Bioessays* **16,** 269–275 (1994).
160. Caceres, M., Ranz, J. M., Barbadilla, A., Long, M. & Ruiz, A. Generation of a widespread Drosophila inversion by a transposable element. *Science* **285,** 415–418 (1999).
161. Gray, Y. H. It takes two transposons to tango: transposable-element-mediated chromosomal rearrangements. *Trends Genet.* **16,** 461–468 (2000).
162. Zhang, J. & Peterson, T. Genome rearrangements by nonlinear transposons in maize. *Genetics* **153,** 1403–1410 (1999).
163. Smit, A. F. Identification of a new, abundant superfamily of mammalian LTR-transposons. *Nucleic Acids Res.* **21,** 1863–1872 (1993).
164. Cordonnier, A., Casella, J. F. & Heidmann, T. Isolation of novel human endogenous retrovirus-like elements with foamy virus-related pol sequence. *J. Virol.* **69,** 5890–5897 (1995).
165. Medstrand, P. & Mager, D. L. Human-specific integrations of the HERV-K endogenous retrovirus family. *J. Virol.* **72,** 9782–9787 (1998).
166. Myers, E. W. *et al.* A whole-genome assembly of Drosophila. *Science* **287,** 2196–2204 (2000).
167. Petrov, D. A., Lozovskaya, E. R. & Hartl, D. L. High intrinsic rate of DNA loss in *Drosophila*. *Nature* **384,** 346–349 (1996).
168. Li, W. H., Ellsworth, D. L., Krushkal, J., Chang, B. H. & Hewett-Emmett, D. Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis. *Mol. Phylogenet. Evol.* **5,** 182–187 (1996).
169. Goodman, M. *et al.* Toward a phylogenetic classification of primates based on DNA evidence complemented by fossil evidence. *Mol. Phylogenet. Evol.* **9,** 585–598 (1998).
170. Kazazian, H. H. Jr & Moran, J. V. The impact of L1 retrotransposons on the human genome. *Nature Genet.* **19,** 19–24 (1998).
171. Malik, H. S. & Eickbush, T. H. NeSL-1, an ancient lineage of site-specific non-LTR retrotransposons from *Caenorhabditis elegans*. *Genetics* **154,** 193–203 (2000).
172. Casavant, N. C. *et al.* The end of the LINE?: lack of recent L1 activity in a group of South American rodents. *Genetics* **154,** 1809–1817 (2000).
173. Meunier-Rotival, M., Soriano, P., Cuny, G., Strauss, F. & Bernardi, G. Sequence organization and genomic distribution of the major family of interspersed repeats of mouse DNA. *Proc. Natl Acad. Sci. USA* **79,** 355–359 (1982).
174. Soriano, P., Meunier-Rotival, M. & Bernardi, G. The distribution of interspersed repeats is nonuniform and conserved in the mouse and human genomes. *Proc. Natl Acad. Sci. USA* **80,** 1816–1820 (1983).
175. Goldman, M. A., Holmquist, G. P., Gray, M. C., Caston, L. A. & Nag, A. Replication timing of genes and middle repetitive sequences. *Science* **224,** 686–692 (1984).
176. Manuelidis, L. & Ward, D. C. Chromosomal and nuclear distribution of the *Hin*dIII 1.9-kb human DNA repeat segment. *Chromosoma* **91,** 28–38 (1984).
177. Feng, Q., Moran, J. V., Kazazian, H. H. Jr & Boeke, J. D. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* **87,** 905–916 (1996).
178. Jurka, J. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proc. Natl Acad. Sci. USA* **94,** 1872–1877 (1997).
179. Arcot, S. S. *et al.* High-resolution cartography of recently integrated human chromosome 19-specific Alu fossils. *J. Mol. Biol.* **281,** 843–856 (1998).
180. Schmid, C. W. Does SINE evolution preclude Alu function? *Nucleic Acids Res.* **26,** 4541–4550 (1998).
181. Chu, W. M., Ballard, R., Carpick, B. W., Williams, B. R. & Schmid, C. W. Potential Alu function: regulation of the activity of double-stranded RNA-activated kinase PKR. *Mol. Cell. Biol.* **18,** 58–68 (1998).
182. Li, T., Spearow, J., Rubin, C. M. & Schmid, C. W. Physiological stresses increase mouse short interspersed element (SINE) RNA expression in vivo. *Gene* **239,** 367–372 (1999).
183. Liu, W. M., Chu, W. M., Choudary, P. V. & Schmid, C. W. Cell stress and translational inhibitors transiently increase the abundance of mammalian SINE transcripts. *Nucleic Acids Res.* **23,** 1758–1765 (1995).
184. Filipski, J. Correlation between molecular clock ticking, codon usage fidelity of DNA repair, chromosome banding and chromatin compactness in germline cells. *FEBS Lett.* **217,** 184–186 (1987).
185. Sueoka, N. Directional mutation pressure and neutral molecular evolution. *Proc. Natl Acad. Sci.*

© 2001 Macmillan Magazines Ltd

USA **85,** 2653–2657 (1988).
186. Wolfe, K. H., Sharp, P. M. & Li, W. H. Mutation rates differ among regions of the mammalian genome. *Nature* **337,** 283–285 (1989).
187. Bains, W. Local sequence dependence of rate of base replacement in mammals. *Mutat. Res.* **267,** 43–54 (1992).
188. Mathews, C. K. & Ji, J. DNA precursor asymmetries, replication fidelity, and variable genome evolution. *Bioessays* **14,** 295–301 (1992).
189. Holmquist, G. P. & Filipski, J. Organization of mutations along the genome: a prime determinant of genome evolution. *Trends Ecol. Evol.* **9,** 65–68 (1994).
190. Eyre-Walker, A. Evidence of selection on silent site base composition in mammals: potential implications for the evolution of isochores and junk DNA. *Genetics* **152,** 675–683 (1999).
191. The International SNP Map Working Group. An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407,** 513–516 (2000).
192. Bohossian, H. B., Skaletsky, H. & Page, D. C. Unexpectedly similar rates of nucleotide substitution found in male and female hominids. *Nature* **406,** 622–625 (2000).
193. Skowronski, J., Fanning, T. G. & Singer, M. F. Unit-length LINE-1 transcripts in human teratocarcinoma cells. *Mol. Cell. Biol.* **8,** 1385–1397 (1988).
194. Boissinot, S., Chevret, P. & Furano, A. V. L1 (LINE-1) retrotransposon evolution and amplification in recent human history. *Mol. Biol. Evol.* **17,** 915–928 (2000).
195. Moran, J. V. Human L1 retrotransposition: insights and peculiarities learned from a cultured cell retrotransposition assay. *Genetica* **107,** 39–51 (1999).
196. Kazazian, H. H. Jr *et al.* Haemophilia A resulting from *de novo* insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature* **332,** 164–166 (1988).
197. Sheen, F.-m. *et al.* Reading between the LINEs: Human genomic variation introduced by LINE-1 retrotransposition. *Genome Res.* **10,** 1496–1508 (2000).
198. Dombroski, B. A., Mathias, S. L., Nanthakumar, E., Scott, A. F. & Kazazian, H. H. Jr Isolation of an active human transposable element. *Science* **254,** 1805–1808 (1991).
199. Holmes, S. E., Dombroski, B. A., Krebs, C. M., Boehm, C. D. & Kazazian, H. H. Jr A new retrotransposable human L1 element from the LRE2 locus on chromosome 1q produces a chimaeric insertion. *Nature Genet.* **7,** 143–148 (1994).
200. Sassaman, D. M. *et al.* Many human L1 elements are capable of retrotransposition. *Nature Genet.* **16,** 37–43 (1997).
201. Dombroski, B. A., Scott, A. F. & Kazazian, H. H. Jr Two additional potential retrotransposons isolated from a human L1 subfamily that contains an active retrotransposable element. *Proc. Natl Acad. Sci. USA* **90,** 6513–6517 (1993).
202. Kimberland, M. L. *et al.* Full-length human L1 insertions retain the capacity for high frequency retrotransposition in cultured cells. *Hum. Mol. Genet.* **8,** 1557–1560 (1999).
203. Moran, J. V. *et al.* High frequency retrotransposition in cultured mammalian cells. *Cell* **87,** 917–927 (1996).
204. Moran, J. V., DeBerardinis, R. J. & Kazazian, H. H. Jr Exon shuffling by L1 retrotransposition. *Science* **283,** 1530–1534 (1999).
205. Pickeral, O. K., Makalowski, W., Boguski, M. S. & Boeke, J. D. Frequent human genomic DNA transduction driven by LINE-1 retrotransposition. *Genome Res.* **10,** 411–415 (2000).
206. Miki, Y. *et al.* Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. *Cancer Res.* **52,** 643–645 (1992).
207. Branciforte, D. & Martin, S. L. Developmental and cell type specificity of LINE-1 expression in mouse testis: implications for transposition. *Mol. Cell. Biol.* **14,** 2584–2592 (1994).
208. Trelogan, S. A. & Martin, S. L. Tightly regulated, developmentally specific expression of the first open reading frame from LINE-1 during mouse embryogenesis. *Proc. Natl Acad. Sci. USA* **92,** 1520–1524 (1995).
209. Jurka, J. & Kapitonov, V. V. Sectorial mutagenesis by transposable elements. *Genetica* **107,** 239–248 (1999).
210. Fraser, M. J., Ciszczon, T., Elick, T. & Bauser, C. Precise excision of TTAA-specific lepidopteran transposons piggyBac (IFP2) and tagalong (TFP3) from the baculovirus genome in cell lines from two species of Lepidoptera. *Insect Mol. Biol.* **5,** 141–151 (1996).
211. Brosius, J. Genomes were forged by massive bombardments with retroelements and retrosequences. *Genetica* **107,** 209–238 (1999).
212. Kruglyak, S., Durrett, R. T., Schug, M. D. & Aquadro, C. F. Equilibrium distribution of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proc. Natl Acad. Sci. USA* **95,** 10774–10778 (1998).
213. Toth, G., Gaspari, Z. & Jurka, J. Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res.* **10,** 967–981 (2000).
214. Ellegren, H. Heterogeneous mutation processes in human microsatellite DNA sequences. *Nature Genet.* **24,** 400–402 (2000).
215. Ji, Y., Eichler, E. E., Schwartz, S. & Nicholls, R. D. Structure of chromosomal duplicons and their role in mediating human genomic disorders. *Genome Res.* **10,** 597–610 (2000).
216. Eichler, E. E. Masquerading repeats: paralogous pitfalls of the human genome. *Genome Res.* **8,** 758–762 (1998).
217. Mazzarella, R. & D. Schlessinger, D. Pathological consequences of sequence duplications in the human genome. *Genome Res.* **8,** 1007–1021 (1998).
218. Eichler, E. E. *et al.* Interchromosomal duplications of the adrenoleukodystrophy locus: a phenomenon of pericentromeric plasticity. *Hum. Mol. Genet.* **6,** 991–1002 (1997).
219. Horvath, J. E., Schwartz, S. & Eichler, E. E. The mosaic structure of human pericentromeric DNA: a strategy for characterizing complex regions of the human genome. *Genome Res.* **10,** 839–852 (2000).
220. Brand-Arpon, V. *et al.* A genomic region encompassing a cluster of olfactory receptor genes and a myosin light chain kinase (MYLK) gene is duplicated on human chromosome regions 3q13-q21 and 3p13. *Genomics* **56,** 98–110 (1999).
221. Arnold, N., Wienberg, J., Ermert, K. & Zachau, H. G. Comparative mapping of DNA probes derived from the V kappa immunoglobulin gene regions on human and great ape chromosomes by fluorescence in situ hybridization. *Genomics* **26,** 147–150 (1995).
222. Eichler, E. E. *et al.* Duplication of a gene-rich cluster between 16p11.1 and Xq28: a novel pericentromeric-directed mechanism for paralogous genome evolution. *Hum. Mol. Genet.* **5,** 899–912 (1996).
223. Potier, M. *et al.* Two sequence-ready contigs spanning the two copies of a 200-kb duplication on human 21q: partial sequence and polymorphisms. *Genomics* **51,** 417–426 (1998).
224. Regnier, V. *et al.* Emergence and scattering of multiple neurofibromatosis (NF1)-related sequences during hominoid evolution suggest a process of pericentromeric interchromosomal transposition. *Hum. Mol. Genet.* **6,** 9–16 (1997).
225. Ritchie, R. J., Mattei, M. G. & Lalande, M. A large polymorphic repeat in the pericentromeric region of human chromosome 15q contains three partial gene duplications. *Hum. Mol. Genet.* **7,** 1253–1260 (1998).
226. Trask, B. J. *et al.* Members of the olfactory receptor gene family are contained in large blocks of DNA duplicated polymorphically near the ends of human chromosomes. *Hum. Mol. Genet.* **7,** 13–26 (1998).
227. Trask, B. J. *et al.* Large multi-chromosomal duplications encompass many members of the olfactory receptor gene family in the human genome. *Hum. Mol. Genet.* **7,** 2007–2020 (1998).
228. van Deutekom, J. C. *et al.* Identification of the first gene (FRG1) from the FSHD region on human chromosome 4q35. *Hum. Mol. Genet.* **5,** 581–590 (1996).
229. Zachau, H. G. The immunoglobulin kappa locus—or—what has been learned from looking closely at one-tenth of a percent of the human genome. *Gene* **135,** 167–173 (1993).
230. Zimonjic, D. B., Kelley, M. J., Rubin, J. S., Aaronson, S. A. & Popescu, N. C. Fluorescence in situ hybridization analysis of keratinocyte growth factor gene amplification and dispersion in evolution of great apes and humans. *Proc. Natl Acad. Sci. USA* **94,** 11461–11465 (1997).
231. van Geel, M. *et al.* The FSHD region on human chromosome 4q35 contains potential coding regions among pseudogenes and a high density of repeat elements. *Genomics* **61,** 55–65 (1999).
232. Horvath, J. E. *et al.* Molecular structure and evolution of an alpha satellite/non-alpha satellite junction at 16p11. *Hum. Mol. Genet.* **9,** 113–123 (2000).
233. Guy, J. *et al.* Genomic sequence and transcriptional profile of the boundary between pericentromeric satellites and genes on human chromosome arm 10q. *Hum. Mol. Genet.* **9,** 2029–2042 (2000).
234. Reiter, L. T., Murakami, T., Koeuth, T., Gibbs, R. A. & Lupski, J. R. The human COX10 gene is disrupted during homologous recombination between the 24 kb proximal and distal CMT1A-REPs. *Hum. Mol. Genet.* **6,** 1595–1603 (1997).
235. Amos-Landgraf, J. M. *et al.* Chromosome breakage in the Prader-Willi and Angelman syndromes involves recombination between large, transcribed repeats at proximal and distal breakpoints. *Am. J. Hum. Genet.* **65,** 370–386 (1999).
236. Christian, S. L., Fantes, J. A., Mewborn, S. K., Huang, B. & Ledbetter, D. H. Large genomic duplicons map to sites of instability in the Prader-Willi/Angelman syndrome chromosome region (15q11-q13). *Hum. Mol. Genet.* **8,** 1025–1037 (1999).
237. Edelmann, L., Pandita, R. K. & Morrow, B. E. Low-copy repeats mediate the common 3-Mb deletion in patients with velo-cardio-facial syndrome. *Am. J. Hum. Genet.* **64,** 1076–1086 (1999).
238. Shaikh, T. H. *et al.* Chromosome 22-specific low copy repeats and the 22q11.2 deletion syndrome: genomic organization and deletion endpoint analysis. *Hum. Mol. Genet.* **9,** 489–501 (2000).
239. Francke, U. Williams-Beuren syndrome: genes and mechanisms. *Hum. Mol. Genet.* **8,** 1947–1954 (1999).
240. Peoples, R. *et al.* A physical map, including a BAC/PAC clone contig, of the Williams-Beuren syndrome-deletion region at 7q11.23. *Am. J. Hum. Genet.* **66,** 47–68 (2000).
241. Eichler, E. E., Archidiacono, N. & Rocchi, M. CAGGG repeats and the pericentromeric duplication of the hominoid genome. *Genome Res.* **9,** 1048–1058 (1999).
242. O'Keefe, C. & Eichler, E. in *Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and the Evolution of Gene Families* (eds Sankoff, D. & Nadeau, J.) 29–46 (Kluwer Academic, Dordrecht, 2000).
243. Lander, E. S. The new genomics: Global views of biology. *Science* **274,** 536–539 (1996).
244. Eddy, S. R. Noncoding RNA genes. *Curr. Op. Genet. Dev.* **9,** 695–699 (1999).
245. Ban, N., Nissen, P., Hansen, J., Moore, P. B. & Steitz, T. A. The complete atomic structure of the large ribosomal subunit at 2.4 angstrom resolution. *Science* **289,** 905–920 (2000).
246. Nissen, P., Hansen, J., Ban, N., Moore, P. B. & Steitz, T. A. The structural basis of ribosome activity in peptide bond synthesis. *Science* **289,** 920–930 (2000).
247. Weinstein, L. B. & Steitz, J. A. Guided tours: from precursor snoRNA to functional snoRNP. *Curr. Opin. Cell Biol.* **11,** 378–384 (1999).
248. Bachellerie, J.-P. & Cavaille, J. in *Modification and Editing of RNA* (ed. Benne, H. G. a. R.) 255–272 (ASM, Washington DC, 1998).
249. Burge, C. & Sharp, P. A. Classification of introns: U2-type or U12-type. *Cell* **91,** 875–879 (1997).
250. Brown, C. J. *et al.* The Human Xist gene—analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell* **71,** 527–542 (1992).
251. Kickhoefer, V. A., Vasu, S. K. & Rome, L. H. Vaults are the answer, what is the question? *Trends Cell Biol.* **6,** 174–178 (1996).
252. Hatlen, L. & Attardi, G. Proportion of the HeLa cell genome complementary to the transfer RNA and 5S RNA. *J. Mol. Biol.* **56,** 535–553 (1971).
253. Sprinzl, M., Horn, C., Brown, M., Ioudovitch, A. & Steinberg, S. Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res.* **26,** 148–153 (1998).
254. Long, E. O. & Dawid, I. B. Repeated genes in eukaryotes. *Annu. Rev. Biochem.* **49,** 727–764 (1980).
255. Crick, F. H. Codon–anticodon pairing: the wobble hypothesis. *J. Mol. Biol.* **19,** 548–555 (1966).
256. Guthrie, C. & Abelson, J. in *The Molecular Biology of the Yeast Saccharomyces: Metabolism and Gene Expression* (eds Strathern, J. & Broach J.) 487–528 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1982).
257. Soll, D. & RajBhandary, U. (eds) *tRNA: Structure, Biosynthesis, and Function* (ASM, Washington DC, 1995).
258. Ikemura, T. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.* **2,** 13–34 (1985).
259. Bulmer, M. Coevolution of codon usage and transfer-RNA abundance. *Nature* **325,** 728–730 (1987).
260. Duret, L. tRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal translation of highly expressed genes. *Trends Genet.* **16,** 287–289 (2000).
261. Sharp, P. M. & Matassi, G. Codon usage and genome evolution. *Curr. Opin. Genet. Dev.* **4,** 851–860 (1994).
262. Buckland, R. A. A primate transfer-RNA gene cluster and the evolution of human chromosome 1. *Cytogenet. Cell Genet.* **61,** 1–4 (1992).
263. Gonos, E. S. & Goddard, J. P. Human tRNA-Glu genes: their copy number and organization. *FEBS Lett.* **276,** 138–142 (1990).
264. Sylvester, J. E. *et al.* The human ribosomal RNA genes: structure and organization of the complete repeating unit. *Hum. Genet.* **73,** 193–198 (1986).

© 2001 Macmillan Magazines Ltd

265. Sorensen, P. D. & Frederiksen, S. Characterization of human 5S ribosomal RNA genes. *Nucleic Acids Res.* **19,** 4147–4151 (1991).
266. Timofeeva, M. *et al.* [Organization of a 5S ribosomal RNA gene cluster in the human genome]. *Mol. Biol. (Mosk.)* **27,** 861–868 (1993).
267. Little, R. D. & Braaten, D. C. Genomic organization of human 5S rDNA and sequence of one tandem repeat. *Genomics* **4,** 376–383 (1989).
268. Maden, B. E. H. The numerous modified nucleotides in eukaryotic ribosomal RNA. *Prog. Nucleic Acid Res. Mol. Biol.* **39,** 241–303 (1990).
269. Tycowski, K. T., You, Z. H., Graham, P. J. & Steitz, J. A. Modification of U6 spliceosomal RNA is guided by other small RNAs. *Mol. Cell* **2,** 629–638 (1998).
270. Pavelitz, T., Liao, D. Q. & Weiner, A. M. Concerted evolution of the tandem array encoding primate U2 snRNA (the RNU2 locus) is accompanied by dramatic remodeling of the junctions with flanking chromosomal sequences. *EMBO J.* **18,** 3783–3792 (1999).
271. Lindgren, V., Ares, A., Weiner, A. M. & Francke, U. Human genes for U2 small nuclear RNA map to a major adenovirus 12 modification site on chromosome 17. *Nature* **314,** 115–116 (1985).
272. Van Arsdell, S. W. & Weiner, A. M. Human genes for U2 small nuclear RNA are tandemly repeated. *Mol. Cell. Biol.* **4,** 492–499 (1984).
273. Gao, L. I., Frey, M. R. & Matera, A. G. Human genes encoding U3 snRNA associate with coiled bodies in interphase cells and are clustered on chromosome 17p11. 2 in a complex inverted repeat structure. *Nucleic Acids Res.* **25,** 4740–4747 (1997).
274. Hawkins, J. D. A survey on intron and exon lengths. *Nucleic Acids Res.* **16,** 9893–9908 (1988).
275. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268,** 78–94 (1997).
276. Labeit, S. & Kolmerer, B. Titins: giant proteins in charge of muscle ultrastructure and elasticity. *Science* **270,** 293–296 (1995).
277. Sterner, D. A., Carlo, T. & Berget, S. M. Architectural limits on split genes. *Proc. Natl Acad. Sci. USA* **93,** 15081–15085 (1996).
278. Sun, Q., Mayeda, A., Hampson, R. K., Krainer, A. R. & Rottman, F. M. General splicing factor SF2/ASF promotes alternative splicing by binding to an exonic splicing enhancer. *Genes Dev.* **7,** 2598–2608 (1993).
279. Tanaka, K., Watakabe, A. & Shimura, Y. Polypurine sequences within a downstream exon function as a splicing enhancer. *Mol. Cell. Biol.* **14,** 1347–1354 (1994).
280. Carlo, T., Sterner, D. A. & Berget, S. M. An intron splicing enhancer containing a G-rich repeat facilitates inclusion of a vertebrate micro-exon. *RNA* **2,** 342–353 (1996).
281. Burset, M., Seledtsov, I. A. & Solovyev, V. V. Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res.* **28,** 4364–4375 (2000).
282. Burge, C. B., Padgett, R. A. & Sharp, P. A. Evolutionary fates and origins of U12-type introns. *Mol. Cell* **2,** 773–785 (1998).
283. Mironov, A. A., Fickett, J. W. & Gelfand, M. S. Frequent alternative splicing of human genes. *Genome Res.* **9,** 1288–1293 (1999).
284. Hanke, J. *et al.* Alternative splicing of human genes: more the rule than the exception? *Trends Genet.* **15,** 389–390 (1999).
285. Brett, D. *et al.* EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Lett.* **474,** 83–86 (2000).
286. Dunham, I. The gene guessing game. *Yeast* **17,** 218–224 (2000).
287. Lewin, B. *Gene Expression* (Wiley, New York, 1980).
288. Lewin, B. *Genes IV* 466–481 (Oxford Univ. Press, Oxford, 1990).
289. Smaglik, P. Researchers take a gamble on the human genome. *Nature* **405,** 264 (2000).
290. Fields, C., Adams, M. D., White, O. & Venter, J. C. How many genes in the human genome? *Nature Genet.* **7,** 345–346 (1994).
291. Liang, F. *et al.* Gene index analysis of the human genome estimates approximately 120,000 genes. *Nature Genet.* **25,** 239–240 (2000).
292. Roest Crollius, H. *et al.* Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence. *Nature Genet.* **25,** 235–238 (2000).
293. The C. elegans Sequencing Consortium. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* **282,** 2012–2018 (1998).
294. Rubin, G. M. *et al.* Comparative genomics of the eukaryotes. *Science* **287,** 2204–2215 (2000).
295. Green, P. *et al.* Ancient conserved regions in new gene sequences and the protein databases. *Science* **259,** 1711–1716 (1993).
296. Fraser, A. G. *et al.* Functional genomic analysis of *C. elegans* chromosome I by systematic RNA interference. *Nature* **408,** 325–330 (2000).
297. Mott, R. EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. *Comput. Appl. Biosci.* **13,** 477–478 (1997).
298. Florea, L., Hartzell, G., Zhang, Z., Rubin, G. M. & Miller, W. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.* **8,** 967–974 (1998).
299. Bailey, L. C. Jr, Searls, D. B. & Overton, G. C. Analysis of EST-driven gene annotation in human genomic sequence. *Genome Res.* **8,** 362–376 (1998).
300. Birney, E., Thompson, J. D. & Gibson, T. J. PairWise and SearchWise: finding the optimal alignment in a simultaneous comparison of a protein profile against all DNA translation frames. *Nucleic Acids Res.* **24,** 2730–2739 (1996).
301. Gelfand, M. S., Mironov, A. A. & Pevzner, P. A. Gene recognition via spliced sequence alignment. *Proc. Natl Acad. Sci. USA* **93,** 9061–9066 (1996).
302. Kulp, D., Haussler, D., Reese, M. G. & Eeckman, F. H. A generalized hidden Markov model for the recognition of human genes in DNA. *ISMB* **4,** 134–142 (1996).
303. Reese, M. G., Kulp, D., Tammana, H. & Haussler, D. Genie—gene finding in *Drosophila melanogaster*. *Genome Res.* **10,** 529–538 (2000).
304. Solovyev, V. & Salamov, A. The Gene-Finder computer tools for analysis of human and model organisms genome sequences. *ISMB* **5,** 294–302 (1997).
305. Guigo, R., Agarwal, P., Abril, J. F., Burset, M. & Fickett, J. W. An assessment of gene prediction accuracy in large DNA sequences. *Genome Res.* **10,** 1631–1642 (2000).
306. Hubbard, T. & Birney, E. Open annotation offers a democratic solution to genome sequencing. *Nature* **403,** 825 (2000).
307. Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **28,** 263–266 (2000).
308. Birney, E. & Durbin, R. Using GeneWise in the Drosophila annotation experiment. *Genome Res.* **10,** 547–548 (2000).
309. The RIKEN Genome Exploration Research Group Phase II Team and the FANTOM Consortium. Functional annotation of a full-length mouse cDNA collection. *Nature* **409,** 685–690 (2001).
310. Basrai, M. A., Hieter, P. & Boeke, J. D. Small open reading frames: beautiful needles in the haystack. *Genome Res.* **7,** 768–771 (1997).
311. Janin, J. & Chothia, C. Domains in proteins: definitions, location, and structural principles. *Methods Enzymol.* **115,** 420–430 (1985).
312. Ponting, C. P., Schultz, J., Copley, R. R., Andrade, M. A. & Bork, P. Evolution of domain families. *Adv. Protein Chem.* **54,** 185–244 (2000).
313. Doolittle, R. F. The multiplicity of domains in proteins. *Annu. Rev. Biochem.* **64,** 287–314 (1995).
314. Bateman, A. & Birney, E. Searching databases to find protein domain organization. *Adv. Protein Chem.* **54,** 137–157 (2000).
315. Futreal, P. A. *et al.* Cancer and genomics. *Nature* **409,** 850–852 (2001).
316. Nestler, E. J. & Landsman, D. Learning about addiction from the human draft genome. *Nature* **409,** 834–835 (2001).
317. Tupler, R., Perini, G. & Green, M. R. Expressing the human genome. *Nature* **409,** 832–835 (2001).
318. Fahrer, A. M., Bazan, J. F., Papathanasiou, P., Nelms, K. A. & Goodnow, C. C. A genomic view of immunology. *Nature* **409,** 836–838 (2001).
319. Li, W. -H., Gu, Z., Wang, H. & Nekrutenko, A. Evolutionary analyses of the human genome. *Nature* **409,** 847–849 (2001).
320. Bock, J. B., Matern, H. T., Peden, A. A. & Scheller, R. H. A genomic perspective on membrane compartment organization. *Nature* **409,** 839–841 (2001).
321. Pollard, T. D. Genomics, the cytoskeleton and motility. *Nature* **409,** 842–843 (2001).
322. Murray, A. W. & Marks, D. Can sequencing shed light on cell cycling? *Nature* **409,** 844–846 (2001).
323. Clayton, J. D., Kyriacou, C. P. & Reppert, S. M. Keeping time with the human genome. *Nature* **409,** 829–831 (2001).
324. Chervitz, S. A. *et al.* Comparison of the complete protein sets of worm and yeast: orthology and divergence. *Science* **282,** 2022–2028 (1998).
325. Aravind, L. & Subramanian, G. Origin of multicellular eukaryotes—insights from proteome comparisons. *Curr. Opin. Genet. Dev.* **9,** 688–694 (1999).
326. Attwood, T. K. *et al.* PRINTS-S: the database formerly known as PRINTS. *Nucleic Acids Res.* **28,** 225–227 (2000).
327. Hofmann, K., Bucher, P., Falquet, L. & Bairoch, A. The PROSITE database, its status in 1999. *Nucleic Acids Res.* **27,** 215–219 (1999).
328. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402 (1997).
329. Wolf, Y. I., Kondrashov, F. A. & Koonin, E. V. No footprints of primordial introns in a eukaryotic genome. *Trends Genet.* **16,** 333–334 (2000).
330. Brunner, H. G., Nelen, M., Breakefield, X. O., Ropers, H. H. & van Oost, B. B. A. Abnormal behavior associated with a point mutation in the structural gene for monoamine oxidase A. *Science* **262,** 578–580 (1993).
331. Cases, O. *et al.* Aggressive behavior and altered amounts of brain serotonin and norepinephrine in mice lacking MAOA. *Science* **268,** 1763–1766 (1995).
332. Brunner, H. G. *et al.* X-linked borderline mental retardation with prominent behavioral disturbance: phenotype, genetic localization, and evidence for disturbed monoamine metabolism. *Am. J. Hum. Genet.* **52,** 1032–1039 (1993).
333. Deckert, J. *et al.* Excess of high activity monoamine oxidase A gene promoter alleles in female patients with panic disorder. *Hum. Mol. Genet.* **8,** 621–624 (1999).
334. Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *J. Mol. Biol.* **147,** 195–197 (1981).
335. Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A genomic perspective on protein families. *Science* **278,** 631–637 (1997).
336. Ponting, C. P., Aravind, L., Schultz, J., Bork, P. & Koonin, E. V. Eukaryotic signalling domain homologues in archaea and bacteria. Ancient ancestry and horizontal gene transfer. *J. Mol. Biol.* **289,** 729–745 (1999).
337. Zhang, J., Dyer, K. D. & Rosenberg, H. F. Evolution of the rodent eosinophil-associated Rnase gene family by rapid gene sorting and positive selection. *Proc. Natl Acad. Sci. USA* **97,** 4701–4706 (2000).
338. Shashoua, V. E. Ependymin, a brain extracellular glycoprotein, and CNS plasticity. *Ann. NY Acad. Sci.* **627,** 94–114 (1991).
339. Schultz, J., Copley, R. R., Doerks, T., Ponting, C. P. & Bork, P. SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res.* **28,** 231–234 (2000).
340. Koonin, E. V., Aravind, L. & Kondrashov, A. S. The impact of comparative genomics on our understanding of evolution. *Cell* **101,** 573–576 (2000).
341. Bateman, A., Eddy, S. R. & Chothia, C. Members of the immunoglobulin superfamily in bacteria. *Protein Sci.* **5,** 1939–1941 (1996).
342. Sutherland, D., Samakovlis, C. & Krasnow, M. A. Branchless encodes a Drosophila FGF homolog that controls tracheal cell migration and the pattern of branching. *Cell* **87,** 1091–1101 (1996).
343. Warburton, D. *et al.* The molecular basis of lung morphogenesis. *Mech. Dev.* **92,** 55–81 (2000).
344. Fuchs, T., Glusman, G., Horn-Saban, S., Lancet, D. & Pilpel, Y. The human olfactory subgenome: from sequence to structure to evolution. *Hum. Genet.* **108,** 1–13 (2001).
345. Glusman, G. *et al.* The olfactory receptor gene family: data mining, classification and nomenclature. *Mamm. Genome* **11,** 1016–1023 (2000).
346. Rouquier, S. *et al.* Distribution of olfactory receptor genes in the human genome. *Nature Genet.* **18,** 243–250 (1998).
347. Sharon, D. *et al.* Primate evolution of an olfactory receptor cluster: Diversification by gene conversion and recent emergence of a pseudogene. *Genomics* **61,** 24–36 (1999).
348. Gilad, Y. *et al.* Dichotomy of single-nucleotide polymorphism haplotypes in olfactory receptor genes and pseudogenes. *Nature Genet.* **26,** 221–224 (2000).
349. Gearhart, J. & Kirschner, M. *Cells, Embryos, and Evolution* (Blackwell Science, Malden, Massachusetts, 1997).
350. Barbazuk, W. B. *et al.* The syntenic relationship of the zebrafish and human genomes. *Genome Res.* **10,** 1351–1358 (2000).
351. McLysaght, A., Enright, A. J., Skrabanek, L. & Wolfe, K. H. Estimation of synteny conservation and genome compaction between pufferfish (Fugu) and human. *Yeast* **17,** 22–36 (2000).
352. Trachtulec, Z. *et al.* Linkage of TATA-binding protein and proteasome subunit C5 genes in mice and humans reveals synteny conserved between mammals and invertebrates. *Genomics* **44,** 1–7 (1997).

© 2001 Macmillan Magazines Ltd

353. Nadeau, J. H. Maps of linkage and synteny homologies between mouse and man. *Trends Genet.* **5,** 82–86 (1989).
354. Nadeau, J. H. & Taylor, B. A. Lengths of chromosomal segments conserved since divergence of man and mouse. *Proc. Natl Acad. Sci. USA* **81,** 814–818 (1984).
355. Copeland, N. G. *et al.* A genetic linkage map of the mouse: current applications and future prospects. *Science* **262,** 57–66 (1993).
356. DeBry, R. W. & Seldin, M. F. Human/mouse homology relationships. *Genomics* **33,** 337–351 (1996).
357. Nadeau, J. H. & Sankoff, D. The lengths of undiscovered conserved segments in comparative maps. *Mamm. Genome* **9,** 491–495 (1998).
358. Thomas, J. W. *et al.* Comparative genome mapping in the sequence-based era: early experience with human chromosome 7. *Genome Res.* **10,** 624–633 (2000).
359. Pletcher, M. T. *et al.* Chromosome evolution: The junction of mammalian chromosomes in the formation of mouse chromosome 10. *Genome Res.* **10,** 1463–1467 (2000).
360. Novacek, M. J. Mammalian phylogeny: shaking the tree. *Nature* **356,** 121–125 (1992).
361. O'Brien, S. J. *et al.* Genome maps 10. Comparative genomics. Mammalian radiations. Wall chart. *Science* **286,** 463–478 (1999).
362. Romer, A. S. *Vertebrate Paleontology* (Univ. Chicago Press, Chicago and New York, 1966).
363. Paterson, A. H. *et al.* Toward a unified genetic map of higher plants, transcending the monocot-dicot divergence. *Nature Genet.* **14,** 380–382 (1996).
364. Jenczewski, E., Prosperi, J. M. & Ronfort, J. Differentiation between natural and cultivated populations of *Medicago sativa* (Leguminosae) from Spain: analysis with random amplified polymorphic DNA (RAPD) markers and comparison to allozymes. *Mol. Ecol.* **8,** 1317–1330 (1999).
365. Ohno, S. *Evolution by Gene Duplication* (George Allen and Unwin, London, 1970).
366. Wolfe, K. H. & Shields, D. C. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387,** 708–713 (1997).
367. Blanc, G., Barakat, A., Guyot, R., Cooke, R. & Delseny, M. Extensive duplication and reshuffling in the arabidopsis genome. *Plant Cell* **12,** 1093–1102 (2000).
368. Paterson, A. H. *et al.* Comparative genomics of plant chromosomes. *Plant Cell* **12,** 1523–1540 (2000).
369. Vision, T., Brown, D. & Tanksley, S. The origins of genome duplications in *Arabidopsis*. *Science* **290,** 2114–2117 (2000).
370. Sidow, A. & Bowman, B. H. Molecular phylogeny. *Curr. Opin. Genet. Dev.* **1,** 451–456 (1991).
371. Sidow, A. & Thomas, W. K. A molecular evolutionary framework for eukaryotic model organisms. *Curr. Biol.* **4,** 596–603 (1994).
372. Sidow, A. Gen(om)e duplications in the evolution of early vertebrates. *Curr. Opin. Genet. Dev.* **6,** 715–722 (1996).
373. Spring, J. Vertebrate evolution by interspecific hybridisation—are we polyploid? *FEBS Lett.* **400,** 2–8 (1997).
374. Skrabanek, L. & Wolfe, K. H. Eukaryote genome duplication—where's the evidence? *Curr. Opin. Genet. Dev.* **8,** 694–700 (1998).
375. Hughes, A. L. Phylogenies of developmentally important proteins do not support the hypothesis of two rounds of genome duplication early in vertebrate history. *J. Mol. Evol.* **48,** 565–576 (1999).
376. Lander, E. S. & Schork, N. J. Genetic dissection of complex traits. *Science* **265,** 2037–2048 (1994).
377. Horikawa, Y. *et al.* Genetic variability in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. *Nature Genet.* **26,** 163–175 (2000).
378. Hastbacka, J. *et al.* The diastrophic dysplasia gene encodes a novel sulfate transporter: positional cloning by fine-structure linkage disequilibrium mapping. *Cell* **78,** 1073–1087 (1994).
379. Tischkoff, S. A. *et al.* Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science* **271,** 1380–1387 (1996).
380. Kidd, J. R. *et al.* Haplotypes and linkage disequilibrium at the phenylalanine hydroxylase locus PAH, in a global representation of populations. *Am. J. Hum. Genet.* **63,** 1882–1899 (2000).
381. Mateu, E. *et al.* Worldwide genetic analysis of the CFTR region. *Am. J. Hum. Genet.* **68,** 103–117 (2001).
382. Abecasis, G. R. *et al.* Extent and distribution of linkage disequilibrium in three genomic regions. *Am. J. Hum. Genet.* **68,** 191–197 (2001).
383. Taillon-Miller, P. *et al.* Juxtaposed regions of extensive and minimal linkage disequilibrium in Xq25 and Xq28. *Nature Genet.* **25,** 324–328 (2000).
384. Martin, E. R. *et al.* SNPing away at complex diseases: analysis of single-nucleotide polymorphisms around APOE in Alzheimer disease. *Am. J. Hum. Genet.* **67,** 383–394 (2000).
385. Collins, A., Lonjou, C. & Morton, N. E. Genetic epidemiology of single-nucleotide polymorphisms. *Proc. Natl Acad. Sci. USA* **96,** 15173–15177 (1999).
386. Dunning, A. M. *et al.* The extent of linkage disequilibrium in four populations with distinct demographic histories. *Am. J. Hum. Genet.* **67,** 1544–1554 (2000).
387. Rieder, M. J., Taylor, S. L., Clark, A. G. & Nickerson, D. A. Sequence variation in the human angiotensin converting enzyme. *Nature Genet.* **22,** 59–62 (1999).
388. Collins, F. S. Positional cloning moves from perditional to traditional. *Nature Genet.* **9,** 347–350 (1995).
389. Nagamine, K. *et al.* Positional cloning of the APECED gene. *Nature Genet.* **17,** 393–398 (1997).
390. Reuber, B. E. *et al.* Mutations in PEX1 are the most common cause of peroxisome biogenesis disorders. *Nature Genet.* **17,** 445–448 (1997).
391. Portsteffen, H. *et al.* Human PEX1 is mutated in complementation group 1 of the peroxisome biogenesis disorders. *Nature Genet.* **17,** 449–452 (1997).
392. Everett, L. A. *et al.* Pendred syndrome is caused by mutations in a putative sulphate transporter gene (PDS). *Nature Genet.* **17,** 411–422 (1997).
393. Coffey, A. J. *et al.* Host response to EBV infection in X-linked lymphoproliferative disease results from mutations in an SH2-domain encoding gene. *Nature Genet.* **20,** 129–135 (1998).
394. Van Laer, L. *et al.* Nonsyndromic hearing impairment is associated with a mutation in DFNA5. *Nature Genet.* **20,** 194–197 (1998).
395. Sakuntabhai, A. *et al.* Mutations in ATP2A2, encoding a Ca2+ pump, cause Darier disease. *Nature Genet.* **21,** 271–277 (1999).
396. Gedeon, A. K. *et al.* Identification of the gene (SEDL) causing X-linked spondyloepiphyseal dysplasia tarda. *Nature Genet.* **22,** 400–404 (1999).
397. Hurvitz, J. R. *et al.* Mutations in the CCN gene family member WISP3 cause progressive pseudorheumatoid dysplasia. *Nature Genet.* **23,** 94–98 (1999).
398. Laberge-le Couteulx, S. *et al.* Truncating mutations in CCM1, encoding KRIT1, cause hereditary cavernous angiomas. *Nature Genet.* **23,** 189–193 (1999).
399. Sahoo, T. *et al.* Mutations in the gene encoding KRIT1, a Krev-1/rap1a binding protein, cause cerebral cavernous malformations (CCM1). *Hum. Mol. Genet.* **8,** 2325–2333 (1999).
400. McGuirt, W. T. *et al.* Mutations in COL11A2 cause non-syndromic hearing loss (DFNA13). *Nature Genet.* **23,** 413–419 (1999).
401. Moreira, E. S. *et al.* Limb-girdle muscular dystrophy type 2G is caused by mutations in the gene encoding the sarcomeric protein telethonin. *Nature Genet.* **24,** 163–166 (2000).
402. Ruiz-Perez, V. L. *et al.* Mutations in a new gene in Ellis-van Creveld syndrome and Weyers acrodental dysostosis. *Nature Genet.* **24,** 283–286 (2000).
403. Kaplan, J. M. *et al.* Mutations in ACTN4, encoding alpha-actinin-4, cause familial focal segmental glomerulosclerosis. *Nature Genet.* **24,** 251–256 (2000).
404. Escayg, A. *et al.* Mutations of SCN1A, encoding a neuronal sodium channel, in two families with GEFS+2. *Nature Genet.* **24,** 343–345 (2000).
405. Sacksteder, K. A. *et al.* Identification of the alpha-aminoadipic semialdehyde synthase gene, which is defective in familial hyperlysinemia. *Am. J. Hum. Genet.* **66,** 1736–1743 (2000).
406. Kalaydjieva, L. *et al.* N-myc downstream-regulated gene 1 is mutated in hereditary motor and sensory neuropathy-Lom. *Am. J. Hum. Genet.* **67,** 47–58 (2000).
407. Sundin, O. H. *et al.* Genetic basis of total colourblindness among the Pingelapese islanders. *Nature Genet.* **25,** 289–293 (2000).
408. Kohl, S. *et al.* Mutations in the CNGB3 gene encoding the beta-subunit of the cone photoreceptor cGMP-gated channel are responsible for achromatopsia (ACHM3) linked to chromosome 8q21. *Hum. Mol. Genet.* **9,** 2107–2116 (2000).
409. Avela, K. *et al.* Gene encoding a new RING-B-box-coiled-coil protein is mutated in mulibrey nanism. *Nature Genet.* **25,** 298–301 (2000).
410. Verpy, E. *et al.* A defect in harmonin, a PDZ domain-containing protein expressed in the inner ear sensory hair cells, underlies usher syndrome type 1C. *Nature Genet.* **26,** 51–55 (2000).
411. Bitner-Glindzicz, M. *et al.* A recessive contiguous gene deletion causing infantile hyperinsulinism, enteropathy and deafness identifies the usher type 1C gene. *Nature Genet.* **26,** 56–60 (2000).
412. The May-Hegglin/Fetchner Syndrome Consortium. Mutations in MYH9 result in the May-Hegglin anomaly, and Fechtner and Sebastian syndromes. *Nature Genet.* **26,** 103–105 (2000).
413. Kelley, M. J., Jawien, W., Ortel, T. L. & Korczak, J. F. Mutation of MYH9, encoding non-muscle myosin heavy chain A, in May-Hegglin anomaly. *Nature Genet.* **26,** 106–108 (2000).
414. Kirschner, L. S. *et al.* Mutations of the gene encoding the protein kinase A type I-α regulatory subunit in patients with the Carney complex. *Nature Genet.* **26,** 89–92 (2000).
415. Lalwani, A. K. *et al.* Human nonsyndromic hereditary deafness DFNA17 is due to a mutation in non-muscle myosin MYH9. *Am. J. Hum. Genet.* **67,** 1121–1128 (2000).
416. Matsuura, T. *et al.* Large expansion of the ATTCT pentanucleotide repeat in spinocerebellar ataxia type 10. *Nature Genet.* **26,** 191–194 (2000).
417. Delettre, C. *et al.* Nuclear gene OPA1, encoding a mitochondrial dynamin-related protein, is mutated in dominant optic atrophy. *Nature Genet.* **26,** 207–210 (2000).
418. Pusch, C. M. *et al.* The complete form of X-linked congenital stationary night blindness is caused by mutations in a gene encoding a leucine-rich repeat protein. *Nature Genet.* **26,** 324–327 (2000).
419. The ADHR Consortium. Autosomal dominant hypophosphataemic rickets is associated with mutations in FGF23. *Nature Genet.* **26,** 345–348 (2000).
420. Bomont, P. *et al.* The gene encoding gigaxonin, a new member of the cytoskeletal BTB/kelch repeat family, is mutated in giant axonal neuropathy. *Nature Genet.* **26,** 370–374 (2000).
421. Tullio-Pelet, A. *et al.* Mutant WD-repeat protein in triple-A syndrome. *Nature Genet.* **26,** 332–335 (2000).
422. Nicole, S. *et al.* Perlecan, the major proteoglycan of basement membranes, is altered in patients with Schwartz-Jampel syndrome (chondrodystrophic myotonia). *Nature Genet.* **26,** 480–483 (2000).
423. Rogaev, E. I. *et al.* Familial Alzheimer's disease in kindreds with missense mutations in a gene on chromosome 1 related to the Alzheimer's disease type 3 gene. *Nature* **376,** 775–778 (1995).
424. Sherrington, R. *et al.* Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease. *Nature* **375,** 754–760 (1995).
425. Olivieri, N. F. & Weatherall, D. J. The therapeutic reactivation of fetal haemoglobin. *Hum. Mol. Genet.* **7,** 1655–1658 (1998).
426. Drews, J. Research & development. Basic science and pharmaceutical innovation. *Nature Biotechnol.* **17,** 406 (1999).
427. Drews, J. Drug discovery: a historical perspective. *Science* **287,** 1960–1964 (2000).
428. Davies, P. A. *et al.* The 5-HT3B subunit is a major determinant of serotonin-receptor function. *Nature* **397,** 359–363 (1999).
429. Heise, C. E. *et al.* Characterization of the human cysteinyl leukotriene 2 receptor. *J. Biol. Chem.* **275,** 30531–30536 (2000).
430. Fan, W. *et al.* BACE maps to chromosome 11 and a BACE homolog, BACE2, reside in the obligate Down Syndrome region of chromosome 21. *Science* **286,** 1255a (1999).
431. Saunders, A. J., Kim, T. -W. & Tanzi, R. E. BACE maps to chromosome 11 and a BACE homolog, BACE2, reside in the obligate Down Syndrome region of chromosome 21. *Science* **286,** 1255a (1999).
432. Firestein, S. The good taste of genomics. *Nature* **404,** 552–553 (2000).
433. Matsunami, H., Montmayeur, J. P. & Buck, L. B. A family of candidate taste receptors in human and mouse. *Nature* **404,** 601–604 (2000).
434. Adler, E. *et al.* A novel family of mammalian taste receptors. *Cell* **100,** 693–702 (2000).
435. Chandrashekar, J. *et al.* T2Rs function as bitter taste receptors. *Cell* **100,** 703–711 (2000).
436. Hardison, R. C. Conserved non-coding sequences are reliable guides to regulatory elements. *Trends Genet.* **16,** 369–372 (2000).
437. Onyango, P. *et al.* Sequence and comparative analysis of the mouse 1-megabase region orthologous to the human 11p15 imprinted domain. *Genome Res.* **10,** 1697–1710 (2000).
438. Bouck, J. B., Metzker, M. L. & Gibbs, R. A. Shotgun sample sequence comparisons between mouse and human genomes. *Nature Genet.* **25,** 31–33 (2000).
439. Marshall, E. Public-private project to deliver mouse genome in 6 months. *Science* **290,** 242–243 (2000).
440. Wasserman, W. W., Palumbo, M., Thompson, W., Fickett, J. W. & Lawrence, C. E. Human-mouse genome comparisons to locate regulatory sites. *Nature Genet.* **26,** 225–228 (2000).
441. Tagle, D. A. *et al.* Embryonic epsilon and gamma globin genes of a prosimian primate (*Galago crassicaudatus*). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J. Mol. Biol.* **203,** 439–455 (1988).
442. McGuire, A. M., Hughes, J. D. & Church, G. M. Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. *Genome Res.* **10,** 744–757 (2000).

© 2001 Macmillan Magazines Ltd

443. Roth, F. P., Hughes, J. D., Estep, P. W. & Church, G. M. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature Biotechnol.* **16,** 939–945 (1998).
444. Cheng, Y. & Church, G. M. Biclustering of expression data. *ISMB* **8,** 93–103 (2000).
445. Cohen, B. A., Mitra, R. D., Hughes, J. D. & Church, G. M. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nature Genet.* **26,** 183–186 (2000).
446. Feil, R. & Khosla, S. Genomic imprinting in mammals: an interplay between chromatin and DNA methylation? *Trends Genet.* **15,** 431–434 (1999).
447. Robertson, K. D. & Wolffe, A. P. DNA methylation in health and disease. *Nature Rev. Genet.* **1,** 11–19 (2000).
448. Beck, S., Olek, A. & Walter, J. From genomics to epigenomics: a loftier view of life. *Nature Biotechnol.* **17,** 1144–1144 (1999).
449. Hagmann, M. Mapping a subtext in our genetic book. *Science* **288,** 945–946 (2000).
450. Eliot, T. S. in *T. S. Eliot. Collected Poems 1909–1962* (Harcourt Brace, New York, 1963).
451. Soderland, C., Longden, I. & Mott, R. FPC: a system for building contigs from restriction fingerprinted clones. *Comput. Appl. Biosci.* **13,** 523–535 (1997).
452. Mott, R. & Tribe, R. Approximate statistics of gapped alignments. *J. Comp. Biol.* **6,** 91–112 (1999).

Supplementary Information is available on *Nature*'s World-Wide Web site (http://www.nature.com) or as paper copy from the London editorial office of *Nature*.

Acknowledgements

Beyond the authors, many people contributed to the success of this work. E. Jordan provided helpful advice throughout the sequencing effort. We thank D. Leja and J. Shehadeh for their expert assistance on the artwork in this paper, especially the foldout figure; K. Jegalian for editorial assistance; J. Schloss, E. Green and M. Seldin for comments on an earlier version of the manuscript; P. Green and F. Ouelette for critiques of the submitted version; C. Caulcott, A. Iglesias, S. Renfrey, B. Skene and J. Stewart of the Wellcome Trust, P. Whittington and T. Dougans of NHGRI and M. Meugnier of Genoscope for staff support for meetings of the international consortium; and the University of Pennsylvania for facilities for a meeting of the genome analysis group.

We thank Compaq Computer Corporations's High Performance Technical Computing Group for providing a Compaq Biocluster (a 27 node configuration of AlphaServer ES40s, containing 108 CPUs, serving as compute nodes and a file server with one terabyte of secondary storage) to assist in the annotation and analysis. Compaq provided the systems and implementation services to set up and manage the cluster for continuous use by members of the sequencing consortium. Platform Computing Ltd. provided its LSF scheduling and loadsharing software without license fee.

In addition to the data produced by the members of the International Human Genome Sequencing Consortium, the draft genome sequence includes published and unpublished human genomic sequence data from many other groups, all of whom gave permission to include their unpublished data. Four of the groups that contributed particularly significant amounts of data were: M. Adams *et al.* of the Institute for Genomic Research; E. Chen *et al.* of the Center for Genetic Medicine and Applied Biosystems; S.-F. Tsai of National Yang-Ming University, Institute of Genetics, Taipei, Taiwan, Republic of China; and Y. Nakamura, K. Koyama *et al.* of the Institute of Medical Science, University of Tokyo, Human Genome Center, Laboratory of Molecular Medicine, Minato-ku, Tokyo, Japan. Many other groups provided smaller numbers of database entries. We thank them all; a full list of the contributors of unpublished sequence is available as Supplementary Information.

This work was supported in part by the National Human Genome Research Institute of the US NIH; The Wellcome Trust; the US Department of Energy, Office of Biological and Environmental Research, Human Genome Program; the UK MRC; the Human Genome Sequencing Project from the Science and Technology Agency (STA) Japan; the Ministry of Education, Science, Sport and Culture, Japan; the French Ministry of Research; the Federal German Ministry of Education, Research and Technology (BMBF) through Projektträger DLR, in the framework of the German Human Genome Project; BEO, Projektträger Biologie, Energie, Umwelt des BMBF und BMWT; the Max-Planck-Society; DFG—Deutsche Forschungsgemeinschaft; TMWFK, Thüringer Ministerium für Wissenschaft, Forschung und Kunst; EC BIOMED2—European Commission, Directorate Science, Research and Development; Chinese Academy of Sciences (CAS), Ministry of Science and Technology (MOST), National Natural Science Foundation of China (NSFC); US National Science Foundation EPSCoR and The SNP Consortium Ltd. Additional support for members of the Genome Analysis group came, in part, from an ARCS Foundation Scholarship to T.S.F., a Burroughs Wellcome Foundation grant to C.B.B. and P.A.S., a DFG grant to P.B., DOE grants to D.H., E.E.E. and T.S.F., an EU grant to P.B., a Marie-Curie Fellowship to L.C., an NIH-NHGRI grant to S.R.E., an NIH grant to E.E.E., an NIH SBIR to D.K., an NSF grant to D.H., a Swiss National Science Foundation grant to L.C., the David and Lucille Packard Foundation, the Howard Hughes Medical Institute, the University of California at Santa Cruz and the W. M. Keck Foundation.

Correspondence and requests for materials should be addressed to E. S. Lander (e-mail: lander@genome.wi.mit.edu), R. H. Waterston (e-mail: bwaterst@watson.wustl.edu), J. Sulston (e-mail: jes@sanger.ac.uk) or F. S. Collins (e-mail: fc23a@nih.gov).

Affiliations for authors: 1, Whitehead Institute for Biomedical Research, Center for Genome Research, Nine Cambridge Center, Cambridge, Massachusetts 02142, USA; 2, The Sanger Centre, The Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1RQ, United Kingdom; 3, Washington University Genome Sequencing Center, Box 8501, 4444 Forest Park Avenue, St. Louis, Missouri 63108, USA; 4, US DOE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, California 94598, USA; 5, Baylor College of Medicine Human Genome Sequencing Center, Department of Molecular and Human Genetics, One Baylor Plaza, Houston, Texas 77030, USA; 6, Department of Cellular and Structural Biology, The University of Texas Health Science Center at San Antonio, 7703 Floyd Curl Drive, San Antonio, Texas 78229-3900, USA; 7, Department of Molecular Genetics, Albert Einstein College of Medicine, 1635 Poplar Street, Bronx, New York 10461, USA; 8, Baylor College of Medicine Human Genome Sequencing Center and the Department of Microbiology & Molecular Genetics, University of Texas Medical School, PO Box 20708, Houston, Texas 77225, USA; 9, RIKEN Genomic Sciences Center, 1-7-22 Suehiro-cho, Tsurumi-ku Yokohama-city, Kanagawa 230-0045, Japan; 10, Genoscope and CNRS UMR-8030, 2 Rue Gaston Cremieux, CP 5706, 91057 Evry Cedex, France; 11, GTC Sequencing Center, Genome Therapeutics Corporation, 100 Beaver Street, Waltham, Massachusetts 02453-8443, USA; 12, Department of Genome Analysis, Institute of Molecular Biotechnology, Beutenbergstrasse 11, D-07745 Jena, Germany; 13, Beijing Genomics Institute/Human Genome Center, Institute of Genetics, Chinese Academy of Sciences, Beijing 100101, China; 14, Southern China National Human Genome Research Center, Shanghai 201203, China; 15, Northern China National Human Genome Research Center, Beijing 100176, China; 16, Multimegabase Sequencing Center, The Institute for Systems Biology, 4225 Roosevelt Way, NE Suite 200, Seattle, Washington 98105, USA; 17, Stanford Genome Technology Center, 855 California Avenue, Palo Alto, California 94304, USA; 18, Stanford Human Genome Center and Department of Genetics, Stanford University School of Medicine, Stanford, California 94305-5120, USA; 19, University of Washington Genome Center, 225 Fluke Hall on Mason Road, Seattle, Washington 98195, USA; 20, Department of Molecular Biology, Keio University School of Medicine, 35 Shinanomachi, Shinjuku-ku, Tokyo 160-8582, Japan; 21, University of Texas Southwestern Medical Center at Dallas, 6000 Harry Hines Blvd., Dallas, Texas 75235-8591, USA; 22, University of Oklahoma's Advanced Center for Genome Technology, Dept. of Chemistry and Biochemistry, University of Oklahoma, 620 Parrington Oval, Rm 311, Norman, Oklahoma 73019, USA; 23, Max Planck Institute for Molecular Genetics, Ihnestrasse 73, 14195 Berlin, Germany; 24, Cold Spring Harbor Laboratory, Lita Annenberg Hazen Genome Center, 1 Bungtown Road, Cold Spring Harbor, New York 11724, USA; 25, GBF - German Research Centre for Biotechnology, Mascheroder Weg 1, D-38124 Braunschweig, Germany; 26, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bldg. 38A, 8600 Rockville Pike, Bethesda, Maryland 20894, USA; 27, Department of Genetics, Case Western Reserve School of Medicine and University Hospitals of Cleveland, BRB 720, 10900 Euclid Ave., Cleveland, Ohio 44106, USA; 28, EMBL European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom; 29, Max Delbrück Center for Molecular Medicine, Robert-Rossle-Strasse 10, 13125 Berlin-Buch, Germany; 30, EMBL, Meyerhofstrasse 1, 69012 Heidelberg, Germany; 31, Dept. of Biology, Massachusetts Institute of Technology, 77 Massachusetts Ave., Cambridge, Massachusetts 02139-4307, USA; 32, Howard Hughes Medical Institute, Dept. of Genetics, Washington University School of Medicine, Saint Louis, Missouri 63110, USA; 33, Dept. of Computer Science, University of California at Santa Cruz, Santa Cruz, California 95064, USA; 34, Affymetrix, Inc., 2612 8th St, Berkeley, California 94710, USA; 35, Genome Exploration Research Group, Genomic Sciences Center, RIKEN Yokohama Institute, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan; 36, Howard Hughes Medical Institute, Department of Computer Science, University of California at Santa Cruz, California 95064, USA; 37, University of Dublin, Trinity College, Department of Genetics, Smurfit Institute, Dublin 2, Ireland; 38, Cambridge Research Laboratory, Compaq Computer Corporation and MIT Genome Center, 1 Cambridge Center, Cambridge, Massachusetts 02142, USA; 39, Dept. of Mathematics, University of California at Santa Cruz, Santa Cruz, California 95064, USA; 40, Dept. of Biology, University of California at Santa Cruz, Santa Cruz, California 95064, USA; 41, Crown Human Genetics Center and Department of Molecular Genetics, The Weizmann Institute of Science, Rehovot 71600, Israel; 42, Dept. of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA; 43, The University of Michigan Medical School, Departments of Human Genetics and Internal Medicine, Ann Arbor, Michigan

© 2001 Macmillan Magazines Ltd

48109, USA; 44, MRC Functional Genetics Unit, Department of Human Anatomy and Genetics, University of Oxford, South Parks Road, Oxford OX1 3QX, UK; 45, Institute for Systems Biology, 4225 Roosevelt Way NE, Seattle, WA 98105, USA; 46, National Human Genome Research Institute, US National Institutes of Health, 31 Center Drive, Bethesda, Maryland 20892, USA; 47, Office of Science, US Department of Energy, 19901 Germantown Road, Germantown, Maryland 20874, USA; 48, The Wellcome Trust, 183 Euston Road, London, NW1 2BE, UK.

† Present addresses: Genome Sequencing Project, Egea Biosciences, Inc., 4178 Sorrento Valley Blvd., Suite F, San Diego, CA 92121, USA (G.A.E.); INRA, Station d'Amélioration des Plantes, 63039 Clermont-Ferrand Cedex 2, France (L.C.).

DNA sequence databases

GenBank, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bldg. 38A, 8600 Rockville Pike, Bethesda, Maryland 20894, USA

EMBL, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

DNA Data Bank of Japan, Center for Information Biology, National Institute of Genetics, 1111 Yata, Mishima-shi, Shizuoka-ken 411-8540, Japan

© 2001 Macmillan Magazines Ltd

Immunology 1979 **37** 753

Information carried by the DNA released by antigen-stimulated lymphocytes

D. JACHERTZ, P. ANKER*, P. A. MAURICE* & M. STROUN† *Institut für medizinische Mikrobiologie und Hygiene, Universität Bern, Switzerland; *Division d'Onco-Hématologie, Département de Médicine, Université de Genève, Genève, Switzerland; †Department of Human Microbiology, Faculty of Medicine, University of Tel Aviv, Tel Aviv, Israel*

Accepted for publication 26 January 1979

Summary. Both antigen-stimulated and non-stimulated human blood lymphocytes release *in vitro* a DNA-containing complex which is not the product of dying or disintegrating cells. Lymphocytes obtained from different PPD or HBs positive or negative donors were incubated with one of these antigens and the DNA released in the culture medium was tested for its information content using, successively, two cell-free systems. The ability of the resulting protein product to bind specifically to the stimulating antigen was examined by immunoadsorption chromatography. Results show that DNA excreted by stimulated lymphocytes was transcribed into an RNA which coded for an antigen-binding protein, whereas DNA released by unstimulated lymphocytes did not. The protein produced in this system, using as template the DNA released after cell stimulation, bound specifically to PPD or HBs Sepharose 4B coated columns, depending on the stimulating antigen and on the cell response to this antigen. After elution from the column the protein sedimented at 19S in a linear sucrose gradient.

INTRODUCTION

DNA release from eukaryote cells has been demonstrated in both amphibian organs (Stroun & Anker, 1972; Stroun, Anker, Gahan & Henri, 1977; Anker & Stroun, 1977) and human lymphocytes, either non-stimulated (Anker, Stroun & Maurice, 1975; 1976) or stimulated with antigens or mitogens (Rogers, Boldt, Kornfeld, Skinner & Valeri, 1972; Sarma & Zubroff, 1973; Olsen & Harris, 1974; Rogers, 1976; Hoessli, Jones, Eisenstadt & Waksman, 1976).

The origin and function of this DNA, which is not produced by dead cells, are unknown, though in the case of lymphocytes preliminary evidence has indicated a possible role in the co-operation process between thymus and bone marrow derived cells during immune interactions (Pelc, Harris & Caldwell, 1972; Harris & Olsen, 1976).

The present study attempts to determine if released DNA from antigen-stimulated lymphocytes contains information for specific antibody synthesis. For this purpose we used successively two cell-free systems previously designed by one of us (D.J.) for studies on cellular DNA: the first aimed to transcribe released DNA into informational RNA (i-RNA) (Jachertz, 1966; Jachertz, Opitz & Opitz, 1972), and the second, to translate this i-RNA into proteins (Jachertz, 1968). Finally the immunological specificity of the proteins produced in this second system was tested by immunoadsorption chromatography.

The data presented are the result of a collaborative study performed in two different laboratories and the various DNA samples tested for biological activity

Correspondence: Dr M. Stroun, Harishonim 7, Savyon, Israel.

0019-2805/79/0800-0753$02.00
© 1979 Blackwell Scientific Publications

Exhibit Q

were sent to one of us (D.J.) under code in order to ensure objective results.

MATERIAL AND METHODS

Antigens

Tuberculin (PPD) was obtained from the Statens Serum Institut, Copenhague, Denmark, and hepatitis B surface antigen A (HBs) was a gift from Dr Gerlich, Hygiene Institut, University of Göttingen, Germany. This last preparation was purified by CsCl equilibrium centrifugation.

Determination of the sensitivity of the different donors to the antigens

(a) *Sensitivity to PPD.* Delayed-type hypersensitivity to PPD was determined by skin testing. The donors were considered negative when no skin reaction appeared 48 h after an injection of 100 units of PPD and when their cultured lymphocytes did not respond to 50 units of PPD/ml of culture fluid (determined by [^{3}H]-thymidine uptake).

(b) *Sensitivity to HBs.* Donors were considered HBs negative if they had no jaundice history and if no HBs antigen nor HBs antibody were detected in their blood. These determinations were performed by the Division of Haematology (Hôpital cantonal, Geneva). A positive donor who had no HBs antigen but had HBs antibody in her blood was chosen in order to ensure the absence of antigen in control cultures.

Lymphocyte culture and separation of the supernatant

Lymphocyte suspensions were prepared by Ficoll–Isopaque centrifugation (Böyum, 1968), and cultured at 10^6/ml in TC 199 medium (Difco), supplemented with 20% pooled AB serum. The medium was buffered with 0·04 M N-2-hydroxyethylpiperazine-N′-2-ethanesulphonic acid, pH 7·6 (Calbiochem). One hundred and twenty units of penicillin (Mycofarm) and 180 units of colimycin (Laboratoire Roger Bellon) per ml were added. The cells were incubated at 37° in air atmosphere (as required with the HEPES buffer) for different periods of time, up to 6 days, in the presence or in the absence of antigen (50 units/ml of PPD and 10^3 particles/ml of HBs). After incubation, the lymphocytes were centrifuged at 200 *g* for 10 min at 4°. The supernatant was then further centrifuged at 48,000 *g* for 1 h and again at 200,000 *g* for 2 h at 4° and the resulting supernatant was collected as already published (Anker *et al.*, 1975). The lymphocytes were counted at time 0 and after each change of medium. Viability was assayed by trypan blue exclusion, only cultures with less than 10% stained cells were used. Sterility of the supernatants was tested.

In vitro *determination of antigen stimulation*

Parallel to the main lymphocyte cultures, 1 ml cultures were run in triplicate for each series in the presence or in absence of the antigens. These cultures were labelled with [^{3}H]-thymidine (1 μCi/ml, 27 Ci/mmol, Radiochemical Centre, Amersham) for 12 h before collecting the main cultures. The cells were pelleted and their TCA-precipitable radioactivity was determined (Munro & Fleck, 1966).

Methods for DNA extraction

(a) *Extraction from the supernatant and from the cultured lymphocytes.* In order to compare nuclear DNA from intact cells with DNA released in their culture medium, the isolation procedure for extracellular DNA (Anker *et al.*, 1976) was used in both cases. In short, the solutions brought to a concentration of 2% sodium lauryl sulphate were submitted twice to phenol extraction and passed through a hydroxyapatite column. The double-stranded DNA, eluted with 0·48 M phosphate buffer was dialysed and run on a Cs_2SO_4 gradient for 72 h at 160,000 *g*.

(b) *Extraction from non-cultured lymphocytes* (high molecular weight DNA). In order to obtain very high molecular weight DNA, the following procedure had to be used. The cells were washed with PBS (pH 7·2; 0·015 M) and resuspended at a concentration of 10^7 cells/ml. The suspension was then added in small amounts (in drops) to a mixture of equal volumes of a solution of 1% sodium dodecylsulphate (SDS), adjusted to pH 7·2, and of 80% freshly distilled phenol. The mixture was agitated manually very gently from time to time at room temperature for several hours (1 to 3) until all cells were solubilized. After centrifugation at 3000 *g* for 20 min, the aqueous phase was separated from the phenolic phase and mixed with the same volume of 0·5 M NaCl. DNA was precipitated by addition of the same volume of ethanol. The net-like precipitate was removed carefully with a platin loop and gently transferred in a 0·15 M solution of sodium chloride. The precipitation by ethanol was repeated four or five times. Solubilization of the precipitate was complete usually after 24 h. Finally the DNA was run in a Cs_2SO_4 gradient (Szybalski, 1968). With this

method, a DNA developing a hyperchromic effect of at least 25% and with a mol. wt of at least 10^8 was obtained; these characteristics are as shown previously (Jachertz, 1966; 1968; 1973; 1976) a prerequisite for biological activity in a cell-free system.

Characterization of the DNA

The amount of DNA was determined by UV absorption at 260 nm and by deoxyribose colouration (Ceriotti, 1952; Giles & Myers, 1963)). The absence of protein contamination was checked (Lowry, Rosebrough, Farr & Randall, 1951). The double-stranded nature of the DNA was ascertained by heat denaturation followed by elution on a hydroxyapatite column (Bernardi, 1965). Sensitivity to pancreatic DNase I (Sigma), pancreatic RNase (Sigma) and pronase (Grade B; Calbiochem) was tested. The molecular weight was estimated by measurements of the DNA strands seen on electron micrographs prepared according to the Kleinschmidt technique (Kleinschmidt, Lang, Jachertz & Zahn, 1962).

Cell-free transcription and translation systems

Two cell-free systems (Fig. 1) previously described (Jachertz, 1966; 1968; 1973; 1976; Jachertz *et al.*, 1972; Jachertz, Trachsel & Hoida-Werner, 1975) were used in the present study.

(a) *S-100 system.* A suspension of white blood cells was sedimented at 1000 ***g*** for 15 min at room temperature. The cells were then resuspended in 1 ml of 0·1 M Tris–HCl buffer (pH 8) containing 4×10^{-2} M KCl and 5×10^{-3} M Mg acetate and frozen in a CO_2–acetone mixture at $-70°$. The frozen suspension was then mechanically homogenized with a pre-cooled 1 ml pipette. During this procedure the homogenate reached a consistence of soft ice. After this freezing and thawing treatment had been repeated ten times all cells were homogenized. The mixture was then transfered to a melting ice-bath until complete thawing and cleared by centrifugation at 30,000 ***g*** for 20 min at 0°. The clear supernatant was then put on the top of a discontinuous sucrose gradient consisting of 1 ml 60% sucrose and 3 ml 30% sucrose. After centrifugation at 140,000 ***g*** for 5 h at 0°, the clear layer of the cell homogenate was removed, using a pre-cooled pipette and transfered in a pre-cooled Kahn tube placed in a melting ice-bath. This S-100 system was stored until use at $-170°$ in liquid nitrogen. The S-100 system was diluted in ten volumes of Tris–HCl buffer pH 8 and measured in a spectrophotometer. The optical density at 275 nm multiplied by a factor of 100 (corrected for the dilution which was necessary for spectrophotometry) gave empirically a first dilution for the use of S-100 in cell-free systems.

(b) *Preparation of ribosomes.* Ribosomes form a thin white pellet on the top of the 60% sucrose layer obtained as just described for the preparation of the S-100 system. After withdrawal of the 30% sucrose layer, the ribosomal pellet was removed with a pipette and suspended in Tris–HCl buffer (0·1 M; pH 8). The suspension was measured in a spectrophotometer as a first empirical guide for the optimal dilution of the ribosomes in the cell-free system (OD at 275 nm × dilution factor × 1000). Furthermore, the ribosomes were purified by sedimentation gradient (Jachertz *et al.*, 1975).

(c) *Cell-free synthesis of i-RNA (first cell-free system).* The cell-free system used for synthesis of i-RNA had the following composition: DNA 10^{-4} μg/ml, S-100 in a concentration of 1 μg N_2/ml, GTP, ATP, CTP, UTP 5×10^{-9} mol of each and 5×10^{-10} mol of [^{3}H]-UTP (Radiochemical Centre Amersham) with a specific activity of about 20 Ci/mmol, and antigen in a previously tested optimal concentration of 10^5 HBs particles or 10^5 PPD molecules/ml as calculated by the average molecular weight. All components of the cell-free system were diluted in 0·1 M Tris–HCl buffer containing 4×10^{-2} M KCl and 5×10^{-3} M Mg acetate.

After incubation in a 37° waterbath for 180–330 s the synthesis of i-RNA was stopped by addition of a solution of 1% SDS and 1% phenol in 0·01 M Tris–HCl buffer pH 7·0. After mixing, the solution was cooled to 0°. After 15 min, the precipitated SDS was separated by centrifugation. The clear supernatant was layered on the top of a discontinuous gradient consisting of 1 ml of a CsCl solution of a density of 1·9 g/ml and 0·5–1 ml of 30% sucrose. After centrifugation at 40,000 r.p.m. for 4 h at 20° in a SW 50·1 rotor of a Spinco L 65 ultracentrifuge, the bottom of the tube was punctured and the CsCl layer withdrawn. The same volume of distilled water was added to allow freezing at $-20°$. An aliquot was used for scintillation counting.

The i-RNA present in the CsCl cushion of the discontinuous gradient just mentioned was purified and fractionated by sedimenting 0·2 ml of i-RNA through 4·8 ml of a Cs_2SO_4 solution of a density of 1·54 g/ml and centrifuged at 40,000 r.p.m. for 3 h 30 min at 20° on a SW 50·1 rotor in a Spinco L 65 centrifuge. The bottom of the tube was then punctured and fractions

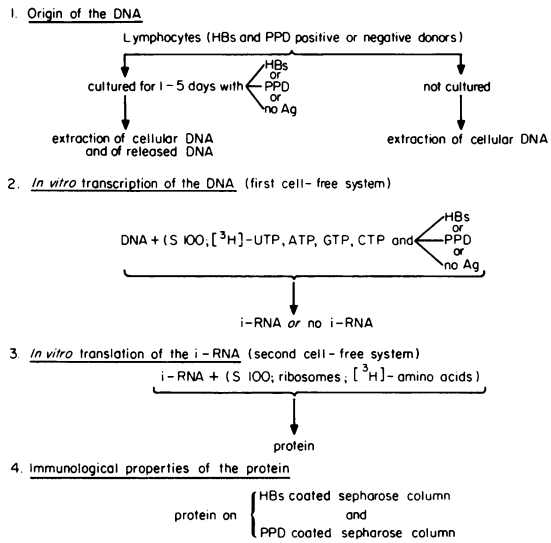


Figure 1. Summary of the methods.

of 0·4 ml collected. An aliquot of 0·1 ml of each fraction was counted by liquid scintillation. The Cs_2SO_4 was autoclaved and the centrifuge tubes were kept in a solution of 1% SDS for 24 h before the run.

(d) *Cell-free synthesis of protein (second cell-free system).* The i-RNA (10^6 molecules) was incubated in 0·1 ml Tris–HCl buffer pH 8·0 (containing 4×10^{-2} M KCl and 5×10^{-3} M Mg acetate) with ribosomes in a concentration of 10^{-3} μg N_2/ml and 4 μg N_2/ml of S-100 System. [^{3}H]-L-valine (20 Ci/mmol, Radiochemical Centre, Amersham) and [^{3}H]-L-leucine (46 Ci/mmol, Radiochemical Centre, Amersham) were added at a concentration of 10^{-9} mol/ml each in order to label the protein. This cell-free system was incubated at 37° for 45 min.

Immuno-adsorption chromatography

Antigen-coated Sepharose 4B was prepared as described by Borries & Koch (1975).

Determination of the molecular weight of the adsorbed protein

The adsorbed proteins were eluted from the Sepharose 4B columns coated with antigen, (Borries & Koch, 1975). In short, the columns were washed with PBS (0·1 M pH 7·2) and when the effluent was free of radioactivity 0·05 M HCl was passed on the column. Fractions of 1 ml were collected. An aliquot of each fraction was counted by liquid scintillation.

The molecular weight of the eluted proteins was determined by zonal centrifugation in a 5–30% sucrose linear gradient.

Figure 1 summarizes the sequence of the procedures used in the present investigation.

RESULTS

Characterization of the different DNAs

Both released and cellular DNA reacted with diphenylamine or with indole in conformity with UV absorption which showed a typical absorption curve. After DNase treatment, over 95% of the acid-insoluble DNA became acid soluble while it remained completely insensitive to either RNase or Pronase. After heat denaturation, over 90% of the DNA was eluted from hydroxyapatite by 0·12 M phosphate buffer.

The only significant difference observed was the size of the molecules. As estimated by E.M. observation, the cellular DNA extracted by the phenol technique with cautious stirring had a mol. wt of about 10^8. Both cellular and extracellular DNAs extracted by the much more drastic procedure had a much lower mol. wt: $2–5 \times 10^6$ for the cellular DNA, and $1–4 \times 10^6$ for the released DNA.

Synthesis of i-RNA *in vitro* upon antigen induction using different cellular and extracellular DNA as template

Figure 2 shows the velocity sedimentation profiles of i-RNA synthesized *in vitro* with high molecular weight cellular DNA from non-cultured lymphocytes as a template. Induction* with both PPD and HBs antigens gave rise to the synthesis of an RNA exceeding 28S. When low molecular weight DNA extracted from cultured cells was used as template, i-RNA with similar profiles as shown in Fig. 2 was formed only in two cases: when PPD was added to the culture medium and to the S-100 cell-free system and only on day 4 of culture (Table 1, A1 and B1). When released DNA was used as template, i-RNA (also with typical profiles as in Fig. 2) was synthesized with all three lymphocyte donors. The lymphocytes of the first donor (PPD^+ HBs^-) cultured in presence of PPD released on the first as well as on day 4 of culture, a DNA from which i-RNA could be transcribed (Table 1, A2). With the

* In this article the term 'induction' is used when the antigen has been added to the cell-free system which yields i-RNA in order to avoid any confusion with antigen stimulation of the lymphocyte cultures.

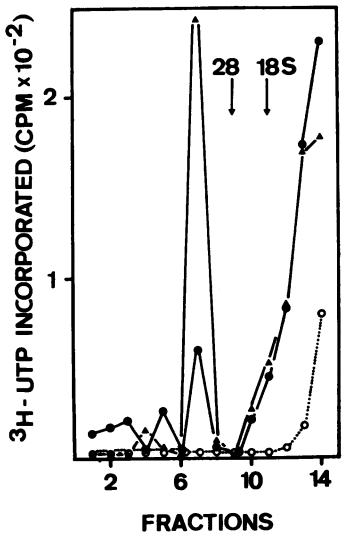


Figure 2. Velocity sedimentation in a Cs_2SO_4 gradient of i-RNA synthesized in a cell-free system containing PPD or HBs antigen and cellular DNA from non-cultured lymphocytes as template. The i-RNA in the antigen-induced S-100 cell-free system was first sedimented through 30% sucrose in a CsCl cushion. The CsCl was withdrawn, diluted with the same volume of distilled water and put on the top of a Cs_2SO_4 solution (density: 1·54 g/ml) and centrifuged at 40,000 r.p.m. for 3 h 30 min at 20°. An aliquot of 0·1 ml of each fraction was counted by liquid scintillation. Arrows indicate the level at which reference bacterial RNA sediments in the same conditions of centrifugation. Sedimentation profile of i-RNA derived from PPD-induced S-100 cell-free system, •; sedimentation profile of i-RNA derived from HBs-induced S-100 cell-free system, ▲; sedimentation profile of i-RNA derived from a non-induced S-100 cell-free system, ○.

first day cultures i-RNA was obtained only when the S-100 cell-free system was induced with PPD while on day 4 i-RNA was obtained when the S-100 cell-free system was induced with either antigen. The lymphocytes of the second donor (PPD^- HBs^-) cultured in presence of PPD, released a DNA which could be transcribed into i-RNA only on day 4 of culture and only when the S-100 cell-free system was induced with PPD (Table 1, B2). The lymphocytes of the third donor (PPD^+ HBs^+) cultured in presence of HBs released on the first as well as on day 4 of culture, a DNA which could be transcribed into i-RNA (Table 1, C2). On the first day, i-RNA was obtained only when the S-100 cell-free system was induced with PPD while on day 4 i-RNA was obtained when the S-100 cell-free system was induced with either antigen.

There was no RNA synthesis in the S-100 system when: (a) no antigen had been added to the lymphocyte cultures; (b) no antigen had been added to the S-100 system; (c) no DNA had been added to the S-100 system.

Antigen-binding properties of the proteins synthesized in a cell-free medium with the different i-RNAs as template

The i-RNA of all series where more than 10^{-15} mol of i-RNA had been obtained, was translated into protein in the second cell-free system. The ability of the labelled protein obtained to specifically bind antigen was tested by passage through antigen-coated Sepharose columns charged with the antigens used to induce the synthesis of i-RNA in the first cell-free system.

Figure 3 shows the elution profile from antigen-coated Sepharose columns of the proteins synthesized in a cell-free system by addition of i-RNA. This i-RNA was obtained after induction with PPD (Fig. 3A) or HBs (Fig. 3B) of a cell-free system containing as template the high molecular weight DNA extracted from the non-cultured lymphocytes. The results show that cellular DNA derived from non-stimulated lymphocytes contains the information for the synthesis of protein binding to both antigens. The i-RNA transcribed from this DNA codes for a protein which binds specifically to the antigen used to induce i-RNA synthesis in the first S-100 cell-free system.

Figure 4 shows the elution profile from antigen-coated Sepharose columns of proteins coded by i-RNA obtained itself in a S-100 cell-free system induced with PPD and in presence of cellular low molecular weight DNA as template. In both cases (PPD^+ and PPD^- lymphocytes) the proteins bind only to the PPD coated column (Fig. 4 A & B).

The elution profiles of the proteins coded by i-RNA obtained from cell-free system containing extracellular DNA are shown in Fig. 5. The proteins derived from the i-RNA obtained from extracellular DNA of lymphocytes (PPD^+ or PPD^-) stimulated with PPD bind only to PPD and not to HBs (Fig. 5 A, B, D). The lymphocytes of a PPD positive donor excreted DNA containing the information for PPD recognition on day 1 (Fig. 5 A). In contrast, a PPD-negative donor excreted such a DNA only after several days (Fig. 5 D).

In the case of the HBs and PPD positive donor whose lymphocytes were stimulated with HBs antigen, the released DNA of the first day yields only PPD induced i-RNA which codes for a PPD binding protein (Fig. 5 E). On day 5, the released DNA contains the information for molecules able to recognize both

Table 1. Incorporation of [^{3}H]-UTP in RNA (mol/ml) after stimulation by PPD or HBs of cell-free systems completed by different cellular DNA and different extracellular DNA obtained from lymphocyte cultures after various times of incubation

Donor	Source of DNA	Culture stimulated by	Cell-free system induced by	Incorporation of [^{3}H]-UTP in RNA harvest of lymphocyte culture after day	
				1	4 or 5*
PPD$^+$ HBs$^-$	Cellular (drastic extraction as extracellular) M W 5×10^6	PPD	PPD	–	$1{\cdot}7 \times 10^{-12}$
			HBs	–	–
	Extracellular M W 2×10^6	PPD	PPD	$1{\cdot}8 \times 10^{-12}$	$1{\cdot}84 \times 10^{-12}$
			HBs	–	$0{\cdot}95 \times 10^{-12}$
PPD$^-$ HBs$^-$	Cellular (drastic extraction as extracellular) M W 5×10^6	PPD	PPD	–	$1{\cdot}7 \times 10^{-12}$
			HBs	–	–
	Extracellular M W 2×10^6	PPD	PPD	–	$1{\cdot}2 \times 10^{-12}$
			HBs	–	–
PPD$^+$ HBs$^+$	Cellular (drastic extraction as extracellular) M W 5×10^6	HBs	PPD	–	–*
			HBs	–	–*
	Extracellular M W 2×10^6	HBs	PPD	$43{\cdot}3 \times 10^{-12}$	$0{\cdot}58 \times 10^{-12}$*
			HBs	–	$1{\cdot}93 \times 10^{-12}$*

* Incorporation on fifth day of culture.

The RNA preparations mentioned above were sedimented through a caesium sulphate gradient and present typical sedimentation profiles as those shown in Fig. 3. Minus sign indicates that there is less than 10^{-15} mol/ml of [^{3}H]-UTP uptake.

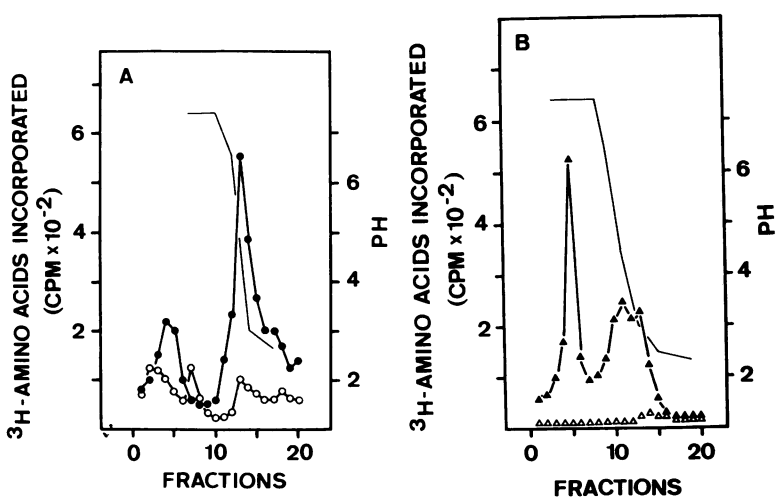


Figure 3. Elution profile from antigen-coated Sepharose columns of proteins synthesized in a cell-free system with i-RNA as template obtained in an antigen induced cell-free system with high molecular weight cellular DNA. The proteins were obtained as described in Material and Methods and put on a Sepharose 4B column coated with PPD or with HBs antigen. The column was then washed with PBS 0·1 M pH 7·2 and, when the effluent was free of radioactivity, 0·05 M HCl was passed on the column. Fractions of 1 ml were collected. An aliquot of 0·1 ml of each fraction was counted by liquid scintillation. A. Elution profile of a protein synthesized with i-RNA obtained from a S-100 cell-free system induced by PPD on a PPD coated column (●) or on a HBs coated column (○). B. Elution profile of a protein synthesized with i-RNA obtained from a S-100 cell-free system induced by HBs on a HBs coated column (▲) or on a PPD coated column (△). The pH gradient is drawn in an uninterrupted line without symbols.

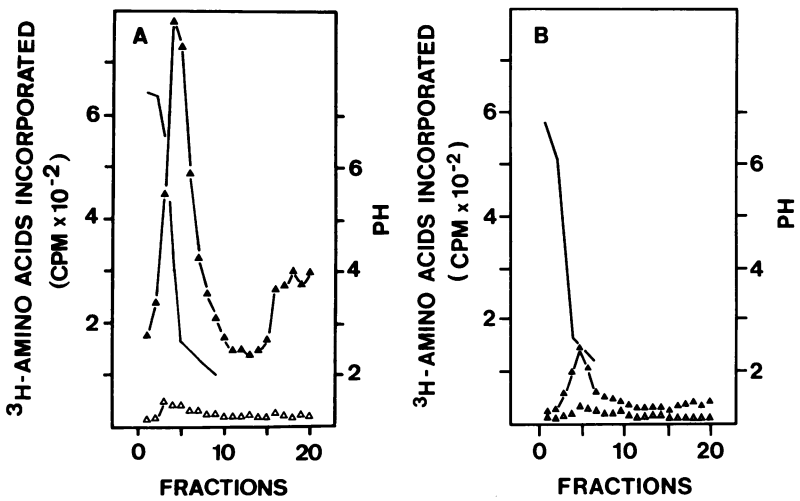


Figure 4. Elution profile from antigen-coated Sepharose columns of proteins synthesized in a cell-free system with i-RNA as template obtained in an antigen induced cell-free system with low molecular weight cellular DNA of lymphocytes cultured in presence of PPD. The proteins obtained were passed through antigen-coated columns as described for Fig. 3. A. Elution profile of a protein synthesized with i-RNA obtained from a S-100 cell-free system induced by PPD. The DNA matrix of the i-RNA came from PPD$^+$ HBs$^-$ lymphocytes cultured 4 days in presence of PPD. ▲, Passage through a PPD coated column; △, passage through an HBs-coated column. B. Elution profile of a protein synthesized with i-RNA obtained from S-100 cell-free system induced by PPD. The DNA matrix of the i-RNA came from PPD$^-$ HBs$^-$ lymphocytes cultured 4 days in presence of PPD. ▲, passage through a PPD coated column; △, passage through an HBs-coated column. The pH gradient is drawn in an uninterrupted line without symbols.

antigens since PPD and HBs i-RNA are synthesized (Fig. 5 F). The PPD induced i-RNA codes for a protein binding to PPD only while the HBs induced i-RNA yields a protein able to recognize PPD and another to recognize HBs (Fig. 5 F).

Finally, the i-RNA produced by the extracellular DNA of a non-stimulated culture was also translated in a cell-free system, in spite of the fact that under 10^{-15} mol of i-RNA had been obtained. The protein formed did not bind to PPD nor to HBs-coated columns.

Table 2 summarizes all the data obtained with excreted DNA.

After elution from the antigen-coated Sepharose 4B columns, the adsorbed proteins sedimented at 19S after a zonal centrifugation in a linear sucrose gradient.

DISCUSSION

The results presented show that DNA released by lymphocytes after antigen stimulation contained information coding for an antigen-binding protein.

i-RNA synthesis in the first S-100 cell-free system

(a) *Cellular DNA*

The results show that high molecular weight cellular DNA of non-cultured lymphocytes can yield i-RNA in a cell-free system stimulated with different antigens. Such results have been obtained with other cellular DNAs as template and with other antigens including several viruses, alcohol dehydrogenase, T^5 as well as K phage receptor particles (Jachertz 1966; 1968; 1973; 1976).

Using the isolation method employed with extracellular DNA, transcription from small mol. wt cellular DNA was observed occasionally and only with cellular DNA from lymphocytes which had been cultured with PPD for 4 days. This is probably due to the isolation procedure used which caused a shearing of the cellular DNA, previous studies having shown that a mol. wt of at least 10^8 is required in the S-100 cell-free system for the cellular DNA to be transcribed into i-RNA after antigen induction.

(b) *Extracellular DNA*

Stimulation of the lymphocytes with antigen seems necessary to obtain an extracellular DNA which can

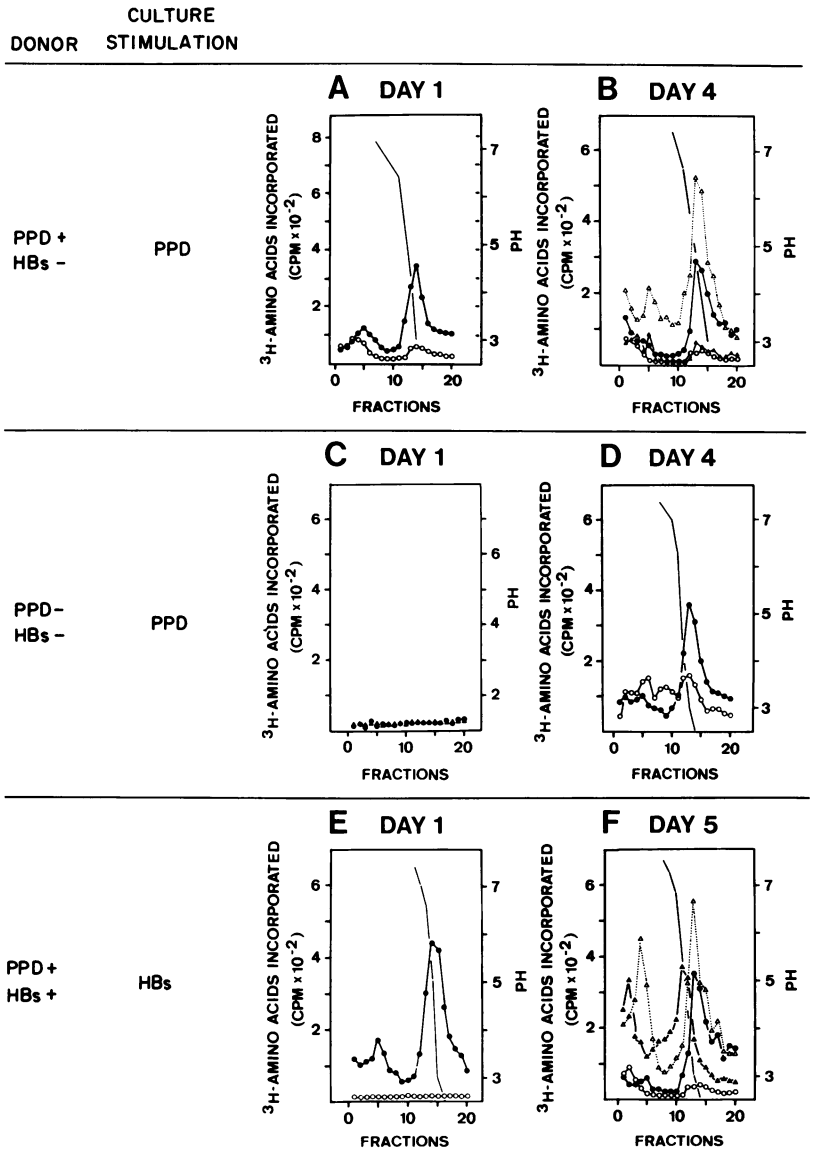


Figure 5. Elution profile from antigen-coated Sepharose columns of protein synthesized in a S-100 cell-free system with i-RNA as template obtained in an antigen-induced cell-free system with low molecular weight DNA released by antigen-stimulated lymphocytes. The proteins obtained were passed through antigen coated columns as described in Fig. 3. ●, Passage through a PPD-coated column of a protein translated from i-RNA obtained in a PPD-induced cell-free system; ○, passage through an HBs-coated column of a protein translated from i-RNA obtained in a PPD-induced cell-free system; ▲, passage through an HBs-coated column of a protein translated from i-RNA obtained in an HBs-induced cell-free system; △, passage through a PPD-coated column of a protein translated from i-RNA obtained in an HBs-induced cell-free system. The pH gradient is drawn in an uninterrupted line without symbols.

Table 2. Summary of the results obtained with DNA released by stimulated lymphocytes concerning i-RNA synthesis and antigen recognition

Donor	Culture stimulated by	i-RNA synthesis induced by	i-RNA synthesis		19S Protein	
			Day 1	Day 4	Day 1	Day 4
PPD$^+$ HBs$^-$	PPD	PPD	+	+	anti-PPD	anti-PPD
		HBs	−	+		anti-PPD
PPD$^-$ HBs$^-$	PPD	PPD	−	+		anti-PPD
		HBs	−	−		
PPD$^+$ HBs$^+$	HBs	PPD	+	+	anti-PPD	anti-PPD
		HBs	−	+		anti-PPD and anti-HBs

+, More than 10^{-12} mol [^{3}H]-UTP incorporation.
−, Less than 10^{-15} mol [^{3}H]-UTP incorporation.

be transcribed into i-RNA in an antigen-induced cell-free system. Indeed, five out of six preparations of DNA released from stimulated lymphocytes were transcribed, whereas no transcription occurred with the DNA released by the three non-stimulated cultures. After 4 or 5 days of culture, no difference in cell counts or cell death rate could be noticed between antigen-stimulated or non-stimulated lymphocytes, thus excluding differences in cell number as the cause of this discrepancy between the two cultures. It should also be stressed that the number of cases in which i-RNA can be transcribed from extracellular DNA increases with the length of the stimulation of the culture.

The need for antigen in the cell-free system to induce i-RNA synthesis is an empirical finding, the mechanism of which has not yet been analysed.

Specificity of the protein translated from i-RNA in the second cell-free system

(a) *Protein translated from i-RNA synthesized with high molecular weight DNA of non-cultured cells as template*
Confirming the earlier experiments of Jachertz (1968; 1973; 1976), the protein translated from i-RNA transcribed from high molecular weight cellular DNA binds specifically to the antigen used to induce the i-RNA synthesis in the first cell-free system.

The different steps of this system have been extensively described and can be summarized as follows: the information content of a given DNA preparation is tested by an indirect method since, in a first step, an i-RNA is transcribed after antigenic induction from the DNA. After the first molecules of i-RNA are synthesized, they are replicated by a RNA-dependent RNA polymerase (Jachertz, 1973). Quantitative determinations suggest that this replication process is a logarithmic one. In a second step, the information content of this i-RNA is translated *in vitro* into protein which is tested for its immunological properties and molecular size. These studies have indicated that the protein translated by i-RNA had the characteristics associated with antibodies. These were immunological specificity, sedimentation velocity, complement fixation after reaction with specific antigens, genetic markers as well as phage and virus neutralization (Jachertz, 1966; 1968; 1973).

(b) *Protein translated from i-RNA synthesized with extracellular DNA or low molecular weight cellular DNA from cultured lymphocytes as template*
The data described in these experiments suggest that the protein translated from i-RNA previously transcribed from extracellular DNA and small cellular DNA of cultured lymphocytes is also an immunologically functional molecule since it binds selectively to an antigen on Sepharose columns.

This antigen recognition seemed to be related to the 'memory' of the lymphocytes, i.e. the antigens they have encountered during their life time. No HBs-binding protein was formed with the DNA released by lymphocytes of the HBs-negative donors, while both the PPD positive donors released a DNA coding for PPD antibody after 24 h of culture. The finding that the lymphocytes of the PPD-negative donor also released a DNA coding for PPD antibody but only on day 4 can be explained by the fact that these lymphocytes need more time to excrete DNA because they encounter this particular antigen for the first time. It

could also be argued that fewer cells are involved in the PPD-negative subject or that this situation reflects differences between primary and secondary response. It seems reasonable, therefore, to assume that after 4 days of culture, the small number of stimulated lymphocytes of the PPD-negative donor have excreted enough DNA coding for PPD antibody to be transcribed which was not the case after 1 day of culture. This increase of specialized DNA could be due to increased synthesis of metabolic extrachromosomal DNA (Pelc, 1964; 1968; 1970; 1972; Harris, 1973) upon antigen stimulation and/or multiplication of the sensitive lymphocytes. This hypothesis also seems to fit with the case of the HBs-positive, PPD-positive donor which released, upon HBs stimulation, a DNA which coded for HBs antibody only on day 5, while it unspecifically released anti-PPD coding DNA on day 1. It appears indeed that in HBs-positive individuals only a very small number of lymphocytes are involved in antigen recognition (Peterson, Dienstag & Purcell, 1975).

The following is proposed to explain the release of DNA from stimulated lymphocytes described here: (a) lymphocytes contain extra copies (metabolic DNA, extrachromosomal DNA) carrying information for antibody; (b) when stimulated by antigen in culture, lymphocytes non-specifically release these copies; (c) antigen stimulation also triggers synthesis of extrachromosomal DNA, coding specifically for antibody directed against this particular antigen. Thus, the DNA subsequently released is 'antigen-specific'.

This would explain why the HBs^+ PPD^+ donor excreted a DNA coding for a PPD-binding protein on the first day of HBs stimulation. In the case of a PPD^+ subject, the first, unspecific DNA release could have statistically more chance to code for PPD antibody than for HBs antibody. The specific excretion (anti-HBs) would occur only later on the result of preferential synthesis of anti-HBs DNA. The absence of anti-HBs antibody synthesis, using DNA released from HBs negative lymphocytes, also argues in favour of such a specificity.

The same model can explain why, in two cases, low molecular weight cellular DNA also yielded i-RNA coding for antibody. Both antibodies were obtained with DNA extracted on the fourth day of culture from cells stimulated with PPD. On the first day of culture, the specialized extrachromosomal molecules would be too small a proportion of the bulk cellular DNA, to have a chance of being transcribed in the cell-free system. After 4 days of stimulation, however, specific DNA synthesis coding for PPD antibody could have increased, resulting in antibody being obtained in the cell-free systems.

Size of DNA and information for antigen recognition

Cellular DNA and extracellular DNA appear to differ only in molecular size and information content per unit length with respect to antibody synthesis. The molecular weight of extracellular DNA is of about 1×10^6, some molecules being as large as 4×10^6. This size would be sufficient to code for H and L chains containing $7{\cdot}5 \times 10^2$ amino acids which implies a minimum of $4{\cdot}5 \times 10^3$ nucleotides of a double stranded molecule, i.e. a molecular weight of about $1{\cdot}75 \times 10^6$. Although the binding protein sedimented at 19S after elution from the column, we have not proved in this study that the protein formed is an IgM molecule and we do not know what proportion of the total protein synthesized binds to the antigen.

Origin and function of the extracellular DNA

Finally, it should be stressed that while the high molecular weight cellular DNA carries information for the synthesis of many different antibodies, only part of this information may be released from the cell. Released DNA is thus qualitatively different from the bulk cellular DNA, and could reflect both the specialization and the memory of the lymphocytes. Transport of DNA between cells has been postulated to explain the co-operation process between thymus and bone marrow derived cells during immune inter-reactions (Pelc *et al.*, 1972; Anker *et al.*, 1975; Rogers, 1976; Hoessli *et al.*, 1976; Harris & Olsen, 1976; Stroun, Anker, Maurice & Gahan, 1977) leading to the differentiation of B cells into antibody-producing plasma cells. Our results provide a molecular basis for this hypothesis.

ACKNOWLEDGMENTS

We thank Professor Gahan and Dr Hoessli for helpful discussion and careful reading of the manuscript. We are also grateful to Mrs Henri and Lederrey for excellent technical assistance.

This work was supported by grants from the 'Ligue Suisse Contre le Cancer' FOR/082/AK/76 (5), the 'Fonds National Suisse de la Recherche Scientifique' No. 3.764.76, and the 'Marthe Seidl-Hentsch Fund'.

REFERENCES

ANKER P. & STROUN M. (1977) Spontaneous extracellular synthesis of DNA released by frog auricles. *Arch. Sci. Genève*, **30**, 263.

ANKER P., STROUN M. & MAURICE P.A. (1975) Spontaneous release of DNA by human blood lymphocytes as shown in an *in vitro* system. *Cancer Res.* **35**, 2375.

ANKER P., STROUN M. & MAURICE P.A. (1976) Spontaneous extracellular synthesis of DNA released by human blood lymphocytes. *Cancer Res.* **36**, 2832.

BERNARDI G. (1965) Chromatography of nucleic acids on hydroxyapatite. *Nature (Lond.)*, **206**, 779.

BORRIES E. & KOCH G. (1975) Isolation of poliovirus-neutralizing antibodies by affinity chromatography. *Virology*, **67**, 356.

BÖYUM A. (1968) Isolation of lymphocytes from human blood. *Scand. J. clin. Invest.* **21**, 31.

CERIOTTI G. (1952) Microchemical determination of deoxyribonucleic acid. *J. biol. Chem.* **198**, 297.

GILES K.W. & MYERS A. (1965) An improved diphenylamine method for the estimation of DNA. *Nature (Lond.)*, **206**, 93.

HARRIS G. (1973) The immune response of spleen explants from primed rabbits to sheep red cells. I. DNA synthesis in the development of antibody producing cells. *Immunology*, **24**, 343.

HARRIS G. & OLSEN I. (1976) Cell division and deoxyribonucleic acid (DNA) synthesis in cultures of stimulated lymphocytes. *Immunology*, **31**, 195.

HOESSLI D.C., JONES A.P., EISENSTADT J.M. & WAKSMAN B.H. (1977) Studies on DNA release by cultured rat lymphoblasts. *Int. Archs Allergy appl. Immun.* **54**, 517.

JACHERTZ D. (1966) Antikörper-Synthesis *in vitro*. II. Stimulierung eines zellfreien Systems. *Z. med. Mikrobiol. Immunol.* **152**, 20.

JACHERTZ D. (1968) Antikörpersynthese *in vitro*. IX. Synthese von informatorischer RNS und Antikörpern in einem zellfreien System aus menschlichen peripheren Leukocyten. *Z. med. Mikrobiol. Immunol.* **154**, 245.

JACHERTZ D. (1973) Flow of information and gene activation during antibody synthesis. *Ann. N.Y. Acad. Sci.* **207**, 122.

JACHERTZ D. (1976) I-RNA: Synthesis and mechanisms of action. In: *Immune RNA in Neoplasia* (Ed. by Mary A. Fink). Academic Press, New York.

JACHERTZ D., OPITZ U. & OPITZ H.G. (1972) Gene amplification in cell-free systems. *Z. Immun.-Forsch.* **144**, 260.

JACHERTZ D., TRACHSEL H. & HOIDA-WERNER A. (1975) Antibody *in vitro* XII. Activity of reconstituted and hybrid recombined ribosomes in cell free synthesis of antibody. *J. Immunogenet.* **2**, 231.

KLEINSCHMIDT A.K., LANG D., JACHERTZ D. & ZAHN R.K. (1962) Darstellung und Länge Messungen des gesamten DNA. Inhaltes von T2 Bacteriophagen. *Biochim. biophys. Acta*, **61**, 857.

LOWRY O., ROSEBROUGH N.J., FARR A.L. & RANDALL R.J. (1951) Protein measurement with the Folin phenol reagent. *J. biol. Chem.* **193**, 265.

MUNRO H. & FLECK A. (1966) Determination of nucleic acids. *Meth. Biochem. Anal.* **14**, 113.

OLSEN I. & HARRIS G. (1974) Uptake and release of DNA by lymphoid tissue and cells. *Immunology*, **27**, 973.

PELC S.R. (1964) Labelling of DNA and cell division in so-called non-dividing tissues. *J. Cell Biol.* **22**, 21.

PELC S.R. (1968) Turnover of DNA and function. *Nature (Lond.)*, **219**, 163.

PELC S.R. (1970) Metabolic DNA and the problem of ageing. *Exp. Gerontol.* **5**, 217.

PELC S.R. (1972) Metabolic DNA in ciliated protozoa, salivary gland chromosomes, and mammalian cells. *Int. Rev. Cytol.* **32**, 327.

PELC S.R., HARRIS G. & CALDWELL I.R. (1972) The relationship between antibody formation and DNA synthesis in mouse spleens during primary and secondary responses to sheep erythrocytes. *Immunology*, **23**, 183.

PETERSON J..M., DIENSTAG J.L. & PURCELL R.H. (1975) Immune response to hepatitis viruses. In: *Viral Immunology and Immunopathology* (Ed. by A. L. Notkins), p. 213. Academic Press, New York.

ROGERS J.C. (1976) Characterization of DNA excreted from phytohemagglutinin stimulated lymphocytes. *J. exp. Med.* **142**, 1249.

ROGERS J.C., BOLDT D., KORNFELD S., SKINNER A. & VALERI P. (1972) Excretion of deoxyribonucleic acid by lymphocytes stimulated with phytohemagglutinin or antigen. *Proc. natn. Acad. Sci. (U.S.A.)*, **69**, 1685.

SARMA D.S.R. & ZUBROFF J. (1973) Synthesis and fragmentation of DNA in phytohaemagglutinin stimulated human peripheral blood lymphocytes. *Immunol. Commun.* **2**, 277.

STROUN M. & ANKER P. (1972) *In vitro* synthesis of DNA spontaneously released by bacteria or frog auricles. *Biochimie*, **54**, 1443.

STROUN M., ANKER P., GAHAN P. & HENRI J. (1977) Spontaneous release of newly synthesized DNA from frog auricles. *Arch. Sci. Genève*, **30**, 229.

STROUN M., ANKER P., MAURICE P.A. & GAHAN P.B. (1977) Circulating nucleic acids in higher organisms. *Int. Rev. Cytol.* **51**, 1.

SZYBALSKI W. (1968) Use of cesium sulfate for equilibrium density gradient centrifugation. *Meth. Enzymol.* **128**, 330.

Newtype COMMUNICATIONS

e-mail: newtypecomm@aol.com
www.newtypecomm.com

445 Fifth Avenue
New York, New York 10016
Phone 212/686-5555
Fax 212/686-5414

STATE OF NEW YORK)
CITY OF NEW YORK :
COUNTY OF NEW YORK)

CERTIFICATION

This is to certify that the following is, to the best of our knowledge and belief, a true and accurate translation into ENGLISH of the attached document(s) relating to:

Cytology journal, Volume 37, Number 3

written in RUSSIAN.

NEWTYPE COMMUNICATIONS, INC.

Sworn to and subscribed before me
this 22nd day of November, 2011

NOTARY PUBLIC

BRIAN G. BROWN
Notary Public, State of New York
No. 01BR6151227
Qualified in Suffolk County
Commission Expires August 14, 2014

Newtype
COMMUNICATIONS

e-mail: newtypecomm@aol.com
www.newtypecomm.com

445 Fifth Avenue
New York, New York 10016
Phone 212/686-5555
Fax 212/686-5414

STATE OF NEW YORK)
CITY OF NEW YORK :
COUNTY OF NEW YORK)

CERTIFICATION

This is to certify that the following is, to the best of our knowledge and belief, a true and accurate translation into ENGLISH of the attached document(s) relating to:

Inset VIII
To the article by V.I. Kazakov, V.M. Bozhkova, V.A. Linde (p. 232)

written in RUSSIAN.

N. S. Lew
NEWTYPE COMMUNICATIONS, INC.

Sworn to and subscribed before me
this 13th day of September, 2012

Brian G. Brown
NOTARY PUBLIC

BRIAN G. BROWN
Notary Public, State of New York
No. 01BR6151227
Qualified in Suffolk County
Commission Expires August 14, 2014

[Handwritten: ISitologiia]

ISSN 0041–3771

RUSSIAN ACADEMY OF SCIENCES

CYTOLOGY

CYTOLOGY

1995. Том 37
1995. Vol. 37

3

Univ. of Minn.
Bio-Medical
Library
8 31 95

DEPARTMENT OF BIOCHEMISTRY, BIOPHYSICS AND CHEMISTRY OF PHYSIOLOGICALLY ACTIVE COMPOUNDS

Journal "Cytology"

founded January 1959

issued 12 times a year

EDITORIAL BOARD:

V. Ya. ALEXANDROV, V. Ya. BRODSKY, Yu. S. CHENTSOV,
I. I. KIKNADZE, E. E. MAKHLIN (Secretary), V. M. MIKHELSON (Associate Editor),
N. N. NIKOLSKY (Editor-in-Chief), A. L. POLENOV, G. D. TUMANISHVILI,
Yu. M. VASILIEV, A. V. ZHIRMUNSKY

[Irrelevant photo caption]

©St. Peterburg Publishing Company RAN "Tsitologiya [Cytology]", 1995

This material may be protected by Copyright law (Title 17 U.S. Code)

© 1995

EXTRACELLULAR DNA IN THE BLOOD OF PREGNANT WOMEN

V. I. Kazakov, V. M. Bozhkov, V. A. Linde, M. A. Repina, V. M. Mikhaylov

Institute of Cytology, Russian Academy of Sciences, and Medical Academy of Postgraduate Physician Education, St. Peterburg

The level of extracellular DNA increases in the blood of women during pregnancy. By means of PCR, the full-size Alu repeats were observed among extracellular blood DNA repeats of pregnant women. Furthermore, with Tc65 type primer the PCR method allowed to observe in the blood DNA fragments flanked by inverted Alu repeats (inter Alu repeats). The presence of such a type of inter Alu repeats was estimated in the blood of women being in the first trimester of pregnancy only, but was not estimated among blood DNA fragments of women of the last trimester of pregnancy. It is discussed which types of cells may serve as a source of extracellular blood DNA (either trophoblast cells, lymphocytes, or decidual cells), the significance of such DNA for pregnancy being appreciated.

It has been shown that extracellular DNA is contained in the blood of humans and animals (Stroun et al., 1977; Fedorov, Yaneva, 1982; Vladimirov et al., 1992). An increase in the content of extracellular DNA in the blood of humans has been described during pathological processes taking place in various types of tissue of the body, especially during certain inflammatory processes of the gastrointestinal tract, during tumor and infectious diseases of viral etiology, and during disseminated lupus (Anker et al., 1975; Leon et al., 1977; Shapiro et al., 1983; Stroun et al., 1987; Goto et al., 1991).

The molecular mass of the extracellular DNA of the blood is on average between $1 \cdot 10^6$ to $15 \cdot 10^6$ Da (Stroun et al., 1977; Fedorov, Yaneva, 1982). According to EM microscopy data, the extracellular DNA of healthy persons is double-strand and linear (Dennin, 1979). According to Vasyukhin et al. (1991), human extracellular DNA contains unique sequences. It has also been shown that the extracellular DNA of the blood during disseminated lupus has sequences capable of forming Z-DNA (Van Helder, 1985).

It is believed that the high-molecular component of extracellular DNA in the blood comes from living cells (Stroun et al., 1977). It has been demonstrated in vitro that certain types of cells, especially lymphocytes, excrete extracellular DNA into their surroundings (Rogers et al., 1972; Rogers, 1976; Stroun et al., 1977; Fedorov, Yaneva, 1982). Furthermore, the blood also contains low-molecular DNA, corresponding to nucleosomes in its mobility. In the blood of rats, its content increases after total X-ray exposure (Belokhvostov et al., 1987; Vladimirov et al., 1992; Tishchenko et al., 1993). According to Tishchenko et al. (1993), the low-molecular DNA in the blood of rats after total irradiation is enriched in GC sequences. It has been conjectured that the low-molecular DNA in the blood is a product of intensified extrachromosomal synthesis

of ring DNA, detected in many types of cells of humans and mammals (Vladimirova et al., 1992). Our attention was drawn to the proposition that the rise in the level of extracellular low-molecular DNA in the blood of rats after irradiation is the consequence of increased activity of Ca/Mg-endonuclease in the cells (Tishchenko et al., 1993). This proposition is in good agreement with data on the intensification of the process of apoptosis in the cells of various tissues of the body after irradiation (Khanson, Komar, 1985).

It is clear from the above that analysis of the extracellular DNA in the blood of humans and animals is of both theoretical and practical interest. It is not ruled out that the nucleotide composition of the extracellular DNA of blood is not so random as not to reflect the peculiarities of the processes of differentiation and cell death taking place in various tissues at each particular moment in the life of the organism. All of this dictated our choice of the blood of pregnant women as the object of our research. According to available data, cellular proliferation, differentiation, and cell death occur in the uterus during pregnancy (Fedorova, Kalashnikova, 1986; Mikhaylov et al., 1989, 1992a, 1992b; Mikhaylov, 1993). It was anticipated that these processes exert an influence on the specifics of the nucleotide composition of the extracellular DNA in the blood of pregnant women.

Material and method

We studied the blood sera of men, nonpregnant women, and women in the first and third trimester of pregnancy and those with late toxicosis of pregnancy. The blood was taken by syringe from the cubital vein under sterile conditions, placed in a centrifuge test tube, and left at room temperature until clotted. Immediately after the formation of a thrombus, it was removed from the walls of the test tube and centrifuged at 400 g for 10 min. The serum was centrifuged yet again at 2000 g for 10 min at 4° C. The serum obtained in this way was kept at -60° C. The use of serum instead of plasma for the analysis of the DNA of the blood can be justified if one observes the conditions for formation of a thrombus at room temperature and immediate removal of the serum from the thrombus (Leon et al., 1977; Shapiro et al., 1983). Such was done in the present investigation.

After this the serum was treated twice with phenol, mixtures of phenol and chloroform (1:1), chloroform, and isoamyl alcohol (24:1) and precipitated with ethanol at -20° C. The DNA preparations were analyzed in 1% agarose gel, and then used as the matrix in a polymerase chain reaction (PCR).

The PCR was carried out using thermophilic DNA polymerase from Thermus thermophilus. We performed 30-35 amplification cycles. The amplified fragments were analyzed in 8% PAAG. The annealing temperature of the primers was 55° C for primers B1 and C2 and 60° C for primer Tc65, the concentration of Mg^{2+} ions being 2-5 mM. The primers for the PCR were synthesized by us with the solid-phase triether phosphite method in the ß-cyanoethyl modification on the Gene assembler instrument from Pharmacia. The sequence of the Tc65 primer was taken from literature sources (Nelson et al., 1989), the sequences of the other primers were found with the help of the Oligo program (Microsoft) for the consensus sequence of Alu repeats of the MF family and designated as B1 and C2 by us.

When the PCR was carried out with the pair of primers B1 and C2, an Alu repeat with a length of 239 base pairs was amplified as a result of the reaction. When only one primer Tc65 was used in the PCR, DNA fragments were amplified that were flanked by two Alu repeats with their terminal 3' regions facing each other (inter-Alu repeats).

The sequences of the primers are:

Tc65—5'—AAGTCGCGGCCGCTTGCAGTGAGCCGAGA—3'
B1—5'—CCTGTAATCCCAGCACTTTGGGAGGC—3'
C2—5'—CCCAGGC$\begin{matrix}T\\G\end{matrix}$GGAGTGC$\begin{matrix}A\\G\end{matrix}$ GTGG—3'

Results and discussion

It was not our goal to study in detail the changes in the concentrations of DNA in the blood serum of pregnant women. From the data of authors using radioimmunological methods of determination (Leon et al., 1977; Shapiro et al., 1983), the concentration of DNA in the blood sera of healthy donors varied from 0 to 0.1 mcg/ml; according to authors using biochemical methods (Anker et al., 1975; Stroun et al., 1987), from 0.1 to 0.6 mcg/ml. Our data on the

concentration of DNA in the blood of men and nonpregnant women corresponded to the literature data.

According to our data findings, during pregnancy there is an increase first of all in the concentration of low-molecular DNA, the increase being most pronounced during gestosis. The size of such DNA is from 150 to 2500 base pairs. There is a belief that the increase in concentration of low-molecular extracellular DNA in the blood is a consequence of increased synthesis of extrachromosomal DNA (Vladimirov et al., 1992). However, in evaluating the data presented, one must consider that a substantial number of endometrial cells perish during pregnancy, especially at the end of pregnancy and in connection with gestosis (Kottsova et al., 1989; Mikhaylov et al., 1992a). According to our data, the death of decidual cells is attended by a gradual loss of nuclear DNA and fragmentation of peripheral parts of the nuclei and cytoplasm and, apparently, this occurs by apoptosis (Mikhaylov et al., 1992a, 1992b; Mikhaylov, 1993). One of the major mechanisms of cell death of the apoptosis type is internucleosomal degradation of DNA due to activation of Ca^{2+} endonuclease (Wyllie, 1980). This does not rule out the possibility of extrachromosomal copying of DNA as the source of the extracellular DNA in the blood. Regardless of whether intensified synthesis of extrachromosomal DNA or apoptotic cell death is the source of the extracellular DNA in the blood, the presence of Alu repeats in it testifies to its nuclear origin (Fig. 1; see insert VIII).

Worthy of attention is the presence of inter-Alu repeats among the DNA of the blood serum of pregnant women. It is known that approximately a third of the Alu repeats of humans belongs to the inverted repeats, and the mean distance between inverted pairs is approximately 50 t.p.n. (Jelinek, Schmid, 1982). The bulk of such inverted repeats apparently contain a unique internal fragment. There are several kinds of inter-Alu repeats. With the help of the Tc65 primer, we detected inter-Alu repeats in the blood of women only in the first trimester of pregnancy (Fig. 2). When the PCR was carried out with the Tc65 primer, only those DNA fragments are amplified in which the spacer DNA is flanked by Alu repeats mutually facing each other by their 3' regions, which makes these fragments even more unique. Thus, if the presence in the blood of pregnant women of an increased quantity of low-molecular DNA fragments is a sign of increased cell death occurring in the uterus in all phases of pregnancy, then the presence of inter-Alu repeats indicates the functioning of other mechanisms for such sequences getting into the serum. It is important that inter-Alu repeats have been detected only in the blood of women in the first trimester of pregnancy. This fact most likely reflects the difference in content of the cellular processes that are characteristic of the early and late stages of pregnancy. While at the end of pregnancy there is observed a death of decidual cells, which explains increase in concentration of low-molecular DNA in the blood; in the initial period of pregnancy when implantation and placentation of the embryo in the uterus occurs, there is decidualization of the endometrium and growth of the embryonal vesicle, i.e., a process of differentiation of the cells of the uterine mucous membrane and cells of the trophoblast.

It is likely that, in the early stages of pregnancy, an excretion of inter-Alu repeats from the cells occurs. Which type of cells can excrete inter-Alu repeats is a matter of great interest. This question also confronts investigators when discussing the sources of high-molecular serum DNA (see the survey: Fedorov, Yaneva, 1982). Excretion of DNA in vitro has been well characterized for lymphocytes (Fedorov, Yaneva, 1982). It is assumed that the lymphocytes might be the source of such DNA during tumor diseases. In the early stages of pregnancy, substantial lymphocyte infiltration is also observed in the uterine mucous membrane.

Thus, in the early stages of pregnancy in humans, cells of the fetus (trophoblasts) and the mother (cells of the endometrium and lymphocytes) may excrete DNA. In view of the above, and also considering the transposonic and recombinogenic nature of the Alu repeats (Tomilin, 1992), it can be conjectured that the inter-Alu repeats discovered by us in the blood serum of pregnant women may play some kind of regulatory role in the early stages of pregnancy. The cloning and sequencing of these fragments is of particular interest. What has been said does not rule out the presence of other inter-Alu repeats in the blood of pregnant women, which can be identified by means of other primers and may have their own features of distribution in the blood in the course of pregnancy.

Bibliography

Belokhvostov A. S., Lebedev S. N., Sherlina S.S. "Changes in the fractional makeup of nucleic acids of the blood serum during radiation injuries", Radiobiologiya, *1987*, vol. 27, No. 4, pp. 505-509.

Vasyukhin V. I., Lipskaya L. A., Tsvetkov A. G., Podgornaya O. I., "DNA coming from human lymphocytes contains sequences homologous to the immunoglobulin gene", Molekulyar. biol., *1991*, vol. 25, No. 2, pp. 405-412.

Vladimirov V. G., Belokhvostov A. S., Sherlina S.S., Vasilyeva I. N., Voskresenskiy A.M., "Content of extracellular DNA in the blood of irradiated rats", Biol. eksperim. biol. med., *1992*, vol. 63, No. 2, pp. 188-190.

Kottsova N. A., Bystrova O. A., Susloparov L. A., Mikhaylov V. M., "Changes in the cellular makeup of the deciduous membrane during the physiological course of pregnancy and one with late toxicosis", Arkh. anat. gistol. embriol., *1989*, vol. 97, No. 11, pp. 76-82.

Mikhaylov V. M., "Ultrastructure of the nuclei during programmed cell death of the decidua", Tsitologiya, *1993*, vol. 35, No. 10, p. 85.

Mikhaylov V. M., Lebedeva G. S., Shteyn G. I., "Changes in the quantity and synthesis of DNA in nuclei of the large decidual cells of the rat in the course of differentiation", Tsitologiya, *1989*, vol. 31, No. 6, p. 677-684.

Mikhaylov V. M., Linde V. A., Rozanov Yu. M., Tatarova N. A., Susloparov L. A., Konycheva Ye. A., "Synthesis and content of DNA in the cells of human decidua during differentiation according to data of the DNA flow cytometry method", Tsitologiya, *1992a*, vol. 34, No. 6, p. 67-73.

Mikhaylov V. M., Rozanov Yu. M., Linde V. A., Tatarova N. A., Bystrova O. A., "The terminal stage of differentiation of cells of the decidua", Abstracts of Papers of the XI Congress of Anatomists, Histologists and Embryologists, Poltava, *1992b*, p. 164.

Tishchenko L. N., Vladimirov V. G., Vasilyeva I. N., "The role of Ca/Mg endonuclease of chromatin in the formation of extracellular DNA of blood", Tsitologiya, *1993*, vol. 35, No. 10, p. 96.

Fedorov N. A., Yaneva N. S., "Expression of DNA by human lymphocytes", Uspekhi sovrem. biol., *1982*, vol. 93, No. 2, pp. 171-182.

Fedorova M. F., Kalashnikova Ye. P., Platsenta i yeye rol' pri beremennosti [The placenta and its role in pregnancy], Meditsina: Moscow, *1986*, 253 pp.

Khanson K. P., Komar V. Ye., Molekulyarnye mekhanizmy radiatsionnoy gibeli kletok [Molecular mechanisms of radiation cell death], Energoatomizdat: Moscow, *1985*, 150 pp.

Anker P., Stroun M., Maurice P. A. Spontaneous release of DNA by human blood lymphocytes as shown in vitro system // Cancer Res. *1975.* Vol. 35. P. 2375—2382.

Dennin R. H. DNA of free and complexed origin in human plasma: concentration and length distribution // Klinische Wochenschrifr. *1979.* Vol. 57. P. 451—456.

Goto M., Kuzashima H., Shibata M., Morishima T., Kimura H., Futamura M., Kito H., Ando T. Detection of viral DNA in neonatal herpes simplex virus infections: frequent and prolonged presence in serum and cerebrospinal fluid // J. Infect. Disease. *1991.* Vol. 164. P. 289—293.

Jelinek W. R., Schmid C. W. Repetitive sequences in eukaryotic DNA and their expression // Ann. Rev. Biochem. *1982.* Vol. 51. P. 813—814.

Leon S. A., Shapiro B., Sklaroff D. M., Yaros M. J. Free DNA in the sera of cancer patients and the effect of therapy // Cancer Res. *1977.* Vol. 37. P. 646—650.

Nelson D., Ladbetter S., Corfo L. Alu polymerase chain reaction: a method for rapid isolation of human-specific sequenses from complex DNA sourses // Proc. Nat. Acad. Sci. USA. *1989.* Vol. 86. P. 6669—6686.

Rogers J. C. Identification of an intracellular precursor of DNA exctracted by human lymphocytes // Proc. Nat. Acad. Sci. USA. *1976.* Vol. 73. P. 3211—3215.

Rogers J. C., Bold D., Kornfeld S., Skinner S. A., Valeri C. L. Excretion of deoxyribonucleic acid by lymphocytes stimulated with phytohemagglutinin or antigen // Proc. Nat. Acad. Sci. USA. *1972.* Vol. 69. P. 1685—1689.

Shapiro B., Chakrabarty M., Cohn E. M., Leon S. A. Determination of circulating DNA levels in patients with benigh or malignant gastrointestinal disease // Cancer. *1983.* Vol. 51. P. 2116—2120.

Stroun M., Anker P., Lyantey J., Lederrey C., Maurice P. A. Isolation and characterisation of DNA from the plasma of cancer patients // Eur. J. Cancer Clin. Oncol. *1987.* Vol. 23. P. 707—712.

Stroun M., Anker P., Maurice P. A., Gahan P. B. Circulating nucleic acids in higher organizms // Int. Rev. Cytol. *1977.* Vol. 51. P. 1—48.

Tomilin N. V. Mechanisms of chromosome destabilization in human cells // Sov. Sci. Rev. D. Physicochem. Biol. *1992.* Vol. 10. P. 39—89.

Van Helden P. D. Potential Z-DNA forming elements in sera DNA from human systemic lupus erythematosus // J. Immunol. *1985.* Vol. 134. P. 177—179.

Wyllie A. H. Glucocorticoid-induced thymocyte apoptosis is associated with endogenous endonuclease activation // Nature. *1980.* Vol. 284. P. 555—556.

Received 6 September 1994

EXTRACELLULAR DNA OF PREGNANT WOMEN BLOOD

V. I. Kazakov, V. M. Bozhkov, V. A. Linde, M. A. Repina, V. M. Mikhailov

Institute of Cytology, Russian Academy of Sciences, and Medical Academy of Postgraduate Education, St. Petersburg

The level of extracellular DNA increases in the blood of women during pregnancy. By means of PCR, the full-size Alu repeats were observed among extracellular blood DNA repeats of pregnant women. Furthermore, with Tc65 type primer the PCR method allowed to observe in the blood DNA fragments flanked by inverted Alu repeats (inter Alu repeats). The presence of such a type of inter Alu repeats was estimated in the blood of women being in the first trimester of pregnancy only, but was not estimated among blood DNA fragments of women of the last trimester of pregnancy. It is discussed which types of cells may serve as a source of extracellular blood DNA (either trophoblast cells, lymphocytes, or decidual cells), the significance of such DNA for pregnancy being appreciated.

To the article by V. I. Kazakov, V. M. Bozhkova, V. A. Linde (p. 232)

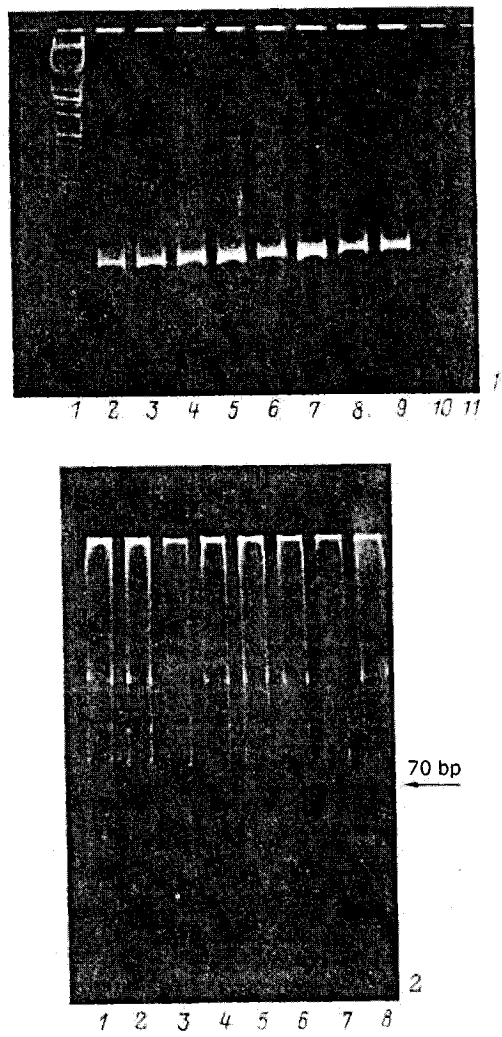


Fig. 1. DNA fragments obtained by carrying out polymerase chain reaction with a pair of primers B1 and C2. Synthesis of a full-length Alu-repeat.

1 – DNA of λ phage digested by PstI; *2* – DNA from blood serum of a male; *3* – DNA from blood serum of a non-pregnant female; *4* – DNA from blood serum of a pregnant woman, I trimester; *5* – DNA from blood serum of a pregnant woman, III trimester; *6* – DNA from blood serum during gestosis; *7* – DNA from a blood thrombus of a pregnant woman, III trimester; *8* – total human DNA; *9* – PalI plasmid with a cloned Alu-repeat; *10* – reaction without template; *11* – reaction without primers.

Fig. 2. DNA fragments obtained by carrying out polymerase chain reaction with the primer Tc65 (inter-Alu-repeat)

1-8 – DNA from blood serum samples of several pregnant women, I trimester.

TSitologiia

ISSN 0041-3771

РОССИЙСКАЯ АКАДЕМИЯ НАУК

ЦИТОЛОГИЯ

CYTOLOGY

1995. Том 37
1995. Vol. 37
3

Univ. of Minn.
Bio-Medical
Library
8 31 95

ОТДЕЛЕНИЕ БИОХИМИИ, БИОФИЗИКИ И ХИМИИ
ФИЗИОЛОГИЧЕСКИ АКТИВНЫХ СОЕДИНЕНИЙ

Журнал «Цитология»

Основан в январе 1959 г.

Выходит 12 раз в год

РЕДАКЦИОННАЯ КОЛЛЕГИЯ:

В. Я. АЛЕКСАНДРОВ, В. Я. БРОДСКИЙ, Ю. М. ВАСИЛЬЕВ,
А. В. ЖИРМУНСКИЙ, И. И. КИКНАДЗЕ,
Е. Е. МАХЛИН (ответственный секретарь),
В. М. МИХЕЛЬСОН (зам. главного редактора),
Н. Н. НИКОЛЬСКИЙ (главный редактор),
А. Л. ПОЛЕНОВ, Г. Д. ТУМАНИШВИЛИ, Ю. С. ЧЕНЦОВ

EDITORIAL BOARD:

V. Ya. ALEXANDROV, V. Ya. BRODSKY, Yu. S. CHENTSOV,
I. I. KIKNADZE, E. E. MAKHLIN (Secretary), V. M. MIKHELSON (Associate Editor),
N. N. NIKOLSKY (Editor-in-Chief), A. L. POLENOV, G. D. TUMANISHVILI,
Yu. M. VASILIEV, A. V. ZHIRMUNSKY

На лицевой стороне обложки: клубеньковая ткань солодки уральской при исследовании с помощью сканирующего электронного микроскопа (см. статью Т. И. Новиковой, Н. Я. Гордиенко в этом номере).

© С.-Петербургская издательская фирма РАН
«Цитология», 1995 г.

This material may be protected by Copyright law (Title 17 U.S. Code)

© 1995

ВНЕКЛЕТОЧНАЯ ДНК В КРОВИ БЕРЕМЕННЫХ ЖЕНЩИН

В. И. Казаков, В. М. Божков, В. А. Линде, М. А. Репина, В. М. Михайлов

Институт цитологии РАН и Медицинская академия последипломного образования врачей, Санкт-Петербург

Показано увеличение концентрации внеклеточной ДНК в крови беременных женщин, наиболее выраженное при гестозе, что может быть связано с гибелью децидуальных клеток, происходящей по механизмам апоптоза. В сыворотке крови женщин I и III триместров беременности при помощи полимеразной цепной реакции обнаружена ДНК, содержащая нефрагментированные Alu-повторы. Кроме того, при помощи праймера Tc65 только в сыворотке крови женщин I триместра беременности обнаружены фрагменты ДНК, фланкированные инвертированными Alu-повторами (интер-Alu-повторами). Обсуждаются возможности экскреции интер-Alu-повторов клетками плода (клетками трофобласта) и матки (клетками эндометрия и лимфоцитами) и регуляторной функции такой ДНК на ранних сроках беременности.

Показано, что в крови человека и животных содержится внеклеточная ДНК (Stroun et al., 1977; Федоров, Янева, 1982; Владимиров и др., 1992). Увеличение концентрации внеклеточной ДНК в крови у людей описано при патологических процессах, протекающих в различных тканях организма, в частности при некоторых воспалительных процессах желудочно-кишечного тракта, при опухолевых и инфекционных заболеваниях вирусной этиологии, при дессиминированной волчанке (Anker et al., 1975; Leon et al., 1977; Shapiro et al., 1983; Stroun et al., 1987; Goto et al., 1991).

Мол. масса внеклеточной ДНК крови в среднем составляет от $1 \cdot 10^6$ до $15 \cdot 10^6$ Да (Stroun et al., 1977; Федоров, Янева, 1982). По данным ЭМ-микроскопии внеклеточная ДНК крови здоровых людей является двухцепочечной и линейной (Dennin, 1979). По данным Васюхина и соавторов (1991), внеклеточная ДНК человека содержит уникальные последовательности. Также показано, что при дессиминированной волчанке во внеклеточной ДНК крови имеются последовательности, способные образовывать Z-ДНК (Van Helder, 1985).

Считается, что высокомолекулярный компонент внеклеточной ДНК крови происходит из живых клеток (Stroun et al., 1977). Показано in vitro, что некоторые типы клеток, особенно лимфоциты, экскретируют в среду высокомолекулярную ДНК (Rogers et al., 1972; Rogers, 1976; Stroun et al., 1977; Федоров, Янева, 1982). Кроме того, в крови содержится также низкомолекулярная ДНК, соответствующая по подвижности нуклеосомам. В крови крыс ее содержание возрастает после общего рентгеновского облучения (Белохвостов и др., 1987; Владимиров и др., 1992; Тищенко и др., 1993). По данным Тищенко и соавторов (1993), низкомолекулярная ДНК крови крыс после общего облучения обогащена ГЦ-последовательностями. Высказана точка зрения о том, что низкомолекулярная ДНК крови есть продукт усиленного экстрахромосомного синтеза кольцевых ДНК, обнаруженных у многих типов клеток человека и млекопитающих (Владимиров и др., 1992). Наше внимание привлекло предположение о том, что повышение уровня внеклеточной низкомо-

лекулярной ДНК в крови крыс после облучения есть следствие усиления активности в клетках Ca/Mg-эндонуклеазы (Тищенко и др., 1993). Это предположение хорошо согласуется с данными об усилении процессов апоптоза в клетках различных тканей организма после облучения (Хансон, Комар, 1985).

Из сказанного ясно, что анализ внеклеточной ДНК крови человека и животных имеет как теоретический, так и практический интерес. Не исключено, что нуклеотидный состав внеклеточной ДНК крови не настолько случаен, чтобы в нем не отразились особенности процессов дифференцировки и клеточной гибели, протекающей в разных тканях в каждый отдельный момент жизни организма. Все это определило выбор в качестве объекта для настоящего исследования крови беременных женщин. По имеющимся данным, в матке при беременности имеют место клеточная пролиферация, дифференцировка и гибель клеток (Федорова, Калашникова, 1986; Михайлов и др., 1989, 1992а, 1992б; Михайлов, 1993). Ожидалось, что эти процессы окажут свое влияние на специфику нуклеотидного состава внеклеточной ДНК крови беременных женщин.

Материал и методика

Исследовали сыворотки крови мужчин, небеременных женщин, женщин I и III триместров беременности и с поздним токсикозом беременности. Кровь брали шприцем из локтевой вены в стерильных условиях, помещали в центрифужную пробирку и оставляли при комнатной температуре до свертывания. Сразу после образования тромба его отделяли от стенок пробирки и центрифугировали при 400*g* в течение 10 мин. Сыворотку центрифугировали еще раз при 2000*g* в течение 10 мин при 4 °С. Полученную таким образом сыворотку хранили при −60 °С. Использование сыворотки вместо плазмы для анализа ДНК крови может быть оправдано, если соблюдены условия формирования тромба при комнатной температуре и немедленного отделения сыворотки от тромба (Leon et al., 1977; Shapiro et al., 1983). Это и было выполнено в настоящем исследовании.

В дальнейшем сыворотку двукратно обрабатывали фенолом, смесями фенола и хлороформа (1 : 1), хлороформа и изоамилового спирта (24 : 1) и осаждали этанолом при −20 °С. Препараты ДНК анализировали в 1 %-ном агарозном геле, а затем использовали в качестве матрицы при проведении полимеразной цепной реакции (ПЦР).

ПЦР проводили с использованием термофильной ДНК-полимеразы из Thermus thermophilus. Осуществляли 30—35 циклов амплификации. Амплифицируемые фрагменты анализировали в 8 %-ном ПААГ. Температура отжига праймеров составляла 55 °С для праймеров B1 и C2 и 60 °С для праймера Tc65, концентрация ионов Mg^{2+} — 2—5 mM. Праймеры для ПЦР были синтезированы нами твердофазным фосфиттриэфирным методом в β-цианоэтильной модификации на приборе Gene assembler фирмы Pharmacia. Последовательность праймера Tc65 взята из литературных источников (Nelson et al., 1989), последовательности остальных праймеров найдены при помощи программы Oligo (Microsoft) по консенсуальной последовательности Alu-повторов MF семейства и обозначены нами как B1 и C2.

При проведении ПЦР с парой праймеров B1 и C2 в результате реакции амплифицировался Alu-повтор длиной 239 пар оснований. При использовании в ПЦР только одного праймера Tc65 амплифицировались фрагменты ДНК, фланкированные двумя Alu-повторами, обращенными друг к другу 3′-концевыми областями (интер-Alu-повторами).

Последовательности праймеров:

Tc65—5′—AAGTCGCGGCCGCTTGCAGTGAGCCGAGA—3′

B1—5′—CCTGTAATCCCAGCACTTTGGGAGGC—3′

C2—5′—CCCAGGC$\begin{matrix}\mathrm{T}\\\mathrm{G}\end{matrix}$GGAGTGC$\begin{matrix}\mathrm{A}\\\mathrm{G}\end{matrix}$ GTGG—3′

Результаты и обсуждение

В работе мы не ставили целью детальное изучение изменений концентраций ДНК в сыворотке крови беременных женщин. По данным авторов, использующих радиоиммунологические методы определения (Leon el al., 1977; Shapiro et al., 1983), концентрация ДНК в сыворотках крови здоровых доноров колебалась от 0 до 0.1 мкг/мл; по данным авторов, использующих биохими-

ческие методы (Anker et al., 1975; Stroun et al., 1987), — от 0.1 до 0.6 мкг/мл. Наши данные о концентрации ДНК в крови мужчин и небеременных женщин соответствовали литературным.

По полученным нами данным, при беременности в первую очередь увеличивается концентрация низкомолекулярной ДНК, причем наиболее выражено увеличение при гестозе. Размеры такой ДНК составляют от 150 до 2500 пар оснований. Существует точка зрения о том, что увеличение концентрации низкомолекулярной внеклеточной ДНК в крови является следствием усиления синтеза экстрахромосомной ДНК (Владимиров и др., 1992). Однако при оценке представленных данных необходимо учитывать, что при беременности гибнет значительное количество клеток эндометрия, особенно в конце беременности и в связи с гестозом (Котцова и др., 1989; Михайлов и др., 1992а). По нашим данным, гибель децидуальных клеток сопровождается постепенной потерей ядерной ДНК и фрагментацией периферических частей ядер и цитоплазмы клеток и, по-видимому, происходит путем апоптоза (Михайлов и др., 1992а, 1992б; Михайлов, 1993). Одним из ведущих механизмов гибели клеток по типу апоптоза является межнуклеосомная деградация ДНК вследствие активации Ca^{2+}-эндонуклеазы (Wyllie, 1980). Сказанное не исключает возможности экстрахромосомного копирования ДНК как источника внеклеточной ДНК крови. Независимо от того, является ли источником внеклеточной ДНК крови усиленный синтез экстрахромосомной ДНК или апоптическая гибель клеток, наличие в ней Alu-повторов свидетельствует о ее ядерном происхождении (рис. 1; см. вкл. VIII).

Обращает на себя внимание присутствие среди ДНК сыворотки крови беременных женщин интер-Alu-повторов. Известно, что среди Alu-повторов человека примерно третья часть принадлежит к инвертированным повторам, и среднее расстояние между инвертированными парами составляет примерно 50 т. п. н. (Jelinek, Schmid, 1982). Основная масса таких инвертированных повторов содержит, по-видимому, уникальный внутренний фрагмент. Существует несколько видов интер-Alu-повторов. При помощи праймера Tc65 мы обнаружили интер-Alu-повторы в крови у женщин только I триместра беременности (рис. 2). При проведении ПЦР с праймером Tc65 амплифицируются только те фрагменты ДНК, в которых спейсерная ДНК фланкирована Alu-повторами, взаимно обращенными друг к другу 3'-областями, что делает эти фрагменты еще более уникальными. Таким образом, если наличие в крови беременных женщин увеличенного количества низкомолекулярных фрагментов ДНК является проявлением усиленной клеточной гибели, имеющей место в матке на всех сроках беременности, то присутствие интер-Alu-повторов указывает на функционирование других механизмов поступления таких последовательностей в кровоток. Важно, что интер-Alu-повторы были обнаружены только в крови женщин I триместра беременности. Этот факт, скорее всего, отражает различие в содержании клеточных процессов, характерных для ранних и поздних стадий беременности. Если в конце беременности наблюдается гибель децидуальных клеток, чем объясняется увеличение концентрации низкомолекулярной ДНК в крови, то в начальном периоде беременности, когда происходят имплантация и плацентация зародыша в матке, имеют место децидуализация эндометрия и рост плодного пузыря, т. е. процесс дифференцировки клеток слизистой матки и клеток трофобласта.

Вероятно, на ранних стадиях беременности происходит экскреция интер-Alu-повторов из клеток. Представляет большой интерес вопрос о том, какого типа клетки могут экскретировать интер-Alu-повторы. Этот вопрос также встает перед исследователями при обсуждении источников высокомолекулярной сывороточной ДНК (см. обзор; Федоров, Янева, 1982). Экскреция ДНК in vitro хорошо охарактеризована для лимфоцитов (Федоров, Янева, 1982). Предполагается, что при опухолевых заболеваниях источником такой ДНК могут

быть лимфоциты. На ранних стадиях беременности в слизистой матки также наблюдается значительная лимфоцитарная инфильтрация.

Таким образом, на ранних стадиях беременности у человека экскретировать ДНК могут клетки плода (трофобласты) и материнского организма (клетки эндометрия и лимфоциты). С учетом вышесказанного, а также с учетом транспозонной и рекомбиногенной природы Alu-повторов (Tomilin, 1992) можно предположить, что обнаруженные нами в сыворотке крови беременных женщин интер-Alu-повторы могут выполнять какую-то регуляторную функцию на ранних сроках беременности. Особый интерес представляют клонирование и секвенирование этих фрагментов. Сказанное не исключает присутствие в крови беременных женщин иных интер-Alu-повторов, которые могут быть выявлены при помощи других праймеров и иметь собственные особенности распределения в крови в течение беременности.

Список литературы

Белохвостов А. С., Лебедев С. Н., Шерлина С. С. Изменения фракционного состава нуклеиновых кислот сыворотки крови при радиационных поражениях // Радиобиология. *1987*. Т. 27, № 4. С. 505—509.

Васюхин В. И., Липская Л. А., Цветков А. Г., Подгорная О. И. ДНК, выходящая из лимфоцитов человека, содержит последовательности, гомологичные иммуноглобулиновому гену // Молекуляр. биол. *1991*. Т. 25, № 2. С. 405—412.

Владимиров В. Г., Белохвостов А. С., Шерлина С. С., Васильева И. Н., Воскресенский А. М. Содержание внеклеточной ДНК в крови облученных крыс // Бюл. эксперим. биол. мед. *1992*. Т. 63, № 2. С. 188—190.

Котцова Н. А., Быстрова О. А., Суслопаров Л. А., Михайлов В. М. Изменения клеточного состава отпадающей оболочки при физиологически протекающей беременности и с поздним токсикозом // Арх. анат. гистол. эмбриол. *1989*. Т. 97, № 11. С. 76—82.

Михайлов В. М. Ультраструктура ядер при программированной гибели клеток децидуа // Цитология. *1993*. Т. 35, № 10. С. 85.

Михайлов В. М., Лебедева Г. С., Штейн Г. И. Изменения количества и синтеза ДНК в ядрах больших децидуальных клеток крыс в ходе дифференцировки // Цитология. *1989*. Т. 31, № 6. С. 677—684.

Михайлов В. М., Линде В. А., Розанов Ю. М., Татарова Н. А., Суслопаров Л. А., Конычева Е. А. Синтез и содержание ДНК в клетках децидуа человека при дифференцировке по данным метода проточной ДНК-цитометрии // Цитология. *1992а*. Т. 34, № 6. С. 67—73.

Михайлов В. М., Розанов Ю. М., Линде В. А., Татарова Н. А., Быстрова О. А. Терминальная стадия дифференцировки клеток децидуа // Тез. докл. XI съезда анатомов, гистологов и эмбриологов. Полтава, *1992б*. С. 164.

Тищенко Л. Н., Владимиров В. Г., Васильева И. Н. Роль Ca/Mg-эндонуклеазы хроматина в образовании внеклеточной ДНК крови // Цитология. *1993*. Т. 35, № 10. С. 96.

Федоров Н. А., Янева Н. С. Экскреция ДНК лимфоцитами человека // Успехи соврем. биол. *1982*. Т. 93, № 2. С. 171—182.

Федорова М. Ф., Калашникова Е. П. Плацента и ее роль при беременности. М.: Медицина, *1986*. 253 с.

Хансон К. П., Комар В. Е. Молекулярные механизмы радиационной гибели клеток. М.: Энергоатомиздат, *1985*. 150 с.

Anker P., Stroun M., Maurice P. A. Spontaneous release of DNA by human blood lymphocytes as shown in vitro system // Cancer Res. *1975*. Vol. 35. P. 2375—2382.

Dennin R. H. DNA of free and complexed origin in human plasma: concentration and length distribution // Klinische Wochenschrifr. *1979*. Vol. 57. P. 451—456.

Goto M., Kuzashima H., Shibata M., Morishima T., Kimura H., Futamura M., Kito H., Ando T. Detection of viral DNA in neonatal herpes simplex virus infections: frequent and prolonged presence in serum and cerebrospinal fluid // J. Infect. Disease. *1991*. Vol. 164. P. 289—293.

Jelinek W. R., Schmid C. W. Repetitive sequences in eukaryotic DNA and their expression // Ann. Rev. Biochem. *1982*. Vol. 51. P. 813—814.

Leon S. A., Shapiro B., Sklaroff D. M., Yaros M. J. Free DNA in the sera of cancer patients and the effect of therapy // Cancer Res. *1977*. Vol. 37. P. 646—650.

Nelson D., Ladbetter S., Corfo L. Alu polymerase chain reaction: a method for rapid isolation of human-specific sequenses from complex DNA sourses // Proc. Nat. Acad. Sci. USA. *1989*. Vol. 86. P. 6669—6686.

Rogers J. C. Identification of an intracellular precursor of DNA exctracted by human lymphocytes // Proc. Nat. Acad. Sci. USA. *1976*. Vol. 73. P. 3211—3215.

Rogers J. C., Bold D., Kornfeld S., Skinner S. A., Valeri C. L. Excretion of deoxyribonucleic acid by lymphocytes stimulated with phytohemagglutinin or antigen // Proc. Nat. Acad. Sci. USA. *1972.* Vol. 69. P. 1685—1689.

Shapiro B., Chakrabarty M., Cohn E. M., Leon S. A. Determination of circulating DNA levels in patients with benigh or malignant gastrointestinal disease // Cancer. *1983.* Vol. 51. P. 2116—2120.

Stroun M., Anker P., Lyantey J., Lederrey C., Maurice P. A. Isolation and characterisation of DNA from the plasma of cancer patients // Eur. J. Cancer Clin. Oncol. *1987.* Vol. 23. P. 707—712.

Stroun M., Anker P., Maurice P. A., Gahan P. B. Circulating nucleic acids in higher organizms // Int. Rev. Cytol. *1977.* Vol. 51. P. 1—48.

Tomilin N. V. Mechanisms of chromosome destabilization in human cells // Sov. Sci. Rev. D. Physicochem. Biol. *1992.* Vol. 10. P. 39—89.

Van Helden P. D. Potential Z-DNA forming elements in sera DNA from human systemic lupus erythematosus // J. Immunol. *1985.* Vol. 134. P. 177—179.

Wyllie A. H. Glucocorticoid-induced thymocyte apoptosis is associated with endogenous endonuclease activation // Nature. *1980.* Vol. 284. P. 555—556.

Поступила 6 IX 1994

EXTRACELLULAR DNA OF PREGNANT WOMEN BLOOD

V. I. Kazakov, V. M. Bozhkov, V. A. Linde, M. A. Repina, V. M. Mikhailov

Institute of Cytology, Russian Academy of Sciences, and Medical Academy of Postgraduate Education, St. Petersburg

The level of extracellular DNA increases in the blood of women during pregnancy. By means of PCR, the full-size Alu repeats were observed among extracellular blood DNA repeats of pregnant women. Furthermore, with Tc65 type primer the PCR method allowed to observe in the blood DNA fragments flanked by inverted Alu repeats (inter Alu repeats). The presence of such a type of inter Alu repeats was estimated in the blood of women being in the first trimester of pregnancy only, but was not estimated among blood DNA fragments of women of the last trimester of pregnancy. It is discussed which types of cells may serve as a source of extracellular blood DNA (either trophoblast cells, lymphocytes, or decidual cells), the significance of such DNA for pregnancy being appreciated.

К ст. В. И. Казакова, В. М. Божкова, В. А. Линде (с. 232)

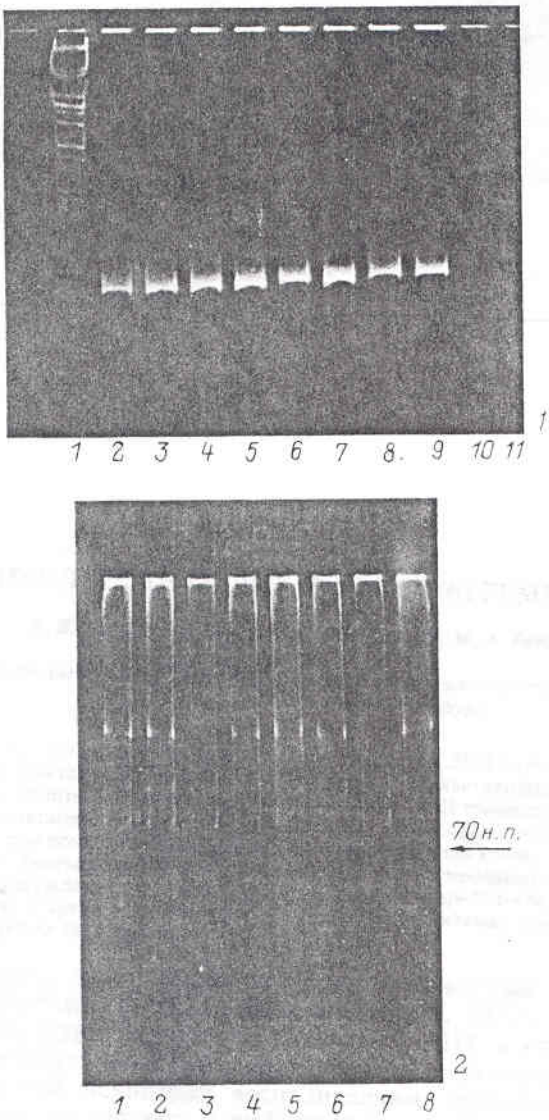


Рис. 1. Фрагменты ДНК, полученные в результате проведения полимеразной цепной реакции с парой праймеров B1 и C2. Синтез полноразмерного Alu-повтора.

1 — ДНК фага λ, рестрицированная по PstI; *2* — ДНК из сыворотки крови мужчин; *3* — ДНК из сыворотки крови небеременной женщины; *4* — ДНК из сыворотки крови женщины I триместра беременности; *5* — ДНК из сыворотки крови женщины III триместра беременности; *6* — ДНК из сыворотки крови при гестозе; *7* — ДНК из тромба крови женщин III триместра беременности; *8* — тотальная ДНК человека; *9* — плазмида PaII с клонированным Alu-повтором; *10* — реакция без матрицы; *11* — реакция без праймера.

Рис. 2. Фрагменты ДНК, полученные в результате проведения полимеразной цепной реакции с праймером Tc65 (интер-Alu-повтор).

1—8 — ДНК из сыворотки крови различных женщин I триместра беременности.

This material may be protected by Copyright law (Title 17 U.S. Code)

Sensitive Detection of Circulating Hepatocellular Carcinoma Cells in Peripheral Venous Blood

Toshiki Komeda, M.D., Yoshihiro Fukuda, M.D., Ph.D., Takehiro Sando, M.D., Ryuichi Kita, M.D., Motonobu Furukawa, M.D., Naoshi Nishida, M.D., Ph.D., Masahiro Amenomori, M.D., and Kazuwa Nakao, M.D., Ph.D.

Background. **This study was performed to develop a sensitive method for the detection of circulating hepatocellular carcinoma (HCC) cells in peripheral blood, in advance of the diagnosis of distant metastasis of HCC by conventional means.**

Methods. **Peripheral blood (5 ml) samples were obtained from 64 patients with HCC and from 48 control subjects (31 patients with benign liver disease, 8 with metastatic liver cancer, and 9 with normal liver function). To identify HCC cells in peripheral blood, liver-specific human alpha-fetoprotein (hAFP) mRNA was amplified from total RNA extracted from whole blood by reverse transcriptase-polymerase chain reaction.**

Results. **Human alpha-fetoprotein mRNA was detected in 23 blood samples from the HCC patients (23/64, 36%), in 17 patients in whom there was no clinical evidence of distant metastasis. In contrast, there were no control patients whose samples showed detectable hAFP mRNA in the peripheral blood. The presence of hAFP mRNA in blood seemed to be correlated with the stage (by TNM classification) of HCC, the serum hAFP value, and the presence of intrahepatic metastasis, portal vein thrombosis, and/or distant metastasis.**

Conclusions. **Reverse-transcriptase polymerase chain reaction is a very sensitive method for detecting circulating HCC cells. With this technique, important information for the management of HCC can be acquired, such as the indications for orthotopic liver transplantation in HCC patients. Moreover, use of this detection method may encourage investigation of the mechanism of metastasis in HCC. *Cancer* 1995;75:2214–9.**

Key words: hepatocellular carcinoma, human alpha-fetoprotein, mRNA, peripheral blood, circulating cancer cells, metastasis, reverse transcriptase-polymerase chain reaction.

Recent progress in the treatment of patients with chronic liver disease has led to a better quality of life for these patients, even those with severe cirrhosis. Nevertheless, hepatocellular carcinoma (HCC) is still a major cause of death in patients with chronic liver disease. One of the difficulties in the management of HCC is its extensive character; intrahepatic metastasis, venous invasion, and distant metastasis. Some patients with HCC who receive curative resection or chemoembolization develop metastatic nodules shortly after these therapies. Furthermore, orthotopic liver transplantation for unresectable HCC yields disappointing results because of a high recurrence rate.[1] In these events, it is assumed that HCC cells have been disseminated in the systemic circulation, and proliferate in an immunosuppressive environment at and after therapy. To improve the prognosis of HCC, it is important to detect such blood-borne circulation of HCC cells before therapy is instituted. Although it has been reported that dislodged HCC cells were identified cytologically in portal venous blood during surgical procedures,[2] this method seems inappropriate to identify the small number of HCC cells in the systemic circulation.

There have been some reports of the employment of the reverse transcriptase-polymerase chain reaction (RT-PCR) technique to detect tumor cells in peripheral blood,[3–8] bone marrow,[9] and lymph nodes.[10] These trials have aimed to amplify tumor (or tissue)-specific gene transcripts that are not detected in these tissues under normal conditions. Free mRNA is so fragile under

Presented in part at the biennial meeting of the International Association for the Study of the Liver, Cancun, Mexico, May 21, 1994.

Supported in part by a Grant-in-Aid from the Ministry of Education of Japan.

From the Department of Medicine and Clinical Science, Kyoto University Graduate School of Medicine, Kyoto, Japan.

Address for reprints: Yoshihiro Fukuda, M.D., Ph.D., Department of Medicine and Clinical Science, Kyoto University Graduate School of Medicine, 54, Kawahara-cho, Shogoin, Sakyo-ku, Kyoto, 606, Japan.

Received October 5, 1994; revision received January 12, 1995; accepted January 12, 1995.

Exhibit S

conditions of abundant RNAse activity that the specific mRNA in blood will indicate the presence of intact cells producing such proteins just before the extraction of RNA, not circulating free mRNA.

The human alpha-fetoprotein (hAFP) gene is known to be transcribed prominently in HCC, but weakly in normal adult hepatocytes.[11,12] Thus, we investigated the presence of hAFP mRNA in the peripheral blood of patients with HCC and controls. Specific polymerase chain reaction (PCR) products were obtained in the samples of some patients with HCC, especially in advanced cases. We describe here the clinical applications of this technique for the detection of circulating HCC cells and for use in the selection of therapy for HCC.

Materials and Methods

Samples

Hep G2 (hepatoblastoma) (RIKEN Cell Bank, Tsukuba, Japan) and PLC/PRF/5 (HCC) cell lines served as positive controls for hAFP mRNA expression. For negative controls, the following cell lines, derived from other human tissues, were prepared; KHC (lung cancer),[13] JK1 (chronic myelogenous leukemia),[14] MM-M1 (multiple myeloma),[15] RAJI (malignant lymphoma), and HL60 (acute promyelocytic leukemia). The cell lines were provided by the Japanese Cancer Research Resources Bank (Tokyo, Japan) unless otherwise noted.

Patients were divided into four groups according to their clinical status: group A, patients with HCC diagnosed by histology, imaging, or other means (n = 64); group B, patients with chronic or acute liver diseases but no evidence of HCC (n = 31); group C, patients with metastatic liver cancer (n = 8); and group D, healthy volunteers and patients without liver disease (n = 9). The staging of HCC in group A was carried out according to the TNM classification of the International Union Against Cancer.[16] Five ml of whole blood was withdrawn from a peripheral vein of each subject into a heparinized tube. As a rule, these materials were sampled for RT-PCR or cytologic study before any invasive therapies were performed. Brief information on each group is given in Table 1.

For 29 patients of group A whose blood did not show hAFP mRNA, we examined their venous blood after subsequent invasive therapies. Peripheral blood samples were obtained 3 days after transcatheter arterial embolization (25 patients) and 30 minutes after percutaneous ethanol injection (4 patients). The same procedures were performed upon these materials.

Detection of hAFP mRNA

Heparinized whole blood was centrifuged and the plasma fraction was removed. The cellular fraction was enriched for mononuclear cells (MNC) or possible tumor cells according to the method described by Oppenheim.[17] Total cellular RNA was then extracted with TRIzol (Life Technologies Inc., Gaithersburg, USA),[18] from 5 μg of which cDNA was synthesized. The reverse transcription reaction solution contained 6 μl of 5 $\times$ first strand buffer, 10 mM dithiothreitol, 1.25 mM each of dCTP, dATP, dGTP, and dTTP, 0.3 μg of random hexamers, and 200 units of SuperscriptII Moloney murine leukemia virus reverse transcriptase (Life Technologies Inc.). The RNA solution was incubated at 65°C for 10 minutes, mixed with the reverse transcription reaction solution (total volume 30 μl), and incubated at 37°C for 60 minutes. The PCR primers for hAFP used in this study are shown in Table 2. The first PCR reaction solution contained 5 μl of the synthesized cDNA solution, 10 μl of 10 $\times$ polymerase reaction buffer, 500 μM each of dCTP, dATP, dGTP, and dTTP, 15 pmol of each external primer (EX-sense and EX-antisense), 4 units of Delta Thermus thermophilus DNA polymerase (TOYOBO, Osaka, Japan), and water. The PCR cycles were: denaturation at 94°C for 1 minute, annealing at 52°C for 1 minute, and primer extension at 72°C for 1 minute; the cycles were repeated 40 times. The PCR product was reamplified with internal primers for nested PCR to obtain higher sensitivity. The first and second PCR components were the same, but for the primer pairs (IN-sense and IN-antisense). The final product was electrophoresed on 1.8% agarose gel and stained with ethidium bromide for the specific band of 101 base pairs.

To examine whether free mRNA in blood could also be detected and result in occurrence of false positivity, 10 μg of total RNA from Hep G2 cells was diluted with 5 ml of whole blood from healthy volunteers. Also to determine the sensitivity of this system, a dilution series was prepared; 1.5×10^4 Hep G2 cells were mixed with 5 ml of control blood and diluted sequentially to 10^{-5} concentration.

Cytologic Study

Buffy coated fractions of peripheral blood from patients with HCC were smeared onto microscopic slides according to a method described previously.[2] The fractions were stained with hematoxylin and eosin and examined microscopically.

Table 1. Profiles of Groups

	No. of patients	Sex (M/F)	Age* (yr)	Serum hAFP† (ng/ml)
Group A	64			
Stage I	9	7/2	60.5 ± 4.3	12.5 (5.5–44.0)
Stage II	23	17/6	61.1 ± 1.9	41.0 (3.0–1189.0)
Stage III	10	7/3	66.9 ± 1.4	80.5 (6.2–3590.0)
Stage IV‡	22	13/9	63.7 ± 1.8	453.0 (6.3–527,309.0)
Group B	31	15/16	58.1 ± 1.6	11.0 (3.0–477.0)
Group C	8	4/4	57.0 ± 4.8	6.0 (3.0–104.0)
Group D	9	4/5	34.1 ± 3.8	Not done

hAFP: human α-fetoprotein.

* Patient's age is expressed as mean ± SE.

† Serum hAFP value is expressed as median (range).

‡ Six patients had distant or lymph node metastases, and all of them were in Stage IV.

Statistical Analysis

The relationship between the presence of hAFP mRNA in peripheral blood and various clinical parameters was examined by chi-square test.

Results

mRNA from Cell Lines

With cDNA derived from Hep G2 and PLC/PRF/5 cells, which cells are known to produce hAFP,[20,21] specific bands for hAFP (101 bp) were observed (data not shown). In contrast, the cell lines that served as negative controls did not yield such bands. It was also impossible to detect free RNA extracted from Hep G2 cells when they were diluted once with control blood.

In the limiting dilution series, 15 Hep G2 cells suspended in 5-ml aliquots of control blood proved to be sufficient for detection (Fig. 1). Although our assay was nonstable when such very sparse samples were examined, good reproducibility was obtained from identical materials.

Table 2. Setting of Primer Pairs

EX-sense	5′-ACTGAATCCAGAACACTGCATAG-3′
EX-antisense	5′-TGCAGTCAATGCATCTTTCACCA-3′
IN-sense	5′-TGGAATAGCTTCCATATTGGATTC-3′
IN-antisense	5′-AAGTGGCTTCTTGAACAAACTGG-3′

PCR products of 176 and 101 base pairs were amplified from hAFP cDNA by external (EX-sense and EX-antisense) and internal (IN-sense and IN-antisense) primer pairs, respectively. The primer pairs were located as follows: EX-sense in exon 1 (nucleotides 90–112), EX-antisense in exon 2 (nucleotides 263–241), IN-sense over exon 1 and exon 2 (nucleotides 122–145), in IN-antisense in exon 3 (nucleotides 222–200). cDNA sequences were based upon those previously reported.[19]

mRNA from Clinical Samples

We detected hAFP mRNA in 23 of the group A samples. The frequency of positive cases in each stage was: 1 of 9 (11%) in Stage I, 2 of 23 (9%) in Stage II, 3 of 10 (30%) in Stage III, and 17 of 22 (77%) in Stage IV. These results showed strong correlation with clinical severity. hAFP mRNA positivity was significantly more frequent in patients with large (exceeding 5 cm) tumors (88%) than in those with small (under 5 cm) tumors (17%) ($P < 0.001$). The frequency of hAFP mRNA positivity was also increased in patients with HCC with intrahepatic metastatic nodules ($P < 0.001$), portal thrombosis ($P < 0.001$), and extrahepatic metastasis ($P < 0.001$) (Table 3). Further, there were more positive cases in patients with high serum hAFP values (exceeding 300 ng/ml, $P = 0.014$) (Fig. 2). With regard to controls, none of the samples from groups B, C, or D showed hAFP mRNA in their peripheral blood.

To confirm the accuracy of this method in clinical cases, we repeated 35 clinical samples independently, 33 of which (94%) showed identical results.

Cytologic Study

We selected two patients whose blood samples were positive for hAFP mRNA and performed a cytologic study of the samples. Atypical cells with large nuclei were found in some samples of the smeared peripheral blood of only one of these patients (data not shown). We examined two other patients in whom hAFP mRNA was undetectable, and found no such atypical cells.

Alterations after Therapy

Eight of 29 (6/25 patients who received transcatheter arterial embolization and 2/4 who received percutane-

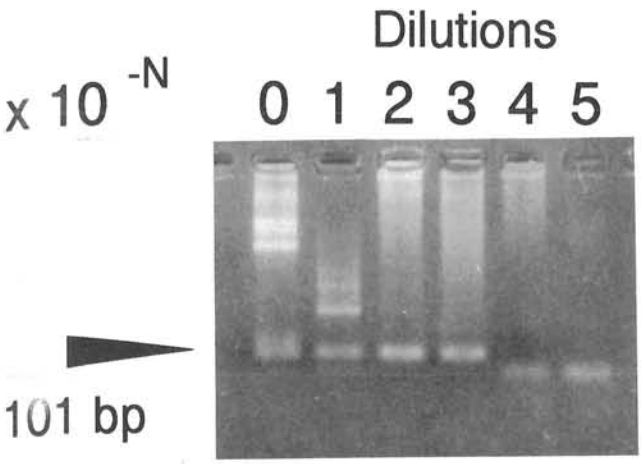


Figure 1. Hep G2 cells (1.5×10^4) were suspended in 5 ml of whole blood obtained from healthy volunteers and sequentially diluted with blood, 10 times, from 10^{-1} to 10^{-5} concentration. Human alpha-fetoprotein cDNA-specific PCR products (101 bp) can be observed at dilutions from 10^0 to 10^{-3}, as indicated.

ous ethanol injection) patients in group A who underwent invasive therapies, and whose samples were originally negative for hAFP mRNA later showed detectable hAFP mRNA in the peripheral blood after these treatments. Six of these patients were examined again 7 days after the treatment; five patients returned to negative and one patient remained positive. The latter patient later showed progression of HCC. Thus far, we have been unable to identify any specific clinical profiles or therapies that would have induced such conversions.

Discussion

The early diagnosis and treatment of HCC now available have improved the prognosis of this disease. However, intrahepatic or distant metastases are detected frequently, even during the course of therapy. Many therapies, including orthotopic liver transplantation, cannot bring about complete remission because of the very extensive character of HCC. If metastasis in the early stage can be determined sensitively, we should be able to select more beneficial therapeutic methods. Moreover, investigation of the metastatic mechanism will be enhanced by comparing the potential for metastasis with the characteristics of the primary tumor. Apart from a cytologic method,[2] which is specific but lacks in sensitivity, there has, thus far, been no way to determine the presence of HCC cells in blood. To this end, we attempted to establish a sensitive method for the detection of circulating HCC cells.

PCR is a very sensitive technique that detects minute volumes of DNA copies. Some researchers have recently combined PCR with reverse transcription and used this method to detect micrometastases of malignant melanoma,[3,6,7] neuroblastoma,[4,9] and prostatic cancer.[5,10] A recent study of HCC has used the human albumin gene as a target.[8] The authors of these studies argued that RT-PCR was useful to detect tissue-specific transcripts from tissues in which the expression of the examined gene was not expected. The results for prostatic cancer suggested that tumor cells were present in the peripheral blood of patients at an early stage in which distant metastasis was not confirmed.[5] In this study, we attempted to find small numbers of cancer cells in the peripheral blood of patients with HCC.

First, we examined the sensitivity of our system with RNA extracted from cultured HCC cells. hAFP-specific PCR products were obtained from 5 μg of these samples without false negative results. Limiting dilution examination revealed that 15 Hep G2 cells suspended in 5 ml of peripheral blood could be detected. This implies that, with our detection system, a single tumor cell among 10^6–10^7 normal MNC can be identified. Of course, there is a chance that hAFP gene may also be transcribed in MNC,[22] but this good reproducibility of the dilution series and repeated study of clini-

Table 3. Relationship Between Detection of hAFP mRNA in the Blood and Various Clinical Parameters

Clinical parameter	No. of cases in which hAFP mRNA was detectable/total no. of cases (%)	
Stage of HCC		
I	1/9 (11)	
II	2/23 (9)	
III	3/10 (30)	
IV	17/22 (77)	
Tumor size		
≥5 cm	15/17 (88)	$P < 0.001$
<5 cm	8/47 (17)	
Serum hAFP value*		
≥300 ng/ml	16/20 (80)	$P = 0.014$
<300 ng/ml	7/44 (16)	
Intrahepatic metastasis		
Present	20/34 (59)	$P < 0.001$
Absent	3/30 (10)	
Portal vein thrombosis		
Present	10/10 (100)	$P < 0.001$
Absent	13/54 (24)	
Distant metastasis		
Present	6/6 (100)	$P < 0.001$
Absent	17/58 (29)	

hAFP: human α-fetoprotein; HCC: hepatocellular carcinoma.

* Serum hAFP value was 50,231.5 ± 27,082.1 (mean ± SE) ng/ml in the group in which hAFP mRNA was detected in the blood and 109.8 ± 37.4 ng/ml in the group in which hAFP mRNA was not detected in the blood.

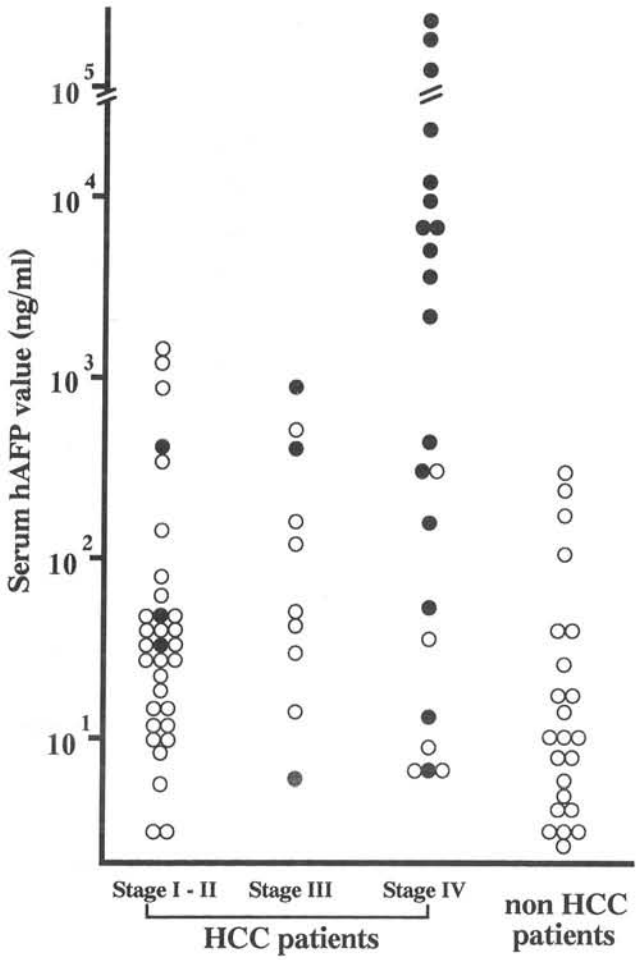


Figure 2. Relationship between the presence of hAFP mRNA in the blood and serum hAFP level and tumor stage in patients with HCC. Patients in whom hAFP mRNA was detected in the blood are indicated by closed circles (●), and undetectable cases are indicated by open circles (○). Detectable cases were more frequent in the advanced stages. The frequency of detectable cases was increased in those patients whose serum hAFP value exceeded 300 ng/ml.

cal samples mean that mRNA from MNC plays no significant role in our assay results. Further, the finding of atypical cells microscopically in patients with detectable hAFP mRNA in the peripheral blood supports this conclusion.

Further, RNA extracted from Hep G2 cells suspended in control blood did not yield hAFP-specific PCR products, indicating that free mRNA released from the liver and circulated in blood would not produce false-positive results. Thus, we assumed that the hAFP mRNA we detected implied the presence of intact cells producing hAFP protein in blood. However, with this method, we cannot distinguish between hAFP mRNA originating from HCC cells and that originating from normal hepatocytes.

We did not find hAFP mRNA in the blood of controls, although some control subjects had elevated serum hAFP values. In contrast, in 36% of the patients with HCC, hAFP mRNA was detectable in the peripheral blood. This result was comparable with the findings in a previous report.[8] The frequency of positive hAFP mRNA in the blood was associated with the severity of HCC, indicated by tumor size, presence of intrahepatic metastasis, portal vein thrombosis, or distant metastasis, as shown in Table 3. There also appeared to be some relationship between our findings and patient's prognosis; however, we have not shown these data, because the observation period seemed insufficient to such conclusions.

In some patients, originally negative amplification changed to positive after interventional therapies. It is possible that the disruption of lobules caused by the therapy was responsible for this conversion. Further study is required, focusing on the interval after treatment at which the changes occur.

Some patients without elevated serum hAFP values showed hAFP mRNA in the blood; it was assumed that the circulating HCC cells, losing their contact with surrounding cells, produced and secreted hAFP or other primitive proteins. Further, we detected hAFP transcripts in the blood of some patients without any distant metastases. It will be of interest to examine whether these circulating HCC cells, which could not be detected by conventional means, will later develop to form metastatic nodules. We are now following these patients carefully for the occurrence of metastasis.

Our findings imply that HCC tumor cells are readily disseminated via the systemic blood flow than was previously thought. Moreover, such dissemination can be induced by therapeutic maneuvers. These tumor cells are transported throughout the body and trapped in the capillary walls of various organs. However, successful implantation and subsequent growth in situ may vary from site to site, perhaps affected by environmental factors. It may be why most of the patients who underwent the appearance of hAFP mRNA in their blood later returned to negative. Those tumor cells driven to disseminate did not have such malignant potential that they could not remain viable outside of the liver and disappeared in time. To understand these mechanisms completely, further investigations of cell adhesion and local antitumor immunity are required.

As discussed above, our results showed that this assay is sensitive enough to detect very small numbers of HCC cells in blood, and is useful for detecting disseminated cancer cells in the earlier stages of the disease. Furthermore, this method will provide new information that can be used in assessing patients and selecting the optimal therapy. For example, indications for total liver resection with orthotopic liver transplantation can be

determined more effectively by our method. Also, considering that some therapies can induce dissemination of HCC cells, the strategy for HCC treatment may be reconstructed (e.g., adjuvant chemotherapy or immunotherapy may be combined with conventional therapies). Further studies are now ongoing in our laboratory, focusing particularly on posttherapeutic changes, to clarify the way in which the cancer cells are driven to disseminate.

References

1. Pichlmayr R, Weimann A, Ringe B. Indications for liver transplantation in hepatobiliary malignancy. *Hepatology* 1994;20: 33S–40S.
2. Yamanaka N, Okamoto E, Fujihara S, Kato T, Fujimoto J, Oriyama T, et al. Do the tumor cells of hepatocellular carcinomas dislodge into the portal venous stream during hepatic resection? *Cancer* 1992;70:2263–7.
3. Smith B, Selby P, Southgate J, Pittman K, Bradley C, Blair GE. Detection of melanoma cells in peripheral blood by means of reverse transcriptase and polymerase chain reaction. *Lancet* 1991;338:1227–9.
4. Mattano LA, Moss TJ, Emerson SG. Sensitive detection of rare circulating neuroblastoma cells by the reverse transcriptase-polymerase chain reaction. *Cancer Res* 1992;52:4701–5.
5. Moreno JG, Croce CM, Fischer R, Monne M, Vihko P, Mulholland SG, et al. Detection of hematogenous micrometastasis in patients with prostate cancer. *Cancer Res* 1992;52:6110–2.
6. Tobal K, Sherman LS, Foss AJE, Lightman SL. Detection of melanocytes from uveal melanoma in peripheral blood using the polymerase chain reaction. *Invest Ophthalmol Vis Sci* 1993;34: 2622–5.
7. Brossart P, Keilholz U, Willhauck M, Scheibenbogen C, Mohler T, Hunstein W. Hematogenous spread of malignant melanoma cells in different stages of diseases. *J Invest Dermatol* 1994;101: 887–9.
8. Hillaire S, Barbu V, Boucher E, Moukhtar M, Poupon R. Albumin messenger RNA as a marker of circulating hepatocytes in hepatocellular carcinoma. *Gastroenterology* 1994;106:239–42.
9. Naito H, Kuzumaki N, Uchino J, Kobayashi R, Shikano T, Ishikawa Y, et al. Detection of tyrosine hydroxylase mRNA and minimal neuroblastoma cells by the reverse transcription-polymerase chain reaction. *Eur J Cancer* 1991;27:762–5.
10. Deguchi T, Doi T, Ehara H, Ito S, Takahashi Y, Nishino Y, et al. Detection of micrometastatic prostate cancer cells in lymph nodes by reverse transcriptase-polymerase chain reaction. *Cancer Res* 1993;53:5350–4.
11. Taketa K. α-Fetoprotein: reevaluation in hepatology. *Hepatology* 1990;12:1420–32.
12. Deutsch HF. Chemistry and biology of α-fetoprotein. *Adv Cancer Res* 1991;56:253–312.
13. Suzuki A, Takahashi T, Okuno Y, Nakamura K, Tashiro H, Fukumoto M, et al. Analysis of abnormal expression of G-CSF gene in a novel tumor cell line (KHC 287) elaborating G-CSF, IL-1 and IL-6 with co-amplification of c-myc and c-ki-ras. *Int J Cancer* 1991;48:428–33.
14. Okuno Y, Suzuki A, Ichiba S, Takahashi T, Nakamura K, Hitomi K, et al. Establishment of an erythroid cell line (JK-1) that spontaneously differentiates to red cells. *Cancer* 1990;66:1544–51.
15. Suzuki A, Takahashi T, Okuno Y, Fukumoto M, Fukui H, Koishihara Y, et al. Estimation of interleukin 6 production by reverse transcriptase-polymerase chain reaction in four human myeloma cell lines. *Leukemia Res* 1991;15:1043–50.
16. Hermanek P, Sobin LH, editors. UICC TNM Classification of malignant tumours. 4th ed. 2nd rev. Berlin: Springer-Verlag, 1992.
17. Oppenheim A, Karsai A, Treisman R, Fibach E, Treves A, Goldfarb A, et al. Beta-thalassemia: analysis of mRNA precursors of a mutant human globin gene with defective splicing using peripheral blood nucleated red blood cells. *Hemoglobin* 1986;10: 573–86.
18. Simms D, Cizdziel PE, Chomczynski P. TRIzolTM: a new reagent for optimal single-step isolation of RNA. *Focus* 1993;15: 99–102.
19. Morinaga T, Sakai M, Wegmann TG, Tamaoki T. Primary structures of human α-fetoprotein and its mRNA. *Proc Natl Acad Sci USA* 1983;80:4604–8.
20. Knowles BB, Howe CC, Aden DP. Human hepatocellular carcinoma cell lines secrete the major plasma proteins and hepatitis B surface antigen. *Science* 1980;209:497–9.
21. Di Bisceglie AM, Dusheiko GM, Paterson AC, Alexander J, Shouval D, Lee CS, et al. Detection of alpha-foetoprotein messenger RNA in human hepatocellular carcinoma and hepatoblastoma tissue. *Br J Cancer* 1986;54:779–85.
22. Chelly J, Concordet JP, Kaplan JC, Kahn A. Illegitimate transcription: transcription of any gene in any cell type. *Proc Natl Acad Sci USA* 1989;86:2617–21.

[CANCER RESEARCH 37, 646–650, March 1977]

Free DNA in the Serum of Cancer Patients and the Effect of Therapy

S. A. Leon, B. Shapiro, D. M. Sklaroff, and M. J. Yaros

Departments of Nuclear Medicine and Radiation Therapy, Division of Radiology, Albert Einstein Medical Center, Philadelphia, Pennsylvania 19141

SUMMARY

A radioimmunoassay for ng quantities of DNA was developed. [^{125}I]Iododeoxyuridine-labeled DNA was used as the antigen, and the serum of a lupus erythematosus patient served as the source of antibody. The level of free DNA in the serum of 173 patients with various types of cancer and in 55 healthy individuals was determined by this radioimmunoassay. DNA concentration in the normal controls had a range of 0 to 100 ng/ml with a mean of 13 ± 3 ng/ml (S.E.). For comparison purposes, the range of 0 to 50 ng/ml was designated as normal, and 93% of controls were found in this range. In the cancer patients, the DNA concentration ranged from zero to μg levels with a mean of 180 ± 38 ng/ml. Fifty % of the patients' values were found in the range of 0 to 50 ng/ml; the other 50% were between 50 and 5000 ng/ml. No correlation could be seen between DNA levels and the size or location of the primary tumor. Significantly higher DNA levels, however, were found in the serum of patients with metastatic disease (mean of 209 ± 39 ng/ml), as compared to nonmetastatic patients (mean 100 ± 30, $p <$ 0.02).

After radiation therapy in lymphoma, lung, ovary, uterus, and cervical tumors, the levels decreased in 66 to 90% of the patients, whereas in glioma, breast, colon, and rectal tumors, the DNA levels decreased only in 16 to 33% of the patients. Generally, the decrease in DNA concentration in the serum correlated with improved clinical condition, such as decrease of tumor size and reduction of pain. Conversely, when DNA levels either increased or remained unchanged, a lack of response to the treatment was noted. Of 17 patients who died within a year, 13 showed DNA levels that remained high or unchanged, whereas only 4 showed lower levels during treatment. Persistent high or increasing DNA levels in the circulation, therefore, may signal a relapse and are probably a poor prognostic sign.

The relatively high percentage (50%) of cancer patients with apparently normal DNA levels would suggest that this test may have low diagnostic value. It should be pointed out, however, that all these patients represent a selected group considered for radiation therapy, usually after surgery and/or chemotherapy. It is possible that a better correlation between DNA levels and cancer will be obtained prior to the initiation of treatment. On the other hand, DNA in the serum may be an important tool for the evaluation of therapy or the comparison of different regimens.

Received July 14, 1976; accepted November 22, 1976.

INTRODUCTION

Tissue and cell injury take place under both normal and pathological conditions. It is expected, therefore, that intracellular material such as DNA may be released into the circulation. Indeed, free DNA has been detected in serum and other fluids such as synovial fluid (8). For example, circulating DNA can be found in disorders such as systemic lupus erythematosus, rheumatoid arthritis, pulmonary embolism or infarct, and cancer (5, 10). In addition, therapeutic procedures may also release DNA as after high-dose corticosteroid treatment or surgery (5, 8). Healthy controls, on the other hand, usually showed lower levels of free DNA in the serum. The results for both disease and control groups are highly variable, depending on the methods used for detection of DNA. In some instances, indirect methods such as hemagglutinin inhibition (10), complement fixation (3), and double diffusion in agarose (8) showed μg levels of DNA. In other cases, using counterimmunoelectrophoresis, only fractions of a μg were detected (5, 15). Considerable controversy exists concerning the normal level of DNA in healthy controls, with reported values ranging from barely detectable levels (3, 5) to μg per ml (10). Similarly, a wide range of levels has been reported for a variety of disorders, including leukemia, solid tumors, and chronic inflammatory diseases such as systemic lupus erythematosus and rheumatoid arthritis (10). In addition to the release of DNA, the common trait characterizing these diseases is lymphoid hyperplasia or sustained immunological stress. Most of the results agree in one respect: generally, the levels of DNA are higher than in healthy controls.

The development of a direct competitive binding test for ng quantities of DNA (12, 13) in our laboratory provided a sensitive probe for the investigation of DNA release. The method is based on the binding between ^{125}I-labeled DNA as the antigen and the serum of a systemic lupus erythematosus patient containing high-avidity antibody. Extraction and purification of the test sample are not necessary, and the determination can be performed directly on the patient's serum. No interference has been found from either RNA or low-molecular-weight nucleotides (12). The small amount of nonspecific binding by the basic component of complement (C1q) can be eliminated by heat inactivation (1). The conformation of DNA in the serum has been reported as single stranded by some investigators (10) and as double stranded by others (2, 7, 8). The antiserum used in our study contains antibodies against both double- and single-stranded DNA, with a relative avidity of 9.6 and 5.6 × 10^5 liters/mole, respectively (11). This antiserum, therefore, is used for the

Downloaded from cancerres.aacrjournals.org on December 15, 2011
Copyright © 1977 American Association for Cancer Research

Exhibit T

determination of total DNA and cannot discriminate between the 2 forms of DNA if present in a mixture.

The purpose of this study was to determine the incidence of elevated DNA levels in cancer patients and whether radiation therapy affects the DNA level.

MATERIALS AND METHODS

Labeling and Preparation of DNA. DNA was labeled by growing *Escherichia coli* in the presence of [^{125}I]iododeoxyuridine as described previously (12, 13). After extraction and purification, the Millipore filtration step (12) was omitted, and the DNA was passed through a methylated albumin-Kieselguhr column (11). With this modification, the nonspecific binding of DNA by serum components was reduced to 2 to 3%.

DNA Antiserum. The serum of a lupus erythematosus patient (L. W.) was used as the antibody in the radioimmunoassay (11, 12).

Radioimmunoassay. The details have been described previously (12). Each sample was tested without antibody, for the determination of nonspecific binding. These counts were subtracted from the counts obtained in the presence of antibody. In the absence of cold DNA (no inhibition control), 45 to 52% of the counts were found in the precipitate. Known amounts of native salmon sperm DNA were added to normal human serum in the range of 0 to 1000 ng/ml, and a standard curve was constructed in the logit transform (12).

Patients and Sera. Blood samples were drawn from patients and normal, healthy controls by venipuncture. The blood was allowed to clot at room temperature within 1 hr, and the serum was separated and frozen until use. Twenty-five μl were taken for the determination of DNA concentration.

Only cancer patients considered for radiation therapy were taken for this study. Samples were drawn before the beginning of the treatment and on biweekly intervals during the treatment. All patients received radiation daily for a minimum of 2 weeks, and some received it for as long as 6 weeks. All patients had confirmed histological diagnosis of cancer. The test was performed without knowledge of their clinical condition, the presence or absence of metastases, or the outcome of surgical procedures. Subsequently, this information was obtained and correlated with the results of the DNA tests.

The normal controls consisted of healthy hospital personnel, matched by age (17 to 61 years) and sex (24 males, 31 females) to the cancer patients.

RESULTS

Normal Values of DNA in the Serum. The concentration of DNA was below detectable levels (0 to 25 ng/ml) in 42 of the 55 sera tested (76%). Another 9 sera contained 25 to 50 ng/ml (16%). Only 4 sera (7%) contained 50 to 100 ng/ml, and none were found above 100 ng/ml (Chart 1). Thus, 93% of the normal sera were found in the range of 0 to 50 ng/ml with a mean of 13 ± 3 ng/ml. For comparison purposes, 0 to 50 ng/ml was designated as the normal range. On repeated tests with subsequent samples from the 4 individuals with a DNA concentration of 50 to 100 ng/ml in the serum, 2 showed lower values in the 0 to 50 ng/ml range. It is unknown at present whether this represents normal fluctuation in DNA concentration in the serum or whether it is due to a latent or subclinical pathological condition. No correlation could be found between the DNA value and the age, sex, and smoking habits of the test subjects. Hopefully, further testing with larger numbers of individuals will clarify these points.

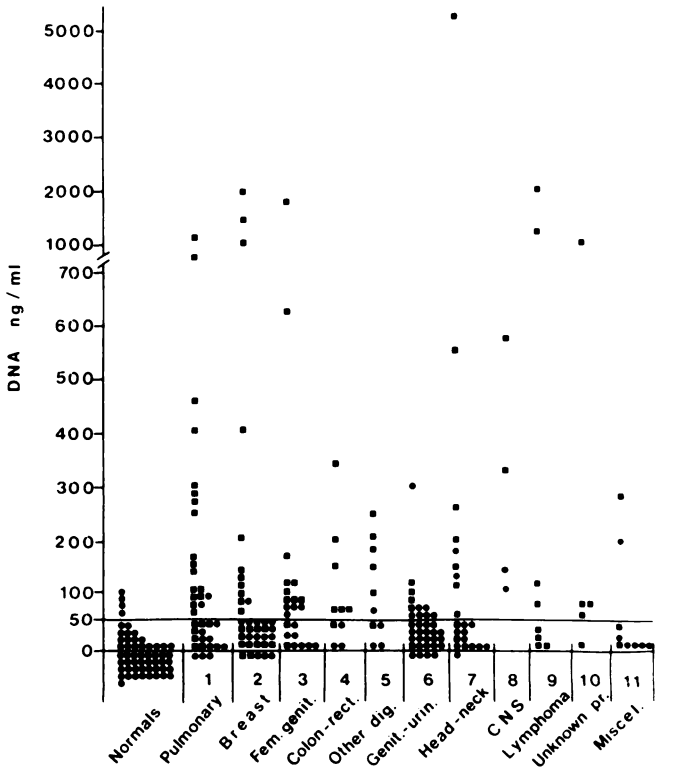


Chart 1. Distribution of DNA levels in the serum of tumor groups (total, 173 patients) and 55 healthy individuals, measured by radioimmunoassay. The values represent the average of duplicate determinations. The reproducibility was within 10 to 15% (12). ■, metastatic; ●, nonmetastatic. *Fem. genit.*, female genital; *Colon-rect.*, colon-rectum; *Other dig.*, other digestive; *Genit.-urin.*, genitourinary; *Unknown pr.*, unknown primary; *Miscel.*, miscellaneous.

DNA Concentration in the Serum of Cancer Patients before Therapy. The 173 patients were classified according to the site of the primary tumor. All types showed very wide ranges, from zero to as high as 5000 ng/ml. The mean for each group (± S.E.) and the p value relative to the mean of the normal controls were: (1) 164 ± 44, $p < 0.002$; (2) 193 ± 79, $p < 0.02$; (3) 182 ± 95, $p < 0.05$; (4) 102 ± 37, $p < 0.01$; (5) 104 ± 28, $p < 0.002$; (6) 41 ± 11, $p < 0.01$; (7) 108 ± 35, $p < 0.005$; (8) 286 ± 106, $p < 0.01$; (9) 493 ± 299, $p = 0.05$; (10) 245 ± 189, p not significant; and (11) 64 ± 39, p not significant. The mean for all cancer patients was 180 ± 38 ng/ml, $p < 0.005$. It can be seen that in some groups of tumors (head-neck, lung, breast, ovary, uterus, lymphosarcoma) very high concentrations of DNA could be found, up to μg levels. In others (digestive tract, genitourinary), the distribution had narrower range (Chart 1). In the central nervous system (glioma of brain or cord), all the patients

Downloaded from cancerres.aacrjournals.org on December 15, 2011
Copyright © 1977 American Association for Cancer Research

showed high values, but the small number of samples (4) precludes further evaluation of this finding. With the possible exception of central nervous system tumors, all groups showed some values in the normal range of 0 to 50 ng/ml. Interestingly, almost all patients with high values (above 100 ng/ml) had metastatic disease. In the low range (0 to 50 mg/ml), on the other hand, both types were found. No correlation could be found between the DNA concentration in the serum and the size, location, or histological composition of the tumor. However, in patients with well-contained, slow-growing tumors, DNA levels tended to be lower than in those with faster-growing, poorly differentiated cells. This correlation was seen clearly when the DNA levels in patients without evidence of metastases were compared to those with proven metastatic disease (Chart 2). The mean DNA concentration in the 1st group was 100 ± 30 ng/ml, whereas the mean calculated for the 2nd group was 209 ± 39 ng/ml ($p < 0.02$). The distribution of normal, healthy individuals is also included in Chart 2 for comparison.

DNA Concentration in the Serum after Radiation Therapy. In some tumors, such as lung, ovary, uterus, cervix, and lymphoma, the majority of the patients (66 to 90%) responded with lower DNA levels after radiation. In others, such as breast, colon, rectum, and glioma, this effect was observed only in a few of the patients treated (16 to 33%) (Table 1). Generally, when the treatment was beneficial, as determined by decrease in tumor size or reduction of pain, a decrease in serum DNA concentration also was observed. Conversely, when the treatment was unsuccessful, DNA levels tended to remain unchanged or even increased. Some representative examples for each type of response are illustrated in Table 2. The results for all types of cancer tested are summarized in Table 3. Clearly, in patients with metastatic disease, the DNA levels are generally higher (see above) and tend to stay high during and after treatment. Relapses occurred more frequently in this group, although this observation has not been evaluated critically because of the small number of patients. Of the 17 patients who died within 1 year, 13 showed DNA levels that increased or remained unchanged during the treatment, whereas the DNA decreased in only 4 cases.

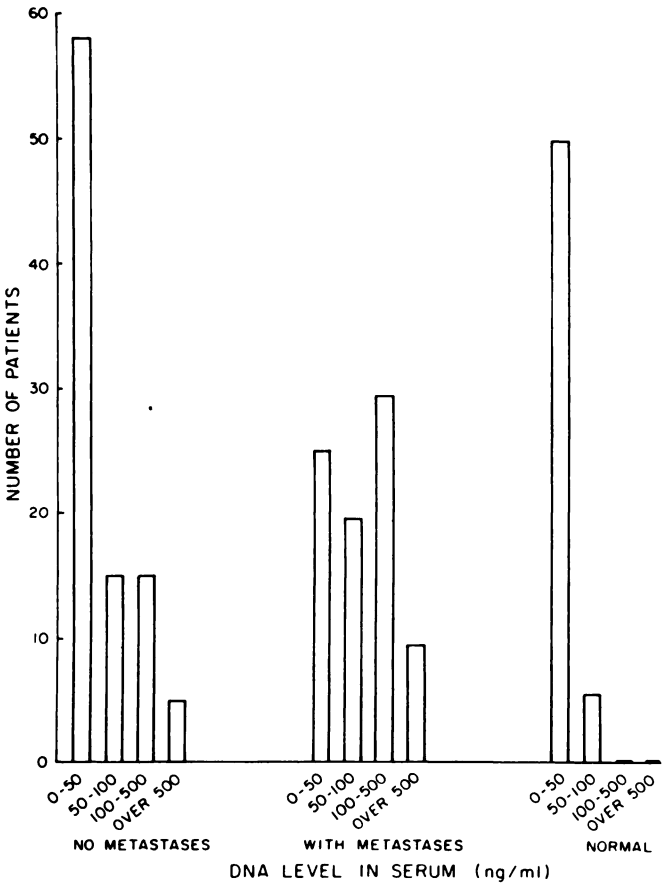


Chart 2. DNA levels in metastatic and nonmetastatic patients, compared to normal controls. The highest concentration in each group ranged between 500 and 5000 ng/ml.

DISCUSSION

The source of free DNA in solution in the serum is unknown at present. In addition to viral DNA, other sources have been reported such as normal lymphocytes (2) and tumor cells (14). The release of DNA from lymphocytes *in vitro* may also take place *in vivo*, although evidence for such a phenomenon is not available. Autolysis of cells after the collection of blood samples and release of DNA can be ruled out on the basis of control experiments. Repeated freezing and thawing of the white cell layer suspended in normal serum did not result in detectable DNA concentration in the supernatant fraction (S. A. Leon and B. Shapiro, unpublished results). The fact that DNA levels in normal individuals in our study are usually zero or very low (25 to 50 ng/ml), in contrast to other reports (3, 8, 10), is probably due to the method used. The choice of [^{125}I]iododeoxyuridine-labeled DNA, with its low affinity for normal globulin, as well as an antiserum containing high-avidity antibody for DNA, results in a very specific and sensitive determination of DNA (12). For example, we had shown previously that low-molecular-weight nucleotides and RNA do not interfere in this direct competitive binding radioimmunoassay, indicating that the antibody is highly specific for DNA (12). For these reasons, we conclude that normal serum contains very little free DNA, in the range of 0 to 50 ng/ml and, in a small number of sera, up to 100 ng/ml. Further study will be required to establish whether fluctuation in DNA levels within this range is a normal physiological phenomenon or due to a latent or subclinical pathological condition.

In cancer patients, higher concentrations of DNA in the serum may be expected if we assume that the source of this material is proliferating tumor cells (14). Another possibility is that the DNA is released from normal cells (lymphocytes) during their interaction with virus or tumor cells (16, 18). Isolation and characterization of the DNA in serum may be required to distinguish between these possibilities.

The persistence of DNA in the circulation raises another question, namely, the function of DNAse I and II in the serum (17). In malignant disease, lower activity of these enzymes has been reported. This may be due to the release of an inhibitor which may be present in both tumors (4) and normal cells (thrombocytes) (6, 9). We are presently investigating this question in our laboratory.

The elevated concentration of DNA in the serum of cancer

Downloaded from cancerres.aacrjournals.org on December 15, 2011
Copyright © 1977 American Association for Cancer Research

Table 1

Radiation therapy and DNA levels in serum of cancer patients

Pulmonary tumors included bronchogenic and alveolar adenocarcinoma; female reproductive tract: ovary, uterus and cervix; colorectal tumors were examined separately from other digestive types: esophagus, pancreas, and gallbladder. The genitourinary types included bladder, prostate, and testicular carcinomas. In the head and neck group, 8 tumors are represented: palate, epiglottis, gingiva, nasopharynx, tonsil, larynx, maxillary sinus, and submaxillary gland. The central nervous system group included glioma of the brain and cord. In the miscellaneous group were included bone giant cell carcinoma, epithelioma, melanoma, kidney, and hepatocarcinoma. Samples during and after treatment were not available for some patients. Levels were considered increased or decreased only if the change was greater than 25%.

		Before therapy					
		No. of patients			Effect of therapy (% no. of patients)		
Type	Total no. of patients	0–50 ng/ml	Over 50 ng/ml	Over 50 ng/ml (%)	Decreased	Increased	No change
1. Pulmonary	33	13	20	61	75	25	0
2. Breast	32	20	12	38	22	66	11
3. Female genitals	19	7	12	63	90	10	0
4. Colon-rectum	9	4	5	56	16	50	33
5. Other digestive	10	4	6	60	50	0	50
6. Genitourinary	29	19	10	34	55	27	18
7. Head-neck	17	9	8	47	55	33	11
8. Central nervous system	4	0	4	100	33	66	0
9. Lymphoma	7	3	4	57	66	33	0
10. Unknown primary	5	1	4	80	0	100	0
11. Miscellaneous	8	6	2	25	0	0	100
Total	173	86	87	50			
12. Normals	55	51	4	7			

Table 2

DNA levels in individual patients

The 1st number in each sequence represents the DNA level before radiotherapy. Subsequent values represent the levels after 2, 4, or more weeks of treatment.

		DNA (ng/ml)		
Patient	Tumor	Increase	Decrease	No change
H. M.	Lung (M)[a]		400–360–110	
G. D.	Lung (M)		250–140–75	
A. M.	Lung (M)		280–95–30	
J. G.	Lung (M)	140–1000–250		
H. S.	Lung (M)	30–190–160		
V. B.	Lung	25–115–120		
F. S.	Lung (M)			35–50–50
J. G. A.	Lung			80–60–80
I. B.	Breast (M)	25–120–120		
I. R.	Breast (M)	125–160–320–300		
R. J.	Breast	40–80–100		
B. M.	Breast (M)	115–210–200		
D. H.	Breast			90–95–60
H. P.	Breast (M)		140–90–80	
A. T.	Cervix (M)		170–100–50	
F. O.	Ovary (M)		1800–0–0	
D. Y.	Ovary (M)		650–50–50	
P. M.	Uterus			75–40–75
J. F.	Rectum (M)			175–200–150
A. G.	Colon (M)	150–1600		
J. G. L.	Rectum		340–40–40	
H. G.	Pancreas (M)			145–60–120
C. B.	Larynx		550–2000–110	
J. A.	Tonsil (M)	260–330–360		
J. M.	Glioma	110–55–350		
E. K.	Lymphosarcoma		2000–90–340	

[a] M, metastatic.

Downloaded from cancerres.aacrjournals.org on December 15, 2011
Copyright © 1977 American Association for Cancer Research

Table 3

DNA levels after radiotherapy in metastatic and nonmetastatic patients

DNA levels were compared before therapy and at 2 weekly intervals during the treatment in 81 metastatic and 92 nonmetastatic patients.

DNA Level	Patients (%)	
	Metastatic	Nonmetastatic
Increased	61	58
No change	13	3
Decreased	26	39

patients seems to correlate with the presence of tumor and/or metastatic disease. In preliminary studies, the DNA levels in patients with malignant lesions of the gastrointestinal tract were significantly higher than in those with inflammatory conditions (carcinoma of the pancreas *versus* pancreatitis; Hodgkin's gastric ulcer *versus* duodenal ulcer) (B. Shapiro, S. A. Leon, E. Cohn, and M. Desai, DNA in the Serum of Patients with Tumors of the Gastrointestinal Tract, in preparation). The diagnostic value of this finding, either for screening of suspected patients or for predicting a relapse, may be questionable at present. The reason for this conclusion is the relatively high number of cancer patients (50%) with DNA levels in the normal range (0 to 50 ng/ml). It should be pointed out, however, that the patients reported in this study by no means represent a random sample of malignant diseases. These patients represent a selected group considered for radiation therapy, in many cases after surgery and/or chemotherapy. It is possible that a prospective study of patients suspected for cancer, before, during, and after treatment, will yield results with better correlation between DNA levels and neoplastic disease. We conclude for the present that high levels of DNA in the serum accompany a pathological condition, but normal levels do not rule out cancer and, probably, other diseases.

After radiation therapy, a surprisingly high percentage of the patients (about 40%) showed decreased or unchanged DNA levels. This was especially obvious in patients who remained in the normal range of 0 to 50 ng/ml. Although radiation should have induced cell death and release of DNA in the circulation, it is equally possible that the arrest of tumor cell proliferation also reduced DNA release, on the assumption that they are the source of this DNA. Another explanation may be that the release of DNAse inhibitor from the tumor cells was reduced. In either case (and possibly both), lower levels of DNA in the serum would result. Although the mechanism is not clear at present, our findings suggest that persistence of DNA in the circulation after therapy correlates with poor response to treatment. In these patients, X-ray visualization or palpation (when possible) did not show a decrease in tumor size; persistence of pain and relapses were observed. Conversely, decrease in DNA levels during treatment seem to be a better prognostic sign, with reduction in tumor size, pain, and other signs of remission. We hope, therefore, that sequential measurements of DNA concentration may be a useful tool for monitoring the effects of therapy. Preliminary studies indicate that similar results are obtained with chemotherapy (R. Bornstein, B. Shapiro, and S. A. Leon, Effect of Chemotherapy on DNA Levels in Cancer Patients, in preparation). This test may be useful, therefore, to compare and evaluate objectively the effectiveness of different treatments or their combinations.

ACKNOWLEDGMENTS

We thank Cynthia Baumel for her technical assistance.

REFERENCES

1. Agnello, V., Winchester, R. J., and Kunkel, H. G. Precipitin Reactions of the C1q Component of Complement with Aggregated γ-globulin and Immune Complexes in Gel Diffusion. Immunology, *19:* 909-919, 1970.
2. Anker, P., Stroun, M., and Maurice, P. A. Spontaneous Release of DNA by Human Blood Lymphocytes *in Vitro*. Cancer Res., *35:* 2375-2382, 1975.
3. Barnett, E. V. Detection of Nuclear Antigens (DNA) in Normal and Pathologic Human Fluids by Quantitative Complement Fixation. Arthritis Rheumat., *11:* 407-417, 1968.
4. Cooper, E. J., Trautmann, M. L., and Laskowski, M. Occurrence and Distribution of an Inhibitor for Deoxyribonuclease in Animal Tissues. Proc. Soc. Exptl. Biol. Med., *73:* 219-222, 1950.
5. Davis, G. L., and Davis, J. S. Detection of Circulating DNA by Counterimmunoelectrophoresis (CIE). Arthritis Rheumat., *16:* 52-58, 1973.
6. Frost, P. G., and Lachmann, P. J. The Relationship of Deoxyribonuclease Inhibitor Levels in Human Sera to the Occurrence of Antinuclear Antibodies. Clin. Exptl. Immunol., *3:* 447-455, 1968.
7. Harbeck, R. J., Hoffmann, A. A., and Carr, R. I. Studies on the Nature of Circulating DNA in SLE. J. Rheumatol., *2:* 194-203, 1975.
8. Hughes, G. R. V., Cohen, S. A., Lightfoot, R. W., Meltzer, J. I., and Christian, C. L. The Release of DNA into Serum and Synovial Fluid. Arthritis Rheumat., *14,* 259-266, 1971.
9. Hughes, G. R. V., and Lachmann, P. J. Systemic Lupus Erythematosus. *In:* P. G. H. Gell, R. R. A. Coombs, and P. J. Lachmann (eds.), Clinical Aspects of Immunology, p. 1131. Oxford: Blackwell Scientific Publications, 1975.
10. Koffler, D., Agnello, V., Winchester, R., and Kunkel, H. G. The Occurrence of Single-stranded DNA in the Serum of Patients with SLE and Other Diseases. J. Clin. Invest., *52:* 198-204, 1973.
11. Leon, S. A., Green, A., Ehrlich, G. E., Poland, M., and Shapiro, B. Avidity of Antibodies in SLE: Relation to Severity of Renal Involvement. Arthritis Rheumat., in press.
12. Leon, S. A., Green, A., Yaros, M. J., and Shapiro, B. Radioimmunoassay for Nanogram Quantities of DNA. J. Immunol. Methods, *9:* 157-164, 1975.
13. Leon, S. A., Shapiro, B., Kollmann, G., and Green, A. A Comparison of Methods for Labelling DNA for Use in the Radioimmunoassay of DNA-antibodies. J. Immunol. Methods, *5:* 1-8, 1974.
14. Rosenberg, B. Possible Mechanisms for the Antitumor Activity of Platinum Coordination Complexes. Cancer Chemotherapy Rept., *59:* 589-598, 1975.
15. Schur, P. H., DeAngelis, D., and Jackson, J. M. Immunological Detection of Nucleic Acids and Antibodies to Nucleic Acids and Nuclear Antigens by Counterimmunoelectrophoresis. Clin. Exptl. Immunol., *17:* 209-218, 1974.
16. Tonietti, G., Oldstone, M. B. A., and Dixon, F. J. The Effect of Induced Chronic Viral Infections on the Immunologic Diseases of New Zealand Mice. J. Exptl. Med., *132:* 89-109, 1970.
17. Wroblewski, F., and Bodansky, O. Presence of Desoxyribonuclease Activity in Human Serum. Proc. Soc. Exptl. Biol. Med., *74:* 443-445, 1950.
18. Yoshida, T. O. High Incidence of Antinuclear Antibodies in the Sera of Nasopharyngeal Cancer Patients. *In:* W. Nakahara, K. Nishioka, T. Hirayama, and Y. Ito (eds.), Recent Advances in Human Tumor Virology and Immunology, pp. 443-460. Baltimore: University Park Press, 1971.

Downloaded from cancerres.aacrjournals.org on December 15, 2011
Copyright © 1977 American Association for Cancer Research

that although point mutations (sometimes in clusters) can be recovered from irradiated spermatogonia, structural rearrangements seldom are. Low mutability combined with selection against aberrations involving chromosome breakage seem to be the principal causes of the low recovery rate of chromosome aberrations from spermatogonia. In the mouse, too, point mutations are transmitted from irradiated stem cell spermatogonia while large deletions and other structural changes are detected only in treated post-stem-cell stages.[7] For translocations, treated spermatids and spermatozoa generally yield the highest levels in the mouse, germinal selection eliminating many of those induced in spermatogonia—indeed translocations that cause male sterility can only be recovered from spermatids or spermatozoa.[8] In man it is almost certain that any de-novo reciprocal X-autosome translocation of paternal origin has originated post-meiotically. Movements which accompany metamorphosis of the spermatid into the sperm head could be a factor favouring structural change.

In the adult female mouse, the principal period of sensitivity to mutation seems to be when the oocyte is maturing in the follicle before ovulation, when point mutations and deletions can be induced.[7] In women this period lasts only 36 hours or so and it is not surprising that so few mutagenic "hits" are scored: human spermatogenesis takes about 64 days, a further 10 or 11 days being required for the passage of spermatozoa through the epididymis and vas deferens. With millions of sperm being required to ensure the fertilisation of a single egg, it would seem inevitable that more mutations might arise in males. The strength or weakness of a paternal age effect might, moreover, depend on the relative proportions of point mutations to chromosomal aberrations for any particular condition, only those diseases due to single-gene defects being expected to show a strong association.

MRC Human Genetics Unit,
Western General Hospital,
Edinburgh EH4 2XU, UK

ANN C. CHANDLEY

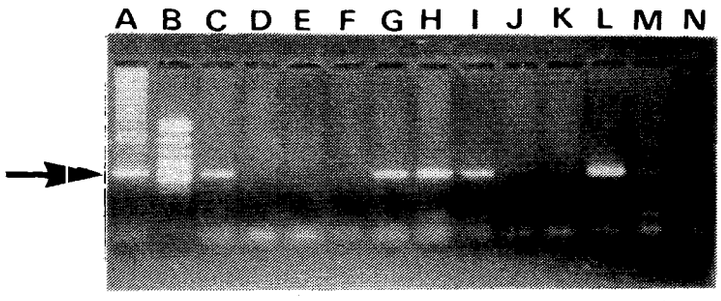
1. Olsen SB, Magenis RE. Preferential paternal origin of de novo structural chromosome rearrangements. In: Daniel A, ed. The cytogenetics of mammalian autosomal rearrangements. New York: Alan R. Liss, 1988: 583–99.
2. Dryja TP, Mukai S, Petersen R, Rapaport JM, Walton D, Yandell DW. Parental origin of mutations of the retinoblastoma gene. *Nature* 1989; **339:** 556–58.
3. Nicholls RD, Knoll JH, Glatt K, et al. Restriction fragment length polymorphisms within proximal 15q and their use in molecular cytogenetics and the Prader-Willi syndrome. *Am J Med Genet* 1989: **33:** 66–77.
4. Jadayel D, Fain P, Upadhyaya M, et al. Paternal origin of new mutations in von Recklinghausen neurofibromatosis. *Nature* 1990; **343:** 558–59.
5. Baker TG. Oogenesis and ovulation. In: Austin CR, Short RV, eds. Germ cells and fertilization. Cambridge: Cambridge University Press. 1972: 14–45.
6. Penrose LS. Parental age and mutation. *Lancet* 1955; ii: 312–13.
7. Russell LB, Hunsicker PR, Cacheiro NLA, Bangham JW, Russell WL, Shelby MD. Chlorambucil effectively induces deletion mutations in mouse germ cells. *Proc Natl Acad Sci (USA)* 1989; **86:** 3704–08.
8. Searle AG, Beechey CV, Evans EP. Meiotic effects in chromosomally derived male sterility of mice. *Ann Biol Anim Biochem Biophys* 1978; **18:** 391–98.

Detection of single-copy fetal DNA sequence from maternal blood

SIR,—We have reported amplification of a Y-specific repeat sequence from the peripheral blood of women pregnant with male fetuses[1] by use of the polymerase chain reaction (PCR).[2] However, before the method can be applied to prenatal diagnosis of non-sex-linked genetic disorders it is essential that single-copy fetal sequences be amplified from maternal blood. We now report the amplification of a single-copy Y-specific sequence from peripheral blood DNA of pregnant women. This sequence constitutes part of a gene which is expressed in testicular tissue.[3] The single-copy nature of the sequence chosen was based on Southern blotting[3] and co-amplification experiments, in which both the Y-sequence and α-antitrypsin sequence were amplified to the same extent.[4] The new primer sequences are:

Y1·5 = CTAGACCGCAGAGGCGCCAT
Y1·6 = TAGTACCCACGCCTGCTCCGG
Y1·7 = CATCCAGAGCGTCCCTGGCTT
Y1·8 = CTTTCCACAGCCACATTTGTC.

Y1·5 and Y1·6 are the external pair of primers and have been used by Lawler et al previously to amplify a 239 base pairs Y-specific fragment.[4] Y1·7 and Y1·8 flank a 198 base pairs sequence internal to Y1·5 and Y1·6 and were designed from published sequences.[3]



PCR amplification of single-copy Y-specific sequence from maternal blood.

A = male positive control.
B = pBR322 digested by *Msp*I (marker).
C–N = Cases 10, 12, 9, 11, 1, 8, 7, 6, 5, 4, 3, and 2. Results after two rounds of amplification (forty followed by twenty-five cycles). Arrow marks position of 198 bp band.

The dual-amplification procedure was as previously described[1] except that the cycle combination used was forty cycles (with primers Y1·5/Y1·6) plus twenty-five cycles (with primers Y1·7/Y1·8). Denaturation was at 94°C for 2 min with annealing at 55°C for 2 min and extension at 72°C for 3 min. Initial decontamination of the reagents was done by incubating the reagents with *Hin*fI (30 units per 100 µl PCR reagent at 37°C for 1 h). With the previous primer combination (Y1·1/Y1·2 followed by Y1·3/Y1·4) amplification above forty cycles plus twenty cycles consistently resulted in false-positive results[1] whereas false positivity is no longer seen with the new primers. Thus our previous difficulty with a narrow window of specific amplification[1] has been overcome. This finding also implies that the suggestion by Nakagome et al[5] about the autosomal cross-reactivity of the previous Y-primers is correct.

We have used this system on a series of 27 pregnant women, at gestational ages of 6–41 weeks (figure, table). Early pregnancy samples (6–11 weeks) were obtained from women under investigation for prenatal diagnosis of possible genetic disorders at the University of Milan. Blood samples were taken before chorionic villus biopsy and the sex of the fetus was determined from cytogenetic analysis of chorionic villus culture. Intermediate and late pregnancy samples (16–41 weeks) were obtained, with informed consent, from women attending an antenatal clinic at the John Radcliffe Hospital; fetal sex was noted at delivery.

15 samples gave rise to a positive PCR signal; 13 of these women had male fetuses, with 2 false-positives (cases 20 and 21), indicating that a single-copy fetal sequence can be amplified from maternal peripheral blood. This is the first report showing that the method has potential for a wide range of single-gene disorders. Failure to

PCR AMPLIFICATION OF A SINGLE-COPY Y-SPECIFIC SEQUENCE FROM MATERNAL BLOOD

Case	Gest (wk)	PCR†	Sex	Case	Gest (wk)	PCR†	Sex
Early pregnancy samples				*Intermediate to late pregnancy samples*			
1	6	+	M	13	16	+	M
2*	9	−	F	14	17	+	M
3*	9	−	M	15	18	−	M
4*	9	+	M	16*	32	+	M
5	9	−	M	17	33	−	F
6*	9	−	F	18*	35	+	M
7*	9	+	M	19	35	−	M
8	10	+	M	20	35	+	F
9	10	−	F	21*	37	+	F
10	10	+	M	22*	37	+	M
11	11	−	F	23*	37	−	F
12	11	−	F	24*	38	+	M
				25*	39	+	M
				26*	41	−	F
				27*	41	+	M

*Cases previously reported [1]
† + = presence of clear band at 198 bp, − = absence of clear band at 198 bp.

Exhibit U

generate a positive signal in 4 women bearing male fetuses (cases 3, 5, 15, and 19) may be due to sampling error. Perhaps this could be overcome by amplifying multiple samples from each patient. The existence of false-positive results suggests that a more stringent anti-contamination regimen may be necessary and emphasises that further development will be needed before the system can be used routinely.

We thank Dr A. F. Markham, ICI Diagnostics, for donating oligonucleotide primers.

Nuffield Department of Surgery, John Radcliffe Hospital, Oxford — Y-M. D. LO

Department of Haematology, John Radcliffe Hospital — P. PATEL

A. Bianchi Bonomi Haemophilia and Thrombosis Centre, University of Milan, Milan, Italy — M. SAMPIETRO

Maternity Department, John Radcliffe Hospital — M. D. G. GILLMER

Nuffield Department of Pathology, John Radcliffe Hospital — K. A. FLEMING

Department of Haematology, John Radcliffe Hospital, Oxford OX3 9DU, UK — J. S. WAINSCOAT

1. Lo Y-MD, Patel P, Wainscoat JS, Sampietro M, Gillmer MDG, Fleming KA. Prenatal sex determination by DNA amplification from maternal peripheral blood. *Lancet* 1989; ii: 1363–65.
2. Saiki RK, Gelfand D, Stoffel S, et al. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* 1988; **239:** 487–91.
3. Arnemann J, Epplen JT, Cooke HJ, Sauermann U, Engel W, Schmidke J. A human Y-chromosomal DNA sequence expressed in testicular tissue. *Nucleic Acids Res* 1987; **15:** 8713–24.
4. Lawler M, McCann SR, Conneally E, Humphries P. Chimaerism following allogeneic bone marrow transplantation: detection of residual host cells using the polymerase chain reaction. *Br J Haematol* 1989; **73:** 205–10.
5. Nakagome Y, Nagafuchi S, Nakahori Y. Prenatal sex determination. *Lancet* 1990; **335:** 291.

Reperfusion injury and lipid peroxidation

SIR,—Dr Davies and his colleagues claim (March 31, p 741) that their study provides evidence on free-radical activity and coronary reperfusion in man. This claim is based on changes in plasma thiobarbituric reactive material (TBA-RM) before and after thrombolysis. However, this index may not be a valid index of free-radical-induced lipid peroxidation.

TBA reactivity is a sensitive and commonly used measure of free-radical-induced lipid peroxidation in biological samples. Samples are heated under acidic conditions with TBA, which reacts with malondialdehyde (MDA), a secondary product formed from decomposition of oxidised lipids. A red chromogen (MDA-TBA) is formed which is measured spectrophotometrically at 532 nm or by fluorimetry. However, the TBA test is notoriously non-specific.[1-5] TBA can react with many non-lipid components, including carbohydrates, pyrimidines, bilirubin, haemoglobin, and albumin, forming chromogens with spectra that overlap the authentic MDA-TBA standard.[2] Changes of such components in the plasma with successful thrombolysis cannot be ignored.

Plasma TBA reactivity is strongly influenced by trace concentrations of metal ions, which promote the decomposition of the unstable lipid peroxides to MDA.[2] Lipid peroxy radicals formed during the decomposition reactions can also initiate artifactual autoxidation of unsaturated lipids during the acid and heating steps.[2] As a consequence of these difficulties many workers have adopted their own versions of the method, which makes comparisons difficult.

Davies and colleagues used the method of Yagi et al,[6] in which phosphotungstic acid is used to precipitate lipoproteins. However, phosphotungstic acid can increase TBA-RM unless the precipitates are thoroughly washed.[7] This may be relevant here since Davies et al report inordinately high levels of plasma TBA-RM. For example, Yagi's group[8] has reported a mean TBA-RM serum level in healthy individuals of 3·42 nmol/ml. With a standardised high-performance liquid chromatography, which eliminates specificity and autoxidation problems, we have observed a mean plasma lipid peroxide concentration in healthy volunteers (n = 30) of 1·4 (SD 0·3) nmol/ml.[2]

Davies et al do not provide a normal range and unconventionally express TBA-RM in nmol/g albumin. For an albumin of 4 g/dl Yagi's reported normal mean corresponds to 85 nmol/g albumin and ours to 35 nmol/g. Davies et al report plasma TBA-RM levels in patients receiving thrombolytic treatment of 300–1200 nmol/g albumin (fig 1). In our own studies of patients at risk of lipid peroxidative stress, including immediate postoperative and post-angioplasty coronary reperfusion, we have not observed plasma lipid peroxide values above 100 nmol/g albumin ([5] and unpublished).

Unfortunately Davies et al correct for volume changes and haemodilution by using plasma albumin. Physiological changes in albumin after surgery or myocardial infarction are well recognised. A lowering of plasma albumin during the study period might be anticipated, irrespective of fluid infusion, and this would increase the "adjusted" plasma TBA-RM. In the absence of the actual TBA-RM values and a clearer justification for the albumin correction scepticism is warranted. Indeed the more interesting observation may have been the decrease in TBA-RM in patients not responding to thrombolysis.

Despite our misgivings thrombolytic therapy does seem to affect "TBA-RM". What this means is far from clear but Davies and colleagues' study does not warrant the claim that it constitutes evidence associating coronary reperfusion injury with lipid peroxidation in man.

Department of Surgery, University of Otago Medical School, Dunedin, New Zealand — CHRIS R. WADE, ANDRE M. VAN RIJ

1. Wong SHY, Knight JA, Hopfer SM, et al. Lipoperoxides in plasma as measured by liquid-chromatographic separation of malondialdehyde-thiobarbituric acid adduct. *Clin Chem* 1987; **33:** 214–20.
2. Wade CR, van Rij AM. Plasma thiobarbituric acid reactivity: reaction conditions and the role of iron, antioxidants and lipid peroxy radicals on the quantitation of plasma lipid peroxides. *Life Sci* 1988; **43:** 1085–93.
3. Hackett C, Linley-Adams M, Lloyd B, Walker V. Plasma malondialdehyde: a poor measure of in vivo lipid peroxidation. (letter). *Clin Chem* 1988; **34:** 208.
4. Wade CR, van Rij AM. Plasma malondialdehyde, lipid peroxides and the thiobarbituric acid reaction. (letter). *Clin Chem* 1989; **35:** 336.
5. Wade CR, van Rij AM. Plasma malondialdehyde and exhaled hydrocarbons: are they reliable indicators of lipid peroxidation and glutathione peroxidase activity? In: Wendel A, ed. Selenium in biology and medicine. Heidelberg: Springer, 1989.
6. Yagi K. A simple fluorometric assay for lipoperoxide in blood plasma. *Biochem Med* 1976; **15:** 212–16.
7. Hendriks T, Assmann RFTA. On the fluorometric assay of circulating lipid peroxides. *Clin Chim Acta* 1988; **174:** 263–70.
8. Suematsu T, Kamada T, Abe H, Kikuchi S, Yagi K. Lipid peroxidation in alcoholic liver disease in humans. *Alcohol Clin Exp Res* 1981; **5:** 427–30.

** This letter has been shown to Dr Davies and colleagues, whose reply follows.—ED. L.

SIR,—The well-known limitations of the TBA assay are well reviewed by Dr Wade and Dr van Rij. The relevance of their observations to our paper is less clear. While it is true that the spectrophotometric determination of thiobarbituric reactive material (TBA-RM) is influenced by factors such as bilirubin and sialic acid, these do not interfere with Yagi's[1] fluorimetric method, which we used to avoid these difficulties. Hydroxyl radical damage to aminoacids, carbohydrates, and nucleic acid bases in vitro can generate TBA-RM indistinguishable both spectrophotometrically and fluorimetrically from that derived from polyunsaturated fatty acids, but the importance of this in vivo is uncertain.[2] Even if some TBA-RM were derived from these sources the results would still reflect oxidative damage, albeit not lipid peroxidation. The necessity for the presence of trace amounts of iron to decompose lipid peroxides in the TBA assay is also well established.[3] However, there is sufficient iron contamination in TBA reagents for the availability of iron not to be a limiting factor—in preliminary experiments the addition of 20 μmol/l Fe^{3+} was without effect. Although we accept that residual traces of phosphotungstic acid might increase TBA-RM[4] this would cause a systematic error and would not affect the changes in TBA-RM from pre to post thrombolysis.

Early report

Presence of fetal DNA in maternal plasma and serum

Y M Dennis Lo, Noemi Corbetta, Paul F Chamberlain, Vik Rai, Ian L Sargent, Christopher W G Redman, James S Wainscoat

Summary

Background The potential use of plasma and serum for molecular diagnosis has generated interest. Tumour DNA has been found in the plasma and serum of cancer patients, and molecular analysis has been done on this material. We investigated the equivalent condition in pregnancy—that is, whether fetal DNA is present in maternal plasma and serum.

Methods We used a rapid-boiling method to extract DNA from plasma and serum. DNA from plasma, serum, and nucleated blood cells from 43 pregnant women underwent a sensitive Y-PCR assay to detect circulating male fetal DNA from women bearing male fetuses.

Findings Fetus-derived Y sequences were detected in 24 (80%) of the 30 maternal plasma samples, and in 21 (70%) of the 30 maternal serum samples, from women bearing male fetuses. These results were obtained with only 10 μL of the samples. When DNA from nucleated blood cells extracted from a similar volume of blood was used, only five (17%) of the 30 samples gave a positive Y signal. None of the 13 women bearing female fetuses, and none of the ten non-pregnant control women, had positive results for plasma, serum or nucleated blood cells.

Interpretation Our finding of circulating fetal DNA in maternal plasma may have implications for non-invasive prenatal diagnosis, and for improving our understanding of the fetomaternal relationship.

Lancet 1997; **350:** 485–87

Nuffield Department of Clinical Biochemistry (Y M D Lo MRCP), **Nuffield Department of Obstetrics and Gynaecology** (P F Chamberlain MD, Vik Rai MRCOG, I L Sargent PhD, Prof C W G Redman FRCP), **and Department of Haematology, John Radcliffe Hospital, University of Oxford, Oxford OX3 9DU, UK** (J W Wainscoat FRCPath); **and Istituto di Medicina Interna e Fisiopatologia Medica, IRCCS, Ospedale Maggiore, Milan, Italy** (N Corbetta)

Correspondence to: Dr Y M Dennis Lo, Department of Chemical Pathology, Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, New Territories, Hong Kong

Introduction

The passage of nucleated cells between mother and fetus is well recognised.[1,2] One important clinical application is the use of fetal cells in maternal blood for non-invasive prenatal diagnosis.[3] This approach avoids the risks associated with conventional invasive techniques, such as amniocentesis and chorionic-villus sampling. Substantial advances have been made in the enrichment and isolation of fetal cells for analysis.[3,4] However, most techniques are time-consuming or require expensive equipment.

There has been much interest in the use of DNA derived from plasma or serum for molecular diagnosis.[5] In particular, there have been reports that tumour DNA can be detected by molecular techniques in the plasma or serum of cancer patients.[6–8] Such reports prompted us to investigate whether fetal DNA can be detected in maternal plasma and serum.

Methods

Patients

Pregnant women attending the John Radcliffe Hospital (Oxford, UK) were recruited before amniocentesis or delivery. We obtained approval of the project from the Central Oxfordshire Research Ethics Committee. Informed consent was obtained in each case. 5–10 mL maternal peripheral blood was collected into one tube containing edetic acid and one plain tube. For women undergoing amniocentesis, maternal blood was always taken before the procedure, and 10 mL amniotic fluid was also collected for fetal sex determination. For women recruited just before delivery, fetal sex was noted at the time of delivery. Control blood samples were also taken from ten non-pregnant women, and the samples were processed in the same way as those obtained from the pregnant women.

Sample preparation

Maternal blood samples were processed 1–3 h after venesection. Blood samples were centrifuged at 3000 *g*, and plasma and serum were carefully removed from the edetic-acid-containing and plain tubes, respectively, and transferred into plain polypropylene tubes. Great care was taken to ensure that the buffy coat or the blood clot was undisturbed when plasma or serum samples, respectively, were removed. After removal of the plasma samples, the red-cell pellet and buffy coat were saved for DNA extraction with a Nucleon DNA extraction kit (Scotlabs, Strathclyde, Scotland, UK). The plasma and serum samples then underwent a second centrifugation at 3000 *g*, and the recentrifuged plasma and serum samples were collected into fresh polypropylene tubes. The samples were stored at −20°C until further processing.

DNA extraction from plasma and serum samples

Plasma and serum samples were processed for PCR by a modified version of Emanuel and Peska's method.[9] 200 μL plasma or serum in a 0·5 mL eppendorf tube was heated at 99°C for 5 min on a heat block. The heated sample was then centrifuged at maximum speed in a microcentrifuge, after which the clear supernatant was collected and 10 μL used for PCR.

Exhibit V

DNA extraction from amniotic fluid

The amniotic-fluid samples were processed for PCR by the method of Rebello and colleagues.[10] 100 μL amniotic fluid was transferred into a 0·5 mL eppendorf tube, and mixed with an equal volume of 10% Chelex-100 (Bio-Rad). After addition of 20 μL mineral oil to prevent evaporation, the tube was incubated at 56°C for 30 min on a heat block. The tube was then vortexed briefly and incubated at 99°C for 20 min. The treated amniotic fluid was stored at 4°C until PCR, and 10 μL was used in a 100 μL reaction.

PCR

The PCR was carried out broadly as described elsewhere[11] with reagents from a GeneAmp DNA Amplification Kit (Perkin Elmer, Foster City, CA, USA). The detection of Y-specific fetal sequence from maternal plasma, serum, and cellular DNA was done as described with primers Y1·7 and Y1·8, designed to amplify a single-copy sequence (DYS14).[12] The Y-specific product was 198 bp. 60 cycles of Hot Start PCR with Ampliwax technology were used on 10 μL maternal plasma or serum, or on 100 ng maternal nucleated blood-cell DNA; each cycle consists of a denaturation step at 94°C for 1 min, and a combined reannealing/extension step at 57°C for 1 min. 40 cycles were used for amplification of amniotic fluid. PCR products were analysed by agarose-gel electrophoresis and ethidium-bromide staining. PCR results were scored before fetal sex was revealed to the investigator.

Results

Serial dilutions of male genomic DNA in 1 μg female genomic DNA were carried out and amplified by the Y-PCR system with 60 cycles of amplification. Positive signals were detected up to the 1/100 000 dilution—ie, the approximate equivalent of a single male cell.

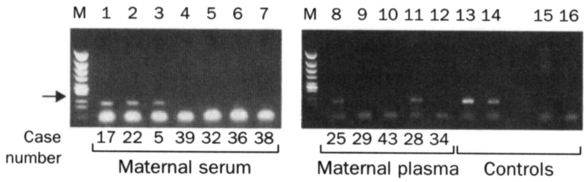
The optimum volume of heated plasma and serum samples for PCR amplification was assessed by subjecting 1 μL, 2 μL, 10 μL, 30 μL, and 50 μL heated plasma or serum samples from male individuals to Y-PCR. The best signal was obtained from 10 μL plasma or serum. Complete inhibition of amplification was reached with 50 μL. Thus, 10 μL heated plasma or serum samples were used for subsequent experiments.

Maternal plasma and serum samples were collected from 43 women who were between 12 and 40 weeks pregnant. There were 30 male and 13 female fetuses. Among the 30 women bearing male fetuses, Y-positive signals were detected in 24 plasma samples and 21 serum samples when 10 μL of the samples was used for PCR (figure and table). When DNA from nucleated blood-cells was used for Y-PCR, positive signals were detected in only five of the 30 cases (table). None of the 13 women bearing female fetuses, and none of the ten non-pregnant control women, had a positive Y signal when plasma, serum, or cellular DNA was amplified.

Discussion

Our results show that fetal DNA is present in maternal plasma and serum. Use of maternal plasma or serum for the detection of fetal DNA for non-invasive prenatal diagnosis may therefore be possible. Ironically, plasma is the material routinely discarded in the early stages of many DNA-extraction protocols, and also after the density-centifugaton step used by many investigators for non-invasive prenatal diagnosis. This is probably one of the reasons why the presence of fetal DNA in maternal plasma has not been explored previously.

Maternal serum is being used by many centres for biochemical screening of chromosomal aneuploidies and neural-tube defects. An approach by which DNA-based diagnosis is done on serum samples could be incorporated into existing screening programmes.



Amplification of fetal Y-chromosomal sequences from maternal plasma and serum

Lanes 13 and 14=Y-PCR on male genomic DNA (positive controls); lane 13 DNA equivalent to ten male cells; lane 14 DNA equivalent of one male cell. Land 15, 1 μg female genomic DNA (negative control). Lane 16, water (negative control). Arrow marks position of 198 bp Y-PCR product. M=molecular weight marker (ϕX174 DNA digested with *Hinc*II).

Our data show that fetal DNA can be detected in as little as 10 μL maternal plasma and serum. The detection rate is much higher than that for DNA from nucleated blood cells extracted from a similar volume of whole blood. This finding suggests a relative enrichment of fetal DNA in maternal plasma and serum, a phenomenon analogous to the relative enrichment of tumour DNA in the plasma and serum of cancer patients.[7,8] The low detection rate of fetal DNA sequences for DNA from nucleated blood cells

Case number	Gestation (weeks)	Y-PCR		
		Plasma	Serum	Blood
Male fetuses				
1	12	−	−	−
2	14	+	−	−
3	15	+	+	+
4	15	+	+	−
5	15	+	+	−
6	15	+	+	−
7	15	+	−	−
8	15	+	−	−
9	15	+	−	−
10	15	+	−	−
11	15	−	−	−
12	15	−	−	−
13	16	+	+	+
14	16	+	+	−
15	16	+	+	−
16	16	+	+	−
17	16	+	+	−
18	16	+	+	−
19	16	+	+	−
20	17	+	+	−
21	17	+	+	−
22	22	−	+	−
23	40	+	+	+
24	40	+	+	+
25	40	+	+	+
26	40	+	+	−
27	40	+	+	−
28	40	+	+	−
29	40	−	+	−
30	40	−	−	−
Female fetuses				
31	15	−	−	−
32	16	−	−	−
33	16	−	−	−
34	16	−	−	−
35	16	−	−	−
36	16	−	−	−
37	17	−	−	−
38	17	−	−	−
39	17	−	−	−
40	18	−	−	−
41	40	−	−	−
42	40	−	−	−
43	40	−	−	−

Amplification of fetal Y-chromosomal sequences from maternal plasma, serum, and blood

resulted from the use of only 100 ng DNA, compared with 1 μg in our previous study.[12] We chose 100 ng because this was the average quantity of DNA extracted from 10 μL whole blood with our genomic DNA-extraction method. This approach allows us to compare the relative detectability of fetal DNA in 10 μL plasma or serum and the cellular component of 10 μL whole blood.

The detection rate of fetal DNA in 10 μL plasma and serum is already high at 80% and 70%, respectively, but these rates can probably be improved. 1 mL maternal blood or serum, for example, will result in a one hundred-fold increase in the absolute amount of fetal genetic material available for analysis. Such a magnitude of increase should lead to a robust and non-invasive system for detection of paternally inherited fetal DNA sequences. However, more work is needed before this can be achieved, since the boiling method we used resulted in a relatively impure DNA extract, which inhibited PCR when volumes much larger than 10 μL were used. We are investigating other DNA-extraction methods for plasma and serum. We believe that concentration methods for plasma and serum DNA could potentially be easier than many fetal-cell isolation methods, such as cell sorting and micromanipulation.

The relative merit of the use of plasma or serum samples requires further investigation. In seven cases—2, 7, 8, 9, 10, 22, and 29—there was discordance in the PCR results obtained from the plasma and serum. A possible explanation for the discrepancies is that the quantity of fetal DNA in these cases may have a limiting effect, and thus sampling errors could be contributing to the observed results. Future research with quantitative PCR assays may elucidate this effect.

In four cases with male fetuses, both plasma and serum were negative for fetal DNA (cases 1, 11, 12, and 30). Three of these women were tested at 15 weeks of pregnancy or earlier (cases 1, 11, and 12). Furthermore, of the seven cases in which there was discordance between plasma and serum samples, all but one were tested before 23 weeks. Taken together, these results suggest that the concentration of fetal DNA increases as gestation progresses, possibly owing to the increase in fetal size. This result is analogous to that of Nawroz and colleagues,[8] who detected in head and neck cancer patients mutant plasma DNA predominantly in those with high tumour load. Future studies should investigate the temporal relation between gestation and the appearance and concentration of fetal DNA in maternal plasma.

As well as sex-linked disorders, techniques for fetal-DNA detection in maternal plasma or serum can also be used to detect many paternally inherited DNA sequences that differ from their maternal counterparts. Clinical examples include fetal rhesus D status assessment[13] and detection of certain haemoglobinopathies.[14] The plasma or serum-based approach might also be applicable to screening for chromosomal aneuploidies (such as Down's syndrome) if there is a quantitative difference in the concentration of fetal DNA in maternal plasma and serum between affected and normal pregnancies; this is a situation analogous to the high concentration of fetal cells detectable in pregnancies that involve aneuploid fetuses.[15] Fetal-cell isolation, however, will nevertheless be necessary for the definitive cytogenetic diagnosis of fetal chromosomal aneuploidies, and for the direct mutational analysis of autosomal recessive disorders caused by a single mutation.

The presence of fetal DNA in maternal plasma and serum has not been previously described. The underlying processes that cause free fetal DNA to be released into the maternal circulation have yet to be explained. Possible mechanisms include cell lysis resulting from physical and immunological damage, and developmentally regulated apoptosis of certain fetal tissues. There are interesting similarities between a growing fetus and a neoplasm: both are immunologically foreign to—and have an extensive vascular interface with—their hosts. Investigation of the variations in fetal-DNA concentrations in conditions in which the placental interface is damaged (eg, pre-eclampsia) would be of clinical and biological interest.

Contributors

Y M Dennis Lo and James S Wainscoat initiated the project. Y M Dennis Lo reviewed literature, supervised the daily practical work and analysis of laboratory results. Noemi Corbetta designed the experimental methods and was the main laboratory worker. Paul F Chamberlain, Vik Rai, Ian L Sargent and Christopher W G Redman designed the clinical module of the project. Paul F Chamberlain and Vik Rai were responsible for the clinical correlation of laboratory results. Ian L Sargent, Christopher W G Redman and James S Wainscoat analysed laboratory results. James S Wainscoat designed the DNA-extraction methods, and supervised the writing of the manuscript. All the authors read and contributed to the writing of the paper.

Acknowledgments

YMDL is supported by the Wellcome Trust. We thank N Manning for assistance in sample collection and S L Wong for helpful discussion.

References

1 Lo YMD, Patel P, Wainscoat JS, Sampietro M, Gillmer MDG, Fleming KA. Prenatal sex determination by DNA amplification from maternal peripheral blood. *Lancet* 1989; ii: 1363–65.

2 Lo YMD, Lo ESF, Watson N, et al. Two-way cell traffic between mother and fetus biologic and clinical implications. *Blood* 1996; **88:** 4390–95.

3 Simpson JL, Elias S. Isolating fetal cells from maternal blood: advances in prenatal diagnosis through molecular technology. *JAMA* 1993; **270:** 2357–61.

4 Cheung MC, Goldberg JD, Kan YW. Prenatal diagnosis of sickle cell anemia and thalassemia by analysis of fetal cells in maternal blood. *Nat Genet* 1996; **14:** 264–68.

5 Mulcahy HE, Croke DT, Farthing MJG. Cancer and mutant DNA in blood plasma. *Lancet* 1996; **348:** 628.

6 Stroun M, Anker P, Maurice P, Lyautey J, Lederrey C, Beljanski M. Neoplastic characteristics of the DNA found in the plasma of cancer patients. *Oncology* 1989; **46:** 318–22.

7 Chen XQ, Stroun M, Magnenat J-L, et al. Microsatellite alterations in plasma DNA of small cell lung cancer patients. *Nat Med* 1996; **2:** 1033–35.

8 Nawroz H, Koch W, Anker P, Stroun M, Sidransky D. Microsatellite alterations in serum DNA of head and neck cancer patients. *Nat Med* 1996; **2:** 1035–37.

9 Emanuel SL, Pestka S. Amplification of specific gene products from human serum. *GATA* 1993; **10:** 144–46.

10 Rebello MT, Hackett G, Smith J, et al. Extraction of DNA from amniotic fluid cells for the early prenatal diagnosis of genetic disease. *Prenat Diagn* 1991; **11:** 41–46.

11 Saiki RK, Gelfand DH, Stoffel S, et al. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* 1988; **239:** 487–91.

12 Lo YMD, Patel P, Sampietro M, Gillmer MDG, Fleming KA, Wainscoat JS. Detection of single-copy fetal DNA sequence from maternal blood. *Lancet* 1990; **335:** 1463–64.

13 Lo YMD, Bowell PJ, Selinger M, et al. Prenatal determination of fetal RhD status by analysis of peripheral blood of rhesus negative mothers. *Lancet* 1993; **341:** 1147–48.

14 Camaschella C, Alfarano A, Gottardi E, et al. Prenatal diagnosis of fetal hemoglobin Lepore-Boston disease on maternal peripheral blood. *Blood* 1990; **75:** 2102–06.

15 Bianchi DW, Williams JM, Pelletier C, Klinger KW, Shuber AP. Fetal cell quantitation in maternal blood samples from normal and aneuploid pregnancies. *Pediatr Res* 1996; **39:** 142 (abstr).

Clin Chem Lab Med 2012;50(5):xxx–xxx © 2011 by Walter de Gruyter • Berlin • Boston. DOI 10.1515/CCLM.2011.765

Mini Review

Fetal nucleic acids in maternal blood: the promises

Yuk Ming Dennis Lo*

Centre for Research into Circulating Fetal Nucleic Acids, Li Ka Shing Institute of Health Sciences and Department of Chemical Pathology, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, New Territories, Hong Kong SAR, PR China

Abstract

Fetal DNA is present at an approximate mean fractional concentration of 10% in the plasma of pregnant women. The detection of paternally-inherited DNA sequences that are absent in the maternal genome, e.g., Y chromosomal sequences for fetal sexing and the *RHD* gene for blood group genotyping, is well established. The recent emergence of single molecule counting technologies, such as digital polymerase chain reaction and massively parallel sequencing has allowed circulating fetal DNA to be used for the non-invasive prenatal diagnosis of fetal chromosomal aneuploidies and monogenic diseases. With large scale clinical validation and further reduction in costs, it is expected that the analysis of circulating fetal DNA will play an increasingly important role in the future practice of prenatal diagnosis.

Keywords: circulating nucleic acids; Down syndrome; next-generation sequencing; noninvasive prenatal diagnosis; plasma DNA; prenatal screening.

Introduction

Prenatal diagnosis is now an established part of modern obstetrics practice. However, the direct analysis of fetal DNA has conventionally relied on the invasive sampling of fetal tissues, e.g., via amniocentesis and chorionic villus sampling, which represents a risk to the fetus. Over the last four decades, much effort has been expended on the development of non-invasive approaches which would obviate such a risk.

In 1997, Lo et al. (1) reported that, during pregnancy, fetal DNA could be detected from the plasma and serum of pregnant women. These investigators showed that in pregnant women carrying male fetuses, Y chromosomal DNA sequences could be detected in their plasma and serum. Subsequent work using real-time quantitative polymerase chain reaction (PCR) has shown that fetal DNA represents a mean of some 3%–6% of the DNA that is present in maternal plasma (2). Subsequent work using more precise measurement technologies has shown that the actual concentration of circulating fetal DNA in maternal plasma might be double this amount, i.e., at a mean fractional concentration of some 10% (3). It has also been demonstrated that fetal DNA is cleared with a half-life of some 16 min following delivery of the baby (4). This latter observation has thus shown that circulating fetal DNA will not be carried from one pregnancy onto a subsequent one. With the above-mentioned observations, the stage has thus been set for exploring the use of fetal DNA in maternal plasma for non-invasive prenatal diagnosis.

Detection of paternally-inherited sequences

The simplest type of clinical applications using fetal DNA in maternal plasma are those that are based on the detection of sequences that the fetus has inherited from its father and which are absent in the genome of its mother. One example of such sequences involves those that are present on the Y chromosome, which are only present in a male fetus. The presence of such sequences in the pregnant mother's plasma would thus indicate the presence of a male fetus (1, 2). Non-invasive fetal sexing is useful for the prenatal diagnosis of sex-linked diseases (5). For such diseases, the disease-causing gene is present on the X chromosome and thus male fetuses, in whom only one X chromosome is present, are typically more susceptible to such disorders. Another disease in which fetal sexing is useful is congenital adrenal hyperplasia in which the use of dexamethasone treatment to reduce the risk of virilization can be rationalized to cases involving female fetuses (6).

A second example of paternally-inherited DNA sequences that have been detected in maternal plasma is the *RHD* gene (7). This application is useful for determining the RhD blood group status of a fetus carried by a RhD-negative pregnant woman. A number of large scale studies have demonstrated the robustness of this approach of fetal *RHD* genotyping (8, 9).

Because of the robustness of fetal sexing and *RHD* genotyping from maternal plasma, such tests are already in diagnostic use in a number of laboratories.

Detection of fetal chromosomal aneuploidies

The prenatal screening for Down syndrome is perhaps the most sought-after goal for research into non-invasive prenatal

*Corresponding author: Y.M. Dennis Lo, Department of Chemical Pathology, Prince of Wales Hospital, 30-32 Ngan Shing Street, Shatin, New Territories, Hong Kong SAR, PR China
E-mail: loym@cuhk.edu.hk
Received June 26, 2011; accepted September 30, 2011

Authenticated | jburke@oblon.com
Download Date | 3/26/12 11:27 AM

Exhibit W

diagnosis. However, the development of this application is much more challenging to develop than the detection of fetal sex and *RHD* genotype. One reason is that chromosome 21, the chromosome that is involved in Down syndrome, is present in both the fetus and the pregnant mother. Furthermore, one has to devise a method for measuring the dosage of the fetal chromosome 21 in the maternal plasma.

Early work for fetal aneuploidy detection from maternal plasma was typically based on the development of fetal-specific nucleic acid markers. Examples of such markers include those based on the differential DNA methylation between fetal and maternal genomic loci (10, 11) and those based on the detection of mRNA transcribed from genes that are only expressed in the fetus (e.g., in the placenta) (12–14). The first demonstrations that maternal plasma nucleic acids could be used for the direct detection of fetal trisomy 18 and trisomy 21 were achieved in 2006 (11) and 2007 (14), respectively. While these studies have shown in principle that the non-invasive detection of fetal chromosomal aneuploidies is achievable, these approaches are challenging to use in a widespread manner because these strategies are typically based on the measurement of allelic ratios of single nucleotide polymorphisms (11, 14) and each polymorphism is only informative for a proportion of the population.

Recently, Papageorgiou et al. (15) described a method based on the detection of differential DNA methylation between the fetus and mother that is independent of genetic polymorphisms. This method is based on the use of methylated DNA immunoprecipitation followed by real-time PCR analysis of multiple markers. It remains to be seen if the immunoprecipitation step and the performance of the multiple markers can be made robust enough for the approach to be used clinically.

An alternative strategy for achieving the detection of fetal chromosomal aneuploidies is through the use of highly precise methods for quantitation. Thus, when comparing a trisomy 21 with a euploid fetal cell, the proportion of the genome that is contributed by chromosome 21 in the former should be 1.5-times higher than that of the latter. Similarly, the proportional representation of DNA sequences derived from chromosome 21 that is present in the plasma of a pregnant woman carrying a trisomy 21 fetus should be higher than that of a woman carrying a euploid fetus. It has been shown that the amount of increase is half that of the fractional concentration of fetal DNA in maternal plasma (16). One approach for reaching the precision required is by methods that would allow single DNA molecules to be counted (16, 17). Indeed, the larger is the number of DNA molecules that are counted, the higher is the precision.

With the advent of massively parallel sequencing (18), one has a method that would allow millions or even billions of DNA molecules to be counted. During massively parallel sequencing, single DNA molecules are spatially separated and then the molecules are clonally amplified and sequenced, or even sequenced directly without amplification. Two groups have shown that massively parallel sequencing would allow fetal trisomy 21 to be detected non-invasively from maternal plasma (19, 20). This observation has recently been confirmed by a number of subsequent studies (21–24). In particular, the studies by Chiu et al. (22) and Ehrich et al. (23) are large scale investigations involving hundreds of samples each.

Early reports suggest that the detection of trisomy 13 and 18 are likely to be more challenging (19–21). However, this issue has recently been solved by the use of particular bioinformatics algorithms (24–26).

Thus, it is expected that the detection of the common fetal chromosomal aneuploidies will be increasingly investigated in the future. Due to the expense associated with massively parallel sequencing, it is likely that this technology will initially be used for pregnant women classified as high risk by conventional screening approaches. However, with the rapid reduction in the costs of massively parallel sequencing, it is possible that this technology would eventually be usable as a first line prenatal test.

Detection of fetal monogenic diseases

With the development of single molecule counting technologies for maternal plasma DNA, new possibilities have also emerged for the non-invasive prenatal diagnosis of fetal monogenic diseases. Thus, one is no longer confined to the detection of DNA sequences that the fetus has inherited from its father but which are absent in its pregnant mother's genome. Instead, one can detect the dosage of a particular mutant gene in the fetus's genome (17, 27). For example, let us consider the scenario of a pregnant woman who is a carrier for an autosomal recessive disorder. The ratio of the mutant and normal gene in her genome is 1:1. Let us consider the scenario that she carries a fetus that is homozygous for the mutant gene. Thus, following the release of the fetal DNA into the maternal plasma, the dosage of the mutant gene should be slightly higher than that for the normal gene. Conversely, if the fetus is homozygous for the normal gene, then the dosage of the normal gene should be slightly higher than that for the mutant gene. Finally, if the fetus is a heterozygous carrier of the disease, similar to its mother, then the dosage of the mutant and normal genes in the mother's plasma should be the same. It has been shown that digital PCR allows one to reach the precision needed for this type of analysis (27). Digital PCR is a technology whereby the DNA template for amplification is adjusted to a concentration whereby most PCR would only contain either a single or no target DNA molecule. When a large number of such amplifications are carried out simultaneously, the number of input template molecules can be measured by counting the number of positive amplification reactions (16, 28). This approach has been successfully realized for the prenatal diagnosis of hemoglobinopathies (27) and hemophilia (29).

With the development of massively parallel sequencing, it has recently been shown that one can deduce a genome-wide genetic map of the fetus using sequencing data from maternal plasma DNA and genetic information from the parents of the fetus (30). Furthermore, with the use of targeted sequencing approaches, in which the power of massively sequencing can be focused on selected genomic regions, one could have a cost-effective way of targeting multiple disease-

Authenticated | jburke@oblon.com
Download Date | 3/26/12 11:27 AM

causing mutations prevalent in a particular population (31). Discussion concerning the social, practical and ethical implications of these developments has started (32) and should be encouraged in the coming years.

Expert opinion

Since the first discovery of fetal DNA in maternal plasma in 1997, this phenomenon has been rapidly translated into clinical applications. Early diagnostic uses have focused on the detection of paternally-inherited DNA sequences, e.g., for fetal sex determination and fetal *RHD* genotyping. With the development of single molecule counting technologies, such as digital PCR and massively parallel sequencing, such applications have extended to the non-invasive prenatal diagnosis of many monogenic diseases, fetal chromosomal aneuploidies and even for fetal whole genome scanning. It is expected that maternal plasma DNA analysis will play an increasingly important role in the future practice of prenatal diagnosis.

Outlook

With a rapid reduction in the costs of massively parallel sequencing, it is expected that DNA sequencing will move to the center of investigation of the field. For the next 5–10 years, the most active area is expected to be in the use of maternal plasma DNA sequencing for the non-invasive prenatal screening of fetal chromosomal aneuploidies. It is likely that initially the technology will be used predominantly for high risk women, i.e., women who have been screened by other techniques, e.g., nuchal translucency and maternal serum biochemistry. However, with the increasing accessibility of sequencing technologies, it is expected that maternal plasma DNA sequencing will gradually move towards a first line investigation.

Highlights

- Fetal DNA is present at an approximate mean fractional concentration of 10% in maternal plasma.
- The detection of paternally-inherited sequences that are not present in the pregnant woman's genome, e.g., for fetal sex determination and *RHD* blood group genotyping, has been well established.
- Recent development of single DNA molecule counting technologies, such as digital PCR and massively parallel sequencing, allows maternal plasma DNA to be analyzed with an unprecedented degree of precision.
- Massively parallel sequencing of maternal plasma DNA has been shown in a number of studies to allow the sensitive and specific detection of fetal trisomy 21, and more recently also for trisomy 18 and 13.
- Molecular counting technologies allow the prenatal diagnosis of monogenic diseases and even allow the entire fetal genome to be scanned non-invasively.
- Large scale validation of these technologies and further reduction in the costs of these technologies are expected in the coming years.
- Discussion of the social, practical and ethical implications of these developments should be conducted for all stakeholders.

Conflict of interest statement

Author's conflict of interest disclosure: The author holds patents and has filed patent applications in the area of non-invasive prenatal diagnosis. Some of these have been licensed to Sequenom. The author has research funding, consultancy and equity relationships with Sequenom. Patents and relationships with Sequenom played no role in the study design; in the collection, analysis and interpretation of data; in the writing of the report; or in the decision to submit the report for publication.

Research funding: University grants Committee Areas of Excellence Scheme (AOE/m-04/06).

Employment or leadership: None declared.

Honorarium: None declared.

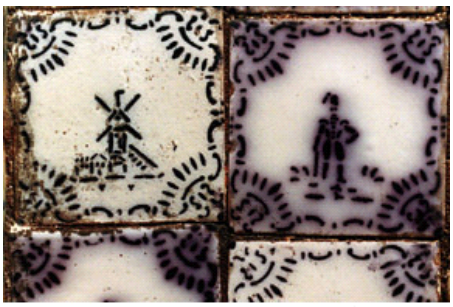
References

1. Lo YM, Corbetta N, Chamberlain PF, Rai V, Sargent IL, Redman CW, et al. Presence of fetal DNA in maternal plasma and serum. Lancet 1997;350:485–7.
2. Lo YM, Tein MS, Lau TK, Haines CJ, Leung TN, Poon PM, et al. Quantitative analysis of fetal DNA in maternal plasma and serum: implications for noninvasive prenatal diagnosis. Am J Hum Genet 1998;62:768–75.
3. Lun FM, Chiu RW, Chan KC, Leung TY, Lau TK, Lo YM. Microfluidics digital PCR reveals a higher than expected fraction of fetal DNA in maternal plasma. Clin Chem 2008;54: 1664–72.
4. Lo YM, Zhang J, Leung TN, Lau TK, Chang AM, Hjelm NM. Rapid clearance of fetal DNA from maternal plasma. Am J Hum Genet 1999;64:218–24.
5. Costa JM, Benachi A, Gautier E. New strategy for prenatal diagnosis of X-linked disorders. N Engl J Med 2002;346:1502.
6. Rijnders RJ, van der Schoot CE, Bossers B, de Vroede MA, Christiaens GC. Fetal sex determination from maternal plasma in pregnancies at risk for congenital adrenal hyperplasia. Obstet Gynecol 2001;98:374–8.
7. Lo YM, Hjelm NM, Fidler C, Sargent IL, Murphy MF, Chamberlain PF, et al. Prenatal diagnosis of fetal RhD status by molecular analysis of maternal plasma. N Engl J Med 1998;339:1734–8.
8. Finning K, Martin P, Summers J, Massey E, Poole G, Daniels G. Effect of high throughput RHD typing of fetal DNA in maternal plasma on use of anti-RhD immunoglobulin in RhD negative pregnant women: prospective feasibility study. Br Med J 2008;336:816–8.
9. Bombard AT, Akolekar R, Farkas DH, VanAgtmael AL, Aquino F, Oeth P, et al. Fetal RHD genotype detection from circulating cell-free fetal DNA in maternal plasma in non-sensitized RhD negative women. Prenat Diagn 2011;31:802–8.
10. Poon LL, Leung TN, Lau TK, Chow KC, Lo YM. Differential DNA methylation between fetus and mother as a strategy for detecting fetal DNA in maternal plasma. Clin Chem 2002;48:35–41.

Authenticated | jburke@oblon.com
Download Date | 3/26/12 11:27 AM

11. Tong YK, Ding C, Chiu RW, Gerovassili A, Chim SS, Leung TY, et al. Noninvasive prenatal detection of fetal trisomy 18 by epigenetic allelic ratio analysis in maternal plasma: theoretical and empirical considerations. Clin Chem 2006;52: 2194–202.
12. Ng EK, Tsui NB, Lau TK, Leung TN, Chiu RW, Panesar NS, et al. mRNA of placental origin is readily detectable in maternal plasma. Proc Natl Acad Sci USA 2003;100:4748–53.
13. Tsui NB, Chim SS, Chiu RW, Lau TK, Ng EK, Leung TN, et al. Systematic microarray-based identification of placental mRNA in maternal plasma: towards non-invasive prenatal gene expression profiling. J Med Genet 2004;41:461–7.
14. Lo YM, Tsui NB, Chiu RW, Lau TK, Leung TN, Heung MM, et al. Plasma placental RNA allelic ratio permits noninvasive prenatal chromosomal aneuploidy detection. Nat Med 2007;13: 218–23.
15. Papageorgiou EA, Karagrigoriou A, Tsaliki E, Velissariou V, Carter NP, Patsalis PC. Fetal-specific DNA methylation ratio permits noninvasive prenatal diagnosis of trisomy 21. Nat Med 2011;17:510–3.
16. Lo YM, Lun FM, Chan KC, Tsui NB, Chong KC, Lau TK, et al. Digital PCR for the molecular detection of fetal chromosomal aneuploidy. Proc Natl Acad Sci USA 2007;104:13116–21.
17. Chiu RW, Cantor CR, Lo YM. Non-invasive prenatal diagnosis by single molecule counting technologies. Trends Genet 2009;25:324–31.
18. Schuster SC. Next-generation sequencing transforms today's biology. Nat Methods 2008;5:16–18.
19. Chiu RW, Chan KC, Gao Y, Lau VY, Zheng W, Leung TY, et al. Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma. Proc Natl Acad Sci USA 2008;105:20458–63.
20. Fan HC, Blumenfeld YJ, Chitkara U, Hudgins L, Quake SR. Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. Proc Natl Acad Sci USA 2008;105:16266–71.
21. Chiu RW, Sun H, Akolekar R, Clouser C, Lee C, McKernan K, et al. Maternal plasma DNA analysis with massively parallel sequencing by ligation for noninvasive prenatal diagnosis of trisomy 21. Clin Chem 2010;56:459–63.
22. Chiu RW, Akolekar R, Zheng YW, Leung TY, Sun H, Chan KC, et al. Non-invasive prenatal assessment of trisomy 21 by multiplexed maternal plasma DNA sequencing: large scale validity study. Br Med J 2011;342:c7401.
23. Ehrich M, Deciu C, Zwiefelhofer T, Tynan JA, Cagasan L, Tim R, et al. Noninvasive detection of fetal trisomy 21 by sequencing of DNA in maternal blood: a study in a clinical setting. Am J Obstet Gynecol 2011;204:205 e1–11.
24. Sehnert AJ, Rhees B, Comstock D, de Feo E, Heilek G, Burke J, et al. Optimal detection of fetal chromosomal abnormalities by massively parallel DNA sequencing of cell-free fetal DNA from maternal blood. Clin Chem 2011;57:1042–9.
25. Fan HC, Quake SR. Sensitivity of noninvasive prenatal detection of fetal aneuploidy from maternal plasma using shotgun sequencing is limited only by counting statistics. PLoS One 2010;5:e10439.
26. Chen EZ, Chiu RW, Sun H, Akolekar R, Chan KC, Leung TY, et al. Noninvasive prenatal diagnosis of fetal trisomy 18 and trisomy 13 by maternal plasma DNA sequencing. PLoS One 2011;6:e21791.
27. Lun FM, Tsui NB, Chan KC, Leung TY, Lau TK, Charoenkwan P, et al. Noninvasive prenatal diagnosis of monogenic diseases by digital size selection and relative mutation dosage on DNA in maternal plasma. Proc Natl Acad Sci USA 2008;105:19920–5.
28. Vogelstein B, Kinzler KW. Digital PCR. Proc Natl Acad Sci USA 1999;96:9236–41.
29. Tsui NB, Kadir RA, Chan KC, Chi C, Mellars G, Tuddenham EG, et al. Noninvasive prenatal diagnosis of hemophilia by microfluidics digital PCR analysis of maternal plasma DNA. Blood 2011;117:3684–91.
30. Lo YM, Chan KC, Sun H, Chen EZ, Jiang P, Lun FM, et al. Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. Sci Transl Med 2010;2:61ra91.
31. Liao GJ, Lun FM, Zheng YW, Chan KC, Leung TY, Lau TK, et al. Targeted massively parallel sequencing of maternal plasma DNA permits efficient and unbiased detection of fetal alleles. Clin Chem 2011;57:92–101.
32. Greely HT. Get ready for the flood of fetal gene screening. Nature 2011;469:289–91.

Authenticated | jburke@oblon.com
Download Date | 3/26/12 11:27 AM



Photographed by Sergio Barbieri, Córdoba, Argentina.
Cúpula de Santo Domingo (detail).
Francia (Pas de Calais),
segunda mitad S.XIX. Medidas: 11 x 11 cm.

Circulating Prostate Cancer Cells Detected by Reverse Transcription-Polymerase Chain Reaction (RT-PCR): What Do They Mean?

Jose G. Moreno, MD, and Leonard G. Gomella, MD

The clinical application of "molecular staging" of prostate cancer, which uses the reverse transcriptase-polymerase chain reaction to detect cells that contain PSA or PSMA in the bloodstream, is controversial.

Background: *Molecular techniques have been developed recently to assess for circulating tumor cells. This "molecular staging" of prostate cancer uses the reverse transcription-polymerase chain reaction (RT-PCR) to detect cells that contain PSA or PSMA in the bloodstream. Currently, the clinical application of this concept is controversial.*
Methods: *The authors discuss the current status of molecular biologic methods to detect circulating prostate cancer cells. They report on the limitations of the technology and the advances that will allow the quantification of these circulating cells.*
Results: *Studies generally indicate an increasing level of PSA RT-PCR positivity as disease advances. However, reports have been significantly diverse, and there is no clear explanation for this disparity.*
Conclusions: *The determination of the "circulating prostate cancer cell load" by RT-PCR or other techniques may prove to be useful in the management of patients with prostate cancer, but questions remain to be answered before we can develop and assess new therapeutic strategies that will advance the treatment of prostate cancer before metastasis becomes evident. A better understanding of the biology of tumor cells present in the circulatory system is also needed.*

Introduction

The clinical significance of circulating prostate cancer cells has long been questioned. In 1961, Jonasson and associates[1] reported transient dissemination of prostate cancer cells into the vena cava as detected by a Papanicolaou stain in patients undergoing a rectal massage or a transurethral resection of the prostate. Interestingly, these cells reached peak detectable levels in the first 10 minutes after the massage and were cleared after 30 minutes. Other circulating cancer cell studies also suggested that these cells perish in the blood stream. Engell's colon cancer study[2] in 1955 concluded that approximately 50% of patients surviving five to nine years had venous tumor cells detected at the time of surgery, implicating that most of these cells were destroyed in the circulation.

These early studies were met with skepticism due to a lack of consensus on what constituted a circulating cancer cell. In 1969, the National Cancer Institute reviewed 81 cytologic photographs and concluded that only four fulfilled the criteria of a "suspicious" cell.[3] In 1975, Salsbury[3] conducted a 20-year literature review, and he concluded that it would be doubtful that circulating cancer cells would carry any prognostic value. Hence, this field subsequently remained relatively dormant until the advent of the molecular PCR and RT-PCR technology.

In 1992, our group revisited the concept that circulating prostate cells could be detected using molecular biologic methods. Using RT-PCR targeted at prostate-specific antigen (PSA) mRNA, we collected peripheral blood of men with newly diagnosed metastatic prostate cancer and demonstrated that there were cells in the circulation that could be detected by molecular methods.[4] Using this technology, other groups later demonstrated that 38.5% with clinically localized disease were PSA RT-PCR positive with a strong statistical correlation to pathologic stage.[5] Others have not been able to corroborate these staging results.[6]

This review examines controversial issues surrounding the significance of circulating prostate cancer cells and updates investigators on the current biotechnological advances that allow us to detect, quantitate, and isolate circulating prostate cancer cells with unprecedented accuracy. These detection methods may significantly change our understanding and treatment of this disease.

Epidemiology and Biology of Prostate Cancer Metastasis

In 1998, approximately 182,000 new prostate cancer cases will be diagnosed, of which approximately 50% will undergo radical prostatectomy based on historic trends.[7] In 22% to 45% of these cases, the operation may fail signaled by a detectable serum PSA, suggesting that up to 41,000 men will harbor persistent cancer.[7,8] Similarly, men undergoing definitive radiotherapy will likely face similar recurrences. Many patients treated with such curative intent will manifest disease recurrence at distant sites such as bone.

The development of metastasis is a complex process that includes the exodus of tumor cells from the organ, survival of cells in the circulation, and establishment of a metastatic deposit. Bone metastasis is the eventual site of spread in over 85% of cases.[9] Non-osseous cancer can invade the skeleton via three mechanisms: direct extension, retrograde venous flow, and the general arterial circulation.[10] Prostate cancer initially invades the cancellous marrow and, at autopsy, the most common metastatic sites in decreasing order are the spine, femur, pelvis, ribs, sternum, and skull.[9,10] These sites mirror the distribution of active red marrow. Hence, the pattern of prostate cancer bone invasion and distribution is most consistent with dissemination through the systemic arterial system.[9,10] Studies in animals demonstrate that the primary requisite for vertebral metastasis is a hematopoietic marrow with access gained through the arterial system.[11,12]

Exhibit X

These animal models also require that cells be injected into the left ventricle. This mechanism of tumor dissemination is used due to the fact that 98% of cancer cells injected into the tail vein are cleared in the pulmonary vascular bed in a single pass.[13] Solid tumor animal studies show that at least 10,000 tumor cells may be required for intravenous injection in order to form a single metastatic deposit.[14,15] Extrapolating this to man, at least 2 cells per mL of blood may be needed in the circulation to form a single metastatic deposit. This is an exquisitely low level of cells to detect by standard pathologic or cytologic means.

These observations led us to develop a method to detect prostate cancer cells in the circulation by using RT-PCR.[4] We demonstrated that prostate cancer cells could be detected in the circulation of patients with metastatic prostate cancer and that women and negative controls failed to detect any prostate cells in the circulation. It should be stressed that these initial observations neither quantitated nor positively characterized the prostate cells as cancerous. However Ts'o and colleagues[16] were recently able to isolate and characterize these as "prostate cancer cells" in the peripheral blood of patients with advanced prostate cancer at a concentration of 1 to 20 cells per 1 mL of whole blood.

The body of literature suggests that when these cells are detected through a properly designed RT-PCR protocol, they are of a malignant phenotype. Since our first 1992 PSA RT-PCR report, there has been a rapid expansion in the literature to use this technique to stage prostate cancer. This attempt to detect cells outside the prostate has been generally referred to as "molecular staging."[6]

Summary of Clinical Studies Using RT-PCR to Detect Circulating Prostate Cancer Cells

Year	Investigators	Normal Controls	Organ Confined	Non-Organ Confined	Advanced Metastatic
1992	Moreno et a[14]	0%	NA	NA	33%
1994	Katz et a[15]	0%	9%	65%	77.8%
1994*	Israeli et al[32]	2%	0%	6.7%	25%
1994	Seiden et al[18]	0%	NA	NA	31%
1995	Ghossein et al[19]	0%	16%	30%	35%
1996	Sokoloff et al[20]	0%	59%	72%	88%
1996	Melchior et al[33]	0%	16%	27%	71%
1997	Olsson et al[22]	0%	12%	27%	71%
1997	Corey et al[34]	0%	22%	16%	46%
1998**	Grasso et al[23]	0%	5%	19%	64%

* PSMA primers

** PSA and PSMA primers

Molecular Staging With RT-PCR

RT-PCR has been used to detect prostate cells in the peripheral blood, lymph nodes, and bone marrow and at the operative field.[6] However, most clinical molecular staging studies have been performed on peripheral blood samples.

The principal steps of molecular staging involve a peripheral phlebotomy, whole blood fractionation, RNA extraction, reverse transcription, PCR amplification, and detection of the PCR products.[6,17] Approximately 6 to 8 mL of peripheral venous blood is initially subjected to centrifugation on a Ficoll density gradient to isolate mononuclear cells. Our work demonstrated that PSA-expressing cells and mononuclear cells cofractionate in the Ficoll-gradient "buffy coat" in clinical samples from patients.[4] In the latest technique, blood is drawn directly into a CPT Vacutainer tube (Becton Dickinson Vacutainer Systems, Franklin Lakes, NJ) that allows fractionation of the blood without a time-consuming Ficoll gradient. Cells of the isolated buffy coat are lysed and mRNA is isolated by chemical separation using a modifications of the guanidinium isothiocyanate-phenol-chloroform extraction, a standard molecular biology technique.[6] RT-PCR is performed by an automated programmable thermal controller that provides the various temperature conditions necessary for unique primers to select PSA mRNA and generate ample quantities of cDNA for molecular analysis. Presence of a specific amplified gene product can be examined by agarose gel electrophoresis and can be verified by gene sequencing, Southern blot analysis, or restriction enzyme digest analysis. The amplification of a specific PSA gene product should correlate with the presence of PSA-expressing prostatic cells in the peripheral circulation.

The sensitivity of detection by RT-PCR has been estimated as 1 cell in 10^7 to 10^8 mononuclear cells.[6] PSA RT-PCR positivity in metastatic prostate cancer cases ranges between 31% and 88%. In negative normal controls, the false-positive rate has been mostly zero, but recent reports have suggested a range between 10% and 20%, suggesting that these highly sensitive assays may detect a potentially problematic background expression of PSA (M. O'Hara, unpublished data, 1998). It should be noted that primers to prostate-specific membrane antigen (PSMA) have also been used in the RT-PCR assay for prostate cancer, with the results not significantly improved over PSA.[6]

The results of published clinical studies on PSA RT-PCR in the peripheral blood are summarized in the Table. In the first clinical report describing this approach, we examined patients with stage D0-3 (occult, nodal, or bone metastasis) to determine if PSA-expressing cells could be detected in the peripheral venous blood. The RT-PCR assay was positive in one third of the cases (4 of 12 cases), implicating the presence of circulating prostatic cells.[18] In control patients (men with clinically diagnosed benign prostatic hypertrophy or women), the RT-PCR assay was negative (0 of 17 cases).[4] Using a modified PSA RT-PCR assay, researchers at Columbia University found that all negative controls tested negative, while 14 (77.8%) of the 18 metastatic cases tested positive.[5] More importantly, they found a statistically significant correlation between pathologic stage and PSA RT-PCR that was positive in 25 (38.5%) of the 65 radical prostatectomy cases.

However, other investigators such as Seiden et al[18] and Ghossein et al[19] have failed to correlate PSA RT-PCR with pathologic stage with metastatic cases testing

positive in 31% to 35% of cases. On the other end of the spectrum, Sokoloff et al[20] found 88% PSA RT-PCR positivity in metastatic cases, but the 59% positive rate in organ-confined patients was excessively high. Based on 82 consecutive radical prostatectomy cases, Ignatoff et al[21] concluded that PSA RT-PCR correlated weakly with serum PSA, Gleason score, and pathologic stage. However, they found a strong positivity rate in metastatic cases and a nearly perfect negative correlation in normal controls.

Overall, most studies report a 0% PSA RT-PCR rate in negative control cases, while the positive rate in the metastatic group is much higher -- between 31% and 88%.[6] Indeed, there is an increase PSA RT-PCR positivity rate with advancing stage from T1-2 to T3-4 and on up to metastatic disease. However, most reports do not yet recommend PSA RT-PCR as a practical staging tool.[6,21] Olsson et al[22] claim that a positive PSA RT-PCR is a significant predictor of recurrence-free survival time. In their series form Columbia University, 14 (7%) of 190 PSA RT-PCR-negative patients had a serum PSA greater than 0.2 ng/mL within five years after radical prostatectomy compared with 36 (28%) of 127 PSA RT-PCR-positive patients. Lastly, it is possible that a combination of primers may improve the overall staging accuracy of RT-PCR. Preliminary work from the Cleveland Clinic suggests that combining RT-PCR for PSA and PSMA may improve the staging accuracy.[23]

Despite some of these promising reports, PSA RT-PCR has fallen short of being predictive enough to be used at the bedside. Several factors have contributed to the inconsistent PSA RT-PCR reports in peripheral blood. Sample handling and timing of the blood draw is critical. Ubiquitous nucleases in our environment make mRNA exquisitely susceptible to degradation such that the starting sample could already be compromised.[6] There have been suggestions in the literature that a variety of surgical procedures ranging from biopsies to radical prostatectomies may lead to hematogenous dissemination of prostatic cells.[6] In an analysis of the effect of transrectal ultrasound prostate biopsy, we found a 9% seroconversion rate, ie, 9% of patients who had no detectable PSA-expressing cells in their peripheral circulation prior to surgery tested RT-PCR positive postoperatively.[24] In a separate study, a significant portion of patients (25%) seroconverted to a positive RT-PCR at three months after surgery.[6]

Another paramount factor is assay technical variability, eg, primer selection, PCR product detection methods, and cross contamination, all of which affect sensitivity and specificity. Even in ideal conditions, if 10 identical PCR reactions are performed in parallel and simultaneously, there can be a threefold difference in the quantity of PCR end products. In a comparison of three independent RT-PCR protocols, Slawin and associates[25] found a large degree of variation in the delectability of PSA in peripheral blood of patients with clinically localized disease. Most PCR experts agree that in a strict sense, PCR reactions cannot be precisely quantitated, which may be the Achilles' heel of this assay.

Technical issues outlined here indicate that an accurate assessment of the clinical utility of the RT-PCR assay requires strict quality control standards and specifications at multiple levels ranging from selection criteria of patients, sample acquisition, storage and processing, and a detailed analysis of RT-PCR assay conditions. Optimization of all parameters from specimen collection and handling through the RT-PCR protocol is currently the focus of several studies. For example, investigators at Baylor College reported that PSA RT-PCR predicated pathologic stage when done postoperatively compared with preoperatively.[26] Ylikoski and colleagues[27] have recently presented a quantitative PSA RT-PCR assay that ambitiously aims to target one of the significant weakness of this system. Our group has participated in a multi-institutional consortium that aims to address many of these difficult technical issues; slow but steady progress is being made.[28]

Isolation, Quantitation, and Characterization of Circulating Prostate Cancer Cells

Both flow cytometry and now RT-PCR have allowed these cells to be discovered in the peripheral circulation. Further quantification and characterization of these cells represent the next challenge in determining the significance of these circulating cells. Ts'o et al[16] at Johns Hopkins University recently isolated and characterized prostate cancer cells from the peripheral blood of eight men with hormone-refractory metastatic prostate cancer. Cells were isolated by fractionating the whole blood over a Percoll gradient, and the isolated cells were identified with PSA immunofluorescent staining and fluorescent *in situ* hybridization. Approximately 1 to 20 prostate cancer cells per 1 mL of blood were identified with positive PSA cytologic staining and chromosome 7 and 8 aneuploidy, which is common for prostate cancer cells but rare for lymphocytes. Similarly, Moss et al[29] reported positive PSA staining cells in fractionated whole blood from prostate cancer patients, which correlated with positive PSA RT-PCR results but not with clinical stage.

Using a combination of density gradient and magnetic cell sorting, Brandt et al[30] identified 10 to 100 prostate cancer cells per 40 mL (both single cells and in clusters) in 10 prostate cancer patients. Positive isolation and identification were performed with cytokeratin and PSA antibodies.

Racila et al[31] reported on a new immunomagnetic enrichment system by which epithelial cancer cells can be isolated, identified, and quantitated in breast cancer and in a limited number of prostate cancer patients. Unlike magnetic bead technology, these workers used a magnetic ferrofluid that may be more effective at capturing cells present at low concentrations. Immunomagnetic separation of epithelial cells from peripheral blood is remarkably fast and simple to perform. Target epithelial cells are captured by nanometer-size iron particles (ferrofluid) linked to an epithelial cell-specific antibody when whole blood is incubated with this ferrofluid. This mixture is subjected to magnetic separation at room temperature. The whole epithelial enrichment process takes one hour and requires no centrifuges or special equipment outside of the proprietary magnet (Immunicon Corp, Huntingdon Valley, Pa). This assay can detect and enumerate 1 epithelial cell in 1 mL of whole human blood with flow cytometry performed on the ferrofluid enriched specimen. Peripheral blood from 30 breast and 3 prostate cancer patients contained a substantially greater number of circulating epithelial cells compared with 13 normal controls. These epithelial cells stained positive with cytokeratin and mucin-1, and their morphology was consistent malignancy. Preliminary work by these investigators has shown a correlation between quantity of circulating prostate cancer cells and disease progression (L. Terstappen, unpublished data, 1998). A newer system under development will allow these cells to be captured with the ferrofluid and directly observed under a fluorescent microscope, thus eliminating the need for a costly flow cytometer.

Conclusions

A growing body of evidence is now validating the existence and importance of circulating prostate cancer cells. These cells were initially detected grossly by flow cytometry and now by RT-PCR. Further enhancements in cell capture may allow quantification and further characterization of these cells that are attempting to establish a metastatic lesion. It is clear that even with PSA, Gleason grade, and sophisticated imaging tools, we are unable to clinically distinguish curable organ-confined prostate cancer from non-organ-confined incurable disease. Since our initial report on the use of PSA RT-PCR for detecting circulating prostate cancer cells, several studies have attempting to use this assay to define pathologic organ confined from non-organ confined disease. Most studies have indicated an increasing level of PSA RT-PCR positivity as one advances in stage. However, reports have been widely disparate, and there is no one clear explanation for this disparity. Many other solid tumors (breast, melanoma, colon cancer, and others) are also being studied using RT-PCR technology.[17] Workers in a number of disciplines studying these different solid tumors may unravel the intricacies of these important new technologies.

Important questions regarding the biology of cells in the circulatory system will need to be answered. In addition, a quantitative reproducible assay needs to be developed that can specifically detect those select prostate cancer cells harboring the machinery that will enable them to survive the circulation and grow at a distant site. Once these two important issues are resolved, we may then have a powerful tool to develop and assess new therapeutic strategies and to enable us to rapidly

advance the treatment of prostate cancer in much the same way human immunodeficiency virus treatment has been advanced with sophisticated blood tests that measure the viral load. One day it may be possible to quantify a circulating prostate cancer cell load. This determination will allow therapeutic decisions long before the PSA shows its characteristic and ominous elevation in patients long before metastasis become evident.

Dr Gomella and Dr Moreno have joint patent rights with Thomas Jefferson University to some RT-PCR technology.

References

1. Jonasson O, Long L, Roberts S, et al. Cancer cells in the circulating blood during operative management of genitourinary tumors. *J Urol*. 1961;85:1.

2. Engell HC. Cancer cells in the blood: a five to nine year follow up study. *Ann Surg*. 1959;149:457-461.

3. Salsbury AJ. The significance of the circulating cancer cell. *Cancer Treat Rev*. 1975;2:55-72.

4. Moreno JG, Croce CM, Fischer R, et al. Detection of hematogenous micrometastasis in patients with prostate cancer. *Cancer Res*. 1992;52:6110-6112.

5. Katz AE, Olsson CA, Raffo AJ, et al. Molecular staging of prostate cancer with the use of an enhanced reverse transcriptase-PCR assay. *Urology*. 1994;43:765-775.

6. Gomella LG, Raj GV, Moreno JG. Reverse transcriptase polymerase chain reaction for prostate specific antigen in the management of prostate cancer. *J Urol*. 1997;158:326-337.

7. Parker SL, Tong T, Bolden S, et al. Cancer statistics, 1997. *CA Cancer J Clin*. 1997;47:5-27.

8. Oesterling JE. Using prostate-specific antigen to eliminate unnecessary diagnostic tests: significant worldwide economic implications. *Urology*. 1995;46:26-33.

9. Jacobs SC. Spread of prostatic cancer to bone. *Urology*. 1983;21:337-344.

10. Thrall JH, Ellis BI. Skeletal metastases. *Radiol Clin North Am*. 1987;25:1155-1170.

11. Arguello F, Baggs RB, Duerst RE, et al. Pathogenesis of vertebral metastasis and epidural spinal cord compression. *Cancer*. 1990;65:98-106.

12. Suzuki T, Shimizu T, Kurokawa K, et al. Pattern of prostate cancer metastasis to the vertebral column. *Prostate*. 1994;25:141-146.

13. Weiss L, Orr FW, Honn KV. Interactions of cancer cells with the microvasculature during metastasis. *FASEB J*. 1988;2:12-21.

14. Fidler IJ. Critical factors in the biology of human cancer metastasis: twenty-eighth G.H.A. Clowes Memorial Award lecture. *Cancer Res*. 1990;50:6130-6138.

15. Liotta LA, Kleinerman J, Saidel GM. Quantitative relationships of intravascular tumor cells, tumor vessels and pulmonary metastases following tumor implantation. *Cancer Res*. 1974;34:997-1004.

16. Ts'o PO, Pannek J, Wang ZP, et al. Detection of intact prostate cancer cells in the blood of men with prostate cancer. *Urology*. 1997;49:881-885.

17. Raj G, Moreno JG, Gomella LG. Detection of circulating cancer cells using RT-PCR technology. *Cancer*. 1998;82:1909-1914.

18. Seiden MV, Kantoff PW, Krithivas K, et al. Detection of circulating tumor cells in men with localized prostate cancer. *J Clin Oncol*. 1994;12:2634-2639.

19. Ghossein RA, Scher HI, Gerald WL, et al. Detection of circulating tumor cells in patients with localized and metastatic prostatic carcinoma: clinical implications. *J Clin Oncol*. 1995;13:1195-1200.

20. Sokoloff M, Patel A, Tso CL, et al. Surgical manipulation does not result in hematogenous dissemination of prostate cancer cells. *J Urol*. 1996;155:528. Abstract.

21. Ignatoff JM. RT-PCR for staging: promising but impractical. *Cont Urol*. 1998;13-22.

22. Olsson CA, de Vries G, Buttyan R, et al. Reverse transcriptase-polymerase chain reaction assays for prostate cancer. *Urol Clin North Am*. 1997;24:367-378.

23. Grasso YZ, Gupta MK, Levin HS, et al. Combined nested RT-PCR assay for prostate-specific antigen and prostate-specific membrane antigen in prostate cancer patients: correlation with pathological stage. *Cancer Res*. 1998;58:1456-1459.

24. Moreno JG, O'Hara M, Long JP, et al. Transrectal ultrasound-guided biopsy causes hematogenous dissemination of prostate cells as determined by RT-PCR. Urology. 1997;49:515-520.

25. Slawin K, O'Hara SM, Song W, et al. Comparison of results obtained using different RT-PCR-PSA assay methods on a single set of clinical specimens. *J Urol*. 1996;155:417. Abstract.

26. Nguyen CV, Dickason R, Song W, et al. RT-PCR for hK2: a new assay for detecting circulating prostate cells. *J Urol*. 1997;157:113. Abstract.

27. Ylikoski A, Sjoroos M, Iitia A, et al. Quantitation RT-PCR assay for detection of prostate specific antigen mRNA. *J Urol*. 1998;159: 295. Abstract.

28. Vessella RL, Lange PH, Blumenstein BA, et al. Multi-center RT-PCR PSA clinical trial for pre-operative staging of prostate cancer. *J Urol*. 1998;159:1124. Abstract

29. Moss TJ, Reseda CA, Logethetis CJ, et al. Immunocytochemical (ICC) analysis: an attractive alternative to reverse-transcriptase-polymerase chain reaction (RT-PCR) in molecular staging of prostate cancer. *J Urol*. 1998;159:1126. Abstract.

30. Brandt B, Junker R, Griwatz C, et al. Isolation of prostate-derived single cells and cell clusters from human peripheral blood. *Cancer Res*. 1996;56:4556-4561.

31. Racila E, Euhus D, Weiss AJ, et al. Detection and characterization of carcinoma cells in the blood. *Proc Natl Acad Sci U S A*. 1998; 95:4589-4594.

32. Israeli RS, Miller WH Jr, Su SL, et al. Sensitive nested reverse transcription polymerase chain reaction detection of circulating prostatic tumor cells: comparison of prostate-specific membrane antigen and prostate-specific antigen-based assays. *Cancer Res*. 1994;54: 6306-6310.

33. Melchior S, Corey E, Ellis W, et al. Highly sensitive reverse transcriptase polymerase chain reaction (RT-PCR) assay for prostate specific antigen (PSA) messenger RNA detects early tumor spread in patients with clinically localized prostate cancer (CaP). *Proc Annu Meet Am Assoc Cancer Res*. 1996;37:A617. Abstract.

34. Corey E, Arfman EW, Oswin MM, et al. Detection of circulating prostate cells by reverse transcriptase-polymerase chain reaction of human glandular kallikrein (hK2) and prostate-specific antigen (PSA) messages. *Urology*. 1997;50:184-188.

From the Department of Urology, Kimmel Cancer Center, Thomas Jefferson University, Philadelphia, Pa.

Address reprint requests to Leonard Gomella, MD, at the Department of Urology, Kimmel Cancer Center, Thomas Jefferson University, 1025 Walnut St, Philadelphia, PA 19107.

Back to Cancer Control Journal Volume 5 Number 6

Plasma DNA K-*ras* Mutations in Patients with Gastrointestinal Malignancies

HUGH E. MULCAHY, JACQUELINE LYAUTEY, CHRISTINE LEDERREY, XU QI CHEN, FRANÇOIS LEFORT, VALERI VASIOUKHIN, PHILIPPE ANKER, ELSPETH M. ALSTEAD, MICHAEL J. G. FARTHING, AND MAURICE STROUN

Digestive Diseases Research Centre, St. Bartholomew's and Royal London School of Medicine and Dentistry, London, England, and Département de Biochimie et de Physiologie Végétale, Université de Genève, Genève, Switzerland

The plasma of normal subjects and those with a variety of acute and chronic diseases is known to contain small quantities of DNA that circulate freely in the bloodstream.[1] Patients with a variety of malignant diseases also have DNA circulating within blood plasma.[2] This DNA is measurable by conventional means; however, even though cancer patients tend to have greater quantities of plasma DNA than patients with nonmalignant diseases,[3] quantitative measurements are neither sensitive nor specific enough for diagnostic purposes.

Since the introduction of the polymerase chain reaction, somatic gene alterations have been found in a wide range of tumors. Mutations of the K-*ras* gene are associated with a number of gastrointestinal malignancies including colorectal cancer,[4] and the highest incidence of K-*ras* alterations occurs in those patients with pancreatic cancer, approximately 90% of whom have an alteration at codon 12 of this gene.[5] In addition to their presence within tumor tissues, K-*ras* alterations have also been detected in the feces of patients with colorectal cancer[6] and within pancreatic juice, bile, duodenal juice, and feces of patients with pancreatic cancer.[7,8] Our aim was to detect K-*ras* alterations in the plasma of patients with colorectal and pancreatic cancer.

We studied 14 patients with colorectal cancer (mean 63 years, range 41–78; 4 females) and 21 patients with pancreatic cancer (mean 60 years, range 41–88; 12 females). In addition, we studied 11 healthy subjects and 3 patients with chronic pancreatitis as negative controls and the SW480 cell line as a positive control. For colorectal cancer specimens, blood was collected prior to surgery, and tumor specimens were taken at the time of operation. The 21 patients with pancreatic cancer had blood taken either before diagnosis (4 patients) or at the time of diagnosis (17 cases). All 21 pancreatic cancer patients had unresectable disease. Diagnosis was made by a combination of ultrasound scanning, computerized tomography, endoscopic retrograde cholangiopancreatography, and pancreatic biopsy or aspirate.

Plasma DNA was extracted from blood samples using simplified technique.[9,10] A number of molecular techniques were used to confirm the presence of K-*ras* alterations. PCR amplification using mutant-specific alleles (MASA-PCR) was used to amplify 10 ng of the first exon of the K-*ras* gene in 100 μL reactions. Thirty-five to 45 cycles (94°C for 1 min, 55°C to 62°C for 90 s, 72°C for 1 min) were performed, followed by a 7-min extension at 72°C. Products were visualized following electrophoresis on 8% polyacrylamide gels. PCR products were then diluted and further

Exhibit Y

amplified with primers complementary to the wild-type gene and to ALA, VAL, SER, ASP, and CYS mutations. MASA-PCR was performed at least twice on DNA samples extracted independently from each plasma sample. Final PCR products were again visualized on polyacrylamide gels and sequenced.

PCR using restriction fragment length polymorphisms (RFLP-PCR) and subsequent sequencing was also used to detect alterations. This included a two-stage restriction assay with 50 μL reactions, a hot-start and initial denaturation of 10 min followed by 30 PCR cycles (96°C for 1 min, 55°C for 1 min, 73°C for 30 s), and a 7-min extension at 73°C. Products were digested with the Mva I restriction enzyme and visualized on a high-resolution electrophoresis system. Products were diluted and reamplified using conditions identical to the first PCR. Following a second Mva I digestion, the products were subjected to direct sequencing to determine the precise mutations.

Furthermore, the first exon of the K-*ras* gene extracted from the plasma DNA specimens of those patients displaying codon 12 K-*ras* mutations was cloned[6] and sequenced in order to confirm the results obtained with the two other techniques.

In relation to colorectal cancer, the wild-type K-*ras* gene was found in all 14 tumor specimens, in all 14 plasma specimens, and in all negative controls. Wild-type K-*ras* was not found in the SW480 cell line, which is homozygous for the GTT mutation. We found mutated K-*ras* in 7 of the 14 colorectal cancers (50%), of which 6 were situated at the second base of codon 12. The most common alteration was from GGT (wild-type glycine) to GTT (valine). Other alterations included GCT (alanine), GAT (aspartic acid), and TGT (cysteine). One tumor contained both a GTT and a GCT mutation. K-*ras* mutations were unrelated to patient age and sex, and tumor site, size, and stage. Of interest is that mutations were found not only in the plasma of patients with advanced disease, but also in patients with stage I and II disease.

Mutations were not detected in the plasma of patients whose primary tumor was negative for K-*ras* or in plasma samples obtained from normal subjects. Six of 7 patients (86%) with tumor DNA K-*ras* alterations displayed similar mutations in DNA extracted from plasma samples. These results were confirmed by performing a second series of experiments using an RFLP-PCR technique on new DNA extractions from corresponding original tumor and plasma samples and by cloning the first exon of the K-*ras* gene of these samples.

For pancreatic cancer patients, as for colorectal cancer patients, PCR analysis and gel electrophoresis showed that the wild-type K-*ras* gene was present in all tumor and plasma specimens, and that only mutant DNA was present in the SW480 cell line. K-*ras* alterations, detected using the RFLP-PCR reaction and sequencing, were not found in DNA extracted from the plasma of 3 patients with chronic pancreatitis or from normal subjects. PCR product sequencing identified mutant K-*ras* in the plasma of 17 of 21 cancer patients (81%). All mutations were situated at the first or second base of codon 12. Similar to colorectal cancer patients, the most common alteration was from GGT (wild-type) to GTT (valine). Alterations from GGT to TGT (cysteine), GAT (aspartic acid), and CGT (arginine) were also found, and a combination of two mutations was seen in the plasma of a single patient. Biopsy tissues were available for 10 patients, and plasma and tumor DNA alterations corresponded in each case in which mutations were present.

Our findings that K-*ras* mutations are present in the plasma of most patients with colorectal and pancreatic cancer whose tumors contain a mutated gene suggest that plasma DNA mutations arise from the primary tumor. The exact mechanism is not yet known. The importance of necrosis, apoptosis, and active release are analyzed and discussed elsewhere in this volume.[11]

Although our results are of interest from a scientific viewpoint, any clinically useful diagnostic test would require further assays to detect additional gene rearrangements such as p53 alterations or changes within microsatellites. Recent work has shown that microsatellite alterations are present in DNA extracted from the plasma and serum of small cell lung cancer and head and neck cancer patients,[12,13] indicating that plasma DNA may harbor a range of genetic mutations. More recently, p53 and microsatellite alterations have also been found in the plasma of colorectal cancer patients,[14,15] suggesting that a simultaneous search for distinctly different mutations within the same specimen would increase the diagnostic range of a plasma DNA test. Aberrant methylation of tumor suppressor genes such as found in plasma of liver[16] or lung[17] cancer patients might also be detected in the plasma of gastrointestinal malignancy patients. More sensitive PCR techniques may also allow even smaller quantities of mutant DNA to be characterized. Further research, focusing on the association between precancerous and early gastrointestinal disease with the appearance of mutant plasma DNA, will help define its role as a diagnostic and screening test for cancer in the future.

REFERENCES

1. KOFFLER, D., V. AGNELLO, R. WINCHESTER & H. KUNKEL. 1973. The occurrence of single stranded DNA in the serum of patients with SLE and other diseases. J. Clin. Invest. **52:** 198–204.
2. LEON, S.A., B. SHAPIRO, D.M. SKLAROFF & M.J. YAROS. 1977. Free DNA in the serum of cancer patients and the effect of therapy. Cancer Res. **37:** 646–670.
3. SHAPIRO, B., M. CHAKRABARTY, E.M. COHN & S.A. LEON. 1983. Determination of circulating DNA levels in patients with benign or malignant gastrointestinal disease. Cancer **51:** 2116–2120.
4. FEARON, E.R. & B. VOGELSTEIN. 1990. A genetic model for colorectal tumorigenesis. Cell **61:** 759–767.
5. ALMOGUERA, C., D. SHIBATA, K. FORRESTER, J. MARTIN & M. PERUCHO. 1988. Most human carcinomas of the exocrine pancreas contain mutant c-K-*ras* genes. Cell **53:** 549–554.
6. SIDRANSKY, D., T. TOKINO, S.R. HAMILTON, K.W. KINZLER, B. LEVIN, P. FROST & B. VOGELSTEIN. 1992. Identification of *ras* oncogene mutations in the stool of patients with curable colorectal tumors. Science **256:** 102–105.
7. KONDO, H., K. SUGANO, N. FUKAYAMA *et al.* 1994. Detection of point mutations in the K-*ras* oncogene in pure pancreatic juice for diagnosis of pancreatic carcinoma. Cancer **73:** 1589–1594.
8. O'MAHONY, S., M. LONGFELLOW, M.J. MCMAHON, A.T.R. AXON & P. QUIRKE. 1995. Detection of c-Ki-*ras* mutations in bile samples from patients with pancreatic and biliary cancers. J. Clin. Pathol. Mol. Pathol. **48:** M316–M318.
9. DAVIS, R., H.E. MULCAHY, S. BUSTIN, E.M. ALSTEAD, S. DORUDI & M.J.G. FARTHING. 1999. Circulating mutant DNA in patients with pancreatic cancer: a simplified extraction and detection procedure. Gastroenterology **114:** A402.
10. ANKER, P., H. MULCAHY, X.Q. CHEN & M. STROUN. 1999. Detection of circulating tumour DNA in the blood (plasma/serum) of cancer patients. Cancer Metastasis Rev. **18:** 65–73.

11. STROUN, M., P. MAURICE, V. VASIOUKHIN, J. LYAUTEY, C. LEDERREY, F. LEFORT, A. ROSSIER, X.Q. CHEN & P. ANKER. 2000. The origin and mechanism of circulating DNA. This volume.
12. CHEN, X.Q., M. STROUN, J.L. MAGNENAT, L.P. NICOD, A.M. KURT, J. LYAUTEY, C. LEDERREY & P. ANKER. 1996. Microsatellite alterations in plasma DNA of small cell lung cancer patients. Nat. Med. **2:** 1033–1035.
13. NAWROZ, H., W. KOCH, P. ANKER, M. STROUN & D. SIDRANSKY. 1996. Microsatellite alterations in serum DNA of head and neck cancer patients. Nat. Med. **2:** 1035–1037.
14. HIBI, K., C.R. ROBINSON, S. BOOKER, L. WU, S.R. HAMILTON, D. SIDRANSKY & J. JEN. 1998. Molecular detection of genetic alterations in the serum of colorectal cancer patients. Cancer Res. **58:** 1405–1407.
15. KOLBLE, K., O.M. ULLRICH, H. PIDDE, B. BARTHEL, J. DIERMANN, B. RUDOLPH, M. DIETEL, P.M. SCHLAG & S. SCHERNECK. 1999. Microsatellite alterations in serum DNA of patients with colorectal cancer. Lab Invest. **79:** 1145–1150.
16. ESTELLER, M., M. SANCHEZ-CESPEDES, M. ROSELL, D. SIDRANSKY, S.B. BAYLIN & J.G. HERMAN. 1999. Detection of aberrant promoter hypermethylation of tumor suppressor genes in serum DNA from non-small cell lung cancer patients. Cancer Res. **59:** 67–70.
17. WONG, I.H., Y.M. LO, J. ZHANG, C.T. LIEW, M.H. NG, N. WONG, P.B. LAI, W.Y. LAU, N.M. HJELM & P.J. JOHNSON. 1999. Detection of aberrant p16 methylation in the plasma and serum of liver cancer patients. Cancer Res. **59:** 71–73.

© 1996 Nature Publishing Group http://www.nature.com/naturemedicine

patients with myelodysplastic syndrome or acute myelogenous leukaemia. *Br. J. Haematol.* **86**, 774–779 (1994).
12. Sorenson, G.D. *et al.* Soluble normal and mutated DNA sequences from single-copy genes in human blood. *Cancer Epidemiol. Biomark. Prev.* **3**, 67–71 (1994).
13. Nagel, S. *et al.* Somatic mutations detected by mini- and microsatellite DNA markers reveal clonal intratumor heterogeneity in gastrointestinal cancers. *Cancer Res.* **55**, 2866–2870 (1995).
14. Nawroz, H., Koch, W., Anker, P., Stroun, M. & Sidransky, D. Microsatellite alterations in serum DNA of head and neck cancer patients. *Nature Med.* **2**, 1036–1038 (1996).

Microsatellite alterations in serum DNA of head and neck cancer patients

HOMAIRA NAWROZ[1], WAYNE KOCH[1], PHILIPPE ANKER[2], MAURICE STROUN[2] & DAVID SIDRANSKY[1]

[1]*Department of Otolaryngology — Head and Neck Surgery, Division of Head and Neck Cancer Research, Johns Hopkins University School of Medicine, 818 Ross Research Building, 720 Rutland Avenue, Baltimore, Maryland 21205-2196*
[2]*Laboratory of Plant Biochemistry and Physiology, Pavillon des Isotopes, Faculty of Science, University of Geneva, 20 boulevard d'Yvoy, 1211 Geneva 4, Switzerland*
Correspondence should be addressed to D.S.

Microsatellite DNA alterations are an integral part of neoplastic progression and are valuable as clonal markers for the detection of human cancers[1–3]. Moreover, recent evidence suggests that senescent tumor cells may release DNA into the circulation, which is subsequently carried by and therefore enriched in the serum and plasma[4,5]. We tested 21 patients with primary head and neck squamous cell carcinoma (HNSCC) by polymerase chain reaction (PCR)-based microsatellite analysis of DNA from lymphocytes and paired serum samples. Patients were scored for alterations as defined by the presence of new alleles (shifts) or loss of heterozygosity (LOH) in serum at each of 12 markers and then compared with primary tumor DNA. Six out of 21 patients (29%) were found to have one or more microsatellite alterations in serum precisely matching those in the primary tumors. All six patients had advanced disease (stage III or IV); five of these patients had nodal metastases, three later developed distant metastases, and four died of disease. Microsatellite analysis of serum represents a novel method for the detection of circulating tumor cell DNA. If these results are confirmed in larger studies, microsatellite markers may be useful in assessing tumor burden in cancer patients.

We retrospectively analyzed lymphocyte, serum and tumor DNA from 21 head and neck cancer patients. These patients were chosen from our tumor bank because all three DNA sources were available for complete analysis, and the serum samples had been collected before surgical resection of head and neck cancer. We selected 12 microsatellite markers to detect shifts or LOH. Eight markers were chosen on 9p, 3p and 17p, because these chromosomal arms show the highest percentage of LOH and appear to harbor tumor suppressor gene loci involved early in the progression of head and neck cancer[6,7] (see the Methods section). Furthermore, two trinucleotide (*D14S50* and *DRPLA*) and two tetranucleotide (*D21S1245* and *FgA*) markers, recognized as being prone to microsatellite instability and located on loci commonly altered in cancers, were used in the study for increased sensitivity in the detection of shifts[2]. LOH was scored if the allele signal was reduced to less than 50% of normal intensity, and shifts were called if there was an obvious new allele compared with normal lymphocyte DNA. Representative alterations (LOH and shift) in tumor and serum as compared with lymphocyte DNA are shown in Fig. 1.

A tumor-specific microsatellite shift, represented by a novel allele after gel electrophoresis, can still be seen when tumor DNA is diluted between 1:500 to 1:1000 with normal DNA (ref. 2). We hypothesized that we might see these shifts derived from primary tumor cell DNA in the serum; however, we were quite surprised to see clear LOH in serum DNA. The first patient that had such provocative results was an 80-year-old man diagnosed with $T_3N_0M_0$ (see Table 1) glottic cancer who underwent a total laryngectomy in January 1993. He had no evidence of disease on 3 months' follow-up. However, in September 1993, he was diagnosed with recurrence of tumor in the right neck with the mass surrounding the right carotid artery, and he died of regional disease in November 1994. The serum DNA of this patient displayed a clear LOH (Fig. 1). This result was unexpected and reminiscent of the clear LOH and shifts seen in the urine of patients diagnosed with bladder cancer[1].

After microsatellite analysis of all specimens was completed, clinical data were correlated with the results. The stage, outcome and microsatellite analyses of serum and tumor DNA of the 21 patients (H1–H21) in the study are listed in Table 1. All six patients (H1–H6) that had microsatellite alterations in the serum DNA demonstrated identical alterations in the primary tumor DNA (Table 2). Four out of six patients displayed alterations in more than one locus, and all of these patients had advanced disease (stage III–IV). In this small group, four patients went on to die from cancer, one patient has terminal cancer with metastases and one patient has no evidence of disease at 3 years' follow-up. Five patients had nodal metastases and three of them later developed distant metastases, one patient to lung and bone and the

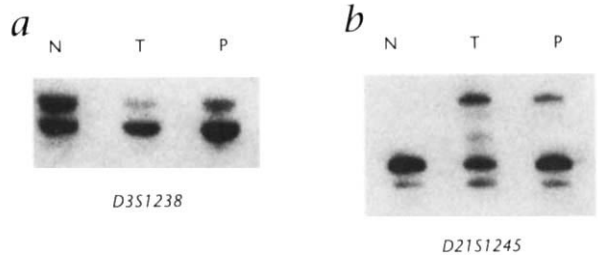


Fig. 1 Autoradiographs from microsatellite analysis. Representative microsatellite analysis of patient serum (P) with corresponding HNSCC tumor (T) and lymphocyte (N) DNA with markers indicated at the bottom. *a*, In patient H2, serum and tumor DNA reveal loss of the upper allele with marker *D3S1238* compared to normal lymphocyte DNA. *b*, In patient H1, serum and tumor DNA exhibit a shift (presence of new top allele) with marker *D21S1245*.

Exhibit Z

© 1996 Nature Publishing Group http://www.nature.com/naturemedicine

other two patients to lung and liver (Table 1).

Conversely, another nine patients (H6–H15) had advanced stage cancer, but no microsatellite alterations in their serum DNA. Six of these patients (H6–H11) had successful resections and were free of disease on long-term follow-up (>1 year); two of them died within 2 years of diagnosis from regional recurrence and one was lost to follow-up. Seven of these patients, including those with a good prognosis, exhibited LOH or shifts in their primary tumor DNA but had no evidence of alterations in serum DNA. Lack of positive findings in serum DNA was also seen in 6 out of 21 patients that had stage I and II cancer, all with good prognoses except for one patient who had a recurrence 7 years later and died. Of note, three patients (H9, H14 and H20) displayed no microsatellite alterations in their primary tumor DNA with the markers we tested. In this small study, the data are statistically significant for a positive serum test as predictor of future distant metastases by the Fisher's exact test ($P = 0.015$); nevertheless, any conclusions of predictive ability or clinical utility require verification with larger populations and well-defined cohorts.

Table 1 Microsatellite analysis and clinical outcome in head and neck cancer patients

ID#	Age/sex	Stage	Tumor LOH/shift		Serum LOH/shift		O/S	Clinical outcome
H1	84/M	$T_3N_1M_0$	+	+	–	+	10	DOD
H2	80/M	$T_4N_0M_0$	+	–	+	–	12	DOD
H3	51/M	$T_3N_2M_0$	+	+	+	+	36	NED
H4	50/M	$T_4N_{2b}M_0$	+	–	+	–	44	AWD, distant mets
H5	63/M	$T_2N_1M_0$	+	+	+	+	20	DOD, distant mets
H6	59/F	$T_3N_1M_0$	+	–	+	–	04	DOD, distant mets
H7	46/M	$T_2N_{2a}M_0$	+	–	–	–	14	NED
H8	85/F	$T_4N_0M_0$	+	–	–	–	34	NED
H9	62/M	$T_3N_3M_0$	–	–	–	–	32	NED
H10	72/F	$T_3N_0M_0$	+	–	–	–	12	NED
H11	41/M	$T_2N_{2b}M_0$	+	–	–	–	36	NED
H12	56/M	$T_3N_{2c}M_0$	+	–	–	–	29	NED
H13	45/M	$T_3N_{2b}M_0$	+	–	–	–	05	DOD
H14	91/M	$T_3N_{2b}M_0$	–	–	–	–	0	LTF
H15	64/M	aud.canal	+	–	–	–	10	DOD
H16	66/M	$T_1N_0M_0$	+	–	–	–	30	NED
H17	46/M	$T_2N_0M_0$	+	+	–	–	36	NED
H18	49/M	$T_1N_0M_0$	+	–	–	–	36	NED
H19	40/M	$T_1N_0M_0$	+	+	–	–	56	NED
H20	61/F	$T_1N_0M_0$	–	–	–	–	88	DOD
H21	76/M	$T_2N_0M_0$	+	–	–	–	44	NED

Clinical data and results of microsatellite analysis. Column 2 shows corresponding age at the time of diagnosis and sex. Stage of HNSCC, according to the tumor-node-metastasis (TNM) staging system of the American Joint Committee on Cancer[20], is listed in column 3. Microsatellite data of tumor and plasma DNA. are indicated by + for presence of LOH or shift and – for absence of these findings. O/S states the overall survival (survival regardless of disease status) period in months for each patient. Clinical outcome abbreviations: DOD, died of disease; AWD, alive with disease; NED, no evidence of disease; mets, metastases; and LTF, lost to follow-up.

Previous studies have proposed that tumor DNA is released into the circulation and is enriched in plasma and serum. In one study, the serum of cancer patients was found to contain approximately 4 times the amount of free DNA as that of normal controls[8]. Using a sensitive radioimmunoassay, others found twice the concentration of DNA in patients with metastatic cancer as in cancer patients without evidence of metastatic disease[9]. We also found elevated serum DNA levels in our cancer patients compared with normal controls (see the Methods). Recently, N-*ras* point mutations have been found in plasma DNA of patients with acute myelogenous leukemia[10], and K-*ras* point mutations have been shown in plasma DNA

Table 2 Detailed microsatellite analysis of patients with alterations in plasma

Patient	*IFNA*	*D9S200*	*D9S161*	*D9S156*	*D3S1284*	*D3S1238*	*D14S50*	*D21S1245*	*CHRNB1*	*D17S786*	*FGA*	*DRPLA*
H1 T	LOH	Ret	Ret				Ret	Shift			Ret	
S	Ret	Ret	Ret				Ret	Shift			Ret	
H2 T		LOH			Ret	LOH	LOH		Ret		Ret	
S		LOH			Ret	LOH	LOH		Ret		Ret	
H3 T		Ret	LOH		Ret	Ret		LOH	Shift			Ret
S		Ret	LOH		Ret	Ret		LOH	Shift			Ret
H4 T		LOH	Ret	Ret	Ret	LOH		LOH		Ret	Ret	Ret
S		Ret	Ret	Ret	Ret	LOH		LOH		Ret	Ret	Ret
H5 T			LOH	LOH	LOH	Ret		Shift	Ret		Ret	Ret
S			LOH	LOH	LOH	Ret		Shift	Ret		Ret	Ret
H6 T	Ret			Ret	Ret	LOH	Ret	Ret	Ret		Ret	LOH
S	Ret			Ret	Ret	LOH	Ret	Ret	Ret		Ret	LOH

Summary of six patients with LOH (loss of heterozygosity) or shifts (new alleles) demonstrated after microsatellite analysis of serum (S) DNA with comparison to analysis in tumor (T). All samples were compared to constitutive normal DNA from lymphocytes (see text). Empty boxes represent noninformative markers and Ret (retention) indicates informative markers without evidence of loss.

© 1996 Nature Publishing Group http://www.nature.com/naturemedicine

of patients with colorectal cancer[11] and pancreatic cancer[12].

In an accompanying study, a much higher frequency of microsatellite alterations in plasma was identified in patients with small cell lung cancer (SCLC)[13]. Our study supports the idea of tumor DNA enrichment in blood serum and plasma. Moreover, serum DNA microsatellite alterations were always identical to alterations in the primary tumor DNA. The higher frequency of plasma alterations in small cell lung cancer may reflect the much higher frequency of clinical metastases in SCLC patients compared with head and neck cancer patents. It is plausible that in bulky head and neck tumors, cell lysis by necrosis or even apoptosis leads to the release of naked DNA into the circulation. For large tumors, this phenomenon may occur more frequently because of local angiogenesis and necrosis. Although a surprising finding, our own studies suggest that tumor DNA readily survives in various bodily fluids including urine, stool and sputum. However, much work needs to be done to understand the enrichment of tumor DNA in serum and plasma.

Head and neck cancer remains a morbid and often fatal disease[14]. Large tumor bulk and tumor extension are predictors of local regional recurrence and poor outcome. We recently demonstrated that molecular detection of occult neoplastic cells in surrounding surgical margins was a strong predictor of local regional recurrence resulting in a significant decrease in overall survival[15]. This small study reveals that serum DNA microsatellite alterations are detectable in patients with advanced stage disease, usually with poor prognosis including future development of distant metastasis. Although shifts and LOH were both seen in this study, theoretically, the detection of new alleles by electrophoresis should be more sensitive for analysis of plasma DNA. Identification of clear LOH strongly favors the hypothesis that tumor DNA is enriched and, in fact, the predominant form of DNA in the plasma. Such analysis of plasma DNA may be useful in follow-up of cancer patients receiving medical or surgical treatment. Although quite preliminary, the findings are provocative and require further investigation of serum and plasma DNA in cancer patients for assessment of tumor burden, metastatic status and overall prognosis.

Methods

Sample collection and DNA isolation. Tumors from HNSCC patients obtained fresh from surgical resection and blood by venipuncture were collected from patients at the Johns Hopkins University Medical Institutions with prior consent. To obtain serum, clotted blood specimens were centrifuged at low speed for 5 min, and the serum was stored at –80 °C before DNA extraction. Tumor tissue was frozen and microdissected as previously described[16]. Lymphocytes, tumor tissue and serum were digested in SDS and proteinase K at 48 °C overnight, followed by phenol/chloroform extraction and ethanol precipitation of DNA (ref. 17). The mean concentration of DNA in our cancer patients was 110 ± 50 ng per ml serum, and 10 µl was usually sufficient for robust microsatellite analysis. The concentration of serum DNA from normal controls ranges from 0 to 100 ng/ml (ref. 8, 9).

PCR amplification. Oligonucleotide markers for microsatellite analysis were obtained from Research Genetics (Huntsville, Alabama) and included *IFNA, D9S156, D9S161, D9S200* on 9p; *D3S1238, D3S1284* on 3p; *D17S786* and *Chrnb1* on 17p. Trinucleotide and tetranucleotide primers used included the following: *D14S50*; (F) 5'–AACACCCCTAATTCACCACT–3' and (R) 5'–ATGATTCCACAAGATGGCAG–3', *D21S1245*; (F) 5'–GTCAGTATTACCCTGTTACCA–3' and (R) 5'–GTTGAGGATTTTTGCATCAGT–3', *DRPLA*; (F) 5'–CACAGTCTCAACACATC–3' and (R) 5'–CCTCCAGTGGGTGGGGAAATGCTC–3', *FgA*; (F) 5'–CCATAGGTTTTGAACTCACAG–3' and (R) 5'–CTTCTCAGATTCCTTCTTGACAC–3'. One primer from each set was end labeled with (γ-^{32}P) ATP (Amersham) using T4 polynucleotide kinase (New England Biolabs). PCR amplification was performed with 30–60 ng DNA as described previously[17,18]. Products were separated in 8% denaturing urea-polyacrylamide-formamide gels[19] followed by autoradiography. Loss of heterozygosity was called if the ratio of one allele was significantly decreased (>50%) in tumor or serum DNA compared with normal DNA from lymphocytes.

Clinical correlation. Clinical outcome data was obtained from the Johns Hopkins Head and Neck Cancer Tumor Registry and by chart review. Fisher's exact test was used to compare results from plasma analysis with clinical outcome parameters.

Acknowledgments

This work was supported by a Lung Cancer Specialized Program of Research Excellence (SPORE) from the National Cancer Institute (CA 58184-01) and Oncor, Inc., in Gaithersburg, Maryland. Under an agreement between Oncor and the Johns Hopkins University, D.S. is entitled to a share of sales royalty received by the university from Oncor. Under that agreement, the university and D. S. also have received Oncor stock which, under university policy, cannot be traded until two years after the first commercial sales of the products related to this research. D. S. also serves as a member of the Scientific Advisory Board of OncorMed, Inc., an Oncor subsidiary, which is commercializing some of Oncor's technology. The terms of this arrangement have been reviewed and approved by the university in accordance with its conflict-of-interest policies.

RECEIVED 15 MAY; ACCEPTED 24 JULY 1996

1. Mao, L. *et al.* Molecular detection of primary bladder cancer by microsatellite analysis. *Science* **271**, 659–662 (1996).
2. Mao, L. *et al.* Microsatellite alterations as clonal markers for the detection of human cancer. *Proc. Natl. Acad. Sci. USA* **91**, 9871–9875 (1994).
3. Merlo, A. *et al.* Frequent microsatellite instability in primary small cell lung cancer. *Cancer Res.* **54**, 2098–2101 (1994).
4. Stroun, M. *et al.* Neoplastic characteristics of the DNA found in the plasma of cancer patients. *Oncology* **46**, 318–322 (1989).
5. Stroun, M., Anker, P., Lyautey, J., Lederrey, C. & Maurice, P. Isolation and characterization of DNA from the plasma of cancer patients. *Eur. J. Clin. Oncol.* **23**, 707–712 (1987).
6. Nawroz, H. *et al.* Allelotype of head and neck squamous cell carcinoma. *Cancer Res.* **54**, 1152–1155 (1994).
7. Califano, J. *et al.* A genetic progression model for head and neck cancer: Implications for field cancerization. *Cancer Res.* **56**, 2488–2492 (1996).
8. Shapiro, B., *et al.* Determination of circulating DNA levels in patients with benign or malignant gastrointestinal disease. *Cancer* **51**, 2116–2120 (1983).
9. Leon, S.A. *et al.* Free DNA in the serum of cancer patients and the effect of therapy. *Cancer Res.* **37**, 646–650 (1977).
10. Vasioukhin, V. *et al.* Point mutations of the N-*ras* gene in the blood plasma: DNA of patients with myelodysplastic syndrome or acute myelogenous leukaemia. *Br. J. Haematol.* **86**, 774–779 (1994).
11. Vasioukhin, V. *et al.* K-*ras* point mutations in the blood plasma DNA of patients with colorectal tumors. in *Challenges of Modern Medicine: Biotechnology Today.* (eds. Verna, R. & Shamoo, A.) 141–150 (Ares-Serano Symposia Publications, Rome, 1994).
12. Sorenson, G.D. *et al.* Soluble normal and mutated DNA sequences from single-copy genes in human blood. *Cancer Epidemiol. Biomark. Prev.* **3**, 67–71 (1994).
13. Chen, X.Q. *et al.* Microsatellite alterations in plasma DNA of small cell lung cancer patients. *Nature Med.* **2**, 1034–1036 (1996).
14. Vokes, E.E., Weichselbaum, R.R., Lippman, S.M. & Hong, W.K. Head and neck cancer [review]. *N. Engl. J. Med.* **328**, 184–194 (1993).
15. Brennan, J. *et al.* Molecular assessment of histopathological staging in squamous cell carcinoma of the head and neck. *N. Engl. J. Med.* **332**, 429–435 (1995).
16. Boyle, J.O. *et al.* The incidence of p53 mutations increases with progression of head and neck cancer. *Cancer Res.* **53**, 4477–4480 (1993).
17. Van der Riet, P. *et al.* Progression of basal cell carcinoma through loss of chromosome 9q and inactivation of single p53 allele. *Cancer Res.* **54**, 25–27 (1994).
18. Sidransky, D. *et al.* Identification of p53 gene mutations in bladder cancers and urine samples. *Science* **252**, 706–709 (1991).
19. Litt, M., Hauge, X. & Sharma, V. Shadow bands seen when typing polymorphic dinucleotide repeats: Some causes and cures. *Biotechniques* **15**, 280–284 (1993).
20. Beahrs, O., Henson, D., Hutter, R. & Myers, M. eds. *Manual for Staging of Cancer.* 3rd edn. 27–62 (Lippincott, Philadelphia, 1988).

Proc. Nat. Acad. Sci. USA
Vol. 69, No. 7, pp. 1685–1689, July 1972

Excretion of Deoxyribonucleic Acid by Lymphocytes Stimulated with Phytohemagglutinin or Antigen

(mitogenesis/pulse labeling/density gradient centrifugation/DNase)

JOHN C. ROGERS,* DAVID BOLDT,† STUART KORNFELD,† SISTER ANN SKINNER,† AND C. ROBERT VALERI*

* Naval Blood Research Laboratory, Chelsea, Massachusetts 02150; and † Departments of Medicine and Biochemistry, Washington University School of Medicine, St. Louis, Missouri 63110

Communicated by Carl V. Moore, May 1, 1972

ABSTRACT When human lymphocytes are cultured in the presence of phytomitogens, 70–90% of the cells undergo blast transformation and synthesize DNA. However, less than 40% of these lymphocytes actually undergo mitosis while 35–90% of the newly synthesized DNA is excreted into the media. The release of DNA by the cells is selective since experiments with [^{14}C]uridine indicate that RNA is not lost into the culture media. DNA excretion occurs under many culture conditions. The excreted DNA has an estimated molecular weight of 3 to 12 $\times$ 10^6 as determined by gel filtration on Sepharose 2B. It forms a single sharp peak at a density of 1.055 g/cm^3 when examined by sucrose density gradient centrifugation, suggesting that the DNA is complexed to protein or lipid.

Although phytohemagglutinin (PHA) is known to be a mitogen, the number of mitoses actually counted after addition of PHA to lymphocyte cultures is relatively small. For example, Nowell found only 10% after 18 hr of incubation with colchicine at the peak of DNA synthesis despite morphologic transformation of up to 90% of these cells (1). Such PHA-transformed lymphocytes are known to be synthesizing DNA, since they are labeled when pulsed with radioactive thymidine and studied by autoradiography. The discrepancy between many cells making DNA and few caught dividing has been attributed to asynchrony of the cultured cells (2).

In PHA-stimulated lymphocyte cultures, one might expect to find a significant increase in cell number resulting from mitosis, as well as a concomitant increase in the total DNA content of the culture. The reports dealing with this question do not consistently support this expectation. Schellekens and Eijsvoogels noted that total cell number declined 17% the first day, while DNA content dropped 25%; by day 3, the cell count was still only 89% of the initial values, and the DNA content was 96% (3). Hirschhorn *et al.* reported a decline in DNA content during a 48-hr culture of PHA-stimulated lymphocytes to 75% of the initial value (4). Clearly, if most PHA-transformed lymphocytes in these experiments underwent mitosis, a large number of daughter cells must have died and lysed. In contrast, Loeb and Agarwal measured a PHA-induced increase in DNA to nearly twice that of unstimulated cultures by 72 hr (5). Their results could be explained if each stimulated lymphocyte underwent a single mitosis and suggest that lymphocyte cultures reported in the former papers were complicated for some reason by excessive cell lysis. However, this simple explanation is confounded by the data of Polgar and Kibrick (6). These authors report experiments in which greater than 90% of lymphocytes stimulated for 24 hr with PHA ultimately became blasts, as judged morphologically; when grown continuously in the presence of [^{3}H]thymidine, a maximum of 63% of the total number of lymphocytes present were labeled, as revealed by autoradiography on day 5, while on day 10 only 6% of the lymphocytes were labeled, yet the number of viable cells on day 10 was 57% of the original number. Cumulative in-

Abbreviations: PHA, phytohemagglutinin—E-PHA, erythroagglutinating and L-PHA, leukoagglutinating phytohemagglutinin, respectively.

TABLE 1. *DNA content of PHA-stimulated cells related to [^{3}H]thymidine uptake and loss*

Day	DNA (μg per culture)	[^{3}H]Thymidine acid-precipitable cpm in cells	[^{3}H]-Thymidine acid-precipitable cpm in medium
0	26.9	0	0
3	22.0 (23.5)	23,600 (305)	0
4	23.1 (26.6)	15,600 (246)	11,000
5	21.8 (22.7)	11,100 (240)	13,000
6	23.5 (21.3)	5,300 (175)	26,000

Quadruplicate cultures of 3.3 $\times$ 10^6 lymphocytes in 2 ml of medium were stimulated with 24 μg of E-PHA. On day 3, 3 μCi [^{3}H]thymidine (6.7 Ci/mmol) was added; at the end of a 4-hr pulse, 8 ml of minimal essential medium was added, the cultures were centrifuged at 500 $\times$ *g* for 5 min, the cell pellets were suspended in 2 ml of medium, and grown for up to 3 more days. On the days noted above, duplicate tubes were harvested. Culture tubes were centrifuged at 500 $\times$ *g* for 10 min. On one set, DNA content was determined by the diphenylamine reaction (11). Acid-precipitable [^{3}H]thymidine cpm were determined on the other set: 0.2 ml of medium was precipitated with 2.8 ml of ice-cold 5% Cl_3CCOOH. The remainder of the medium was removed and the cells were washed with 10 ml of ice-cold 0.9% NaCl. The cell pellet was precipitated with 3 ml of 5% Cl_3CCOOH and sonicated for 15 sec with a Biosonix IV sonicator at low intensity, 30% maximum. Cl_3CCOOH precipitates were centrifuged at 27,000 $\times$ *g*, 4°, for 20 min, washed with 2 ml of 5% Cl_3CCOOH, digested in 0.5 ml of Nuclear Chicago Solubilizer, and counted in 10 ml of Bray's solution. Values obtained for identical cultures not stimulated with PHA are in parentheses.

Exhibit AA

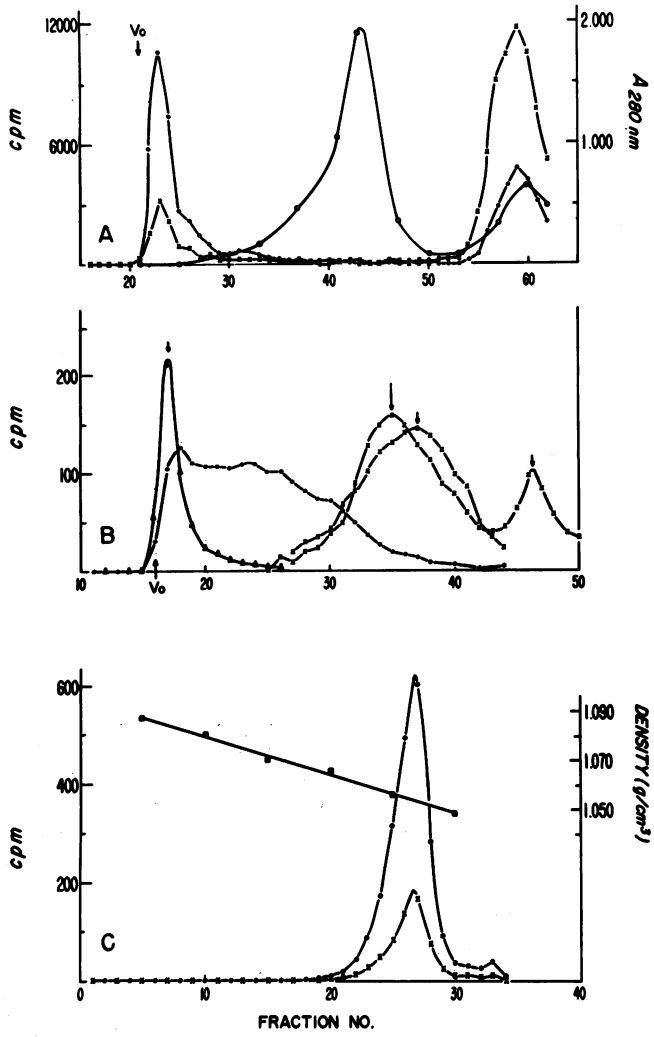


FIG. 1. (*A*) Sepharose 6B gel filtration chromatography. Three cultures containing 2 × 10^6 lymphocytes per culture were stimulated with E-PHA, pulsed with [^{3}H]thymidine and [^{14}C]-uridine on day 3, and harvested on day 6 as in Table 3; the cells in these cultures lost 85% of their ^{3}H-labeled DNA between day 3 and day 6, 80% of which was recovered in acid-precipitable form in the culture medium. The 6 ml of medium was made 0.1% with sodium dodecyl sulfate, applied to a Sepharose 6B column, 2.2 × 55 cm, and eluted at room temperature with Tris–NaCl–EDTA buffer [0.2 M NaCl–0.01 M Tris·HCl (pH 7.4)–0.001 M EDTA]. 3-ml Fractions were collected. All acid-precipitable ^{3}H and ^{14}C that had been applied eluted in the front peak. ^{3}H and ^{14}C eluted in fractions 55–62 were not acid-precipitable and represent traces of the thymidine and uridine pulse that were not washed out on day 3. Fractions 22–24 were pooled. ^{3}H cpm (●——●); ^{14}C cpm (×——×); $A_{280\ nm}$ (○——○). *Arrow: V_0.*

(*B*) Sepharose 2B gel filtration chromatography. A 0.9 × 60-cm column of Sepharose 2B equilibrated with Tris–NaCl–EDTA–0.1% sodium dodecyl sulfate was calibrated at 23–24° with ^{3}H-labeled *E. coli* DNA (gift of Dr. W. Sly) to determine the void volume and with ^{14}C-labeled HeLa cell rRNA, 32S (molecular weight 2.1 × 10^6), 28S (1.7 × 10^6), and 18S (0.7 × 10^6) (15) (gifts of Dr. A. Kumar) as molecular weight markers. 0.5 ml of the pooled fractions (22–24) made 0.1% with sodium dodecyl sulfate was chromatographed and collected in 0.9-ml fractions. Symbols are as in (*A*). *Arrows* (*left* to *right*) designate *E. coli* DNA, 32S RNA, 28S RNA, and 18S RNA.

(*C*) Sucrose density gradient equilibrium centrifugation. 0.5 ml of the pooled fraction (1*A*) made 0.1% with sodium dodecyl sulfate was layered on 17 ml of a 5–20% linear gradient of sucrose in Tris–NaCl–EDTA–sodium dodecyl sulfate and centrifuged at 4° in a SW27.1 rotor at 27,000 rpm for 16 hr. 0.5-ml Fractions were collected from the bottom of the tube. Density, g/cm^3 (■——■); other symbols are as in (*A*).

corporation of the [^{3}H]thymidine into the cells in acid-precipitable form also peaked on day 5, but then declined so that on day 11 the cumulative counts approached those of unstimulated cells. The authors' interpretation of this work was that many PHA-transformed cells do not synthesize DNA and do not divide, and that most daughter cells produced by mitosis (those cells labeled with [^{3}H]thymidine) die.

These confusing data motivated the following experiments in an attempt to clarify the fate of DNA synthesized by lymphocytes stimulated with phytomitogens and antigens.

MATERIALS AND METHODS

All experiments described were performed on human peripheral-blood lymphocytes that were isolated as described (7) and cultured in Eagle's minimal essential medium (pH 7.4) supplemented with 20% autologous serum, 4 mM glutamine, 50 units/ml of penicillin, and 50 μg/ml of streptomycin. The cultures were maintained at 37° in a 5% CO_2 atmosphere. Erythroagglutinating (E-) and leukoagglutinating hemagglutinin (L-PHA) from *Phaseolus vulgaris* were prepared from Difco PHA-P according to the method of Weber *et al.* (8). Pokeweed mitogen was obtained from Grand Island Biological Co. and reconstituted according to the supplier's instructions. Concanavalin-A was purchased from Calbiochem. RNase was obtained from Sigma Chemical Co. and DNase (RNase free) from Worthington Biochemicals. [*methyl*-^{3}H]Thymidine and [^{14}C]uridine were purchased from New England Nuclear Co. ^{3}H-labeled *Escherichia coli* DNA was a gift of Dr. W. Sly. The mitotic index was obtained by exposure of the cultures for 4 hr to 0.02 μg/ml colcemid at 37°. Metaphase chromosomes were prepared by the method of Mendelsohn *et al.* (9). To obtain accurate cell counts of lymphocytes cultured in the presence of E-PHA, the aggregated cells were dispersed by the addition of fetuin (5 mg/ml), a potent haptene inhibitor of the binding of E-PHA to lymphocytes (10).

RESULTS AND DISCUSSION

Loss of labeled DNA

When lymphocytes were incubated with E-PHA over a 6-day culture period, the number of viable cells remained within ±20% of the initial count. At least 95% of the cells counted on any day were viable as judged by trypan blue exclusion; by day 3, about 80% of the lymphocytes were transformed into blasts. The amount of DNA per culture, as measured by the diphenylamine reaction (11), also remained constant. However, cells pulse labeled with [^{3}H]thymidine on day 3 (the peak time of thymidine incorporation), washed, and cultured again for 3 more days lost 35–90% of the acid-precipitable [^{3}H]thymidine incorporated on day 3 (Table 1). The acid-precipitable [^{3}H]thymidine lost from the cells can be accounted for in the culture medium. The mitotic index was determined at daily intervals on these cultures; an increase in the number of mitoses occurred with the maximum being 3% after 4 hr of exposure to colcemid on day 3.

Lymphocytes stimulated with Concanavalin-A, *Phaseolus vulgaris* L-PHA, and mumps antigen (data in Table 2) as well as with pokeweed mitogen, *Lens culinaris* phytohemag-

glutinin (12), and PPD (purified protein derivative) antigen (data not presented) similarly excreted DNA into the culture medium.‡ When lymphocyte cultures are maintained for 3 days or more in the absence of phytomitogens or antigens, a small percentage of the cells spontaneously begin to synthesize DNA. Much of the [^{3}H]thymidine incorporated into DNA by these lymphocytes is also excreted into the medium (Table 2).

TABLE 2. *Effect of different mitogens on DNA excretion*

Mitogen	Acid-precipitable cpm: Day of pulse 3	Day of pulse 5	Day experiment completed 6	Day experiment completed 8
E-PHA (36 μg):				
Cells	52,600	—	24,300	—
Medium	0	—	16,200	—
L-PHA (30 μg):				
Cells	34,200	—	16,700	—
Medium	0	—	9,570	—
Con-A (20 μg):				
Cells	115,000	—	39,100	—
Medium	215	—	52,700	—
Mumps antigen (0.005 ml):				
Cells	—	45,900	—	14,000
Medium	—	0	—	22,700
None:				
Cells	3,040	—	713	—
Medium	0	—	302	—
None:				
Cells	—	6,950	—	1,770
Medium	—	0	—	7,630

8.5 × 10^6 lymphocytes in 2 ml of medium were cultured as described in Table 1. The cells were pulsed with 3 μCi of [^{3}H]-thymidine (6.7 Ci/mmol) for 4 hr on the days indicated. After the cells were washed and resuspended in fresh medium containing the appropriate mitogen, the cultures were incubated for 3 more days. Acid-precipitable counts in cells and medium were determined as in Table 1.

Selective DNA excretion

To confirm that intracellular, acid-precipitable [^{3}H]thymidine was in DNA, the DNA from cells pulsed and harvested on day 3 was extracted by a slight modification of the method outlined by Gottlieb *et al.* (13). Nucleic acid precipitated by cold ethanol after one phenol extraction was treated with alkali (1 N NaOH, 22°, and 18 hr) and then precipitated with Cl_3CCOOH. More than 90% of the radioactivity initially present was recovered in the precipitate, suggesting that the [^{3}H]thymidine was in DNA. The selective nature of DNA excretion was determined by experiments in which stimulated cells were pulsed on day 3 with both [^{3}H]thymidine and [^{14}C]uridine. The results of five consecutive experiments performed on lymphocytes from different individuals are outlined in Table 3. In each instance the cultures were pulsed on day 3 and then followed for 3 more days to compare the fate of the two isotopes. About 10–15% of the acid-precipitable ^{3}H was lost from the cells during the first 24 hr after the pulse label (experiments 3–5). In contrast, during this same 24-hr period, the amount of intracellular acid-precipitable ^{14}C increased in each instance. The explanation for this finding is that there is a large intracellular pool of ^{14}C-labeled acid-soluble nucleotides present after the 4-hr pulse on day 3 (data not presented); by day 4 the radioactive nucleotides in this pool have been used by the cells and converted to an acid-precipitable form. No such pool of ^{3}H is found. The acid-precipitable ^{14}C counts shown in Table 3 are expressed as being either alkali labile (presumably RNA) or alkali stable (presumably DNA). While most of the acid-precipitable ^{14}C on day 3 was in RNA, and while the number of counts in RNA increased by day 4, much of the ^{14}C incorporated into acid-precipitable form during the 24 hr after the pulse was alkali stable, and presumably DNA. The apparent conversion by lymphocytes of [^{14}C]uridine to [^{14}C]-thymidine and its incorporation into DNA has been noted (14).

Between days 4 and 6, the ^{14}C counts in cellular DNA declined, while the total ^{14}C counts in DNA (cells + media) increased, indicating the continued conversion of RNA degradation products into DNA. During this time a significant amount of ^{14}C-labeled DNA appears in the culture media while essentially no [^{14}C]RNA is released. These data suggest that the loss of isotope from RNA is due to normal turnover by cells, which must be viable in order to convert the RNA degradation products into DNA. Further, if loss of DNA were due to cell lysis an equal amount of RNA should be released into the medium. As shown in Table 3, essentially no acid-precipitable RNA was released by the cells. These observations are confirmed by the data in Table 4. Treatment for 2 hr with DNase solubilized 70% of both ^{3}H and ^{14}C radioactivity from the medium, while RNase and alkaline hydrolysis were without effect.

In an attempt to exclude the possibility that DNA excretion occurred because the culture medium was deficient in some factor necessary for complete DNA synthesis, other experiments were performed in medium TC-199, TC-199 buffered with 0.01 M HEPES (*N*-2-hydroxyethylpiperazine-*N'*-2-ethane sulfonic acid; pH 7.6), or TC-199 supplemented with thymine (5 μg/ml); all gave results similar to those experiments with minimal essential medium. Use of [^{3}H]deoxycytidine rather than [^{3}H]thymidine does not alter the pattern of DNA excretion. When mouse L-cells growing in log phase in minimal essential medium under conditions similar to those of lymphocyte cultures were pulsed for 2 hr with [^{3}H]thymidine, washed, and resuspended in fresh medium, none of the radioactivity incorporated during the pulse was lost from the cells during the following 72 hr.

Characterization of excreted DNA

As demonstrated in Fig. 1*A*, all acid-precipitable ^{3}H and ^{14}C radioactivity present in the culture medium eluted in a peak at the void volume on Sepharose 6B gel filtration chromatography. The medium applied to this column was taken

‡ While we did not determine the percent blast transformation in the cultures stimulated with PPD and mumps antigen, others have found that in such cultures about 2% of the lymphocytes are blast-like at 48 hr, while up to 50% of the cells present at day 6 are transformed (16). This increase was thought to represent multiple-cell divisions by a small number of antigen-responsive cells, although recruitment of other populations of lymphocytes could not be excluded (16).

TABLE 3. *Preferential loss of DNA from lymphocytes pulsed with [^3H]thymidine and [^{14}C]uridine*

Exp.	Day 3			Day 4			Day 6		
		^{14}C			^{14}C			^{14}C	
	^3H	Alkali stable	Alkali labile	^3H	Alkali stable	Alkali labile	^3H	Alkali stable	Alkali labile
1. Cells	23,200	782	5,090		—	—	13,800	7080	7190
Medium	0	0	0		—	—	8,320	1310	161
2. Cells	28,800	791	7,200		—	—	6,340	3380	4830
Medium	0	0	0		—	—	20,600	3050	160
3. Cells	23,700	2230	8,760	20,400	6560	12,200	9,910	7340	9600
Medium	0	0	0	1,380	325	0	11,200	3400	21
4. Cells	55,400	2750	10,400	46,500	7340	14,000	24,100	7410	7570
Medium	0	0	0	3,650	285	249	20,800	2780	484
5. Cells	104,000	3700	17,800	84,500	9660	21,600	17,500	7130	7650
Medium	0	0	0	12,500	950	558	67,000	7000	514

2 to 8 × 10^6 lymphocytes in 2 ml of medium were cultured with 36 μg of E-PHA (Exp. 1, 3, and 4) or 20 μg of concanavalin A (Exp. 2 and 5), pulsed for 4 hr on day 3 with 3 μCi of [^3H]thymidine (6.7 Ci/mmol) and 1 μCi of [^{14}C]uridine (59.8 mCi/mmol), and washed and resuspended as in Table 1. On appropriate days duplicate tubes were harvested. Cl_3CCOOH precipitates were dissolved in 1 ml of 1 N NaOH and incubated at room temperature (23°). At the end of 18 hr, the tubes were neutralized with HCl, precipitated with Cl_3CCOOH, and centrifuged at 37,000 × *g* for 20 min. 1 ml of the supernatant (alkali labile radioactivity) was counted in 15 ml of Bray's solution. The pellet (alkali-stable radioactivity) was dissolved and counted as in Table 1; cpm are corrected for quenching and channel crossover.

from 3 cultures containing a total of 6 × 10^6 cells. These cells lost 85% of their ^3H-labeled DNA between days 3 and 6 (data not shown). Since 1 × 10^6 lymphocytes contain roughly 8 μg of DNA (4), lysis of 85% of the cells in these cultures would have released about 40 μg of DNA. A DNA concentration of 3 μg/ml present in the 12 ml encompassing the front peak in Fig. 1*A* would have been detected by measurement of the absorption at 260 nm; no significant A_{260} or A_{280} accompanied this peak, even though it contained 80% of the labeled DNA lost from the cells. This observation is added evidence that cell lysis alone cannot explain the loss of DNA into the culture medium. The amount of DNA actually lost must be fairly small compared to the total amount of DNA in the cells, which would explain the apparent paradox noted in Table 1; namely, much of the acid-precipitable [^3H]-thymidine left the cells, but the total DNA in the cells measured colorimetrically remained constant.

TABLE 4. *Treatment of acid-precipitable ^3H and ^{14}C radioactivity in the culture medium with DNase, RNase, and alkaline hydrolysis*

	Acid-soluble cpm		Acid-precipitable cpm	
	^3H	^{14}C	^3H	^{14}C
$^3H/^{14}C$ alone	7	7	1670	528
$^3H/^{14}C$ + RNase	5	5	1560	516
$^3H/^{14}C$ + DNase	7	5	1680	580
$^3H/^{14}C$ + Mg^{++} + DNase	1260	357	551	187
$^3H/^{14}C$ alkaline hydrolysis	21	15	1560	526
E. coli DNA alone	32	—	3140	—
E. coli DNA + RNase	22	—	3070	—
E. coli DNA + Mg^{++} + DNase	1950	—	1130	—

For enzyme incubations, 1 ml of acid-precipitable $^3H/^{14}C$ radioactivity from the Sepharose 6B front peak (Fig. 1*A*), or 5 μl of ^3H-labeled *E. coli* DNA (7.5 μg/ml) in 1 ml of Tris–NaCl–EDTA buffer, were incubated at 37° for 2 hr alone, with 5 μg of RNase, or with 5 μg of DNase in 5 mM $MgCl_2$. For alkaline hydrolysis, 1 ml of $^3H/^{14}C$ plus 0.2 ml of 5 N NaOH were incubated at 22° for 18 hr, then the solution was neutralized with 5 N HCl. At the end of the incubation period, 0.1 ml of bovine serum albumin, 50 mg/ml, and 0.9 ml of 12% Cl_3CCOOH were added. Acid precipitate and supernatant were prepared and counted as in Table 3.

Gel filtration chromatography on Sepharose 2B was used to estimate the molecular weight of DNA found in the culture medium (Fig. 1*B*). The column was calibrated with ^{14}C-labeled HeLa cell rRNA; elution volume was determined with ^3H-labeled *E. coli* DNA. DNA from lymphocyte-culture media eluted in a broad heterogeneous peak with a molecular weight ranging between 3 and 12 × 10^6. This heterogeneity may result from partial hydrolysis of the excreted DNA by DNase activity in the culture media.§ When this excreted DNA was examined by sucrose density gradient centrifugation (Fig. 1*C*) it formed a single sharp peak at a density of 1.055 g/cm^3. A single peak formed at this density even when centrifugation conditions were varied as follows: (*A*) 18 hr in the same system as in Fig. 1*C* except at 20° rather than 4°; (*B*) 3.5 hr in the SW50.1 rotor, 49,000 rpm, at 4° on a 5–20% linear gradient of sucrose in 0.1M NaCl–0.01 M Tris (pH 7.4)–0.001 M EDTA–0.1% sodium dodecyl sulfate; (*C*) 3.5 hr in the same system as in (*B*), except on a linear 5–10% sucrose gradient. We therefore conclude that this position represents density equilibrium. This type of behavior by such a large particle would be consistent only if DNA were complexed to protein or lipid. Further characterization of this particle is in progress.

§ Evidence that the medium contains DNase activity was provided by demonstrating that when *E. coli* DNA was added to the cultures it was rapidly degraded to an acid-soluble form. The reason why the excreted DNA was not degraded to an acid-soluble form in the medium may be that it is complexed to protein or lipid.

These data can most logically be interpreted as follows: Some lymphocytes in a culture stimulated with PHA synthesize DNA and undergo mitosis; if up to 3% of the cells divide every 4 hr during a 60-hr period, the number may reach 30–40% of the total cells present. However, since some cells may undergo more than one division during this time (16), the percentage of the original cells actually undergoing mitosis is probably less than 40%. Some cells die, although not necessarily those synthesizing DNA. Since these two variables are not in balance one would expect measurement of cell number and DNA content to give inconstant results, as has been true in the literature. Up to 70% of the cells become blast-like in appearance and synthesize DNA, but these cells excrete the DNA without necessarily undergoing mitosis. The quantity of DNA they make must be small in relation to total DNA in the culture.

While the explanation for the excretion of DNA by these cells is obscure, a number of possibilities can be considered: These could be cells prevented from successfully completing the steps necessary for mitosis by some deficiency in the culture medium. Thus, frustrated in their mitotic attempts, they might excrete the unnecessary DNA and return to a resting state. Alternatively, these cells could have a purpose for DNA synthesis other than mitosis (17). The DNA might be used within the cell to amplify a yet undefined function and then excreted when no longer useful. The observation that lymphoblasts in PHA-stimulated cultures revert to small, resting lymphocytes after 6–9 days (6) is consistent with such an explanation. Gene amplification has been demonstrated to occur during oocyte development in amphibia (18), and it may occur in antibody-synthesizing plasma cells (19). Recently, Lerner *et al.* have found that about 0.5% of the DNA in a long-term lymphocyte cell line is associated with the cytoplasmic membrane (20). Whether this species of DNA is related to the DNA excreted by the PHA-stimulated lymphocytes is unknown. Finally, it is possible that the DNA is excreted as a message directed to an unknown recipient. Since lymphocytes that undergo spontaneous transformation excrete DNA in a manner similar to lymphocytes stimulated with PHA and antigen, it is likely that this phenomenon is an intrinsic property of the lymphocyte rather than a special function induced by phytohemagglutinin or antigens.

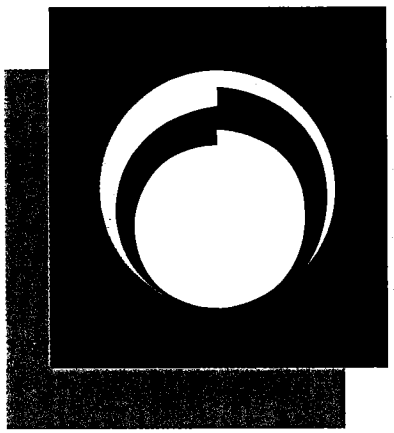
While this manuscript was in preparation, an abstract appeared by Sarma and Rutman describing findings similar to ours (21). These investigators found that PHA-stimulated lymphocytes incorporated labeled thymidine into small fragments that were then converted to high molecular weight material. Subsequently, the labeled thymidine reappeared in light fragments that were then released into the medium in an acid-precipitable form.

We thank Drs. Boyce Burge and Philip W. Robbins for allowing the use of their Beckman L2-65 centrifuge, Dr. A. Kumar for generously providing [^{14}C]rRNA, Drs. Howard Green and Burge for most helpfully discussing the manuscript, and Dr. David Botstein for his valuable advice. This work was supported in part by the U.S. Navy and in part by grants from the United States Public Health Service (HE-00022, 5 T01 AM 05089, and R01-CA-08769) and the American Cancer Society (PRA-31). The opinions or assertions contained herein are those of the authors, and are not to be construed as official or reflecting the views of the Navy Department or Naval Service at large. S. K. is the recipient of a Research Career Development Award from the National Institutes of Health.

1. Nowell, P. C. (1960) *Cancer Res.* **20**, 462–466.
2. Hirschhorn, K. & Ripps, C. S. (1965) in *Isoantigens and Cell Interactions*, ed. Palm, J. (Wistar Institute Press, Philadelphia), pp. 57–63.
3. Schellekens, P. Th. A. & Eijsvoogel, V. P. (1968). *Clin. Exp. Immunol.* **3**, 571–584.
4. Hirschhorn, R., Grossman, J., Troll, W. & Weissman, G. (1971) *J. Clin. Invest.* **50**, 1206–1217.
5. Loeb, L. A. & Agarwal, S. S. (1971) *Exp. Cell Res.* **66**, 299–304.
6. Polgar, P. R. & Kibrick, S. (1970) *Nature* **225**, 857–858.
7. Mendelsohn, J., Skinner, S. A. & Kornfeld, S. (1971) *J. Clin. Invest.* **50**, 818–826.
8. Weber, T., Nordman, C. T. & Gräsbeck, R. (1967) *Scand J. Haematol.* **4**, 77–80.
9. Mendelsohn, J., Moore, D. E. & Salzman, N. P. (1968) *J. Mol. Biol.* **32**, 101–112.
10. Kornfeld, R. & Kornfeld, S. (1970) *J. Biol. Chem.* **245**, 2536–2545.
11. Tedesco, T. A. & Mellman, W. J. (1967) *Exp. Cell Res.* **45**, 230–232.
12. Howard, I. K. & Sage, H. J. (1969) *Biochemistry* **8**, 2436–2441.
13. Gottlieb, A. A., Taylor, L. & Seinsheiner, F. (1970) *Biochemistry* **9**, 4322–4328.
14. Donahue, H. A., Moore, V. G. & Little, J. R. (1970) *J. Immunol.* **105**, 592–603.
15. Kumar, A. & Warner, J. R. (1972) *J. Mol. Biol.* **63**, 233–246.
16. Marshall, W. H., Valentine, F. T. & Lawrence, H. S. (1969) *J. Exp. Med.* **130**, 327–343.
17. Pelc, S. R. (1968) *Nature* **219**, 162–163.
18. Brown, D. D. & Blackler, A. W. (1972) *J. Mol. Biol.* **63**, 75–83.
19. Little, J. R. & Donahue, H. A. (1970) *Proc. Nat. Acad. Sci. USA* **67**, 1299–1306.
20. Lerner, R. A., Meinke, W. & Goldstein, D. A. (1971) *Proc. Nat. Acad. Sci. USA* **68**, 1212–1216.
21. Sarma, D. S. R. & Rutman, J. (1972) *Fed. Proc.* **31**, 607.

Robbins
PATHOLOGIC BASIS OF DISEASE

5th Edition



Ramzi S. Cotran, M.D.

Frank Burr Mallory Professor of Pathology
Harvard Medical School
Chairman, Departments of Pathology
Brigham and Women's Hospital
The Children's Hospital
Boston, Massachusetts

Vinay Kumar, M.D.

Vernie A. Stembridge Chair in Pathology
Southwestern Medical School
The University of Texas
Southwestern Medical School
Dallas, Texas

Stanley L. Robbins, M.D.

Visiting Professor of Pathology
Harvard Medical School
Senior Pathologist
Brigham and Women's Hospital
Boston, Massachusetts

Managing Editor

Frederick J. Schoen, M.D., Ph.D.

Associate Professor of Pathology
Harvard Medical School
Vice-Chairman, Department of Pathology
Brigham and Women's Hospital
Boston, Massachusetts

W.B. SAUNDERS COMPANY
A Division of Harcourt Brace & Company
Philadelphia London Toronto Montreal Sydney Tokyo

Exhibit AB

W.B. SAUNDERS COMPANY
A Division of
Harcourt Brace & Company

The Curtis Center
Independence Square West
Philadelphia, Pennsylvania 19106

Library of Congress Cataloging-in-Publication Data

Cotran, Ramzi S.
Robbins pathologic basis of disease. —5th ed. / Ramzi S. Cotran, Stanley L. Robbins, Vinay Kumar.
p. cm.
Includes bibliographical references and index.
ISBN 0-7216-5032-5
1. Pathology. I. Robbins, Stanley L. (Stanley Leonard).
II. Kumar, Vinay. III. Title. IV. Title: Pathologic basis of disease.
[DNLM: 1. Pathology. QZ 4 C845r 1994]
RB111.R62 1994
616.07—dc20
DNLM/DLC 94-2629

Robbins Pathologic Basis of Disease, 5th edition ISBN 0-7216-5032-5

Copyright © 1994, 1989, 1984, 1979, 1974 by W.B. Saunders Company.

All rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopy, recording, or any information storage and retrieval system, without permission in writing from the publisher.

Printed in the United States of America.

Last digit is the print number: 9 8 7 6 5 4 3 2 1

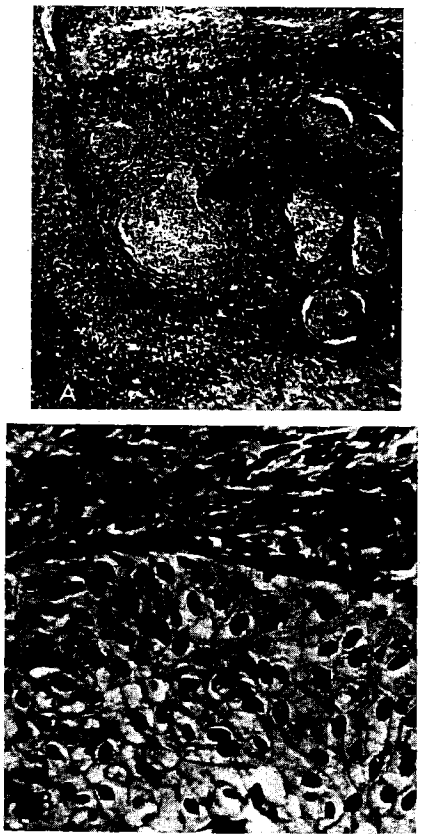


Figure 23–47. Brenner tumor. *A*, Low-power micrograph showing spindle cell component and nests of epithelial cells *(arrow)*. *B*, High-power detail of characteristic epithelial nests.

cinomas extend through the capsule of the tumor to seed the peritoneal cavity, massive ascites is common. Characteristically, the ascitic fluid is filled with diagnostic exfoliated tumor cells. The peritoneal seeding that these malignancies produce is quite distinctive: They tend to seed all serosal surfaces diffusely with 0.1- to 0.5-cm nodules of tumor. These surface implants rarely invade deeply into the underlying parenchyma of the organ. The regional nodes are often involved, and metastases may be found in the liver, lungs, gastrointestinal tract, and elsewhere. Metastasis across the midline to the opposite ovary is discovered in about half the cases by the time of laparotomy and heralds a progressive downhill course to death within a few months or years.

Because ovarian carcinomas often remain undiagnosed until very large, many patients are first seen with lesions that are no longer confined to the ovary. This is perhaps the primary reason for the relatively poor 5- and 10-year survival rates for these patients, compared with rates in cervical and endometrial carcinoma. For these reasons, specific biochemical markers for tumor antigens or tumor products in the plasma of these patients are being sought vigorously. One such marker is a high-molecular-weight glycoprotein present in more than 80% of serous and endometrioid carcinomas, known as CA-25, but whether its use will influence outcome is unproven.

Germ Cell Tumors

Germ cell tumors constitute 15 to 20% of all ovarian tumors.[74] The vast majority are benign cystic teratomas, but the remainder, which are found principally in children and young adults, have a higher incidence of malignant behavior and pose problems in histologic diagnosis and in therapy. They bear a remarkable homology to germ tumors in the male testis (see Chapter 22) and arise from germ cell differentiation in a similar manner (Fig. 23–48).

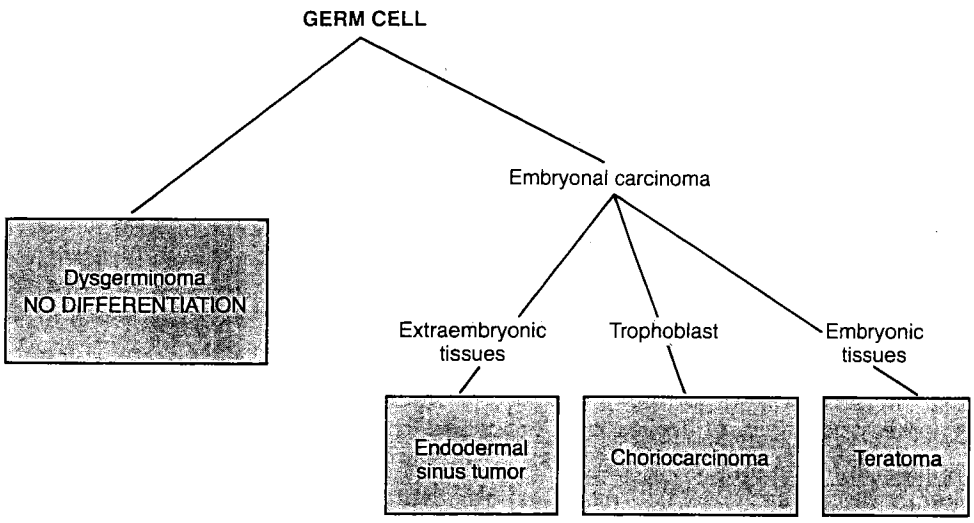


Figure 23–48. Histogenesis and interrelationships of tumors of germ cell origin.

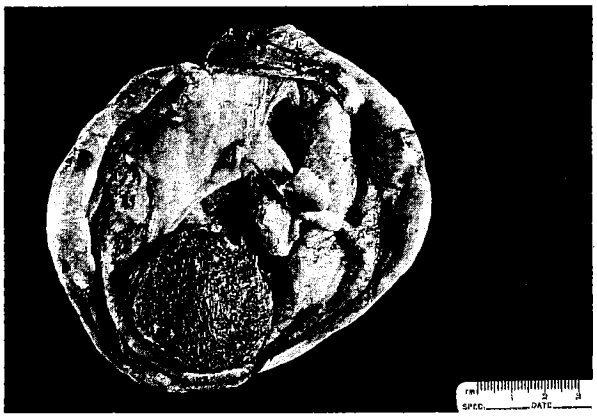


Figure 23–49. Opened mature cystic teratoma (dermoid cyst) of the ovary. Hair *(bottom)* and a mixture of tissues are evident.

Figure 23–50. Dermoid cyst. Low-power view of skin, underlying sebaceous glands, and cartilage *(lower left)*.

Teratomas

Teratomas are divided into three categories: (1) mature (benign), (2) immature (malignant), and (3) monodermal or highly specialized.

MATURE (BENIGN) TERATOMAS. The vast majority of benign teratomas are cystic and are better known in clinical parlance as dermoid cysts. These neoplasms are invariably benign and are presumably derived from the ectodermal differentiation of totipotential cells. Cystic teratomas are usually found in young women during the active reproductive years.[74]

MORPHOLOGY. Benign teratomas are bilateral in 10 to 15% of cases. Characteristically, they are unilocular cysts containing hair and cheesy sebaceous material (Fig. 23–49). On section they reveal a thin wall lined by an opaque, gray-white, wrinkled, apparent epidermis. From this epidermis, hair shafts frequently protrude. Within the wall, it is common to find tooth structures and areas of calcification.

Histologically, the cyst wall is composed of stratified squamous epithelium with underlying sebaceous glands, hair shafts, and other skin adnexal structures (Fig. 23–50). In most cases, structures from other germ layers can be identified, such as cartilage, bone, thyroid tissue, and other organoid formations (Fig. 23–51). Dermoid cysts are sometimes incorporated within the wall of a mucinous cystadenoma. **About 1% of the dermoids undergo malignant transformation of any one of the component elements—e.g., thyroid carcinoma, melanoma, but most commonly squamous cell carcinoma.**

In rare instances, a teratoma is solid but is composed entirely of benign-looking heterogeneous collections of tissues and organized structures derived from all three germ layers. These tumors presumably have the same histogenetic origin as dermoid cysts but lack preponderant differentiation into ectodermal derivatives. **These neoplasms may be initially difficult to differentiate from the malignant, immature teratomas, which almost always are largely solid.**

Figure 23–51. Mature cystic teratoma of the ovary, opened to illustrate several abortive tooth structures *(arrows)* and a darker area of thyroid substance (T).

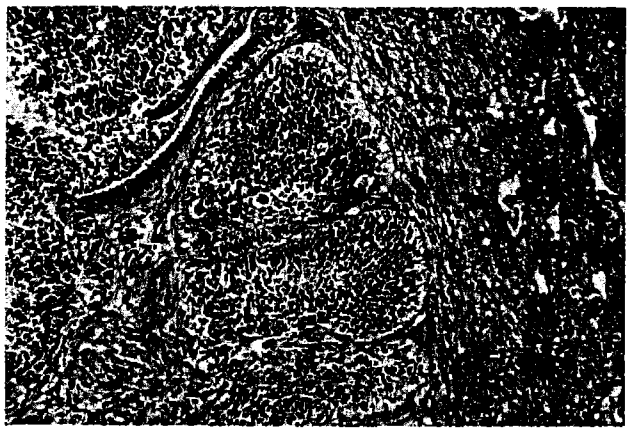


Figure 23–52. Immature teratoma of the ovary, illustrating primitive neuroepithelium.

The origin of teratomas has been a matter of fascination for centuries. Some common beliefs blamed witches, nightmares, or adultery with the devil. The current parthenogenetic theory suggests origin from a meiotic germ cell. The karyotype of all benign ovarian teratomas is 46,XX. From the results of chromosome-banding techniques and the distribution of electrophoretic variants of enzymes in the normal and teratoma cells, Linder and coworkers[86] suggested that tumors arise from an ovum after the first meiotic division. Other derivations have been proposed.[87]

IMMATURE MALIGNANT TERATOMAS. These are rare tumors that differ from benign teratomas in that the component tissue resembles that observed in the fetus or embryo rather than adult. The tumor is found chiefly in prepubertal adolescents and young women, the mean age being 18.[88]

MORPHOLOGY. The tumors are bulky and have a smooth external surface. On section they have a solid or predominantly solid structure. There are areas of necrosis and hemorrhage. Hair, grumous material, cartilage, bone, and calcification may be present. Microscopically, there are varying amounts of immature tissue differentiating toward cartilage, glands, bone, muscle, nerve, and others. An important risk for subsequent extraovarian spread is the histologic grade of tumor (I through III), which is based on the proportion of tissue (in histologic sections) containing immature neuroepithelium (Fig. 23–52).

Immature teratomas grow rapidly and frequently penetrate the capsule with spread or metastases. Stage I tumors, however, particularly those with low-grade (grade I) histology, have an excellent prognosis. Higher grade tumors confined to the ovary are generally treated with prophylactic chemotherapy. Most recurrences develop in the first 2 years, and absence of disease beyond this period carries an excellent chance of cure.

MONODERMAL OR SPECIALIZED TERATOMAS. The specialized teratomas are a remarkable, rare group of tumors, the most common of which are struma ovarii and carcinoid. They are always unilateral, although a contralateral teratoma may be present. Struma ovarii is composed entirely of mature thyroid tissue. Interestingly these thyroidal neoplasms may hyperfunction, producing hyperthyroidism. The ovarian carcinoid, which presumably rises from intestinal epithelium in a teratoma, might in fact be functioning, particularly in large (greater than 7 cm) tumors, producing 5-hydroxytryptamine and the carcinoid syndrome. Primary ovarian carcinoid can be distinguished from metastatic intestinal carcinoid, the latter virtually always bilateral. Even more rare is the strumal carcinoid, a combination of struma ovarii and carcinoid in the same ovary. Primary carcinoids are uncommonly (less than 2%) malignant.

Dysgerminoma

The dysgerminoma[89] is best remembered as the ovarian counterpart of the seminoma of the testis. Similar to the latter, it is composed of large vesicular cells having a cleared cytoplasm, well-defined cell boundaries, and centrally placed regular nuclei. Relatively uncommon tumors, the dysgerminomas account for about 2% of all ovarian cancers yet form about half of malignant germ cell tumors. They may occur in childhood, but 75% occur in the second and third decades. Some occur in patients with gonadal dysgenesis, including pseudohermaphroditism. Most of these tumors have no endocrine function. A few produce elevated levels of chorionic gonadotropin and may have syncytiotrophoblastic giant cells on histologic examination.

MORPHOLOGY. Usually unilateral (80 to 90%), they most frequently are solid tumors ranging in size from barely visible nodules to masses that virtually fill the entire abdomen. On cut surface, they have a yellow-white to gray-pink appearance and are often soft and fleshy. Histologically, the dysgerminoma cells are dispersed in sheets or cords separated by scant fibrous stroma. As in the seminoma, the fibrous stroma is infiltrated with mature lymphocytes and occasional granulomas. Occasionally, small nodules of dysgerminoma are encountered in the wall of an otherwise benign cystic teratoma; conversely, a predominantly dysgerminomatous tumor may contain a small cystic teratoma.

All dysgerminomas are malignant, but the degree of histologic atypia is variable, and only about one-third are aggressive. Thus a unilateral tumor

that has not broken through the capsule and has not spread has an excellent prognosis (up to 96% cure rate) after simple salpingo-oophorectomy. These neoplasms are extremely radiosensitive, and even those that have extended beyond the ovary can generally be controlled by radiotherapy. Overall survival exceeds 80%.

Endodermal Sinus (Yolk Sac) Tumor

This tumor is rare but is the second most common malignant tumor of germ cell origin. It is thought to be derived from a multipotential embryonal carcinoma by selection and differentiation toward yolk sac structure[82] (see Fig. 23–48). Similar to the yolk sac, the tumor is rich in alpha-fetoprotein and alpha-antitrypsin. Its characteristic histologic feature is a glomerulus-like structure composed of a central blood vessel enveloped by germ cells within a space similarly lined by germ cells (Schiller-Duval body). Similar structures are observed in the yolk sac of the rat placenta. Conspicuous intracellular and extracellular hyaline droplets are present in all tumors, and some of these can be stained for alpha-fetoprotein by immunoperoxidase techniques.

Most patients are children or young women presenting with abdominal pain and a rapidly developing pelvic mass. The tumors usually appear to involve a single ovary but grow rapidly and aggressively. These tumors were once almost uniformly fatal within 2 years of diagnosis, but combination chemotherapy has measurably improved the outcome.

Choriocarcinoma

More commonly of placental origin, the choriocarcinoma, similar to the endodermal sinus tumor, is an example of extraembryonic differentiation of malignant germ cells. It is generally held that a germ cell origin can be certified only in the prepubertal girl because after this age, an origin from an ovarian ectopic pregnancy cannot be excluded.

Most ovarian choriocarcinomas exist in combination with other germ cell tumors, and pure choriocarcinomas are extremely rare. Histologically they are identical with the more common placental lesions, described later. These ovarian primaries are ugly tumors that generally have metastasized widely through the bloodstream to the lungs, liver, bone, and other viscera by the time of diagnosis. As with all choriocarcinomas, they elaborate high levels of chorionic gonadotropins that are sometimes helpful in establishing the diagnosis or highlighting recurrences. In contrast to choriocarcinomas arising in placental tissue, those arising in the ovary are generally unresponsive to chemotherapy and are highly fatal.

Other Germ Cell Tumors

These include (1) embryonal carcinoma, another highly malignant tumor of primitive embryonal elements, histologically similar to tumors arising in the testes (see Chapter 22); (2) polyembryoma, a malignant tumor containing so-called embryoid bodies; and (3) mixed germ cell tumors containing various combinations of dysgerminoma, teratoma, endodermal sinus tumor, and choriocarcinoma.

Sex Cord–Stromal Tumors

These ovarian neoplasms are derived from the ovarian stroma, which in turn is derived from the sex cords of the embryonic gonad. Because the undifferentiated gonadal mesenchyme eventually produces structures of specific cell type in both male (Sertoli and Leydig) and female (granulosa and theca) gonads, tumors emulating all of these cell types can be identified in the ovary.[90] Moreover, because some of these cells normally secrete estrogens (theca cells) or androgens (Leydig cells), their corresponding tumors may be either feminizing (granulosa–theca cell tumors) or masculinizing (Leydig cell tumors).

Granulosa-Theca Cell Tumors

This designation embraces ovarian neoplasms composed of varying proportions of granulosa and theca cell differentiation. These tumors are composed almost entirely of granulosa cells or a mixture of granulosa and theca cells. Collectively these neoplasms account for about 5% of all ovarian tumors. Although they may be discovered at any age, approximately two-thirds occur in postmenopausal women.

MORPHOLOGY. Granulosa cell tumors are usually unilateral and vary from microscopic foci to large, solid, and cystic encapsulated masses. Tumors that are endocrinologically active have a yellow coloration to their cut surfaces, produced by contained lipids. The pure thecomas are solid, firm tumors.

The granulosa cell component of these tumors takes one of many histologic patterns. The small, cuboidal-to-polygonal cells may grow in anastomosing cords, sheets, or strands (Fig. 23–53). In occasional cases, small, distinctive, gland-like structures filled with an acidophilic material recall immature follicles (Call-Exner bodies). When these structures are evident, the diagnosis is rendered considerably more simple. The thecoma component consists of clusters or sheets of cuboidal-to-polygonal cells. In some tumors, the granulosa or theca cells may appear more plump with ample cytoplasm characteristic of luteinization (i.e., luteinized granulosa–theca cell tumors).

Granulosa–theca cell tumors have clinical importance for two reasons: (1) their potential elaboration of large amounts of estrogen and (2) the small but distinct hazard of malignancy in the

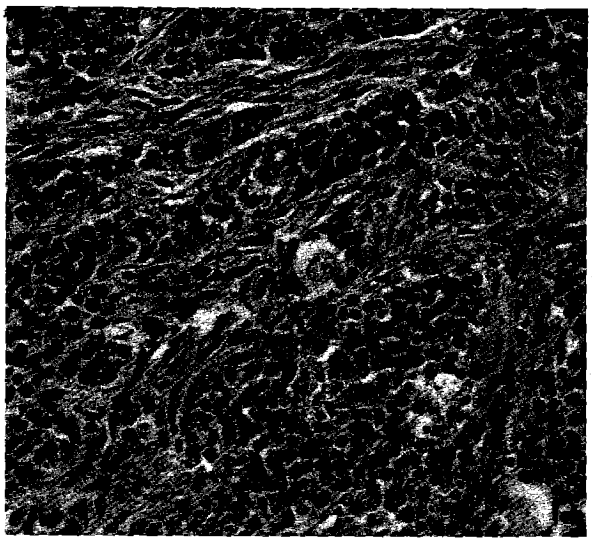


Figure 23–53. Granulosa cell tumor.

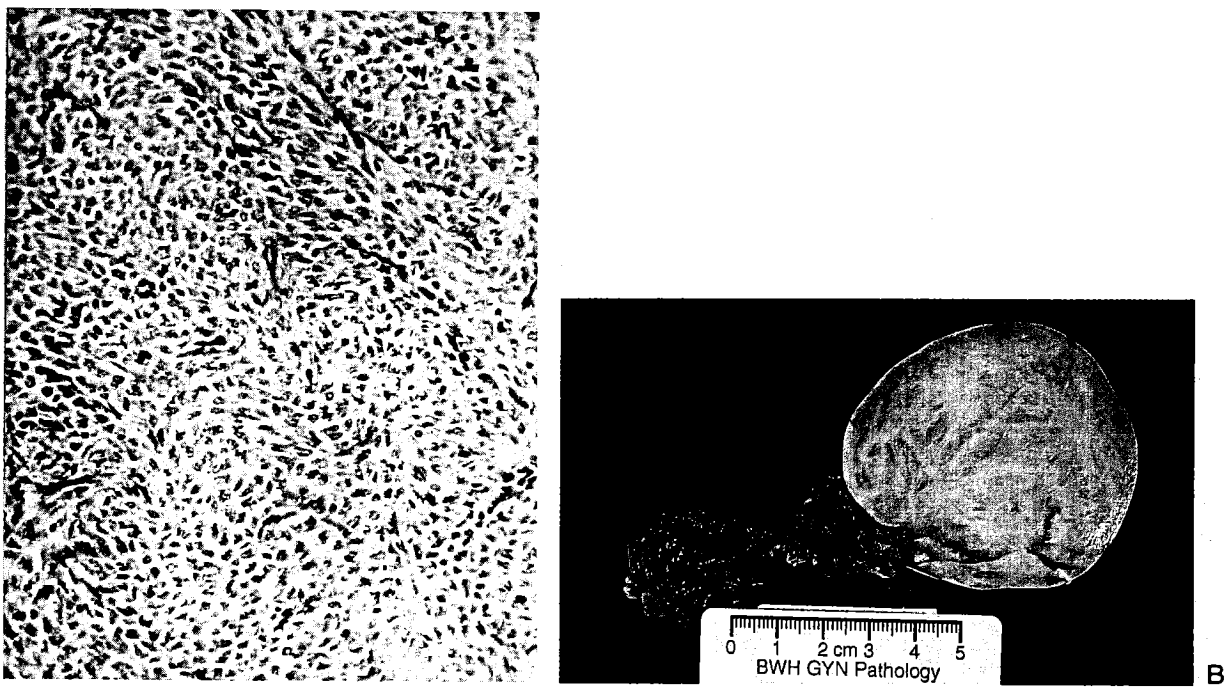
granulosa cell forms. Functionally active tumors in young girls (juvenile granulosa cell tumors) may produce precocious sexual development in prepubertal girls. In adult women, they may be associated with endometrial hyperplasia, cystic disease of the breast, and endometrial carcinoma. About 10 to 15% of patients with steroid-producing tumors eventually develop an endometrial carcinoma. Occasional granulosa cell tumors produce androgens, masculinizing the patient.

The additional clinical significance of these tumors lies in the fact that all are potentially malignant. It is difficult, from the histologic evaluation of granulosa cell tumors, to predict their biologic behavior.[74] The estimates of clinical malignancy (recurrence, extension) range from 5 to 25%. In general, malignant tumors pursue an indolent course in which local recurrences may be amenable to surgical therapy. Recurrences within the pelvis and abdomen may appear many years (10 to 20) after removal of the original tumor. The 10-year survival rate is approximately 85%. Tumors composed predominantly of theca cells are almost never malignant.

Thecoma-Fibromas

Tumors arising in the ovarian stroma that are composed of either fibroblasts (fibromas) or more plump spindle cells with lipid droplets (thecomas) are relatively common and account for about 4% of all ovarian tumor types (Fig. 23–54*A*). Because many tumors contain a mixture of these cells, they are termed fibroma-thecomas. Pure thecomas are very rare, but tumors in which these cells predominate may be hormonally active. Most are composed principally of fibroblasts and are for practical purposes hormonally inactive.

Fibroma-thecomas of the ovary are unilateral in about 90% of cases and usually are solid, spherical or slightly lobulated, encapsulated, hard, gray-white masses covered by glistening, intact ovarian serosa (Fig. 23–54*B*). Histologically, they are com-



A B

Figure 23–54. *A*, Thecoma-fibroma composed of plump, differentiated stromal cells with thecal appearance. (Courtesy of Dr. A. Hertig and the Armed Forces Institute of Pathology.) *B*, Large bisected fibroma of ovary apparent as a white, firm mass *(right)*. The fallopian tube is attached.

posed of well-differentiated fibroblasts with a more or less scant collagenous connective tissue interspersed between the cells. Areas of thecal differentiation may be identified and can be confirmed by fat stains. This exercise, however, is considered unnecessary on clinical grounds.

In addition to the relatively nonspecific findings of pain and pelvic mass, the tumors may be accompanied by two curious associations. The first is ascites, found in about 40% of cases, in which the tumors measure more than 6 cm in diameter. Uncommonly, there is also hydrothorax, usually only of the right side. This combination of findings, i.e., ovarian tumor, hydrothorax, and ascites, is designated Meigs' syndrome. The genesis is unknown. The second association is with the basal cell nevus syndrome, described in Chapter 26. Rarely, cellular tumors with mitotic activity and increased nuclear cytoplasmic ratio are identified; because they may pursue a malignant course, they are termed fibrosarcomas.[91]

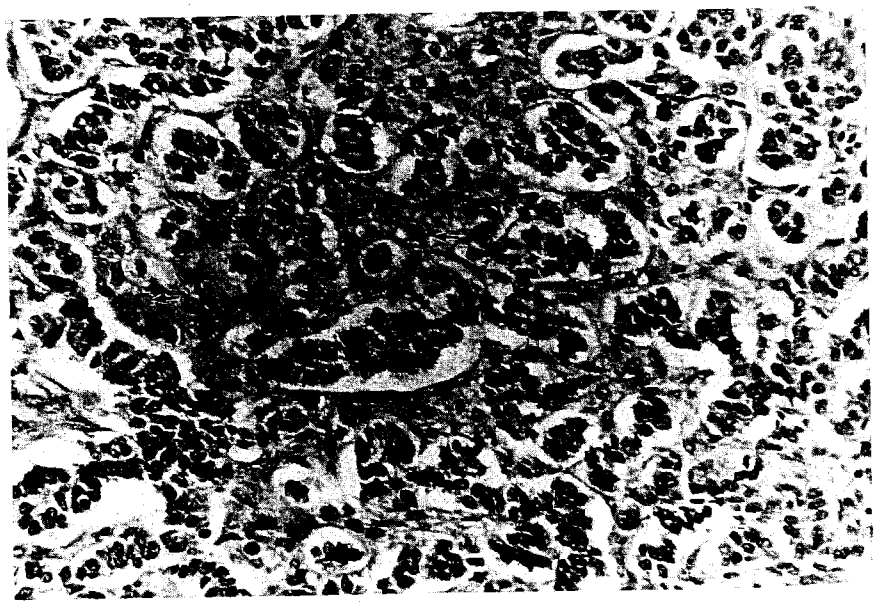
Sertoli-Leydig Cell Tumors (Androblastoma)

These tumors recapitulate, to a certain extent, the cells of the testis at various stages of development.[92] They commonly produce masculinization or at least defeminization, but a few have estrogenic effects. They occur in women of all ages, although the peak incidence is in the second and third decades. The embryogenesis of such male-directed stromal cells remains a puzzle. These tumors are unilateral and resemble granulosa–theca cell neoplasms.

The cut surface is usually solid and varies from gray to golden brown in appearance (Fig. 23–55*A*). Histologically the well-differentiated tumors exhibit tubules composed of Sertoli cells or Leydig cells interspersed with stroma (see Fig. 23–55*B*). The intermediate forms show only outlines of immature tubules and large eosinophilic Leydig cells. The poorly differentiated tumors have a sarcomatous pattern with a disorderly disposition of epithelial cell cords. Leydig cells may be absent. Heterologous elements, such as mucinous glands, bone, and cartilage, may be present in some tumors.

The incidence of recurrence or metastasis by Sertoli-Leydig cell tumors is less than 5%. These neoplasms may block normal female sexual development in children and may cause defeminization

A



B

Figure 23–55. Sertoli cell tumor. *A*, Gross photograph illustrating characteristic golden-yellow appearance of tumor; *B*, photomicrograph showing well-differentiated Sertoli cell tubules. A few larger cells (Leydig cells) are noted between the tubular structures *(arrows)*. (Courtesy of Dr. William Welch, Brigham and Women's Hospital, Boston.)

of women, manifested by atrophy of the breasts, amenorrhea, sterility, and loss of hair. The syndrome may progress to striking virilization, that is, hirsutism, male distribution of hair, hypertrophy of the clitoris, and voice changes.

Other Sex Cord–Stromal Tumors

Normally, the ovarian hilus contains clusters of polygonal cells arranged around vessels (hilar cells). *Hilus cell tumors (pure Leydig cell tumor)* are derived from these cells, are rare, are unilateral, and are characterized histologically by large, lipid-laden cells with distinct borders. A typical cytoplasmic structure characteristic of Leydig cells—Reinke crystalloids—is usually present. Typically, the patients present clinically with evidence of masculinization, hirsutism, voice changes, and clitoral enlargement. The tumors are unilateral. The most consistent laboratory finding is an elevated 17-ketosteroid excretion level unresponsive to cortisone suppression. Treatment is surgical excision. True hilus cell tumors are almost always benign. Occasionally histologically identical tumors occur in the cortical stroma *(nonhilar Leydig cell tumors)*.

In addition to Leydig cell tumors, the stroma rarely may give rise to tumors composed of pure luteinized cells, producing small benign tumors rarely more than 3 cm in diameter. The tumor may produce the clinical effects of androgen, estrogen, or progestogen stimulation.

As mentioned previously, the ovary in pregnancy may exhibit microscopic nodular proliferations of theca cells in response to gonadotropins. Rarely a frank tumor may develop—termed *pregnancy luteoma*—which closely resembles a corpus luteum of pregnancy. These tumors have been associated with virilization in the pregnant patient and in their respective female infants.

Gonadoblastoma is an uncommon tumor thought to be composed of germ cells and sex cord–stroma derivatives. It occurs in individuals with abnormal sexual development and in gonads of indeterminate nature. Eighty per cent are phenotypic females, and 20% are phenotypic males with undescended testicles and female internal secondary organs. Microscopically, the tumor consists of nests of a mixture of germ cells and sex cord derivatives resembling immature Sertoli and granulosa cells. A coexistent dysgerminoma occurs in 50% of the cases. The prognosis is excellent if the tumor is completely excised.[93]

Metastatic Tumors

The most common "metastatic" tumors of the ovary are probably derived from tumors of müllerian origin: the uterus, fallopian tube, contralateral ovary, or pelvic peritoneum. Although technically metastases, they are accepted as part of the "field effect" of müllerian neoplasia. The most common extramüllerian primaries are the breast and gastrointestinal tract, including stomach, biliary tract, and pancreas. A classic example of metastatic gastrointestinal neoplasia to the ovaries is termed Krukenberg's tumor, which are bilateral metastases composed of mucin-producing, signet-ring cancer cells, most often of gastric origin (Fig. 23–56).[94]

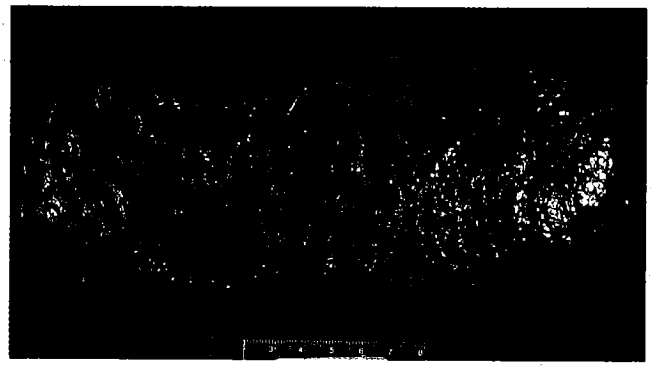


Figure 23–56. Krukenberg's tumor (metastatic gastric carcinoma) of the ovary.

GESTATIONAL AND PLACENTAL DISORDERS

Diseases of pregnancy and pathologic conditions of the placenta are important causes of intrauterine or perinatal death, congenital malformations, intrauterine growth retardation, maternal death, and a great deal of morbidity for both mother and child.[6,95] Here we discuss only a limited number of disorders in which knowledge of the morphologic lesions contributes to an understanding of the clinical problem.

PLACENTAL INFLAMMATIONS AND INFECTIONS

Infections may occur in the placenta (placentitis, villitis),[96] in the fetal membranes (chorioamnionitis), and in the umbilical cord (funisitis).[97] They reach the placenta by two pathways: (1) ascending infection through the birth canal and (2) hematogenous (transplacental) infection. Ascending infections are by far the most common and are most often bacterial; in many such instances, localized infection of the membranes by an organism produces premature rupture of membranes and entry for the organisms. Sexual intercourse has been implicated in enhancing ascending infections.[95] The amniotic fluid may be cloudy with purulent exudate, and the chorion-amnion histologically contains a leukocytic polymorphonuclear infiltration

with accompanying edema and congestion of the vessels. The infection frequently elicits a fetal response with umbilical cord vasculitis. Extension into the villous space may produce a septic intervillous thrombus.

Uncommonly, bacterial infections of the placenta and fetal membranes may arise by the hematogenous spread of bacteria directly to the placenta. Histologically, the villi are most often affected (villitis). Classically, TORCH (toxoplasma and others [syphilis, tuberculosis, listeriosis], rubella, cytomegalovirus, herpes simplex) should be considered, although the cause is usually obscure and may involve immunologic phenomena.[7] (See also Chapter 10.)

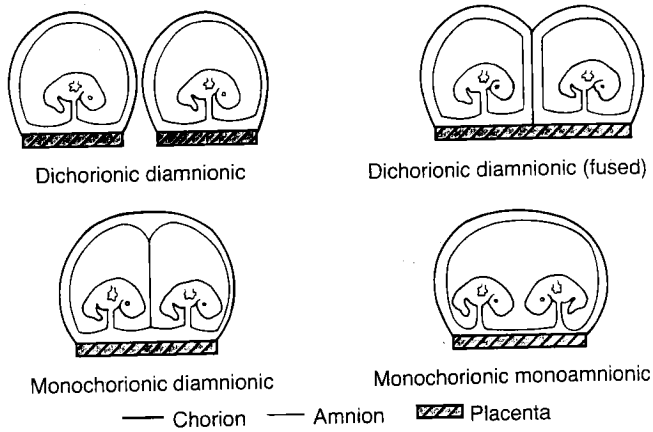


Figure 23–57. Diagrammatic representation of the various types of twin placentation and their membrane relationships. (Adapted with permission from Gersell, D., et al.: *In* Kurman, R. (ed): Blaustein's Pathology of the Female Genital Tract. New York, Springer-Verlag, 1987.)

PLACENTAL ABNORMALITIES AND TWIN PLACENTAS

Abnormalities in placental shape, structure, and implantation are not uncommon. Accessory placental lobes, bipartite placenta (placenta made up of two equal segments), and circumvallate placenta (having an extrachorial part) are examples of abnormalities that have limited clinical significance.

Placenta accreta[97] is caused by partial or complete absence of the decidua with adherence of the placenta directly to the myometrium. It is important for two reasons: (1) Postpartum bleeding, often life-threatening, occurs because of failure of placental separation; (2) in up to 60% of cases, it is associated with placenta previa, a condition in which the placenta implants in the lower uterine segment or cervix, often with serious antepartum bleeding and premature labor. Many cases of placenta previa–associated accreta occur in cesarean section scars.

Twins arise from fertilization of two ova (dizygotic) or from division of one fertilized ovum (monozygotic). There are three basic types of twin placentas (Fig. 23–57):[98] dichorionic diamnionic (which may be fused); monochorionic diamnionic; and monochorionic monoamnionic. Monochorionic placentas imply monozygotic (identical) twins, and the time at which splitting occurs determines whether one or two amnions are present. Dichorionic gestation may occur with either monozygotic or dizygotic twins and is not specific.

SPONTANEOUS ABORTION

Ten to 15% of recognized pregnancies terminate in spontaneous abortion. Studies, however, using highly sensitive immunoassay of chorionic gonadotropin to detect pregnancy identified an additional 22% loss of presumably fertilized and implanted ova in otherwise healthy women.[99] The causes for this early loss of pregnancy are still mysterious, but as one good writer said, "There's many a slip 'twixt implantation and the crib."[100]

The causes of recognized spontaneous abortion are both fetal and maternal. Defective implantation inadequate to support fetal development and death of the ovum or fetus in uteri because of some genetic or acquired abnormality constitute the major origins of spontaneous abortion. Numerous studies have indicated chromosomal abnormalities in more than one-half of spontaneous abortuses.[101]

Maternal influences, which are less well understood, include inflammatory diseases, both localized to the placenta and systemic; uterine abnormalities; and possibly trauma. The role of trauma is generally overemphasized and must be considered a rare-to-exceptional trigger of spontaneous abortion. Toxoplasma, mycoplasma, listeria, and viral infections have also been implicated as causes of abortion.

The morphologic changes usually seen in endometrial curettage specimens depend, of course, on the interval between fetal death and passage of the products of conception.[102] Generally there are focal areas of decidual necrosis with intense neutrophilic infiltrations, thrombi within decidual blood vessels, and considerable amounts of hemorrhage, both recent and old, within the necrotic decidua. Placental villi may be markedly edematous and devoid of blood vessels. The changes encountered in the ovum or fetus are highly variable. In many spontaneous abortions, no fetal products can be identified, but when they are present, they should be carefully examined for anomalies that would suggest specific genetic or karyotypic defects. Chromosomal studies are recommended in

(1) habitual or recurrent abortion and (2) when there is a malformed fetus.

ECTOPIC PREGNANCY

Ectopic pregnancy is the term applied to implantation of the fetus in any site other than normal uterine location. The most common site is within the tubes (approximately 90%).[103] The other sites are the ovary, the abdominal cavity, and the intrauterine portion of the fallopian tube (cornual pregnancy). Ectopic pregnancies occur about once in every 150 pregnancies. The most important predisposing condition in 35 to 50% of patients is PID with chronic salpingitis. Other factors are peritubal adhesions owing to appendicitis or endometriosis, leiomyomas, and previous surgery. Fifty per cent, however, occur in tubes that are apparently normal. Intrauterine devices (IUDs) may also increase risk.

Ovarian pregnancy is presumed to result from the rare fertilization and trapping of the ovum within the follicle just at the time of its rupture. Abdominal pregnancies may develop when the fertilized ovum drops out of the fimbriated end of the tube. In all these abnormal locations, the fertilized ovum undergoes its usual development with the formation of placental tissue, amniotic sac, and fetus, and the host implantation site develops decidual changes.

MORPHOLOGY. In tubal pregnancy, the placenta is poorly attached to the wall of the tube. Intratubal hemorrhage may thus occur from partial placental separation without tubal rupture (Fig. 23–58). Tubal pregnancy is the most common cause of hematosalpinx and should always be suspected when a tubal hematoma is present. More often, the placental tissue invades the tubal wall and causes tubal rupture and intraperitoneal hemorrhage. Less commonly, the tubal pregnancy may undergo spontaneous regression and resorption of the entire gestation. Still less commonly, the tubal pregnancy is extruded through the fimbriated end into the abdominal cavity (tubal abortion).

The clinical course of ectopic pregnancy is punctuated by the onset of severe abdominal pain about 6 weeks following a previous normal menstrual period, when rupture of the tube leads to a pelvic hemorrhage. In such cases, the patient may rapidly develop a shock-like state with signs of an acute abdomen, and early diagnosis becomes critical. Chorionic gonadotropin assays, ultrasound studies, and laparoscopy may be helpful. Endometrial biopsy specimens may or may not disclose decidual changes but, excluding the extremely rare dual pregnancy, do not exhibit chorionic villi. It must be remembered that rupture of a tubal pregnancy constitutes a medical emergency.

Figure 23–58. Tubal pregnancy with marked dilatation and rupture of distal end of tube by the contained pregnancy and subsequent hemorrhage.

TOXEMIA OF PREGNANCY (PRE-ECLAMPSIA AND ECLAMPSIA)

Toxemia of pregnancy refers to a symptom complex characterized by hypertension, proteinuria, and edema (pre-eclampsia). It occurs in about 6% of pregnant women, usually in the last trimester and more commonly in primiparas than in multiparas. Certain of these patients become more seriously ill, developing convulsions; this more severe form is termed eclampsia. Patients with eclampsia develop disseminated intravascular coagulation (DIC) with lesions in the liver, kidneys, heart, placenta, and sometimes the brain. There is no absolute correlation between the severity of eclampsia and the magnitude of the anatomic changes.

PATHOGENESIS. The many theories on the nature of toxemia of pregnancy[104] are beyond our scope, but three events that seem to be of prime importance in this disorder are addressed: placental ischemia, hypertension, and DIC (Fig. 23–59).

The causes of the initial events of toxemia are unknown, but evidence points to an abnormality of

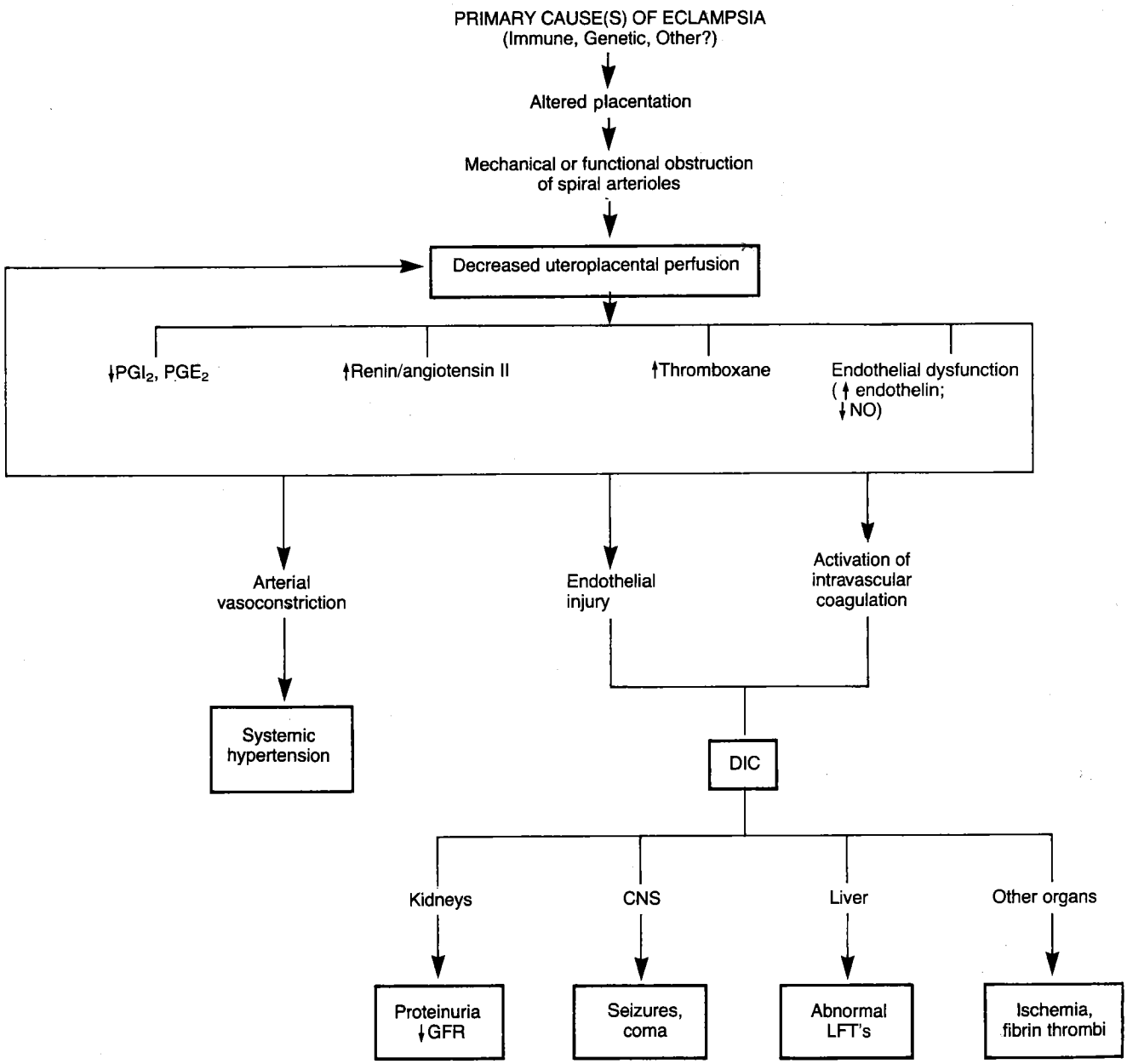


Figure 23–59. Proposed sequence of events in the pathogenesis of toxemia of pregnancy. The main features are (1) decreased uteroplacental perfusion; (2) increased vasoconstrictors and decreased vasodilators, resulting in local and systemic vasoconstriction; and (3) disseminated intravascular coagulation (DIC). (Adapted from Friedman, S.A.: Obstet. Gynecol. *71*:122, 1988 (reprinted by permission of the American College of Obstetricians and Gynecologists); and Khong, T.Y., et al.: Br. Med. J. 93:1049, 1986.)

placentation.[105] This may involve an abnormality in both trophoblast invasion and the development of the physiologic alterations in the placental vessels required to perfuse the placental bed adequately. Immunologic, genetic, and other acquired factors have been postulated as the cause of such abnormality. The net effect is a shallow implantation with incomplete conversion of decidual vessels to vessels adequate for the pregnancy state.[105] The resultant perfusion abnormality produces placental ischemia, the basis for the toxemic placenta. It is thought that this decreased uteroplacental perfusion induces the stimulation of vasoconstrictor substances (thromboxane, angiotensin, endothelin) and the inhibition of vasodilator influences (PGI_2, PGE_2, nitric oxide) from the ischemic placenta. The resultant DIC, hypertension, and organ damage ensue (see Fig. 23–59).

As to the pathogenesis of DIC in toxemia, endothelial damage, abnormalities in the level and activities of coagulation factors, and primary platelet alteration may play a role.[106] For example, during toxemia, the placental ischemia leads to a higher output of thromboplastic substances, and antithrombin III levels are reduced. The characteristic lesions in eclampsia are in large part due to

thrombosis of arterioles and capillaries throughout the body, particularly in the liver, kidneys, brain, pituitary, and placenta.

The mechanism of toxemic hypertension appears to involve the renin-angiotensin mechanism and prostaglandins.[107] Normal pregnant women develop a resistance to the vasoconstrictive and hypertensive effects of angiotensin, but women with toxemia lose such resistance, developing a tendency to hypertension. Prostaglandins of the E series, produced in the uteroplacental vascular bed during pregnancy, are thought to mediate the normal resistance of pregnant women to angiotensin, and prostaglandin production is indeed decreased in the placenta of toxemic women. Thus the increase in angiotensin hypersensitivity, characteristic of toxemia, may be due to decreased synthesis of prostaglandin by the toxemic placenta. There is also evidence that renin production by the toxemic placenta is increased, another potentially vasoconstrictive event. Figure 23–59 presents a hypothetical schema for the pathogenesis of eclampsia.

MORPHOLOGY. The liver lesions, when present, take the form of irregular, focal, subcapsular, and intraparenchymal hemorrhages. On histologic examination, there are fibrin thrombi in the portal capillaries with foci of characteristic peripheral hemorrhagic necrosis.

The **kidney** lesions are variable. Glomerular lesions are diffuse, at least when assessed by electron microscopy. They consist of striking swelling of endothelial cells, the deposition of fibrinogen-derived amorphous dense deposits on the endothelial side of the basement membrane, and mesangial cell hyperplasia. Immunofluorescent studies confirm the abundance of fibrin in glomeruli. In the more well-defined cases, fibrin thrombi are present in the glomeruli and capillaries of the cortex. When the lesion is far advanced, it may produce complete destruction of the cortex in the pattern referred to as bilateral renal cortical necrosis (see Chapter 20).

The **brain** may have gross or microscopic foci of hemorrhage along with small-vessel thromboses. Similar changes are often found in the **heart** and the **anterior pituitary.** The **placenta** is the site of variable changes, most of which reflect ischemia and vessel injury. (1) Placental infarcts, which occur in normal full-term placentas, are larger and more numerous. (2) There is increased frequency of retroplacental hematomas. (3) There is evidence of increased villous ischemia: formation of prominent syncytial knots, thickening of trophoblastic basement membrane, and villous hypovascularity. (4) A characteristic finding in the walls of uterine vessels is striking fibrinoid necrosis and intramural lipid deposition (acute atherosis) (Fig. 23–60).

Figure 23–60. Acute atherosis of uterine vessels in eclampsia. Note fibrinoid necrosis of the vessel walls, subendothelial macrophages *(arrows)*, and perivascular lymphocytic infiltrate. (Courtesy of Dr. Drucilla J. Roberts.)

CLINICAL COURSE. Pre-eclampsia usually starts after the 32nd week of pregnancy but begins earlier in patients with hydatidiform mole or pre-existing kidney disease or hypertension. The onset is usually insidious, characterized by hypertension and edema, with proteinuria following within several days. Headaches and visual disturbances are common. Eclampsia is heralded by central nervous system involvement, including convulsions and eventual coma. Mild and moderate forms of toxemia can be controlled by bed rest, a balanced diet, and antihypertensive agents, but induction of delivery is the only definitive treatment of established pre-eclampsia and eclampsia. Proteinuria and hypertension usually disappear within 1 or 2 weeks after delivery except in patients in whom these findings predated the pregnancy.

GESTATIONAL TROPHOBLASTIC DISEASE

Gestational trophoblastic disease constitutes a spectrum of tumors and tumor-like conditions characterized by proliferation of pregnancy-associated trophoblastic tissue of progressive malignant potential. The lesions include the hydatidiform mole (complete and partial), the invasive mole, and

the frankly malignant choriocarcinoma.[62,108] Gestational trophoblastic disease is important for the following reasons:

- The hydatidiform mole is a common complication of gestation, occurring about once in every 1000 to 2000 pregnancies in the United States and curiously far more commonly in the Far East.
- It has become possible, by monitoring the circulating levels of human chorionic gonadotropin, to determine the early development of persistent trophoblastic disease.
- Choriocarcinoma, once a dreaded and uniformly fatal complication, is now highly responsive to chemotherapy in most cases.[109]

Hydatidiform Mole (Complete and Partial)

Hydatidiform mole is characterized by cystic swelling of the chorionic villi, accompanied by variable trophoblastic proliferation. The most important reason for the correct recognition of true moles is that they are the most common precursors of choriocarcinoma.[108] Most patients present in the fourth or fifth month of pregnancy with vaginal bleeding and with a uterus that is usually, but not always, larger than expected for the duration of pregnancy. These moles can occur at any age during active reproductive life, but the risk is higher in pregnant women in their teens or between the ages of 40 and 50 years. For poorly explained reasons, the incidence varies considerably in different regions of the world: 1 in 1000 pregnancies in the United States but 10 in 1000 in Indonesia.[110]

TYPES AND PATHOGENESIS. Two types of benign, noninvasive moles—complete and partial—can be differentiated by histologic, cytogenetic, and flow cytometric studies (Table 23–4).[111] In complete (or classic mole), all or most of the villi are edema-

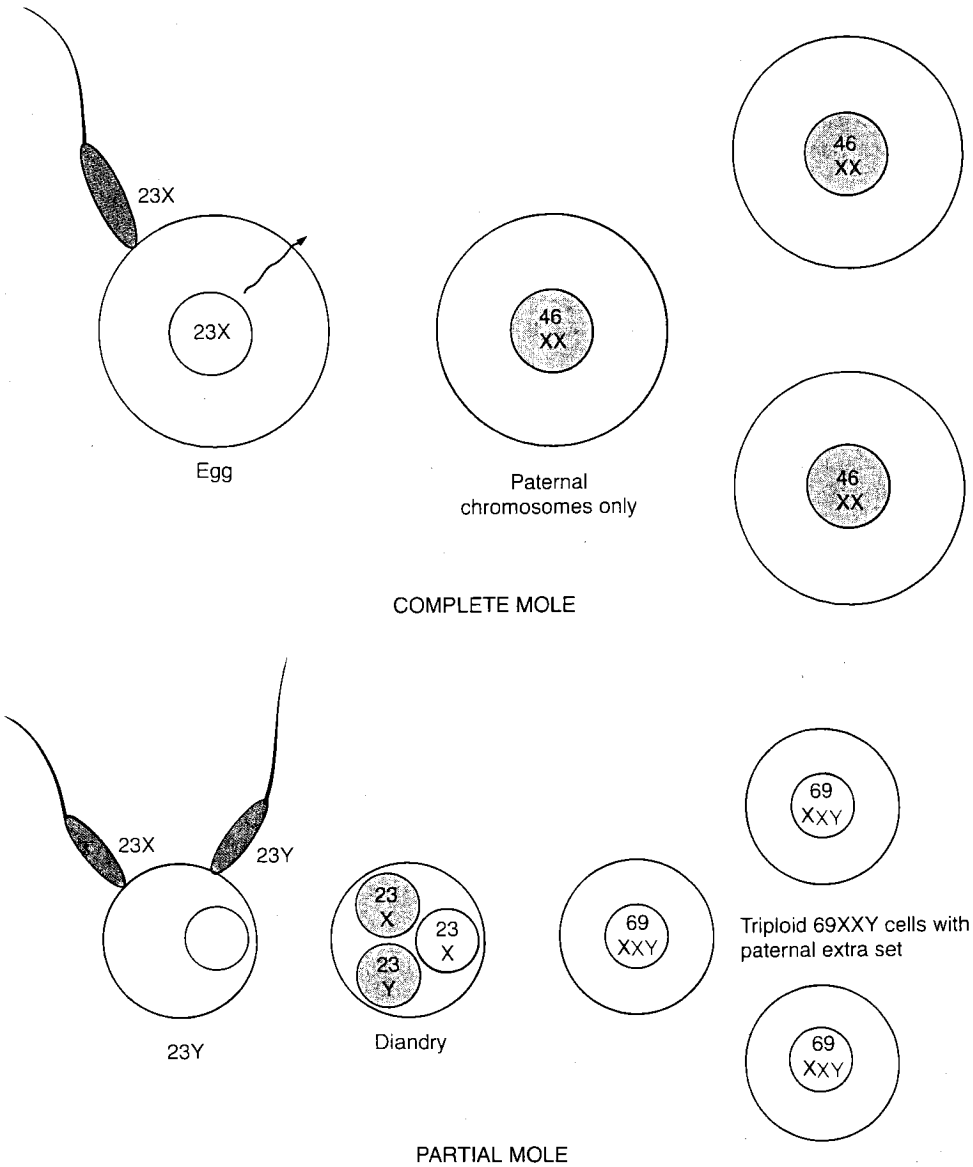


Figure 23–61. Patterns of fertilization to account for chromosomal origin of complete (46XX) and triploid partial moles (XXY). In a complete mole, a single sperm fertilizes an egg that has lost its chromosomes. Partial moles are due to fertilization of an egg by two sperm—one 23X and one 23Y. (Adapted with permission from Szulman, A.E.: Syndromes of hydatidiform moles: Partial vs. complete. J. Reprod. Med. 29:788, 1984.)

Table 23–4. FEATURES OF COMPLETE VERSUS PARTIAL HYDATIDIFORM MOLE

FEATURE	COMPLETE MOLE	PARTIAL MOLE
Karyotype	46,XX (46,XY)	Triploid
Villous edema	All villi	Some villi
Trophoblast proliferation	Diffuse; circumferential	Focal; slight
Atypia	Often present	Absent
Serum HCG	Elevated	Less elevated
HCG in tissue	++++	+
Behavior	2% choriocarcinoma	Rare choriocarcinoma

HCG = Human chorionic gonadotropin.

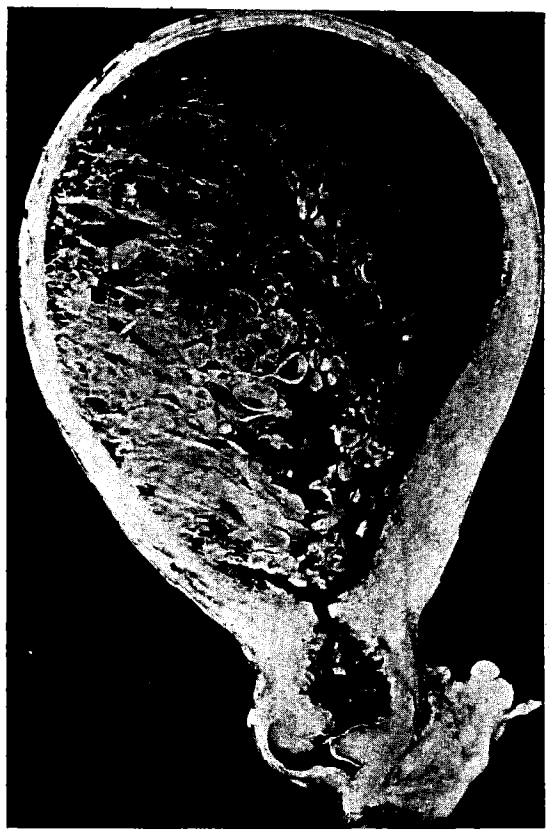


Figure 23–62. Hydatidiform mole. The uterus is filled with the classic mass of grape-like clusters. (Courtesy of Dr. A. Hertig. With permission from Anderson, W.A.D.: Textbook of Pathology. St. Louis, C.V. Mosby Co., 1971.)

tous, and there is diffuse trophoblast hyperplasia. Cytogenetic studies of these moles show that more than 90% have a 46,XX diploid pattern, all derived from the sperm (a phenomenon called androgenesis). They are presumed to be derived from fertilization, by a single sperm of an egg that has lost its chromosomes (Fig. 23–61). The remaining 10% are from the fertilization of such an empty egg by two sperm (46,XX and 46,XY). In both circumstances, embryonic development does not occur, and thus complete moles show no fetal parts.

In partial moles, some of the villi are edematous, and other villi show only minor changes; the trophoblastic proliferation is focal. In these moles, the karyotype is triploid (e.g., 69,XXY) or even occasionally tetraploid (92,XXXY). The moles result from fertilization of an egg with two sperm, one bearing 23,X and the other 23,Y (see Fig. 23–61). The embryo is viable for weeks, and thus fetal parts may be present when the resultant mole is aborted. In contrast to complete moles, partial moles are rarely followed by choriocarcinoma.

MORPHOLOGY. In most instances, moles develop within the uterus, but they may occur in any ectopic site of pregnancy. When discovered, usually in the fourth or fifth month of gestation, the uterus is usually larger (but may be normal sized, or even smaller) than anticipated for the duration of the pregnancy. The uterine cavity is filled with a delicate, friable mass of thin-walled, translucent, cystic, grape-like structures (Fig. 23–62). Careful dissection may disclose a small, usually collapsed amniotic sac. Fetal parts are frequently seen in partial moles but are never found in complete moles (unless a twin pregnancy).

In **partial moles** (Fig. 23–63*A*), the villous edema involves only a proportion of villi, and the trophoblastic proliferation is focal and slight.

Microscopically, the **complete mole** shows hydropic swelling of most chorionic villi and virtual absence or inadequate development of vascularization of villi. The central substance of the villi is a loose, myxomatous, edematous stroma, and they may be covered by a layer of chorionic epithelium, both cytotrophoblast and syncytial trophoblast (see Fig. 23–63*B*). At the opposite end of the spectrum are moles having similar cystic dilation of villi, accompanied, however, by circumferential and striking proliferation of the chorionic epithelium to produce sheets and masses of both cytotrophoblast and syncytial trophoblast. Histologic grading does not predict the outcome;[112] therefore all moles should be carefully followed by monitoring of the human chorionic gonadotropin levels.

CLINICAL COURSE. Most patients have abnormal uterine bleeding that usually begins early in the course of the pregnancy and is accompanied by the passage of a thin, watery fluid and bits of tissue seen as small, grape-like masses. The uterine enlargement is more rapid than anticipated. Ultrasound examination permits a definitive diagnosis in most cases.

In the classic case, quantitative analysis of human chorionic gonadotropin shows levels of hormone greatly exceeding those produced by a normal pregnancy of similar age. Serial hormone de-

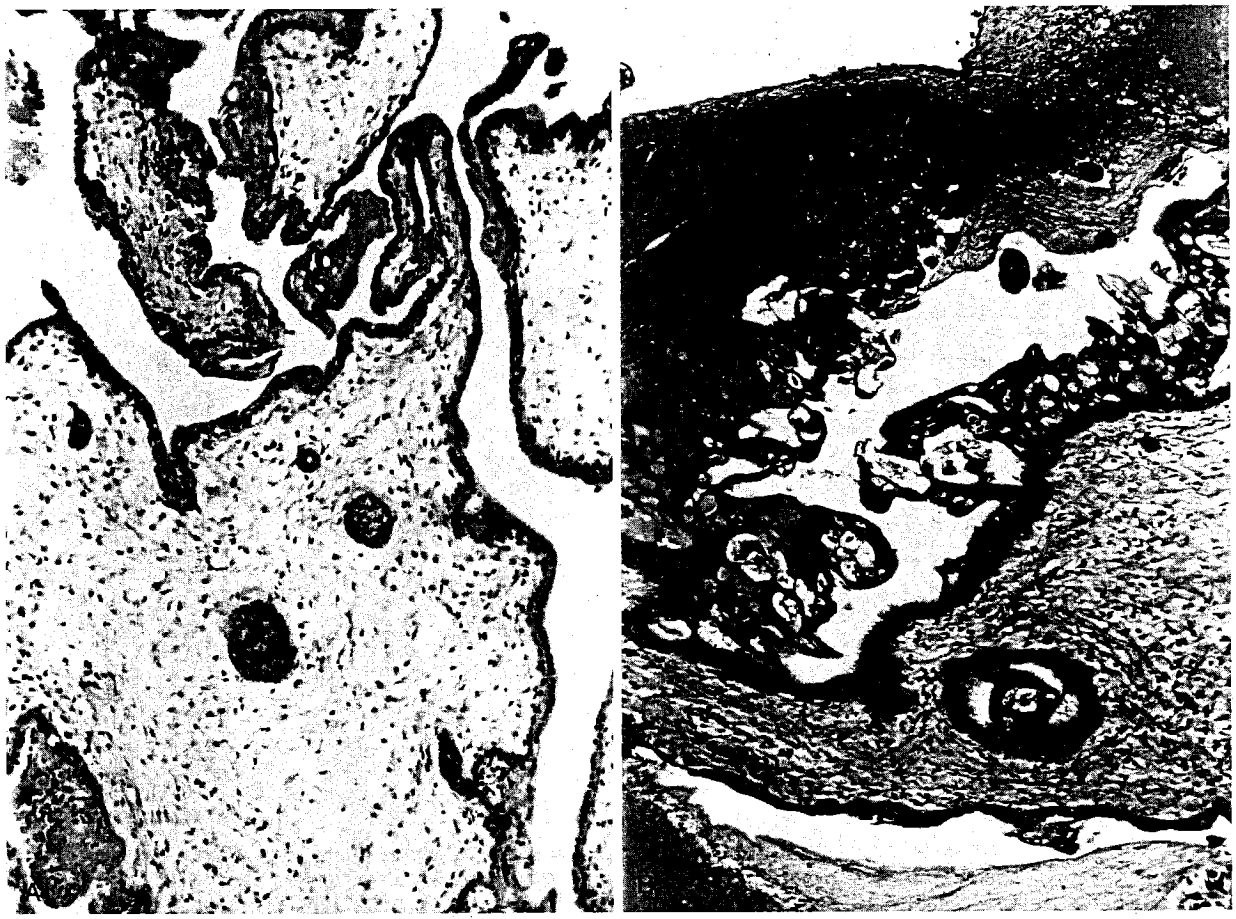


Figure 23–63. *A*, Partial mole showing villous edema and minimal focal epithelial hyperplasia. *B*, Complete mole with widespread trophoblast hyperplasia. (Courtesy of Dr. J. Lage.)

termination indicates a rapidly mounting level that climbs faster than the usual normal single or even multiple pregnancy.

Once the diagnosis is made, the mole must be removed by thorough curettage. In patients not desirous of further childbearing, hysterectomy may be performed. The course following curettage alone depends on the malignant potential of the removed uterine contents. From many studies, it is clear that 80 to 90% of these moles remain benign and give no further difficulty. Ten per cent develop into invasive moles and 2.5% into choriocarcinoma.[113]

Invasive Mole

This is defined as a mole that penetrates and may even perforate the uterine wall. There is invasion of the myometrium by hydropic chorionic villi, accompanied by proliferation of both cytotrophoblast and syncytiotrophoblast. The tumor is locally destructive and may invade parametrial tissue and blood vessels. Hydropic villi may embolize to distant sites, such as lungs and brain, but do not grow in these organs as true metastases, and even before the advent of chemotherapy, they eventually regressed unless fatal hemorrhage occurred. Clinically the tumor is manifested by vaginal bleeding and irregular uterine enlargement. It is always associated with a persistent elevated human chorionic gonadotropin level and varying degrees of luteinization of the ovaries. The tumor responds well to chemotherapy. Although invasive mole is biologically benign, rupture of the uterus may lead to hemorrhage.

Choriocarcinoma

Gestational choriocarcinoma is an epithelial malignancy of trophoblastic cells derived from any form of previously normal or abnormal pregnancy.[108] Although most cases arise in the uterus, ectopic pregnancies provide extrauterine sites of origin. Choriocarcinoma is a rapidly invasive, widely me-

Figure 23–64. Low-power view of choriocarcinoma of uterine wall showing invasion of underlying myometrium.

tastasizing malignancy, but once identified, it responds well to chemotherapy.

INCIDENCE. This is an uncommon condition that arises in 1 in 20,000 to 30,000 pregnancies in the United States. It is much more common in some Asian countries, for example, 1 in 2500 pregnancies in Ibadan. It is preceded by several conditions: 50% arise in hydatidiform moles, 25% in previous abortions, approximately 22% in normal pregnancies, and the rest in ectopic pregnancies and genital and extragenital teratomas. About 1 in 40 hydatidiform moles may be expected to give rise to a choriocarcinoma, in contrast to 1 in approximately 150,000 normal pregnancies.

MORPHOLOGY. Classically, the choriocarcinoma is a soft, fleshy, yellow-white tumor, with a marked tendency to form large pale areas of ischemic necrosis, foci of cystic softening, and extensive hemorrhage. Histologically, it is a purely epithelial cellular malignancy that does not produce chorionic villi and that grows, as do other cancers, by the abnormal proliferation of both cytotrophoblast and syncytiotrophoblast (Figs. 23–64 and 23–65).

It is sometimes possible to identify anaplasia within such abnormal proliferation replete with abnormal mitoses. The tumor invades the underlying myometrium, frequently penetrates blood vessels and lymphatics, and in some cases extends out onto the uterine serosa and adjacent structures. In its rapid growth, it is subject to hemorrhage, ischemic necrosis, and secondary inflammatory infiltration.

In fatal cases, metastases are found in the lungs, brain, bone marrow, liver, and other organs.

Occasionally, metastatic choriocarcinoma is discovered without a detectable primary in the uterus (or ovary), presumably because the primary has undergone total necrosis.

CLINICAL COURSE. Classically, the uterine choriocarcinoma does not produce a large, bulky mass. It becomes manifest only by irregular spotting of a bloody, brown, sometimes foul-smelling fluid. This discharge may appear in the course of an apparently normal pregnancy, after a miscarriage, or after a curettage.

Sometimes the tumor does not appear until months later. Usually, by the time the tumor is discovered locally, radiographs of the chest and bones already disclose the presence of metastatic lesions. The human chorionic gonadotropin titers are elevated to levels above those encountered in hydatidiform moles. Occasional tumors, however, produce little hormone, and some tumors have become so necrotic as to become functionally inactive.

Widespread metastases are characteristic of these tumors. Favored sites of involvement are the lungs (50%) and vagina (30 to 40%), followed in descending order of frequency by the brain, liver, and kidney.

The treatment of trophoblastic neoplasms depends on the type and stage of tumor and includes evacuation of the contents of the uterus, surgery, and chemotherapy. The results of chemotherapy for gestational choriocarcinoma are spectacular and have resulted in up to 100% cure or remission in

Figure 23–65. High-power detail of choriocarcinoma illustrating the two types of epithelial cells—cytotrophoblast and syncytiotrophoblast.

all patients except some who had high-risk metastatic trophoblastic disease. Many of the cured patients have had normal subsequent pregnancies and deliveries. By contrast, nongestational choriocarcinomas are much more resistant to therapy.

Placental Site Trophoblastic Tumor

This rare tumor is characterized by the presence of proliferating trophoblastic tissue deeply invading the myometrium and is composed largely of an *intermediate trophoblast.* These cells are usually mononuclear, rather than syncytial, but are larger and have more abundant cytoplasm than the regular cytotrophoblast; in contrast to syncytiotrophoblast (which produces human chorionic gonadotropin), intermediate trophoblast cells are immunoreactive for human placental lactogen. This lesion differs from choriocarcinoma in the absence of cytotrophoblastic elements and in the low level of human chorionic gonadotropin production. The tumors are locally invasive, but many are self-limited and subject to cure by curettage. Malignant variants, however, have been reported; they are distinguished by a high mitotic index, extreme cellularity, extensive necrosis, local spread, or even widespread metastases. About 10% result in disseminated metastases and death.[108,114]

1. McLean, J.M.: Embryology and anatomy of the female genital tract and ovary. *In* Fox, H. (ed.): Haines and Taylor's Obstetrical and Gynaecologic Pathology, 4th ed. Edinburgh, Churchill Livingstone, 1987, pp. 1–50.
2. Parmley, T.: Embryology of the female genital tract. *In* Kurman, R. (ed.): Blaustein's Pathology of the Female Genital Tract, 4th ed. New York, Springer-Verlag, 1994 (in press).
3. Donahoe, P.K.: The diagnosis and treatment of infants with intersex abnormalities. Pediatr. Clin. North Am. *34*:1333, 1987.
4. Kurman, R. (ed.): Blaustein's Pathology of the Female Genital Tract, 4th ed. New York, Springer-Verlag, 1994 (in press).
5. Fox, H. (ed.): Haines and Taylor's Obstetrical and Gynaecologic Pathology, 4th ed. Edinburgh, Churchill Livingstone, 1987.
6. Benirschke, K., and Kaufmann, P.: Pathology of the Human Placenta, 2nd ed. New York, Springer-Verlag, 1990.
7. Danforth, D.N. (ed.): Obstetrics and Gynecology, 5th ed. Philadelphia, Harper & Row, 1988.
8. Buchsbaum, H. (ed.): Gynecology and Obstetrics. Hagerstown, MD, Harper & Row, 1988.
9. Knapp, R.C., and Berkowitz, R.S. (eds.): Gynecologic Oncology, 2nd ed. New York, McGraw-Hill, 1993.
10. Holmes, K.K., et al. (eds.): Sexually Transmitted Diseases. New York, McGraw-Hill, 1984.
11. Koutsky, L.A. et al.: The frequency of unrecognized type 2 herpes simplex virus infection among women. Implications for the control of genital herpes. Sex. Transm. Dis. *17*:2, 1990.
12. Prober, C.G.: Herpetic vaginitis in 1993. Clin. Obstet. Gynecol. 36:177–187, 1993.
13. Wolner-Hanssen, P., et al.: Clinical manifestations of vaginal trichomoniasis. J.A.M.A. *261*:4, 1989.
14. Mann, M.S., et al.: Vulvar vestibulitis: Significant clinical variables and treatment outcome. Obstet. Gynecol. *79*:1, 1992.
15. Wilkinson, E.J.: Normal histology, and nomenclature of the vulva and malignant neoplasms, including VIN. Dermatol. Clin. *10*:283–296, 1992.
16. Report of the ISSVD Terminology Committee. Proceedings of the VIII World Congress (1986), Stockholm, Sweden. J. Reprod. Med. *31*:973, 1986.
17. Wilkinson, E.J.: Premalignant and malignant diseases of the vulva. *In* Kurman, R. (ed.): Blaustein's Pathology of the Female Genital Tract, 4th ed. New York, Springer-Verlag, 1994 (in press).
18. Crum, C.P.: Vulvar intraepithelial neoplasia: Histology and associated viral changes. Contemp. Issues Surg. Pathol. *9*:119, 1987.
19. Lynch, P.J.: Condyloma acuminata. Clin. Obstet. Gynaecol. *28*:142, 1985.
20. Gissman, L., et al.: Human papilloma virus type 6 and 11 DNA sequences in genital and laryngeal papillomas. Proc. Natl. Acad. Sci. (U.S.A.) *80*:560, 1983.
21. Zaino, R.J.: Carcinoma of the vulva, urethra, and Bartholin's glands. Contemp. Issues Surg. Pathol. *9*:119, 1987.
22. Crum, C.P.: Carcinoma of the vulva: Epidemiology and pathogenesis. Obstet. Gynecol. *79*:3, 1992.
23. Leibowitch, M., et al.: The epithelial changes associated with squamous cell carcinoma of the vulva: A review of the clinical, histological and viral findings in 78 women. Br. J. Obstet. Gynaecol. *97*:1135, 1990.
24. Michael, H., and Roth, L.M.; Paget's disease, skin appendage tumors, and cysts of the vulva. Contemp. Issues Surg. Pathol. *9*:25, 1987.
25. Chung, A. F., et al.: Malignant melanoma of the vulva: A report of 44 cases. Obstet. Gynecol. *45*:6, 1975.
26. Hilborne, L.H., and Fu, Y.S.: Intraepithelial, invasive and metastatic neoplasms of the vagina. Contemp. Issues Surg. Pathol. *9*:181–208, 1987.
27. Herbst, A.L.: Clear cell adenocarcinoma and the current status of DES-exposed females. Cancer *48*:484, 1981.
28. Scully, R.E., and Welch, W.R.: Pathology of the female genital tract after prenatal exposure to diethylstilbestrol. *In* Herbst, A.L., and Bern, H.A. (eds.): Developmental Effects of Diethylstilbestrol in Pregnancy. New York, Thieme-Stratton, 1981, pp. 26–45.
29. Copeland, L.J., et al.: Sarcoma botryoides of the female genital tract. Obstet. Gynecol. *66*:262, 1985.
30. Cancer Manual, 8th ed. Boston, American Cancer Society, 1990, p. 390.
31. Coppleson, M., and Reid, B.: Preclinical Carcinoma of the Cervix Uteri. Oxford, Pergamon, 1967.
32. Winkler, B., and Crum, C.P.: *Chlamydia trachomatis* infection of the female genital tract. Pathogenetic and clinicopathologic correlations. Pathol. Annu. *5*:193–221, 1984.
33. Kiviat, N.B., et al.: Histopathology of endocervical infection by Chlamydia trachomatis, herpes simplex virus, Trichomonas vaginalis, and Neisseria gonorrheae. Hum. Pathol. *21*:831–837, 1990.
34. Silverberg, E., and Lubera, J.: Cancer statistics. Cancer *39*:1, 1989.
35. Koutsky, L.A., et al.: A cohort study of the risk of cervical intraepithelial neoplasia grade 2 or 3 in relation to papillomavirus infection. N. Engl. J. Med. *327*:1272–1278, 1992.
36. Herrero, R., et al.: Sexual behavior, venereal diseases, hygiene practices and invasive cervical cancer in a high-risk population. Cancer *65*:380–386, 1990.
37. zur Hausen, H., and Schneider, A.: The role of human papilloma viruses in human urogenital cancer. *In* Salzman, N., and Howley, P. (eds.): The Papovaviridae. New York, Plenum Press, 1987, pp. 245–263.
38. Lorincz, A.T., et al.: Human papillomavirus infection of

the cervix: Relative risk associations of 15 common anogenital types. Obstet. Gynecol. *79*:328–337, 1992.
39. Crum, C.P., and Nuovo, G.J.: Genital Papillomaviruses and Related Neoplasms. New York, Raven Press, 1991.
40. Crook, T., and Vousden, K.H.: Properties of p53 mutations detected in primary and secondary cervical cancers suggest mechanisms of metastasis and involvement of environmental carcinogens. Embo J. *11*:3935–3940, 1992.
41. Wright, T., et al.: Precursors of cervical carcinoma. *In* Kurman, R. (ed.): Blaustein's Pathology of the Female Genital Tract, 4th ed. New York, Springer-Verlag, 1994 (in press).
42. Richart, R.M.: Cervical intraepithelial neoplasia. *In* Sommers, S.C.: Pathology Annual. New York, Appleton-Century-Crofts, 1973, pp. 301–328.
43. Crum, C.P., et al.: Cervical papillomaviruses segregate within morphologically distinct precancerous lesions. J. Virol. *54*:675–681, 1985.
44. The Bethesda System for reporting cervical/vaginal cytologic diagnoses. Report of the 1991 Bethesda Workshop. Am. J. Surg. Pathol. *16*:914–916, 1992.
45. Ostor, A.G.: Natural history of cervical intraepithelial neoplasia: A critical review. Int. J. Gynecol. Pathol. *12*:186–192, 1993.
46. Van Nagell, J.R., et al.: Small cell cancer of the uterine cervix. Cancer *40*:2243–2249, 1977.
47. Rotman, M., et al.: Prognostic factors in cervical carcinoma: Implications in staging and management. Cancer *48*:560, 1981.
48. Smotkin, D., et al.: Human papillomavirus DNA in adenocarcinoma and adenosquamous carcinoma of the uterine cervix. Obstet. Gynecol. *68*:241–244, 1986.
49. Rotterdam, H.: Chronic endometritis. A clinicopathologic study. Pathol. Annu. *13*(Part 2):209, 1978.
50. Kiviat, N.B., et al.: Endometrial histopathology in patients with culture-proved upper genital tract infection and laparoscopically diagnosed acute salpingitis. Am. J. Surg. Pathol. *14*:167–175, 1990.
51. Fox, H., and Buckley, C.H.: Current concepts of endometriosis. Clin. Obstet. Gynecol. *11*:279, 1984.
52. Dahlenbach-Hellweg, G.: Histopathology of the Endometrium, 4th ed. New York, Springer-Verlag, 1993.
53. Ober, W.B.: Effects of oral and intrauterine administration of contraceptives on the uterus. Hum. Pathol. *8*:513, 1977.
54. Hertig, A.T., and Sommers, S.C.: Genesis of endometrial carcinoma. 1. Study of prior biopsies. Cancer *2*:964, 1949.
55. Kurman, R.J., et al.: The behavior of endometrial hyperplasia. A long-term study of untreated hyperplasia in 170 patients. Cancer *56*:403, 1985.
56. renczy, A., and Gelfand, M.: The biologic significance of cytologic atypia in progestogen-treated endometrial hyperplasia. Am. J. Obstet. Gynecol. *160*:126–131, 1989.
57. Winkler, B., et al.: Pitfalls in the diagnosis of endometrial neoplasia. Obstet. Gynecol. *64*:185–194, 1984.
58. Corley, D., et al.: Postmenopausal bleeding from unusual endometrial polyps in women on chronic tomoxifen therapy. Obstet. Gynecol. *79*:111–116, 1992.
59. Fletcher, J.A., et al.: Clonal 6p21 rearrangement is restricted to the mesenchymal component of an endometrial polyp. Genes Chromosom. Cancer *5*:260–263, 1992.
60. Fletcher, J.A., et al.: Chromosome aberrations in uterine smooth muscle tumors: Potential diagnostic relevance and cytogenetic instability. Cancer Res. *50*:4092–4097, 1990.
61. Brinton, L.A., et al.: Reproductive, menstrual, and medical risk factors for endometrial cancer: Results from a case-control study. Am. J. Obstet. Gynecol. *167*:1317–1325, 1992.
62. Silverberg, S.G., and Kurman, R.J.: Atlas of Tumor Pathology. Tumors of the Uterine Corpus and Gestational Trophoblastic Disease. Washington, DC, Armed Forces Institute of Pathology, 1991, pp. 219–287.
63. Connelly, P.J., et al.: Carcinoma of the endometrium. III. Analysis of 865 cases of adenocarcinoma and adenoacanthoma. Obstet. Gynecol. *59*:569, 1982.
64. Hendrickson, M.R., et al.: Adenocarcinoma of the endometrium: Analysis of 256 cases with disease limited to uterine corpus. Analysis of prognostic variables. Gynecol. Oncol. *12*:373, 1982.
65. Cowles, T.A., et al.: Comparison of clinical and surgical staging in patients with endometrial carcinoma. Obstet. Gynecol. *66*:413, 1985.
66. Christopherson, W.M., and Richardson, M.: Uterine mesenchymal tumors. Pathol. Annu. *16*:215, 1981.
67. Gelsinger, K.R., et al.: Malignant mixed müllerian tumors. Cancer *54*:1281, 1987.
68. Chu, M.T., et al.: Expression of various antigens by different components of uterine mixed Müllerian tumors. Acta Pathol. *38*:35, 1988.
69. Silverberg, S.G., et al.: Carcinosarcoma (malignant mixed mesodermal tumor of the uterus). Int. J. Gynecol. Pathol. *9*:1, 1990.
70. Buscema, J., et al.: Epithelioid leiomyosarcoma. Cancer *47*:1192, 1986.
71. Young, R.H., and Scully, R.E.: Ovarian pathology in infertility. *In* Kraus, F.T. (ed.): Pathology of Reproductive Failure. Baltimore, Williams & Wilkins, 1991, pp. 104–139.
72. McKenna, T.J.: Pathogenesis and treatment of polycystic ovary syndrome. N. Engl. J. Med. *318*:558, 1988.
73. Russel, P., and Bannatyne, P.: Surgical Pathology of the Ovaries. Edinburgh, Churchill Livingstone, 1989.
74. Young, R.H., et al.: The ovary. *In* Sternberg S., et al. (eds.): Diagnostic Surgical Pathology. New York, Raven Press, 1994.
75. Piver, M.S., et al.: Epidemiology and etiology of ovarian cancer. Semin. Oncol. *18*:177–184, 1991.
76. Cramer, D.W., et al.: Factors affecting the association of contraceptives and ovarian cancer. N. Engl. J. Med. *307*:1047, 1982.
77. Bast, R.C., et al.: Malignant transformation of ovarian epithelium. J. Natl. Cancer Inst. *84*:557–558, 1992.
78. Bell, D.A., et al.: Peritoneal implants of ovarian serous borderline tumors. Cancer *62*:2212–2222, 1988.
79. Watkin, W., et al.: Mucinous carcinoma of the ovary. Cancer *69*:208–212, 1992.
80. Snyder, R.R., et al.: Endometrial proliferative and low malignant potential tumors of the ovary. Am. J. Surg. Pathol. *12*:661, 1988.
81. Eifel, P., et al.: Simultaneous presentation of carcinoma involving the ovary and uterine corpus. Cancer *50*:163–170, 1982.
82. Teilum, G.: Special Tumors of Ovary and Testis and Related Extragonadal Lesions. Philadelphia, W.B. Saunders Co., 1976.
83. Shevchuk, M.M., et al.: Histogenesis of Brenner tumors. I. Histology and ultrastructure. II. Histochemistry and CEA. Cancer *46*:2607, 1980.
84. Silva, E.G., et al.: Ovarian carcinomas with transitional cell carcinoma pattern. Cancer *93*:457–465, 1990.
85. Altaras, M.M., et al.: The role of cancer antigen CA 125 in the management of various epithelial carcinomas. Gynecol. Oncol. *30*:26, 1988.
86. Linder, D., et al.: Pathogenetic origin of benign ovarian teratomas. N. Engl. J. Med. *292*:63, 1975.
87. Mutter, G.L.: Teratoma genetics and stem cells: A review. Obstet. Gynecol. Surv. *42*:661, 1987.
88. Norris, H.J., et al.: Immature (malignant) teratoma of the ovary. A clinical and pathological study of 58 cases. Cancer *37*:2359, 1976.
89. Gordon, T., et al.: Dysgerminoma. A review of 158 cases from the Emil Novak ovarian tumor registry. Obstet. Gynecol. *58*:497, 1981.

90. Young, R.H., and Scully, R.: Ovarian sex cord-stromal tumors. Recent progress. Int. J. Gynecol. Pathol. *1*:101, 1982.
91. Prat, J., and Scully, R.E.: Cellular fibromas and fibrosarcomas of the ovary. Cancer *47*:2663, 1981.
92. Roth, L.M., et al.: Sertoli-Leydig cell tumors: A clinicopathologic study of 34 cases. Cancer *48*:187, 1981.
93. Scully R.E.: The ovary. *In* Wolfe, H. (ed.): Endometrial Pathology. New York, Springer-Verlag, 1986.
94. Young, R.H., and Scully, R.E.: Metastatic tumors of the ovary. *In* Kurman, R.J. (ed.): Blaustein's Pathology of the Female Genital Tract, 4th ed. New York, Springer-Verlag, 1994 (In press).
95. Naeye, R.L.: Disorders of the Placenta Fetus and Neonate: Diagnosis and Clinical Significance. St. Louis, Mosby-Year Book, 1992, pp. 113–114.
96. Grossman, J.H.: Infections affecting the placenta. *In* Lavery, J.P. (ed.): The Placenta. Rockville, MD, Aspen Publishing, 1987, pp. 131–134.
97. Morrison, J.E.: Placenta accreta: A clinicopathologic review of 67 cases. Obstet. Gynecol. Annu. *7*:107, 1978.
98. Lage, J.M., et al.: The twin placenta. *In* Lavery, J.P. (ed.): The Placenta. Rockville, MD, Aspen Publishing, 1987, p. 67.
99. Wilcox, A.J.: Incidence of early loss of pregnancy. N. Engl. J. Med. *319*:189, 1988.
100. Little, A.B.: There's many a slip 'twixt implantation and the crib. N. Engl. J. Med. *319*:241, 1988.
101. Kalousek, D.K., and Lau, A.E.: Pathology of spontaneous abortion. *In* Dimmick, J.E., and Kalousek, D.K. (eds.): Developmental Pathology of the Embryo and Fetus. Philadelphia, J.B. Lippincott Co., 1992, p. 62.
102. Rushton, D.I.: Examination of products of conception from previable human pregnancies. J. Clin. Pathol. *34*:819, 1981.
103. Hockberger, R.S.: Ectopic pregnancy. Obstet. Gynecol. Emerg. *5*:481, 1987.
104. Weiner, G.P.: The clinical spectrum of pre-clampsia. Am. J. Kidney Dis. *9*:312, 1987.
105. Khong, T.Y., et al.: Inadequate maternal vascular response to placentation in pregnancies complicated by pre-eclampsia and by small for gestational age infants. B.M.J. *93*:1049, 1986.
106. Ferris, T.F.: Pregnancy, preeclampsia and the endothelial cell. N. Engl. J. Med. *325*:1439–1440, 1991.
107. Friedman, S.A.: Pre-clampsia: A review of the role of prostaglandins. Obstet. Gynecol. *71*:122, 1988.
108. Lage, J.M., and Young, R.H.: Pathology of trophoblastic disease. *In* Clement, P.B., and Young, R.H. (eds.): Tumors and Tumor-like Conditions of the Uterine Corpus and Cervix, New York, Churchill Livingstone, 1993, pp. 419–475.
109. Berkowitz, R.S., and Goldstein, D.P.: Management of molar pregnancy and gestational trophoblastic tumors. *In* Knapp, R.C., and Berkowitz, R.S. (eds.): Gynecologic Oncology. New York, McGraw-Hill, 1993, pp. 328–340.
110. Bracken, M.B., et al.: Epidemiology of hydatidiform mole and choriocarcinoma. Epidemiol. Rev. *6*:52, 1984.
111. Lage, J.M., et al.: Hydatidiform moles. Application of flow cytometry to diagnosis. Am. J. Clin. Pathol. *89*:596, 1988.
112. Genest, D.R., et al.: A clinico-pathologic study of 153 cases of complete hydatidiform mole (1980–1990): Histologic grade lacks prognostic significance. Obstet. Gynecol. *78*:402–409, 1991.
113. Lurain, J.R., et al.: Natural history of hydatidiform mole after primary evacuation. Am. J. Obstet. Gynecol. *145*:591, 1983.
114. Finkler, N.J., et al.: Clinical experience with placental site trophoblastic tumors at the New England Trophoblastic Disease Center. Obstet. Gynecol. *71*:854, 1988.

Downloaded from jmg.bmj.com on March 6, 2012 - Published by group.bmj.com

Review article

Journal of Medical Genetics (1975). **12**, 230.

Transplacental passage of blood cells

JIM SCHRÖDER

From the Folkhälsan Institute of Genetics, 00101 Helsinki 10, Finland

Passage of erythrocytes between mother and fetus is a well-documented phenomenon. In 1946, Naeslund and his collaborators (Naeslund and Nylin; Hedenstedt and Naeslund) showed that radioactively labelled erythrocytes, and elliptiocytes, when injected into the maternal circulation before delivery, could be found postnatally in the blood of the infant. Similar results have been obtained with various agglutination tests (von Muralt, 1967; Eimer and Weiland, 1969) and immunofluorescence (Fischer and Müller, 1967).

Since the introduction of the acid elution method for recognizing cells carrying fetal haemoglobin (Kleihauer *et al*, 1957), the occurrence of fetal erythrocytes in maternal blood has been studied in detail (Zipursky *et al*, 1959; Wimhöfer *et al*, 1962; Zilliacus, 1963; Clayton *et al*, 1964; Cohen *et al*, 1964; Cohen and Zuelzer, 1967; Betke and Nierhaus, 1968; Bartsch, 1972). Today, it is generally known that Rh immunization is caused by fetal erythrocytes from Rh-positive fetuses which enter the circulation of Rh-negative mothers (Woodrow and Finn, 1966; Clarke, 1967; Bartsch, 1972), and that this immunization can be prevented by administration of anti-D gammaglobulin of Rh^- mothers carrying Rh^+ fetuses (Clarke, 1967; Pollack *et al*, 1971a; 1971b; Bartsch, 1972; Eklund and Nevanlinna, 1973).

During the last 10 years attempts have been made to demonstrate the passage of leucocytes between mother and fetus. Desai and Creger (1963) reported maternal leucocytes in the cord blood after normal pregnancies. Their results were later challenged by others, who found maternal leucocytes in the umbilical cord only in connection with malformations of fetus and placenta (Kadowaki *et al*, 1965; Turner *et al*, 1966; Olding, 1972). However, quinacrine mustard fluorescence of human chromosomes has been used to show that such a passage may occur in normal pregnancies (Schröder, 1974).

Fetal leucocytes seem to find their way into the maternal blood in most pregnancies (Walknowska *et al*, 1969; de Grouchy and Trébuchet, 1971; Schröder and de la Chapelle, 1972; Grosset *et al*, 1974), but seldom provoke the production of HL-A antibodies in the mother (Schröder *et al*, 1974).

In the mouse maternal erythrocytes and lymphocytes have also been demonstrated in the circulation of the offspring by several different methods (Finegold and Michie, 1961; Tuffrey *et al*, 1969a; 1969b; Barnes and Tuffrey, 1970).

It is my intention in this paper to summarize what is currently known of the passage of blood cells between mother and fetus in man, and to discuss whether there is any difference in the mechanisms by which erythrocytes and leucocytes cross the placenta. Moreover, I shall present results on the same subject in an animal system, the mouse.

Man

Erythrocytes

Fetus to mother. The occurrence of fetal erythrocytes in maternal blood during pregnancy and after delivery has been studied by the acid elution technique of Kleihauer *et al* (1957) and by immunofluorescence (Cohen *et al*, 1964). Both methods are reliable, and allow recognition of a low proportion of fetal cells among the maternal cells. The method of Kleihauer is based on the fact that cells containing Hb A (the haemoglobin of adults) become ghosted and unstainable in acid buffer, while cells containing Hb F (fetal haemoglobin) are resistant to this treatment and remain strongly stainable (Fig. 1). The use of immunofluorescence in this connection is based on the incubation of maternal cells with fluorescein-conjugated antisera directed against the ABO antigens of the fetus. Fetal erythrocytes

Received 23 September 1974.

Exhibit AC

Downloaded from jmg.bmj.com on March 6, 2012 - Published by group.bmj.com

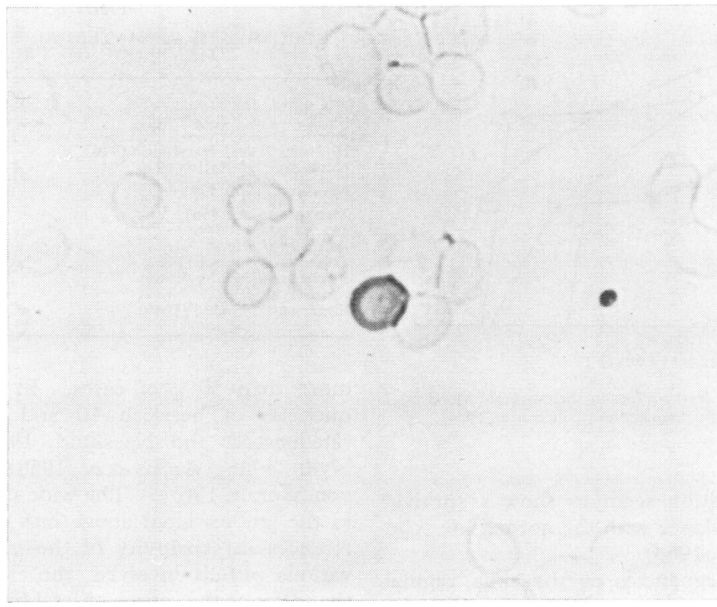


FIG. 1. Blood smear from a pregnant women showing one Hb F positive cell.

have been found in the maternal blood even during the first trimester. However, the proportion is usually less than 1 fetal cell per 50 000 maternal cells. Different authors have found somewhat different frequencies of mothers with fetal red cells in their blood at this stage of pregnancy (Clayton *et al*, 1964; Cohen *et al*, 1964; Betke and Nierhaus, 1968; Bartsch, 1972), but the values generally seem to be below 10%. During the second and third trimesters there is a rise in the frequency, since about 20–40% of mothers are positive immediately before delivery (Table I). At this stage more than 1 fetal cell per 50 000 maternal cells is found in 5–10% of the cases. The presence of fetal erythrocytes in the maternal circulation seems to be related to the blood groups of mother and fetus, since in ABO-compatible pregnancies fetal cells are detected twice as often as in ABO-incompatible ones (Clayton *et al*, 1964; Cohen *et al*, 1964; Woodrow and Finn, 1966; Cohen and Zuelzer, 1967; Betke and Nierhaus, 1968; Bartsch, 1972). Immediately after delivery fetal erythrocytes are found in 20–50% of all women (Table I). The difference in the frequency of mothers with fetal erythrocytes between ABO-compatible and ABO-incompatible mother/child pairs persists after delivery (Clayton *et al*, 1964; Cohen *et al*, 1964; Cohen and Zuelzer, 1967). Fetal erythrocytes are also found more often in the mother's blood in Rh-compatible than in Rh-incompatible pregnancies (Cohen and Zuelzer, 1967; Betke and Nierhaus, 1968).

TABLE I

OCCURRENCE OF FETAL ERYTHROCYTES IN THE MATERNAL CIRCULATION

Reference	% of Mothers with Fetal Erythrocytes in their Blood	
	During Pregnancy	At Term or after Delivery
Zipursky *et al* (1959)	—	21.0
Wimhöfer *et al* (1962)	—	15.4
Clayton *et al* (1964)	15.0	47.0
Cohen *et al* (1964)	17.0	50.0
Woodrow and Finn (1966)	3.5	19.8
Cohen and Zuelzer (1967)	—	47.0
Betke and Nierhaus (1968)	—	19.0
Bartsch (1972)	13.6	40.6

Erythrocytes have a life span of about 100–200 days. Where fetal erythrocytes are present in the maternal blood at delivery, and the blood groups of mother and child are incompatible, these cells are rapidly eliminated from the maternal blood (Cohen and Zuelzer, 1967). Where such incompatibility exists, agglutinated fetal erythrocytes have been found in the maternal blood after delivery (Zilliacus, 1963). In homospecific pregnancies (Fig. 2), on the other hand, fetal erythrocytes that have entered

Downloaded from jmg.bmj.com on March 6, 2012 - Published by group.bmj.com

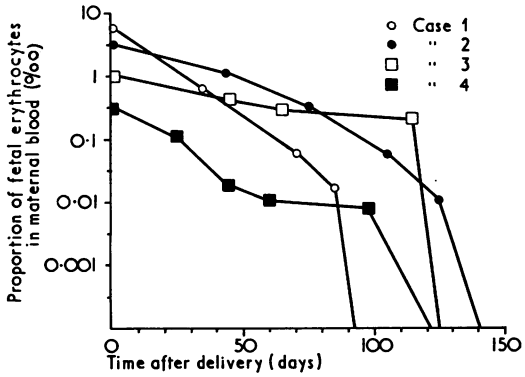


FIG. 2. Survival of fetal erythrocytes in the maternal blood after delivery in four homospecific pregnancies (Cohen *et al*, 1964).

the maternal circulation seem to show a survival time that is in accordance with the normal life span of RBC (Cohen *et al*, 1964).

In complicated pregnancies, eg, toxaemia, vaginal bleeding, etc, the proportion of fetal red cells in the maternal blood is considerably higher than in uncomplicated ones (Bartsch, 1972). Higher proportions are also found after abortions, amniocentesis, and Caesarean sections (Wimhöfer *et al*, 1962; Bartsch, 1972).

Thus fetal erythrocytes occur not infrequently in the maternal circulation during pregnancy. In most cases, however, the number of these cells is less than 1 per 50 000 maternal cells. The frequency of mothers with fetal erythrocytes and the proportion of such cells rise somewhat at delivery, and the survival time of fetal RBC in the maternal blood seems to be closely related to the ABO and Rh compatibility between mother and fetus. In some homospecific pregnancies fetal erythrocytes can still be found in the maternal blood up to 3 months after delivery.

Mother to fetus. All studies indicate (Table II) passage of maternal erythrocytes into the circulation of the fetus in some normal pregnancies, although the frequency is higher in complicated ones (toxicosis, Caesarean section, etc). Data on the subject are not consistent. The results of Naeslund (1950/1951), using ^{32}P, indicated such passage in 5% of all pregnancies, while his results with ^{59}Fe (1951) gave positive results in over 50% of cases. Cohen and Zuelzer (1965) found maternal cells in only 2.8% of infants when heel blood was studied, but in more than 10% in umbilical cord samples. The results of Mengert *et al* (1955), Fischer and Müller (1967), and Eimer and Weiland (1969) indicate passage in

TABLE II

OCCURRENCE OF MATERNAL ERYTHROCYTES IN THE BLOOD OF THE FETUS

Reference	% of Children with Maternal Erythrocytes
Hedenstedt and Naeslund (1946)	66
Naeslund and Nylin (1946)	16
Naeslund (1950/1951)	60 5
Mengert *et al* (1956)	80
Macris *et al* (1958)	12
Zarou *et al* (1964)	19
Cohen and Zuelzer (1967)	2.8
Fischer and Müller (1965)	58
von Muralt (1967)	10
Eimer and Weiland (1969)	66

more than 50% of cases. In other studies frequencies of between 10 and 66% were found (Hedenstedt and Naeslund, 1946; Naeslund and Nylin, 1946; Macris *et al*, 1958; Zarou *et al*, 1964; von Muralt, 1967). The wide range of frequencies in the studies cited above may be due to the differences in sensitivity of the methods used or to various pitfalls involved, the criteria for choosing the series or the source of fetal blood. The amount of maternal blood found in the infant forms 1–10% of the total fetal blood volume (Macris *et al*, 1958; Cohen and Zuelzer, 1965; Fischer and Müller, 1967; von Muralt, 1967). Maternal erythrocytes seem to occur in the blood of more than 10% of all infants without giving any clinical symptoms. When there has been copious bleeding from mother to fetus the infant often shows raised haemoglobin and haematocrit values, with decreased Hb F but raised Hb A values, cyanosis, and dyspnoea (von Muralt, 1967). After the elimination of the maternal cells, however, the symptoms subside (von Muralt, 1967).

The maternal red cells have been reported to disappear rapidly from the blood of the infant. With agglutination tests, Eimer and Weiland (1969) failed to detect maternal RBC in the blood of newborns 40 hours after delivery. Von Muralt (1967) found maternal cells 2–3 days after delivery, but none when the test was repeated 3–6 months later.

Leucocytes

Fetus to mother. After 10 years of intensive research the question of fetal erythrocytes in the maternal blood is fairly well documented, while the question whether fetal leucocytes are present in the maternal circulation is still somewhat controversial (Jacobs and Smith, 1969; Walknowska *et al*, 1969; de Grouchy and Trébuchet, 1971; Schröder and de la Chapelle, 1972).

Using chromosome studies, Walknowska *et al* (1969) were able to show that cells with a male

Downloaded from jmg.bmj.com on March 6, 2012 - Published by group.bmj.com

karyotype may be present in the circulation of mothers carrying a male fetus. These cells have been interpreted as fetal lymphocytes that have entered the maternal blood after passage through the placenta, since they have responded to PHA (phytohaemagglutinin) stimulation in 3-day cultures and entered mitosis. Walknowska *et al* (1969) found XY mitoses in blood cultures from 21 out of 30 pregnant women; 19 boys and 2 girls were born (Table III). Of the nine cases in which no male cells were found, six girls and three boys were born. The proportion of fetal mitoses in the 19 cases was between 0.2–1.5% of the mothers' mitoses, with a mean value of 0.5% (Table III, Fig. 3).

Since these values are much higher than those reported for fetal erythrocytes in the maternal circulation, Walknowska *et al* (1969) suggest that they do not reflect the situation *in vivo*. They suggested that preferential stimulation of fetal cells in the blood of mothers with PHA was responsible.

Similar values have been reported by de Grouchy and Trébuchet (1971), who studied the occurrence of fetal lymphocytes in the maternal blood by the same method as Walknowska *et al* (1969). Of 21 pregnant mothers, 12 had XY mitoses in their blood cultures (Table III). Of these, eight delivered boys and four girls. The remaining nine mothers had no XY mitoses in blood cultures made during pregnancy, and six of them gave birth to girls (Table III). The proportion (0.6%) of fetal mitoses found by de Grouchy and Trébuchet (1971) in maternal blood cultured during pregnancy is in good accord with the figures of Walknowska *et al* (1969) (Fig. 3). Occasional false positive cases were found by both teams and were attributed to earlier pregnancies of male fetuses, with persistence of fetal leucocytes in the maternal blood until the next pregnancy.

TABLE III

NUMBER OF XY MITOSES IN THE CIRCULATION OF PREGNANT WOMEN

Reference	No. of Women	Sex of Child	No. of Cells Studied	
			Total	No. of XY Cells
Walknowska *et al* (1969)	19	Male	7399	32 (0.5%)
	3		1771	0
	6	Female	2945	0
	2		1095	2 (0.3%)
De Grouchy and Trébuchet (1971)	8	Male	3766	21 (0.6%)
	3		2409	0
	6	Female	4002	0
	4		2204	5 (0.3%)

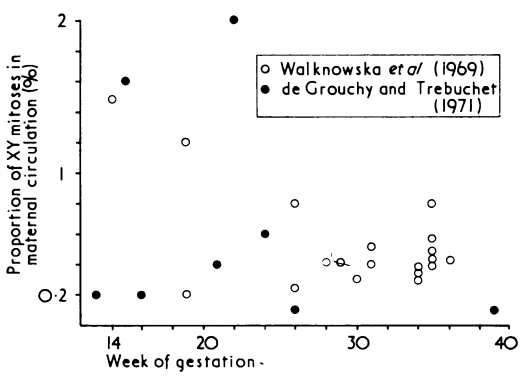


FIG. 3. Prevalence of XY mitoses in lymphocyte cultures of maternal blood at various stages of pregnancy.

Since both studies (Walknowska *et al*, 1969; de Grouchy and Trébuchet, 1971) are based on the finding of a single or a few XY mitosis among hundreds of maternal mitoses, the reliability of the method may be questioned. The criteria of a Y chromosome vary from one laboratory to another, and for adequate identification the morphology of the cell studied has to be clear. Jacobs and Smith (1969) have shown that occasional XY mitoses can be found in the blood even in newborn girls. Furthermore the method is very time-consuming, and only fetal cells capable of responding to PHA by mitotic division can be recognized.

The distal part of the Y chromosome shows bright fluorescence after QM staining in metaphase (Fig. 4) and interphase (Fig. 5) (Caspersson *et al*, 1970; Pearson *et al*, 1970). Male WBC can easily be distinguished from female WBC at both metaphase and interphase without time-consuming chromosome analysis. This method has been used successfully in clinical diagnoses (Caspersson *et al*, 1970; Polani and Mutton, 1971; Forsius *et al*, 1972).

The reliability of the method for detecting a low proportion of male interphase cells among female interphase cells was tested by making artificial mixtures of male and female lymphocytes (Schröder and de la Chapelle, 1972). The proportion of male WBC was 100, 10, 1, and 0%, and the slides were analysed blindly. The results (Table IV) indicate that the method can be used with confidence, provided very stringent criteria are set for a cell to be regarded as Y-positive.

The frequency of lymphocytes containing a Y-body (the Y chromosome at interphase) was studied in 25 cases between the fourth and ninth months of pregnancy. False-positive cases of the type proposed by Walknowska *et al* (1969) and de Grouchy and Trébuchet (1971) were excluded by choosing

Downloaded from jmg.bmj.com on March 6, 2012 - Published by group.bmj.com

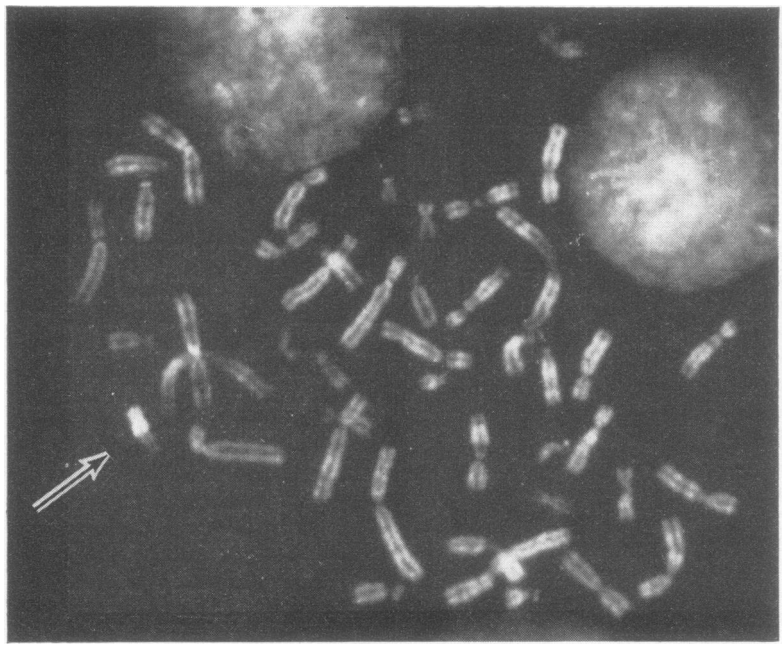


FIG. 4. A mitosis stained with QM from a blood culture of a mole with a large Y chromosome (arrow).

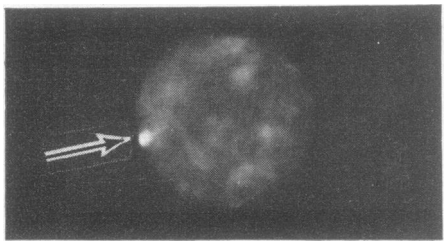


FIG. 5. A PHA-stimulated lymphocyte of a male showing a Y-body (arrow) after QM staining.

TABLE IV

PREVALENCE OF INTERPHASE NUCLEI WITH A Y-BODY IN ARTIFICIAL MIXTURES OF PHA-STIMULATED LYMPHOCYTES AND UNSTIMULATED LYMPHOCYTES FROM A NORMAL MALE AND A NORMAL FEMALE

(Schröder and de la Chapelle, 1972)

Proportion of Male Cells in Mixture (%)	No. of Cells Studied	Stimulated Lymphocytes		Unstimulated Lymphocytes	
		No. of Cells with Y-body	%	No. of Cells with Y-body	%
100	1000	351	35	410	41
10	1000	32	3.2	37	3.7
1	1000	4	0.4	3	0.3
0	1000	0	0	0	0

TABLE V

PRESUMPTIVE Y-BODIES IN PHA-STIMULATED LYMPHOCYTES FROM 25 WOMEN IN THE 4th–9th MONTH OF PREGNANCY

(Schröder and de la Chapelle, 1972)

No. of Women	Sex of Child	No. of Cells Studied	
		Total	With Presumptive Y-body
9	Male	49 500	45 (0.09%)
2		10 000	0
9	Female	46 600	0
5		20 500	22 (0.11%)

TABLE VI

POSSIBLE REASONS FOR APPARENT DISCREPANCIES BETWEEN MATERNAL LYMPHOCYTE Y-BODY FINDINGS AND SEX OF CHILD

Presumptive Y-bodies (female child)	Misinterpretation due to exceptionally bright autosomal fluorescence in mother or child Previous abortion of male fetus
No Y-bodies (male child)	Short or non-fluorescent Y Absence or very low incidence of fetal cells in maternal blood

Downloaded from jmg.bmj.com on March 6, 2012 - Published by group.bmj.com

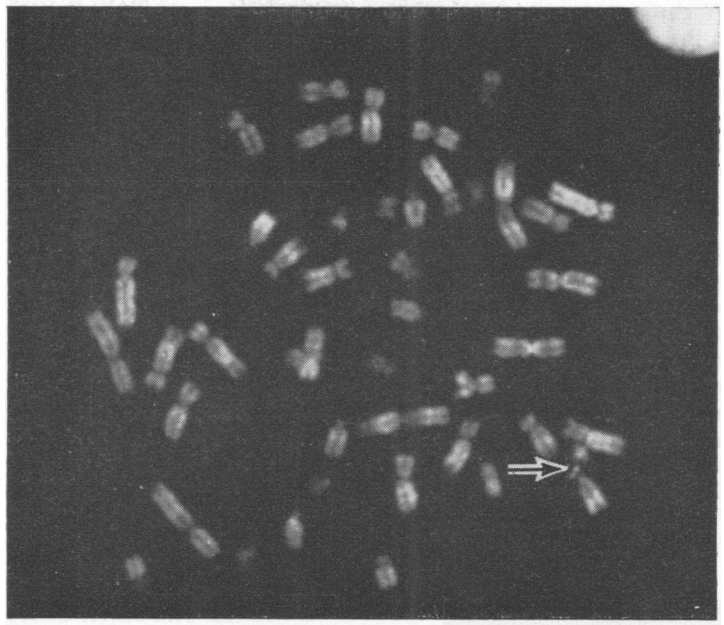


FIG. 6. A metaphase from a female showing brilliant fluorescence in the short arm of one chromosome No. 13 (arrow).

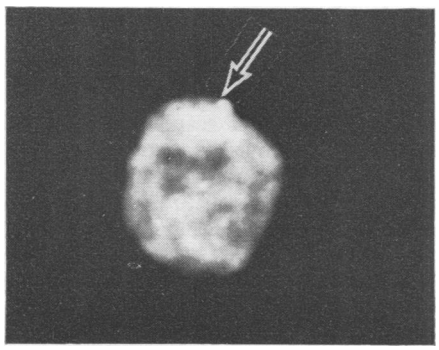


FIG. 7. An interphase cell from the female of Figure 6. Brilliantly fluorescent autosomal spot indicated by arrow.

TABLE VII

OCCURRENCE OF Y-BODY CONTAINING CELLS IN 86 MOTHERS BETWEEN 14th AND 18th WEEK OF PREGNANCY FROM 5- TO 6-DAY LYMPHOCYTE CULTURES
(Grosset *et al*, 1974)

No. of Women	Sex of Child	Y-body Containing Cells
42	Male	Present
3		Absent
9	Female	Present
32		Absent

only primiparae or women who had earlier given birth to a girl. Of 11 mothers who gave birth to a boy, nine had Y-body containing cells in their blood during pregnancy; a mean proportion of 0.09% (Table V). In one mother Y-positive cells were found as early as the fourth month of pregnancy. However, false-positive cases seem to occur. In five out of 14 pregnancies where the sex of the fetus was female, the mothers had 'Y-body' positive cells in their blood (Table V). Retrospective studies suggested reasons for these discrepancies (Table VI). The most common reason for misinterpretation was brilliant fluorescence of one of the maternal autosomes (Fig. 6), which often looked like a Y-body at interphase (Fig. 7). Grosset *et al* (1974) using an almost identical method have obtained similar results (Table VII).

Since the Y chromosome can be seen as a Y-body in about 30–50% of the interphase cells (Schröder and de la Chapelle, 1972), the values cited above for fetal lymphocytes in the maternal blood during pregnancy are in good accord with those obtained in mitotic studies (Walknowska *et al*, 1969; de Grouchy and Trébuchet, 1971). Zimmerman and Schmickel (1971) have tried to use the brilliant

Downloaded from jmg.bmj.com on March 6, 2012 - Published by group.bmj.com

fluorescence of the Y chromosome at interphase to detect male WBC in the circulation of women during pregnancy, but without success. Polani and Mutton (1971) have shown Y-positive cells in the blood of non-pregnant women. The numbers given by these authors indicate that 1–5% of the interphase WBC of normal women contain a Y-body-like structure during interphase. These values are much higher than we have found in pregnant women (Schröder and de la Chapelle, 1972), and our impression is that Y-body-like structures are very rare in interphase cells of normal women, being found only in women with an autosome that fluoresces brilliantly. The criterion for a Y-body seems to vary between different laboratories. The Y-body can be used as a criterion for a fetal WBC in the maternal circulation if very stringent criteria are adopted for its identification (Fig. 8), and if all mothers with brilliant autosomal fluorescence are excluded. This is also indicated by our later data (Schröder *et al*, 1974).

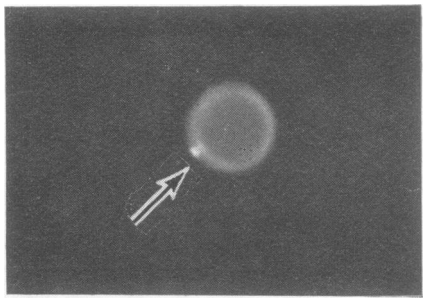


FIG. 8. An unstimulated lymphocyte from the circulation of a mother immediately after delivery of a boy. The lymphocyte contains a Y-body (arrow).

In a study where the frequency of fetal WBC in the maternal blood was evaluated after delivery, a coded experiment was performed where all mothers with brilliant autosomal fluorescence were excluded to avoid false positive results. Altogether 240 000 cells were studied from 24 mothers, the observer being unaware of the sex of the child, and only one false-positive cell was found (Table VIII).

The different fetal WBC types that occur in the blood of mothers have been evaluated by examining unstimulated leucocytes from 46 primiparae immediately after delivery of a boy (Schröder *et al*, 1974). Fetal lymphocytes and granulocytes could be found in 30% of the mothers, lymphocytes only in 37%, granulocytes only in 7%, and no fetal WBC at all in 26% (Table IX). The proportion of fetal cells varied between 0.01 and 0.2%.

TABLE VIII

PREVALENCE OF Y-BODY CONTAINING CELLS IN MATERNAL BLOOD AT DELIVERY
(Schröder *et al*, 1974)

Code No. of Mother	Sex of Child	No. of Y-body Containing Cells per 5000 Cells Examined	
		Lymphocytes	Granulocytes
1	Male	4	3
2	Male	4	1
3	Male	2	1
4	Male	0	2
5	Male	0	1
6	Male	3	0
7	Male	4	0
8	Male	3	0
9	Male	1	0
10	Male	1	0
11	Male	1	0
12	Male	1	0
13	Male	1	0
14	Female	1	0
15–19 (five cases)	Male	0	0
20–24 (five cases)	Female	0	0

10 000 leucocytes were studied from 24 mothers after delivery. The observer was unaware of the sex of the child.

TABLE IX

FETAL LEUCOCYTES IN THE BLOOD OF 46 WOMEN AFTER THE BIRTH OF A MALE CHILD
(Schröder *et al*, 1974)

Y-bodies in Interphase Cells	No. of Women (n = 46)
Lymphocytes and granulocytes	14
Lymphocytes only	17
Granulocytes only	3
No Y-bodies	12

Fetal cells capable of responding to PHA and undergoing mitotic division seem, in at least some cases, to be present in maternal blood after delivery, since in three cases studied one unequivocal XY mitosis was found (Schröder *et al*, 1974). However, this study was very preliminary and does not give a true picture of the prevalence of the phenomenon.

Twenty mothers were studied at different intervals after delivery and the results indicate that fetal granulocytes disappear from the maternal circulation within 1 week after delivery, while fetal lymphocytes can be found up to 1 year post partum (Fig. 9) (Schröder *et al*, 1974).

About 22% of all mothers produce HL-A antibodies against the paternally derived fetal HL-A antigens (Schröder *et al*, 1974). A correlation exists between the presence of HL-A antibodies in the maternal serum, and the absence of fetal cells in the maternal blood (Schröder *et al*, 1974). HL-A typings in the parents, and from the umbilical cords of the infants (Tiilikainen *et al*, 1974) have given the following results. The maternally derived fetal HL-A antigens could generally be typed without

Downloaded from jmg.bmj.com on March 6, 2012 - Published by group.bmj.com

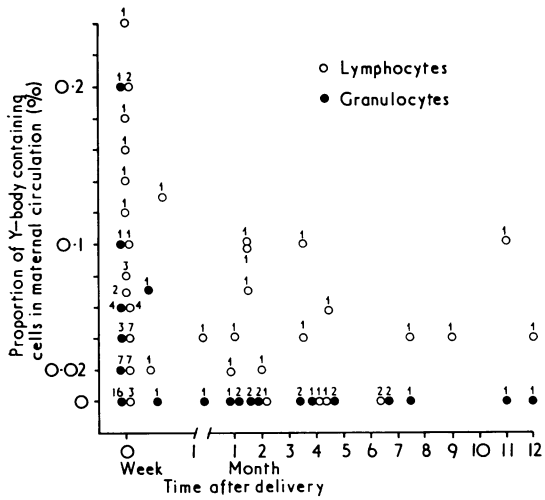


FIG. 9. Prevalence of Y-body-containing cells in the circulation of women at various intervals after birth of a male infant. The figures above the circles and squares indicate the numbers of mothers tested. (Schröder *et al* 1974)

difficulty from cord blood lymphocytes, while these cells often failed to demonstrate paternally derived HL-A antigens. These could generally be demonstrated on a later occasion after delivery, or after incubation of the cells under tissue culture conditions (Tiilikainen *et al*, 1974). It is possible that this 'masking' of the paternally derived fetal HL-A antigens is of importance in maintaining the immunological tolerance between mother and fetus.

It can be concluded that passage of leucocytes from fetus to mother is a common phenomenon, occurring in most pregnancies. The results obtained by independent methods are in good accord with each other (Walknowska *et al*, 1969; de Grouchy and Trébuchet, 1971; Schröder and de la Chapelle, 1972; Schröder *et al*, 1974). A correlation exists between the absence of fetal cells in the maternal blood, and the presence of HL-A antibodies in the maternal serum. A 'masking' of paternally derived fetal HL-A antigens may be of importance in the maintenance of tolerance between fetus and mother.

Mother to fetus. The question of whether maternal leucocytes reach the circulation of the fetus is still controvesial. Desai and Creger (1963) were able to find fluorescent cells in the circulation of six fetuses after injecting quinacrine-labelled leucocytes into the circulation of nine pregnant women immediately before delivery. The maternal cell types found were platelets, granulocytes, and lymphocytes. Lymphocytes could be demonstrated in three of the cases. Olding (1972) studied mitoses in the cord blood of 14 male infants after PHA stimulation, and found no maternal cells. XX/XY chimaerism in the lymphopoietic tissues of newborn boys has been reported by others (Kadowaki *et al*, 1965; Taylor and Polani, 1965; Turner *et al*, 1966; Benirschke and Sullivan, 1969), but generally in connection with severe immunological disease.

These data suggest that no maternal mitoses are present in PHA cultures from cord blood in normal pregnancies, their presence being an exceptional occurrence connected with developmental malformations of the fetus. As chromosome studies are very time-consuming, however, rather few cells have been examined.

By using quinacrine mustard fluorescence, and mother-and-son combinations where one or more of the mother's chromosomes had a brilliantly fluorescent region not present in the child's, Schröder (1974) studied large numbers of mitoses from PHA cultures of cord blood of nine newborn boys and one girl. In all cases the infants were delivered after normal uncomplicated deliveries, and in one case 0.07% of maternal mitoses were found, four out of 5853 mitoses (Table X).

TABLE X

CORD BLOOD MITOSES FROM 10 INFANTS AFTER PHA STIMULATION
(Schröder, 1974)

Code No.	No. of Mitoses Studied	No. of Maternal Cells
1–9	21 845	0
10	5853	4
Total	27 698	4 (0.01%)

Mitotic cells were studied after QM staining from the umbilical cords of 10 infants, to ascertain whether maternal cells had passed through the placenta. Nine of the infants were boys, and the Y chromosome and fluorescent autosomal markers were used to differentiate between fetal and maternal cells.

Maternal WBC have frequently been observed in the blood of fetuses in normal pregnancies (Desai and Creger, 1963). These cells seldom seem able to undergo mitosis in PHA cultures (Kadowaki *et al*, 1965; Taylor and Polani, 1965; Turner *et al*, 1966; Benirschke and Sullivan, 1969; Olding, 1972). If, however, a large number of cells are examined by adequate methods maternal mitoses can be found in at least some cases (Schröder, 1974).

Trophoblasts. In light of the above data it is evident that a transplacental passage of blood cells is a common, or even normal physiological event in man. What is still unanswered is why these cells so seldom cause an immune reponse. Actually, it is even more surprising that trophoblasts—fetal

Downloaded from jmg.bmj.com on March 6, 2012 - Published by group.bmj.com

cells in the placental villi which are in immediate contact with the maternal circulation—are tolerated. The explanation was advanced that trophoblastic villi are deficient in blood group antigens, and this has been verified (Thiede *et al*, 1965; Gross, 1966). After enzyme digestion, however, transplantation antigens have been possible to demonstrate in trophoblasts (Currie *et al*, 1968), so the 'absence' of antigens is probably a consequence of suppression by the mucoprotein layer coating these cells.

Trophoblasts have been found in the arteries of the uterus during pregnancy and after delivery (Beer and Billingham, 1971). In normal pregnancies there seems to be a chronic shedding of multicellular fragments or sprouts of syncytiotrophoblasts from placental villi into the maternal blood stream at a rate of 100 000 cells per day from the second month of pregnancy onward (Iklé, 1961). The majority of these cells are probably destroyed by enzymes in the blood stream, but some survivors are filtered out in the capillary bed of the lungs, where they gradually disappear.

Passage of trophoblasts in the reverse direction has also been reported. Iklé (1961) demonstrated trophoblastic cells in the umbilical veins of fetuses at various stages of pregnancy, and Salvaggio *et al* (1960) found them in cord blood after delivery.

Despite speculation the question of what significance such passage could have is still unanswered.

Mouse

Fetus to mother. No reports exist on the occurrence of fetal erythrocytes in the maternal circulation in the mouse. This is probably due to lack of suitable markers by which the phenomenon can be studied.

Whether fetal lymphocytes enter the maternal circulation of the mouse is totally unknown. However, a study has been initiated to answer this question. In matings between the laboratory mouse (*Mus musculus*), which has 40 acrocentric chromosomes, and the tobacco mouse (*Mus posciavinus*), which has 26 chromosomes, many of them are metacentric (Gropp *et al*, 1970) fetal mitoses are identifiable in lymphocyte cultures of maternal blood. The results, which are still tentative, indicate that such a passage of lymphocytes is very unusual, or does not occur at all (Schröder and Andersson, 1975). However, the method has the limitation that only fetal cells that have responded to PHA-stimulation by mitosis can be identified.

Mother to fetus. Since the methods used to study the occurrence of maternal cells in the haemopoietic tissues of the offspring are not always sensitive enough to distinguish between erythrocytes and leucocytes, I shall consider both under the same heading.

The results presented in the literature are somewhat confusing. Finegold and Michie (1961) hold the view that maternal cells can seldom be found in the blood of the offspring. They injected erythrocytes labelled with radioactive chromium into the circulation of pregnant female mice, and found labelled cells in only two out of 279 fetuses studied. Barnes and Tuffrey (1970) also used chromium labelled erythrocytes, and the fetal blood smears were studied by autoradiography. However, their figures differ considerably from those of Finegold and Michie (1961). Maternal erythrocytes were found in the majority of the cases after both normal deliveries and Caesarean sections.

Tuffrey *et al* (1969a) reported a high frequency of maternal cells in the haemopoietic tissues of 4- to 60-day-old mice that had developed in the uterus of foster mothers after transplantation at the blastocyst stage. A chromosome marker (T_6) was used to differentiate between maternal and fetal mitoses. Similar observations were later reported by the same authors (Tuffrey *et al*, 1969b) after normal pregnancies, in which they used matings between CFW and CBA T_6T_6 mice. However, Billington *et al* (1969) were unable to confirm these results, and suggested that the findings of Tuffrey *et al* (1969a; 1969b) might be due to misinterpretation of the T_6 marker.

In a later study, Barnes and Tuffrey (1970), using autoradiography, have found maternal lymphocytes in the circulation of newborn mice. However, Seller (1970), working with mice with macrocytic anaemia and using the T_6 marker, was unable to find any evidence of a traffic of blood cells from mother to fetus.

As the studies mentioned above indicate, it is still uncertain whether passage of cells occurs between mother and fetus in mice. The conflicting reports are probably due to pitfalls in the methods used.

When a small number of cells has to be recognized among a large number of other cells, the T_6 marker method and radioactive labelling of cells do not seem to be sensitive enough.

Using matings between the laboratory mouse and tobacco mouse we have also looked for maternal cells in the haemopoietic tissues of the offspring (Schröder and Andersson, 1975). So far we have found no sign of large-scale passage of blood cells from mother to fetus.

Downloaded from jmg.bmj.com on March 6, 2012 - Published by group.bmj.com

DISCUSSION

Correlation between erythrocyte and leucocyte data

Fetus to mother. The data concerning fetal erythrocytes and leucocytes in the maternal circulation during pregnancy and after delivery have been considered separately in the previous sections. All the results indicate that both RBC and WBC of the fetus are present in the maternal circulation at these times. Fetal erythrocytes are supposed to pass the placenta as a result of a bleeding (Zilliacus, 1963; Clayton *et al*, 1964; Cohen *et al*, 1964; Cohen and Zuelzer, 1967; Betke and Nierhaus, 1968; Bartsch, 1972), since they are not motile. The fact that fetal erythrocytes are found more frequently in the mother's blood when the fetus or placenta has been subject to mechanical violence also supports this theory (Wimhöfer *et al*, 1962; Bartsch, 1972). On the other hand, it is well known that WBC are capable of active movement. Hence it has been suggested that fetal lymphocytes cross the placental barriers in most, if not all, pregnancies (Walknowska *et al*, 1969; de Grouchy and Trébuchet, 1971; Schröder and de la Chapelle, 1972; Grosset *et al*, 1974).

Light might be shed on these questions by counting both white and red fetal blood cells in the circulation in a large series of mothers at different stages of pregnancy. Such a study is under way at the Folkhälsan Institute of Genetics, Helsinki. Until such counts are available, however, we may try to relate the erythrocyte and lymphocyte findings from data in the literature.

Time of occurrence, frequency, and proportion. Fetal erythrocytes have been demonstrated in the maternal circulation even during the first trimester (Clayton *et al*, 1964; Cohen *et al*, 1964; Betke and Nierhaus, 1968; Bartsch, 1972). However, they have generally been found in less than 1 fetal cell per 50 000 maternal cells, and then only in about 10% of women. Fetal lymphocytes have been found in the mother's blood during the first trimester in almost all cases studied at this stage (Tables III, VII, Walknowska *et al*, 1969; de Grouchy and Trébuchet, 1971; Grosset *et al*, 1974). The ratio of fetal to maternal cells is about 1/500. During the second and third trimesters fetal erythrocytes are found in 30–40% of mothers, but still in a low proportion, mostly less than 1 fetal cell/50 000 maternal cells. The figures for fetal lymphocytes in the maternal circulation during the second and third trimesters are of magnitude 1 fetal cell/1000 maternal cells or higher, and fetal cells are found in more than 80% of cases (Walknowska *et al*, 1969; de Grouchy and Trébuchet, 1971; Schröder and de la Chapelle, 1972).

If, in contrast to what is indicated by experimental data so far fetal erythrocytes and lymphocytes are assumed to pass into the maternal circulation in normal relative proportions, two theoretical explanations would be possible; (1) fetal lymphocytes are tolerated better than fetal erythrocytes in the mother's blood or (2) the frequency of fetal mitoses in cultures of the mother's blood does not reflect the frequency of fetal WBC in the maternal circulation.

(*1*) Fetal erythrocytes occur twice as frequently in the maternal blood in blood-group-compatible pregnancies than in blood-group-incompatible ones (see above). In blood-group-incompatible pregnancies fetal erythrocytes seem to be eliminated rapidly, but in homospecific pregnancies the survival values of the fetal red cells in the mother's blood are in good accord with the normal life span of the cells. In about 95% of all primigravidae the HL-A antigens of the fetus are different from those of the mother, and in 60% of these cases at least two different HL-A antigens are involved (Tiilikainen *et al*, 1974). Since at least 70% of the mothers have fetal leucocytes in their blood at the same time, it is surprising that only 22% of them produce HL-A antibodies (Schröder *et al*, 1974). On the other hand, a negative correlation exists between the occurrence of HL-A antibodies directed against fetal HL-A antigens in the mother's serum, and the proportion of fetal leucocytes in the maternal blood (Schröder *et al*, 1974). This correlation can be shown immediately after delivery, but is even stronger later.

Fetal lymphocytes are probably better tolerated in the maternal blood than fetal erythrocytes. This may to some extent be due to a masking of paternally derived fetal HL-A antigens during pregnancy (Tiilikainen *et al*, 1974).

A greater tolerance is not in itself enough to explain the great difference in the relative proportion of fetal erythrocytes and lymphocytes in the maternal circulation during pregnancy and after delivery.

(*2*) Walknowska *et al* (1969) suggested that the high rates of XY mitoses in maternal blood cultures did not reflect the proportion of fetal WBC *in vivo*. Preferential stimulation of fetal cells by PHA in cultures of maternal blood would exaggerate the proportion of fetal mitoses. This would also explain the large numbers of fetal mitoses reported by de Grouchy and Trébuchet (1971), and the high proportion of Y-body containing cells in the mother's blood in 5- to 6-day PHA cultures found by Grosset *et al* (1974). On the other hand, when testing the theory of preferential stimulation of fetal

Downloaded from jmg.bmj.com on March 6, 2012 - Published by group.bmj.com

cells, Schröder and de la Chapelle (1972) and Schröder *et al* (1974) have found cultured and uncultured interphase cells in roughly similar numbers.

Validity of mitotic findings. Walknowska *et al* (1969) and de Grouchy and Trébuchet (1971) used conventional methods of chromosome staining. The use of such methods for this purpose has been criticized (Jacobs and Smith, 1969). Although methodological difficulties may possibly lead to misinterpretations all data indicate that fetal leucocytes do occur in the maternal circulation during pregnancy, in much higher relative proportions than fetal erythrocytes.

Mother to fetus. Very different values have been given for the occurrence of maternal erythrocytes in the blood of their newborn infants (Naeslund and Nylin, 1946; Macris *et al*, 1958; Cohen and Zuelzer, 1965). However, reports of high proportions of maternal erythrocytes have been based on counts of cells labelled with radioactive isotopes (Naeslund and Nylin, 1946; Naeslund, 1950/1951; Mengert *et al*, 1955; Zarou *et al*, 1964). These studies are open to criticism because free isotope molecules may pass through the placenta and be taken up by fetal cells. If all but the studies based on agglutination tests, immunofluorescence, or the use of marker cells (sickle cells, elliptiocytes) are disregarded, maternal cells have been found in 10–80% of all newborns (Hedensted and Naeslund, 1946; Mengert *et al*, 1955; Macris *et al*, 1958; Cohen and Zuelzer, 1965; Fischer and Müller, 1967; von Muralt, 1967; Eimer and Weiland, 1969). Even in normal pregnancies maternal leucocytes occur in the blood of newborns (Desai and Creger, 1963; Schröder, 1974), but they have generally been reported in connection with malformations of the fetus and placenta (Kadowaki *et al*, 1965; Taylor and Polani, 1965; Turner *et al*, 1966; Benirschke and Sullivan, 1969). This might be due to the fact that most teams have studied only cells from PHA cultures. Since only a proportion of the leucocytes (T lymphocytes) respond to PHA by entering mitosis, these results give a very limited picture of the situation *in vivo*. With Atebrin-labelled leucocytes Desai and Creger (1963) were able to show passage of maternal cells through the placenta in the majority of their cases. A further factor limiting chromosome studies is their time-consuming nature, for which reason only small numbers of cells have been counted (Turner *et al*, 1966; Benirschke and Sullivan, 1969; Olding, 1972).

If a correlation is sought between the data for erythrocytes and lymphocytes the facts mentioned above have to be taken into consideration. In the light of these one cannot exclude the possibility that lymphocytes and erythrocytes pass from mother to fetus to the same extent, and in the same normal relative proportions.

Biological importance

What is the biological importance of this passage of blood cells between mother and fetus? Is it a pathological phenomenon followed by antibody formation and immunological disease, or is it of importance in the immunological tolerance between mother and fetus? Or is the occurrence of these cells in the mother's or infant's blood merely a secondary consequence of immunological tolerance? These questions cannot be fully answered at the moment, but if the phenomenon were of basic importance for the tolerance between mother and fetus, one would expect it to be wide-spread among other mammals.

The passage of blood cells has been studied in detail in the mouse (Finegold and Michie, 1961; Billington *et al*, 1969; Tuffrey *et al*, 1969a; 1969b; Barnes and Tuffrey, 1970; Seller, 1970), but the question is so far unanswered. However, these studies involved the use of cytological markers. In the rat Beer *et al* (1971; 1972) and Beer and Billingham (1971, 1973) have been able to induce tolerance or runt disease in the offspring by giving living cellular homografts to females 1–2 weeks before mating. This they have taken to reflect a transmission of cells from mother to fetus. Here, however, the constellation has been a pathologic one, and caution is called for in drawing parallels with normal pregnancies.

In conclusion, the passage of blood cells between mother and fetus is understandable in many pathological conditions, but the importance of the phenomenon in normal pregnancies is not yet clear.

Practical implications

Since the recognition that the passage of erythrocytes from Rh^+ fetuses to Rh^- mothers is of importance in Rh immunization, programmes have been developed in several countries to avoid Rh toxaemia (Clarke, 1967; Bartsch, 1972; Eklund and Nevanlinna, 1973).

How can the occurrence of fetal lymphocytes in the maternal circulation be turned to useful purposes by the clinician? A very tempting thought would be to use fetal cells in the maternal blood for antenatal diagnosis instead of amniocentesis. So far the method has not proved sensitive enough for the purpose (Schröder and de la Chapelle, 1972).

Downloaded from jmg.bmj.com on March 6, 2012 - Published by group.bmj.com

The main difficulty lies in the low frequency of fetal lymphocytes in the mother's blood and the persistence of fetal lymphocytes from earlier pregnancies. A possible method to enrich fetal cells in blood samples from the mother would perhaps be to destroy the maternal cells by HL-A antibodies directed against maternal, but not fetal HL-A antigens (Tiilikainen *et al*, 1970). Another method which seems possible on theoretical grounds at least, is to use a cell separator of the type described by Hulett *et al* (1973). With this device different lymphocyte populations have been separated successfully from each other.

Although some data already exist on the passage of blood cells between mother and fetus, many of the most important questions are still unanswered. From a combination of cytological and immunological methods, it seems realistic to expect the answer to most of these questions in the near future.

The writing of this article was financed by the Finnish National Research Council for Natural Science. I wish to thank Professor Albert de la Chapelle for reading the manuscript, and for continuous support during the work.

References

Barnes, R. D. and Tuffrey, M. (1970). Maternal cells in the newborn. *Advances in the Biosciences*, **6,** 457–473.

Bartsch, F. K. (1972). Fetale Erythrozyten im mütterlichen Blut und Immunoprophylaxe der Rh-Immunisierung. *Acta Obstetricia et Gynecologica Scandinavica*, Supplement 20.

Beer, A. E. and Billingham, R. E. (1971). Immunobiology of mammalian reproduction. *Advances in Immunology*, **14,** 1–83.

Beer, A. E. and Billingham, R. E. (1973). Maternally acquired runt disease. *Science*, **179,** 240–243.

Beer, A. E. Billingham, R. E., and Hoerr, R. A. (1971). Elicitation and expression of transplantation immunity in the uterus. *Transplantation Proceedings*, **3,** 609–611.

Beer, A. E., Billingham, R. E., and Yang, S. L. (1972). Maternally induced transplantation immunity, tolerance, and runt disease in rats. *Journal of Experimental Medicine*, **135,** 808–826.

Benirschke, K. and Sullivan, M. M. (1969). The human placenta in relationship to the development of chimerism. In *The Foetoplacental Unit*, p. 49, ed. by A. Pecile and C. Finzi. Excerpta Medica Foundation, Amsterdam.

Betke, K. and Nierhaus, K. (1968). Fetales Haemoglobin und fetale Erythrozyten im mütterlichen Blut. *Münchener Medizinische Wochenschrift*, **110,** 509–516.

Billington, W. D., Kirby, D. R. S., Owen, J. J. T., Ritter, M. A., Burtonshaw, M. D., Evans, E. P., Ford, C. E., Gauld, I. K., and McLaren, A. (1969). Placental barrier to maternal cells. *Nature*, **224,** 704–706.

Caspersson, T., Gahrton, G., Lindsten, J., and Zech, L. (1970). Identification of the Philadelphia chromosome as a number 22 by quinacrine mustard fluorescence analysis. *Experimental Cell Research*, **63,** 238–240.

Clarke, C. A. (1967). Prevention of Rh-haemolytic disease. *British Medical Journal*, **4,** 7–12.

Clayton, E. M., Feldhaus, W. D., and Whitacre, F. E. (1964). Fetal erythrocytes in the maternal circulation of pregnant women. *Obstetrics and Gynecology*, **23,** 915–919.

Cohen, F. and Zuelzer, W. W. (1965). The transplacental passage of maternal erythrocytes into the fetus. *American Journal of Obstetrics and Gynecology*, **93,** 566–569.

Cohen, F. and Zuelzer, W. W. (1967). Mechanisms of Isoimmunization. II. Transplacental passage and postnatal survival of fetal erythrocytes in heterospecific pregnancies. *Blood*, **30,** 796–804.

Cohen, F., Zuelzer, W. W., Gustafson, D. C., and Evans, M. M. (1964). Mechanisims of Isoimmunization. I. The transplacental passage of fetal erythrocytes in homospecific pregnancies. *Blood*, **23,** 621–646.

Currie, G. A., Doorninck, W. van, and Bagshawe, K. D. (1968). Effect of neuraminidase on the immunogenicity of early mouse trophoblast. *Nature*, **219,** 191–192.

De Grouchy, J. and Trébuchet, C. (1971). Transfusion foeto-maternelle de lymphocytes sanguins et détection du sexe du foetus. *Annales de Génétique*, **14,** 133–137.

Desai, R. G. and Creger, W. P. (1963). Maternofetal passage of leukocytes and platelets in man. *Blood*, **21,** 665–673.

Eimer, H. and Weiland, A. (1969). Untersuchung über die Placentapassage mütterlicher Erythrocyten. *Archiv für Gynäkologie*, **208,** 113–122.

Eklund, J. and Nevanlinna, H. R. (1973). Rh prevention: A report and analysis of a national programme. *Journal of Medical Genetics*, **10,** 1–7.

Finegold, M. and Michie, D. (1961). Experiments on the maternal-foetal barrier in the mouse. *Journal of Embryology and Experimental Morphology*, **9,** 618–622.

Fischer, K. and Müller, V. (1967). Zum Nachweis mütterlicher Erythrocyten im Neugeborenenblut (Immunfluorescenztechnik). *Monatsschrift für Kinderheilkunde*, **115,** 147–148.

Forsius, H., Kaski, U., Schröder, J., and Chapelle, A. de la (1972) Is there a common psychopathology of XYY boys? A clinical report on three cases of XYY and one of XY/XYY. *Acta Paedopsychiatrica*, **39,** 28–41.

Gropp, A., Tettenborn, U., and Lehman, E. von (1970). Chromosomenvariation vom Robertson'schen Typus bei der Tabakmaus, *M. poschiavinus*, und ihren Hybriden mit der Laboratoriummaus. *Cytogenetics*, **9,** 9–23.

Gross, S. J. (1966). Human blood group A substance in human endometrium and trophoblast localized by chromatographed rabbit antiserum. *American Journal of Obstetrics and Gynecology*, **95,** 1149–1159.

Grosset, L., Barrelet, V., and Odartchenko, N. (1974). Antenatal foetal sex determination in maternal blood during early pregnancy. *American Journal of Obstetrics and Gynecology*, **120,** 60–63.

Hedenstedt, S. and Naeslund, J. (1946). Investigations of the permeability of the placenta with the help of elliptocytes. *Acta Medica Scandinavica*, Suppl. 170, 126–134.

Hulett, H. R., Bonner, W. A., Sweet, R. G., and Herzenberg, L. A. (1973). Development and application of a rapid cell sorter. *Clinical Chemistry*, **19,** 813–816.

Iklé, A. (1961). Trophoblastzellen im strömenden Blut. *Schweizerische Medizinische Wochenschrift*, **91,** 943–945.

Jacobs, P. A. and Smith, P. G. (1969). Practical and theoretical implications of fetal/maternal lymphocyte transfer. *Lancet*, **2,** 745.

Kadowaki, J., Thompson, R. I., Zuelzer, W. W., Wooley, P. V., Brough, A. J., and Gruber, D. (1965). XX/XY lymphoid chimaerism in congenital immunological deficiency syndrome with thymic alymphoplasia. *Lancet*, **2,** 1152–1156.

Kleihauer, E., Braun, H., and Betke, K. (1957). Demonstration von fetalem Hämoglobin in den Erythrocyten eines Blutausstriches. *Klinische Wochenschrift*, **35,** 637–638.

Macris, N. T., Hellman, L. M., and Watson, R. J. (1958). The transmission of transfused sickle-trait cells from mother to fetus. *American Journal of Obstetrics and Gynecology*, **76,** 1214–1218.

Mengert, W. F., Rights, C. S., Bates, C. R., Reid, A. F., Wolf, G. R. and Nabors, G. C. (1955). Placental transmission of erythrocytes. *American Journal of Obstetrics and Gynecology*, **69,** 678–683.

Muralt, G. von (1967). Materno-fetale Blutung und 'Respiratory Distress Syndrom'. *Pädiatrische Fortbildungskurse für die Praxis*, **21,** 90–102.

Naeslund, J. (1946). Undersökningar över placentas permeabilitet med tillhjälp av bl. a. blodgruppsbestämningar, radioaktivt märkta blodkroppar och elliptocyter. *Nordisk Medicin*, **29,** 589–592.

Naeslund, J. (1950/1951). Studies on placental permeability with radioactive isotopes of phosphorus and iron. *Acta Obstetrica et Gynecologica Scandinavica*, **30,** 231–246.

Downloaded from jmg.bmj.com on March 6, 2012 - Published by group.bmj.com

Naeslund, J. and Nylin, G. (1946). Investigations on the permeability of the placenta with the aid of red blood corpuscles tagged with radio-active phosphorous. *Acta Medica Scandinavica*, Suppl. 170, pp. 390–398.

Olding, L. (1972). The possibility of materno-foetal transfer of lymphocytes in man. *Acta Paediatrica Scandinavica*, **61**, 73–75.

Pearson, P. L., Bobrow, M., and Vosa, C. G. (1970). Technique for identifying Y chromosomes in human interphase nuclei. *Nature*, **226**, 78–79.

Polani, P. and Mutton, D. E. (1971). Y-fluorescence of interphase nuclei, especially circulating lymphocytes. *British Medical Journal*, **1**, 138–142.

Pollack, W., Ascari, W. Q., Crispen, J. F., O'Connor, R. R., and Ho, T. Y. (1971a). Studies on Rh prophylaxis. II. Rh immune prophylaxis after transfusion with Rh-positive blood. *Transfusion*, **11**, 340–344.

Pollack, W., Ascari, W. Q., Kochesky, R. J., O'Connor, R. R., Ho, T. Y., and Tripodi, D. (1971b). Studies on Rh prophylaxis. I. Relationship between doses of anti-Rh and size of antigenic stimulus. *Transfusion*, **11**, 333–339.

Salvaggio, A. T., Nigogosyan, G., and Mack, H. C. (1960). Detection of trophoblast in cord blood and fetal circulation. *American Journal of Obstetrics and Gynecology*, **80**, 1013–1021.

Schröder, J. (1974). Passage of leukocytes from mother to fetus. *Scandinavian Journal of Immunology*, **3**, 369–373.

Schröder, J. and Andersson, L. C. (1975). Lack of permeability of mouse placenta to maternal and foetal cells. *Nature*, **253**, 453–455.

Schröder, J. and Chapelle, A. de la (1972). Fetal lymphocytes in the maternal blood. *Blood*, **39**, 153–162.

Schröder, J., Tiilikainen, A., and Chapelle, A. de la (1974). Fetal leukocytes in the maternal circulation after delivery. *Transplantation*, **17**, 346–354.

Seller, M. J. (1970). Lack of porosity of the mouse placenta to maternal cells. *Nature*, **225**, 1254–1255.

Taylor, A. I. and Polani, P. E. (1965). XX/XY mosaicism in man. *Lancet*, **1**, 1226.

Thiede, H. A., Choate, J. W., Gardner, H. H., and Santay, H. (1965). Immunofluorescent examination of the human chorionic villus for blood group A and B substance. *Journal of Experimental Medicine*, **121**, 1039–1050.

Tiilikainen, A., Kaakinen, A., and Amos, D. B. (1970). The separation of genetically different lymphocyte populations in vitro. *Transplantation*, **10**, 361–365.

Tiilikainen, A., Schröder, J., and Chapelle, A. de la (1974). Fetal leukocytes in the maternal circulation after delivery. *Transplantation*, **17**, 355–360.

Tuffrey, M., Bishun, N. P., and Barnes, R. D. (1969a). Porosity of the mouse placenta to maternal cells. *Nature*, **221**, 1029–1031.

Tuffrey, M., Bishun, N. P., and Barnes, R. D. (1969b). Porosity of the placenta to maternal cells in normally derived mice. *Nature*, **224**, 701–704.

Turner, J. H., Wald, N., and Quinlivan, W. L. G. (1966). Cytogenetic evidence concerning possible transplacental transfer of leukocytes in pregnant women. *American Journal of Obstetrics and Gynecology*, **95**, 831–833.

Walknowska, J., Conte, F. A., and Grumbach, M. M. (1969). Practical and theoretical implication of fetal/maternal lymphocyte transfer. *Lancet*, **1**, 1119–1122.

Wimhöfer, H., Schneider, J., and Leidenberger, F. (1962). Untersuchungen über die Einschwemmung fetaler Erythrozyten in den mütterlichen Kreislauf bei Spontangeburten und geburt shilflichen Eingriffen. *Geburtshilfe und Frauenheilkunde*, **22**, 589–597.

Woodrow, J. C. and Finn, R. (1966). Transplacental haemorrage. *British Journal of Haematology*, **12**, 297–309.

Zarou, D. M., Lichtman, H. C., and Hellman, L. M. (1964). The transmission of chromium-51 tagged maternal erythrocytes from mother to fetus. *American Journal of Obstetrics and Gynecology*, **88**, 565–571.

Zilliacus, H. (1963). Agglutinated incompatible fetal erythrocytes in the maternal circulation. *American Journal of Obstetrics and Gynecology*, **86**, 1093–1095.

Zimmerman, A. and Schmickel, R. (1971). Fluorescent bodies in maternal circulation. *Lancet*, **1**, 1305.

Zipursky, A., Hull, A., White, F. D., and Israels, L. G. (1959). Foetal erythrocytes in the maternal circulation. *Lancet*, **1**, 451–452.

Downloaded from jmg.bmj.com on March 6, 2012 - Published by group.bmj.com

Transplacental passage of blood cells.

J Schröder

J Med Genet 1975 12: 230-242
doi: 10.1136/jmg.12.3.230

Updated information and services can be found at:
http://jmg.bmj.com/content/12/3/230.citation

These include:

References Article cited in:
http://jmg.bmj.com/content/12/3/230.citation#related-urls

Email alerting service Receive free email alerts when new articles cite this article. Sign up in the box at the top right corner of the online article.

Notes

To request permissions go to:
http://group.bmj.com/group/rights-licensing/permissions

To order reprints go to:
http://journals.bmj.com/cgi/reprintform

To subscribe to BMJ go to:
http://group.bmj.com/subscribe/

This material may be protected by Copyright law (Title 17 U.S. Code)

Determination of Circulating DNA Levels in Patients with Benign or Malignant Gastrointestinal Disease

BERNARD SHAPIRO, MD, MILANKUMAR CHAKRABARTY, MD,*
EDWIN M. COHN, MD,* AND SHALOM A. LEON, PHD

Previous studies showed that patients with neoplasms of various types and origins have abnormally high concentration of DNA in their serum. The current work compares circulating DNA levels in patients with benign or malignant disease of the gastrointestinal tract and determines the diagnostic value of such measurements. DNA was quantitated by radioimmunoassay capable of detecting 25 ng/ml, and as a simple and noninvasive test, it could be a useful addition to other diagnostic procedures. The GI tract was chosen because it affords a comparison of benign, precancerous, and malignant lesions of the same organ. Of the 386 patients studied prospectively, 48% had benign disease and mean DNA levels (±SE) of 118 ± 14 ng/ml, whereas 52% had malignant disease and 412 ± 63 ng DNA/ml. The difference was statistically significant ($P < 0.001$). The DNA assay showed the highest sensitivity for pancreas carcinoma: 90% of the patients had DNA levels above 100 ng/ml, chosen as the upper normal limit. Simultaneous measurements of both DNA and carcinoembryonic antigen (CEA) resulted in increased sensitivity and specificity, even when either marker alone had low sensitivity (gastric carcinoma). The results indicate that serum DNA concentration is markedly elevated in malignancy, and moderately elevated in benign disease, as compared with normal controls. These findings may have diagnostic and prognostic value.

Cancer **51:2116–2120, 1983.**

EARLIER STUDIES in our laboratory and elsewhere have shown that the concentration of free soluble DNA in the circulation of normal healthy subjects is below 25 ng/ml in 93% of the individuals tested.[1] On the other hand, in certain pathologic conditions such as systemic lupus erythematosus, rheumatoid arthritis, pulmonary embolism, and malignancy, abnormally high serum DNA levels have been found.[1–5] Cancer patients have the highest levels, especially in the presence of metastatic disease as compared to nonmetastatic lesions.[1] Both the source of this DNA and the mechanism for its release are unknown currently. It has been postulated that cell and tissue destruction associated with acute or chronic inflammation, as well as proliferative disorders such as cancer, may be responsible for this phenomenon.[2] Based on this rationale, we attempted to measure cell destruction and possible DNA release induced by ionizing radiation in cancer patients undergoing radiotherapy. Surprisingly, in many cases, DNA levels declined after radiotherapy especially when the treatment was beneficial.[1] However, there was no correlation between DNA levels and other clinical features such as tissue of origin, size of tumor, or stage of disease. Our previous study included a large variety of tumors and some patients were treated also with other modalities, such as surgery or chemotherapy.[1] It is possible that this large number of variables is responsible for the lack of correlation mentioned above. Consequently, we initiated this prospective investigation, focusing on both malignant and benign disorders of the GI tract. This system was chosen because a spectrum of benign, precancerous, and malignant lesions is known to exist.[6] The purpose of this study was twofold: (1) to compare DNA levels on the basis of organ and tissue of origin; and (2) to assess the diagnostic potential of simultaneous determinations of circulating DNA and other markers associated with tumor proliferation such as carcinoembryonic antigen (CEA) and α-fetoprotein (AFP).[7–11]

Presented in part at the 28th Annual Meeting, Society of Nuclear Medicine, Las Vegas, Nevada, June 16–19, 1981.

From the Division of Nuclear Medicine, Department of Radiology, and the Gastroenterology Section,* Department of Medicine, Albert Einstein Medical Center, Philadelphia, Pennsylvania.

Supported by a grant (PDT-163) from the American Cancer Society.

Address for reprints: S. A. Leon, PhD, Albert Einstein Medical Center, Philadelphia, PA 19141.

The authors thank R. Adler, P. Servi, and V. Peterson for excellent technical assistance and B. Graubard for preparation of the manuscript. The cooperation of Hoffman-LaRoche, Inc. in testing normal individuals is gratefully acknowledged.

Accepted for publication March 2, 1982.

0008–543X/83/0601/2116 $1.05 © American Cancer Society

Exhibit AD

Materials and Methods

Patients

In this investigation we studied 398 consecutive patients admitted to our institution. Their symptoms were one or more of the following: abdominal pain, sustained weight loss, jaundice, and rectal bleeding. After physical examination, while routine laboratory tests and other diagnostic procedures were carried out (cholecystography, barium gastrointestinal roentgenography, computerized tomography scanning, ultrasonography, biopsy, or laparotomy), serum DNA levels were also determined. DNA tests were performed without knowledge of the results of other studies. Subsequently, the data were compiled, analyzed, and final diagnosis was established for 386 patients. In the remaining 12 cases, final diagnosis could not be established for various reasons such as ambiguous test results, and refusal or denial of surgery; therefore, they were excluded from this report. In the malignant disease group only patients with positive tissue diagnosis were included in the study.

Blood Samples

The sera from blood samples obtained by venipuncture for admission tests were stored frozen at −20°C until used. Immediately before the DNA assay, the sera were inactivated at 56°C for 30 minutes in order to eliminate DNA binding by complement component C1q.[12]

Assay for DNA

The radioimmunoassay for DNA has been described in detail elsewhere.[13] Briefly, the assay is based on the competition of DNA in the test sample with ^{125}I-iododeoxyuridine-labeled DNA for the anti-DNA antibody obtained from the serum of an SLE (Systemic Lupus Erythematosus) patient.[14] A standard calibration curve is constructed, using normal human serum containing pure unlabeled DNA, in the range of 0–1000 ng/ml. The ^{125}I-radioactivity in both control and unknown samples is quantitated in a gamma-scintillation counter with >85% efficiency for ^{125}I. The sensitivity of the assay is 25 ng DNA/ml (the lowest concentration distinguishable from zero), with a reproducibility of 8–10% in the range of 100–1000 ng/ml and 18% for 25–100 ng/ml.[13,15]

Assays for other Tumor Markers

Commercially available reagents and kits were used: carcinoembryonic antigen (CEA) (Hoffmann-LaRoche), and α-fetoprotein (AFP) (Princeton Laboratories) according to the instructions supplied by the manufacturers.

TABLE 1. DNA Levels in 386 Patients with Gastrointestinal Disease

	Malignant	Benign	Normal controls
No. of patients	199	187	88
Mean ± SE (ng/ml)	412 ± 63	118 ± 14	14 ± 3
Range (ng/ml)	0–10,000	0–1,150	0–100
	Malignant *vs.* benign		Benign *vs.* normal
P value	<0.001		<0.001

Statistical Analysis

To determine the significance of differences in mean levels of the given tumor markers, we used the nonparametric Kolmogorov-Smirnov test.[16,17] The significance of variation in sensitivity and specificity was tested by conventional chi-square tests with Yates' correction.[17,18]

Results

The results of DNA assays in the current series of 386 patients with gastrointestinal disorders are summarized in Table 1. Clearly, most normal controls showed low levels of DNA in the circulation. Taken as a group, patients with benign GI disease had significantly higher DNA levels than the controls ($P < 0.001$). The highest levels of DNA were found in patients with malignant GI disease ($P < 0.001$, relative to the benign group), although some patients showed normal DNA values.

The distribution of DNA values according to some disease sites is shown on Figure 1. A comparison of patients with malignant or benign lesions of the colorectum showed that the former had significantly higher DNA levels ($P < 0.05$). On the other hand, when we compared 17 patients with stomach carcinoma (mean ± SE, 376 ± 105 ng DNA/ml) and 21 patients with benign gastric ulcer (197 ± 65 ng/ml), the difference was not significant, $P > 0.05$. Most satisfactory results were obtained for carcinoma of the pancreas (Fig. 1): DNA levels were among the highest in individual patients, also reflected in the mean for this group. Setting normal limits at 0–99 ng/ml, the difference between pancreas carcinoma and pancreatitis was statistically significant, $P < 0.05$. The only benign lesion which induced abnormally high DNA levels in *all* patients was pancreatic pseudocyst. The values had a range and mean not significantly different from carcinoma of the pancreas, $P > 0.05$ (Table 2). The reason for this finding is unknown currently.

Recently, we conducted a blind study through the auspices of the Mayo Clinic-National Cancer Institute Serum Diagnostic Bank. A total of 61 patients (30 malignant and 31 benign) with GI tract disease were tested

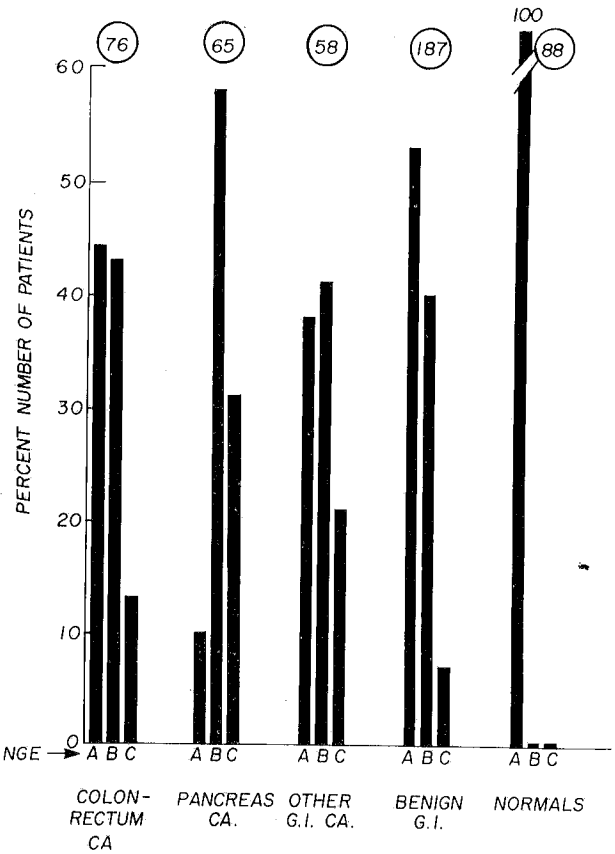


FIG. 1. Range distribution of DNA levels (in ng/ml). A, 0–99; B, 100–500; C, over 500. The circled numbers indicate the number of patients in each group. The mean ± SE and *P* values relative to benign disease are: colon–rectum, 270 ± 62, $P < 0.05$; pancreas, 646 ± 171, $P < 0.001$; other, 329 ± 66, $P < 0.05$. For details see: benign group, Table 2; other malignancies consisted of 17 patients with stomach carcinoma, two stomach lymphosarcoma, two lymphosarcoma and liver metastasis, one reticulum cell sarcoma of small bowel, seven gallbladder carcinoma, two duodenum, four esophagus, eight liver, one ileum–cecum junction, one carcinoid of ileum, two Hodgkin's lymphoma, two non-Hodgkin's lymphoma, three ampulla of Vater, six unknown primary with metastases to liver or abdominal lymph nodes.

for serum DNA levels. The results were analyzed statistically (Methods), with upper normal limits set at 25 ng/ml. The test discriminated between the two groups, with a sensitivity of 80% and specificity of 72% ($P < 0.05$). In our series of 386 patients, the best results were obtained with a normal limit set at 100 ng/ml, probably due to the large variety of pathologic conditions included in the study.

In order to assess the diagnostic potential of simultaneous determinations of markers associated with neoplastic disease, we compared the serum levels of DNA and CEA in 77 patients (63 with cancer, 14 with benign disease) and 41 normal controls (Table 3). Both assays showed significantly higher levels in patients with malignancy than in patients with benign lesions ($P < 0.05$) or the normal controls ($P < 0.001$ for DNA and $P < 0.01$ for CEA). However, we found no correlation between the levels of the two substances in any of the groups (Table 3).

A major advantage in measuring both DNA and CEA became apparent when the sensitivity and specificity were calculated for individual disease sites (Table 4). In most cases, the sensitivity increased quite dramatically. For example, in carcinoma of the colon–rectum the value of 55% for DNA and 69% for CEA increased to 90% with both markers. A most encouraging observation was that for all cancer patients the sensitivity increased from 63% for DNA and 71% for CEA to 88% when both assays were performed. The specificity declined only slightly (79% to 72%) relative to benign disease. These results are significant, with *P* values ranging from <0.01 to 0.001 (Table 4).

Discussion

Several possibilities have been proposed concerning soluble DNA in the circulation. DNA can be released spontaneously from normal lymphocytes *in vitro,*[19] as well as from tissues or organs damaged in disorders such as SLE, rheumatoid arthritis, pulmonary embolism, and cancer.[2,3] Certain therapeutic procedures may induce high circulating DNA levels, such as high-dose cortcosteroid treatment, surgery,[2,20] or radiation (Leon and Shapiro, in preparation). Thus, cell and tissue damage resulting from excessive proliferative processes (lymphoid hyperplasia, tumor spread), blood vessel obstruction, or sustained immunologic stress may result in abnormal serum DNA levels.[3] In contrast, in the normal healthy subject, DNA is either released at lower rates or is rapidly degraded and eliminated from the circulation, as shown in experimental animals.[21] The degradation

TABLE 2. DNA Levels in Benign Gastrointestinal Disease

Disease	No. of patients	Mean ± SE (ng/ml)	Range (ng/ml)
Hepatitis*	50	166 ± 34	0–750
Cirrhosis	31	120 ± 24	0–530
Chronic or acute gastritis	8	106 ± 39	25–270
Esophagitis	5	74 ± 19	55–130
Gastric or duodenal ulcer	21	197 ± 65	0–1000
Ulcerative colitis	17	53 ± 14	0–185
Chronic or acute pancreatitis	17	142 ± 35	0–500
Pancreatic pseudocyst	8	415 ± 121	185–1000
Cholecystitis, cholelithiasis	17	320 ± 91	30–1100
Other†	13	231 ± 191	95–450
Totals	187	118 ± 14	0–1100

* Thirty-five patients with infectious hepatitis (A, B, or non-A non-B) and 15 with chronic active hepatitis (alcohol or drug abuse).

† Six patients with Crohn's disease, and four with colon polyp, one each with polyp of splenic flexure, pneumatosis cystoides intestinalis, and aneurysm of iliac artery.

of DNA is probably mediated by the normally present DNAse I.[22] In certain pathologic conditions, this enzyme may be inhibited resulting in the accumulation of DNA in the circulation. In this connection, we recently demonstrated the presence of a powerful inhibitor of DNAse in the circulation of cancer patients.[15]

Consistent with such a mechanism, in this investigation we found quantitative differences in DNA levels in the circulation of normals and patients with benign or malignant disease of the GI tract. The highest levels were observed in cancer patients, especially in carcinoma of the pancreas. Patients with benign lesions of the same organ usually showed lower DNA levels. An obvious exception was pancreatic pseudocyst which induced abnormally high levels in all patients. In these cases, other diagnostic procedures such as CAT or ultrasound scanning are needed to discriminate between these possibilities. Considering that the diagnosis of pancreas carcinoma is quite difficult,[18] the measurement of DNA levels may be useful. Since the DNA assay is a simple and inexpensive laboratory test, we feel that it may be a valuable tool for this purpose.

As a tumor marker, DNA compared favorably with CEA and showed higher sensitivity for some malignancies. The determination of both substances resulted in higher sensitivity than either marker alone, and the specificity was only slightly decreased. Thus, the two tests seemed to complement each other, presumably because the DNA assay may yield normal values in cancer, whereas abnormally high levels of CEA have been found in smokers[23] and in benign conditions.[24,25]

In light of these findings, the total lack of correlation between DNA and CEA levels in the circulation of patients tested for both, suggests that they reflect different aspects of the neoplastic process. It is possible that the levels of the two markers do not increase or decrease *simultaneously.* A similar lack of temporal correlation was observed recently in sequential determinations of DNA and another substance associated with malignancy, the DNA-binding protein C3DP.[15]

In many previous studies, the quantitation of DNA has been controversial due to assay methods.[1] The development of a direct competitive binding test for nanogram quantities of DNA has provided a sensitive probe for the investigation of DNA release.[13,14] The antibody used in our assay recognizes both single- and double-stranded DNA, with relatively high avidity, resulting in high sensitivity (25 ng DNA/ml). Extraction and purification of the test sample are unnecessary, and the determination can be performed directly on the patient's serum.[13]

In conclusion, we found that serum DNA levels were very low (usually zero) in normal healthy controls, elevated in patients with benign lesions, and the highest levels were measured in malignant disease. Serum DNA determinations may have diagnostic value, especially in conjunction with other tumor markers such as CEA. Specifically for pancreatic carcinoma, the DNA assay showed high sensitivity. Sequential determinations of DNA levels may have prognostic significance.

TABLE 3. Measurements of DNA and CEA Levels in 77 Patients and 41 Normal Controls

	Malignant	Benign	Normal
No. of patients	63	14	41
DNA Mean ± SE (ng/ml)	271 ± 58	80 ± 27	14 ± 3
Range (ng/ml)	0–2500	0–370	0–100
Significance			
malignant *vs* benign	$P < 0.05$		
benign *vs* normal		$P < 0.001$	
CEA Mean ± SE (ng/ml)	45.0 ± 11	7.5 ± 1	3.5 ± 0.2
Range (ng/ml)	3.1–250	2.1–13.6	1.5–9.0
Significance			
malignant *vs* benign	$P < 0.05$		
benign *vs* normal		$P < 0.01$	
Correlation coefficient (r), CEA *vs* DNA	0.142	0.009	0.109

TABLE 4. Sensitivity and Specificity of DNA and CEA Tests

Tumor	No. of patients	Percent No. of patients DNA (+)*	CEA (+)†	Either (+)	Both (−)
Colon–rectum	29	55	69	90	10
Pancreas	14	93	78	93	7
Stomach	9	33	55	67	33
Cecum–ileum	3	100	66	100	0
Gallbladder	3	33	100	100	0
Ampulla	2	100	100	100	0
Unknown primary	3	100	100	100	0
Cancer total	63	63	71	88	12
Benign total	14	21	21	28	72
Normal total	41	0	12	12	88
Sensitivity (% tests positive in presence of malignancy)		63 $P < 0.01$	71 $P < 0.001$	88 $P < 0.001$	
Specificity (% tests negative in absence of malignancy)		79	79	72	

* Levels higher than 100 ng/ml.
† Levels higher than 5 ng/ml.

REFERENCES

1. Leon SA, Shapiro B, Sklaroff DM, Yaros MJ. Free DNA in the serum of cancer patients and the effect of therapy. *Cancer Res* 1977; 37:646–650.
2. Davis GL, Davis JS. Detection of circulating DNA by counter-immuno-electrophoresis (CIE). *Arthritis Rheum* 1973; 16:52–58.
3. Koffler D, Agnello V, Winchester R, Kunkel HG. The occurrence of single-stranded DNA in the serum of patients with SLE and other diseases. *J Clin Invest* 1973; 52:198–204.
4. Leon SA, Ehrlich GE, Shapiro B, Labbate A. Free DNA in the serum of rheumatoid arthritis patients. *J Rheum* 1977; 4:139–143.
5. Leon SA, Revach M, Ehrlich GE, Adler R, Petersen V, Shapiro

B. DNA in synovial fluid and the circulation of patients with arthritis. *Arthritis Rheum* 1981; 24:1142–1150.

6. Barthold SW, Jonas AM. Morphogenesis of early 1,2-dimethylhydrazine-induced lesions and latent period reduction of colon carcinogenesis in mice by a variant of *Citrobacter freundii*. *Cancer Res* 1977; 37:4352–4360.

7. Reynoso G, Keane M, Kanopka S. Proficiency testing for the radio-immunoassay of CEA. *Am J Clin Pathol* 1977; 68:170–174.

8. Zamcheck N. Immunology, tumor markers, and pancreatic cancer. *J Surg Oncol* 1975; 7:155–165.

9. McIntire KR, Waldmann TA, Go VLW, Moertel CG, Ravry M. Simultaneous radio-immunoassay for CEA and AFP in neoplasms of the gastrointestinal tract. *Ann Clin Lab Sci* 1974; 4:104–108.

10. Hirai H. A collaborative clinical study of CEA in Japan. *Cancer Res* 1977; 37:2267–2274.

11. Johnson PJ, Portmann B, Williams R. AFP concentrations measured by RIA in diagnosing and excluding hepatocellular carcinoma. *Br Med J* 1978; 2:661–663.

12. Agnello V, Winchester RJ, Kunkel HG. Precipitin reactions of the C1q component of complement with aggregated γ-globulin and immune complexes in gel diffusion. *Immunology* 1970; 19:909–919.

13. Leon SA, Green A, Yaros MJ, Shapiro B. Radioimmunoassay for nanogram quantities of DNA. *J Immunol Methods* 1975; 9:157–164.

14. Leon SA, Green A, Ehrlich GE, Poland M, Shapiro B. Avidity of antibodies in SLE: Relation to severity of renal involvement. *Arthritis Rheum* 1977; 20:23–29.

15. Leon SA, Shapiro B, Servi P, Parsons RG. A comparison of DNA and DNA-binding protein levels in malignant disease. *Eur J Cancer* 1981; 17:533–538.

16. Campbell RC. Statistics for Biologists. Cambridge: University Press, 1967; 56–61.

17. Sokal RR, Rohlf FJ. Introduction to Biostatistics. San Francisco: W. H. Freeman & Co., 1973; 217–224.

18. DiMagno EP, Malagelada JR, Taylor WF, Go VLW. A prospective comparison of current diagnostic tests for pancreatic cancer. *N Engl J Med* 1977; 297:737–742.

19. Anker P, Stroun M, Maurice PA. Spontaneous release of DNA by human blood lymphocytes *in vitro*. *Cancer Res* 1975; 35:2375–2382.

20. Hughes GRV, Cohen SA, Lightfoot RW, Meltzer JI, Christian CL. The release of DNA into serum and synovial fluid. *Arthritis Rheum* 1971; 14:259–266.

21. Dorsch CA, Chia D, Levy L, Barnett EV. Persistence of DNA in the circulation of immunized rabbits. *J Rheum* 1975; 2:161–166.

22. Wroblewski F, Bodansky O. Presence of DNAse activity in human serum. *Proc Soc Exp Biol* 1950; 74:443–445.

23. Hansen HJ, Snyder JJ, Miller E *et al.* Carcinoembryonic antigen (CEA) assay: A laboratory adjunct in the diagnosis and management of cancer. *Human Pathol* 1974; 5:139–147.

24. Rule AH, Straus, E, Vandevoorde JP, Janowitz HD. Tumor-associated (CEA-reacting) antigen in patients with inflammatory bowel disease. *N Engl J Med* 1972; 287:24–26.

25. Moore TL, Kantrowitz PA, Zamcheck N. Carcinoembryonic antigen (CEA) in inflammatory bowel disease. *JAMA* 1972; 222:944–947.

Isolating Fetal Cells From Maternal Blood

Advances in Prenatal Diagnosis Through Molecular Technology

Joe Leigh Simpson, MD, Sherman Elias, MD

Objectives.—To review the rationale for and progress toward the goal of isolating and analyzing fetal cells circulating in maternal blood, and to explore the feasibility of this method in providing noninvasive prenatal cytogenetic diagnosis.

Data Sources.—Critical review of data published since the first report (1969) of fetal metaphases in maternal blood. Emphasis is placed on data since the demonstration by polymerase chain reaction (PCR) in 1989 and 1990 that fetal cells indeed exist in maternal blood.

Data Synthesis.—Clinical evaluations have not yet been conducted, but it is already clear that molecular technologies have allowed the unequivocal demonstration of fetal cells in maternal blood. Using PCR, our own group and others have demonstrated Y sequences and single gene sequences (eg, hemoglobin $Lepore_{Boston}$) in maternal blood. Thus, fetal DNA sequences indeed exist in maternal blood. Among the various candidate cells, the most promising appear to be fetal nucleated red blood cells. We isolated nucleated red blood cells on the basis of flow-sorting for the transferrin receptor and glycophorin-A. Enriched samples were then subjected to fluorescence in situ hybridization with chromosome-specific probes. This approach allowed us to detect trisomy 21 and trisomy 18, work later confirmed by others.

Conclusions.—Isolating and analyzing fetal cells from maternal blood is clearly possible. Several key biologic questions remain—the optimal cells for isolation, frequency of cells in maternal blood, timing during gestation for maternal blood sampling, and the likelihood of persistence of fetal cells after delivery. Clinical evaluations planned by the National Institute of Child Health and Human Development will determine the sensitivity and specificity of this method and its precise role in prenatal cytogenetic diagnosis.

(*JAMA*. 1993;270:2357-2361)

NONINVASIVE approaches to identify trisomic fetuses in utero (prenatal diagnosis) are desirable. Ultrasonography and serum analyte screening are already available.[1] However, neither is optimally sensitive, and the former requires considerable on-site expertise that is only available in selected referral centers. Molecular technologies have now made possible an entirely new approach—isolation and analysis of fetal cells circulating in maternal blood. This article updates this topic and inevitably reflects our previous publications.[1-3]

From the Department of Obstetrics and Gynecology University of Tennessee, Memphis (Dr Simpson and Dr Elias).

Reprint requests to the Department of Obstetrics and Gynecology, University of Tennessee, 853 Jefferson Ave, Memphis, TN 38103 (Dr Simpson).

HISTORICAL CONSIDERATIONS

Recovering fetal cells from maternal blood for prenatal diagnosis was first reported by Walknowska et al,[4] who in 1969 found "XY" metaphases in maternal blood of pregnant women carrying a male fetus. De Grouchy and Trubuchet[5] also recovered XY metaphases in pregnant women carrying male fetuses, and Y-chromatin–positive cells in maternal blood of women carrying male fetuses were claimed by several groups.[6-10] Herzenberg and colleagues[11,12] next used flow-sorting techniques to enrich for what were considered HLA-A2 lymphocytes, which were presumed fetal because that antigen was present in the father of each of those families, but not in the mother. For example, HLA-A2–positive cells sorted from the blood of an HLA-A2–negative woman were presumed to have been fetal in origin. However, confirmation was based solely on Y-chromatin analysis, a technique known to be characterized by less than 100% specificity. Vigorous efforts by several groups led to general disappointment, and by the mid 1980s few groups were still active.

After an interval of relative inactivity, various groups rekindled excitement by using polymerase chain reaction (PCR) to show that fetal cells indeed exist in maternal blood. The first group to do so was Lo et al,[13,14] who used nested primer PCR to amplify for a Y sequence. Women carrying a male fetus proved far more likely to show a Y-specific signal (hybridization) than those carrying a female fetus. Our own group[15] and others[16-20] later confirmed the results of Lo et al.[13,14]

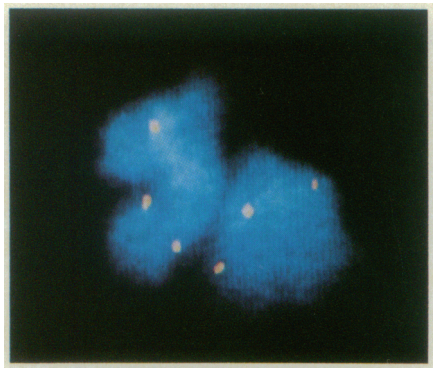
CANDIDATE FETAL CELLS

Trophoblasts

Trophoblasts are obvious candidate cells because of their intimate relation-

Downloaded from jama.ama-assn.org at American Medical Association on 14 December 2010

Exhibit AE



Results of hybridization in case 6 (Table), using a chromosome 21–specific probe set (Klinger et al[39]) to fetal nucleated red blood cells. Three hybridization signals are visualized in each nucleus. Digital imaging microscopy was performed with a Zeiss epifluorescence microscope equipped with a cooled charged couple device camera (Photometrics PM512, Tucson, Ariz), which was controlled by a personal computer. Gray scale images were captured with filter sets for DAPI (Sigma Chemical Co, St Louis, Mo) and fluorescein (pseudocolored and merged for display). Reproduced with permission from Elias et al,[21] © The Lancet Ltd, 1992.

ship with the uterus. Given the fact that trophoblasts necessarily invade the uterus, some cells would be expected to find their way into the maternal circulation. Covone et al[21] were the first to claim trophoblasts had been isolated from peripheral blood of pregnant women at varying stages of gestation, based on separation with monoclonal antibody against H315. Unfortunately, the same group later concluded that H315-positive cells sorted from maternal blood were not fetal trophoblasts, but rather maternal cells to which H315 antigen had been adsorbed.[22]

More recently, other groups again began to generate monoclonal antibodies against trophoblasts. Of 6000 monoclonal antibodies generated from placental tissue, Mueller et al[17] concluded that five were specific for fetal tissue. After maternal blood from pregnant women was exposed to two of the five monoclonal antibodies, the presumptively isolated fetal cells were subjected to PCR to detect presence or absence of Y sequences. Fetal sex was correctly identified in seven of seven males and six of seven females. No further data have been presented since the original publication in 1990.

Bruch et al[23] used monoclonal antibodies (GB17, GB21, and GB25) directed either against syncytiotrophoblasts alone or against both syncytiotrophoblasts and cytotrophoblasts (GB25); however, these antibodies were not specific for trophoblasts, and by electron microscopic analysis, the cells had the morphologic appearance of leukocytes. The possibility that trophoblast antigens were being adsorbed onto maternal cells was thus not excluded. Nonetheless, flow-sorted cells subjected to PCR for Y-specific sequences yielded the predicted Y signal in two of the three samples tested. Using cells separated by immunomagnetic lymphocyte depletion and flow-sorting, Cacheux et al[24] detected 47,XYY cells by fluorescence in situ hybridization (FISH). Of 1387 nuclei, 14 (1%) showed two Y-specific domains; 45 (3.24%) nuclei had one Y-specific domain.

Lymphocytes

The fetal cells sorted from maternal blood by Herzenberg and colleagues[11,12] were considered lymphocytes. Couples studied were those in which the father but not the mother was HLA-A2–positive. To elaborate, leukocytes from maternal blood were first separated by Ficoll-Hypaque gradient and then subjected to fluorescence-activated cell sorting for HLA-A2–positive cells. Fluorescein-stained cells were scored visually for presence or absence of Y chromatin. Among 12 pregnancies that resulted in male infants, five showed Y-chromatin cells (range, 0.3% to 1.6% of sorted cells). Among the seven mothers delivered of infants whose lymphocytes failed to react with anti–HLA-A2 antiserum, none showed Y-chromatin cells. That lymphocytes can be isolated was later shown by molecular techniques. Yeoh et al[25] flow-sorted maternal blood for cells characterized by paternal (and hence fetal) HLA antigens. PCR demonstrated paternal (and hence presumably fetal) HLA alleles in maternal blood.

In our own studies,[26,27] many metaphases were analyzed from pregnancies known to have male or anepuloid fetuses; metaphases were invariably 46,XX. Overall, extant data could be explained by postulating that fetal lymphocytes were present, but unresponsive to mitogens. The function of these cells in the maternal circulation is speculative, but plausibly related to immune signaling.

Granulocytes

A single group has claimed to isolate fetal granulocytes. Wessman et al[28] used Ficoll-Paque density gradient centrifugation followed by FISH with a Y-specific probe that recognizes heterochromatic Yq. Y-specific cells were recovered from eight pregnant women, seven of whom gave birth to male infants. The sole error was said to have involved "lymphocyte-like cells," leading the authors to conclude total accuracy "in each case where granulocytes with a Y-specific signal were detected." That an unusually high number of fetal granulocytes were obtained without fetal cell enrichment was surprising.

Nucleated Red Blood Cells (Erythroblasts)

The candidate cells our group[15,29-31] is pursuing are fetal nucleated red blood cells (NRBCs) or erythroblasts. Bianchi et al[16] were the first to focus on these cells, initially flow-sorting on the basis of transferrin receptor (CD71) positivity. Of eight samples that showed a Y sequence by PCR, six were derived from pregnancies in which women were carrying male fetuses.

Selection of this candidate cell arose because earlier studies by our group had suggested that the rapidly dividing cells obtained from newborn or fetal umbilical blood were erythroblasts (ie, NRBCs).[32] Indeed, circulating NRBCs comprise about 10% of the RBCs in the 11-week fetus and 0.5% in the 19-week fetus. Nucleated red blood cells are very rare in peripheral adult blood. Thus, almost any NRBC present in maternal blood could be presumed to be of fetal origin.

Our initial experience showed that flow-sorting on the basis of CD71 alone proved ineffective, a conclusion to which Bianchi et al[33] later independently agreed. We now sort for fetal nucleated erythrocytes on the basis of four criteria: cell size, cell granularity, CD71, and glycophorin-A, the major sialoglycoprotein of the erythroid cell membrane.[29] The following is our rationale: (*a*) CD71 is present on early erythrocytes (and also monocytes and certain activated lymphocytes), but not on mature erythrocytes; and (*b*) glycophorin-A is present in the erythrocyte lineage and absent in leukocytes. Thus, cells that bind both CD71 and glycophorin-A antibodies are likely to be early (fetal) erythrocytes.

Our group sorts with an EPICS Elite flow cytometer (Coulter Cytometry, Hialeah, Fla) that contains standard instrument computer and electronics. The unit is configured with an air-cooled 15-mW argon laser (model 2201-20FLE, Cyonics Division of Unighase, San Jose, Calif) emitting at 488 nm. Elite 2.1 cytometry software was used. Cells were analyzed in a single pass at a rate of 2500 to 5000 cells per second. Dual fluorescent cells of appropriate size and granularity were sorted directly into 1.5-mL Eppendorf tubes that contained 1 mL of buffer. The cells recovered are morphologically identical to NRBCs in umbilical cord blood.

Employing the same nested primers as did Lo et al,[13] we found that PCR identified male fetuses in 12 (100%) of flow-sorted samples and female fetuses in five (83%) of six samples.[15] Bianchi and colleagues[33] also sort not only on the basis of CD71 positivity, but presence of

Downloaded from jama.ama-assn.org at American Medical Association on 14 December 2010

either glycophorin-A or the thrombospondin receptor (CD36).[34] Ganshirt-Ahlert et al[35] use a triple-density gradient, followed by magnetic-activated cell sorting for recovering transferrin receptor antibody–labeled cells.[36]

DETECTING MENDELIAN TRAITS

Although likely to be exceptional until separation of a "pure" population of fetal cells is possible, mendelian traits have already been detected by analysis of separated fetal cells. Camaschella et al[20] studied total DNA obtained from maternal blood in three pregnancies at risk for fetal β-thalassemia/hemoglobin $Lepore_{Boston}$. Hemoglobin $Lepore_{Boston}$ is a hemoglobinopathy caused by a hybrid δβ gene that is the result of a 7-kilobase deletion in the β-globin gene cluster. Using PCR to amplify hemoglobin $Lepore_{Boston}$–specific DNA fragments from maternal blood of women whose partners carried the $Lepore_{Boston}$ gene, Camaschella et al[20] correctly identified this hemoglobinopathy in two fetuses. In a third, it was not present.

Using a similar approach, Lo et al[37] identified fetal rhesus (RhD) status in pregnancies involving sensitized RhD-negative women married to heterozygous men. Identifying inheritance of the RhD allele from the father would necessitate use of various diagnostic procedures, but identifying inheritance of the RhD-negative allele would not because both mother and fetus would be RhD-negative. Without flow-sorting, cells from 21 RhD-negative pregnant women were subjected to PCR for RhD DNA sequences. Of 11 women who showed a positive PCR signal, seven had an RhD-positive fetus. Of 10 who did not show a PCR signal, only two had an RhD-positive fetus. Overall, eight of the 10 women who carried an RhD-positive signal showed a signal. Further experience should allow fetal RhD status to be determined without an invasive procedure.

Detecting a gene present in the father but not the mother could also be useful in excluding autosomal recessive traits. Suppose the mother has hemoglobin SS and the father AS. If blood from this homozygous mother (*SS*) showed a normal paternal allele (*A*) when subjected to PCR, the fetus would have to be heterozygous (*AS*).

DETECTING FETAL ANEUPLOIDY

Diagnosis of fetal aneuploidy will probably require technologies other than PCR alone, although quantitative PCR is a theoretical option. At present, the best approach seems to involve FISH, using chromosome-specific DNA probes. The background frequency of euploid cells with three domains is accepted as about 1%; thus, a significantly higher proportion of cells with three domains should indicate that there are fetal trisomic cells present in maternal blood. Using this approach, our group was the first to report success in detecting fetal aneuploidy from a blood sample taken prior to the chorionic villus sample (CVS), first for trisomy 18[29] and then for trisomy 21.[30]

In collaboration with Katherine Klinger, PhD (Integrated Genetics, Framingham, Mass), we have now studied 69 maternal blood samples.[38] In some cases, a 20-mL blood sample was obtained after prenatal diagnosis of fetal aneuploidy by first-trimester CVS, second-trimester CVS, or second-trimester amniocentesis. Samples obtained after an invasive procedure are of limited value because fetal cells could pass into the maternal circulation solely as a result of bleeding from fetal circulation into maternal circulation. In more informative cases, a 20-mL maternal blood sample was drawn prior to the CVS or amniocentesis.

After samples were flow-sorted for NRBCs as described, coded slides were sent in our original work to Integrated Genetics in batches of two to six samples. The chromosome-specific probes used for FISH analysis included HL 10-10-8a, a repetitive sequence specific for the X centromere; pWe7.1 and pWe5.0, overlapping mid- to long-arm regions specific for chromosome 18; 519, a 20-kilobase sequence on the distal long arm of chromosome 21; and pDP97, a repeat sequence that recognizes the centromere alphoid repeat *DYZ3* of the Y chromosome. The number of suitable nuclei in a single blood sample ranged widely (300 to 50 000), probably because only 10% to 30% of sorted nuclei become attached to slides during the process of slide preparation. It was sometimes not possible to use all five chromosome-specific probes. If a high probability of finding fetal aneuploid cells existed (ie, CVS or amniocentesis had already shown fetal trisomy, or ultrasonographic evidence of fetal anomalies existed), probes for chromosomes 21 and 18 were preferentially used.

Our frequent changes in protocol render calculations concerning sensitivity and specificity inappropriate. However, our initial observations are worth highlighting, particularly the results from women who carried chromosomally abnormal fetuses (Table; Figure). Other groups later independently confirmed the detection of autosomal trisomy in samples obtained after an invasive procedure[40] and in samples obtained prior to amniocentesis.[36]

Analyses of Flow-Sorted Maternal Blood From Eight Women Who Carried Chromosomally Abnormal Fetuses*

Case	Procedure and Timing in Relationship to Venipuncture	Gestation, wk	Fetal Karyotype	Trisomic Cells/ Total Cells (%)
1	After CVS[29]	11	47,XX,+21	24/62 (38.7)
2	After amniocentesis	19	47,XX,+21	21/113 (18.6)
3	Before CVS[29]	10	47,XY,+18	25/295 (8.5)
4	Before CVS	11	47,XX,+21	3/104 (2.9)
5	After amniocentesis	17	47,XX,+21	4/50 (8.0)
6	Before CVS[30]	10	47,XX,+21	74/100 (74.0)
7	After CVS	11	47,XXY	6/94 (6.4)
8	Before CVS	11	47,XXY	0/93 (0.0)

*CVS indicates chorionic villus sampling at 10 to 12 weeks; amniocentesis was done at 16 to 18 weeks; modified from Elias et al[36]; in 61 other cases, the fetal karyotype was normal, and trisomic cells were not found in similar percentage.

KEY BIOLOGIC QUESTIONS

Although clinical evaluations are slated to begin soon, there is still little understanding concerning several key biologic questions that can affect sensitivity and specificity. These questions need to be addressed before fetal cell isolation can be used as either a definitive or complementary diagnostic test.

Optimal Cell Type(s) for Isolation

As discussed, we believe that the best candidate fetal cells are NRBCs. We are especially concerned that lymphocytes are poor candidates because these cells could persist from prior pregnancies. Granulocytes are largely unexplored as candidate cells. Trophoblasts remain theoretically attractive, but unequivocal proof of a specific monoclonal antibody is still lacking.

Frequency of Fetal Cells in Maternal Blood

Estimates of the ratio of fetal NRBCs to maternal NRBCs have varied widely. Factors that might influence the ratio include gestational age at sampling and specific fetal cell type. Indeed, a steady state of fetal cells in maternal blood may or may not exist. Even bimanual pelvic examination (squeezing the uterus prior to blood sampling) could cause transient

Downloaded from jama.ama-assn.org at American Medical Association on 14 December 2010

passage of fetal cells in maternal blood. Our own work estimates that the ratio of fetal NRBCs to nucleated maternal blood cells is $1:1\times10^7$ to $1:1\times10^8$, based on flow cytometry data.[29] Using an avidin-biotin–based immunoaffinity system, Hall and Williams[41] developed a purification method using an antibody directed against CD34; fetal cells were present in the range $1:4.75\times10^6$ to $1:6\times10^7$ of the nucleated cell fraction of maternal blood. Given the rarity of fetal cells in maternal blood, enrichment by either fluorescence-activated cell sorting, magnetic-activated cell sorting,[35] or other techniques undoubtedly will be required for fetal diagnoses.

An unanswered question is whether the frequency of fetal cells in maternal blood of women carrying aneuploid fetuses is similar to that of women carrying euploid fetuses. That 74% of cells sorted from our patient who carried 47,XX,+21[30] were of that chromosomal complement raises the hypothesis that the placenta in some aneuploid pregnancies allows relatively more fetal cells into the maternal circulation than is true for euploid pregnancies. Thus, it is not unreasonable to postulate that cells from aneuploid fetuses show higher frequencies than cells from normal fetuses. Of course, this would be salutary for diagnosis.

Timing of Maternal Blood Sampling

It remains unclear when would be the optimal time during pregnancy to obtain a maternal blood sample that would yield an adequate number of fetal cells for prenatal diagnosis. In summarizing several of their studies in which flow-sorting for NRBCs based on CD71 positivity was combined with PCR for Y-specific sequences, Bianchi et al[42] concluded that these cells were unlikely to be present after 16 weeks' gestation. However, using a combination of triple-density gradient followed by magnetic-activated cell sorting for transferrin receptor antibody–labeled cells, Ganshirt-Ahlert et al[36] reported that NRBCs could be enriched in all three trimesters. Our own data indicate that NRBCs are present from at least 10 to 18 weeks' gestation, with no obvious diminution in frequency after 16 weeks. Overall, we believe that isolation of fetal cells from maternal blood for prenatal diagnosis may be accomplished in the first trimester as well as in the second trimester. Possibly the proportion of fetal cells in maternal blood relative to fetal mass is highest in early pregnancy; however, given the increasing fetal mass with increasing gestation, the absolute number in the maternal blood remains relatively constant.

Do Rh or ABO Incompatibilities Affect Screening Efficacy?

Concern has been raised that Rh or ABO incompatibilities between the fetus and mother could affect the detection of fetal NRBCs from the maternal circulation. Actually, maternal/fetal ABO incompatibility should probably not be a factor because fetal RBCs contain fewer reactive A or B sites than adult cells (6000 to 7000 per fetal cell vs 26 000 per adult cell); such sites are also farther apart.[43] It has also been suggested that fetal RBC antigens are not as strong and also possibly not as specific as adult antigens, offering additional protection against interaction with maternal antibody.[44]

Persistence of Fetal Cell From Prior Pregnancies

Fetal NRBCs from a prior pregnancy could persist in the maternal circulation and be a potential source of diagnostic error. A false-positive result could then occur if such cells persisted after delivery of a chromosomally abnormal offspring or chromosomally abnormal spontaneous abortion. The latter scenario would be particularly troublesome because the frequency of chromosomal abnormalities among first-trimester spontaneous abortions is high (50% to 60%), and women can become pregnant very soon after a spontaneous abortion. Indeed, we hypothesize that one major advantage for targeting fetal cells in the erythroid lineage is their relatively short life span. Although little is known concerning the maturation time (ie, interval between multipotential stem cell and mature anucleated RBC) during 10 and 19 weeks' gestation, morphologic observations in adult bone marrow suggest four stages of NRBCs separated by three mitotic divisions. Erslev[45] estimated that the total intramedullary maturation time (multipotential stem cell to reticulocyte) is approximately 5 days in the adult. Thus, NRBCs from prior pregnancies should probably not persist and interfere with results of analysis of NRBCs flow-sorted from maternal blood. Nonetheless, this concern must be addressed formally.

NECESSITY FOR CLINICAL EVALUATIONS

In 1987, our group at the University of Tennessee, Memphis, and colleagues at the University of California, San Francisco, were funded by the National Institute of Child Health and Human Development (NICHD) to investigate the feasibility of isolating fetal cells from the maternal peripheral blood for prenatal diagnosis (phase I). Based on encouraging preliminary data from these groups and others, the NICHD has announced a request for proposals for a phase II clinical evaluation to test the methods developed in phase I. The phase II clinical evaluation will include a systematic study of the variables involved in the use of flow-cytometry–sorted fetal cells from maternal blood and in the use of FISH with chromosome-specific DNA probes. Results of these studies, soon to begin, will be compared with the results obtained from amniocentesis or CVS on the same women. Overall, this will produce information on sensitivity and specificity that can be compared with that of traditional prenatal diagnostic tests. The NICHD proposal calls for up to three centers to recruit at least 1000 women per center. A major focus will be to determine the sensitivity and specificity for this diagnostic method. Concurrently, information will be gathered concerning attitudes and preferences of pregnant women toward a first-trimester noninvasive technique of prenatal diagnosis (ie, fetal cells in maternal blood) and toward pregnancy termination if abnormal fetuses are diagnosed.

COULD SORTING FETAL CELLS FROM MATERNAL BLOOD PERMIT DEFINITIVE CYTOGENETIC DIAGNOSIS?

Ideally, the sensitivity, specificity, positive predictive value, and negative predictive value will prove sufficient for this noninvasive approach to serve as the definitive diagnostic test for fetal aneuploidy. Even if this stringent requirement is not met, we believe that this method could still play an important role in prenatal screening for fetal aneuploidy, either as an independent test or as an added benefit in combination with other tests. For example, our group as well as others showed that for women younger than 35 years the combination of maternal serum alpha-fetoprotein, human chorionic gonadotropin, and unconjugated estriol in the second trimester is an effective screening test for fetal Down syndrome and other aneuploidies.[46] For all ages, the sensitivity (detection rate) is estimated as about 60% at a "cost" of 5% of women requiring amniocentesis for definitive diagnosis. If it proves cost-effective to recover fetal cells from maternal blood, this approach could be combined with maternal serum markers to further increase the identification of the proportion of women at increased risk for fetal chromosomal abnormalities. These women could then proceed to either CVS or amniocentesis for definitive prenatal diagnosis. That is, one might entertain a four-part screen (or, with ultrasonog-

Downloaded from jama.ama-assn.org at American Medical Association on 14 December 2010

raphy, a five-part screen) to improve detection rates. Isolating even a few fetal trisomic cells could have great significance, but we suspect failure to find such cells might not connote normalcy. In this situation, information from serum analytes would remain paramount.

This study was supported in part by NIH contracts HD8-2903 and HD8-2904.

The authors thank David Page, MD, Whitehead Institute, Massachusetts Institute of Technology, Boston, for contribution of the pDP97 clone.

References

1. Simpson JL, Elias S. Techniques for prenatal diagnosis of genetic disorders. *Infer Repro Med Clin North Am.* In press.
2. Elias S, Simpson JL. Prenatal diagnosis using fetal cells isolated from maternal blood. In: Grudzinskas JG, ed. *Screening for Down's Syndrome.* London, England: Cambridge University Press. In press.
3. Simpson JL, Elias S. Molecular technologies for the advancement of prenatal genetic diagnosis: isolating fetal cells from maternal blood. In: *Proceedings of the VIII World Congress on Human Reproduction, Bali, Indonesia: April 3-9, 1993.* Park Ridge, NJ: Parthenon Press. In press.
4. Walknowska J, Conte FA, Grumback MM. Practical and theoretical implications of fetal/maternal lymphocyte transfer. *Lancet.* 1969;1:1119-1122.
5. De Grouchy J, Trubuchet C. Transfusion foetomaternelle de lymphocytes sanguins et detection du sexe du foetus. *Ann Genet.* 1971;14:133-137.
6. Schroder J, de la Chapelle A. Fetal lymphocytes in maternal blood. *J Hematol.* 1972;39:153-162.
7. Schroder J, Tiilikainen A, de la Chapelle A. Fetal leukocytes in the maternal circulation after delivery. *Transplantation.* 1974;17:346-360.
8. Grossett L, Barrelet V, Odartchenko N. Antenatal fetal sex determination from maternal blood during early pregnancy. *Am J Obstet Gynecol.* 1974; 120:60-63.
9. Siebers JW, Knauf I, Hillemans HG. Antenatal sex determination in blood from pregnant women. *Hum Genet.* 1975;28:273-280.
10. Zilliacus R, de la Chapelle A, Schroder J, Tiilikainen A, Kohne E, Kleihaue E. Transplacental passage of fetal blood cells. *Scand J Haematol.* 1975;15:333-338.
11. Herzenberg LA, Bianchi DW, Schroder J, Cann HM, Iverson GM. Fetal cells in the blood of pregnant women: detection and enrichment by fluorescence-activated cell sorting. *Proc Natl Acad Sci U S A.* 1979;76:1453-1455.
12. Iverson GM, Bianchi DW, Cann HM, Herzenberg LA. Detection and isolation of fetal cells from maternal blood using the fluorescence-activated cell sorter (FACS). *Prenat Diagn.* 1981;1:61-73.
13. Lo Y-MD, Wainscot JS, Gilmer MDG, Patel P, Sampietro M, Fleming KA. Prenatal sex determination by DNA amplification from maternal peripheral blood. *Lancet.* 1989;2:1363-1365.
14. Lo Y-MD, Pate P, Sampietro M, Gillmer MDG, Fleming KA, Wainscoat JS. Detection of single-copy fetal DNA sequence from maternal blood. *Lancet.* 1990;335:1463-1464.
15. Wachtel SS, Elias S, Price J, et al. Fetal cells in the maternal circulation: isolation by multiparameter flow cytometry and confirmation by PCR. *Hum Reprod.* 1991;6:1466-1469.
16. Bianchi DW, Flint AF, Pizzimenti MF, Knoll JH, Latt SA. Isolation of fetal DNA from nucleated erythrocytes in maternal blood. *Proc Natl Acad Sci U S A.* 1990;87:3279-3283.
17. Mueller OW, Hawes CS, Wright AE, et al. Isolation of fetal trophoblast cells from peripheral blood of pregnant women. *Lancet.* 1990;336:197-200.
18. Holzgreve W, Ganshirt-Ahlert D, Dohr A, Helmer L, Miny P. Magnetically activated cell sorting for the isolation of fetal cells from maternal circulation. *Prenat Diagn.* 1992;12:S17.
19. Wilson RD, Langlois S. Non-invasive prenatal sex determination for X-linked genetic disease by DNA amplification from maternal blood. *Prenat Diagn.* 1992;12:S177.
20. Camaschella C, Alfarno A, Gattardi E, et al. Prenatal diagnosis of fetal haemoglobin Lepore-Boston disease on maternal peripheral blood. *Blood.* 1990;75:2102-2106.
21. Covone AE, Johnson PM, Mutton D, Adinolfi M. Trophoblast cells in peripheral blood from pregnant women. *Lancet.* 1984;2:841-843.
22. Covone AE, Kozma R, Johnson PM, Latt SA, Adinolfi M. Analysis of peripheral maternal blood samples for the presence of placental-derived cells using Y-specific probes and McAb H315. *Prenat Diagn.* 1988;8:591-607.
23. Bruch JF, Metezeau D, Garcia-Fonknechten N, et al. Trophoblast-like cells sorted from peripheral maternal blood using flow cytometry: a multiparametric study involving transmission electron microscopy and fetal amplification. *Prenat Diagn.* 1991; 11:787-798.
24. Cacheux V, Milesi-Fluet C, Tachdjian G, et al. Detection of 47,XYY trophoblast fetal cells in maternal blood by fluorescence in situ hybridization after using immunomagnetic lymphocyte depletion and flow cytometry sorting. *Fetal Diagn Ther.* 1992; 7:190-194.
25. Yeoh SC, Sargent IL, Redman CWG, Wordsworth BP, Thein SL. Detection of fetal cells in maternal blood. *Prenat Diagn.* 1991;11:117-123.
26. Tharapel AT, Jaswaney V, Dockter M, et al. Can fetal cells in maternal blood be selected through cytogenetic means? *Am J Hum Genet.* 1989; 45(suppl):271.
27. Tharapel AT, Jaswaney VL, Dockter ME, et al. Failure to detect fetal metaphases in lymphocytes flow sorted by maternal-fetal HLA differences. *Fetal Diagn Ther.* 1993;8:95-101.
28. Wessman M, Ylinen K, Knuutila S. Fetal granulocytes in maternal venous blood detected in in situ hybridization. *Prenat Diagn.* 1992;12:993-1000.
29. Price J, Elias S, Wachtel SS, et al. Prenatal diagnosis using fetal cells isolated from maternal blood by multiparameter flow cytometry. *Am J Obstet Gynecol.* 1991;165:1731-1737.
30. Elias S, Price J, Dockter M, Wachtel S, Tharapel A, Simpson JL. First trimester prenatal diagnosis of trisomy 21 in fetal cells from maternal blood. *Lancet.* 1992;340:1033.
31. Simpson JL, Elias S. Isolating fetal erythroblasts from maternal blood with identification of fetal trisomy by fluorescent in situ hybridization. *Prenat Diagn.* 1992;12:S34.
32. Tipton RE, Tharapel AT, Change HT, et al. Rapid chromosome analysis using spontaneously dividing cells derived from umbilical cord blood (fetal and neonatal). *Am J Obstet Gynecol.* 1990; 163:1546-1548.
33. Bianchi DW, Klinger KW. Prenatal diagnosis through the analysis of fetal cells in the maternal circulation. In: Milunsky A, ed. *Genetic Disorders and the Fetus.* 3rd ed. Baltimore, Md: The Johns Hopkins University Press; 1992:759-770.
34. Bianchi DW, Zickwolf GK, Yih MC, et al. Erythroid-specific antibodies enhance detection of fetal nucleated erythrocytes in maternal blood. *Prenat Diagn.* 1993;13:293-300.
35. Ganshirt-Ahlert D, Burschyk M, Garritsen HS, et al. Magnetic cell sorting and the transferrin receptor as potential means of prenatal diagnosis from maternal blood. *Am J Obstet Gynecol.* 1992;166: 1350-1355.
36. Ganshirt-Ahlert D, Borjesson-Stoll M, Burschyk M, et al. Noninvasive prenatal diagnosis: triple density gradient, magnetic activated cell sorting and FISH prove to be an efficient and reproducible method for detection of fetal aneuploidies from maternal blood. *Am J Hum Genet.* 1992;51:A48.
37. Lo Y-MD, Bowell PJ, Selinger M, et al. Prenatal determination of fetal RhD status by analysis of peripheral blood of rhesus negative mothers. *Lancet.* 1993;341:1147-1148.
38. Elias S, Price J, Klinger K, et al. Prenatal diagnosis of aneuploidy using fetal cells isolated from maternal blood. *Am J Hum Genet.* 1992;51:A4.
39. Klinger K, Landes G, Shook, D, et al. Rapid detection of chromosome aneuploidies in uncultured amniocytes by using fluorescence in-situ-hybridization (FISH). *Am J Hum Genet.* 1992;51:55-65.
40. Bianchi DW, Mahr A, Zickwolf GL, Houseal TW, Flint AF, Klinger KW. Detection of fetal cells with 47,XY,+21 karyotype in maternal peripheral blood. *Hum Genet.* 1992;90:368-370.
41. Hall JM, Williams SJ. Isolation and purification of CD34+ fetal cells from maternal blood. *Am J Hum Genet.* 1992;51:A257.
42. Bianchi DW, Stewart JE, Garbert MF, Lucotte G, Flint A. Possible effect of gestational age on the detection of fetal nucleated erythrocytes in maternal blood. *Prenat Diagn.* 1991;11:523-528.
43. Voak D, Williams MA. An explanation for the failure of the direct antiglobulin test to detect erythrocyte sensitization in ABO haemolytic disease of the newborn and observations on pinocytosis of IgG anti-A antibodies by infant (cord) red cells. *Br J Haematol.* 1971;20:9-23.
44. Tovey GH. Recent advances in the study of Rh factors. *Med Press (Lond).* 1952;228:300-306.
45. Erslev AJ. Production of erythrocytes. In: Williams WJ, Beutler E, Erslev AM, Lictman MA, eds. *Hematology.* 4th ed. New York, NY: McGraw-Hill International Book Co; 1990:389.
46. Phillips OP, Elias S, Shulman LP, Andersen RN, Morgan CD, Simpson JL. Maternal serum screening for fetal Down syndrome in women less than 35 years of age using alpha fetoprotein, hCG, and unconjugated estriol: a prospective 2-year study. *Obstet Gynecol.* 1992;80:353-358.

Downloaded From: jama.ama-assn.org at American Medical Association on 14 December 2010

Detection of melanoma cells in peripheral blood by means of reverse transcriptase and polymerase chain reaction

BARBARA SMITH PETER SELBY JENNIFER SOUTHGATE
KENNETH PITTMAN CHRISTOPHER BRADLEY G. ERIC BLAIR

Only small numbers of cells from solid tumours are needed for haematogenous metastasis. Detection is difficult because existing techniques are not sensitive enough. We have used reverse transcriptase to make complementary DNA from peripheral blood messenger RNA, and the polymerase chain reaction (PCR) to amplify cDNA specific for a gene actively transcribed only in the tumour tissue type.

We prepared cDNA from peripheral blood of seven patients with malignant melanoma, four patients with other metastatic cancers, and four healthy subjects, as well as from several melanoma-derived cell lines. PCR was used to amplify the gene for tyrosinase, a tissue-specific gene in melanocytes. Since normal melanocytes are not thought to circulate in peripheral blood, detection of tyrosinase transcription in peripheral blood should indicate the presence of circulating cancer cells. The method was highly sensitive and could detect a single melanoma cell from a cell line in 2 ml normal blood. Blood samples from four of the seven patients with malignant melanoma gave positive results, whereas all eight control subjects gave negative results.

This method does not depend on the characterisation of cancer-specific genetic abnormalities and can be applied to any cancer for which tissue-specific genes can be identified, including epithelial cancers. It could prove useful in the diagnosis of primary or metastatic cancers, in assessing prognosis, and in detecting residual disease after treatment.

Lancet 1991; **338**: 1227–29.

Introduction

Haematogenous spread of cancer influences outcome of the disease for most patients. However, little is known about the nature of this process. Implantation of cancer cells from peripheral blood is believed to be an inefficient process,[1] and for the common solid tumours few cancer cells are detected in peripheral blood by routine microscopy. Detailed study of the numbers, appearance, and biology of this important subpopulation of tumour cells has been limited by the methods available. Detection of circulating tumour cells depended on the availability of specific antibodies to tumour-specific cell-surface antigens and the use of immunocytological techniques. An example of this method is the detection of neuroblastoma cells in blood.[2]

Since the numbers of tumour cells in peripheral blood are very small, techniques for their detection need to be not only highly sensitive but also specific. The polymerase chain reaction (PCR)[3] could enable detection of cancer cells in the peripheral blood by amplification of specific DNA sequences. It can be applied directly, however, only when cancer-specific abnormalities in DNA—for example, the translocation forming the Philadelphia chromosome[4] or point mutations in the *RAS* gene[5]—have been characterised. Such abnormalities have not yet been found for most common solid cancers. We therefore tried an alternative approach; we used reverse transcriptase to prepare complementary DNA (cDNA) from peripheral blood messenger RNA and targeted tissue-specific gene transcription to identify cancer cells. Only genes that are being actively expressed should be represented in the cDNA. Since we would not expect epithelial or mesenchymal cells or melanocytes normally to be present in the peripheral blood, genes specific to these tissues might be suitable targets for the detection of malignant circulating cells.

We have evaluated this approach in malignant melanoma. This tumour has a wide range of metastatic sites, most of which are reached haematogenously. Normal melanocytes are not believed to circulate, and a range of melanin biosynthetic enzymes can act as tissue-specific proteins. We chose tyrosinase, an important enzyme in melanin biosynthesis, because the genomic organisation[6] and the cDNA sequence[7] are known. The tyrosinase gene contains several introns, so for DNA amplification by PCR we selected primers represented on different exons of the gene. PCR would be expected to amplify cDNA between the primers, but genomic DNA either would not be amplified or would generate a PCR product of much greater length.

Methods

Primer sequences were devised from the published sequence for human tyrosinase cDNA[7] (fig 1). The outer primers produce a PCR fragment of 284 bp and the nested primers a fragment of 207 bp.

β-globin primers were devised from the published sequence:[8] Glo4 = ACCCAGAGGTTCTTTGAGTC (sense) and Glo5 = TCTGATAGGCAGCCTGCACT (anti-sense). They are located either side of intron 2 and produce a fragment of 283 bp.

Four established human melanoma-derived cell lines were used—the melanotic SK-mel-19 and SK-mel-23[9] and the amelanotic SK-mel-28[9] and M5[10] cell lines. All cultures were maintained on an equal mixture of Dulbecco's modified Eagle's medium and RPMI 1640, supplemented with 5% fetal calf serum (Sera-Lab). Cultures were transferred by means of trypsin (0·25%) and 5 mmol/l edetic acid (EDTA).

ADDRESSES: **Yorkshire Cancer Research Campaign Institute for Cancer Studies, St James's University Hospital, Leeds** (B. Smith, BSc, Prof P. Selby, FRCP, J. Southgate, PhD, K. Pittman, MB, C. Bradley, MD), **and Department of Biochemistry and Molecular Biology, University of Leeds, Leeds, UK** (G. E. Blair, PhD). Correspondence to Prof Peter Selby, Yorkshire Cancer Research Campaign Institute for Cancer Studies, St James's University Hospital, Beckett Street, Leeds LS9 7TF, UK.

Exhibit AF

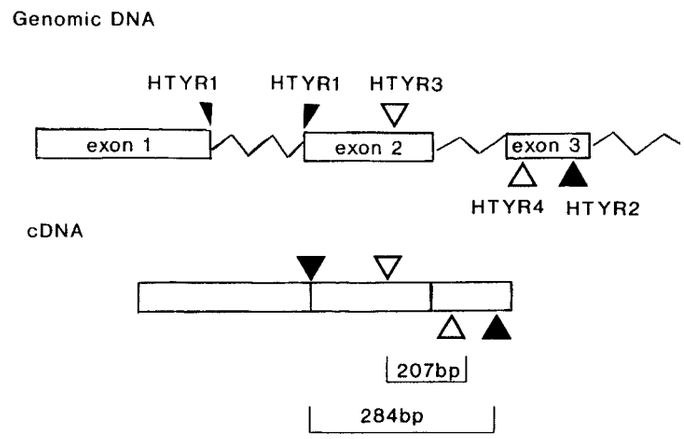


Fig 1—Diagrammatic representation of exons 1–3 of human tyrosinase gene and cDNA showing relative positions of outer primers (▼) and nested primers (▽).

HTYR1 sequence comprises end of exon 1 and start of exon 2. WWW = long intervening sequences (introns) not represented in cDNA or PCR products. Genomic DNA is not amplified.
HTYR1 = TTGGCAGATTGTCTGTAGCC (outer, sense).
HTYR2 = AGGCATTGTGCATGCTGCTT (outer, anti-sense).
HTYR3 = GTCTTTATGCAATGGAACGC (nested, sense)
HTYR4 = GCTATCCCAGTAAGTGGACT (nested, anti-sense).

Samples of blood from seven patients with malignant melanoma, four patients with other cancers, and four healthy controls were analysed by reverse transcription and PCR in a "blind" trial. 2–10 ml blood was centrifuged at 1000 *g* for 5 min, the plasma was discarded, and the samples stored at −70°C. Total RNA was extracted by the guanidinium thiocyanate/caesium chloride method[11] from the blood samples and from about 10^7 cells from the cultured lines. For reverse transcription, 10 µl RNA was heated at 90°C for 4 min, cooled rapidly, and diluted to 20 µl with a mixture containing final concentrations of: 1 × PCR buffer (10 mmol/l "tris"-HCl pH 8·4, 50 mmol/l potassium chloride, 100 µg/ml gelatin), 1 mmol/l each of dATP, dCTP, dGTP, and TTP, 8 mmol/l magnesium chloride, 25 pmol HTYR2, 20 units 'RNAguard' (Pharmacia), and 4 units murine moloney leukaemia virus reverse transcriptase (Pharmacia). After incubation at 37°C for 1 h, half the sample (10 µl) was diluted to 50 µl containing final concentrations of: 1 × PCR buffer, 200 µmol/l of each dNTP, 1·6 mmol/l magnesium chloride, 150 pmol HTYR1 and HTYR2, 0·1% 'Triton X-100', and 1 unit *Taq* DNA polymerase (Promega). Each sample was overlaid with oil and heated at 95°C for 5 min; 30 cycles PCR were then carried out (95°C for 65 s, 55°C for 65 s, 72°C for 50 s). For reamplification with the nested primers HTYR3 and HTYR4, 5 µl of a 1 in 100 dilution were amplified in 25 µl reaction volume for a further 30 cycles. To minimise contamination, all components of the reactions (except RNA and DNA) were prepared in a classs 100 cabinet (Cleansphere', Gelman Sciences) and positive displacement pipettes were used.

PCR products were analysed by electrophoresis on 2% agarose gels, followed by ethidium bromide staining. Blood RNA integrity was checked by reverse transcription PCR with primers for human β globin mRNA (Glo4 and Glo5).

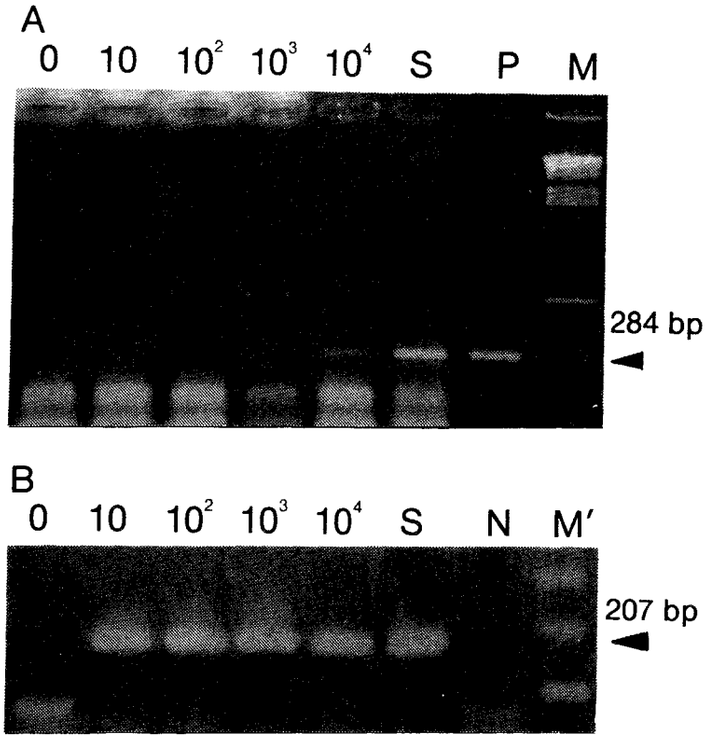


Fig 2—Detection of tyrosinase mRNA in 10 ml normal (subject 15) blood samples spiked with dilutions of SK-mel-23 cells (0–10⁴).

A = PCR with HTYR1 and HTYR2 alone (30 + 30 cycles).
B = PCR with HTYR1 and HTYR2 followed by reamplification with HTYR3 and HTYR4 (30 cycles each)
S = SK-mel-23 positive control; P = Pmel34 DNA; N = no RNA control, M = λ/*Hin*D3 marker; M' = 123 bp ladder marker

DETECTION OF TYROSINASE AND β GLOBIN mRNA IN CANCER PATIENTS AND HEALTHY SUBJECTS

Patient	Sex	Age (yr)	Sites of diseases	Tyrosinase mRNA	β globin mRNA
Malignant melanoma					
1	M	34	L groin nodes; R supraclavicular node; sternum	−	ND
2	M	47	Abdominal nodes; peritoneum; stomach	+	ND
3	F	52	L groin and nodes; subcutaneous tissue	+	ND
4	M	34	R axilla; R chest wall; brain	−	ND
5	M	65	R foot; R groin; R leg in transit metastases	−	+
6	F	61	L thigh; L groin; lung	+	+
7	M	35	Liver	+	+
*Other cancer**					
8	M	64	Liver	−	ND
9	M	59	Liver	−	+
10	F	48	Pelvis	−	+
11	F	53	Liver	−	+
Healthy					
12	M	36	..	−	+
13	F	30	..	−	+
14	M	34	..	−	+
15	F	39	..	−	+

*Metastatic colerectal carcinoma in subjects 8 and 11, metastatic gastric carcinoma in subject 9, and advanced transitional cell carcinoma of the bladder in subject 10. ND = not done because sample was too small

Results

The human cDNA clones Pmel34[7] (tyrosinase) and JW102[12] (β globin) were used to confirm that PCR with each set of primers gave bands of expected sizes. With outer primers alone and 30 cycles of PCR, tyrosinase mRNA from the equivalent of 50 cells or more was easily detected in the melanoma cell lines SK-mel-19, SK-mel-23, and SK-mel-28. The level of detection was greatly improved by further amplification with nested primers, and RNA equivalent to less than 1 cell could be detected. No tyrosinase mRNA was detected in the cell line M5, one of the two amelanotic lines tested. However, when the method was used to detect SK-mel-23 cells added to 10 ml samples of normal blood before RNA preparation, the sensitivity was lower; the minimum number of melanoma cells detected with outer primers alone was 10^4, even after reamplification for an extra 30 cycles. However, at least 10 SK-mel-23 cells per 10 ml

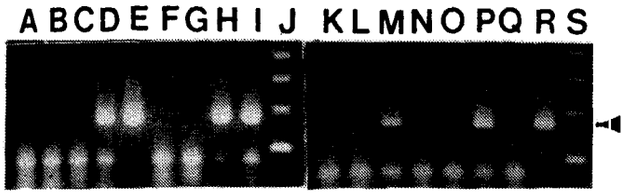


Fig 3—Detection of tyrosinase mRNA in blood of cancer patients.

Lanes D, E, M, and P = positive malignant melanoma patients; lanes C, F, and G and N (same patient) = negative malignant melanoma patients; lanes A, B, K, L, O, and Q = negative controls (2 healthy subjects and 4 other cancers); lanes H, I, and R = positive controls (tyrosinase RNA and cDNA).

blood could be detected after reamplification with nested primers (fig 2).

To determine the level of detection more accurately, individual SK-mel-19 or SK-mel-28 cells were micromanipulated under an inverted microscope with a fine, heat-drawn capillary and different numbers were added to 2 ml normal blood samples. As few as 4 SK-mel-19 cells (1 per 5×10^5 white blood cells) and 1 SK-mel-28 cell (1 per 2×10^6 white blood cells) could be detected.

Tyrosinase mRNA was detected in blood samples from four of the seven patients with malignant melanoma but in none of the healthy controls or patients with other cancers (see table, fig 3). Use of 40 rather than 30 cycles in the second round of PCR gave some false-positive results (not shown).

To confirm that the PCR product was homologous with tyrosinase cDNA, we used restriction enzyme digestion. cDNA from patient 3 was amplified on a large scale and digested with three restriction enzymes (*Hin* F1, *Pvu* 2, and *Dde* 1), each of which has one restriction site within the amplified sequence. All enzymes produced fragments of about the predicted sizes. A small proportion of uncut DNA could be explained by errors introduced into the restriction sites by the *Taq* polymerase.

Discussion

We have developed a non-isotopic method to detect very small numbers of melanoma cells in human blood. The results so far are encouraging in patients with metastatic melanoma and we have not yet had any false-positive results in other subjects with the experimental conditions described. The sensitivity of the test is very high and is greater than that reported for other methods.[2,13] The difference in detection levels between cell lines could reflect different extents of transcription of the tyrosinase gene. Although the number of melanoma cells circulating in patients cannot be directly measured, the encouraging results suggest that this test may be clinically relevant. Further technical developments may increase sensitivity. It is not clear whether the false-positive results seen after 40 cycles in the second round of PCR are due to very low levels of contamination, as reported, for example, by Lo et al,[14] or to "illegitimate transcription"[15] of the tyrosinase gene. Our method takes only 2–3 days and avoids the use of radioactive isotopes.

We believe the test should be evaluated in large numbers of patients. If its usefulness is confirmed, the test may have important applications in malignant melanoma, and perhaps in many other cancers. The diagnosis of a primary cancer may be possible by the demonstration of inappropriate cells in the peripheral blood, thus avoiding the use of more invasive diagnostic methods. Detection of such cells in patients with widespread disease at presentation seems likely. It is not clear whether the test would be useful in patients with apparently localised disease in whom peripheral blood cells may be few or absent. Once the diagnosis has been made, the test could be useful in assessment of prognosis. It seems likely that the presence of peripheral blood cells will increase the probability of metastases. Only extensive evaluation will determine whether the presence of tumour cells in the peripheral blood is always ultimately associated with metastases or whether implantation need not be inevitable. The continued presence of circulating cells after treatment is likely to be associated with a poor outcome. The test might help to identify patients who need systemic therapy after surgery, or those whose drug regimens should be changed or discontinued because of lack of effect, thus sparing unnecessary toxicity.

We propose to explore these ideas further in malignant melanoma and other cancers. They can be readily extended to any cancer for which tissue-specific genes can be identified.

We thank Dr C. Goding for Pmel34; Dr J. Old for JW102; Dr P. Cox for SK-mel-19 and SK-mel-28 cell lines; and Dr I. Hart for SK-mel-23 cell line.

REFERENCES

1. Tarin D, Price JE, Kettlewell MGW, Souter RG, Vass ACR, Crossley B. Clinicopathological observations on metastases in man studied in patients treated with peritoneovenous shunts. *Br Med J* 1984; **288:** 749–51.
2. Moss TJ, Sanders DG. Detection of neuroblastoma cells in blood. *J Clin Oncol* 1990; **8:** 736–40.
3. Saiki RK, Gelfand DH, Stoffel S, et al. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* 1988; **239:** 487–91.
4. Sawyers CL, Timson L, Kawasaki ES, Clark SS, Witte ON, Champlin R. Molecular relapse in chronic myelogenous leukemia patients after bone marrow transplantation detected by polymerase chain reaction. *Proc Natl Acad Sci USA* 1990; **87:** 563–67.
5. Lyons J. Analysis of ras gene point mutations by PCR and oligonucleotide hybridization. In: Innis MA, Gelfand DH, Sninsky JJ, White TJ, eds. PCR protocols: a guide to methods and applications. New York: Academic Press, 1990: 386–91.
6. Tomita Y, Takeda A, Okinaga S, Tagami H, Shibahara S. Human oculocutaneous albinism caused by single base insertion in the tyrosinase gene. *Biochem Biophys Res Commun* 1989; **164:** 990–96.
7. Kwon BS, Haq AK, Pomerantz SH, Halaban R. Isolation and sequence of a cDNA clone for human tyrosinase that maps at the mouse c-albino locus. *Proc Natl Acad Sci USA* 1987; **84:** 7473–77.
8. Lawn RM, Efstratiadis A, O'Connell C, Maniatis T. The nucleotide sequence of the human β-globin gene. *Cell* 1980; **21:** 647–51.
9. Carey TE, Takahashi T, Resnick LA, Oettgen HF, Old LJ. Cell surface antigens of human malignant melanoma: mixed hemadsorption assays for humoral immunity to cultured autologous melanoma cells. *Proc Natl Acad Sci USA* 1976; **73:** 3278–82.
10. Liao SK, Dent PB, McCulloch PB. Characterization of human malignant melanoma cell lines. I. Morphology and growth characteristics in culture. *J Natl Cancer Inst* 1975; **54:** 1037–44.
11. Sambrook J, Fritsch EF, Maniatis T. Molecular cloning, a laboratory manual. Cold Spring Harbor: Cold Spring Harbor Laboratory Press, 1989.
12. Wilson JT, Wilson LB, de Riel JK, et al. Insertion of synthetic copies of human globin genes into bacterial plasmids. *Nucleic Acids Res* 1978; **5:** 563–81.
13. Naito H, Kuzumaki N, Uchino J, et al. Detection of tyrosine hydroxylase mRNA and minimal neuroblastoma cells by the reverse transcription-polymerase chain reaction. *Eur J Cancer* 1991; **27:** 762–65.
14. Lo Y-MD, Patel P, Wainscoat JS, Sampietro M, Gillmer MDG, Fleming KA. Prenatal sex determination by DNA amplification from maternal peripheral blood. *Lancet* 1989; ii: 1363–65.
15. Chelly J, Concordet JP, Kaplan JC, Khan A. Illegitimate transcription: transcription of any gene in any cell type. *Proc Natl Acad Sci USA* 1989; **86:** 2617–21.