

Methods for Genomic Partitioning

Emily H. Turner, Sarah B. Ng,
Deborah A. Nickerson, and Jay Shendure

Department of Genome Sciences, University of Washington, Seattle,
Washington 98195-5065; email: emilyt@u.washington.edu, sarahng@u.washington.edu,
debnick@u.washington.edu, shendure@u.washington.edu

Annu. Rev. Genomics Hum. Genet. 2009.
10:263–84

The *Annual Review of Genomics and Human Genetics*
is online at genom.annualreviews.org

This article's doi:
10.1146/annurev-genom-082908-150112

Copyright © 2009 by Annual Reviews.
All rights reserved

1527-8204/09/0922-0263\$20.00

Key Words

sequencing, genomic enrichment, multiplex analysis, exon capture,
hybrid capture, resequencing

Abstract

The emergence of massively parallel DNA sequencing platforms has made resequencing an affordable approach to study genetic variation. However, the cost of whole genome resequencing remains too high to apply to large numbers of human samples. Genomic partitioning methods allow enrichment for regions of interest at a scale that is matched to the throughput of the new sequencing platforms. We review general categories of methods for genomic partitioning including multiplex PCR, capture-by-circularization, and capture-by-hybridization. Parameters that are relevant to the performance of any given method include multiplexity, specificity, uniformity, input requirements, scalability, and cost. The successful development of genomic partitioning strategies will be key to taking full advantage of massively parallel sequencing, at least until resequencing of complete mammalian genomes becomes widely affordable.

GWAS: genome-wide association studies

PCR: polymerase chain reaction

SNP: single nucleotide polymorphism

INTRODUCTION

The identification of genomic sequence variation underlying specific phenotypes lies at the core of human and model organism genetics. The most comprehensive and straightforward approach for genotype evaluation would be the complete sequencing of each genome of interest. To date, high costs make this largely impractical; instead, alternative approaches are taken to focus on variation within particular genomic intervals. In human genetics, for example, family-based linkage analyses of Mendelian phenotypes have been an enormously successful paradigm. Combined with sequence analysis, these studies have defined the molecular basis of over 2000 clinical disorders (47). The genetic architectures of complex phenotypes, such as quantitative traits and common diseases, have proven more elusive. Databases of common sequence variants underlie the recent wave of successful genome-wide association studies (GWAS) (34). However, the follow-up to each of these GWAS studies frequently involves extensive regional sequencing in a population of phenotyped individuals to search for the causative variants (31). Furthermore, it is increasingly recognized that the heritable contribution to phenotypic variation in complex traits may arise from a combination of common, low-penetrance alleles and rare, high-penetrance alleles (62, 64). Whereas the initial approach to common variants involves GWAS, the identification of rare variants underlying complex traits has relied on the full sequencing of one or more candidate genes in populations of phenotyped individuals (13). The identification of sequence variation is also central to cancer genetics, as the recurrent observation of functional, non-synonymous somatic mutations in tumors is frequently the means by which a candidate gene is implicated in oncogenesis (22).

Here, we define resequencing as the identification of sequence variation in individuals of a species for which a canonical reference genome is available. Conventional resequencing pipelines rely on PCR amplification of each region of interest, followed by bidirectional

Sanger sequencing. Sequences of interest might be sets of exons, full genes, or larger intervals. As PCR amplification from a diploid genome yields a mixture of products derived from each haploid equivalent, heterozygous variants are identified in Sanger sequencing traces as mixed peaks (61), whereas insertion-deletions require a more complex analysis (8). Regions that are too large to be amplified and sequenced as a single PCR product can be “tiled” with multiple pairs of PCR primers, yielding a set of overlapping amplicons that collectively cover the target. Although targeted resequencing pipelines have grown increasingly sophisticated, their costs are directly tied to those of the underlying technologies, namely PCR amplification and Sanger sequencing. Their expense remains a major bottleneck for many studies.

An alternative approach that was applied extensively for the discovery and genotyping of common variants in the human population relies on sequence determination by hybridization to a resequencing microarray with features designed to detect variation (20, 25, 30, 53). Full regions or selected positions are interrogated by amplifying and pooling long-range PCR products as input material for array-based resequencing. However, this approach has not been widely applied for targeted resequencing within individual studies. An exception involves the high-density SNP arrays used in GWAS studies that do not necessarily require a complexity reduction prior to array hybridization (24). However, these are generally focused on positions of common variation that are amenable to genotyping and for which an assay has been designed and validated.

A major development in genomics recently has been the successful proof-of-concept, commercialization, and widespread adoption of several alternative approaches to DNA sequencing (6, 28, 46, 59). A common thread among this crop of “second-generation” or “next-generation” sequencing platforms is that all rely on the concept of massively parallel, iterative sequencing of a dense array of DNA features (45, 57, 58). By the end of 2008, well over 500 such instruments from several vendors

were in use (compared to less than 10 in 2005). These platforms vary widely in terms of performance and cost, but several generalizations can be made. The primary motivator for their adoption is a major reduction in cost-per-base, to several orders of magnitude below that of high-throughput Sanger sequencing pipelines. Currently, the key drawbacks include shorter read-lengths and lower raw accuracies, as compared to conventional Sanger sequencing. An additional consideration is the amount of sequencing analyzed in a single run; the minimal unit of sequencing is on the order of 10^8 bases, as opposed to 10^3 bases with Sanger sequencing.

For genomic resequencing, the utility of short, low-accuracy reads is greatly enhanced by the availability of an assembled reference genome for a given species. Individual reads only need to contain sufficient information to be uniquely mappable to that reference genome in order to be useful. As the canonical genome sequences of most major model organisms and our own species are complete, significant interest has emerged for applying these new sequencing technologies for resequencing. These platforms have effectively rendered “whole genome genetics” feasible, first for bacterial organisms (33, 46, 59), and more recently for model organisms with larger genomes such as *Caenorhabditis elegans* (29, 56). The genome of an individual human, James Watson, was resequenced using the 454 platform at a cost of ~\$1 million (67), and three more human genomes were resequenced using the Illumina/Solexa platform (6, 38, 66) at a reagent cost of ~\$250,000 per genome (6). By comparison, J. Craig Venter’s genome (37) was sequenced with conventional Sanger technology for >\$10 million. The cost of complete human genome resequencing remains high but is dropping rapidly as technical advances with the various platforms are made. It remains an open question, however, how far we are from the so-called \$1000 genome (58), and whether this goal will be achieved by extensions of second-generation platforms or other approaches such as real-time (43) or nanopore sequencing (10).

Complete genome resequencing is not always necessary, however, as investigators are often interested in identifying germline variants or somatic mutations in a particular subset of the genome. Meaningful inferences of genotype-phenotype associations necessarily require the analysis of multiple individuals. For at least the next few years, it is probable (though far from certain) that the routine resequencing of complete human genomes will continue to be prohibitively expensive in the context of studies requiring even modest sample sizes. Studies are therefore likely to initially focus on the sequencing of specific subsets of the human genome across multiple individuals. Examples of genomic subsets that may be highly relevant in the context of a specific study include: (a) positions of common variation, primarily consisting of millions of scattered SNPs; (b) a specific megabase-scale region of the genome that has been implicated in a particular disease through family-based linkage or GWAS analysis; (c) specific candidate genes belonging to a disease pathway; and (d) the full complement of protein-coding DNA sequences (~1% of the human genome). Although “fixed content” genotyping arrays are appropriate for the first example (7), the others are best approached with cost-effective DNA resequencing. However, these subsets generally total to megabases, raising the critical question of how they can be efficiently isolated from non-target sequences. Although performing PCR reactions remains an option, its scale (10^2 to 10^4 bases per reaction unit) is poorly matched to the granularity of next-generation sequencing platforms (10^8 bases per reaction unit). To address this need, significant effort has been dedicated over the past several years to the development of general genomic partitioning methods for the selection and amplification of complex subsets of a mammalian-scale genome (23). It is anticipated that the success of such methods will depend on their synergy with new sequencing technologies, and will facilitate their application to linkage studies, association studies, medical resequencing, cancer resequencing, and other areas.

Genomic partitioning:

methods to enrich a sequence library for specific regions of a genome

Table 1 Summary of genomic partitioning strategies

	Summary	Source of specificity	Multiplex	Scalability	References
Multiplex PCR and related methods	Amplification with multiple primer pairs followed by enrichment steps	Hybridization + enzymatic	10^2 – 10^3	Moderate	(21, 48, 65)
Capture by circularization	Circularization directly on/from genomic DNA followed by enrichment steps	Hybridization + enzymatic	10^4 – 10^5	Excellent	(6a, 15, 16, 35, 55)
Solution-based hybrid selection	Shotgun library hybridized to biotinylated probes in solution	Hybridization	10^4 – 10^5	Excellent	(4, 23a, 50)
Array-based hybrid selection	Shotgun library hybridized to programmable microarray	Hybridization	10^5 – 10^6	Moderate	(2, 32, 51)

Here we review the variety of methods that have been developed in recent years for carrying out genomic partitioning at scales that move well beyond uniplex PCR (Table 1). We begin by describing key performance metrics to consider in evaluating and comparing these strategies. The methods themselves have been divided into three categories. First, we consider methods based around PCR amplification, including pooling of uniplex PCR products and several innovative approaches to multiplex PCR that circumvent its usual limitations. Second, we describe two strategies that rely on target circularization as their primary means of multiplex capture. Finally, we discuss both aqueous-phase and solid-phase methods for capture-by-hybridization.

PERFORMANCE METRICS

Each method described here aims to partition or enrich a genomic DNA sample for sequences derived from all target regions of interest, with the downstream goal of identifying heterozygous or homozygous variants in those targets relative to a reference genome sequence. The fold-enrichment, calculated as the ratio of abundances of the target sequences postenrichment vs pre-enrichment, provides a summary measure of performance. Its theoretical maximum is the ratio of the size of the genome to the aggregate size of the targets. In addition, at least eight relevant performance metrics should

be considered—capture specificity, uniformity, completeness, allelic bias, multiplexity, input requirements, scalability, and cost. In looking at postenrichment sequencing data, capture specificity (related to degree of enrichment) is measured as the fraction of sequence reads that map to targeted regions. Uniformity refers to the relative abundance of individual targets after enrichment, and along with capture specificity is a critical metric in determining the amount of sequencing required to adequately cover the targets. Completeness (related to uniformity) can be defined as the fraction of targets (or target bases) detectably captured by a given strategy. Allelic bias is present if there is a random or systematic nonuniformity in the relative efficiency with which the two alleles of a heterozygous target are captured and amplified. A lack of allelic bias is critical, as preferential capture of one allele over the other will significantly impair heterozygote calling. The remaining metrics—multiplexity, scalability, input requirements, and cost—collectively describe the general practicality of a method. We can define multiplexity as the total number of discontinuous subsequences that can be simultaneously targeted (which is relevant, for example, when targeting large numbers of exons). Input requirements are the amount and quality of genomic DNA required to carry out a given targeting method. Scalability refers to the fact that certain methods may be more amenable than others to high-throughput sample processing.

Multiplex PCR: PCR using more than one pair of primers in the same reaction; multiple reactions carried out in a single volume

Capture specificity: the fraction of molecules in an enriched library that correspond to targets

Uniformity: the relative coverage of targets after genomic partitioning

Allelic bias: unequal capture of two alleles of a heterozygous variant

Scalability: the ease of expansion to studies involving large numbers of samples

Finally, provided that two methods perform adequately well with respect to other performance metrics or when large sample sizes are being considered, cost becomes the critical arbiter. As discussed below, current implementations of genomic partitioning methods suffer from a significant deficit in at least of one of these parameters, underscoring the need for further technological development in this area.

At baseline, for a diploid genome and using a sequencing method where individual reads are derived from single DNA molecules, a reasonably high level of coverage, i.e., the number of times a given target base must be sampled to sensitively and specifically detect variants, is required not only to reduce errors but also to ensure that both alleles are adequately sampled. For example, we simulated a 10-Mb diploid genome with 10,000 heterozygous variants and sampled it in silico at increasing levels of coverage with shotgun microreads (36-bp reads; 1% per-base error rate). The sensitivity of variant detection was assessed using the *maq* tool (39) with a calling threshold set such that the false positive rate was approximately 1 per 100,000 bp. As shown in **Figure 1**, achieving 95% sensitivity for variant detection required at least 15x mean-fold coverage.

One caveat is that the simulation assumed uniformly random sampling of a single contiguous target. However, the required amount of sequence coverage will also depend upon the performance of a given enrichment method. In practice, systematic or random bias in a given capture method results in undersampling of some targets and oversampling of others. The relevance of the former is obvious, as more sequencing would be required to adequately cover undersampled targets, but the latter is also relevant, as oversampled targets essentially waste sequencing capacity. In addition to uniformity, anything short of 100% capture specificity means that not all sequencing reads are derived from the aggregate target. What are the effects of capture specificity and uniformity on the total amount of sequencing required to achieve an effective level of sequence coverage? To address this, we

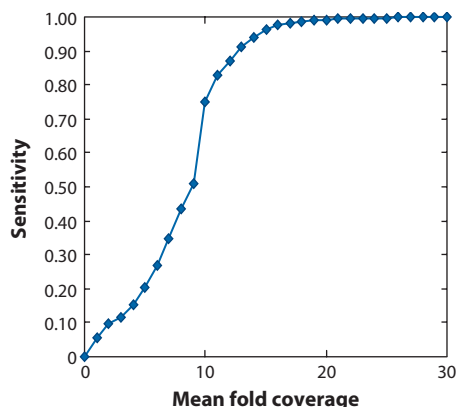


Figure 1

Sequence coverage vs sensitivity of heterozygote detection. Simulation-based estimates of the sensitivity of heterozygote detection at increasing levels of sequence coverage. We simulated a 10-Mb diploid genome as a random sequence containing 10,000 heterozygous variants (1 variant per 1000 bases), and sampled it in silico at increasing levels of coverage with shotgun microreads (36-bp reads; 1% error rate). The sensitivity of variant detection at increasing levels of mean-fold coverage was assessed using the *maq* tool (39) with a calling threshold set such that the false positive rate was approximately 1 per 100,000 bp.

simulated varying levels of capture specificity and uniformity, and estimated how much coverage would be required to achieve at least 15x coverage of at least 80% of targets. As shown in **Figure 2**, both uniformity and capture specificity are of key importance. For example, a method that achieved good uniformity (95% of targets within a 16-fold range), but only 30% capture specificity, would require ~115x mean-fold coverage (blue X in **Figure 2**). A method that had 90% capture specificity, but poor uniformity (95% of targets to within a 128-fold range), would require ~102x mean-fold coverage (black X). However, a method that exhibited both good capture specificity (90%) and good uniformity (95% of targets within a 16-fold range) would only require 38x mean-fold coverage to achieve 15x coverage of 80% of targets (red X). As performance with respect to capture specificity and uniformity largely determines how much sequence coverage will be required,

Coverage: the number of sequence reads covering a given position

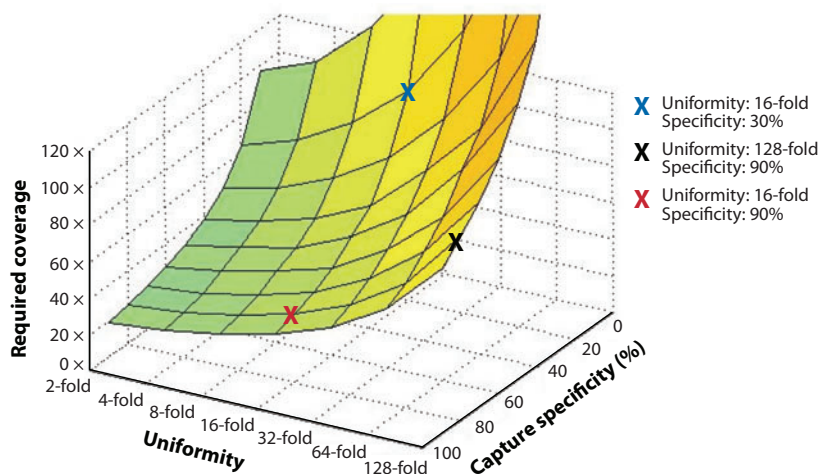


Figure 2

Impact of uniformity and capture specificity on the required amount of sequencing. Here we show simulation-based estimates of the amount of sequencing that will be required to achieve 15x coverage of 80% of targets, at varying levels of uniformity and capture specificity. Uniformity was modeled as a log-normal distribution. X-axis values show the range in which 95% of target abundances are expected to fall (e.g., “16-fold” → 95% of abundances fall within a 16-fold range). Y-axis values show capture specificity (i.e., the fraction of bases that map to targets). For each combination of uniformity and capture specificity, the Z-axis shows the estimated total amount of sequencing (mean-fold coverage relative to the target size) required to achieve 15x coverage of 80% of targets. See text for annotation of colored Xs.

their improvement is critical to reducing downstream sequencing costs associated with a given genomic partitioning method.

UNIPLEX PCR, MULTIPLEX PCR, AND RELATED METHODS

Pooling PCR Products

PCR enables the targeted amplification of regions with lengths compatible with individual Sanger reads, and therefore has served as an effective front-end for Sanger-based resequencing. With second-generation sequencing platforms, uniplex PCR is unlikely to continue in this role. The key difficulty is easily illustrated with an example. In 2006, Sjöblom and colleagues sequenced the full coding regions of all well-annotated human genes in genomic DNA (gDNA) derived from 22 tumors (60). Practically, this involved the use of 135,483 primer pairs and an equivalent number of reactions per tumor to amplify regions corresponding to

120,839 exons. Conventional DNA sequencing of each amplicon totaled to ~21 Mb per tumor genome. Had a second-generation platform been used instead, sequencing costs could have been as low as \$1000 per sample, a price at which the study could likely have been extended to thousands of samples. However, the need to perform, normalize, and pool 135,483 PCR amplifications per sample would have remained as a substantial and expensive impediment.

With much more limited target sets, several groups have recently reported using PCR as a front-end for massively parallel DNA sequencing. For example, Thomas et al. relied on the sensitivity of pyrosequencing (46) to detect low-abundance mutations in five exons of the EGFR gene in lung tumor samples (63). Eleven targets with an average length of 100 bp were amplified by PCR. Amplicons from each of 33 samples were pooled and sequenced on the 454 platform. Within the pooled amplicons from each sample, mutations were detected at frequencies as low as 0.3%. Besides confirming previously

Massively parallel DNA sequencing:

one of several high-throughput, short-read sequencing technologies, generating millions of reads ~30 to ~300 nucleotides in length at a much lower per-base cost than conventional Sanger sequencing

known SNPs, deep pyrosequencing detected indels previously missed by Sanger sequencing. This report demonstrated the strength of massively parallel sequencing over conventional resequencing for detecting low-abundance mutations in tumor samples, where normal tissue may also be present and only a low fraction of tumor cells may carry a clinically significant mutation. This follows from sequencing large numbers of single molecules individually rather than sequencing a mixture of molecules once.

To resequence a 136-kb region of human 8q24 (implicated by GWAS in breast, colon, and prostate cancer), Yeager et al. generated and pooled PCR amplicons, each 2.0 to 5.5 kb in length, tiling over the full region of interest with >150-bp overlaps (69). Primer pairs were carefully designed and tested against several thermal cycling routines to ensure successful amplification of each locus. Pooling of amplicons from each individual at an equimolar ratio was followed by sequencing on the 454 platform. Mean coverage over the target region was 50x, with reasonably good uniformity. At positions with at least 20x coverage, genotypes were called by classification based on the proportion of observations. Overall, across 79 individuals, 442 novel variants (i.e., not in dbSNP) were called. The global genotyping completion rate at polymorphic positions was 93.5%. Comparing SNPs called by these data to SNPs at positions of common variation called by array genotyping, overall concordance per locus was 99.45%.

A study by Craig et al. (14) extended the pooling approach to incorporate sample-identifying barcodes, such that sequencing libraries derived from multiple individuals could be mixed and sequenced simultaneously. Two libraries were prepared for each individual, the first derived from 10 × 5 kb amplicons (covering a 50-kb region), and the second from 14 × 5 kb amplicons (covering a 70-kb region). The 5-kb regions were individually amplified from each of 46 HapMap DNA samples and pooled by sample for library construction. Adaptors containing unique barcode sequences were ligated to fragmented PCR products, providing a sample-specific identifier for each

library molecule. The 6-base barcodes were designed to tolerate at least one sequencing error while maintaining correct index identification. The barcoded libraries derived from all individuals were then pooled and coamplified with a set of universal primers as part of the final steps of library construction. Single-end sequencing was performed on an Illumina Genome Analyzer (5), with individual reads including both the 6-bp barcode and 35 bp of sample-derived sequence. Uniformity across the targeted region for a given individual was quite good, with a 1.5- to 2.0-fold difference between amplicons with the most and fewest reads. However, the representation of each barcode was more variable, with an 11-fold difference between the most and least abundant index (fivefold after optimization). The authors took a Bayesian approach to polymorphism discovery, with a Sanger sequencing–defined set of polymorphisms from the ENCODE project available as a set of true positives. Varying thresholds demonstrated a tradeoff between false positive and false negative rates. At the most lenient calling threshold, the false positive rate was 88% and the false negative rate was 9%. At the most stringent threshold, the false positive rate was 11% and the false negative rate was 91%.

Multiplex PCR and Related Methods

Although compatible with massively parallel sequencing, uniplex PCR with pooling of products does not represent a viable long-term solution for genomic partitioning. A more promising category of methods includes multiplex PCR and its derivatives. First reported in 1988 (12), multiplex PCR has been widely applied in molecular diagnostics, e.g., for pathogen identification and in forensic studies. However, it is difficult to perform on more than a few dozen targets per reaction (18), and even modest levels of multiplexing require significant optimization. The main challenges with multiplex PCR include the formation of primer-dimers, nonuniform amplification of targets, and high rates of mispriming events (17).

dbSNP: a public database of known polymorphisms maintained at NCBI

HapMap: haplotype map of the human genome

Multitemplate PCR:

PCR using a single primer pair to amplify a population of molecules sharing common adaptor sequences. Used in many methods reviewed here

Introducing universal sequences to the 5' ends of each primer pair and performing a two-stage PCR (i.e., with initial cycles directed at amplification from specific loci, followed by switching to the 5' universal sequences for continued amplification via multitemplate PCR) substantially reduces artifacts (11, 40). Several groups have recently reported on novel methods that attempt to circumvent limitations on multiplex PCR and improve its relevance in the context of genomic partitioning applications. Four of these approaches are summarized here.

As the formation of primer-dimers is a major issue for multiplex PCR, the immobilization of primers to a solid substrate can limit primer-dimer formation by physically isolating each primer pair (1, 54). With the MegaPlex PCR method (48), for example, an initial in-solution multiplex PCR is carried out as a relatively nonspecific enrichment for targets of interest. This is followed by several cycles of solid-phase PCR, in which pairs of microbead-immobilized chimeric primers are used. Specifically, the immobilized chimeric primers are designed to append universal sequences to the 5' ends of amplicons. Subsequent solution-phase multitemplate PCR is driven by a single primer pair corresponding to the appended universal sequences. As a proof of concept, coamplification of 50 or 75 targets was carried out by this method and evaluated by both microarray and 454 pyrosequencing. Most targets were recovered within a 100-fold range and less than 10% of products were observed to be primer-dimers artifacts.

A second approach in this area was recently developed by RainDance Technologies, and involves the use of emulsion PCR (68) and microfluidics to compartmentalize individual PCR primer-pairs within the context of a single reaction (K. Brown, personal communication). Specifically, primer pairs corresponding to each target are separately emulsified into water-in-oil droplets, and then the emulsions are pooled to make a primer droplet library. The mixture of emulsions contains many primer pairs, but only a single primer pair within any given droplet. A microfluidics platform is

applied to fuse each primer-containing droplet with droplets from an emulsion containing genomic DNA, dNTPs, and polymerase. The emulsion is then thermocycled as a single PCR reaction, although individual primer pairs are effectively compartmentalized from one another. The key advantage of the approach is that this allows PCR reactions involving different primer pairs to saturate within individual compartments without directly competing with one another, leading to an expectation of high uniformity. In proof-of-concept experiments aimed at 384-plex amplification and with readout on an Illumina Genome Analyzer, it was demonstrated that 383 of 384 targets were detectably amplified. Capture specificity was observed at 50%–80%, and sequence coverage over the targeted exons did not appear to be overly dependent on primer T_m , amplicon GC content, or amplicon length. Overall uniformity was high, with 89% of targeted amplicons within fivefold relative coverage, and 98% within tenfold relative coverage.

Varley & Mitra described a multiplex amplification strategy termed “Nested Patch PCR”, illustrated in **Figure 3** (65). The method relies on two rounds of target-specific enrichment. First, primer pairs are designed against each target, and the mixture of primers used for ten cycles of multiplex PCR. The primers contain uracil bases in place of thymine, such that postamplification exposure to uracil DNA glycosylase, endonuclease VIII, and exonuclease I effectively removes the primer regions from amplicons. For the second round of selection, Nested Patch adaptors are used. These consist of a double-stranded universal segment and a single-stranded overhang that is target specific. Hybridization and ligation of Nested Patch adaptors to primer-depleted amplicons is followed by multitemplate PCR amplification with primers corresponding to the universal sequences. Because this ligation is dependent on sequences immediately internal to the original primers used in the limited multiplex PCR, the Nested Patch adaptors confer additional specificity. The authors also incorporated sample-specific barcodes into the Nested Patch

adaptors such that products from multiple capture reactions could be pooled and sequenced together. Proof-of-concept experiments were performed by coamplifying 94 exons per reaction and sequencing amplification products on the 454 platform. Of the total number of targets, 96% were detectably amplified, with 75% falling within a 50-fold abundance range, and 90% of all reads mapped to one of the 94 targets. Seven of seven predicted polymorphisms were verified by conventional resequencing.

With the Gene-Collector method, a limited multiplex PCR is followed by a circularization step to confer additional specificity. In the recent report describing this method (21), 170 unique primer pairs were designed to target the full coding sequences of 10 human cancer genes; amplicon lengths ranged from 160 to 800 bp. Eight cycles of 170-plex PCR were performed. Then, as illustrated in **Figure 4**, 170 “collector” oligos were introduced, each of which could template the ligase-driven self-circularization of PCR products flanked by one of the 170 intended primer pairings. After exonuclease digestion of uncircularized material (e.g., primer-dimers or nonspecific products, which would not be expected to circularize efficiently), randomly primed rolling circle amplification was carried out, and its products converted to a shotgun sequencing library. An estimated 58% of recovered products were from expected targets. Additionally, enriched products were highly uniform in abundance, with 96% of all targets estimated to be within fourfold of the average abundance by quantitative PCR.

CAPTURE BY CIRCULARIZATION

With the Gene-Collector method, circularization is an effective means of boosting specificity, but is secondary to an initial multiplex PCR enrichment. In this section, we review two promising methods with which the primary capture event involves the multiplex circularization of targets (or copies of targets) via DNA ligase (15, 55). In both approaches, the circularization step serves two purposes:

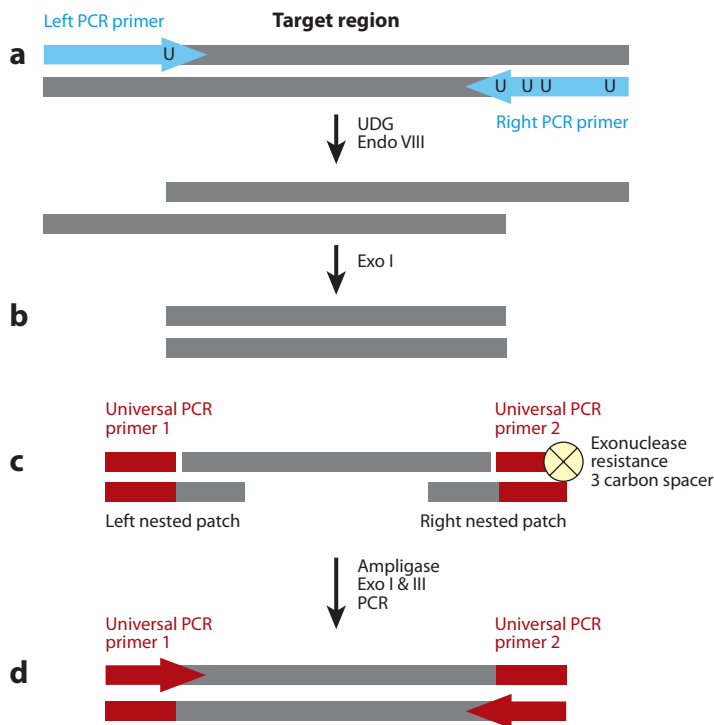


Figure 3

Nested Patch PCR. (a and b) Primer pairs were designed for each target and used to perform an initial enrichment via limited cycles of multiplex PCR. Because the thymines were replaced with uracils in the primers but not the PCR reaction, treatment with UDG and Endo VIII, followed by end-repair, removed the primers while leaving the amplified target sequences intact. (c) Nested Patch primers were designed where each has a single-stranded internal sequence specific to the target amplicon, and an external double-stranded universal adaptor. The ligation of the Nested Patch primers confers additional specificity dependent on the newly exposed flanking sequences of the amplified targets. As the Nested Patch primers are exonuclease resistant at their 5' end, further specificity was achieved by exonuclease digestion. (d) Multitemplate PCR with primers directed at the universal component of the Nested Patch primers allowed further amplification. Image adapted from Varley & Mitra (65) with author's permission.

(a) linking each target to a universal sequence that can subsequently be used to prime PCR amplification of all targets with a single pair of primers; (b) protecting captured targets from the activity of exonucleases, which are used to reduce the concentration of background products that might otherwise interfere with the postcapture multitemplate PCR amplification. The key advantage of these approaches is that, like hybridization-based capture, they may be compatible with a much higher degree of

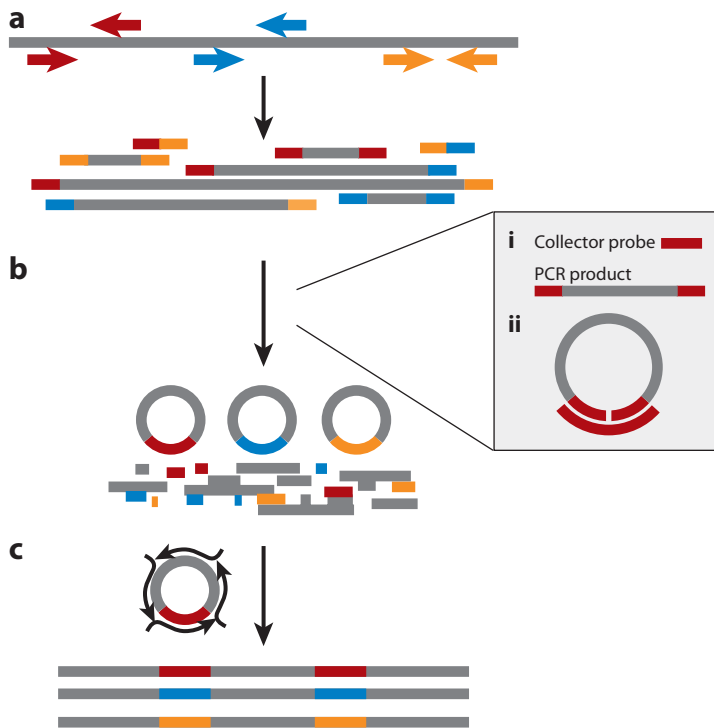


Figure 4

Gene-Collector method. (a) 170 primer pairs (colored bars) were designed against coding sequences in 10 genes, and used to perform multiplex PCR with a limited number of cycles as an initial enrichment. (b) Next, 170 “collector” probes, each containing segments complementary to a specific primer pair, were used to template circularization of intended products by a DNA ligase. Nonspecific products involving incorrect primer pairings and primer-dimers were not efficiently circularized, permitting removal by exonuclease digestion. (c) Randomly primed rolling circle amplification yielded concatamers that were converted to a shotgun sequencing library. Image adapted from Fredriksson et al. (21) with author’s permission.

multiplexing than methods that rely on multiplex PCR for initial enrichment. They also exhibit much greater capture specificity than hybridization-based capture (discussed below), and scale well as they are aqueous-phase reactions.

Selective Circularization

This method, developed by Dahl and colleagues (15, 16), achieves capture via the ligase-driven circularization of targeted restriction fragments from genomic DNA, with a twist that one end of a given restriction fragment

can be effectively trimmed to a desired position by endonucleolytic cleavage of an invasive flap structure (Figure 5). The capture reaction includes the following components: (a) a mixture of “selector” oligonucleotides (one per target), each consisting of a common sequence flanked by target-specific sequences; (b) a “vector” oligonucleotide, the reverse complement of the common sequence internal to each selector oligonucleotide; (c) genomic DNA restriction digested with one or several enzymes (or a mix of genomic DNA that has been subjected to different digests); and (d) Taq DNA ligase and Taq DNA polymerase. The boundaries of selected targets depends in part on the expected pattern(s) of restriction digestion of the reference genome. For a subset of targets, the selector oligo is designed to capture one strand of an expected restriction fragment (RF) in its entirety, via hybridization of the RF to the 5’ and 3’ single-stranded, target-specific overhangs of a given selector-vector hybrid molecule. This is followed by ligation at both ends via Taq ligase to yield a circular molecule that is essentially the full RF and the common vector oligo. For a second subset of targets, one overhang of each selector is again designed to correspond to the 3’ end of targeted RFs, but the other overhang is designed to hybridize to a region internal to the 5’ end of targeted RFs. After hybridization of RFs to selector-vector hybrids, the endonucleolytic activity of Taq polymerase recognizes and cleaves the resulting branched structure (44). Ligation at both ends again yields a circular molecule, here with a partial RF and the common vector oligo. After the capture reaction, exonucleases are used to remove uncircularized material (e.g., incorrect ligations that may not necessarily have resulted in circularization). A single multitemplate PCR with a universal primer pair directed at the common vector sequence is then used to amplify all circularized targets in parallel. In a recent report using this method in the context of cancer resequencing, 177 exons from 10 genes were targeted via 425 selectors, with individual targets ranging in size from 138 bp to 238 bp and totaling to ~49 kb of genomic sequence (16).

RF: restriction fragment

Five restriction enzymes were required to adequately target the desired sequences, with some overlap between targets of different selectors in this panel. Products of capture from six genomic DNA samples were sequenced by 454 pyrosequencing. The length of captured amplicons was approximately matched to the read-length of this sequencing platform, such that additional steps to construct a shotgun sequencing library were not required. A high specificity is expected [conferred by the hybridization of the RF and selector, the activity of *Taq* polymerase (flap cleavage), *Taq* ligase (at both sides of the selector), and the exonucleases] and was observed as ~90% of all reads mapped to targets. As with other enzymatic methods, nonuniformity was significant, with the large majority of reads falling within a 100-fold range. However, the nonuniformity is highly systematic and reproducible across samples. It therefore can be substantially mitigated through empirical adjustment of the concentrations of individual selectors, which allows the majority of amplicons to fall within a tenfold range (H. Ji, personal communication).

Gapped Molecular Inversion Probes

A padlock probe, or molecular inversion probe (MIP), is a single-stranded oligonucleotide consisting of two target-complementary arms separated by a linker sequence (36, 49). The target-complementary arms are designed such that the 5' and 3' ends of the padlock probe are immediately adjacent when hybridized to the expected target sequence. If the 5' end is phosphorylated, DNA ligase will join the two ends, resulting in a circularized padlock probe that is catenated to the target. Following circularization, exonuclease digestion reduces the concentration of uncircularized probe species by several orders of magnitude. PCR or rolling circle amplification (RCA) with primer(s) directed against the linker sequence can be used to amplify circularized probes in a multitemplate PCR (26). For detection of the presence/absence of a target sequence, the reaction is sensitive in that only a single hybridization event is required,

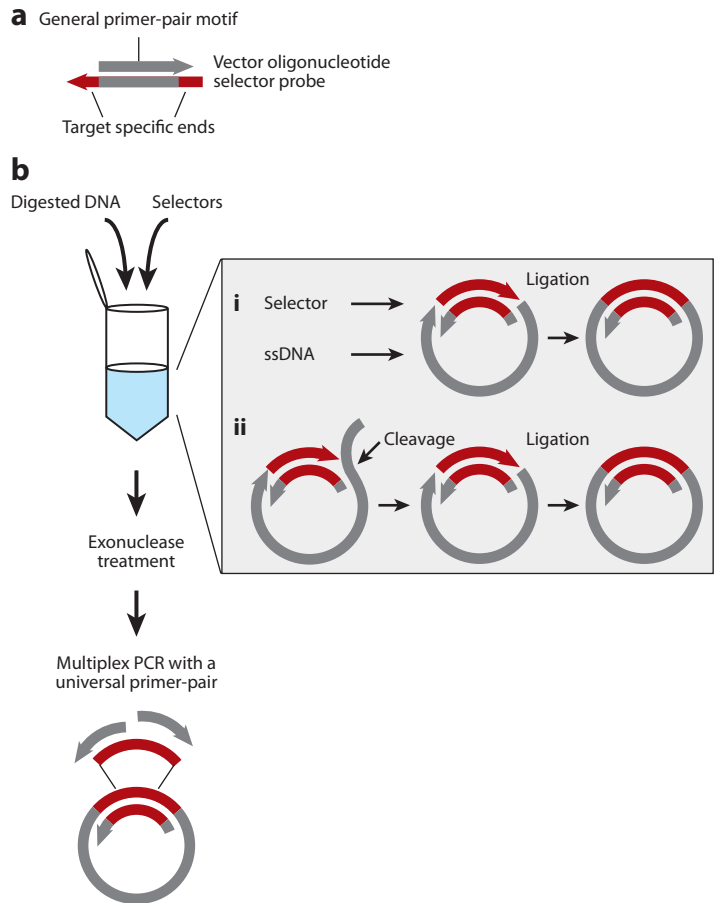


Figure 5

Capture by selective circularization. (a) Selector probes are generated with a shared internal sequence, flanked by target-specific ends. The vector oligo sequence is complimentary to the shared internal sequence. (b) Single-stranded DNA targets, obtained by restriction digestion and denaturation of genomic DNA, are circularized with the selectors. With some designs (i), the selector probe hybridizes to sequences at the 5' and 3' ends of targeted restriction fragments and acts as a template for ligase-driven circularization of the target to the vector oligo. For other designs (ii), the selector probe hybridizes to the 3' end of the target and an internal sequence, such that there is a 5' flap. Invasive cleavage by *Taq* polymerase removes the flap and circularization is completed with ligase. Following circularization, linear DNA is removed by exonuclease, and multitemplate PCR is carried out via a single primer pair directed at the common vector sequence. Image adapted from Dahl et al. (15) with authors' permission.

but is also highly specific, as the ends of the probe must be brought into immediate proximity with no mismatches near the join site in order to be efficiently ligated. The rapid kinetics of the intramolecular padlock reaction favor

MIP: molecular inversion probe

RCA: rolling circle amplification

target-probe hybridization over probe-probe interactions; as a result, the padlock probe reaction can be highly multiplexed. The initial method was adopted for genotyping, either with allele-specific padlock probes (3, 19, 26), or by the introduction of a single base-pair gap between the targets of the arms, such that a single-base gap-fill (with polymerase) was required in addition to ligation (26, 41). Extending the latter concept, Hardenbol et al. demonstrated that over 10,000 SNPs could be genotyped in parallel via a padlock probe scheme requiring single base gap-fills at interrogated positions and four-color readout on microarrays (27).

To adapt this approach for genomic partitioning, Shendure and colleagues explored

MIP designs where the targeting arms of each MIP flanked full exons (**Figure 6**), rather than single nucleotide variants (55). In the experiment described in Porreca et al. (55), a set of 55,000 MIPs was designed to target 55,000 exons as well as two base-pairs adjacent to the splice junctions, with an aggregate target-size of 6.7 Mb. To mitigate the high cost of column-based oligonucleotide synthesis at the desired scale, the 55,000 required MIPs were obtained as a complex mixture of 100-mers by synthesis on and release from the surface of an Agilent microarray. After amplification via 15-bp universal sequences at each end, the 100-mers were converted to 70-mer MIPs through a series of restriction digestions. Each 70-mer MIP consisted of unique 20-bp targeting sequences flanking a common 30-bp linker. The individual targets ranged in length from 60 to 191 bp. With the amplification step, we estimate that the yield of one programmable array is sufficient to support at least 1000 independent capture reactions. Following hybridization to genomic DNA, gap-filling and circularization, and exonuclease treatment, capture products were rolling circle amplified, converted into shotgun sequencing library, and sequenced on the Illumina Genome Analyzer. Analysis of the resulting data demonstrated that: (a) specificity was high, as ~98% of reads that could be confidently mapped to a location overlapping with one of the 55,000 targets; (b) completeness and specificity were poor, as only ~10,000 of the 55,000 targets were detectably captured, and the abundance with which individual targets were observed ranged over several logs; and (c) genotyping accuracy was high at homozygous positions, but low at heterozygous positions, likely secondary to stochastic effects with poor capture efficiency.

We have subsequently observed that simple optimizations markedly improve the performance of this strategy (64a). These include: (a) increasing hybridization and gap-fill time; (b) increasing MIP and ligase concentration; (c) replacement of rolling circle amplification steps with multitemplate PCR with primers directed at the linker; and (d) direct sequencing

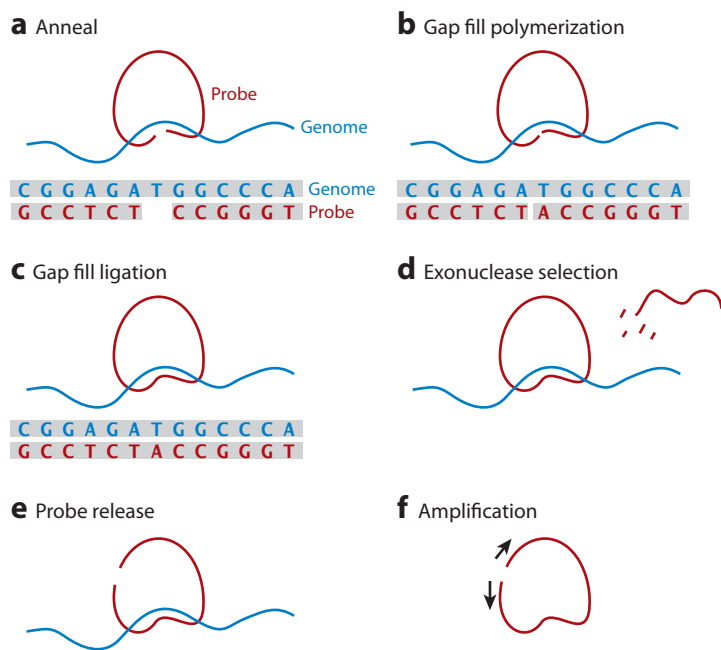


Figure 6

Gapped molecular inversion probes. (a) Probes are designed with a target-specific sequence at the ends, and an internal sequence that is common to all MIPs. Probes hybridize to single-stranded genomic DNA, leaving a gap over the target region. The gap can range from a single nucleotide for SNP genotyping, as in References 26, 27, to several hundred nucleotides for exon capture, as in References 35, 55. (b) Polymerase is used to fill in nucleotides over the gap. (c) Ligase completes circularization of the molecular inversion probe. (d) Exonuclease treatment removes linear DNA. (e) In some versions of this protocol, the probe is linearized. (f) Multitemplate PCR of the probes is carried out with universal primers directed at the common backbone. Image adapted with permission from MacMillan Publishers Ltd: *Nature Biotechnology* 2003 (26).

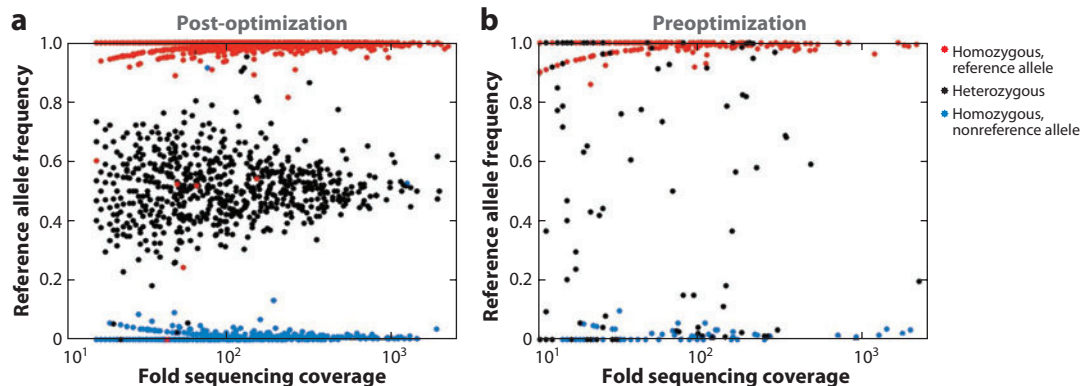


Figure 7

Reduced allelic bias after optimization of MIP capture conditions. (a) Post-optimization. MIP-based capture directed at 55,000 exons (6.7-Mb aggregate target), followed by shotgun Solexa sequencing was performed on a HapMap sample with an optimized protocol in which over 50,000 of the targets were detectably captured. Protocol optimizations are described in the main text. Shown here are 3733 positions with $\geq 15\times$ coverage, *maq* quality scores ≥ 70 (39), and a known HapMap genotype for this individual. Fold-sequencing coverage (x-axis, log scale) is plotted against the frequency of the reference genome allele in our resequencing data (y-axis). Colors indicate HapMap genotypes. Red, homozygous, reference allele; black, heterozygous; blue, homozygous, nonreference allele. (b) Preoptimization. For comparison, we also show results generated using the same set of MIPs and the same genomic DNA sample, prior to the optimization of reaction conditions. These preoptimization data are also presented as **Figure 4** in Porreca et al. (55).

of captured products. With these modifications to the protocol and the same set of 55,000 targeting oligos, we observed that high target specificity was maintained, while major improvements were made with respect to uniformity. Approximately 50,000 of 55,000 (91%) targets were detectably captured (vs 18% with the original protocol), and 33% of targeted bases were captured to abundances within a tenfold range. The improved uniformity was accompanied by improved performance with respect to allelic bias, as shown graphically in **Figure 7**. With the streamlined direct sequencing protocol, using probes with more stringent design constraints, we performed exon capture on 13,000 targets in 16 HapMap individuals. Specificity remained high, with $>99\%$ of reads mapping to one of the targets; completeness was 98%. Uniformity was improved significantly, with 58% of targets captured within a 10-fold range, and 88% of targets captured within a 100-fold range. Variant calling to HapMap genotypes was also highly reproducible, with high concordance for homozygous (99.8%) and heterozygous (99.3%) genotypes.

Advantages of capture-by-circularization with gapped MIPs include the following: (a) the reaction is highly specific, with $>98\%$ of mappable reads derived from targets; (b) the reaction multiplexes to at least 300,000 independent targets, and higher complexities are likely possible (E.T., S.N. & J.S., unpublished observations); (c) like multiplex PCR and other capture-by-circularization methods, the reaction is performed directly on genomic DNA rather than on a shotgun library, making it compatible with lower amounts of starting material; and (d) captured amplicons can be directly sequenced, bypassing the need for constructing shotgun libraries altogether, and making automated high-throughput capture feasible. The main limitation of this technique remains uniformity, which although greatly improved, compares poorly with capture-by-hybridization methods for enrichment. The challenge of poor uniformity could potentially be overcome in several ways. Because the nonuniformity is systematically reproducible with respect to individual targets, MIPs could potentially be grouped into sets based on

similar capture efficiencies, or adjusted to normalizing concentrations in the same reaction. Also, Krishnakumar et al. (35) recently described a modified probe generation procedure that lengthened the linker backbone, and demonstrated that these long backbone probes yielded much greater capture uniformity when tested on a limited set of targets, and were capable of capturing targets up to 500 bp long.

CAPTURE-BY-HYBRIDIZATION

Another approach to genomic partitioning is to rely on the hybridization of shotgun genomic DNA libraries to a complex mixture of capture probes, which may potentially be in solution or tethered to a solid substrate such as a microbead or a glass surface. Advantages of hybridization-based capture include the possibility of much greater degrees of multiplexing without significant interference, and potentially greater tolerance for polymorphisms that overlap with the capture probes themselves compared to protocols based on extension or ligation by an enzyme. An unsurprising disadvantage is that hybridization-based capture tends to be much less specific than enzymatic capture methods, due to cross-hybridization of similar sequences, for example. However, as described below, this is offset by the fact that hybridization-based methods have generally resulted in significantly better uniformity than enzymatic methods. In this section, we review strategies that have been described for both solution-based and array-based capture-by-hybridization.

Solution-Phase Hybrid Selection

Bashiardes and colleagues described a modified version of genomic DNA-based cDNA selection protocols (42, 52) directed at performing hybridization-based targeted capture of shotgun library fragments corresponding to BAC-sized genomic regions (4). The method takes advantage of widely available BAC clone libraries as a source material for the capturing agent. In brief, an adaptor-flanked shotgun library is generated from genomic DNA of

interest, and the library is hybridized in solution to biotinylated DNA that is derived from a BAC corresponding to the region of interest. Streptavidin beads are used to pull down target-probe hybrids, followed by washing to remove nonspecifically bound molecules. Captured targets are then eluted and amplified with primers directed at the common adaptors prior to sequencing. In the described protocol, an adaptor-flanked shotgun library was generated from pooled DNA from 14 individuals via restriction digestion followed by linker ligation. Two rounds of in-solution selection were performed against the biotinylated BAC (~70 h each), with PCR amplification of the captured material after each round of selection. The target region (a ~150-kb region corresponding to the BAC) was enriched 1000-fold after the first round and 10,000-fold (total) after the second round. On Sanger sequencing of 2119 clones generated from recovered material after two rounds of selection, 52% mapped unambiguously to the 150-kb target sequence. Nonrepetitive regions of the target were covered ~threefold, and both established and novel variants were successfully detected. Within the pool, 69 previously known variants and 100 putative novel variants were discovered, as well as several small indels.

In solution hybridization-based capture has also been successfully applied to the enrichment of target DNA sequences from ancient sources (50). The challenge in this context is that contamination of ancient remains by microbial DNA prevents acquisition of a pure sample for shotgun sequencing. To enrich for specific Neanderthal sequences from contaminated sources, Noonan et al. evaluated direct hybrid selection from metagenomic ancient DNA libraries using capture probes derived from modern human DNA (**Figure 8**). Specifically, biotinylated capture probes were generated by PCR amplification from human genomic DNA of regions corresponding to targets. These were hybridized against PCR amplicons derived from metagenomic ancient DNA libraries followed by pulldown of heteroduplexes via streptavidin-coated beads.

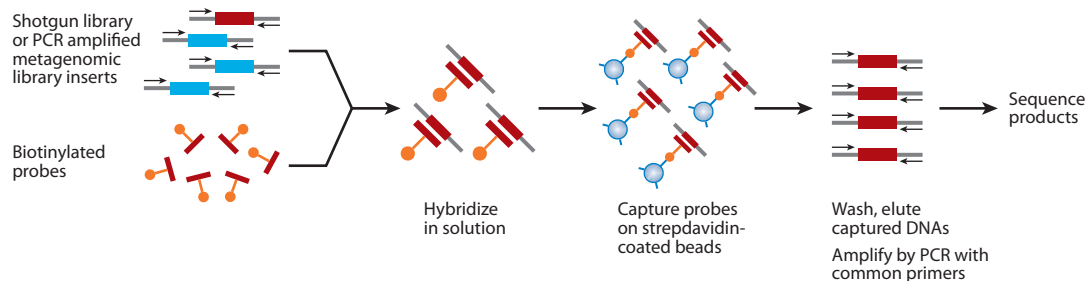


Figure 8

In solution hybrid selection. Target DNA is prepared as an in vitro shotgun library, with common adaptors flanking genomic DNA fragments. The library is hybridized in solution to a set of biotinylated probes. After hybridization, biotinylated probes are captured with streptavidin beads. Beads are washed to remove any nonspecific, unbound library molecules. Multitemplate PCR with primers directed at the common adaptors is used to amplify eluted target molecules before high-throughput sequencing. Adapted from Noonan et al. (50). Images reprinted with permission from AASS.

Parallel capture of 29 of 35 human targets was demonstrated, with the caveat that these sequences were already known to be present in the Neanderthal library via sequencing of the library prior to enrichment. Remarkably, the authors also demonstrated capture of 5 of 96 targets using human capture probes corresponding to highly conserved regions against a Pleistocene cave bear metagenomic library. In this case, the targets were not known to be present prior to enrichment and sequencing.

Andreas Gnirke, Chad Nusbaum, and colleagues at the Broad Institute have recently developed a method for in solution hybrid selection that relies on long RNA molecules as capture probes (23a). First, as with the gapped MIP approach described above, a library of DNA capture probe precursors is generated by synthesis and release from a programmable microarray. Each of these includes 170 bp of target-specific sequence and common flanking sequence. A T7 promoter is appended, such that in vitro transcription can drive synthesis of large amounts of biotinylated RNA capture probes (aka baits) in the same orientation. The mixture of RNA baits is hybridized at high concentration in solution against a shotgun genomic DNA library. After hybridization, streptavidin-coated beads are used to pull down RNA-DNA hybrids, followed by amplification from universal primers and sequencing. Key strengths

of this approach include the following: (a) because the RNA baits are single-stranded and present in only one orientation, a high concentration and molar excess can drive the kinetics of hybridization; (b) relatively low amounts of input genomic DNA (0.5–3 µg) are sufficient; (c) the reaction is solution-based and therefore more scalable, i.e., automatable, than solid-phase array-based hybridization methods; (d) allelic bias may be reduced with long capture probes relative to short probes; (e) the RNA baits can be prepared in large batches that can be quality controlled for use in production-scale settings; (f) the approach can be applied to many short, discontinuous targets or long contiguous regions; and (g) high specificity has been demonstrated with 85%–90% of post-enrichment sequences overlapping with targets. Their initial experiments made use of as many as 22,000 targeting oligos per reaction, though higher levels of multiplexing will likely be feasible. This approach has been licensed to Agilent for development as a commercial product.

Array-Based Hybrid Selection

In 2007, several research groups reported on genomic partitioning methods that made use of programmable oligonucleotide arrays (2, 32, 51, 55). These arrays contain customizable sets of long, single-stranded probes (60 to 200 bp), and are available from vendors such as

Nimblegen and Agilent at complexities up to several million features. The availability of programmable high-density microarrays as relatively affordable consumables provides the possibility of cost-effectively evaluating different capture methods for tens to hundreds of thousands of distinct targets per reaction. Several groups have developed approaches that use complex oligonucleotide libraries released from the surface of programmable microarrays for use in a solution-phase capture reaction (including the gapped molecular inversion probe (55) and the RNA-DNA hybrid selection methods described above). However, in this section we focus on reports from several groups that apply the programmable microarray itself as a selective substrate for solid-phase capture-by-hybridization.

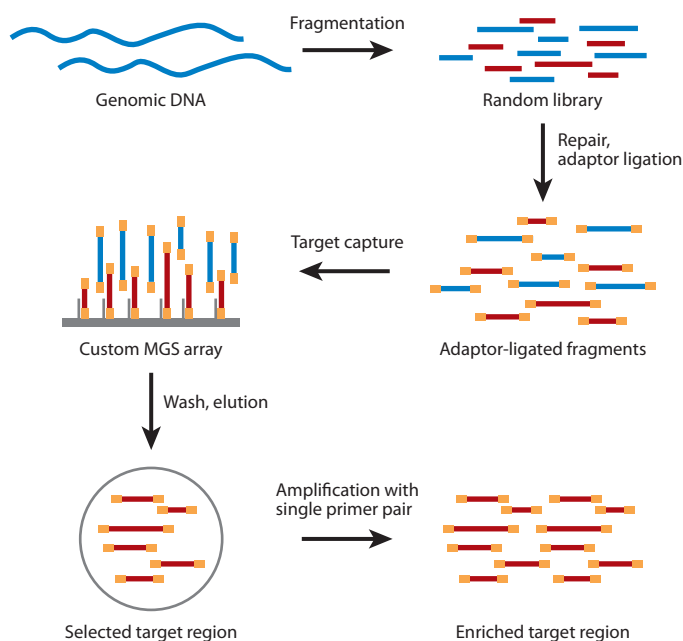


Figure 9

On array hybrid selection. In vitro shotgun libraries are generated from genomic DNA, with common adaptors flanking each fragment. The library is hybridized to oligos tethered on a high-density programmable microarray. Unbound molecules are washed from the array, followed by heat-based elution of specifically hybridized material. Multitemplate PCR with primers directed at the common adaptors is used to amplify eluted target molecules before high-throughput sequencing. Image adapted with permission from Macmillan Publishers Ltd: *Nature Methods* 2007 (51).

Three initial reports of hybridization capture on programmable microarrays (2, 32, 51) involved the use of Nimblegen products and share similar protocols (schematized in **Figure 9**). In brief, $\sim 20 \mu\text{g}$ of genomic DNA is sheared to form a complex mixture of double-stranded fragments (i.e., 250 to 1000 bp in length). Following end-repair, common adaptors are ligated that contain universal priming sequences. The Nimblegen programmable microarrays contain $\sim 385,000$ surface-tethered, single-stranded long oligos (>60 bp) with sequences designed from the reference human genome to tile region(s) of interest at high density (i.e., 1 to 10 bp spacing) for isothermal hybridization, while excluding nonunique or repetitive sequences from consideration. After hybridization for ~ 65 h at 42°C , and a set of wash steps, heat-based elution at 95°C is carried out to recover specifically hybridized material. Universal primers corresponding to the common adaptors are used for PCR amplification, after which the target-enriched shotgun library can be sequenced.

Albert et al. (2) designed and evaluated several capture arrays, one focused on capturing 6726 discontinuous exons and adjacent sequences from 660 genes (total target size of 5 Mb), and the remainder focused on contiguous intervals of varying sizes at the BRCA1 locus (200 kb, 500 kb, 1 Mb, 2 Mb, and 5 Mb) with the same array format but different densities of probe spacing. With three replicates of the exon-focused array, sequencing data (65 Mb to 115 Mb of sequence generated from post-enrichment libraries by 454 sequencing) showed relatively consistent performance, with 65% to 77% of reads mapping to targets, and 93% to 96% of targets overlapped by at least 1 read. For capture directed at a contiguous region (200 kb to 5 Mb), the fraction of reads mapping to the target appeared to be correlated with the size of the target, i.e., 14% for a 200-kb target vs 64% for a 5-Mb target. However, given that the 200-kb target is 25-fold smaller than the 5-Mb target, the calculated fold-enrichment for the 200-kb target is actually better.

Okou et al. (51) targeted sequences near the FMR1 locus—a 50-kb region in one array design, and 304 kb of unique sequence with another array design. Capture and recovery were carried out similarly to Albert et al. (2), but the resequencing was performed with an independent microarray (70), rather than with a massively parallel sequencing platform. Enrichment was instead measured by qPCR at ~1000-fold. The array-based resequencing results were quite good, with a call-rate of 99.1% over 20 replicates, and an accuracy of 99.81% at segregating, genotyped sites in HapMap samples.

Hodges et al. (32) targeted the full set of well-annotated human protein-coding sequences and adjacent splice sites using a set of seven Nimblegen 385K microarrays with 60- to 90-bp probes. Each of these arrays targeted 6 to 8 Mb of coding sequence, with a tiling density of roughly one probe per 20 bp. Approximately 20 µg of genomic DNA was used as input for each of the seven array-based capture reactions. Eluted material was sequenced at high throughput with either the 454 pyrosequencing platform or the Illumina Genome Analyzer. In initial experiments, shotgun genomic libraries were targeted with relatively large fragment sizes (500 to 600 bp). Although successful, in that 55% to 85% of reads mapped to targets, a substantial fraction of these were adjacent to targets rather than within a target. Additional experiments were performed that used shotgun libraries sheared to a smaller size range (100 to 200 bp). Although this resulted in a substantial reduction in capture specificity (to ~29%), a much greater fraction of sequenced bases were now within target, rather than adjacent to target. Given the high input material requirements, the group primarily used genomic DNA that had been subjected to whole genome amplification (WGA) prior to hybridization, but a non-WGA sample was also evaluated against one of the same array designs and performed similarly.

Based on the above reports as well as our own experiences, we can identify key advantages

and disadvantages of hybridization on programmable microarrays as a strategy for genomic partitioning.

Some of these examples relate to the use of hybridization as the capture method, and others to the use of arrays. A first advantage is that there is greater flexibility (than the BAC-based approach, for example), in that either a long contiguous region or many short, discontinuous regions can be targeted with the same level of effort. Moreover, the aggregate size of the target(s) of a programmable array can range from hundreds of kilobases to tens of megabases with the same protocol, simply by varying the probe spacing. Another advantage is that as high-density programmable microarrays are available for approximately \$500 to \$1000, there is a relatively low cost to evaluating any given array design, or to optimizing an array by iterating through new designs. Finally, hybridization-based methods including on-array capture have generally been observed to perform much better with respect to uniformity of target capture, a critical parameter in determining the overall amount of sequencing that will be required for variant discovery across the full target.

A potential disadvantage is that from a practical perspective, array-based genomic partitioning may be more difficult to scale to hundreds or thousands of samples than an aqueous-phase reaction. This potential difficulty is somewhat mitigated by the substantial throughputs that have been achieved for array processing for high-throughput SNP genotyping in the context of GWAS studies. A related disadvantage is that although there is a relatively low cost to optimize an array design, the cost of the arrays is still quite substantial when one is considering scaling to many samples. Moreover, the array cost does not necessarily scale linearly with the size of the aggregate target. In contrast, in-solution capture reactions might make use of relatively small amounts of targeting oligos, or libraries of targeting oligos can potentially be amplified, reducing per-sample costs. Potential workarounds for this concern

WGA: whole genome amplification

include multiplexing more than one sample on the same array via barcoded libraries (S.N., E.T. & J.S., unpublished observations), or reusing arrays (51), though this carries a nontrivial risk for contaminating new samples with old samples on the array. A third disadvantage is that methods relying on hybridization alone generally have lower capture specificity compared to other approaches, which increases the overall sequencing requirements and offsets the advantage of greater target uniformity. The origins of nonspecific capture have not been examined in detail in this context, but possible sources include straightforward cross-hybridization as well as “daisy-chaining” between adaptor sequences that are present in both orientations during array hybridization (i.e., one library molecule hybridizes specifically to the array, then another library molecule hybridizes to it via complementarity between the adaptor sequences, etc.). The extent of nonspecific capture may be correlated with the composition and size of the aggregate target. For example, targeting of smaller genomic subsets may result in significantly lower capture specificity than larger genomic subsets. In Albert et al., for example, 14% of postenrichment reads mapped to a 200-kb target, while 64% of reads mapped to a 5-Mb target. Target sizes between 200 kb and 5 Mb exhibited intermediate capture specificities [Table S3 in (2)]. Other aspects of library preparation can significantly affect performance of array-based enrichment. For example, in Hodges et al. (32), longer genomic fragments were associated with increased capture specificity, but this increase came at a cost of more bases that were target-adjacent rather than target-within sequences. The relevance of this depends on whether one is targeting long, contiguous regions or short, discontinuous exons. Allelic bias is a theoretical concern if multiple variants directly overlap a probe and reduce hybridization efficiency; this has generally not been observed. One related point is that large deletions overlapping the probes may be difficult to detect when sequencing rather than array hybridization (51) is used as the readout.

SUMMARY AND FUTURE DIRECTIONS

Massively parallel DNA sequencing provides a cost-effective means of identifying genetic variation. However, for at least the time being, most investigators interested in taking advantage of these technologies for human resequencing will use a genomic partitioning method as a “front-end” to focus their finite resources. In the past several years, much progress has been made in the development of diverse approaches that meet this need. We cannot offer any clear-cut answer on which of these strategies is the best, in part because they are continuing to evolve and improve. There are tradeoffs inherent to the selection of any given method for genomic partitioning. For example, capture-by-circularization methods are highly specific but relatively less uniform, whereas capture-by-hybridization methods have shown higher uniformity but generally lower specificity. Improvements in performance with respect to these and other metrics (e.g., cost, scalability) and less tangible aspects (e.g., flexibility, access) may be important determinants of the extent to which individual strategies are adopted. One’s choice of method may also depend on the scale of a given study, both with respect to the size of the aggregate target and the number of samples to be analyzed. For example, multiplex PCR and derivative methods, combined with sequence-based barcodes, may be most useful in the context of studies focused on a small number of targets in a large number of individuals. Array-based capture-by-hybridization may be most appropriate for a large aggregate target with limited numbers of samples. For large aggregate targets and many samples, solution-based methods that multiplex and scale well (e.g., capture-by-circularization or solution-based hybrid selection) may be the best choice. Cost is difficult to estimate for each method and may also depend on the scale of application. For example, oligo libraries released from programmable microarrays are expensive and available from only a small number of vendors, but are very cost-effective in this

context when amortized over a large number of samples.

A closing point is that these methods may well be rendered obsolete before they achieve widespread use.

With second-generation sequencing platforms, costs are dropping quickly while read lengths and accuracies continue to improve. The practical implementation of real-time (43) or nanopore sequencing (10) methods may be imminent. If, for example, the all-inclusive costs of whole genome human resequencing were to drop to less than \$1000, the demand for genome partitioning methods might be expected to

significantly wane as increasing numbers of investigators can afford whole genome resequencing of all samples of interest. Nonetheless, these methods may continue to be useful if investigators choose to continue to sequence genomic subsets in larger numbers of samples (i.e., 100 complete genomes for \$100,000 vs 1% of 10,000 genomes for the same cost). Another example where genomic partitioning might continue to be useful is tumor resequencing, where 1000x coverage of all coding sequences (1% of the human genome) in a heterogeneous sample might be more informative than 10x coverage of the whole genome.

DISCLOSURE STATEMENT

The authors are not aware of any biases that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The authors acknowledge support, in part, by grants from the U.S. National Institutes of Health (NIH) National Heart Lung and Blood Institute (RO1 HL094976 to D.A.N. and J.S.) and the NIH National Human Genome Research Institute (NHGRI) (R21 HG004749 to J.S.). E.H.T. is supported by a training fellowship from the NIH NHGRI (T32 HG00035). S.B.N. is supported by the Agency for Science, Technology and Research, Singapore.

LITERATURE CITED

1. Adessi C, Matton G, Ayala G, Turcatti G, Mermod JJ, et al. 2000. Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. *Nucleic Acids Res.* 28:E87
2. Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, et al. 2007. Direct selection of human genomic loci by microarray hybridization. *Nat. Methods* 4:903–5
3. Antson DO, Isaksson A, Landegren U, Nilsson M. 2000. PCR-generated padlock probes detect single nucleotide variation in genomic DNA. *Nucleic Acids Res.* 28:E58
4. Bashiardes S, Veile R, Helms C, Mardis ER, Bowcock AM, Lovett M. 2005. Direct genomic selection. *Nat. Methods* 2:63–69
5. Bentley DR. 2006. Whole-genome resequencing. *Curr. Opin. Genet. Dev.* 16:545–52
6. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456:53–59
7. Bhangale TR, Rieder MJ, Nickerson DA. 2008. Estimating coverage and power for genetic association studies using near-complete variation data. *Nat. Genet.* 40:841–43
8. Bhangale TR, Stephens M, Nickerson DA. 2006. Automating resequencing-based detection of insertion-deletion polymorphisms. *Nat. Genet.* 38:1457–62
9. Bitinaite J, Rubino M, Varma KH, Schildkraut I, Vaisvila R, Vaiskunaite R. 2007. USER friendly DNA engineering and cloning method by uracil excision. *Nucleic Acids Res.* 35:1992–2002
10. Branton D, Deamer DW, Marziali A, Bayley H, Benner SA, et al. 2008. The potential and challenges of nanopore sequencing. *Nat. Biotechnol.* 26:1146–53

11. Brownie J, Shawcross S, Theaker J, Whitcombe D, Ferrie R, et al. 1997. The elimination of primer-dimer accumulation in PCR. *Nucleic Acids Res.* 25:3235–41
12. Chamberlain JS, Gibbs RA, Rainer JE, Nguyen PN, Thomas C. 1988. Deletion screening of the Duchenne muscular dystrophy locus via multiplex DNA amplification. *Nucleic Acids Res.* 16:11141–56
13. Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH. 2004. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* 305:869–72
14. Craig DW, Pearson JV, Szelling S, Sekar A, Redman M, et al. 2008. Identification of genetic variants using bar-coded multiplexed sequencing. *Nat. Methods* 5:887–93
15. Dahl F, Gullberg M, Stenberg J, Landegren U, Nilsson M. 2005. Multiplex amplification enabled by selective circularization of large sets of genomic DNA fragments. *Nucleic Acids Res.* 33:e71
16. Dahl F, Stenberg J, Fredriksson S, Welch K, Zhang M, et al. 2007. Multigene amplification and massively parallel sequencing for cancer mutation discovery. *Proc. Natl. Acad. Sci. USA* 104:9387–92
17. Edwards MC, Gibbs RA. 1994. Multiplex PCR: advantages, development, and applications. *PCR Methods Appl.* 3:S65–75
18. Fan JB, Chee MS, Gunderson KL. 2006. Highly parallel genomic assays. *Nat. Rev. Genet.* 7:632–44
19. Faruqi AF, Hosono S, Driscoll MD, Dean FB, Alsmadi O, et al. 2001. High-throughput genotyping of single nucleotide polymorphisms with rolling circle amplification. *BMC Genomics* 2:4
20. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449:851–61
21. Fredriksson S, Baner J, Dahl F, Chu A, Ji H, et al. 2007. Multiplex amplification of all coding sequences within 10 cancer genes by Gene-Collector. *Nucleic Acids Res.* 35:e47
22. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, et al. 2004. A census of human cancer genes. *Nat. Rev. Cancer* 4:177–83
23. Garber K. 2008. Fixing the front end. *Nat. Biotechnol.* 26:1101–4
- 23a. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, et al. 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.* 27:182–89
24. Gunderson KL, Steemers FJ, Lee G, Mendoza LG, Chee MS. 2005. A genome-wide scalable SNP genotyping assay using microarray technology. *Nat. Genet.* 37:549–54
25. Hacia JG, Brody LC, Chee MS, Fodor SP, Collins FS. 1996. Detection of heterozygous mutations in BRCA1 using high density oligonucleotide arrays and two-colour fluorescence analysis. *Nat. Genet.* 14:441–47
26. Hardenbol P, Baner J, Jain M, Nilsson M, Namsaraev EA, et al. 2003. Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nat. Biotechnol.* 21:673–78
27. Hardenbol P, Yu F, Belmont J, Mackenzie J, Bruckner C, et al. 2005. Highly multiplexed molecular inversion probe genotyping: over 10,000 targeted SNPs genotyped in a single tube assay. *Genome Res.* 15:269–75
28. Harris TD, Buzby PR, Babcock H, Beer E, Bowers J, et al. 2008. Single-molecule DNA sequencing of a viral genome. *Science* 320:106–9
29. Hillier LW, Marth GT, Quinlan AR, Dooling D, Fewell G, et al. 2008. Whole-genome sequencing and variant discovery in *C. elegans*. *Nat. Methods* 5:183–88
30. Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, et al. 2005. Whole-genome patterns of common DNA variation in three human populations. *Science* 307:1072–79
31. Hirschhorn JN, Daly MJ. 2005. Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* 6:95–108
32. Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, et al. 2007. Genome-wide in situ exon capture for selective resequencing. *Nat. Genet.* 39:1522–27
33. Holt KE, Parkhill J, Mazzoni CJ, Roumagnac P, Weill FX, et al. 2008. High-throughput sequencing provides insights into genome variation and evolution in *Salmonella* Typhi. *Nat. Genet.* 40:987–93
34. IHC. 2005. A haplotype map of the human genome. *Nature* 437:1299–320
35. Krishnakumar S, Zheng J, Wilhelmy J, Faham M, Mindrinos M, Davis R. 2008. A comprehensive assay for targeted multiplex amplification of human DNA sequences. *Proc. Natl. Acad. Sci. USA* 105:9296–301

36. Landegren U, Schallmeiner E, Nilsson M, Fredriksson S, Baner J, et al. 2004. Molecular tools for a molecular medicine: analyzing genes, transcripts and proteins using padlock and proximity probes. *J. Mol. Recognit.* 17:194–97
37. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, et al. 2007. The diploid genome sequence of an individual human. *PLoS Biol.* 5:e254
38. Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, et al. 2008. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* 456:66–72
39. Li H, Ruan J, Durbin R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* 18(11):1851–58
40. Lin Z, Cui X, Li H. 1996. Multiplex genotype determination at a large number of gene loci. *Proc. Natl. Acad. Sci. USA* 93:2582–87
41. Lizardi PM, Huang X, Zhu Z, Bray-Ward P, Thomas DC, Ward DC. 1998. Mutation detection and single-molecule counting using isothermal rolling-circle amplification. *Nat. Genet.* 19:225–32
42. Lovett M, Kere J, Hinton LM. 1991. Direct selection: a method for the isolation of cDNAs encoded by large genomic regions. *Proc. Natl. Acad. Sci. USA* 88:9628–32
43. Lundquist PM, Zhong CF, Zhao P, Tomaney AB, Peluso PS, et al. 2008. Parallel confocal detection of single molecules in real time. *Opt. Lett.* 33:1026–28
44. Lyamichev V, Brow MA, Dahlberg JE. 1993. Structure-specific endonucleolytic cleavage of nucleic acids by eubacterial DNA polymerases. *Science* 260:778–83
45. Mardis ER. 2008. Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.* 9:387–402
46. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376–80
47. McKusick VA. 2007. Mendelian inheritance in man and its online version, OMIM. *Am. J. Hum. Genet.* 80:588–604
48. Meuzelaar LS, Lancaster O, Pasche JP, Kopal G, Brookes AJ. 2007. MegaPlex PCR: a strategy for multiplex amplification. *Nat. Methods* 4:835–37
49. Nilsson M, Malmgren H, Samiotaki M, Kwiatkowski M, Chowdhary BP, Landegren U. 1994. Padlock probes: circularizing oligonucleotides for localized DNA detection. *Science* 265:2085–88
50. Noonan JP, Coop G, Kudaravalli S, Smith D, Krause J, et al. 2006. Sequencing and analysis of Neanderthal genomic DNA. *Science* 314:1113–18
51. Okou DT, Steinberg KM, Middle C, Cutler DJ, Albert TJ, Zwick ME. 2007. Microarray-based genomic selection for high-throughput resequencing. *Nat. Methods* 4:907–9
52. Parimoo S, Patanjali SR, Shukla H, Chaplin DD, Weissman SM. 1991. cDNA selection: efficient PCR approach for the selection of cDNAs encoded in large chromosomal DNA fragments. *Proc. Natl. Acad. Sci. USA* 88:9623–27
53. Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, et al. 2001. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* 294:1719–23
54. Pemov A, Modi H, Chandler DP, Bavykin S. 2005. DNA analysis with multiplex microarray-enhanced PCR. *Nucleic Acids Res.* 33:e11
55. Porreca GJ, Zhang K, Li JB, Xie B, Austin D, et al. 2007. Multiplex amplification of large sets of human exons. *Nat. Methods* 4:931–36
56. Sarin S, Prabhu S, O'Meara MM, Pe'er I, Hobert O. 2008. *Caenorhabditis elegans* mutant allele identification by whole-genome sequencing. *Nat. Methods* 5:865–67
57. Shendure J, Ji H. 2008. Next-generation DNA sequencing. *Nat. Biotechnol.* 26:1135–45
58. Shendure J, Mitra RD, Varma C, Church GM. 2004. Advanced sequencing technologies: methods and goals. *Nat. Rev. Genet.* 5:335–44
59. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, et al. 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 309:1728–32
60. Sjoblom T, Jones S, Wood LD, Parsons DW, Lin J, et al. 2006. The consensus coding sequences of human breast and colorectal cancers. *Science* 314:268–74
61. Stephens M, Sloan JS, Robertson PD, Scheet P, Nickerson DA. 2006. Automating sequence-based detection and genotyping of SNPs from diploid samples. *Nat. Genet.* 38:375–81

62. Stratton MR, Rahman N. 2008. The emerging landscape of breast cancer susceptibility. *Nat. Genet.* 40:17–22
63. Thomas RK, Nickerson E, Simons JF, Janne PA, Tengs T, et al. 2006. Sensitive mutation detection in heterogeneous cancer specimens by massively parallel picoliter reactor sequencing. *Nat. Med.* 12:852–55
64. Topol EJ, Frazer KA. 2007. The resequencing imperative. *Nat. Genet.* 39:439–40
- 64a. Turner EH, Lee C, Ng SB, Nickerson DA, Shendure J. 2009. Massively parallel exon capture and library-free resequencing across 16 genomes. *Nat. Methods* 6:315–16
65. Varley KE, Mitra RD. 2008. Nested Patch PCR enables highly multiplexed mutation discovery in candidate genes. *Genome Res.* 18:1844–50
66. Wang J, Wang W, Li R, Li Y, Tian G, et al. 2008. The diploid genome sequence of an Asian individual. *Nature* 456:60–65
67. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, et al. 2008. The complete genome of an individual by massively parallel DNA sequencing. *Nature* 452:872–76
68. Williams R, Peisajovich SG, Miller OJ, Magdassi S, Tawfik DS, Griffiths AD. 2006. Amplification of complex gene libraries by emulsion PCR. *Nat. Methods* 3:545–50
69. Yeager M, Xiao N, Hayes RB, Bouffard P, Desany B, et al. 2008. Comprehensive resequence analysis of a 136 kb region of human chromosome 8q24 associated with prostate and colon cancers. *Hum. Genet.* 124:161–70
70. Zwick ME, McAfee F, Cutler DJ, Read TD, Ravel J, et al. 2005. Microarray-based resequencing of multiple *Bacillus anthracis* isolates. *Genome Biol.* 6:R10