

An integrated semiconductor device enabling non-optical genome sequencing

Jonathan M. Rothberg¹, Wolfgang Hinz¹, Todd M. Rearick¹, Jonathan Schultz¹, William Mileski¹, Mel Davey¹, John H. Leamon¹, Kim Johnson¹, Mark J. Milgrew¹, Matthew Edwards¹, Jeremy Hoon¹, Jan F. Simons¹, David Marran¹, Jason W. Myers¹, John F. Davidson¹, Annika Branting¹, John R. Nobile¹, Bernard P. Puc¹, David Light¹, Travis A. Clark¹, Martin Huber¹, Jeffrey T. Branciforte¹, Isaac B. Stoner¹, Simon E. Cawley¹, Michael Lyons¹, Yutao Fu¹, Nils Homer¹, Marina Sedova¹, Xin Miao¹, Brian Reed¹, Jeffrey Sabina¹, Erika Feierstein¹, Michelle Schorn¹, Mohammad Alanjary¹, Eileen Dimalanta¹, Devin Dressman¹, Rachel Kasinskas¹, Tanya Sokolsky¹, Jacqueline A. Fidanza¹, Eugeni Namsaraev¹, Kevin J. McKernan¹, Alan Williams¹, G. Thomas Roth¹ & James Bustillo¹

The seminal importance of DNA sequencing to the life sciences, biotechnology and medicine has driven the search for more scalable and lower-cost solutions. Here we describe a DNA sequencing technology in which scalable, low-cost semiconductor manufacturing techniques are used to make an integrated circuit able to directly perform non-optical DNA sequencing of genomes. Sequence data are obtained by directly sensing the ions produced by template-directed DNA polymerase synthesis using all-natural nucleotides on this massively parallel semiconductor-sensing device or ion chip. The ion chip contains ion-sensitive, field-effect transistor-based sensors in perfect register with 1.2 million wells, which provide confinement and allow parallel, simultaneous detection of independent sequencing reactions. Use of the most widely used technology for constructing integrated circuits, the complementary metal-oxide semiconductor (CMOS) process, allows for low-cost, large-scale production and scaling of the device to higher densities and larger array sizes. We show the performance of the system by sequencing three bacterial genomes, its robustness and scalability by producing ion chips with up to 10 times as many sensors and sequencing a human genome.

DNA sequencing and, more recently, massively parallel DNA sequencing^{1–4} has had a profound impact on research and medicine. The reductions in cost and time for generating DNA sequence have resulted in a range of new sequencing applications in cancer^{5,6}, human genetics⁷, infectious diseases⁸ and the study of personal genomes^{9–11}, as well as in fields as diverse as ecology^{12,13} and the study of ancient DNA^{14,15}. Although *de novo* sequencing costs have dropped substantially, there is a desire to continue to drop the cost of sequencing at an exponential rate consistent with the semiconductor industry's Moore's Law¹⁶ as well as to provide lower cost, faster and more portable devices. This has been operationalized by the desire to reach the \$1,000 genome¹⁷.

To date, DNA sequencing has been limited by its requirement for imaging technology, electromagnetic intermediates (either X-rays¹⁸, or light¹⁹) and specialized nucleotides or other reagents²⁰. To overcome these limitations and further democratize the practice of sequencing, a paradigm shift based on non-optical sequencing on newly developed integrated circuits was pursued. Owing to its scalability and its low power requirement, CMOS processes are dominant in modern integrated circuit manufacturing²¹. The ubiquitous nature of computers, digital cameras and mobile phones has been made possible by the low-cost production of integrated circuits in CMOS.

Leveraging advances in the imaging field—which has produced large, fast arrays for photonic imaging²²—we sought a suitable electronic sensor for the construction of an integrated circuit to detect the hydrogen ions that would be released by DNA polymerase²³ during sequencing by synthesis, as opposed to a sensor designed for the detection of photons. Although a variety of electrochemical detection methods have been studied^{24,25}, the ion-sensitive field-effect transistor (ISFET)^{26,27} was most applicable to our chemistry and scaling requirements because of

its sensitivity to hydrogen ions, and its compatibility with CMOS processes^{28–31}. Previous attempts to detect both single-nucleotide polymorphisms (SNPs)³² and DNA synthesis³³ as well as sequence DNA electronically³⁴ have been made. However, none of them produced *de novo* DNA sequence, addressed the issue of delivering template DNA to the sensors, or scaled to large arrays. In addition, previous efforts in ISFETs were limited in the number of sensors per array, the yield of working independent sensors and readout speed^{35,36}, and encountered difficulty in exposing the sensors to fluids while protecting the electronics³⁷.

Here, we overcome previous limitations with electronic detection and enable the production of chips with a large number of fast, uniform, working sensors. Our focus has been on the development of these ion chips, as well as the biochemical methods, supporting instrumentation and software needed to enable *de novo* DNA sequencing for applications requiring millions to billions of bases (Supplementary Fig. 1). A typical 2-h run using an ion chip with 1.2 M sensors generates approximately 25 million bases. The performance of the ion chips and overall sequencing platform is demonstrated through whole-genome sequencing of three bacterial genomes. The scalability of our chip architecture is demonstrated by producing chips with up to 10 times the number of sensors and producing a low-coverage sequence of the genome of Gordon Moore, author of Moore's law¹⁶.

A CMOS integrated circuit for sequencing

We have developed a simple, scalable ISFET sensor architecture using electronic addressing common in modern CMOS imagers (Supplementary Fig. 2). Our integrated circuit consists of a large array of sensor elements, each with a single floating gate connected to an underlying ISFET (Fig. 1a). For sequence confinement we rely on a

¹Ion Torrent by Life Technologies, Suite 100, 246 Goose Lane, Guilford, Connecticut 06437, USA.

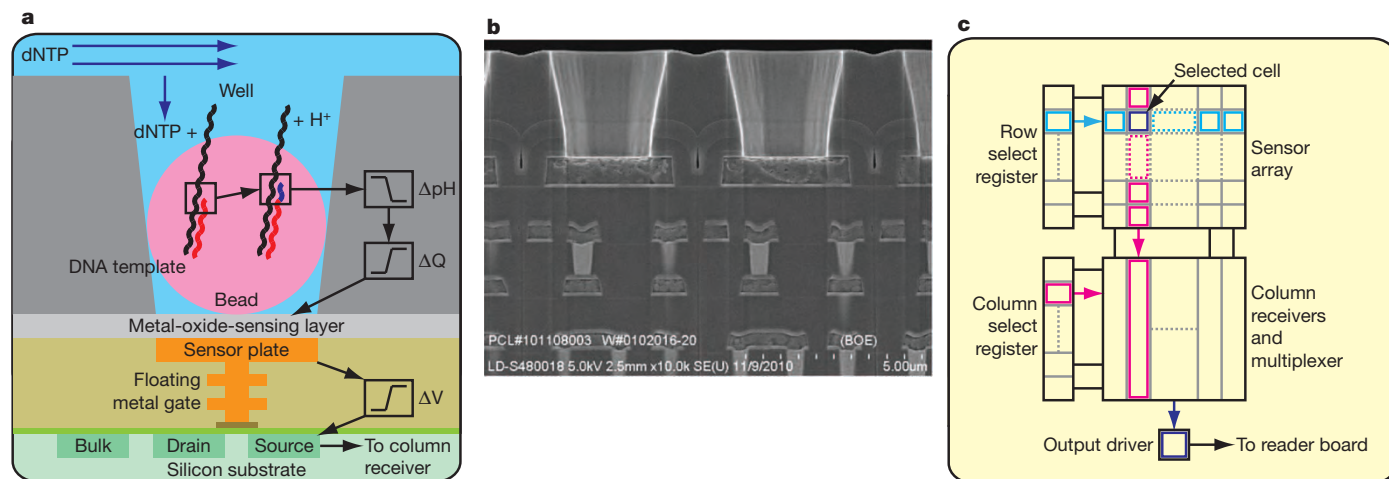


Figure 1 | Sensor, well and chip architecture. **a**, A simplified drawing of a well, a bead containing DNA template, and the underlying sensor and electronics. Protons (H^+) are released when nucleotides (dNTP) are incorporated on the growing DNA strands, changing the pH of the well (ΔpH). This induces a change in surface potential of the metal-oxide-sensing layer, and a change in potential (ΔV) of the source terminal of the underlying field-effect

transistor. **b**, Electron micrograph showing alignment of the wells over the ISFET metal sensor plate and the underlying electronic layers. **c**, Sensors are arranged in a two-dimensional array. A row select register enables one row of sensors at a time, causing each sensor to drive its source voltage onto a column. A column select register selects one of the columns for output to external electronics.

$3.5\text{-}\mu\text{m}$ -diameter well formed by adding a $3\text{-}\mu\text{m}$ -thick dielectric layer over the electronics and etching to the sensor plate (Fig. 1b). A tantalum oxide layer provides for proton sensitivity (58 mV pH^{-1} ; ref. 38). High-speed addressing and readout are accomplished by the semiconductor electronics integrated with the sensor array (Fig. 1c). The sensor and underlying electronics provide a direct transduction from the incorporation event to an electronic signal. Unlike light-based sequencing technology, we do not use the elements of the array to collect photons and form a larger image to detect the incorporation of a base; instead we use each sensor to independently and directly monitor the hydrogen ions released during nucleotide incorporation.

Ion chips are manufactured on wafers (Fig. 2a), cut into individual die (Fig. 2b) and robotically packaged with a disposable polycarbonate flow cell that isolates the fluids to regions above the sensor array and away from the supporting electronics to provide convenient sample loading as well as electrical and fluidic interfaces to the sequencing instrument (Fig. 2c). Chips were designed and fabricated with 1.5 M, 7.2 M and 13 M ISFETs (Supplementary Fig. 3). On the basis of the placement of the flow cell on the sensor array, 1.2 M, 6.1 M and 11 M

wells and sensors are exposed to fluids, with 99.9% of the sensors sensitive to pH and usable for DNA sequencing (Supplementary Fig. 4). Increasing the numbers of sensors per chip was first achieved by increasing the die area, from $10.6\text{ mm} \times 10.9\text{ mm}$ to $17.5\text{ mm} \times 17.5\text{ mm}$, and then by increasing the density of the sensors by reducing the number of transistors per sensor from three to two. Chip density is limited by the selection of the CMOS node and the number of transistors per sensing element. Using a $0.35\text{ }\mu\text{m}$ CMOS node the minimum spacing for a three-transistor sensor is $5.1\text{ }\mu\text{m}$ and for a two-transistor sensor it is $3.8\text{ }\mu\text{m}$ (Supplementary Fig. 5). To understand further the limits on density, we show that $1.3\text{ }\mu\text{m}$ wells are readily manufactured, can be aligned to sensors, enable the generation of high-quality sequence (Supplementary Fig. 6) and can, using a 110 nm node, be fabricated with a spacing as small as $1.68\text{ }\mu\text{m}$ (Supplementary Fig. 7).

Sequencing on a semiconductor device

The all-electronic detection system used by the ion chip simplifies and greatly reduces the cost of the sequencing instrument (Supplementary

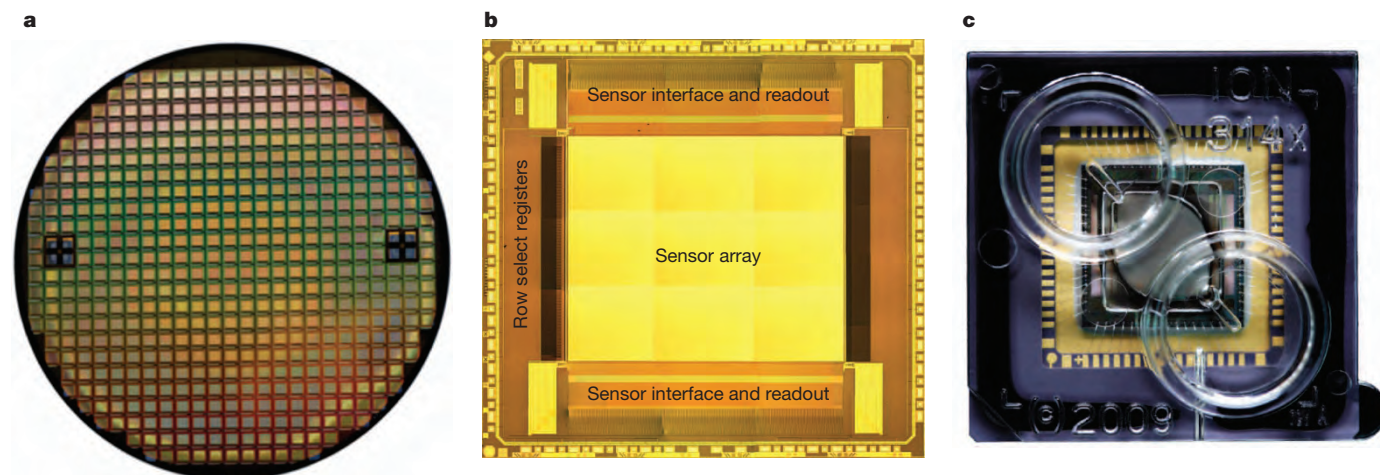


Figure 2 | Wafer, die and chip packaging. **a**, Fabricated CMOS 8'' wafer containing approximately 200 individual functional ion sensor die. **b**, Unpackaged die, after automated dicing of wafer, with functional regions

indicated. **c**, Die in ceramic package wire bonded for electrical connection, shown with moulded fluidic lid to allow addition of sequencing reagents.

Fig. 8). The instrument has no optical components, and is comprised primarily of an electronic reader board to interface with the chip, a microprocessor for signal processing, and a fluidics system to control the flow of reagents over the chip (Supplementary Fig. 9).

Genomic DNA is prepared for sequencing as described in Supplementary Methods. Briefly, DNA is fragmented, ligated to adapters, and adaptor-ligated libraries are clonally amplified onto beads. Template-bearing beads are enriched through a magnetic-bead-based process. Sequencing primers and DNA polymerase are then bound to the templates and pipetted into the chip's loading port. Individual beads are loaded into individual sensor wells by spinning the chip in a desktop centrifuge. A 2 μm acrylamide bead was chosen to deliver sufficient copies of the template to the sensor well to achieve a high signal-to-noise ratio (SNR) (800 K copies, SNR, 10; Supplementary Methods and Supplementary Fig. 10), while well depth was selected to allow only a single bead to occupy a well.

In ion sequencing, all four nucleotides are provided in a stepwise fashion during an automated run (Supplementary Methods). When the nucleotide in the flow is complementary to the template base directly downstream of the sequencing primer, the nucleotide is incorporated into the nascent strand by the bound polymerase. This increases the length of the sequencing primer by one base (or more, if a homopolymer stretch is directly downstream of the primer) and results in the hydrolysis of the incoming nucleotide triphosphate, which causes the net liberation of a single proton for each nucleotide incorporated during that flow. The release of the proton produces a shift in the pH of the surrounding solution proportional to the number of nucleotides incorporated in the flow (0.02 pH units per single base incorporation). This is detected by the sensor on the bottom of each well, converted to a voltage and digitized by off-chip electronics (Fig. 3). The signal generation and detection occurs over 4 s (Fig. 3b). After the flow of each nucleotide, a wash is used to ensure nucleotides do not remain in the well. The small size of the wells allows diffusion into and out of the well on the order of a one-tenth of a second and eliminates the need for enzymatic removal of reagents¹.

Signal processing and base calling

To change raw voltages into base calls, signal-processing software converts the raw data into measurements of incorporation in each well for each successive nucleotide flow using a physical model. Sampling the signal at high frequency relative to the time of the incorporation signal allows signal averaging to improve the SNR. The physical model takes into consideration diffusion rates, buffering effects and polymerase rates (Supplementary Fig. 11). The model is

applied and fit to the raw trace from each well and the incorporation signals are extracted. A base caller corrects the signals for phase and signal loss, normalizes to the key, and generates corrected base calls for each flow in each well to produce the sequencing reads (Fig. 3c and Supplementary Fig. 12).

Next, each read is sequentially passed through two signal-based filters to exclude low-accuracy reads. The first filter measures the fraction of flows in which an incorporation event was measured. When this value is unusually large (greater than 60% of the first 60 flows) the read is not clonal. The second filter measures the extent to which the observed signal values match those predicted by the phasing model. When there is poor agreement (median absolute difference more than 0.06 over the first 60 flows) between the two, it corresponds to higher error rates. Lastly, per-base quality values are predicted using an adaptation of the Phred method³⁹ that quantifies the concordance between the phasing model predictions and the observed signal. These *ab initio* scores track closely with post-alignment derived quality scores, and are used to trim back low-quality sequence from the 3' end of a read (Supplementary Fig. 13).

Sequencing bacterial genomes

Bacterial genome sequencing and signal processing was performed as described earlier. We succeeded in sequencing all three genomes five-fold to tenfold in individual runs using the small ion chip, covering 96.80% to 99.99% of each genome, with genome-wide consensus accuracies as high as 99.99% (Table 1 and Supplementary Fig. 14). *Escherichia coli* sequencing with three successively larger ion chips produced 46 to over 270 megabases of sequence (Table 1).

To characterize run quality, we aligned each read to the corresponding reference genome (Supplementary Fig. 15). The per-base accuracy was observed to be $99.569\% \pm 0.001\%$ within the first 50 bases and $98.897\% \pm 0.001\%$ within the first 100 bases (Supplementary Fig. 16a). This accuracy is similar at 50 bases and higher at 100 bases than light-based methods using modified nucleotides (1.1% versus 5% error⁴⁰). The per-base accuracy in calling a homopolymer of length 5 is $97.328\% \pm 0.023\%$ (Supplementary Fig. 16b) and higher than pyrosequencing-based sequencing methods^{1,41}. For each genome, the observed distribution of per-base coverage matches closely with the theoretical Poisson distribution reflecting the uniform nature of the coverage (Supplementary Fig. 17). The distribution of coverage was also relatively unbiased across GC content (Supplementary Fig. 18).

Ion sequencing technology has allowed the routine acquisition of 100-base read lengths, and perfect read lengths exceeding 200 bases (Supplementary Fig. 19). At present, 20–40% of the sensors in a given

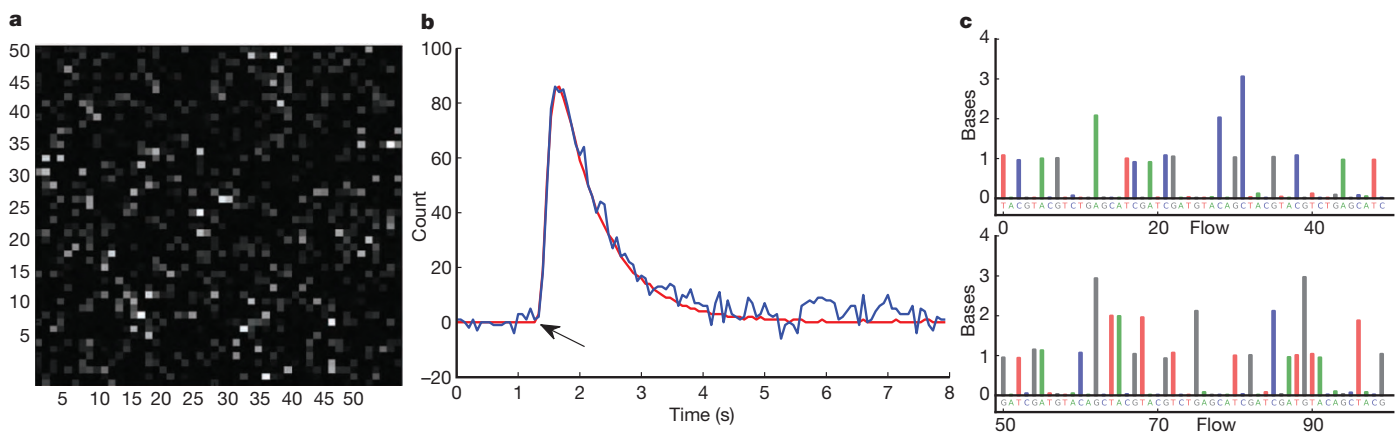


Figure 3 | Data collection and base calling. **a**, A 50×50 region of the ion chip. The brightness represents the intensity of the incorporation reaction in individual sensor wells. **b**, 1-nucleotide incorporation signal from an individual sensor well; the arrow indicates start of incorporation event, with the physical

model (red line) and background corrected data (blue line) shown. **c**, The first 100 flows from one well. Each coloured bar indicates the corresponding number of bases incorporated during that nucleotide flow.

Table 1 | *Vibrio fischeri*, *E. coli*, *Rhodospseudomonas palustris* and *Homo sapiens*

	<i>V. fischeri</i>	<i>R. palustris</i>	<i>E. coli</i>	<i>E. coli</i>	<i>E. coli</i>	<i>H. sapiens</i>
GC content	38%	65%	51%	51%	51%	41%
Genome size	4.2 Mb	5.5 Mb	4.7 Mb	4.7 Mb	4.7 Mb	2.9 Gb
Number of runs x ion chip size	1 × 1.2 M	1 × 1.2 M	1 × 1.2 M	1 × 6.1 M	1 × 11 M	1,601 × 1.2 M 267 × 6.1 M 28 × 11.1 M
Fold coverage	6.2-fold	6.9-fold	11.3-fold	36.2-fold	58.4-fold	10.6-fold
Coverage	96.80%	99.64%	99.99%	100.00%	100.00%	99.21%
Reads ≥21 bases	261,313	444,750	507,198	1,852,931	2,594,031	366,623,578
Reads ≥50 bases	233,049	399,360	487,420	1,698,852	2,343,880	306,042,650
Reads ≥100 bases	156,391	160,726	400,743	1,012,918	1,779,237	139,624,090
Mapped bases	26.0 Mb	37.8 Mb	47.6 Mb	169.6 Mb	273.9 Mb	30.2 Gb

Coverage shows percentage of genome covered based on one or more reads mapping to each base of the reference genome. Reads align with 98% or greater accuracy.

run yield mappable reads. The gap between the number of sensors on a chip and the number yielding sequence is primarily the result of incomplete loading of the chip, poor amplification of a fragment onto the bead, and lack of clonality of the template. With continued improvements in loading and template preparation, along with improvements in signal processing and base calling, it is expected that the percentage of sensors yielding reads, the average read length and read accuracy will all improve significantly, as it has for other sequencing technologies^{1–4,9–11}.

'Post-light' sequencing of G. Moore

To illustrate the scalability of semiconductor sequencing we produced whole-genome sequence data from an individual, G. Moore⁴² (Fig. 4). Written consent was provided by G. Moore to sequence and publish his genome and resulting findings. Reads from his genome were deposited in the Sequence Read Archive (SRA) under accession number ERP000682. The mean coverage of the G. Moore genome was 10.6-fold (Table 1). The degree to which the observed distribution of reads conforms to a Poisson distribution is indicative of a general lack of bias in coverage depth (Fig. 4b).

We found 2,598,983 SNPs in the G. Moore genome, of which 3.08% were found to be novel, consistent with previous reports^{4,9,11} (Supplementary Methods). To confirm the accuracy of our analysis, we also sequenced the G. Moore genome using ABI SOLiD Sequencing⁴³ to 15-fold coverage and validated 99.95% of the heterozygous and 99.97% of the homozygous genotypes (Supplementary Tables 1 and 2).

We used the Online Mendelian Inheritance in Man database⁴⁴ and the 23andMe functional SNP collection (<https://www.23andme.com>) to identify a subset of validated SNPs known to be involved in human disease and interesting phenotypes (Supplementary Table 3). We also examined the G. Moore sequence for the 7,693 deletions and inversions discovered by the 1000 Genomes Consortium and computationally found 3,413 of them in the G. Moore genome at a 99.94% positive predictive value (Supplementary Methods, Supplementary Table 4 and Supplementary Fig. 20). To determine G. Moore's maternal ancestry, reads were also mapped to human mitochondrial DNA⁴⁵ for a mean coverage of 732-fold. G. Moore's mitochondria belong to haplogroup H, the most common in Europe⁴⁶.

Discussion

We have demonstrated the ability to produce and use a disposable integrated circuit fabricated in standard CMOS foundries to perform, for the first time, 'post-light' genome sequencing of bacterial and human genomes. With fifty billion dollars spent per year on CMOS semiconductor fabrication and packaging technologies, our goal was to leverage that investment to make a highly scalable sequencing technology. Using the G. Moore genome we demonstrated the feasibility of sequencing a human genome. The G. Moore genome sequence required on the order of a thousand individual ion chips

comprising about one billion sensors. By demonstrating the ability to make larger and denser arrays, use fewer transistors per sensor, and sequence from wells as small as 1.3 μm, our work suggests that readily available CMOS nodes should enable the production of one-billion-sensor ion chips and low-cost routine human genome sequencing.

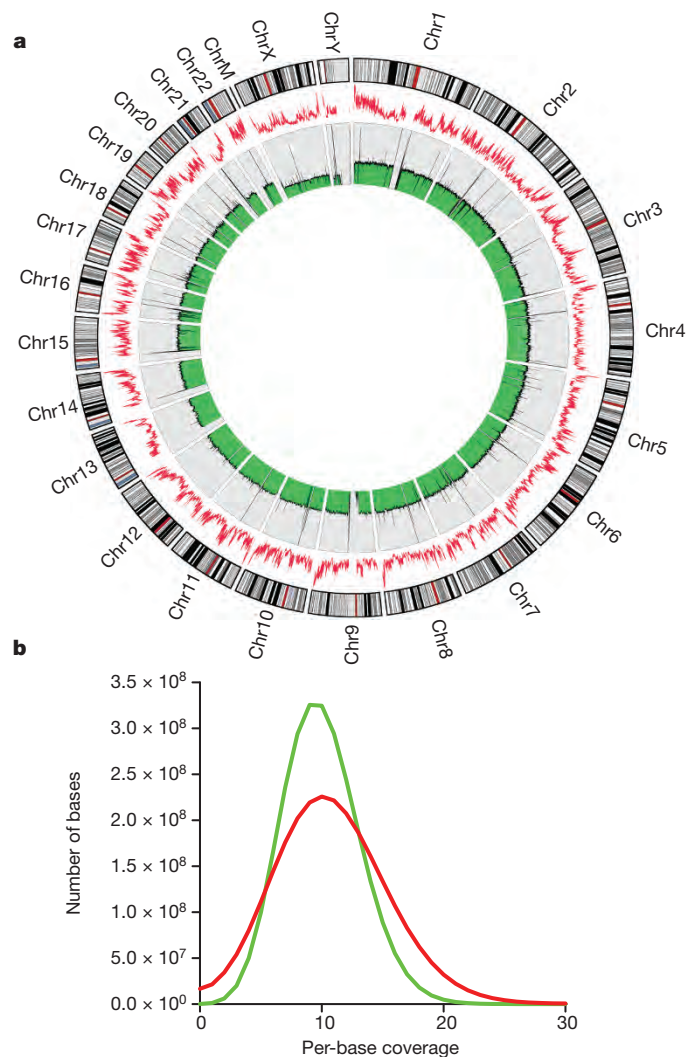


Figure 4 | *G. Moore* genome. **a**, Circular genome plot. The average sequencing coverage (green) and average GC content (red) within 100-kb intervals is shown. **b**, Distribution of the observed per-base coverage depth along the genome (red) compared with the distribution expected from random coverage (green).

Received 8 March; accepted 26 May 2011.

- Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
- Shendure, J. *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728–1732 (2005).
- Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
- Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78–81 (2010).
- Thomas, R. K. *et al.* Sensitive mutation detection in heterogeneous cancer specimens by massively parallel picoliter reactor sequencing. *Nature Med.* **12**, 852–855 (2006).
- Ley, T. J. *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456**, 66–72 (2008).
- Ng, S. B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nature Genet.* **42**, 30–35 (2010).
- Andries, K. *et al.* A diarylquinoline drug active on the ATP synthase of *Mycobacterium tuberculosis*. *Science* **307**, 223–227 (2005).
- Wheeler, D. A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872–876 (2008).
- Lupski, J. R. *et al.* Whole-genome sequencing in a patient with Charcot–Marie–Tooth neuropathy. *N. Engl. J. Med.* **362**, 1181–1191 (2010).
- Schuster, S. C. *et al.* Complete Khoisan and Bantu genomes from southern Africa. *Nature* **463**, 943–947 (2010).
- Venter, J. C. *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66–74 (2004).
- Sogin, M. L. *et al.* Microbial diversity in the deep sea and the underexplored “rare biosphere”. *Proc. Natl Acad. Sci. USA* **103**, 12115–12120 (2006).
- Noonan, J. P. *et al.* Genomic sequencing of Pleistocene cave bears. *Science* **309**, 597–599 (2005).
- Green, R. E. *et al.* Analysis of one million base pairs of Neanderthal DNA. *Nature* **444**, 330–336 (2006).
- Moore, G. E. Cramming more components onto integrated circuits. *Electronics* **38**, 114–117 (1965).
- Davies, K. *The \$1,000 Genome* (Free Press, 2010).
- Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl Acad. Sci. USA* **74**, 5463–5467 (1977).
- Smith, L. M. *et al.* Fluorescence detection in automated DNA sequence analysis. *Nature* **321**, 674–679 (1986).
- Metzker, M. L. Sequencing technologies—the next generation. *Nature Rev. Genet.* **11**, 31–46 (2010).
- Wanlass, F. M. Low stand-by power complementary field effect circuitry. U.S. patent 3,356,858: (1967).
- Theuwissen, A. J. P. CMOS image sensors: state-of-the-art. *Solid-State Electron.* **52**, 1401–1406 (2008).
- Sakurai, T. & Husimi, Y. Real-time monitoring of DNA polymerase reactions by a micro ISFET pH sensor. *Anal. Chem.* **64**, 1996–1997 (1992).
- Fritz, J., Cooper, E. B., Gaudet, S., Sorger, P. K. & Manalis, S. R. Electronic detection of DNA by its intrinsic molecular charge. *Proc. Natl Acad. Sci. USA* **99**, 14142–14146 (2002).
- Drummond, T. G., Hill, M. G. & Barton, J. K. Electrochemical DNA sensors. *Nature Biotechnol.* **21**, 1192–1199 (2003).
- Bergveld, P. Development of an ion-sensitive solid-state device for neurophysiological measurements. *IEEE Trans. Biomed. Eng.* **17**, 70–71 (1970).
- Bergveld, P. Thirty years of ISFETOLOGY. What happened in the past 30 years and what may happen in the next 30 years. *Sens. Actuators B Chem.* **88**, 1–20 (2003).
- Yeow, T., Haskard, M., Mulcahy, D., Seo, H. & Kwon, D. A very large integrated pH-ISFET sensor array chip compatible with standard CMOS processes. *Sens. Actuators B Chem.* **44**, 434–440 (1997).
- Bausells, J., Carrabina, J., Errachid, A. & Merlos, A. Ion-sensitive field-effect transistors fabricated in a commercial CMOS technology. *Sens. Actuators B Chem.* **57**, 56–62 (1999).
- Milgrew, M., Hammond, P. & Cumming, D. The development of scalable sensor arrays using standard CMOS technology. *Sens. Actuators B Chem.* **103**, 37–42 (2004).
- Hizawa, T., Sawada, K., Takao, H. & Ishida, M. Fabrication of a two-dimensional pH image sensor using a charge transfer technique. *Sens. Actuators B Chem.* **117**, 509–515 (2006).
- Purushothaman, S., Toumazou, C. & Ou, C. P. Protons and single nucleotide polymorphism detection: A simple use for the ion sensitive field effect transistor. *Sens. Actuators B Chem.* **114**, 964–968 (2006).
- Pourmand, N. *et al.* Direct electrical detection of DNA synthesis. *Proc. Natl Acad. Sci. USA* **103**, 6466–6470 (2006).
- Sakata, T. & Miyahara, Y. DNA sequencing based on intrinsic molecular charges. *Angew. Chem.* **118**, 2283–2286 (2006).
- Milgrew, M. J., Riehle, M. O. & Cumming, D. R. S. A large transistor-based sensor array chip for direct extracellular imaging. *Sens. Actuators B Chem.* **111–112**, 347–353 (2005).
- Milgrew, M. J. & Cumming, D. R. S. Matching the transconductance characteristics of CMOS ISFET arrays by removing trapped charge. *Electron Devices. IEEE Trans. Electron Devices* **55**, 1074–1079 (2008).
- Hammond, P. & Cumming, D. Encapsulation of a liquid-sensing microchip using SU-8 photoresist. *Microelectron. Eng.* **73–74**, 893–897 (2004).
- Mikolajick, T., Kühnhold, R. & Rysse, H. The pH-sensing properties of tantalum pentoxide films fabricated by metal organic low pressure chemical vapor deposition. *Sens. Actuators B Chem.* **44**, 262–267 (1997).
- Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8**, 186–194 (1998).
- Claesson, M. J. *et al.* Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. *Nucleic Acids Res.* **38**, e200 (2010).
- Huse, S. M., Huber, J. A., Morrison, H. G., Sogin, M. L. & Mark Welch, D. Accuracy and quality of massively-parallel DNA pyrosequencing. *Genome Biol.* **8**, R143 (2007).
- Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
- McKernan, K. J. *et al.* Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res.* **19**, 1527–1541 (2009).
- Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33**, D514–D517 (2005).
- Andrews, R. M. *et al.* Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nature Genet.* **23**, 147 (1999).
- Kloss Brandstätter, A. *et al.* HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. *Hum. Mutat.* **32**, 25–32 (2011).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We want to thank G. Moore for his willingness to participate in this study. We thank G. Fergus, M. Jain, J. Kole, L. Stevens and the ION team for supporting our efforts, and H. Peckman, V. Tadigotla, D. Holloway and S. McLaughlin for help on the variant analysis, and M. Ross of the Broad Institute for help on quality scores. This research was supported, in part, by a grant from the National Human Genome Research Institute (NHGRI), RFA-HG-08-008, Revolutionary Genome Sequencing Technologies—The \$1000 Genome. Grant number: R01 HG005094.

Author Contributions J.M.R. conceived the technology, supervised the project and wrote the manuscript with input from co-authors. K.J., M.J.M. and J.B. designed chips. J.F.D., M.A., D.L., J.W.M., J.F.S., E.N., M.S., X.M., A.B., T.A.C., M.H., I.B.S., B.R., J.S., E.F., M.S., J.A.F., K.J.M. and J.H.L. developed methods. M.D., J.T.B., M.E., J.H., N.H., T.M.R., B.P.P., S.E.C., M.L., Y.F. and A.W. wrote software and analysed data. W.H., J.S., W.M., D.M., J.R.N. and G.T.R. designed the instrument. E.D., D.D., R.K. and T.S. sequenced the human sample.

Author Information Sequences for *Homo sapiens*, *Escherichia coli*, *Vibrio fischeri* and *Rhodospseudomonas palustris* were deposited in the Sequence Read Archive (SRA) under accession numbers ERP000682, ERP000541, ERP000542 and ERP000543 respectively. Reprints and permissions information is available at www.nature.com/reprints. This paper is distributed under the terms of the Creative Commons Attribution-Non-Commercial-Share Alike Licence, and is freely available to all readers at www.nature.com/nature. The authors declare competing financial interests: details accompany the full-text HTML version of this paper at www.nature.com/nature. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to J.M.R. (Jonathan.Rothberg@LifeTech.com).

Supplemental Figures

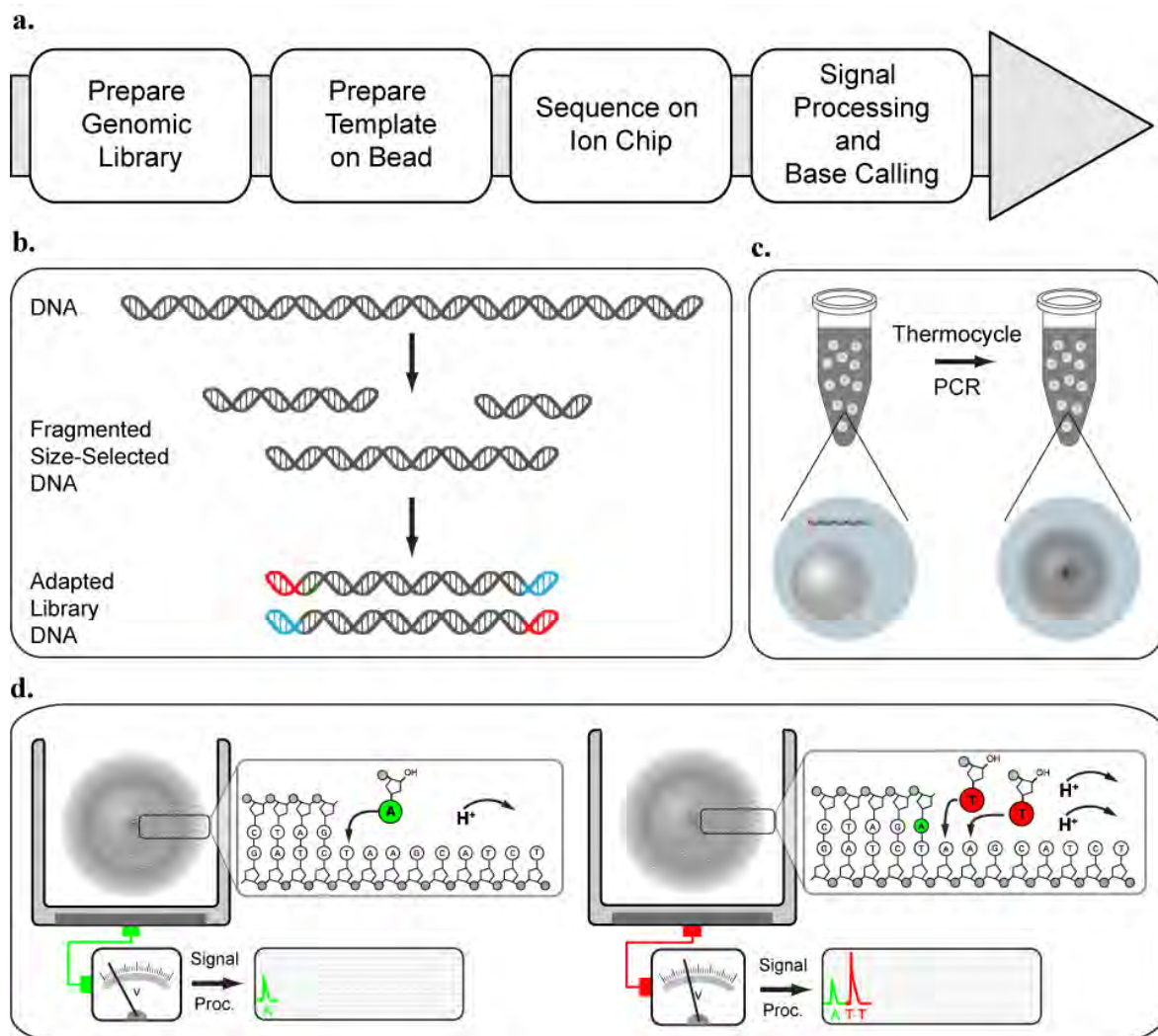


Figure S1 Process overview

a. Overview of ion sequencing work flow. **b.** Prepare genomic library, DNA is fragmented, sized, and forward and reverse adapters ligated. **c.** Amplify Template on bead, adapter-ligated libraries are clonally amplified onto beads. A magnetic bead-based enrichment process selects template-carrying beads. **d.** Sequence on ion chip, sequencing primers and DNA polymerase are bound to the template-carrying beads, beads are pipetted into the chip's loading port. The chip is installed in the sequencing instrument; all four nucleotides cyclically flowed in an automated 2-hour run. Signal processing, software converts the raw data into measurements of incorporation in each well for each successive nucleotide flow. After bases are called, each read is passed through a filter to exclude low-accuracy reads and per-base quality values are predicted.

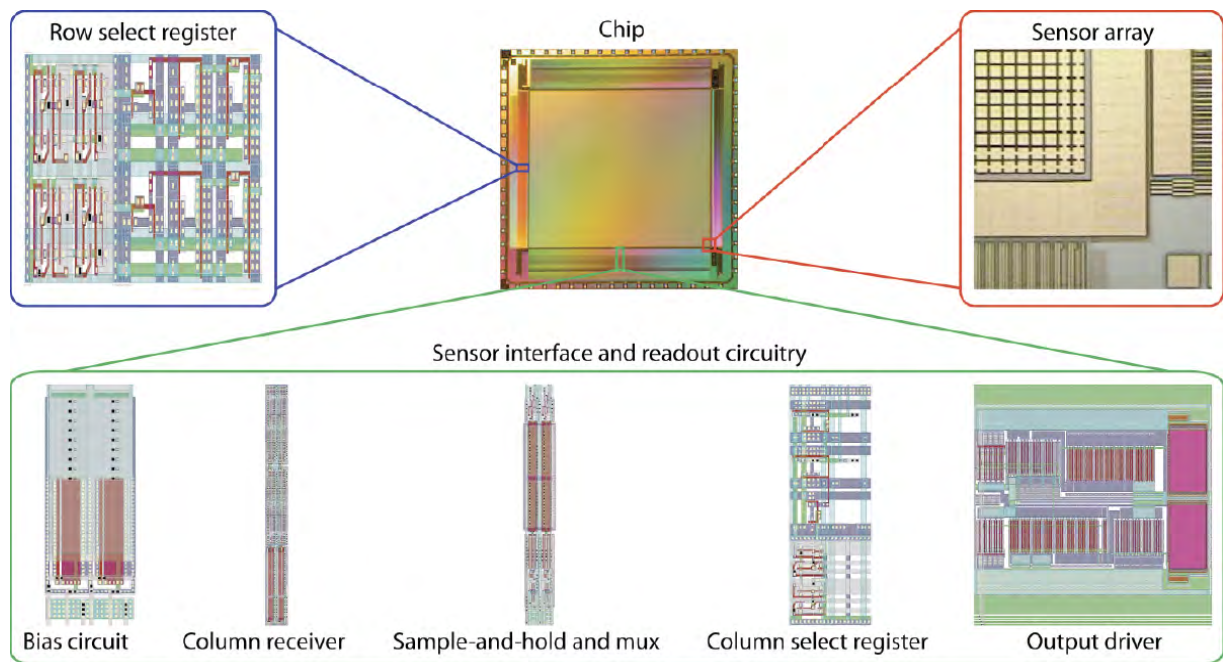


Figure S2 Chip architecture

Functional blocks of the ion chip. **Chip**, with **Row select registers**, to sequentially address each row, and **Sensor array**, showing close up of the individual metal floating plates, **Sensor interface and readout circuitry**, containing **Bias circuits**, to set the operating current, **Column receiver** to set the operating voltage, **Sample-and-hold and mux**, to capture the output voltage, **Column select register** to sequentially address each column, and **Output driver** to transmit voltages off-chip for external data acquisition.

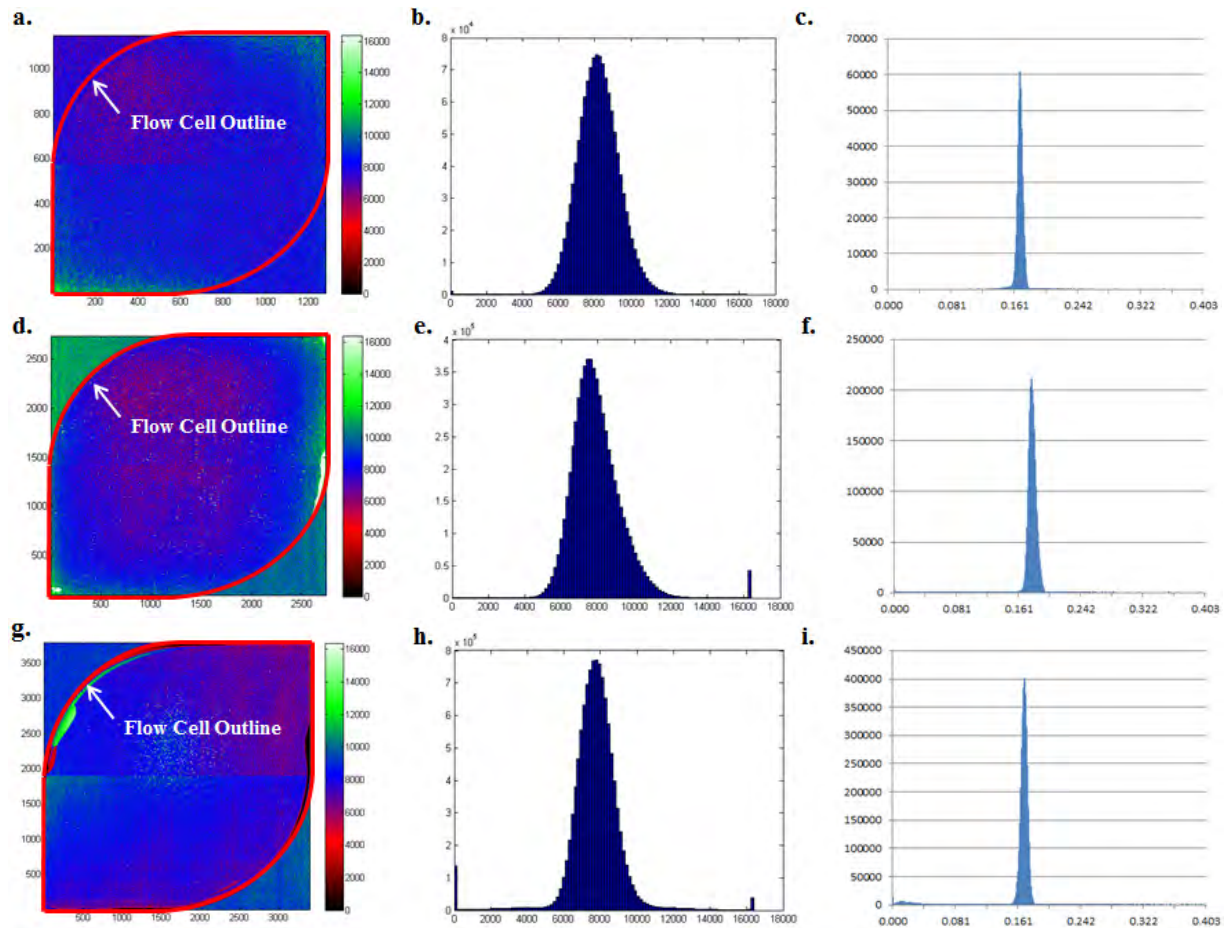


Figure S4 Large, uniform, electronically working, pH responsive sensor arrays

a, Image of the measured voltage of each sensor of the 1.5 M ISFET chip. The red outline indicates the edges of the flow cell and defines the central region with 1.2 M fluid accessible sensors. **b**, Histogram of every fluid accessible sensor's voltage. The extent of the x-axis indicates the minimum and maximum voltages that can be measured. More than 99% of the sensors are within the detection limits of the hardware. **c**, Histogram of pH response (delta pH change) for fluid accessible sensors. A known pH buffer was flowed over the chip and pH change measured at every well. More than 99% of the wells fall within a very tight range of the expected response and hence can work for DNA sequencing. **d, e, f**, Image, histogram, and pH response for the 7.2 M ISFET, 6.1 M accessible sensors. More than 99% of the accessible sensors are within the detection limits, of those more than 99% respond to pH. **g, h, i**, Image, histogram, and pH response for the 13 M ISFET, 11 M accessible sensors. More than 98% of the sensors are within detection limits, of those more than 94% respond to pH. Sensors that don't correctly respond to pH changes are obscured by glue used in the attachment of the flow cell to the sensor.

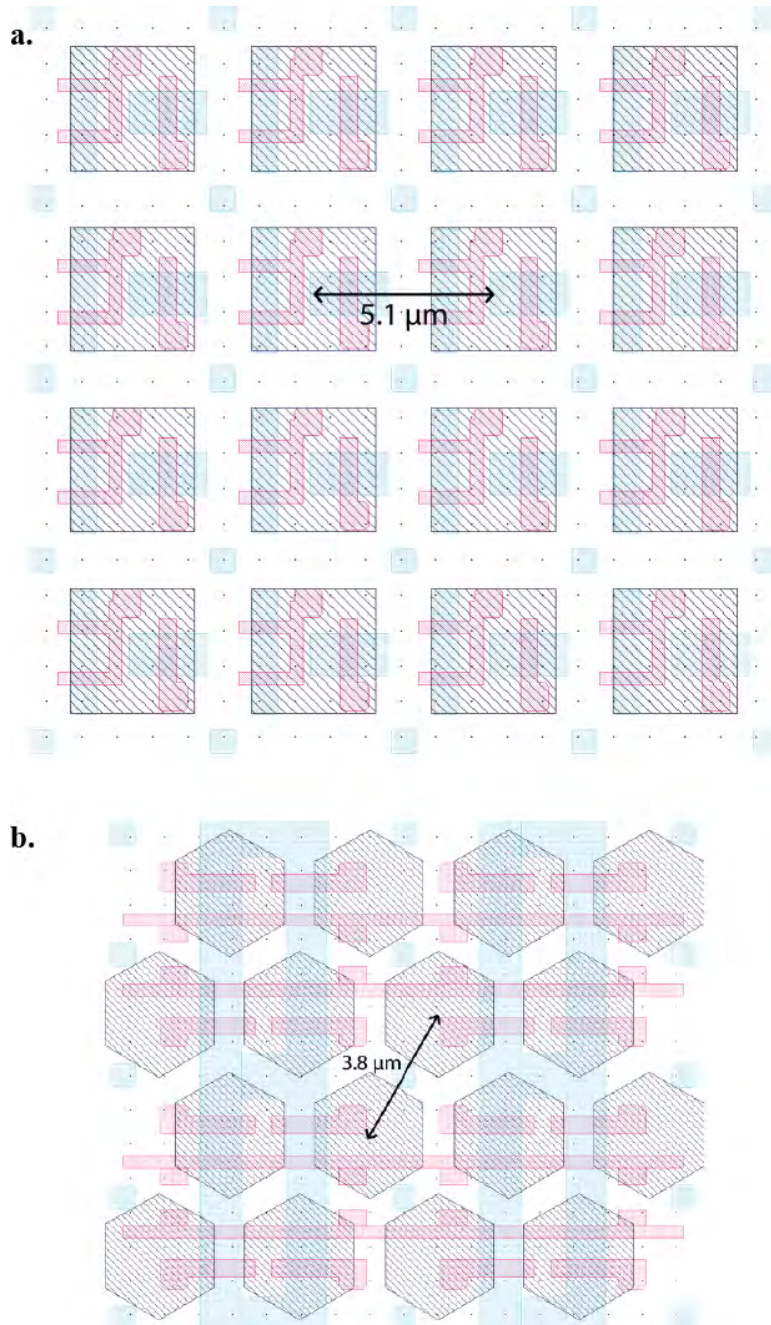


Figure S5 3-transistor and 2-transistor sensor

Layout of a 4 x 4 region of the 3-transistor and 2-transistor sensor, using 0.35 μm CMOS design rules. **a.** 3-transistor sensor array, orthogonal Manhattan arrangement, 5.1 μm center-to-center pitch. **b.** 2-transistor sensor array, 3.8 μm center-to-center pitch.

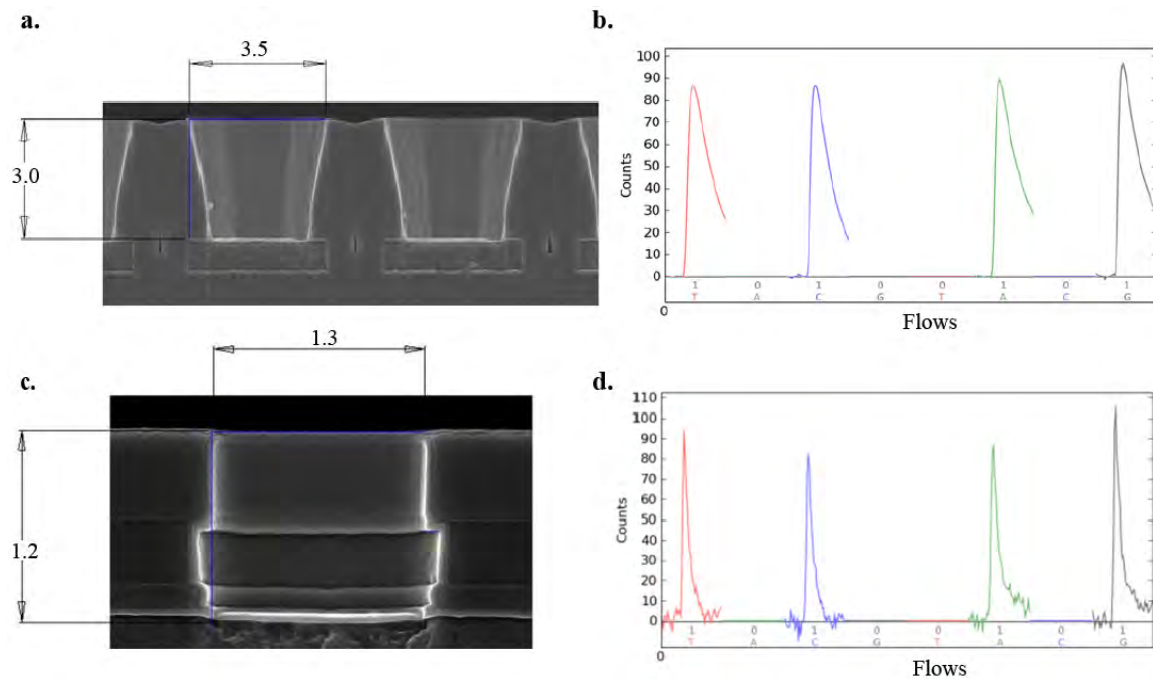


Figure S6 3.5 and 1.3 μm well SEMs and sequencing traces

a, Scanning electron micrograph cross-section of a 3.5 μm top diameter well. **b**, Background-subtracted consensus sequencing trace from the same 3.5 μm well size. **c**, Scanning electron micrograph cross-section from a 1.3 μm diameter well. **d**, Background-subtracted consensus sequencing trace from the same 1.3 μm smaller well size. In the consensus sequencing traces the X axis indicates which of the 4 nucleotides is flowed (A,T,C or G) and the normalized magnitude of the sequencing signal (0-mer or 1-mer). The signal is measured in counts proportional to the voltage detected at the sensor. The polymerase incorporates dTTP (red), dCTP (blue), dATP (green) and dGTP (black), while sequencing the first four bases of a template (TCAG). For the smaller well size a 1 μm diameter bead was utilized. Library, template preparation, and loading were all done as described in the Supplementary Methods.

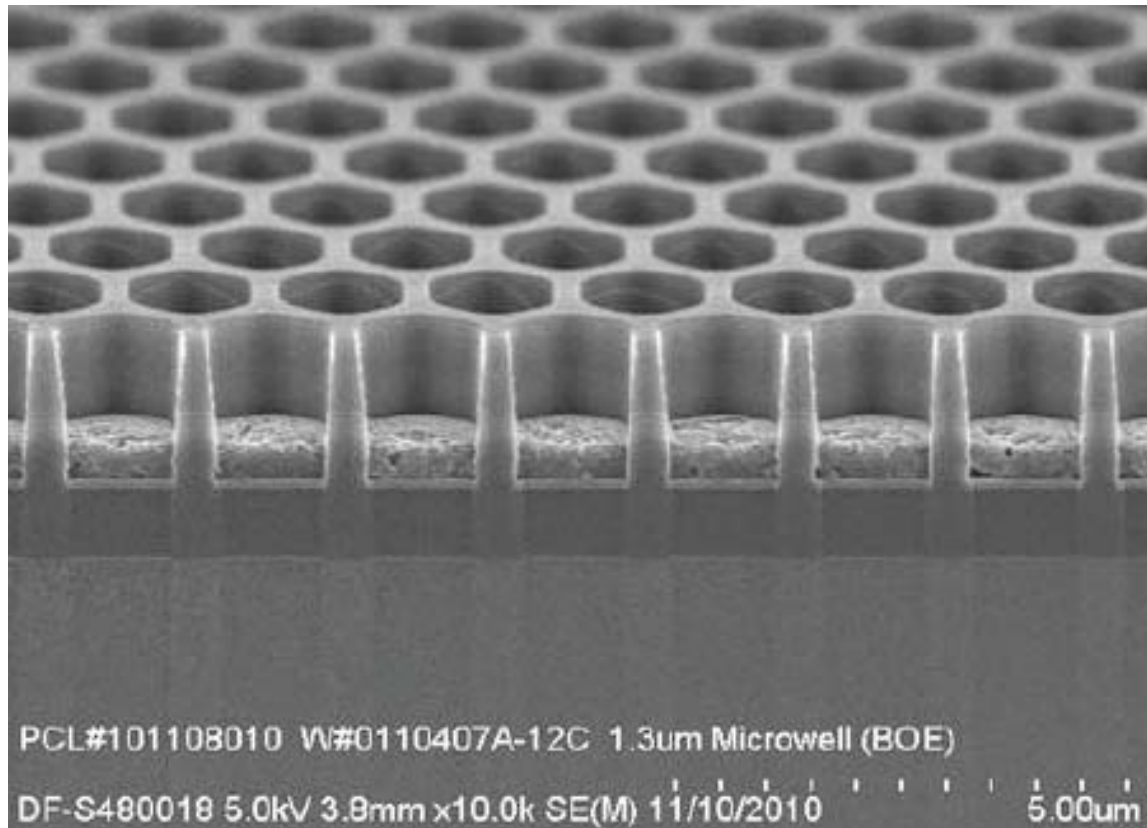


Figure S7 110 nm CMOS node

A scanning electron micrograph of a large array of 1.3 μm wells on a 1.68 μm pitch fabricated in the 110 nm CMOS technology node. This geometry allows for a 2-transistor ion chip of 165 M sensors and supporting electronics to fit on a 23.7 mm x 20.0 mm die. For example reducing the sensor's transistor count to one and increasing the die size to 31.7 mm x 25.8 mm would enable a 1.1 billion-sensor chip (0.5 μm well on a 0.8 μm pitch).

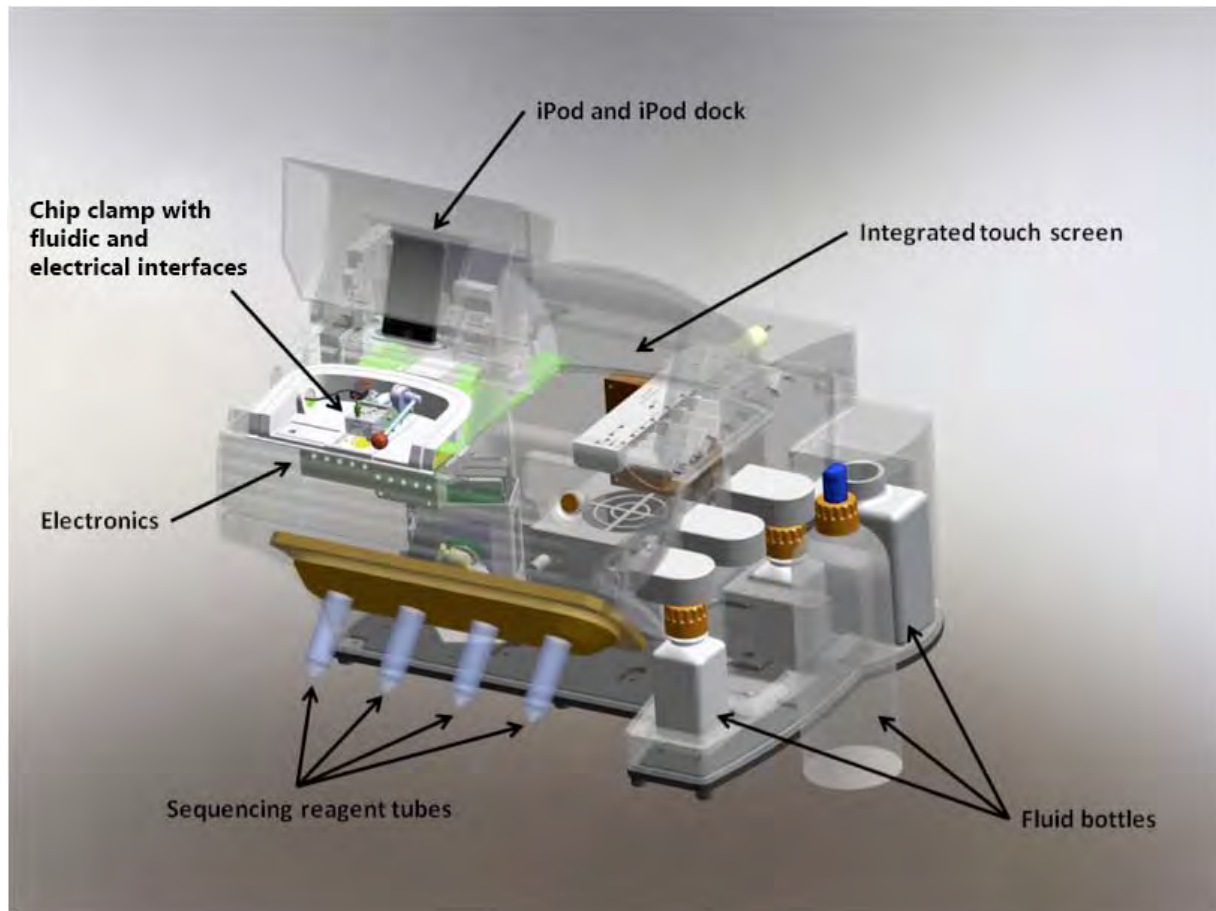


Figure S8 Ion sequencing instrument

Instrument overview. A **Chip clamp** supports the ion chip and provides both **fluidic** and **electronic interfaces**. A fluidic system delivers nucleotides from the four **Sequencing reagent tubes** or wash solutions from the **Fluid bottles** to the ion chip. Reader board **Electronics** sequentially address each sensor in the array, and simultaneously controls the reagent delivery. The **Integrated touch screen** allows for setting up and starting sequencing runs and the **iPod and iPod Dock** allows for remote monitoring of the sequencing machine.

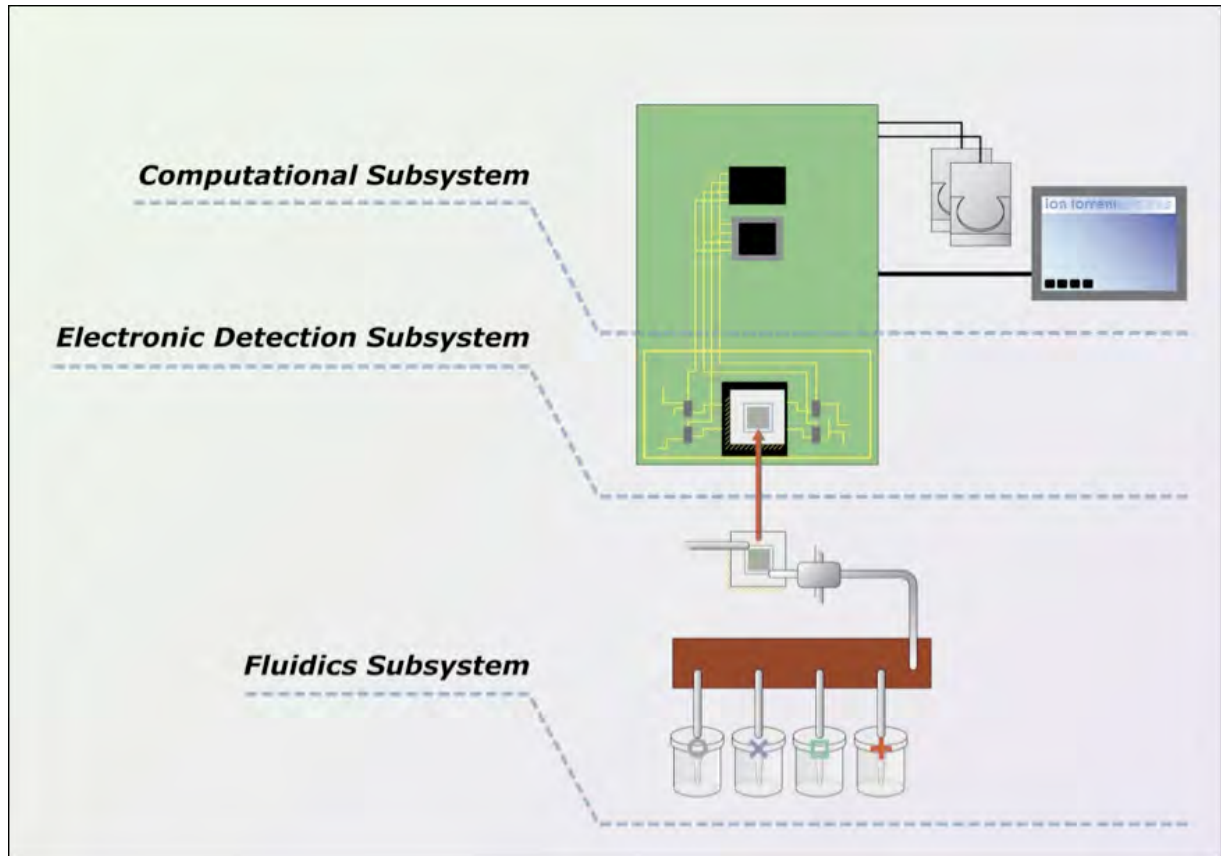


Figure S9 Subsystem architecture

The instrument is comprised of three subsystems. *Computational Subsystem*, simultaneously drives valves for fluidic control, initiates and manages data collection, and stores raw data. *Electronics Detection Subsystem*, collects data from the semiconductor-sequencing chip. *Fluidics subsystem*, for supply of nucleotides and washing reagents to the chip.

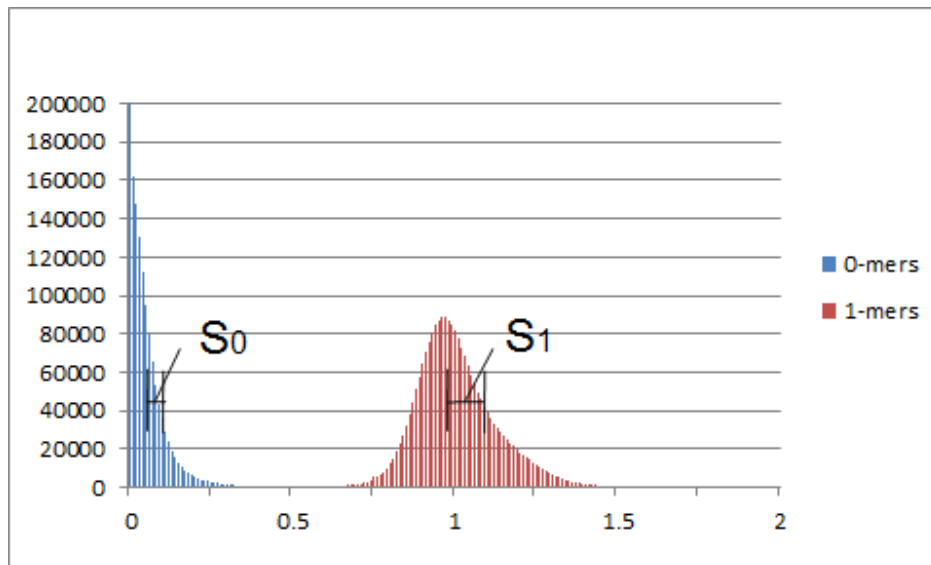


Figure S10 Signal to noise ratio

Signal to Noise Ratio (SNR) is calculated using the known expected 0-mer and 1-mer incorporation signals in the key portion of the library sequence. From the 1.2 M Chip *E. coli* run in Table 1 the resulting SNR is 10.1. SNR is calculated using the formula:

$$\text{SNR} = (M1 - M0) / [(S1 + S0) * 0.5]$$

Where $S0$ is the standard deviation of the 0-mer signals, $S1$ is the standard deviation of the 1-mer signals, $M0$ is the mean of the 0-mer signals, and $M1$ is the mean of the 1-mer signals.

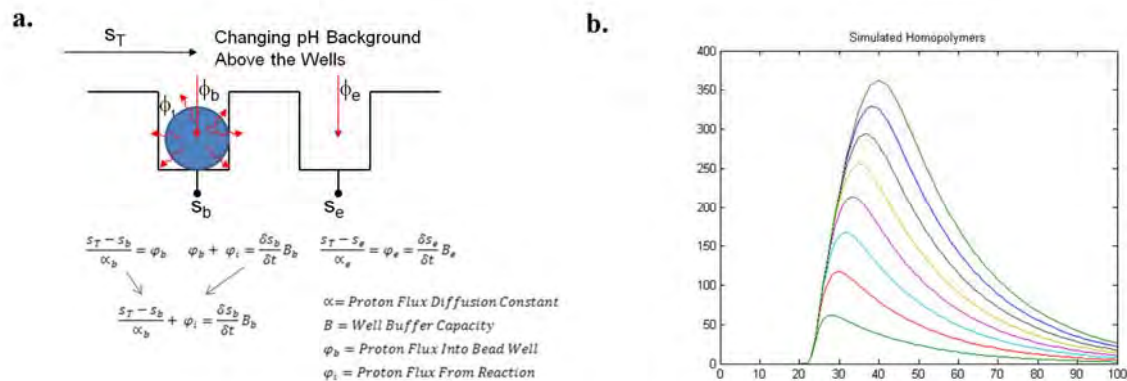


Figure S11 Physical model

a. The physical model of the well relates the measured signal in wells containing beads (S_b) and empty wells (S_e) to the flux of protons between those wells and the bulk fluid (φ_b and φ_e) as well as the flux of protons generated by the incorporation reaction in the bead-containing wells φ_i . The flux of protons between the wells and the bulk is assumed to be proportional to both the concentration gradient and to the difference between the measured signal and an unknown bulk signal S_T . **b.** The proton flux from incorporation, φ_i is simulated from a model of the reaction that combines a Michaelis–Menten kinetic model with the diffusion of the dNTP reactant into the well (colored line represent increases in the number of bases in a homopolymer).

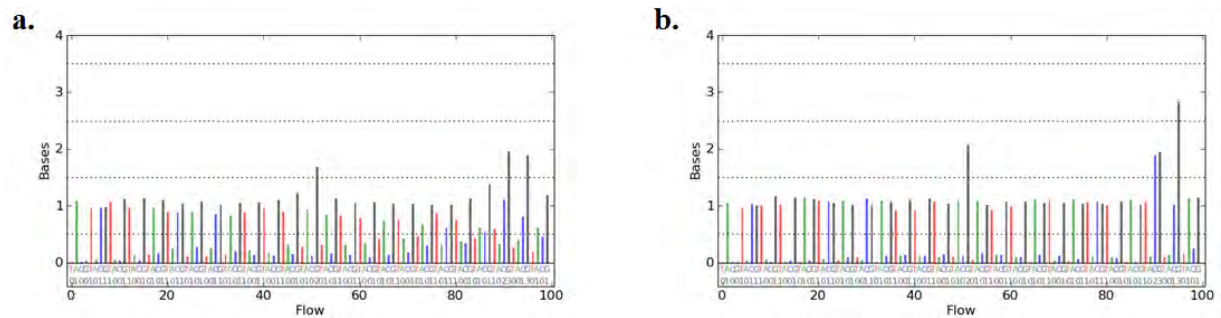


Figure S12 Phase correction

a, Raw measures of incorporation from the first 100 nucleotide flows are shown. These measures are the output of the physical model, prior to correcting for phase and signal loss. Phase errors are evident, especially in the expected 0-mer flows. We estimate the phase and signal loss parameters using these raw measurements and expected incorporations. **b**, The estimates of the magnitude of signal loss and de-phasing are then used in an algorithm that reconstructs the in-phase signal and simultaneously estimates the associated base calls. Bases are called from the corrected measurements simply by rounding each measurement to the nearest integer.

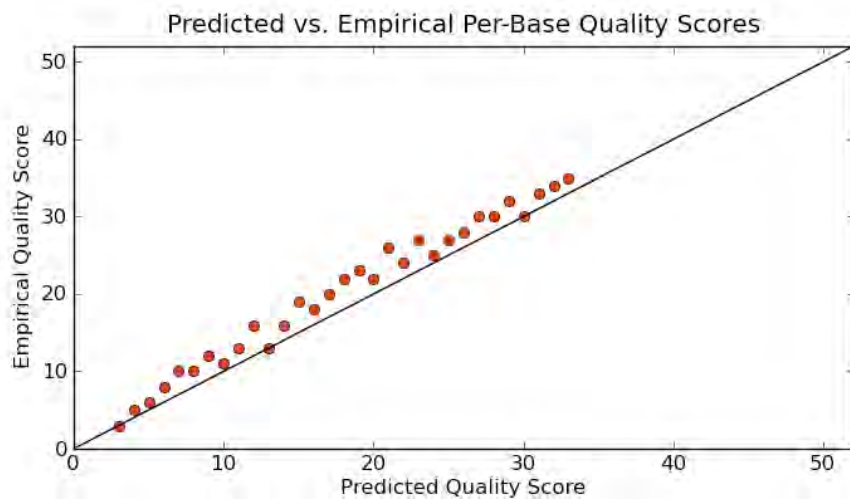


Figure S13 Quality scores

The relationship between observed and predicted quality using 1.2 M ion chip *E. coli* data (Table 1). A perfectly calibrated set of quality scores would lie along the diagonal. The quality scores are systematically slightly under-predicting the actual quality. To compute the empirical accuracy for a predicted quality score the corresponding base calls are evaluated by comparison to a reference sequence. The number of bases for which the accuracy is typically computed is on the order of millions of bases, so error bars are smaller than the circle used to plot the score.

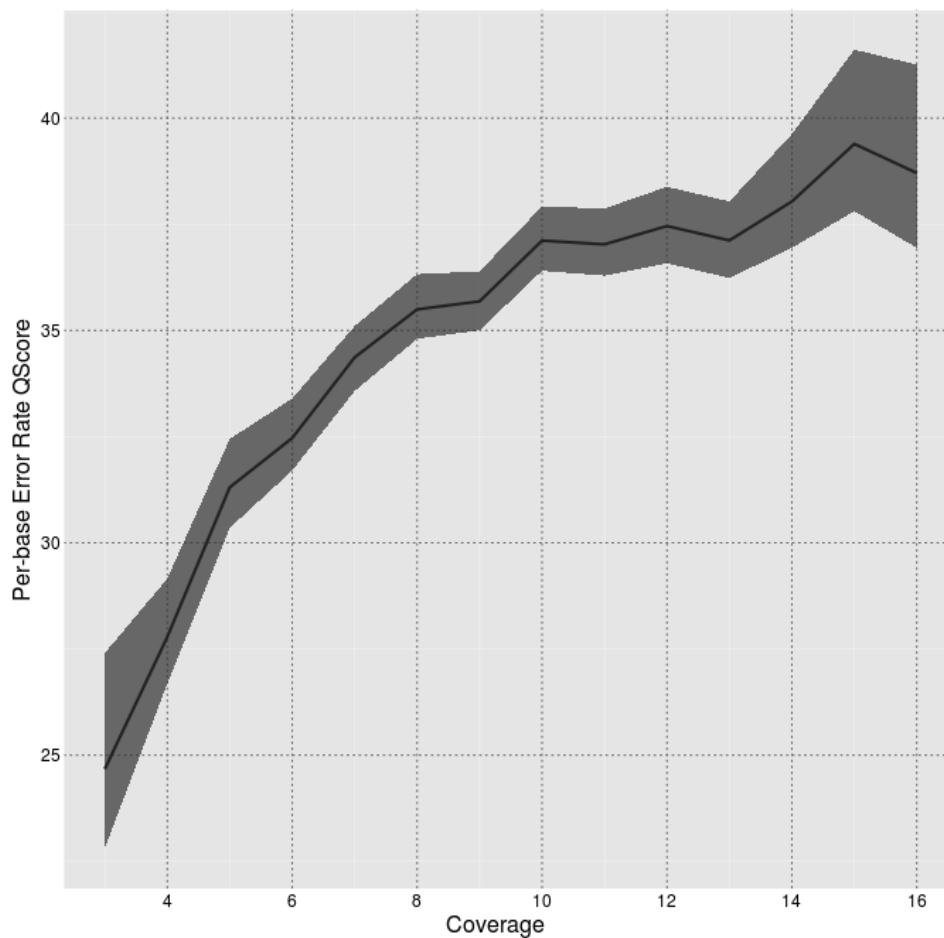


Figure S14 Consensus accuracy vs. coverage depth

Plot of error rate of consensus sequence at different levels of coverage depth for the 1.2 M ion chip *E. coli* data (Table 1). Evaluating across all positions in the genome, the overall consensus coverage was 99.996% with a consensus accuracy of 99.97%. The 1228 observed errors in the consensus sequence break down into 20 transitions, 24 transversions, 13 insertions and 1171 deletions, 10 of the deletions are 2bp in length, the rest are all single-base deletions. Results for the 6.1 M ion chip were 99.94% consensus accuracy with 2852 errors: 2 transitions, 2 transversions, 1 insertion, and 2847 1 base deletions. Results for the 11 M ion chip were 99.99% consensus accuracy with 415 errors: 1 transition, 1 transversion, 1 insertion and 412 1 base deletions.

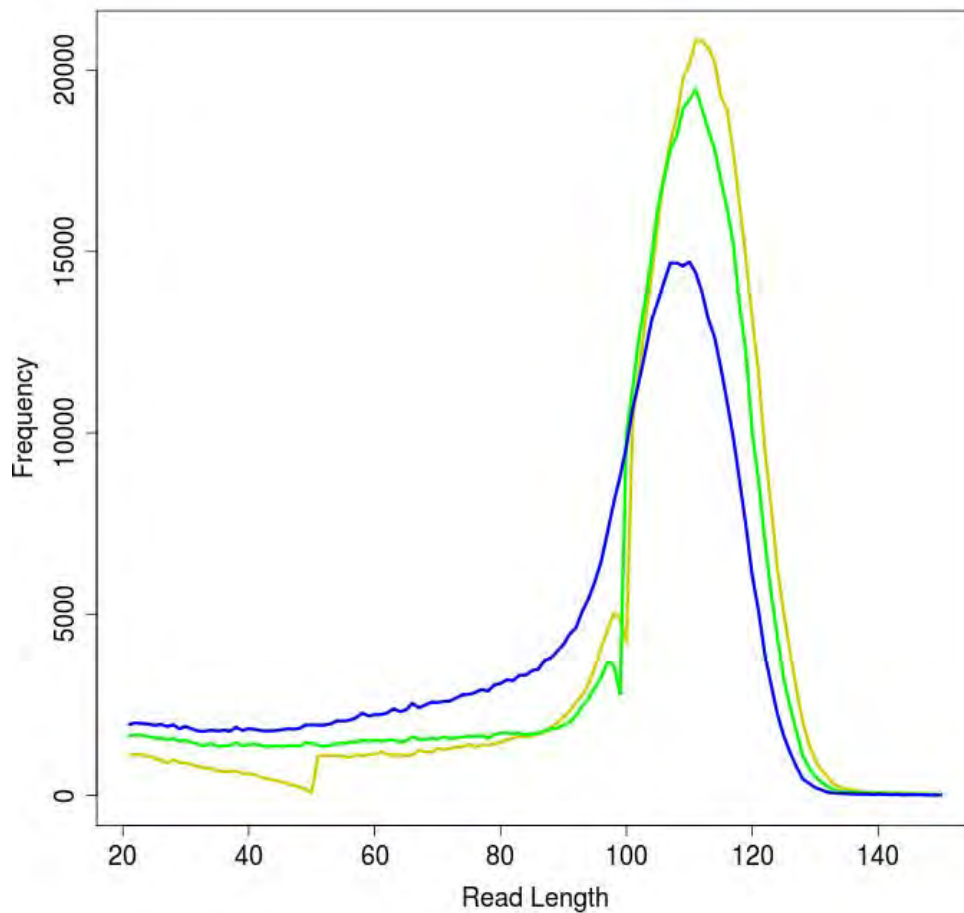


Figure S15 Distribution of read length

Distribution of aligned read lengths at 98% (yellow line), 99% (Green line), and 100% accuracy (blue line) from the 1.2 M ion chip *E. coli* dataset (Table 1). The read length at a given accuracy is defined as the maximal position in the read at which the total accuracy exceeds that value. Reads shorter than 21 bases are not considered.

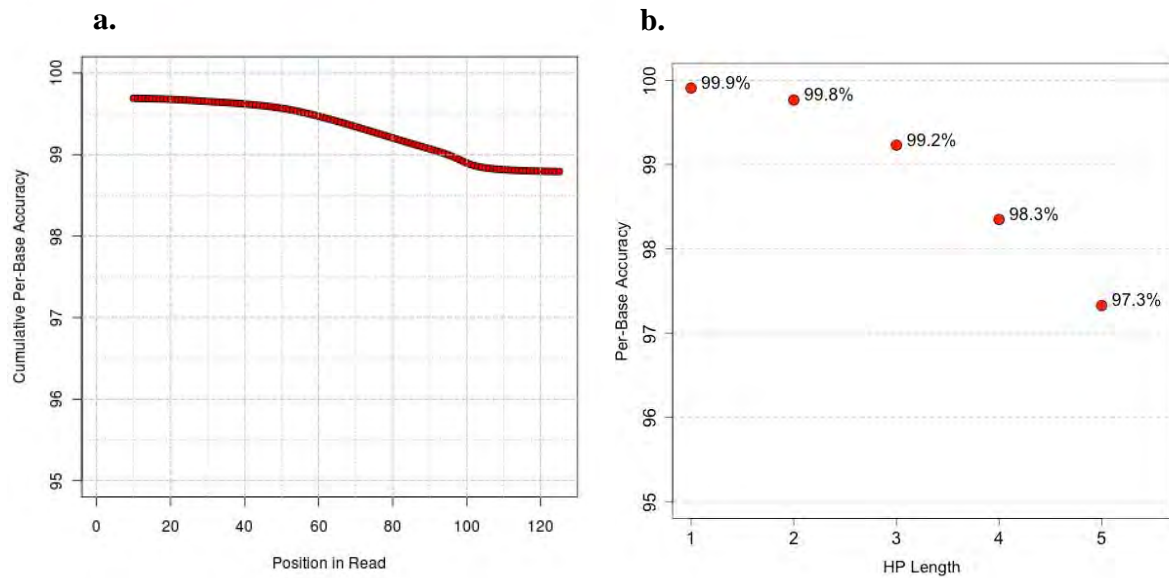


Figure S16 Single read accuracy

a. Single reads aligned to high quality reference, with the average read accuracy as a function of base position using the 1.2 M chip *E. coli* data (Table 1). Taking the first 50 bases of every read, the per-base accuracy is 99.569% \pm 0.001% and taking the first 100 bases it is 98.897% \pm 0.001%. **b.** Summary of the same dataset, showing the per-base error rate stratified by the length of the homopolymer (HP) in which the bases are located. For bases in homopolymer runs of length 5 the per-base accuracy is 97.328% \pm 0.023%.

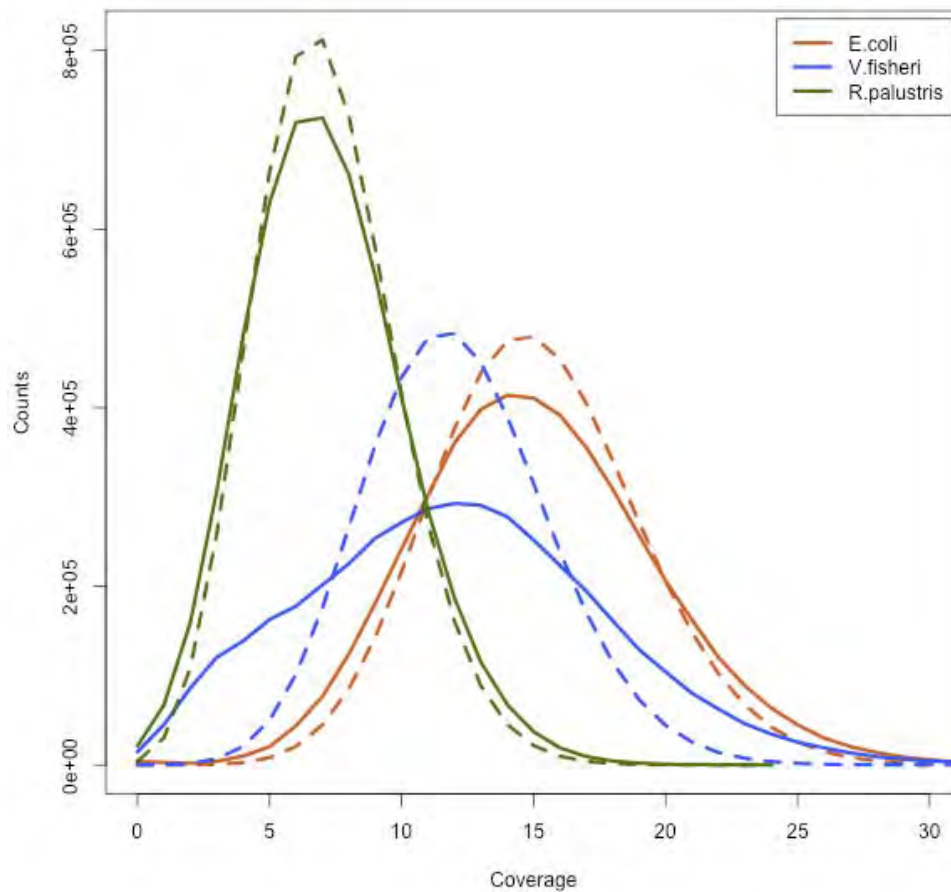


Figure S17 Uniform coverage in bacterial genomes

For each bacterial genome from Table I, the observed distribution of per-base coverage depth (solid lines) is compared with the theoretical distribution that would result from a Poisson process with the same mean (dashed lines). For *E. coli* (red) and *R. palustris* (green) the correspondence between the expected and observed distributions is excellent, reflective of the uniform and random nature of the coverage. For *V. fisheri* (blue) the observed over-dispersion is expected because its two chromosomes are maintained at different copy numbers in the cell and genomic library construction captures finer-scale copy number variation related to the origin of replication⁴⁸.

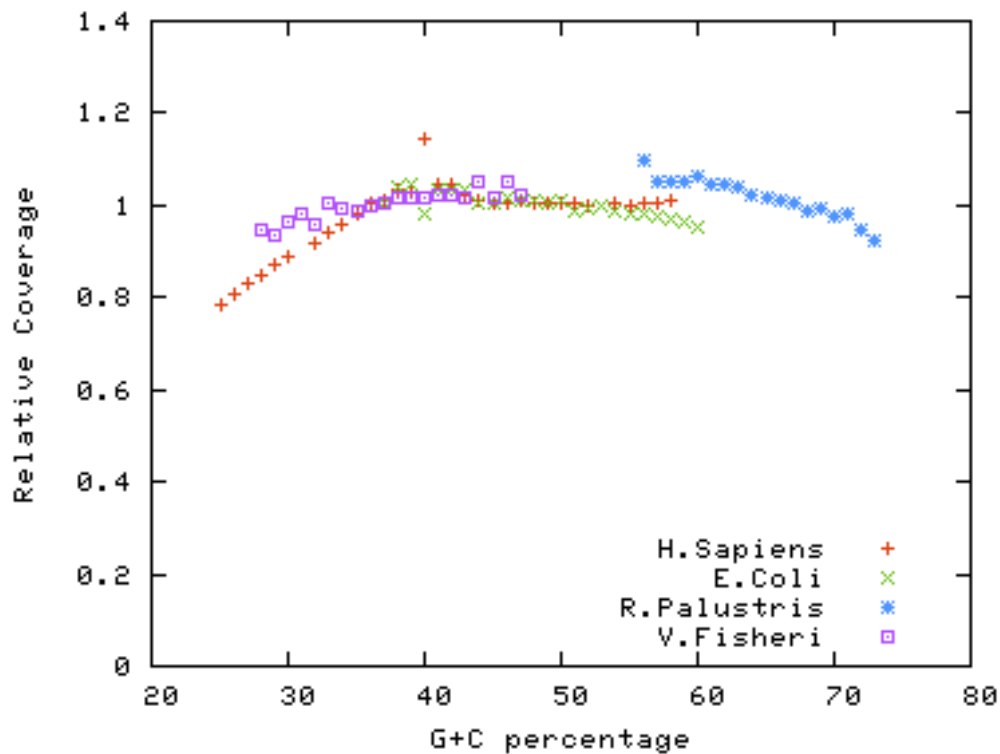
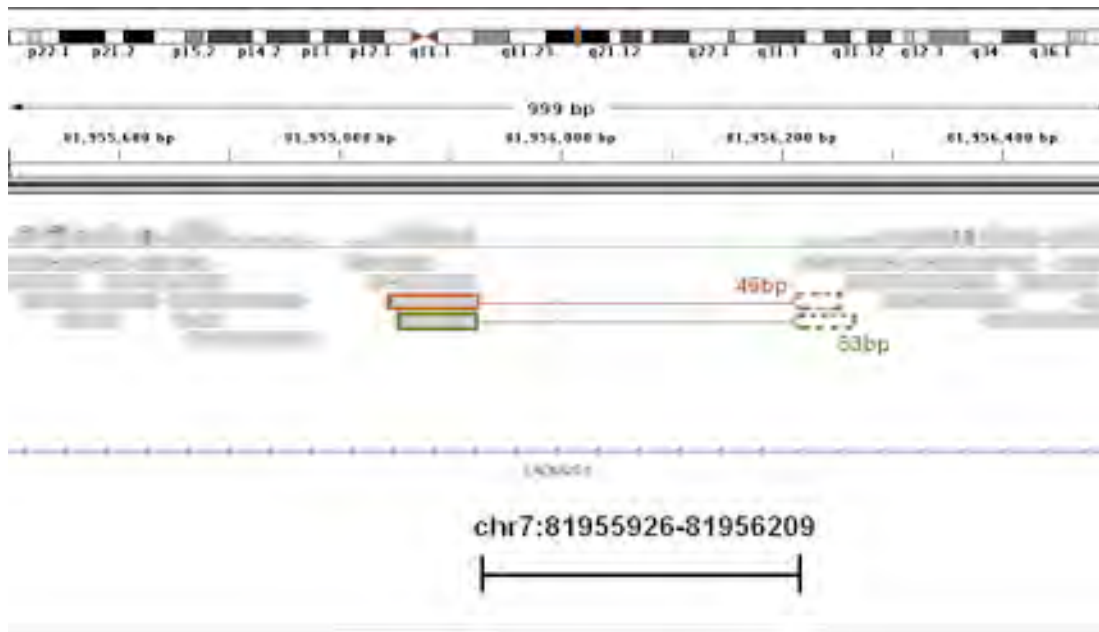
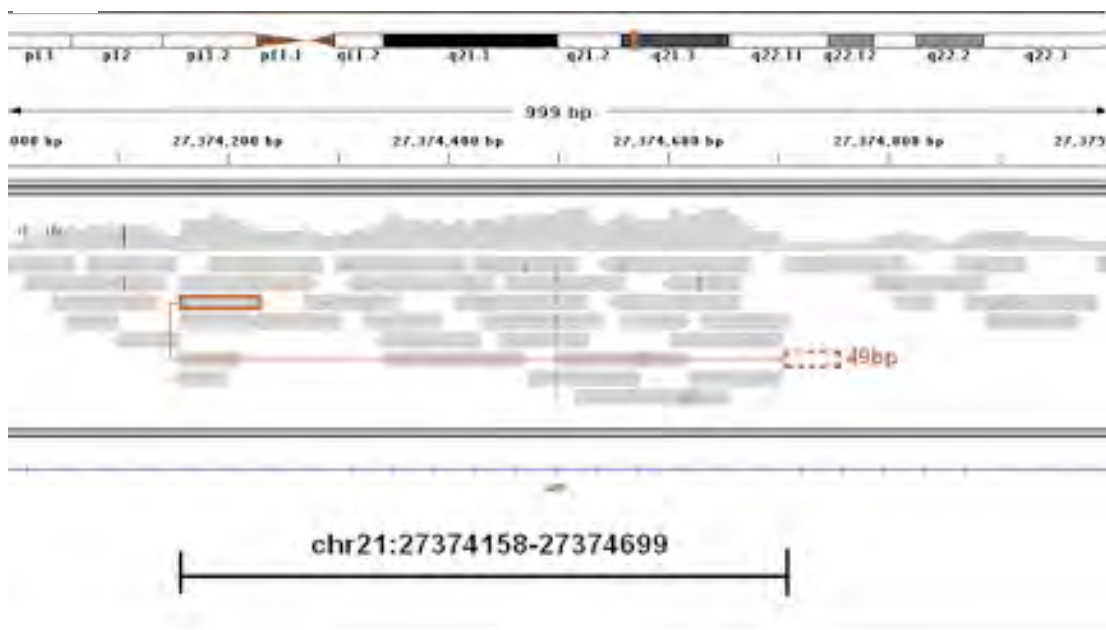


Figure S18 Uniform coverage across range of GC content

Sequencing coverage vs G+C% shows a very even distribution across the four genomes sequenced: *H. sapiens* (red), *E. coli* (green), *R. palustris* (blue) and *V. fisheri* (magenta). For each genome, G+C% is computed and used to group all non-overlapping 100-bp windows. The mean number of reads mapped per window is calculated for each group and is divided by the global mean. The resulting ratio is plotted against G+C%. Bias towards over- or under-representation as a function of GC content shows up as a deviation of the ratio from a value of 1.

a.**b.****Figure S20 Deletions & inversion examples.****a**, Screen shots of a deletion and **b**, an inversion.⁴⁹

Supplemental Tables

Ion Genotype	SOLiD Genotype	Total	percent same genotype	percent in dbSNP132	Transition/Transversion Ratio
het	het	1,061,797	99.95%	97.05%	2.2
het	hom	32,017	-	96.81%	2.0
het	not called	202,888	-	85.70%	2.1
hom	het	68,826	-	96.66%	1.9
hom	hom	1,077,756	99.97%	99.61%	2.1
hom	not called	155,699	-	92.08%	2.1
not called	het	900,709	-	60.66%	1.2
not called	hom	141,014	-	79.66%	1.3

Table S2 Confirmation of ion genotype by SOLiD sequencing

A comparison between variant SNP calls between Ion and SOLiD datasets. The first two columns show the genotype call in Ion and SOLiD data respectively. The third column shows the total number of SNPs corresponding to each row. In cases where both datasets call the same type of SNP (heterozygote or homozygous variant) the proportion for which the genotype call is the same is shown in column 4. The last two columns show the proportion of variants in each row that are also present in dbSNP132⁵⁰ and the transition/transversion ratio, respectively.

Source	dbSNP ID	Gene name	Phenotype	Reference Alleles*	Variant Alleles
23&Me	rs10195871	BCL11A	Adult subjects with this genotype tend to produce some fetal hemoglobin, which may reduce the severity of sickle cell anemia or thalassemia in people with these diseases	6	7
23&Me	rs10427255		Slightly lower odds of having the photic sneeze reflex	0	14
23&Me	rs1051730	CHRNA3	Likely to smoke one more cigarette per day on average than the typical amount	5	2
23&Me	rs12913832	HERC2	In Europeans, 56% chance of brown eyes; 37% chance of green eyes; 7% chance of blue eyes.	7	5
23&Me	rs1953558		Typical sensitivity to the sweaty smell of isovaleric acid	5	6
23&Me	rs2153271	BNC2	Typical amount of freckling	13	3
23&Me	rs363050	SNAP25	Non-verbal IQ performance three points higher on average	1	24
23&Me	rs4481887		Moderately higher odds of smelling asparagus in one's urine	7	7
23&Me	rs4988235	MCM6	Likely to be lactose tolerant due to lactase persistence. Higher adult lactase enzyme levels	2	6
23&Me	rs6060371	UQCC	On average 0.3 - 0.7 centimeters taller than typical height	0	17
23&Me	rs713598	TAS2R38	Can taste certain bitter flavors	4	6
OMIM	rs1045644		Lowered risk for deafness (Ménière's Disease)	5	9
OMIM	rs10970979		Increased risk of mental retardation	7	10
OMIM	rs16910526	CLEC7A	Fungal nail infection risk (onychomycosis)	6	3
OMIM	rs17673268		Increased risk of mental retardation	12	10
OMIM	rs4775765		Risk of Weill-Marchesani Syndrome	0	8
OMIM	rs497116	CASP12	Sepsis susceptibility	0	15

* Allele present in the Human reference genome hg19

Table S3 Selected Moore genome variants with phenotypic or disease annotation

For illustrative purposes the Online Mendelian Inheritance in Man database, and the 23andMe functional SNP collection was used to identify a small subset of validated SNPs involved in human disease and interesting phenotypes.

Supplemental Methods

Ion sequencing

Gordon Moore provided written consent for the publication and release of his genetic sequence data in a personally identifiable manner. In addition he elected to obtain access to his own sequence information, as well as elected to have Life Technologies identify an expert to assist him in understanding this information.

Genomic DNA (Lofstrand, Gaithersburg, MD) from *Vibrio fischeri* (str. ES114), *Rhodospseudomonas palustris* (CGA009), and *Escherichia coli* (str. K12 substr. DH10B) was obtained from American Type Culture Collection (Manassas, VA) bacterial stocks. Genomic DNA was reconstituted in TE Buffer at a concentration of 0.3 ug/ul.

Human whole blood was drawn by a certified phlebotomist in 4 ml sodium citrate coated collection tubes and frozen at -80C until utilized. DNA extraction was conducted on individual 4 ml aliquots of the blood, using the Qiagen FlexiGene DNA Kit (Valencia, CA) and associated manufacturer's protocol. The resultant genomic DNA was precipitated in isopropanol, dried under vacuum and stored as a lyophilized pellet until required. Prior to utilization, the genomic DNA was suspended in Qiagen Flexigen Buffer FG3 (Hydration buffer) at roughly 0.3 µg DNA/µl.

For each genome, 5 µg of the suspended DNA was converted into a genomic library for subsequent sequencing by following the process described in the Ion Fragment Library Kit (Life Technologies, Carlsbad, CA). Briefly, genomic DNA was fragmented via sonication to an average insert size ranging from 100-160 bases in length. Unique forward and reverse adapters were ligated to the inserts. The resulting template pool was size selected to remove unincorporated primers on a Sage Biosciences PippinPrep (Beverly, MA); the final libraries ranged in size from 160 to 220 bases, with a median size of 180-205 bp.

Size selected libraries were clonally amplified as described in the Ion Template 314™ Kit (Life Technologies). Briefly, the genomic library was added to the PCR reaction mix at a limiting dilution. Beads containing DNA oligos (2 µm IonSphere acylamide beads, Life Technologies) were emulsified along with the template molecules and then subjected to PCR amplification. Following amplification, the emulsions were broken to release the beads from the oil, and template-carrying beads were separated by magnetic bead enrichment (Dynabeads M-280 Streptavidin, Dynal Corporation, Oslo, Norway) from beads without template.

Enriched template-carrying beads were then primed (annealing buffer PBS + 0.2% Tween-20 + 0.02% sodium azide) followed by the addition of BST polymerase (Life Technologies), and then loaded into the ion chip (fabricated by Plessey Semiconductor, Plymouth, UK, & X-Fab Semiconductors foundries AG, Erfurt, Germany) for subsequent sequencing on the ion instrument. For all three chip sizes approximately 10 M enriched beads are loaded.

Sequencing was done using the Ion Sequencing kit, according to the Ion Torrent user guide (Life Technologies), using all natural nucleotides as supplied in the kit. Nucleotides are used at a final concentration of 50 μ M and sourced from MyChem LLC (San Diego, CA). The nucleotides are obtained by chromatographically separating the 4 nucleosides from hydrolyzed DNA followed by chemical phosphorylation of the purified nucleosides. Further purification after phosphorylation includes 3 cycles of reverse phase affinity and ion exchange chromatography

Sequencing was done using the Ion Sequencing kit, according to the Ion torrent user guide (Life Technologies), using all natural nucleotides as supplied in the kit. Nucleotides are used at a final concentration of 50 μ M. Nucleotides as supplied in the kits have been purified by ion exchange and reverse phase chromatography to remove any contaminating bases. The wash solution used between nucleotide additions is 6.4 mM MgCl₂, 13 mM NaCl, 0.1% Triton-100 at pH 7.5.

Data capture for all three chip sizes were obtained without any changes to the instrument (Ion Chip Sequencing Protocol, Life Technology). The instrument automatically detects the size of the chip being used, and adjusts the fluidics, and signal processing, and data analysis accordingly.

SOLiD sequencing

1 μ g of human DNA purified from blood was sheared to generate a fragment library primarily in accordance to manufactures instructions SOLiD Fragment Library Constructions reagents #4443713 (Life Technologies, Carlsbad, CA). DNA was amplified in emulsion PCR and sequenced on a SOLiD 4 instrument according to manufacturers instructions. Paired end 50 bp x 35 bp reads were used with Exact Call Chemistry to generate 15-fold coverage. SOLiD data was analyzed and variant called using Nimbus Informatics⁵¹ on the Amazon EC2 cloud with BFAST. Specifically, BFAST was employed for color-space alignment to hg19⁵², SRMA for color-space local realignment⁵³, and Samtools to establish the final variant calls⁵⁴.

Mapping ion sequence to reference

Sequencing reads for each genome were mapped to their corresponding genome reference: *Vibrio fischeri* (NC_006840), *Escherichia coli* (NC_000913.2), *Rhodopsuedomonas palustris* (NC_005296.1), and the human genome reference (NCBI Build 37). We utilized our own mapping software, TMAP, to identify high quality mappings as well as our own and open-source software to perform variant detection and validation on the human sample. TMAP was run with default settings. When a read aligned to multiple locations with equally scoring alignments one was selected at random.

For the Human genome, coverage was evaluated relative to NCBI Human genome build 37 (hg19) whose total size including the 22 autosomes, 2 sex chromosomes and the mitochondrial genome is 3,095,693,981 bases. Of these, 2,835,965,256 bases were covered at least once. Excluding 237,019,316 bases consisting of N's in runs of length 100 or greater, the coverage of the known genome is 2,835,965,256 / (3,095,693,981 - 237,019,316) or 99.21%.

SNP variation

SNPs were called in the Moore genome by running uniquely-mapped reads through samtools pileup⁵⁴ resulting in 2,598,983 variants of which 96.92% are also present in dbSNP132. The variants found break down into 1,296,702 heterozygotes and 1,302,281 homozygotes of which the fractions found also in dbSNP132 are 95.3% and 98.6% respectively. The Moore genome was independently sequenced on the SOLiD platform as a form of validation. Supplementary Table S1 summarizes 2,138,669 which are validated by the SOLiD sequencing. The variants called only in one of the two platforms tend to be found in dbSNP at much lower frequencies and tend to have more deviant Transition/Transversion ratios (Supplementary Table S2).

Structural variation

We collected 7,565 deletions and 128 inversion calls from 1000 Genomes consortium data⁵⁵ and localized on hg19⁵⁶. Flanking sequences surrounding the deletion or inversion junction, 125 bp on each side, were spliced together and used to map ion reads with TMAP's global mapping. We filtered out reads mapped below an alignment score of 10 on either side of junctions, reads with lower alignment scores across junctions than to normal human genome references, reads with no alignment to the half-junctions in the human genome, and reads associated with regions with alignment depth greater than 25. In total 16,907 reads passed the filters, typing 3,413 structural variants. As a control, the same procedure was repeated using random length-matched genomic regions instead of real structural variation calls, leading to only 2 reads mapped to 2 of the simulated constructs, from which we estimate a nominal positive predictive value of 99.94% (fraction of predictions estimated to be correct). A typed variant with N supporting reads is called heterozygous if there are at least $N/5$ reads mapped across its breakpoints on the normal reference, or, in case of a deletion only, there are at least $N/5$ mapped across its center. With this definition, 64.5% of the 3,413 variants are found to be heterozygous. 84.8% of the 3,413 variants are supported by at least two independent reads. Supplementary Fig. S20 shows a typical deletion (upper panel) and an inversion (lower panel) along with their supporting reads.

The resulting 3,413 structural variants typed (3,391 deletions and 22 inversions) were associated with the nearest RefSeq gene⁵⁷ if the gene is within 10kb (Supplementary Table S4). Each variant is annotated to indicate whether it sits outside of the gene (upstream or downstream), overlaps with an intron, or overlaps with an exon. Introns and exons are further separated by whether they are closer to the 5' or 3' end of the gene.

SI References

- 48 Srivastava, P. & Chattoraj, D. K. Selective chromosome amplification in *Vibrio cholerae*. *Molecular Microbiology* **66**, 1016-1028 (2007).
- 49 Robinson, J. T. *et al.* Integrative genomics viewer. *Nature Biotechnology* **29**, 24-26 (2011).
- 50 Database of Single Nucleotide Polymorphisms (dbSNP Build ID:132). National Center for Biotechnology Information National Library of Medicine. Available from <http://www.ncbi.nlm.nih.gov/SNP/> (2011).
- 51 Nimbus Informatics. <http://nimbusinformatics.com/public-website/> (2011).
- 52 Homer, N., Merriman, B. & Nelson, S. F. BFAST: an alignment tool for large scale genome resequencing. *PLoS One* **4**, e7767, doi:10.1371/journal.pone.0007767 (2009).
- 53 Homer, N. & Nelson, S. F. Improved variant discovery through local re-alignment of short-read next-generation sequencing data using SRMA. *Genome Biol* **11**, R99, doi:10.1186/gb-2010-11-10-r99 (2010).
- 54 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
- 55 1000 Genomes. <http://www.1000genomes.org/home> (2011).
- 56 Genome Reference Consortium (GRCh37. hg19 - NCBI Build 37.1). (2009).
- 57 Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35**, D61-65 (2007).