Published in final edited form as:
Nature. 2008 November 6; 456(7218): 53–59. doi:10.1038/nature07517.

Accurate Whole Human Genome Sequencing using Reversible Terminator Chemistry

David R. Bentley[1], Shankar Balasubramanian[2,*], Harold P. Swerdlow[1,†], Geoffrey P. Smith[1], John Milton[1,§], Clive G. Brown[1,§], Kevin P. Hall[1], Dirk J. Evers[1], Colin L. Barnes[1,2], Helen R. Bignell[1], Jonathan M. Boutell[1], Jason Bryant[1], Richard J. Carter[1], R. Keira Cheetham[1], Anthony J. Cox[1], Darren J. Ellis[1], Michael R. Flatbush[3], Niall A. Gormley[1], Sean J. Humphray[1], Leslie J. Irving[1], Mirian S. Karbelashvili[3], Scott M. Kirk[3], Heng Li[4], Xiaohai Liu[1,2], Klaus S. Maisinger[1], Lisa J. Murray[1], Bojan Obradovic[1], Tobias Ost[1], Michael L. Parkinson[1], Mark R. Pratt[3], Isabelle M. J. Rasolonjatovo[1], Mark T. Reed[3], Roberto Rigatti[1], Chiara Rodighiero[1], Mark T. Ross[1], Andrea Sabot[1], Subramanian V. Sankar[3], Aylwyn Scally[4], Gary P. Schroth[3], Mark E. Smith[1], Vincent P. Smith[1], Anastassia Spiridou[1], Peta E. Torrance[1], Svilen S. Tzonev[3], Eric H. Vermaas[3], Klaudia Walter[4], Xiaolin Wu[1], Lu Zhang[3], Mohammed D. Alam[3], Carole Anastasi[1], Ify C. Aniebo[1], David M. D. Bailey[1], Iain R. Bancarz[1], Saibal Banerjee[3], Selena G. Barbour[1], Primo A. Baybayan[3], Vincent A. Benoit[1], Kevin F. Benson[1], Claire Bevis[1], Phillip J. Black[1], Asha Boodhun[1], Joe S. Brennan[1], John A. Bridgham[3], Rob C. Brown[1], Andrew A. Brown[1], Dale H. Buermann[3], Abass A. Bundu[1], James C. Burrows[3], Nigel P. Carter[4], Nestor Castillo[3], Maria Chiara E. Catenazzi[1], Simon Chang[3], R. Neil Cooley[1], Natasha R. Crake[1], Olubunmi O. Dada[1], Konstantinos D. Diakoumakos[1], Belen Dominguez-Fernandez[1], David J. Earnshaw[1,2], Ugonna C. Egbujor[1], David W. Elmore[3], Sergey S. Etchin[3], Mark R. Ewan[3], Milan Fedurco[5], Louise J. Fraser[1], Karin V. Fuentes Fajardo[1], W. Scott Furey[2], David George[3], Kimberley J. Gietzen[6], Colin P. Goddard[1], George S. Golda[3], Philip A. Granieri[3], David E. Green[1], David L. Gustafson[3], Nancy F. Hansen[7], Kevin Harnish[1], Christian D. Haudenschild[3], Narinder I. Heyer[1], Matthew M. Hims[1], Johnny T. Ho[3], Adrian M. Horgan[1], Katya Hoschler[1], Steve Hurwitz[3], Denis V. Ivanov[3], Maria Q. Johnson[3], Terena James[1], T. A. Huw Jones[1], Gyoung-Dong Kang[1], Tzvetana H. Kerelska[3], Alan D. Kersey[1], Irina Khrebtukova[3], Alex P. Kindwall[3], Zoya Kingsbury[1], Paula I. Kokko-Gonzales[1], Anil Kumar[1], Marc A. Laurent[6], Cynthia T. Lawley[6], Sarah E. Lee[1], Xavier Lee[3], Arnold K. Liao[3], Jennifer A. Loch[1], Mitch Lok[3], Shujun Luo[3], Radhika M. Mammen[1], John W. Martin[3], Patrick G. McCauley[1], Paul McNitt[3], Parul Mehta[1], Keith W. Moon[3], Joe W. Mullens[3], Taksina Newington[1], Zemin Ning[4], Bee Ling Ng[4], Sonia M. Novo[1], Michael J. O'Neill[3], Mark A. Osborne[1,2], Andrew Osnowski[1], Omead Ostadan[3,6], Lambros L. Paraschos[3], Lea Pickering[1], Andrew C. Pike[1], Alger C. Pike[3], D. Chris Pinkard[3], Daniel P. Pliskin[3], Joe Podhasky[3], Victor J. Quijano[3], Come Raczy[1], Vicki H. Rae[1], Stephen R. Rawlings[1], Ana Chiva Rodriguez[1], Phyllida M. Roe[1], John Rogers[1], Maria C. Rogert Bacigalupo[1], Nikolai Romanov[1], Anthony Romieu[5], Rithy K. Roth[3], Natalie J. Rourke[1], Silke T. Ruediger[1], Eli Rusman[3], Raquel M. Sanches-Kuiper[1], Martin R. Schenker[1], Josefina M. Seoane[3], Richard J. Shaw[1], Mitch K. Shiver[3], Steven W. Short[3], Ning L. Sizto[3], Johannes P. Sluis[3], Melanie A. Smith[1], Jean Ernest Sohna Sohna[1], Eric J. Spence[3], Kim

Correspondence and requests for materials should be addressed to David.R. Bentley (dbentley@illumina.com)..

[*] Inventors and founders of Solexa Ltd

[†] Present addresses

The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK

[§] Oxford Nanopore Technologies, Begbroke Science Park, Sandy Lane, Kidlington. OX5 1PF

[‡] Pronota, NV, VIB Bio-Incubator, Technologiepark 4, B-9052 Zwijnaarde / Ghent, Belgium

All authors at Illumina (see affiliations) are employees of Illumina Inc., a public company that develops and markets systems for genetic analysis.

Supplementary Information accompanies the paper on www.nature.com/nature.'

Stevens[1], **Neil Sutton**[1], **Lukasz Szajkowski**[1], **Carolyn L. Tregidgo**[1], **Gerardo Turcatti**[5], **Stephanie vandeVondele**[1], **Yuli Verhovsky**[3], **Selene M. Virk**[3], **Suzanne Wakelin**[3], **Gregory C. Walcott**[3], **Jingwen Wang**[1], **Graham J. Worsley**[1], **Juying Yan**[3], **Ling Yau**[3], **Mike Zuerlein**[3], **Jane Rogers**[4], **James C. Mullikin**[7], **Matthew E. Hurles**[4], **Nick J. McCooke**[1,‡], **John S. West**[3], **Frank L. Oaks**[3], **Peter L. Lundberg**[3], **David Klenerman**[2,*], **Richard Durbin**[4], and **Anthony J. Smith**[1]

[1] Illumina Cambridge Ltd., (Formerly Solexa Ltd) Chesterford Research Park, Little Chesterford, Nr Saffron Walden, Essex. CB10 1XL, UK [2] Department of Chemistry, University of Cambridge, The University Chemical Laboratory, Lensfield Road, Cambridge, CB2 1EW, UK [3] Illumina Hayward, (Formerly Solexa Inc) 23851 Industrial Bvld, Hayward, CA 94343, USA [4] The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK [5] Manteia Predictive Medicine S.A. Zone Industrielle, Coinsins, CH-1267, Switzerland [6] Illumina Inc. ,Corporate Headquarters, 9883 Towne Centre Drive, San Diego, CA 92121, USA [7] National Human Genome Research Institute, National Institutes of Health, 41 Center Drive, MSC 2132, 9000 Rockville Pike, Bethesda, MD 20892-2132, USA

DNA sequence information underpins genetic research, enabling discoveries of important biological or medical benefit. Sequencing projects have traditionally employed long (400-800 bp) reads, but the existence of reference sequences for the human and many other genomes makes it possible to develop new, fast approaches to re-sequencing, whereby shorter reads are compared to a reference to identify intraspecies genetic variation. We report an approach that generates several billion bases of accurate nucleotide sequence per experiment at low cost. Single molecules of DNA are attached to a flat surface, amplified *in situ* and used as templates for synthetic sequencing with fluorescent reversible terminator deoxyribonucleotides. Images of the surface are analysed to generate high quality sequence. We demonstrate application of this approach to human genome sequencing on flow-sorted X chromosomes and then scale the approach to determine the genome sequence of a male Yoruba from Ibadan, Nigeria. We build an accurate consensus sequence from >30x average depth of paired 35-base reads. We characterise four million SNPs and four hundred thousand structural variants, many of which are previously unknown. Our approach is effective for accurate, rapid and economical whole genome re-sequencing and many other biomedical applications.

DNA sequencing yields an unrivalled resource of genetic information. We can characterise individual genomes, transcriptional states and genetic variation in populations and disease. Until recently, the scope of sequencing projects was limited by the cost and throughput of Sanger sequencing. The raw data for the 3 billion base (3 gigabase, Gb) human genome sequence, completed in 20041, was generated over several years for ~$300 million using several hundred capillary sequencers. More recently an individual human genome sequence has been determined for ~$10 million by capillary sequencing2. Several new approaches at varying stages of development aim to increase sequencing throughput and reduce cost3-6. They increase parallelisation dramatically by imaging many DNA molecules simultaneously. One instrument run produces typically thousands or millions of sequences that are shorter than capillary reads. Another human genome sequence was recently determined using one of these approaches7. However, much bigger improvements are necessary to enable routine whole human genome sequencing in genetic research.

We describe a massively parallel synthetic sequencing approach that transforms our ability to use DNA and RNA sequence information in biological systems. We demonstrate utility by re-sequencing an individual human genome to high accuracy. Our approach delivers data at very high throughput and low cost, and enables extraction of genetic information of high biological value, including single nucleotide polymorphisms (SNPs) and structural variants.

DNA sequencing using reversible terminators and clonal single molecule arrays

We generated high density single molecule arrays of genomic DNA fragments attached to the surface of the reaction chamber (the flowcell) and used isothermal ‘bridging’ amplification to form DNA ‘clusters’ from each fragment. We made the DNA in each cluster single-stranded and added a universal primer for sequencing. For paired read sequencing, we then converted the templates to double-stranded DNA and removed the original strands, leaving the complementary strand as template for the second sequencing reaction (fig 1a-c). To obtain paired reads separated by larger distances, we circularised DNA fragments of the required length (e.g. 2kb +/− 0.2kb) and obtained short junction fragments for paired end sequencing (fig 1d).

We sequenced DNA templates by repeated cycles of polymerase-directed single base extension. To ensure base-by-base nucleotide incorporation in a stepwise manner, we used a set of four reversible terminators, 3′-O-azidomethyl 2′-deoxynucleoside triphosphates (A, C, G and T) each labelled with a different removable fluorophore (fig S1a)8. The use of 3′-modified nucleotides allowed the incorporation to be driven essentially to completion without risk of over-incorporation. It also enabled addition of all four nucleotides simultaneously rather than sequentially, minimising risk of mis-incorporation. We engineered the active site of 9°N DNA polymerase to improve the efficiency of incorporation of these unnatural nucleotides9. After each cycle of incorporation, we determined the identity of the inserted base by laser-induced excitation of the fluorophores and imaging. We added tris(2-carboxyethyl)phosphine (TCEP) to remove the fluorescent dye and side-arm from a linker attached to the base and simultaneously to regenerate a 3′ hydroxyl group ready for the next cycle of nucleotide addition (fig S1b). The Genome Analyzer (GA1) was designed to perform multiple cycles of sequencing chemistry and imaging to collect the sequence data automatically from each cluster on the surface of each lane of an 8-lane flowcell (fig S2).

To determine the sequence from each cluster, we quantified the fluorescent signal from each cycle and applied a base-calling algorithm. We defined a quality (Q) value for each base call (scaled as by the phred algorithm10) that represents the likelihood of each call being correct (fig S3). We used the Q-values in subsequent analyses to weight the contribution of each base to sequence alignment and detection of sequence variants (e.g. SNP calling). We discarded all reads from mixed clusters and used the remaining ‘purity filtered’ (PF) reads for analysis. Typically we generated 1-2 billion bases (gigabases, Gb) of high quality PF sequence per flow cell from ~60 million single 35-base reads, or 2-4 Gb in a paired read experiment (table S1).

To demonstrate accurate sequencing of human DNA, we sequenced a human bacterial artificial chromosome (BAC) clone (bCX98J21) that contained 162,752 bp of the major histocompatibility complex on human chromosome 6 (accession AL662825.4, previously determined using capillary sequencing by the Wellcome Trust Sanger Institute). We developed a fast global alignment algorithm ELAND that aligns a read to the reference only if the read can be assigned a unique position with 0, 1 or 2 differences. We collected 0.17 Gb of aligned data for the BAC from one lane of a flowcell. Approximately 90% of the 35-base reads matched perfectly to the reference, demonstrating high raw read accuracy (fig S4). To examine consensus coverage and accuracy, we used 5 Mb of 35-base PF reads (30-fold average input depth of the BAC) and obtained 99.96% coverage of the reference. There was one consensus miscall, at a position of very low coverage (just above our cut-off threshold), yielding an overall consensus accuracy of >99.999%.

Detecting genetic variation of the human X chromosome

For an initial study of genetic variation, we sequenced flow-sorted X chromosomes of a Caucasian female (CEPH NA07340). We generated 278 million paired 30-35 bp PF reads and aligned them to the human genome reference sequence. We carried out separate analyses of the data using two alignment algorithms, ELAND (see above) or MAQ11. Both algorithms place each read pair where it best matches the reference and assign a confidence score to the alignment. In cases where a read has two or more equally likely positions (i.e. in an exact repeat), MAQ randomly assigns the read pair to one position and assigns a zero alignment quality score (these reads are excluded from SNP analysis). ELAND rejects all non-unique alignments, which are mostly in recently inserted retroposons (see fig S5). MAQ therefore provides an opportunity to assess the properties of a dataset aligned to the entire reference, whereas ELAND effectively excludes ambiguities from the short read alignment before further analysis.

We obtained comprehensive coverage of the X chromosome from both analyses. With MAQ, 204 million reads aligned to 99.94% of the X chromosome at an average depth of 43x. With ELAND, 192 million reads covered 91% of the reference sequence, showing what can be covered by unique best alignments. These results were obtained after excluding reads aligning to non-X sequence (impurities of flow sorting) and apparently duplicated read pairs (table S2). We reasoned that these duplicates (~10% of the total) arose during initial sample amplification.

The sampling of sequence fragments from the X chromosome is close to random. This is evident from the distribution of mapped read depth in the MAQ alignment in regions where the reference is unique (fig 2a): the variance of this distribution is only 2.26 times that of a Poisson distribution (the theoretical minimum). Half of this excess variance can be accounted for by a dependence on GC content. However, the average mapped read depth only falls below 10x in regions with GC content less than 4% or greater than 76%, comprising in total just 1% of unique chromosome sequence and 3% of coding sequence (fig 2b).

We identified 92,485 candidate SNPs in the X chromosome using ELAND (fig S6). Most calls (85%) match previous entries in the public database dbSNP. Heterozygosity (π) in this dataset is 4.3×10^{-4} (i.e. 1 substitution per 2.3 kb), close to a previously published X chromosome estimate (4.7×10^{-4})12. Using MAQ we obtained 104,567 SNPs, most of which were common to the results of the ELAND analysis. The differences between the two sets of SNP calls are largely the consequence of different properties of the alignments as described earlier. For example, most of the SNPs found only by the MAQ-based analysis were at positions of low or zero sequence depth in the ELAND alignment (fig S6c).

We assessed accuracy and completeness of SNP calling by comparison to genotypes obtained for this individual using the Illumina HumanHap550 BeadChip (HM550). The sequence data covered >99.8% of the 13,604 genotyped positions and we found excellent agreement between sequence based SNP calls and genotyping data (99.52% or 99.99% using ELAND or MAQ, respectively)(table S3). There was complete concordance of all homozygous calls and a low level of ‘undercalling’ (denoted as ‘GT>Seq’ in table 1) at a small number of the heterozygous sites, caused by inadequate sampling of one of the two alleles. The depth of input sequence influences the coverage and accuracy of SNP calling. We found that reducing the read depth to 15x still gives 97% coverage of genotype positions and only 1.27% of the heterozygous sites are undercalled. We observed no other types of disagreement at any input depth (fig S7).

We detected structural variants (defined as any variant other than a single base substitution) as follows. We found 9,747 short insertions/deletions ('short indels', defined here as less than the length of the read) by performing a gapped alignment of individual reads (fig S8). We identified larger indels based on read depth and/or anomalous read pair spacing, similar to previous approaches13-15. We detected 115 indels in total, 77 of which were visible from anomalous read pair spacing (see tables S4 and S5). We developed Resembl, an extension to the Ensembl browser16, to view all variants (fig S9; see also fig 4). Inversions can be detected when the orientation of one read in a pair is reversed (e.g. fig S10). In general, inversions occur as the result of non-allelic homologous recombination, and are therefore flanked by repetitive sequence that can compromise alignments. We found partial evidence for other inversion events, but characterisation of inversions from short read data is complex because of the repeats and requires further development.

Sequencing and analysis of a whole human genome

Our X chromosome study enabled us to develop an integrated set of methods for rapid sequencing and analysis of whole human genomes. We sequenced the genome of a male Yoruba from Ibadan, Nigeria (YRI; sample NA18507). This sample was originally collected for the HapMap project17,18 through a process of community engagement and informed consent19 and has also been studied in other projects20,21. We were therefore able to compare our results with publicly available data from the same sample. We constructed two libraries: one of short inserts (~200 bp) with similar properties to the previous X chromosome library and one with long inserts (~2 kb) to provide longer-range read pair information (see fig S11 for size distributions). We generated 135 Gb of sequence (~4 billion paired 35-base reads; see table S6) over a period of 8 weeks (Dec'07–Jan'08) on six GA1 instruments averaging 3.3Gb per production run (see table S1 for example). The approximate consumables cost (based on full list price of reagents) was $250,000. We aligned 97% of the reads using MAQ and found 99.9% of the human reference (NCBI build 36.1) was covered with one or more reads at an average of 40.6-fold depth. Using ELAND, we aligned 91% of the reads over 93% of the reference sequence at sufficient depth to call a strong consensus (>three Q30 bases). The distribution of mapped read depth was close to random, with slight overdispersion as seen for the X chromosome data. We observed comprehensive representation across a wide range of GC content, dropping only at the very extreme ends, but with a different pattern of distribution compared to the X (see fig S12).

We identified ~4 million SNPs, with 74% matching previous entries in dbSNP (fig 3). We found excellent agreement of our SNP calls with genotyping results: sequence-based SNP calls covered almost all of the 552,710 loci of HM550, with >99.5% concordance of sequencing vs. genotyping calls (tables 1 and S7a). The few disagreements were mostly undercalls of heterozygous positions (GT>Seq) in areas of low sequence depth, providing us with a false negative rate of <0.35% from the ELAND analysis - see table 1). The other disagreements (0.09% of all genotypes) included errors in genotyping plus apparent tri-allelic SNPs (table S7). The main cause of genotype error (0.05% of all genotypes) is the existence of a second 'hidden' SNP close to the assayed locus that disrupts the genotyping assay, leading to loss of one allele and an erroneous homozygous genotype (figs S13, S14).

To examine the accuracy of SNP calling in more detail, we compared our sequence-based SNP calls with 3.7M genotypes (HM-All) generated for this sample during the HapMap project (table 1 and S7b)20 and found excellent concordance. Disagreements included sequence-based undercalls of heterozygous positions in regions of low read depth. The slightly higher level of other disagreements (0.76%) seen in this analysis compared to that of the HM550 data (0.09%) is in line with the higher level of underlying genotype error rate of 0.7% for the HapMap data20. To refine this analysis further, we generated a set of 530,750 very high confidence reference genotypes comprising concordant calls in both the HM550 and HM-All genotype datasets. Comparing the results of the MAQ analysis to this high confidence set (see table 1), we found 130 heterozygote undercalls (i.e. a false negative rate of 0.025%). There were also 130 heterozygote overcalls, but most of these are likely genotype errors as 82 have a nearby ‘hidden’ SNP and 3 have a nearby indel. A further 41 are tri-allelic loci, leaving at most 4 potential wrong calls by sequencing (i.e. false positive rate of 4/529,589 positions). Finally we selected a subset of novel SNP calls from the sequence data and tested them by genotyping. We found 96.1% agreement between sequence and genotype calls (table S8). However, the 47 disagreements included 10 correct sequencing calls (genotyping undercalls due to hidden SNPs) and 7 sequencing undercalls. On this basis, therefore, the false positive discovery rate for the 1M novel SNPs is 2.4% (30/1206). For the entire dataset of 4M SNPs detected in this analysis, the false positive and negative rates both average <1%.

This Yoruba genome contains significantly more polymorphism than a genome of European descent. The autosomal heterozygosity (π) of NA18507 is 9.94×10^{-4} (1 SNP per 1006 bp), higher than previous values for Caucasians (7.6×10^{-4}, ref [12]). Heterozygosity in the pseudoautosomal region 1 (PAR1) was substantially higher (1.92×10^{-3}) than the autosomal value. PAR1 (2.7Mb) at the tip of the short arm of X and Y undergoes obligatory recombination in male meiosis, which is equivalent to 20x the autosome average. This illustrates clearly the correlation between recombination and nucleotide diversity. By contrast, the 0.33 Mb PAR2 region has a much lower recombination rate than PAR1; we observed that heterozygosity in PAR2 is identical to that of the autosomes in NA18507. Heterozygosity in coding regions is lower (0.54×10^{-3}) than the total autosome average, consistent with the model that some coding changes are deleterious and are lost as the result of natural selection22. Nevertheless, the 26,140 coding SNPs (fig S15) include 5,361 non-conservative amino acid substitutions plus 153 premature termination codons (table S9), many of which are expected to affect protein function.

We performed a genome-wide survey of structural variation in this individual and found excellent correlation with variants that had been reported in previous studies, as well as detecting many new variants. We found 0.4 million short indels (1-16 bp) (fig S16), most of which are length polymorphisms in homopolymeric tracts of A or T. Half of these events are corroborated by entries in dbSNP, and 95 of 100 examined were present in amplicons sequenced from this individual in ENCODE regions, confirming the high specificity of this method of short indel detection. For larger structural variants (detected by anomalously spaced paired ends) we found that some were detected by both long and short insert datasets (fig S17a), but the majority were unique to one or other dataset. We observed two reasons for this: first, small events (<400 bp) are within the normal size variance of the long insert data; second, nearby repetitive structures can prevent unique alignment of read pairs (see fig S17b, c). In some cases, the high resolution of the short insert data permits detection of additional complexity in a structural rearrangement that is not revealed by the long insert data. For example where the long insert data indicates a 1.3kb deletion in NA18507 relative to the reference, the short insert data reveal an inversion accompanied by deletions at both breakpoints (fig 4). We carried out *de novo* assembly of reads in this region and constructed a single contig that defines the exact structure of the rearrangement (data not shown).

We discovered 5,704 structural variants ranging from 50 bp to >35 kb where there is sequence absent from NA18507 compared to the reference. We observed a steadily decreasing number of events of this type with increasing size, except for two peaks (fig

S18). Most of the events represented by the large peak at 300-350 bp contain a sequence of the AluY family. This is consistent with insertion of SINEs that are present in the reference but missing from the genome of NA18507. Similarly, the second, smaller peak at 6-7kb is the consequence of insertion of L1Hs elements in many cases. We found good correspondence between our results and the data of Kidd *et al.*23, who reported 148 deletions of <100 kb in this individual on the basis of abnormal fosmid paired end spacing. We found supporting evidence for 111 of these events. We detected a further 2,345 indels in the range 60-160 bp which are sequences present in NA18507 and absent from the reference (fig S19). One example is shown in fig S20. The 'singleton' reads on either side of the event, which have partners that do not align to the reference, form part of a *de novo* assembly that precisely defines the novel sequence and breakpoint (fig S21).

Effect of sequence depth on coverage and accuracy

We investigated the impact of varying input read depth (and hence cost) on SNP calling using chromosome 2 as a model. SNP discovery increases with increasing depth: essentially all homozygous positions are detected at 15x, whereas heterozygous positions accumulate more gradually to 33x (fig 5a). This effect is influenced by the stringency of the SNP caller. To call each allele in this analysis we required the equivalent of two high quality Q30 bases (as opposed to three used in full depth analyses). Homozygotes could be detected at read depth of 2x or higher, whereas heterozygote detection required at least double this depth for sampling of both alleles. Missing calls (not covered by sequence) and discordances between sequence based SNP calls and genotype loci (mostly undercalls of heterozygotes due to low depth) progressively reduced with increasing depth (fig 5b). We observed very few other types of discordance at any depth; and many of these are genotyping errors as described above.

Concluding remarks

Reversible terminator chemistry is a defining feature of this sequencing approach, enabling each cycle to be driven to completion while minimising mis-incorporation. The result is a system that generates accurate data at very high throughput and low cost. We determined an accurate whole human genome sequence in eight weeks to an average depth of ~40x. We built a consensus sequence, optimised methods for analysis, assessed accuracy and characterised the genetic variation of this individual in detail.

We assessed accuracy relative to genotype data over the entire fraction of the human sequence where SNP calling was possible (>90%). We established very low false positive and negative rates for the ~4M SNPs detected (<1% overcalls and undercalls). This compares favourably with previous individual genome analyses which reported a 24% undercalling of heterozygous positions2,7.

Paired reads were very powerful in all areas of the analysis. They provided very accurate read alignment and thus improved the accuracy and coverage of consensus sequence and SNP calling. They were essential for developing our short indel caller, and for detecting structural variants. Our short insert paired read dataset introduced a new level of resolution in structural variation detection, revealing thousands of variants in a size range not characterised previously. In some cases we determined the exact sequence of structural variants by *de novo* assembly from the same paired read dataset. Interpreting events that are embedded in repetitive sequence tracts will require further work.

Massively parallel sequencing technology makes it feasible to consider whole human genome sequencing as a clinical tool in the near future. Characterising multiple individual genomes will enable us to unravel the complexities of human variation in cancer and other diseases and will pave the way for the use of personal genome sequences in medicine and healthcare. Accuracy of personal genetic information from sequence will be critical for life-changing decisions.

In addition to the large scale genomic projects exemplified by the present study and others15,24-26, the system described here is being used to explore biological phenomena in unprecedented detail, including transcriptional activity, mechanisms of gene regulation and epigenetic modification of DNA and chromatin27-32. In the future, DNA sequencing will be the central tool for unravelling how genetic information is used in living processes.

Methods Summary

DNA and sequencing

DNA samples (NA07340 and NA18507) and cell line (GM07340) were obtained from Coriell Repositories, Camden NJ. DNA samples were genotyped on the HM550 array and the results compared to publicly available data to confirm their identity before use. Methods for DNA manipulation, including sample preparation, formation of single molecule arrays, cluster growth and sequencing were all developed during this study and formed the basis for the standard protocols now available from Illumina, Inc. All sequencing was performed on Illumina GA1s equipped with a one-megapixel camera. All PF read data are available for download from the Short Read Archive at NCBI.

Analysis software

Image analysis software and the ELAND aligner are provided as part of the Genome Analyzer analysis software. SNP and structural variant detectors will be available as future upgrades of the analysis pipeline. The Resembl extension to Ensembl is available on request. The MAQ (Mapping and Assembly with Qualities) aligner is freely available for download from http://maq.sourceforge.net

Data access

Sequence data are freely available from the short read archive, accession SRA000271: ftp://ftp.ncbi.nih.gov/pub/TraceDB/ShortRead/SRA000271 Links to Resembl displays for X and human data, plus information on other available data are provided at http://www.illumina.com/iGenome

A detailed Methods section can be found as part of the Supplementary Information.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors gratefully acknowledge the valuable advice and support of Alan Williamson, Tim Rink, Steve Benkovic, John Berriman, John Todd, Robert Waterston, Sam Eletr, William Jack, Matt Cooper, Tom Brown, Colin Reece and Ron Cook during this work, to Elliott Margulies for assistance with data analysis, and the contributions of the administrative and support staff at all the institutions. This research was supported in part by The Wellcome Trust (to H.L., A.Sc., K.W., N.P.C, B.N.L., J.R., M.E.H. and R.D.), the Biotechnology and Biological Sciences Research Council (BBSRC)(to S.B. and D.K.), the BBSRC Applied Genomics LINK Programme (to A.Sp. and C.L.B.) and the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health (to N.F.H. and J.C.M.).

References

1. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature. 2004; 431:931–45. [PubMed: 15496913]
2. Levy S, et al. The diploid genome sequence of an individual human. PLoS Biol. 2007; 5:e254. [PubMed: 17803354]
3. Margulies M, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005; 437:376–80. [PubMed: 16056220]
4. Shendure J, et al. Accurate multiplex polony sequencing of an evolved bacterial genome. Science. 2005; 309:1728–32. [PubMed: 16081699]
5. Harris TD, et al. Single-molecule DNA sequencing of a viral genome. Science. 2008; 320:106–9. [PubMed: 18388294]
6. Lundquist PM, et al. Parallel confocal detection of single molecules in real time. Optics Letters. 2008; 33:1026–1028. [PubMed: 18451975]
7. Wheeler DA, et al. The complete genome of an individual by massively parallel DNA sequencing. Nature. 2008; 452:872–6. [PubMed: 18421352]
8. Milton J, et al. Modified nucleotides. PCT International publication number WO 2004/018497. 2004
9. Smith GP, et al. Modified polymerases for improved incorporation of nucleotide analogues. PCT International publication number WO 2005/024010. 2005
10. Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 1998; 8:186–94. [PubMed: 9521922]
11. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 2008; 19 [PubMed: 18714091]
12. The International SNP Map Working Group. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature. 2001; 409:928–33. [PubMed: 11237013]
13. Tuzun E, et al. Fine-scale structural variation of the human genome. Nat Genet. 2005; 37:727–32. [PubMed: 15895083]
14. Korbel JO, et al. Paired-end mapping reveals extensive structural variation in the human genome. Science. 2007; 318:420–6. [PubMed: 17901297]
15. Campbell PJ, et al. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat Genet. 2008; 40:722–9. [PubMed: 18438408]
16. Hubbard T, et al. The Ensembl genome database project. Nucleic Acids Res. 2002; 30:38–41. [PubMed: 11752248]
17. The International HapMap Consortium. A haplotype map of the human genome. Nature. 2005; 437:1299–320. [PubMed: 16255080]
18. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007; 449:851–61. [PubMed: 17943122]
19. The International HapMap Consortium. The International HapMap Project. Nature. 2003; 426:789–96. [PubMed: 14685227]
20. The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007; 447:799–816. [PubMed: 17571346]
21. Redon R, et al. Global variation in copy number in the human genome. Nature. 2006; 444:444–54. [PubMed: 17122850]
22. Cargill M, et al. Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet. 1999; 22:231–8. [PubMed: 10391209]
23. Kidd JM, et al. Mapping and sequencing of structural variation from eight human genomes. Nature. 2008; 453:56–64. [PubMed: 18451855]
24. Hillier LW, et al. Whole-genome sequencing and variant discovery in C. elegans. Nat Methods. 2008; 5:183–8. [PubMed: 18204455]
25. Hodges E, et al. Genome-wide in situ exon capture for selective resequencing. Nat Genet. 2007
26. Porreca GJ, et al. Multiplex amplification of large sets of human exons. Nat Methods. 2007; 4:931–6. [PubMed: 17934468]

27. Barski A, et al. High-resolution profiling of histone methylations in the human genome. Cell. 2007; 129:823–37. [PubMed: 17512414]

28. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. Science. 2007; 316:1497–502. [PubMed: 17540862]

29. Mikkelsen TS, et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature. 2007; 448:553–60. [PubMed: 17603471]

30. Boyle AP, et al. High-Resolution Mapping and Characterization of Open Chromatin across the Genome. Cell. 2008; 132:311–22. [PubMed: 18243105]

31. Lister R, et al. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell. 2008; 133:523–36. [PubMed: 18423832]

32. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008

33. Fedurco M, Romieu A, Williams S, Lawrence I, Turcatti G. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. Nucleic Acids Res. 2006; 34:e22. [PubMed: 16473845]

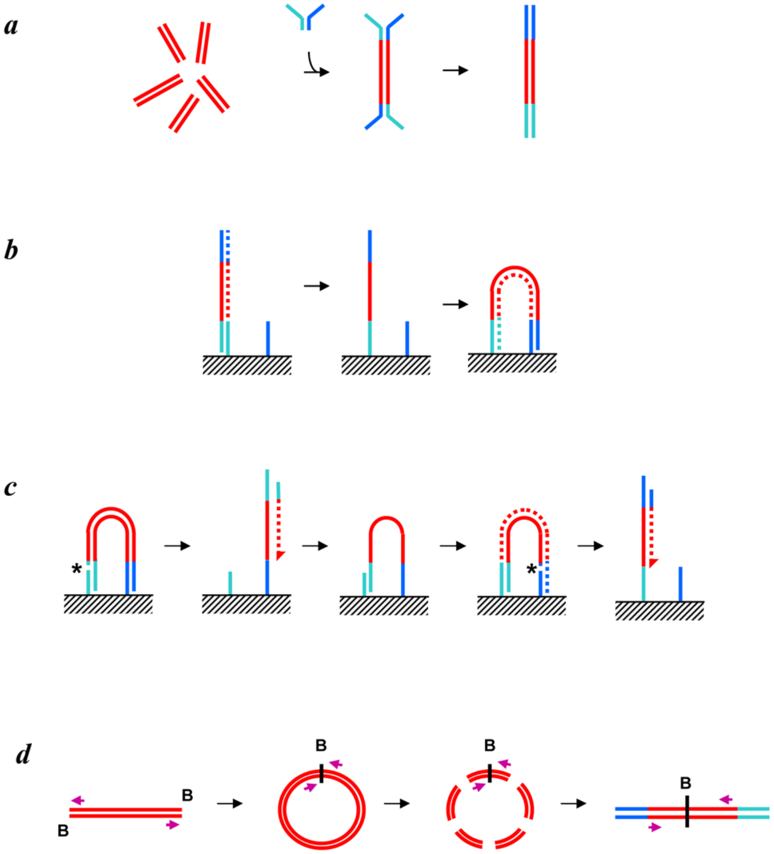
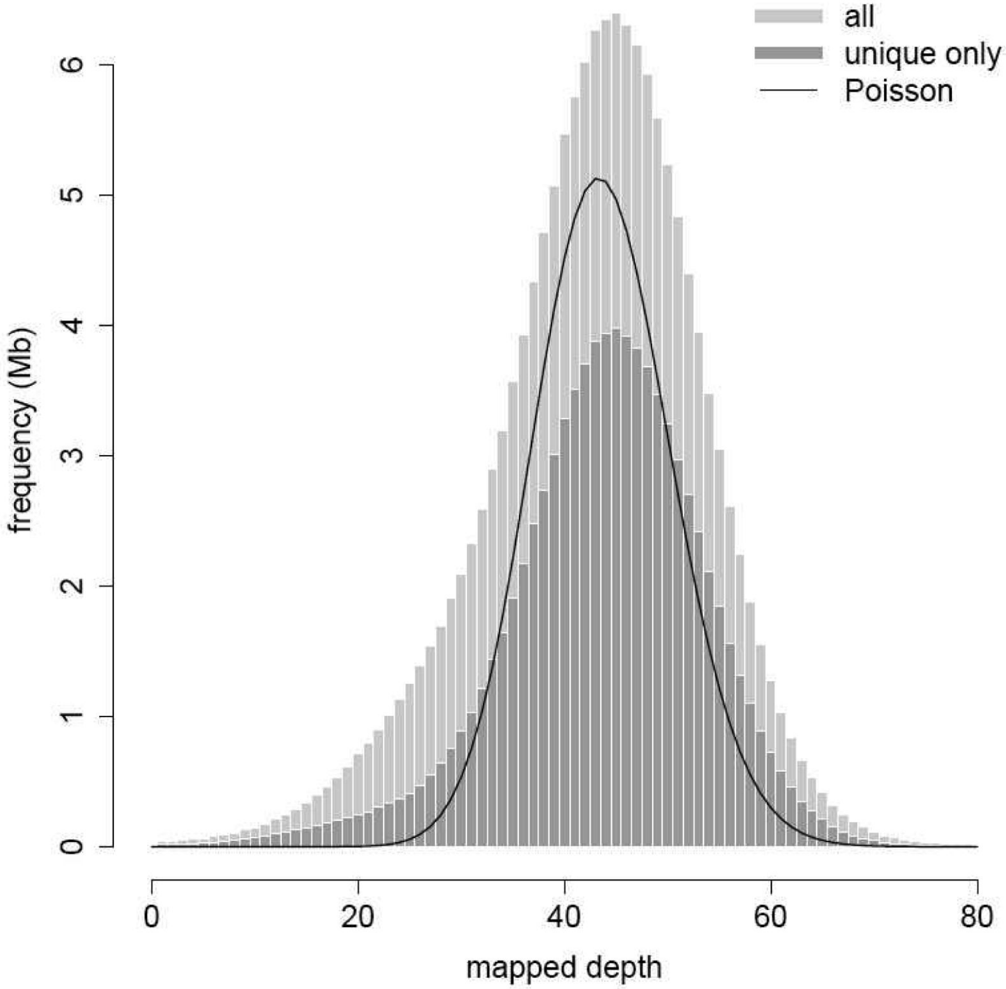


Figure 1. Sample preparation

a. DNA fragments are generated e.g. by random shearing and joined to a pair of oligonucleotides in a forked adapter configuration. The ligated products are amplified using two oligonucleotide primers, resulting in double-stranded blunt-ended material with a different adapter sequence on either end. ***b***. formation of clonal single molecule array. DNA fragments prepared as in ***a*** are denatured and single strands are annealed to complementary oligonucleotides on the flowcell surface (hatched in the figure). A new strand (dotted) is copied from the original strand in an extension reaction that is primed from the 3′ end of the surface-bound oligonucleotide, and the original strand is then removed by denaturation. The adapter sequence at the 3′ end of each copied strand is annealed to a new surface bound complementary oligonucleotide, forming a bridge and generating a new site for synthesis of a second strand (shown dotted). Multiple cycles of annealing, extension and denaturation in isothermal conditions result in growth of clusters each ~1micron in physical diameter. This follows the basic method outlined in ref [33] ***c.*** The DNA in each cluster is linearised by cleavage within one adapter sequence (gap marked by an asterisk) and denatured, generating single stranded template for sequencing by synthesis to obtain a sequence read (read 1)(the sequencing product is shown dotted). To perform paired-read sequencing, the products of read 1 are removed by denaturation, the template is used to generate a bridge, the second strand is re-synthesised (shown dotted), and the opposite strand is then cleaved (gap marked by an asterisk) to provide the template for the second read (read 2). ***d***. Long range paired end sample preparation. To sequence the ends of a long (e.g. >1 kb) DNA fragment, the ends of each fragment are tagged by incorporation of biotinylated (B) nucleotide and then circularised, forming a junction between the two ends. Circularised DNA is randomly fragmented and the biotinylated junction fragments are recovered and used as starting material in the standard sample preparation procedure illustrated in ***a*** above. The orientation of the sequence reads relative to the DNA fragment is tracked in the figure by magenta arrows. When aligned to the reference sequence, these reads are oriented with their 5′ ends towards each other (in contrast to the short insert paired reads produced as shown in ***a-c***). See fig S17a for examples of both. Turquoise and blue lines represent oligonucleotides and red lines represent genomic DNA. Note that all surface-bound oligonucleotides are attached to the flowcell by their 5′ ends. Dotted lines indicate newly synthesized strands during cluster formation or sequencing. See supplementary methods for details.


all
unique only
Poisson
frequency (Mb)
0
1
2
3
4
5
6
0
20
40
60
80
mapped depth

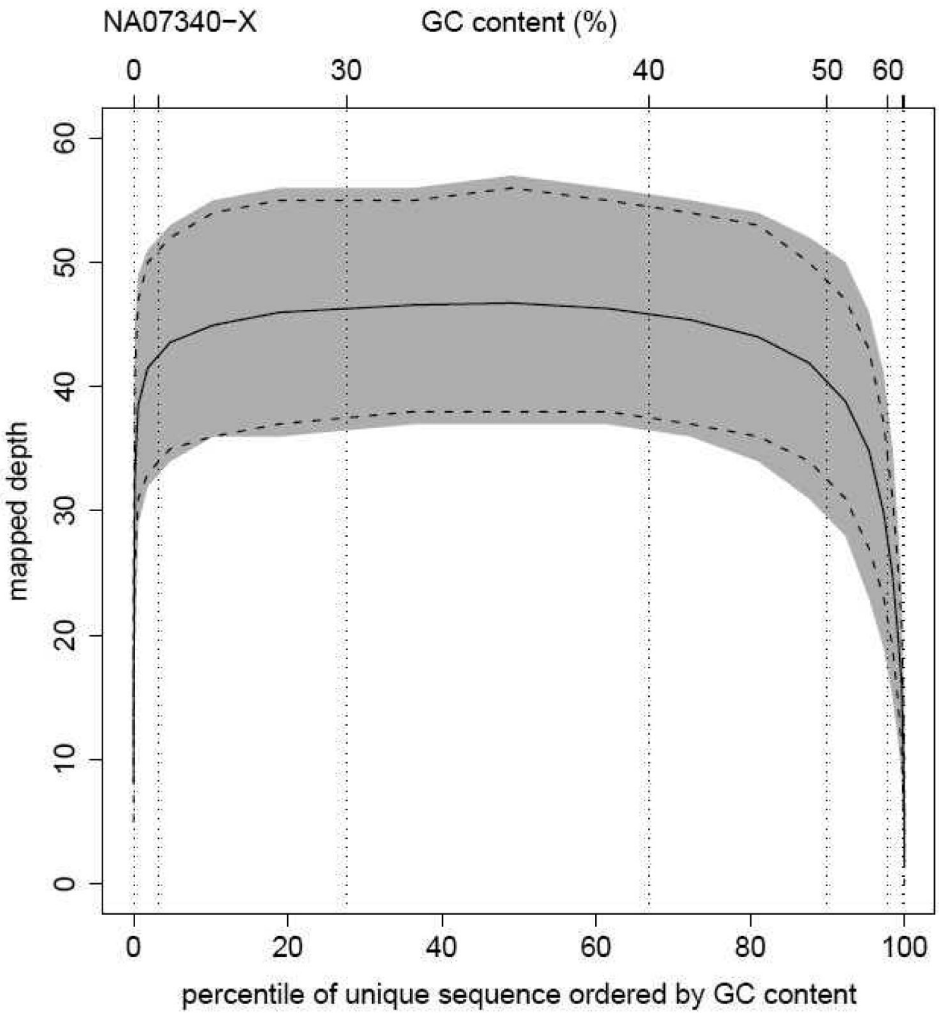


Figure 2. X chromosome data

a. Distribution of mapped read depth in the X chromosome dataset, sampled at every 50th position along the chromosome and displayed as a histogram ('all'). An equivalent analysis of mapped read depth for the unique subset of these positions is also shown ('unique only'). The solid line represents a Poisson distribution with the same mean. ***b***. Distribution of X chromosome uniquely mapped reads as a function of GC content. Note that the x axis is % GC content and is scaled by percentile of unique sequence. The solid line is average mapped depth of unique sequence; the grey region is the central 80% of the data (10th to 90th centiles); the dashed lines are 10th and 90th centiles of a Poisson distribution with the same mean as the data.

a

Call	ELAND		MAQ	
	SNPs	In dbSNP	SNPs	In dbSNP
	n	%	n	%
Homozygote	1,417,320	90.1%	1,503,420	90.8%
Heterozygote	2,411,022	63.9%	2,635,776	63.8%
All	3,828,342	73.6%	4,139,196	73.6%

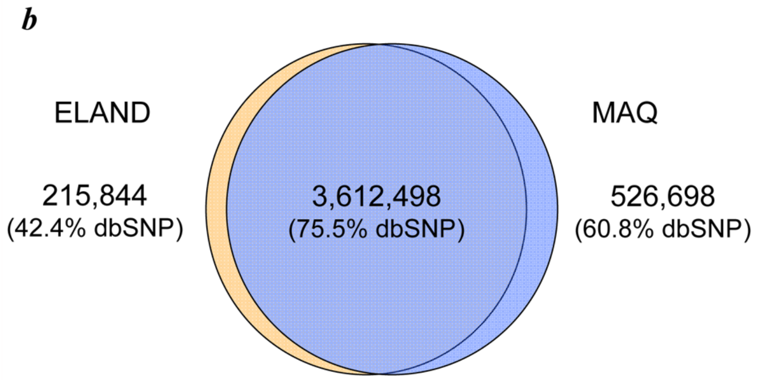


Figure 3.

SNPs identified in the human genome sequence of NA18507. ***a.*** number of SNPs detected by class and % in dbSNP (release 128). Results from ELAND and MAQ alignments are reported separately. ***b.*** Overlap of SNPs detected in each analysis reveals extensive overlap. The % of NA18507 SNP calls that match previous entries in dbSNP is lower than that of our X chromosome study (see fig S6). We expect this because individual NA07340 (from the X study) was also previously used for discovery and submission of SNPs to dbSNP during the HapMap project, in contrast to NA18507.

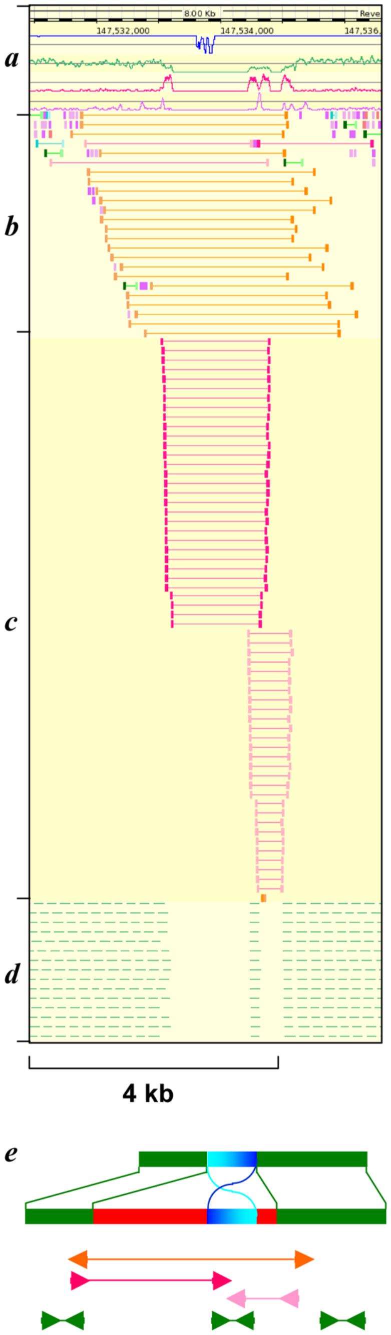


Figure 4.

Homozygous complex rearrangement detected by anomalous paired reads. The rearrangement involves an inversion of 369 bp (blue-turquoise bar in the schematic) flanked by deletions (red bars) of 1206 and 164 bp, respectively, at the left and right hand breakpoints. ***a***. summary tracks in the Resembl browser, denoting scale, simulated alignability of reads to reference (blue plot), actual aligned depth of coverage by NA18507 reads (green plot), density of anomalous reads indicating structural variants (red plot; peaks denote ‘hotspots’), density of singleton reads (pink plot). ***b***. anomalous long insert read pairs (orange lines denote DNA fragment, blocks at either end denote each read); the data indicate loss of ~1.3kb in NA18507 relative to the reference. ***c.*** anomalous short insert pairs of two types (red and pink) indicate an inverted sequence flanked by two deletions. ***d.*** normal short insert read pair alignments (each green line denotes the extent of the reference that is covered by the short fragment, including the two reads). ***e.*** The schematic depicts the arrangement of normal and anomalous read pairs relative to the rearrangement. Top line: structure of NA18507, second line: structure of reference sequence. Green bars denote sequence that is collinear in the reference and NA18507. The turquoise-blue bar illustrates the inverted segment. Red bars indicate the sequences present in the reference but absent in NA18507. Arrows denote orientation of reads when aligned to the reference. Note that the display in ***a-d*** is a composite of screen shots of the same window, overlapped for display purposes in this figure.

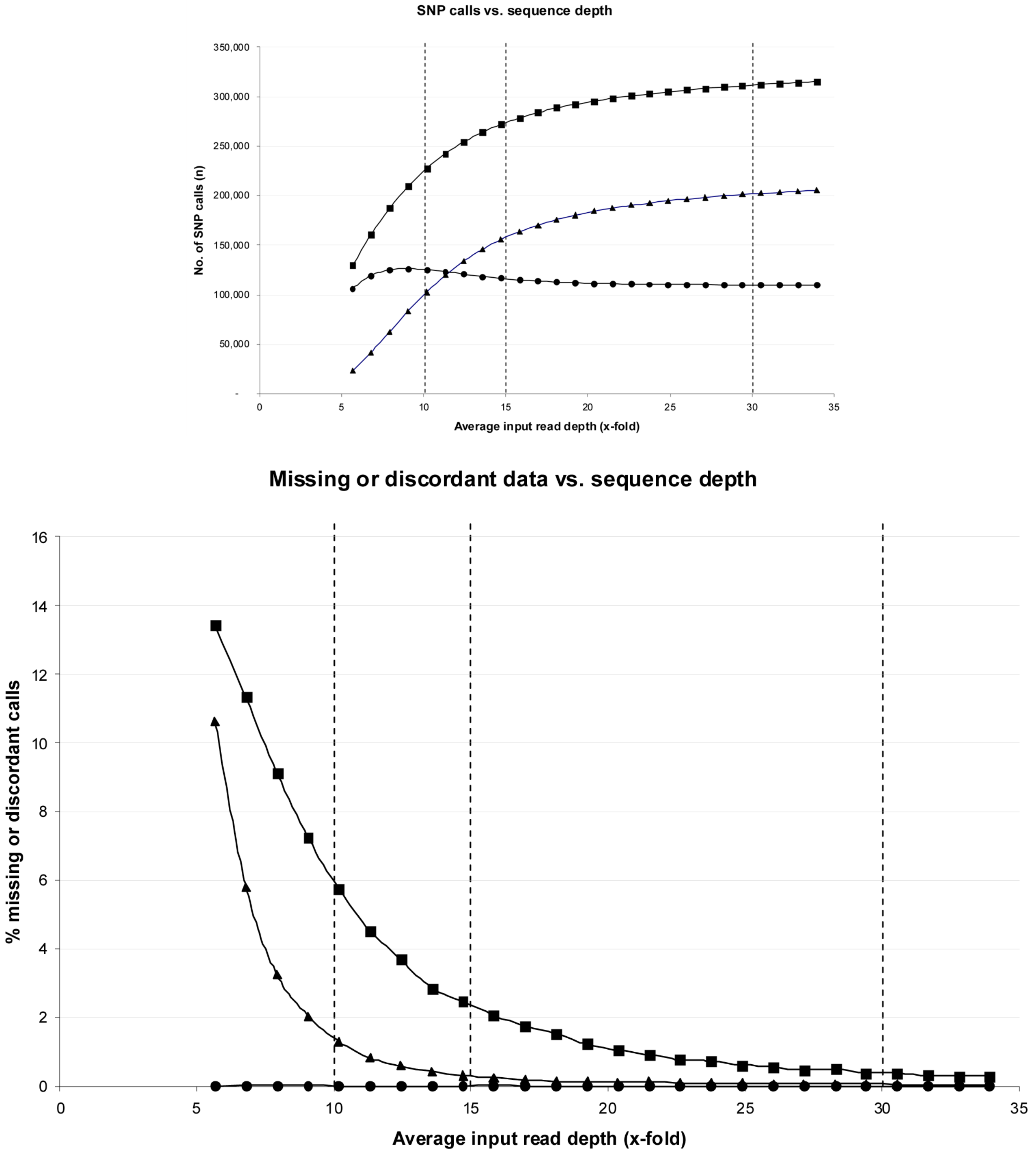


Figure 5.

Effect of sequence depth on coverage and accuracy of human genome sequencing. ELAND alignments were used for this analysis. ***a***. Accumulation of sequence-based SNP calls, including all SNPs (squares), heterozygous SNPs (triangles) and homozygous SNPs (circles) with increasing input read depth. ***b***. Decrease in genotype positions not covered by sequence (squares), heterozygote undercalls in sequence data relative to genotype data (triangles) and discordant SNP calls compared to genotypes (circles) with increasing input read depth. Vertical dotted lines indicate various input read depths (10x, 15x, 30x haploid genome).

Table 1

Comparison of human genome NA18507 SNP calls made from sequence vs. genotype data

	ELAND			MAQ				
Study	X	human	human	X	human	human	human	human
SNP panel	HM550	HM550	HM-All	HM550	HM550	HM-All	Combined	
SNPs (n)	13,604	552,710	3,699,592	13,604	552,710	3,699,592	530,750	
	%	%	%	%	%	%	%	n
Covered by sequence	99.77	99.60	99.24	99.91	99.74	99.29	99.78	529589
Concordant calls	99.52	99.57	98.80	99.99	99.90	99.12	99.94	529285
All disagreements	0.48	0.43	1.20	0.01	0.1	0.88	0.06	304
GT>Seq	0.48	0.35	0.46	0.01	0.03	0.15	0.02	130
Seq>GT	0	0.05	0.52	0	0.05	0.54	0.02	130
Other discordances	0	0.03	0.22	0	0.02	0.2	0.01	44

SNP panels referred to are ‘HM550’ (Illumina Infinium HumanHap550 BeadChip) and ‘HM-All’ (Complete data from phase 1 and phase 2 of the International HapMap Project). ‘Combined’ is a set of concordant genotypes from both sets (HM550 and HM-All) (see text). GT>Seq denotes a heterozygous genotyping SNP call where there is a homozygous sequencing SNP call (one of the two alleles); Seq>GT denotes the converse, i.e. a heterozygous sequencing SNP call where there is a homozygous genotyping call. Other discordances are differences in the two SNP calls that cannot be accounted for by one allele being missing from one call.