doi:10.1038/nature10242

Supplemental Figures

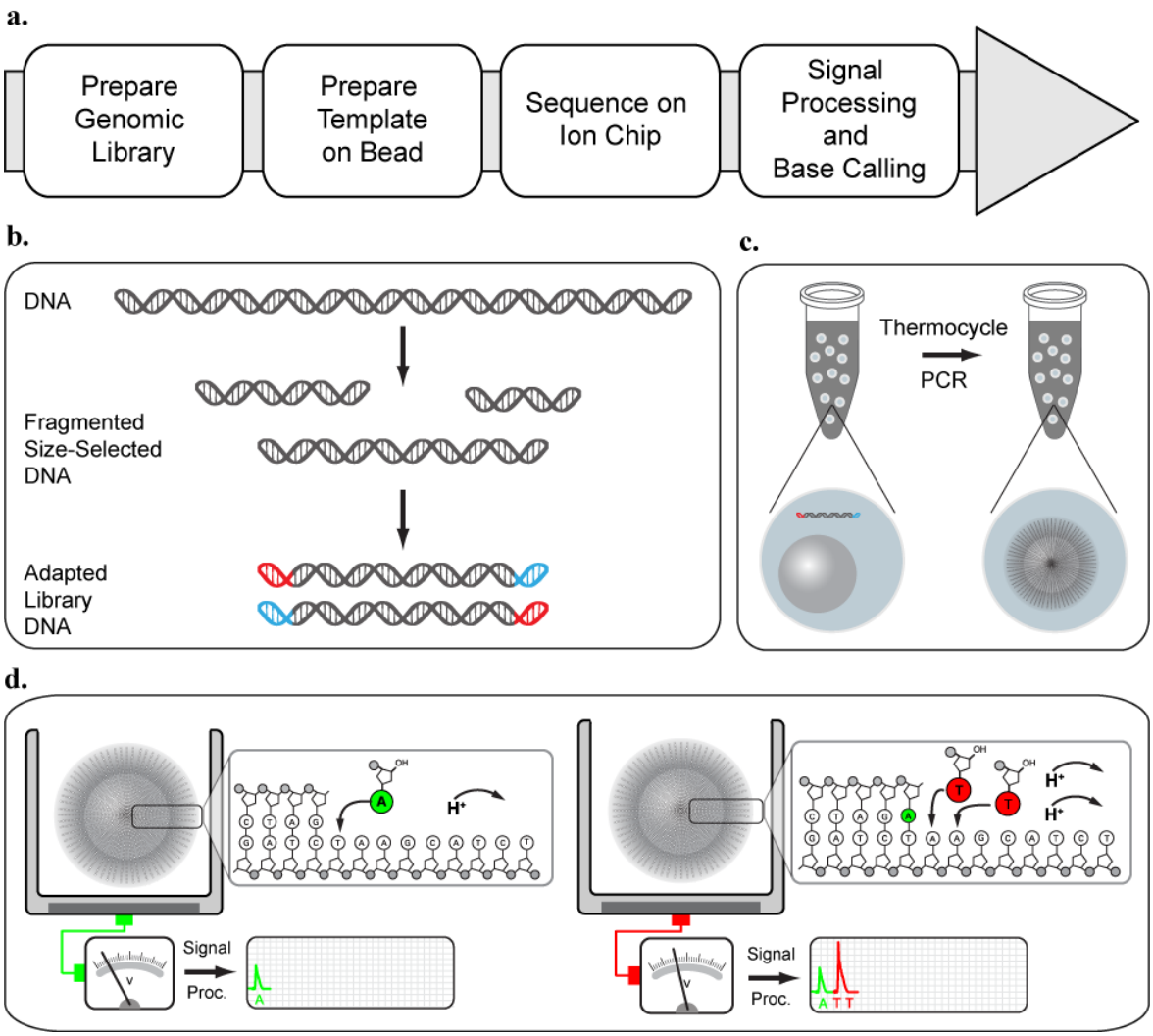


Figure S1 Process overview

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.

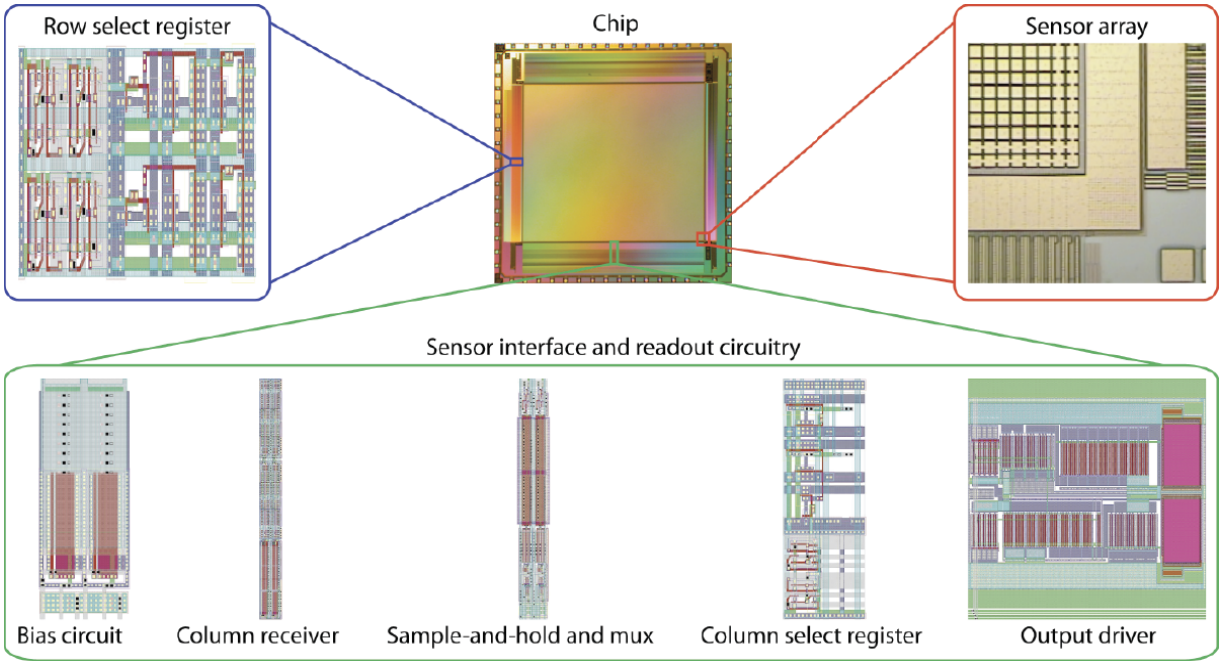


Figure S2 Chip architecture

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.

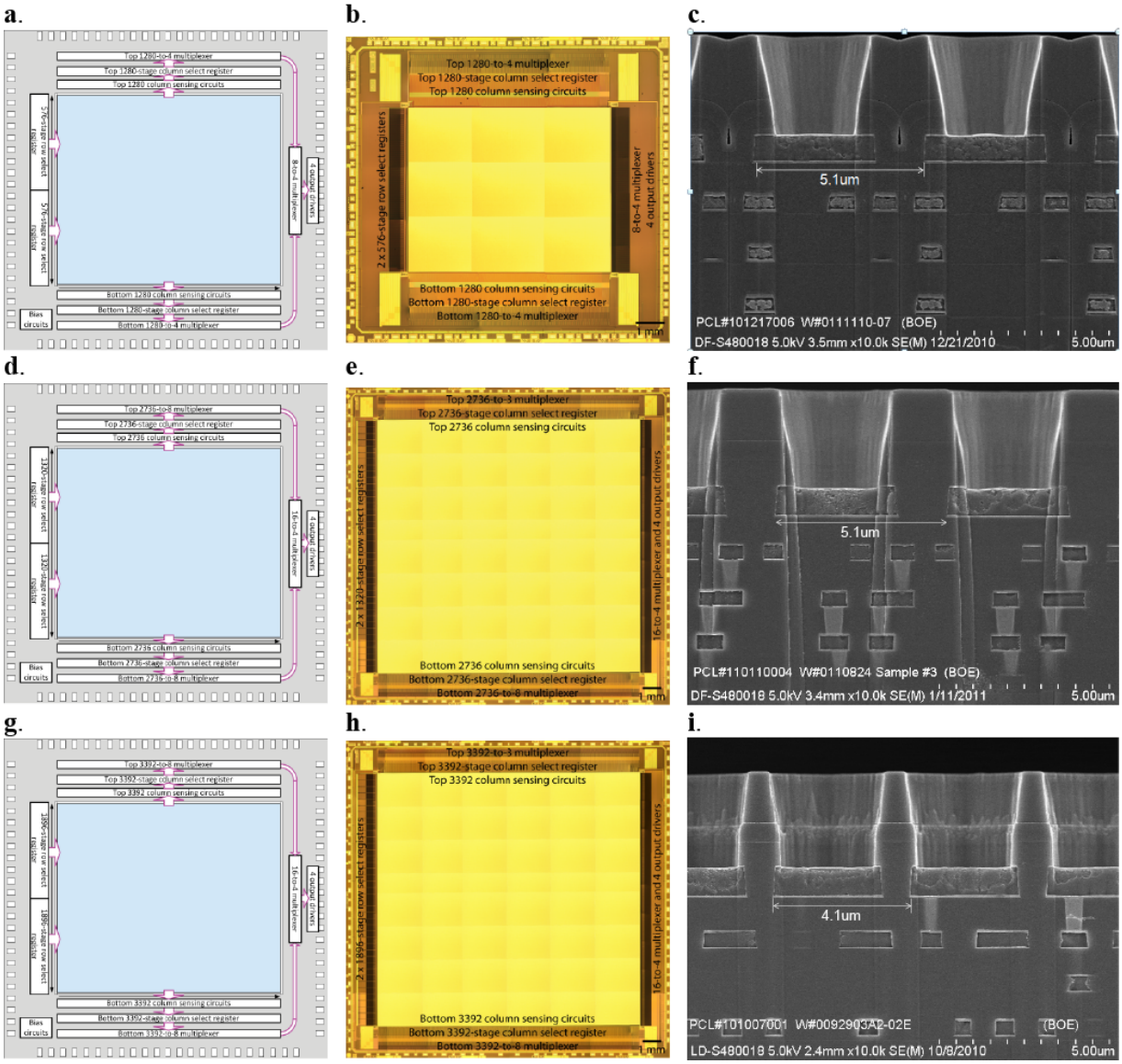


Figure S3 Block diagram, die, and well cross section

Three views of three ion chips are shown, block diagram, photograph of single die, and scanning electron micrograph cross-sectional images of two adjacent wells aligned to the underlying electronic structure. **a**, A chip with 1.5 M ISFETs has 1.2 M fluid-accessible sensors, **b,** measures 10.6 mm x 10.9 mm and **c**, has a 5.1 μm center-to-center pitch. **d**, A chip with 7.2 M ISFETs has 6.1 M fluid-accessible sensors, **e**, measures 17.5 mm x 17.5 mm and **f**, has a 5.1 μm center-to-center pitch. **g**, A chip with 13 M ISFETs has 11 M fluid-accessible sensors, **h**, measures 17.5 mm x 17.5 mm and **i,** has a 3.8 μm center-to-center pitch.

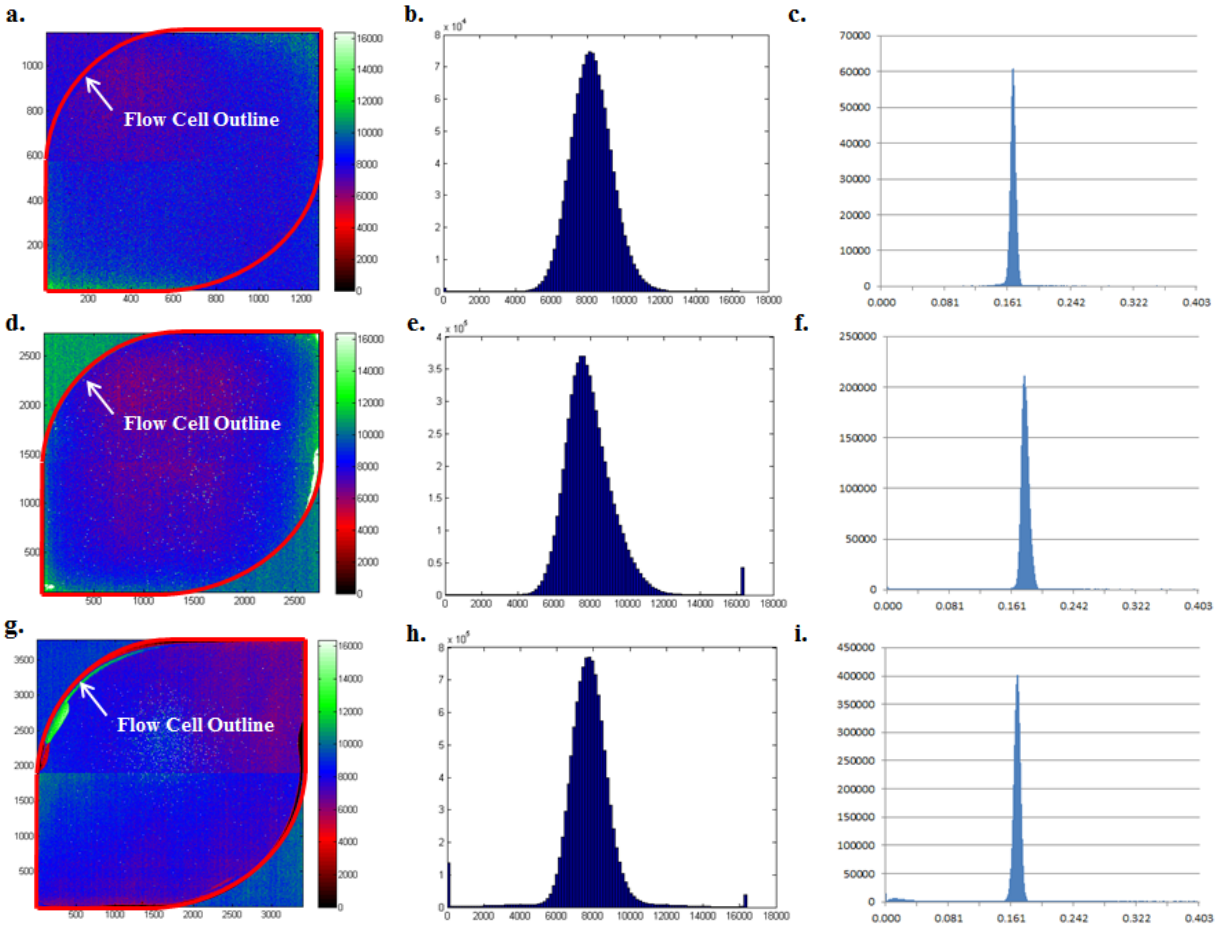


Figure S4 Large, uniform, electronically working, pH responsive sensor arrays

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.

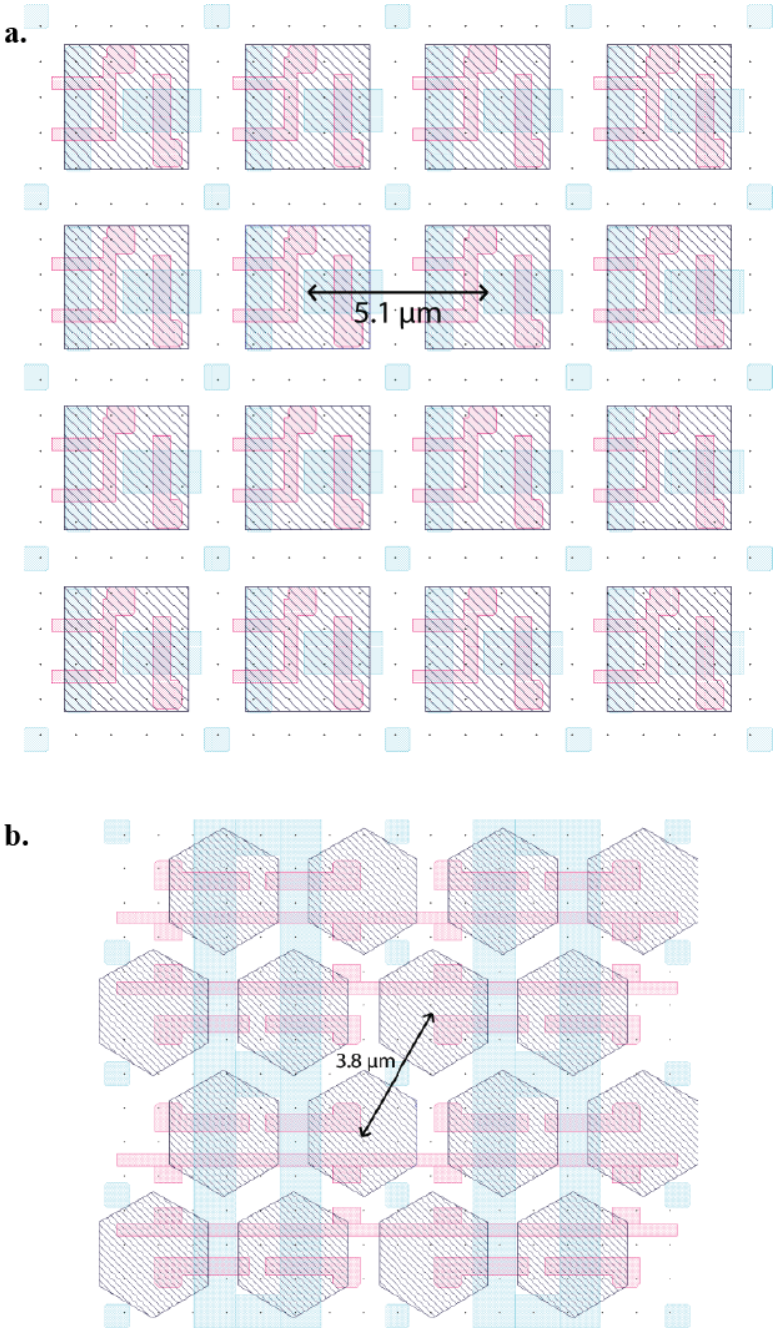


Figure S5 3-transistor and 2-transistor sensor

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.

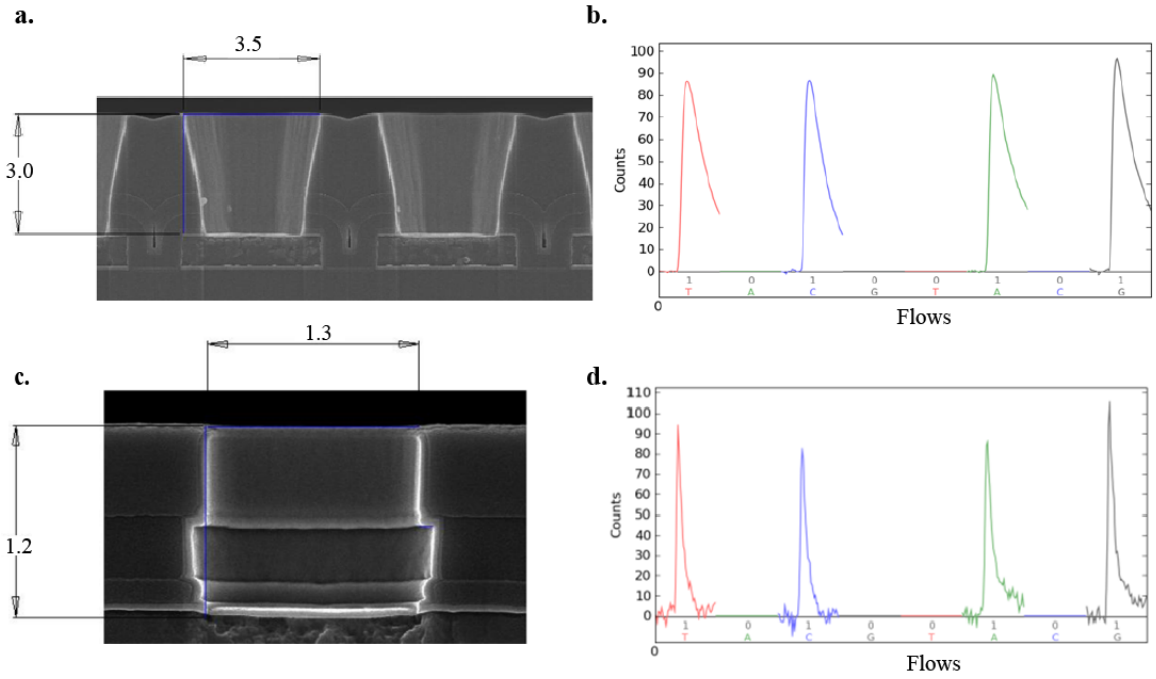


Figure S6 3.5 and 1.3 μm well SEMs and sequencing traces

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.

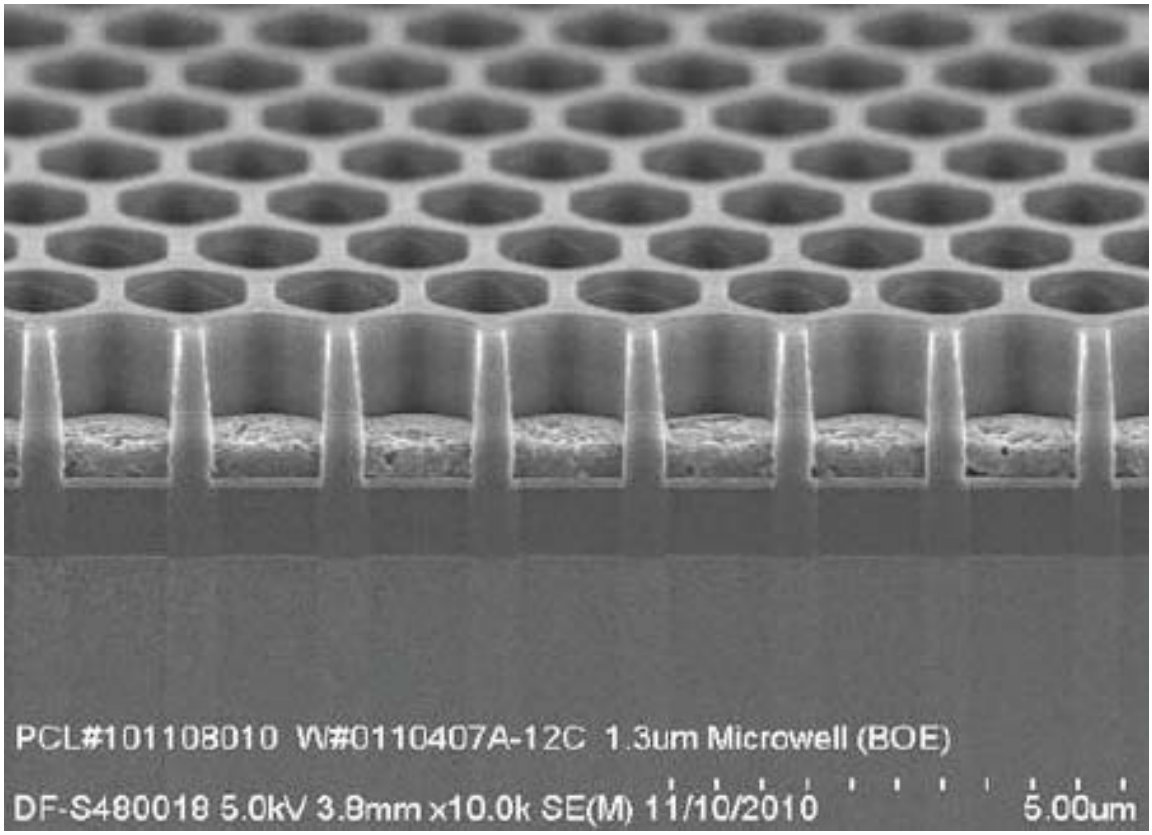


Figure S7 110 nm CMOS node

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).

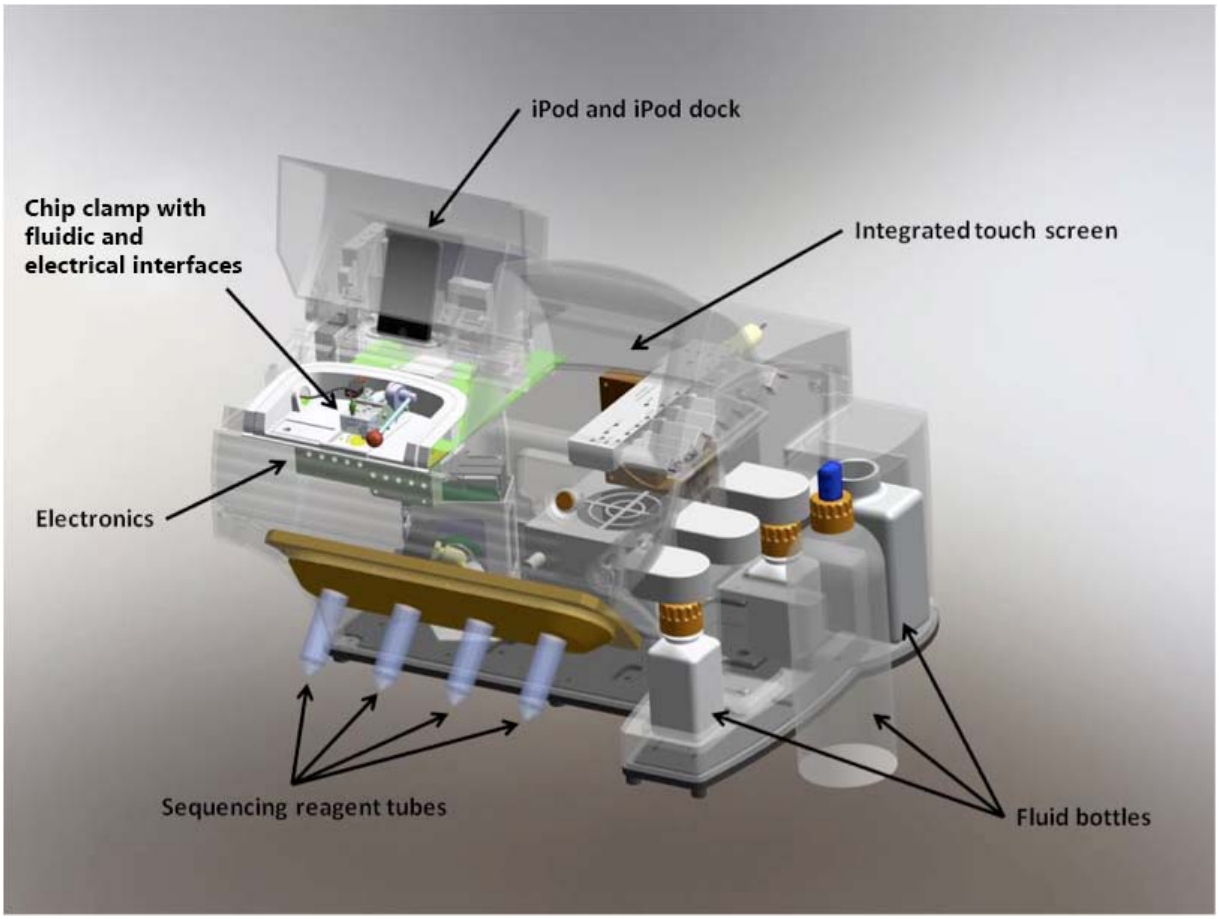


Figure S8 Ion sequencing instrument

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.

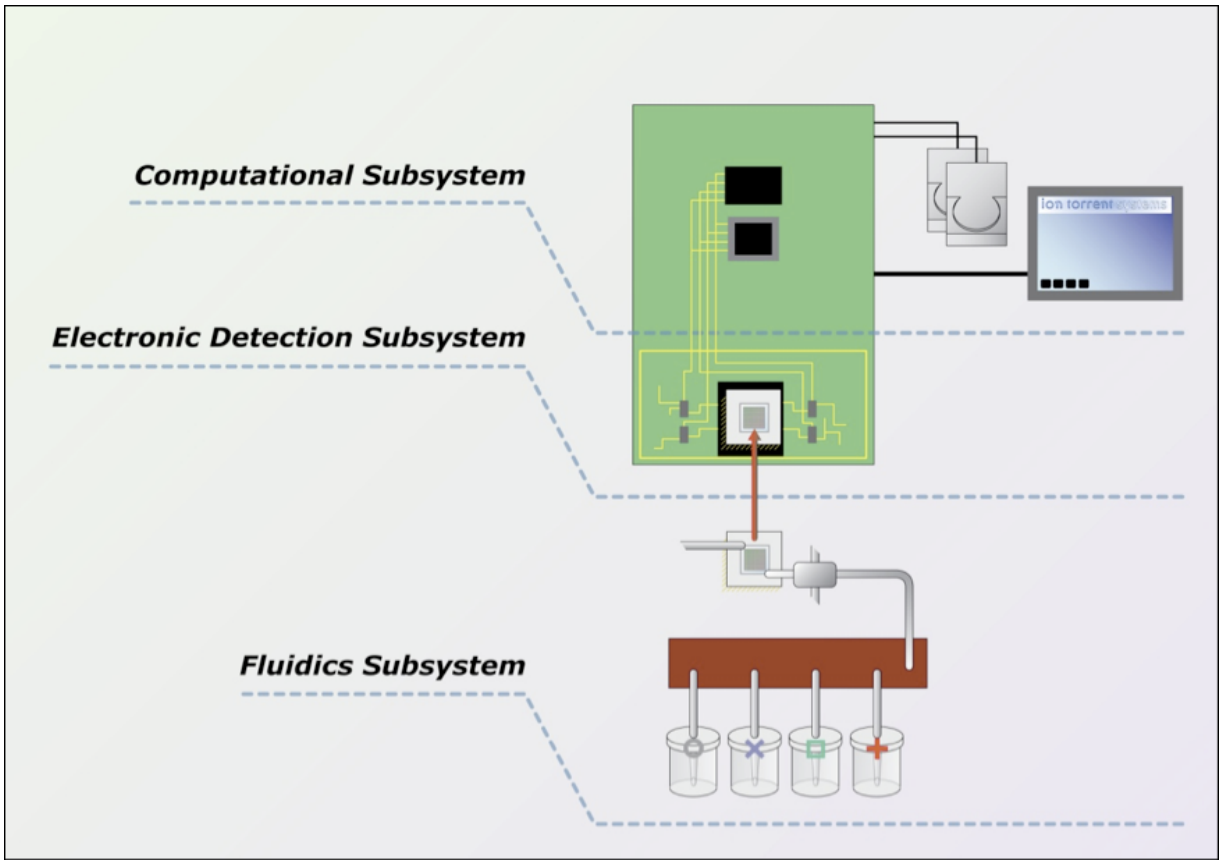


Figure S9 Subsystem architecture

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.

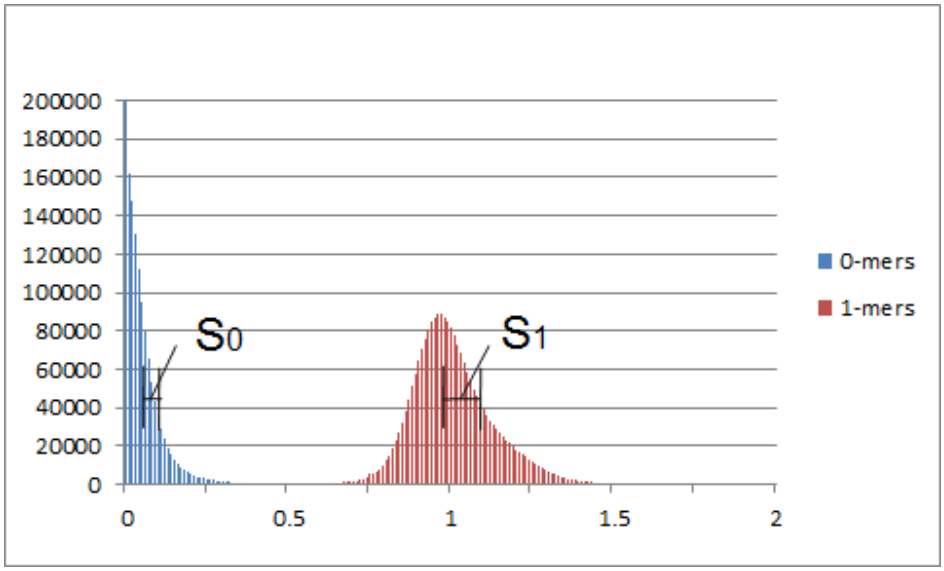


Figure S10 Signal to noise ratio

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:

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.

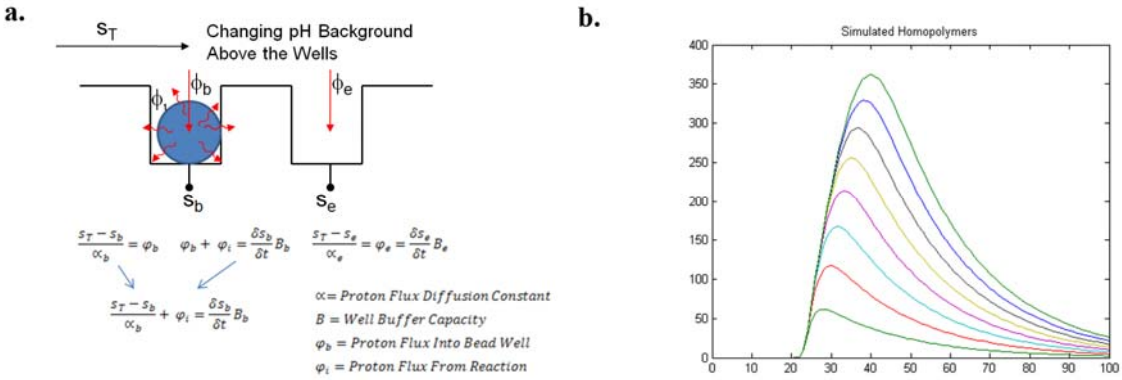


Figure S11 Physical model

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).

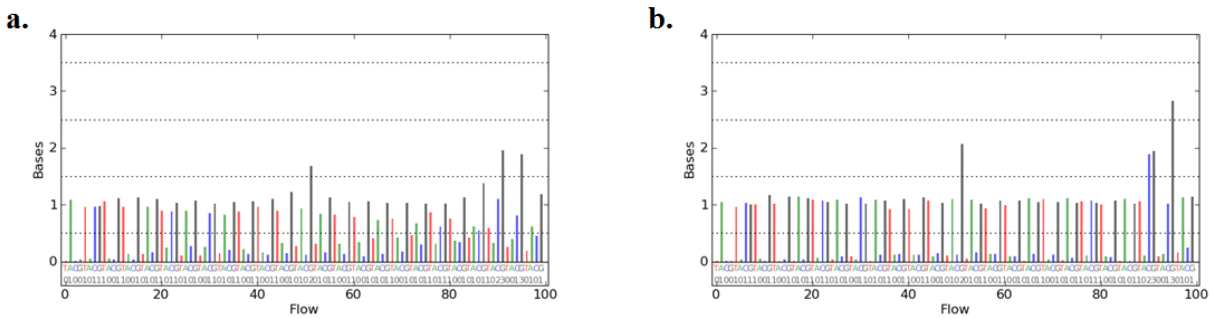


Figure S12 Phase correction

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.

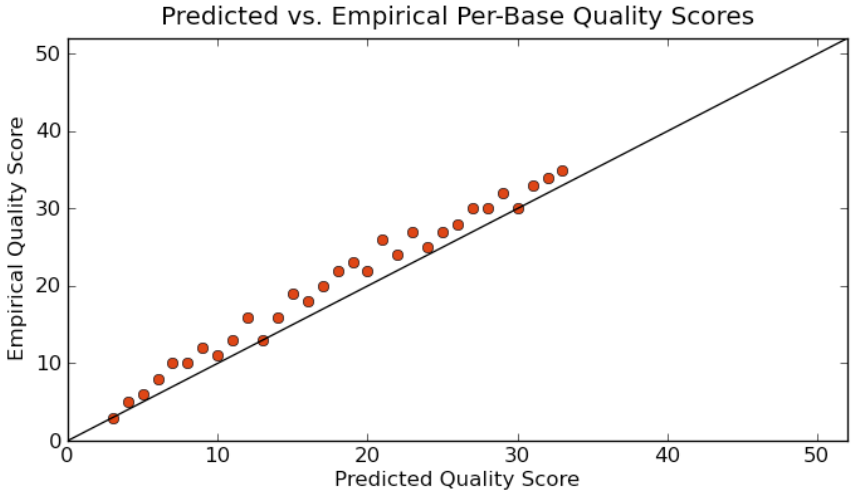


Figure S13 Quality scores

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 then the circle used to plot the score.

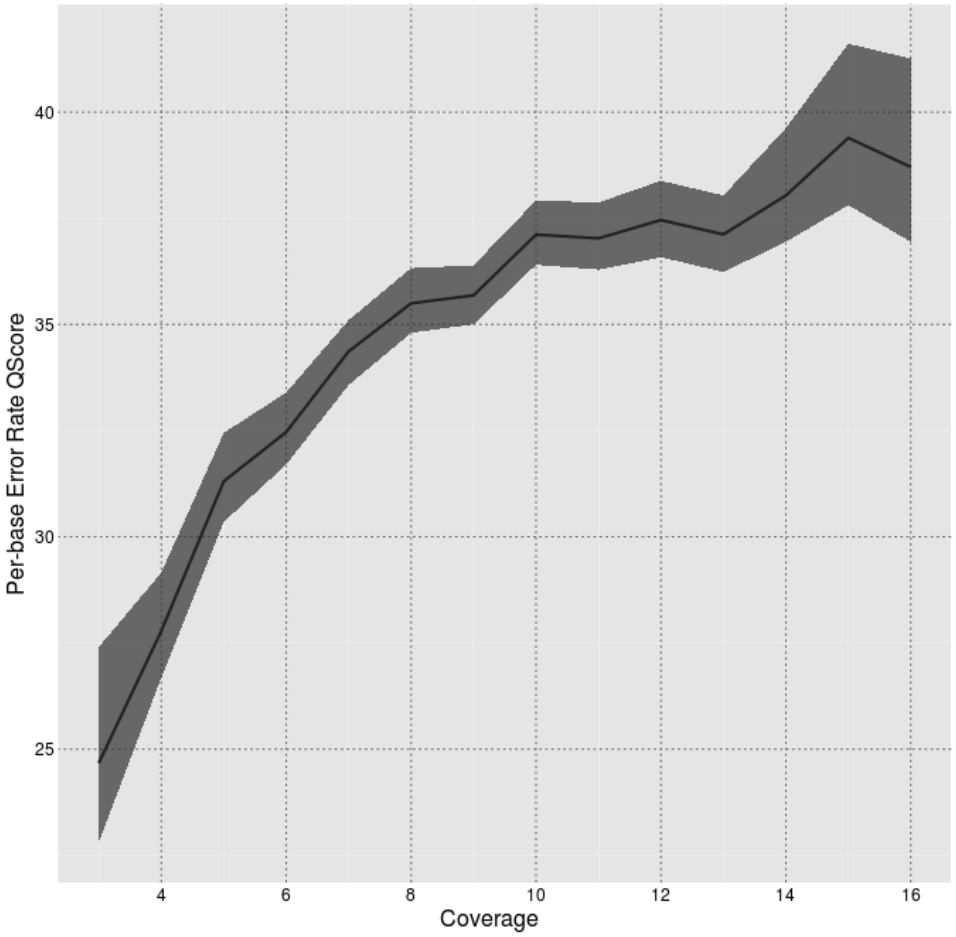


Figure S14 Consensus accuracy vs. coverage depth

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.

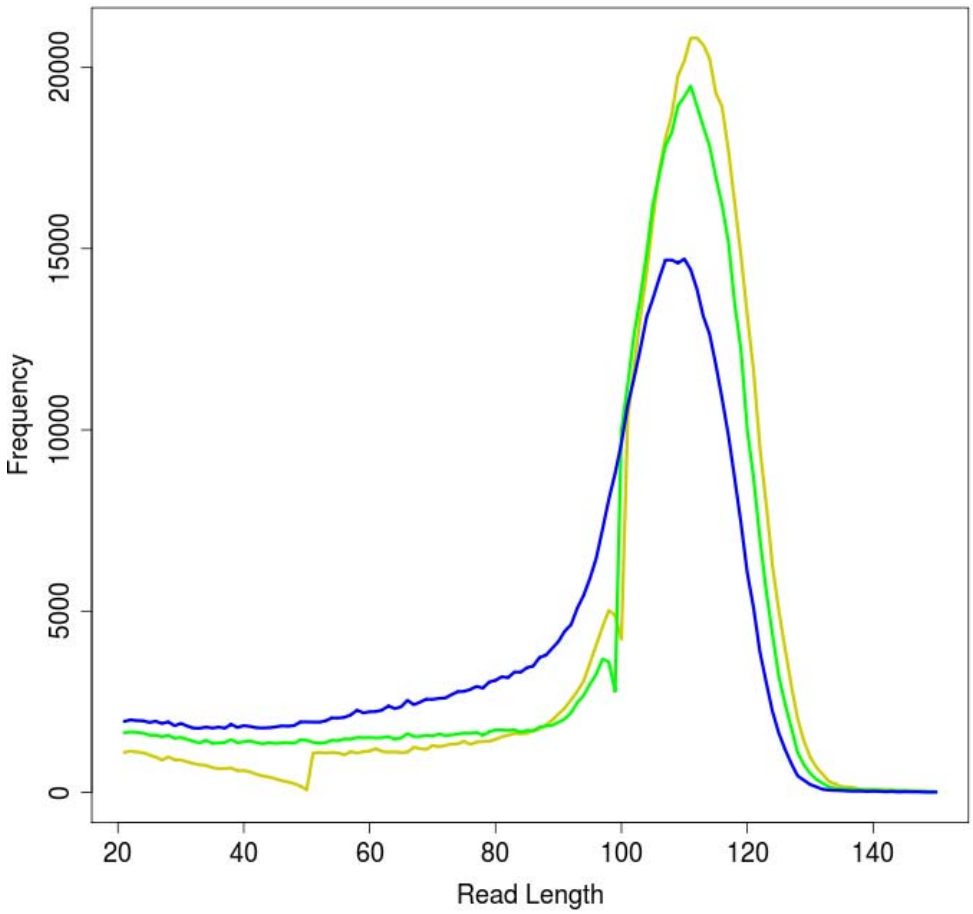


Figure S15 Distribution of read length

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 then 21 bases are not considered.

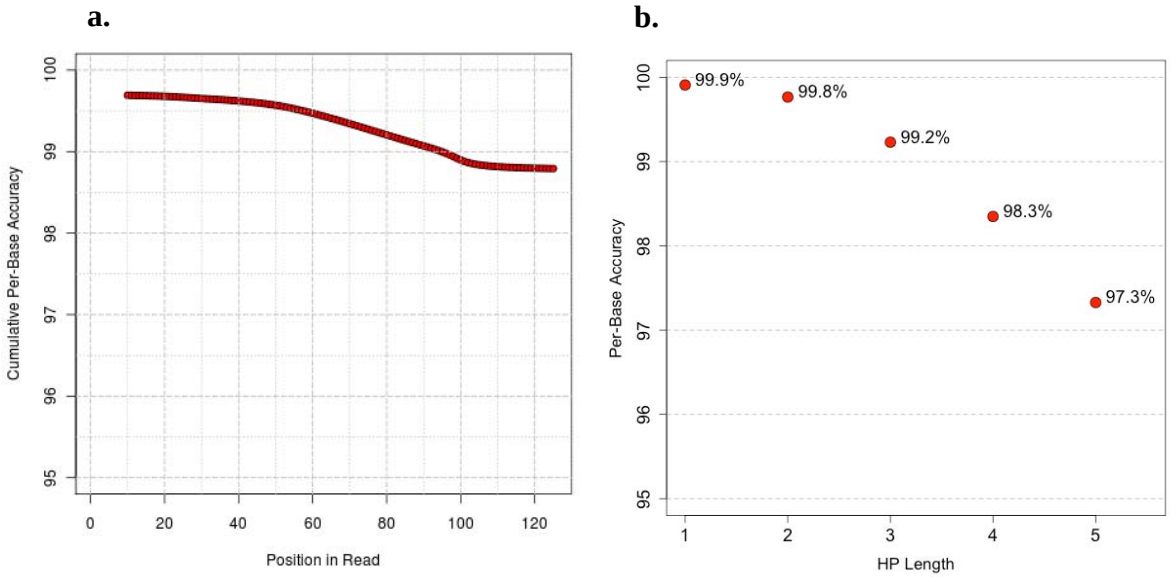


Figure S16 Single read accuracy

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% ±0.001% and taking the first 100 bases it is 98.897% ±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% ±0.023%.

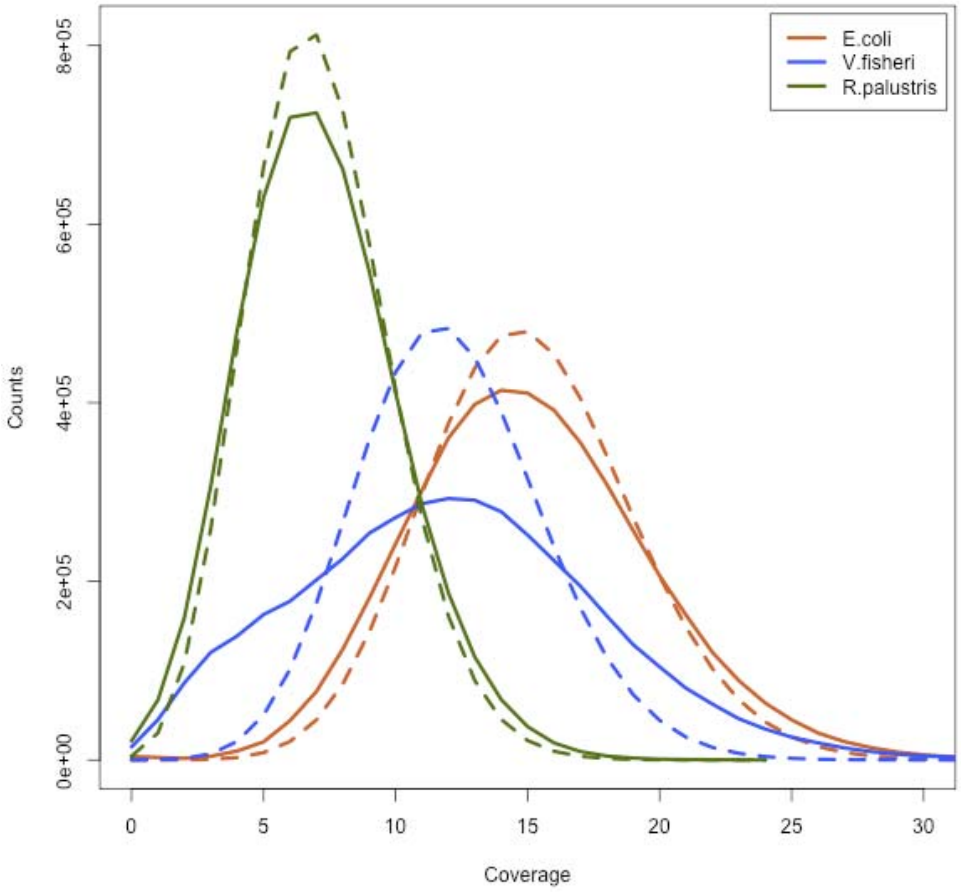


Figure S17 Uniform coverage in bacterial genomes

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 replication [48].

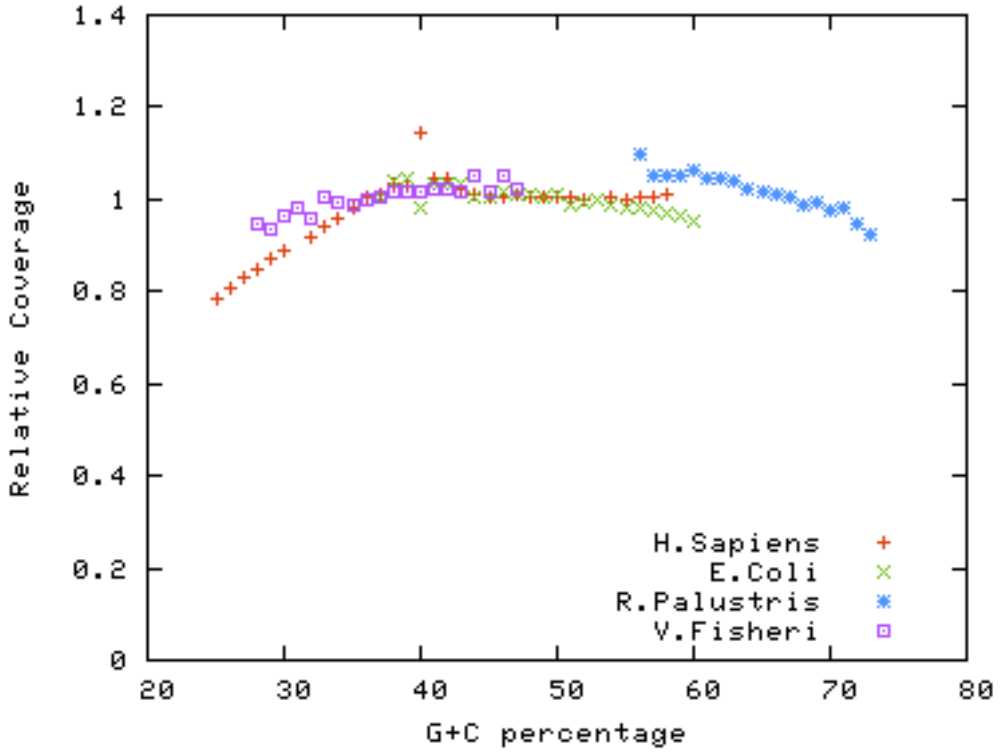


Figure S18 Uniform coverage across range of GC content

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.

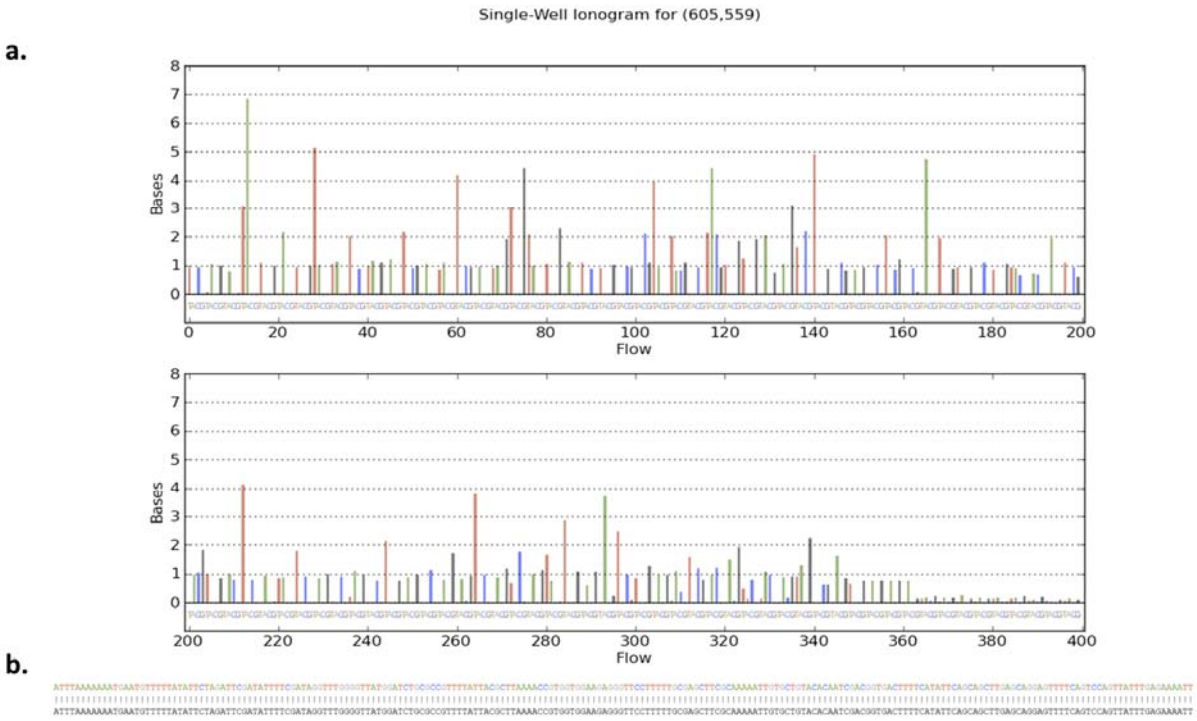
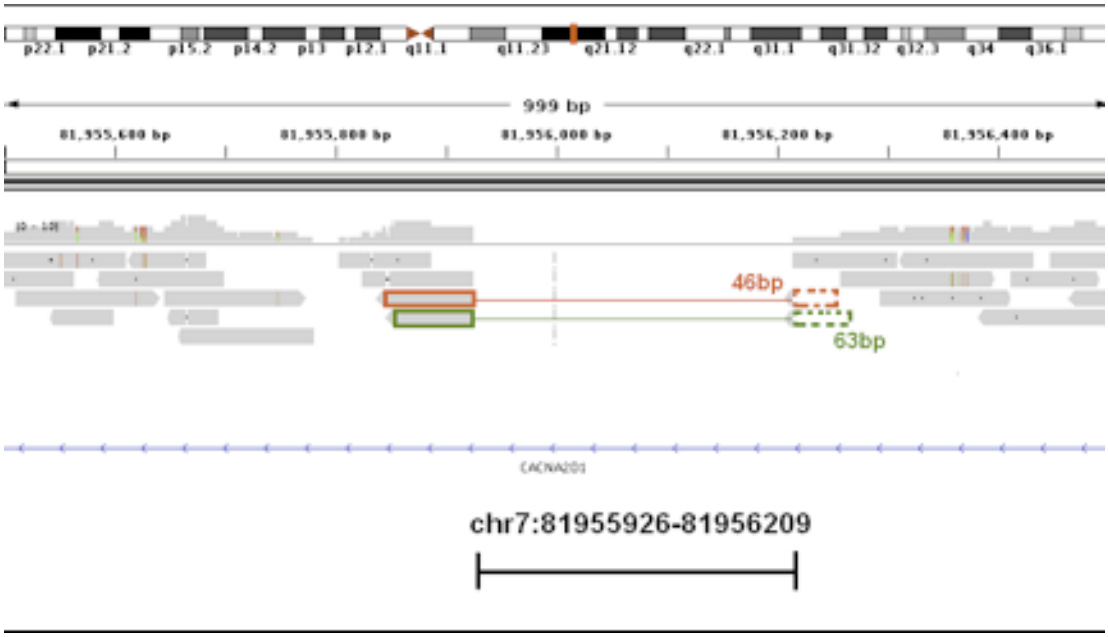


Figure S19 212 base perfect read

a, Single *E. coli* read with 400 flows (100 cycles). Each bar in the plot indicates the corresponding number of bases incorporated during that nucleotide flow. Bars are color-coded according to nucleotide; Red: 'T', Green: 'A', Blue: 'C', Grey: 'G'. The bars show the output signals after normalization to a key sequence, and correction for phase and signal loss. (Uncertainty in the base call is represented by a quality score and is not captured in this representation, see Supplementary Fig. S13). **b**, The alignment between the read (212 bases shown in color) and the genomic reference (grey).

a.



b.

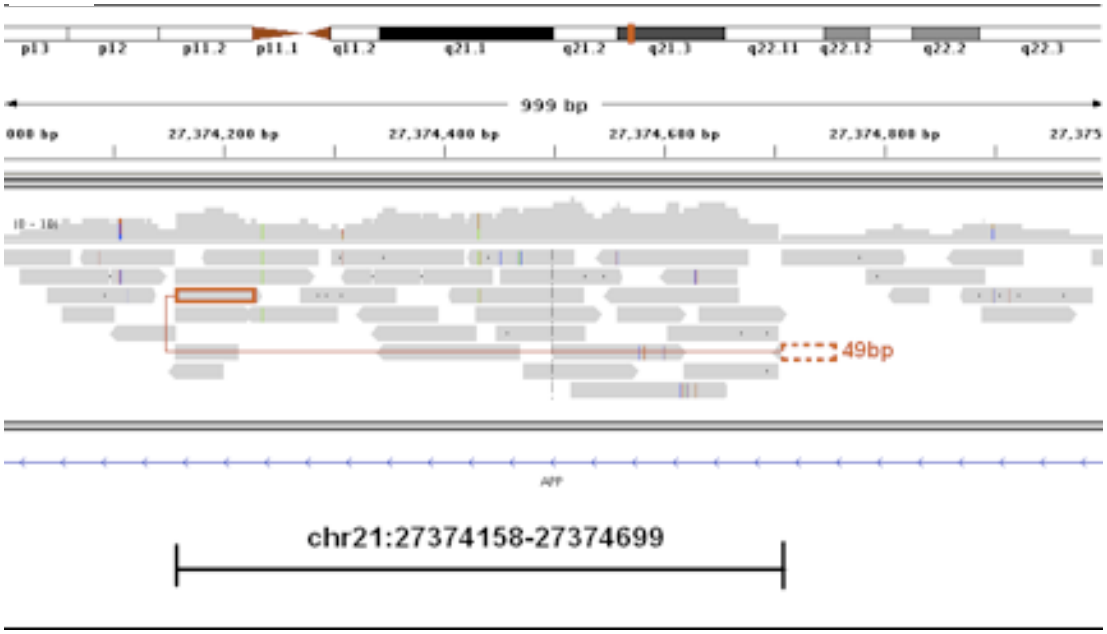


Figure S20 Deletions & inversion examples.

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 dbSNP132[50] 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), *Rhodopseudomonas 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 μM 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 μm. 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 where 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 μg 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 Informatics[51] on the Amazon EC2 cloud with BFAST. Specifically, BFAST was employed for color-space alignment to hg19[52], SRMA for color-space local realignment[53], and Samtools to establish the final variant calls[54].

Mapping ion sequence to reference

Sequencing reads for each genome were mapped to their corresponding genome reference: *Vibrio fisheri* (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 pileup[54] 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 data[55] and localized on hg19[56]. 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 gene[57] 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).