nature biotechnology

Single-molecule sequencing of an individual human genome

Dmitry Pushkarev[1,2], Norma F Neff[1,2] & Stephen R Quake[1]

Recent advances in high-throughput DNA sequencing technologies have enabled order-of-magnitude improvements in both cost and throughput. Here we report the use of single-molecule methods to sequence an individual human genome. We aligned billions of 24- to 70-bp reads (32 bp average) to ~90% of the National Center for Biotechnology Information (NCBI) reference genome, with 28× average coverage. Our results were obtained on one sequencing instrument by a single operator with four data collection runs. Single-molecule sequencing enabled analysis of human genomic information without the need for cloning, amplification or ligation. We determined ~2.8 million single nucleotide polymorphisms (SNPs) with a false-positive rate of less than 1% as validated by Sanger sequencing and 99.8% concordance with SNP genotyping arrays. We identified 752 regions of copy number variation by analyzing coverage depth alone and validated 27 of these using digital PCR. This milestone should allow widespread application of genome sequencing to many aspects of genetics and human health, including personal genomics.

There is broad interest in using human genome sequencing to better understand human genetic variation and genome-related diseases, such as cancer, and ultimately to guide discoveries and decisions about the health of individuals. Since the publication of the first rough draft consensus human genomes[1,2], there have been several reports of individual human genome sequences[3–9]. Even using current next-generation technologies, however, sequencing of a human genome has required at least 35–40 machine runs with many instruments operating in parallel. Reagent costs are substantial, estimated at $250,000–$500,000 per genome (**Supplementary Table 1**). Here we demonstrate that genome science is rapidly advancing to the point where individual instruments can achieve throughput that just a few years ago required the facilities of large genome centers.

Since the first demonstration of single-molecule sequencing in 2003 (ref. 10), there has been rapid progress in the field. Published results of various forms of single-molecule sequencing[11–13] have shown improvements in throughput of about tenfold per year over the last few years, and no fundamental limits have yet been reached (**Supplementary Fig. 1** and **Supplementary Table 2**). Single-molecule sequencing is an attractive approach due to its simplicity and the lack of cloning or amplification in sample preparation. High densities of unamplified single molecules on a surface can be extended asynchronously, thereby allowing substantial flexibility in the kinetics of sequencing chemistry. Previous reports of single-molecule sequencing have been proofs of principle[11–13], and their sequencing throughput has not been competitive with alternative approaches. Generally, read lengths have been relatively short and error rates have been dominated by deletions; it has not been clear whether the resulting sequence quality is suitable for human genome sequencing applications.

The Heliscope Single Molecule Sequencer (Helicos Biosciences) is the first commercial release of a single-molecule sequencing instrument. It allows one to follow ~1 billion individual molecules as they are sequenced over the course of a week—a throughput that is practical for human genome sequencing. There have been several technical improvements to the platform since the reported sequencing of a viral genome[12], including more than a 1,000-fold improvement in parallelism, a new generation of sequencing reagents that allows digital measurement of homopolymer sequences, and a new software algorithm, IndexDP, for performing alignments to the entire human genome.

We used two of the instrument's 50 flow-cell channels to resequence the *Staphylococcus aureus* genome as a calibration of sequencer performance. About nine million reads per channel were uniquely aligned to the reference genome (**Supplementary Figs. 2** and **3**), and from this data we were able to determine the raw sequencing error rates. Errors in the raw reads were dominated by deletions (2%), followed by insertions (1.2%) and substitutions (0.38%). Mapping to the reference genome resulted in complete genome coverage and the identification of three SNPs (**Supplementary Table 3**).

We evaluated the performance of the Heliscope for human genome sequencing by determining the genome sequence of a male of European descent (hereafter referred to as Patient Zero or P0). We generated at least 6× human genome coverage of sequence per week-long run. Each full run consisted of 50 channels distributed across two flow cells. We combined data from four instrument runs, during which 172 of the 200 channels were loaded with P0 genomic DNA. Sequence data were mapped to the NCBI 36 reference human genome (hg18) using the open-source aligner IndexDP; 63% of the raw reads were aligned (**Fig. 1a**), yielding a total useful coverage of 28×.

IndexDP is designed to perform efficiently in the presence of deletion errors by allowing insertions or deletions in the seeds, whereas software designed for other short-read technologies (such as ELAND[7], MAQ[14] and SHRiMP[15]) constructs seeds with the assumption that the dominant

[1]Department of Bioengineering, Stanford University and Howard Hughes Medical Institute, Stanford, California, USA. [2]These authors contributed equally to this work. Correspondence should be addressed to S.R.Q (quake@stanford.edu).

Received 10 June; accepted 31 July; published online 10 August 2009; doi:10.1038/nbt.1561

errors are substitutions. Although we did not directly compare IndexDP to other short-read mappers, it is expected to perform better on data generated by the Heliscope Single Molecule Sequencer. Approximately 90% of the reference genome sequence was covered with uniquely mapped reads (2.5 GB out of 2.77 GB). The distribution of coverage depth was close to a Poisson distribution (**Fig. 1b**).

Because IndexDP does not call variant bases and because existing variant callers have been designed specifically for other sequencing platforms, we developed an algorithm to perform variant base-calling on the P0 genome. This algorithm, called UMKA, uses alignment quality scores, accounts for the specific sequencing error profile introduced by the Heliscope and a naive prior probability distribution about the distribution of variation in the human genome. UMKA selects the most probable diploid base call for each position in the genome and also returns a PHRED-like quality score that is the absolute value of the logarithm of the expected error probability at that location. As the naive priors do not use information from the NCBI SNP database (dbSNP) or other catalogs of human variation, we were able to test the performance of UMKA and the quality of the genome assembly by comparing our sequence data to independent experimental measurements at sites of known variation and have confidence in the extrapolation of the results.

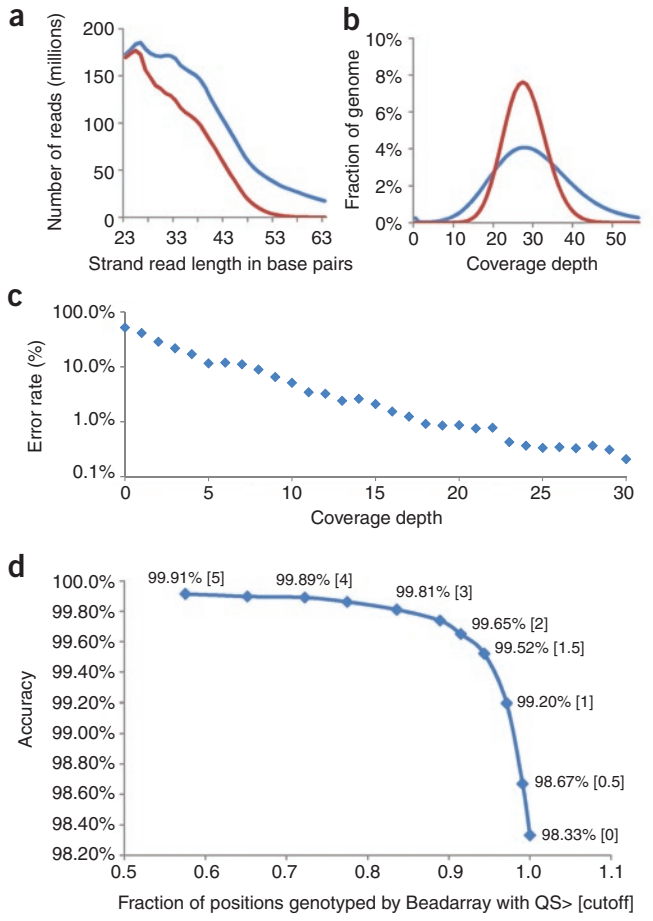

Figure 1 P0 genome sequencing metrics. (**a**) Read length distributions for raw reads (blue) and uniquely aligned reads (red) from Helicos single-molecule sequencing of the genome of Patient Zero (P0). Filtered reads tend to be shorter because a larger proportion of the long reads are instrument artifacts related to the base addition order. (**b**) Coverage depth for sequence data of the P0 genome, computed over repeat masked regions (ENSEMBL, blue) compared to theoretical Poisson limit (red). (**c**) Error rate as a function of sequence coverage depth. Above 30× coverage, sampling noise from the limited number of BeadArray results begins to dominate the error rate, and error rate measurements are not accurate. Error rates are defined as concordance with independent measurement of SNPs using the Illumina Human610-Quad SNP BeadArray (see Online Methods for details). (**d**) Quality score (QS) tradeoffs between sensitivity and accuracy. High sensitivity is obtained by using a QS threshold of 0, which results in calls for all comparison BeadArray locations, with an accuracy of 98.3%. Raising the QS threshold to 1 results in 97% of comparison BeadArray locations being called, thereby lowering the sensitivity but increasing the accuracy of those calls to 99.2%. Numbers next to each data point indicate accuracy (percentages) and cutoff score (in brackets).

One important application of human genome resequencing is to measure the genotype of an individual at sites of known variation. This allows one to determine whether the individual is carrying an allele for a genetic disorder, to determine ancestry or to profile SNPs for pharmacogenomic purposes. We analyzed the accuracy of UMKA's base calling in this context by comparing the sequence data to independent measurements using the Illumina Human610-Quad SNP BeadArray. We used UMKA to call SNPs in the P0 genome and analyzed the results as a function of both coverage (**Fig. 1c**) and quality score (**Fig. 1d**). The base-calling error rate is defined as the fraction of positions that are not in concordance with variant calls made using the BeadArray (Online Methods). As the quality threshold was varied, we found that UMKA was able to call 100% of locations in the comparison pool with 98.3% accuracy and 97% of SNPs with 99.0% accuracy. This is slightly better than the results of the leukemia tumor genome, which found all high-quality SNPs in a reference pool with 94% accuracy[5], and is similar to results obtained for the Asian[6,8,9] and Yoruban[7] genomes.

Whole-genome data are also used to discover new sources of genetic and phenotypic variation. These data are useful in association studies to discover disease alleles and also in efforts to understand the fundamental distribution of human variation. We analyzed the ability of UMKA to discover novel variation by comparing SNPs in the P0 genome to those in the dbSNP database (build 129). Previous individual human genome sequencing studies have found that dbSNP contains validated entries for >70% of the SNPs in an individual genome[3–9]. Similarly, we found that ~76% of SNPs in the P0 genome are listed in dbSNP as validated (**Fig. 2a**). As annotated SNPs are relatively rare (occurring at the part-per-thousand level), the chances that a SNP called in the P0 genome will be both a false positive and also annotated as validated in dbSNP are very low. Therefore, the high proportion of annotated SNPs in P0 suggests a low false-positive rate in SNP discovery.

By examining the proportion of P0 SNPs validated in dbSNP as a function of quality score, we found that it stays relatively constant over a large range of quality scores, and then declines substantially as the quality score falls below 2.8 (**Fig. 2b**). We interpret this as a rapid increase in false-positive SNP calling below that threshold. At a quality score threshold of 2.8, there are 2,805,471 SNPs. The concordance of this set with the BeadArray data is 99.8%, and comparison of the overlap leads to an estimated 4.3% false-negative rate (Online Methods). We randomly selected 100 SNPs with a quality score >2.8 and resequenced them with Sanger sequencing. All of them agreed with the UMKA prediction, thus establishing the false-positive rate as being <1%. The choice of a quality-score cutoff is somewhat arbitrary, and, even at a less stringent threshold of 1.9, the number of SNPs increases to 3,263,470 and there remains substantial agreement in SNP overlap between the P0 genome and those of other males of European descent such as Craig Venter[1] and James Watson[4] (**Fig. 2c**).

Structural variation is an important source of variation in the human genome, and it appears to be dominated by copy number variation (CNV)[6]. A few studies have predicted CNV using small datasets of high-throughput sequencing coverage[16,17], but thus far their application to

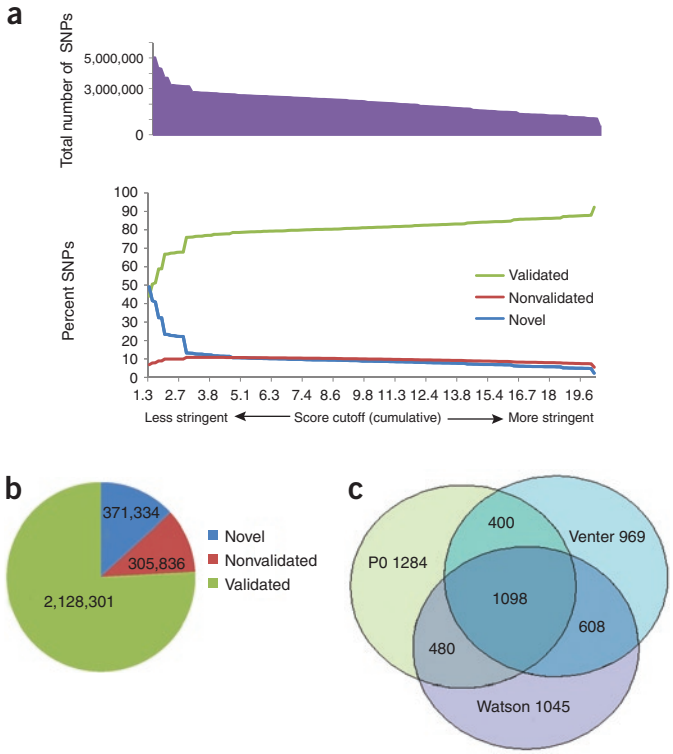

Figure 2 SNP discovery in P0. (**a**) SNP distribution in the P0 genome as a function of quality score. Putative SNPs are 'validated' or 'nonvalidated' if they are annotated as such in dbSNP. Putative SNPs not found in dbSNP are 'novel'. SNPs with larger quality scores are called with higher confidence. A substantial decrease in the proportion of validated SNPs is seen as the quality score drops below 2.8, suggesting that 2.8 is a reasonable threshold for identifying high quality SNPs. (**b**) Distribution of high-quality SNP calls (quality score >2.8) for the P0 human genome. Validated, nonvalidated and novel SNPs are defined as in **a**. (**c**) Overlap in SNP locations between the genomes of P0, James Watson and Craig Venter (in thousands). In this figure the quality-score cutoff was moved to the second plateau in **a** (QS >1.9), increasing the sensitivity and resulting in a total of 3,263,470 SNPs in the P0 genome. This is due to a further 389,736 novel SNPs, 18,495 unvalidated SNPs and 49,768 validated SNPs. The ratio of validated to novel SNPs can be used to estimate that this improvement in sensitivity comes at a cost of an increased overall false-positive rate (from 1% to 10%). Even with this less restrictive cutoff, the SNP proportions shared with Venter and Watson remain consistent.

whole human genomes has had limited sensitivity[4,18], finding many fewer copy number variants than other approaches[6]. We used a binning strategy to determine the density of reads over the genome, from which we were able to make direct estimates of CNV at a resolution of 1 kb (**Fig. 3**). We did not attempt to detect small indels by this method, although it should be possible in principle. CNV detection using this method is, of course, only possible in regions of the reference genome where reads can be uniquely aligned. We detected 752 regions of CNV totaling 16 MB, 54% of which were previously annotated in the Database of Genomic Variants[19]. We used digital PCR[20] to independently validate 27 CNV regions (selected to include 0×, 0.5×, ≥2× copy number variants and both novel and previously annotated regions), 25 of which had quantitative agreement with the predicted CNV (**Supplementary Table 4**).

The individual genome data reported here and previously[3–9] represent important technological advances but are incomplete approximations for various reasons. First, there is both systematic and biological variation of the genome across tissue types within an individual. Post-mitotic cells suffer gradual disregulation of the genome at rates that vary according to tissue type[21], and cells of the immune system reprogram their genomes in specific ways[22]. Second, genome coverage is not complete, and highly repetitive regions are generally not represented. Third, haplotype phasing is difficult to measure and has limited analysis[5,6]. Fourth, structural variation is not determined exhaustively, and there is little independent confirmation. In some cases paired-end reads have been used to show that individual structural variation may be predominantly deletions of various sizes[6–9], but verification and estimates of completeness are lacking. Fifth, when SNPs are measured independently, there are nonzero false-positive and false-negative rates. None of the published individual genomes[3–9] has claimed exhaustive SNP determination or even 100% concordance with independent SNP genotyping; this is due to a trade-off between cost, total coverage and desired accuracy in variant base-calling. We expect that such trade-offs will continue to be important and will strongly depend on the biological questions being asked. For example, when novel SNP discovery is required, it may be useful to accept a lower accuracy in order to sequence more individuals.

More generally, genomes sequenced using short-read technologies provide a wealth of knowledge about the geography of the genome and how that geography varies between individuals. Currently, only a tiny fraction of the data can be interpreted in the context of human traits, but in principle such data could be generated for virtually any known trait and exploited in personalized medicine. Even the approximate measurements of structural variation available today (such as the copy number variants described here) are opening new avenues of genomic research and changing our understanding of human variation. That such measurements can now be performed in virtually any lab using a single commercial instrument serves to democratize access to the fruits of the genome revolution and may enable rapid and widespread adoption of individual genome sequencing in various scientific and medical contexts.

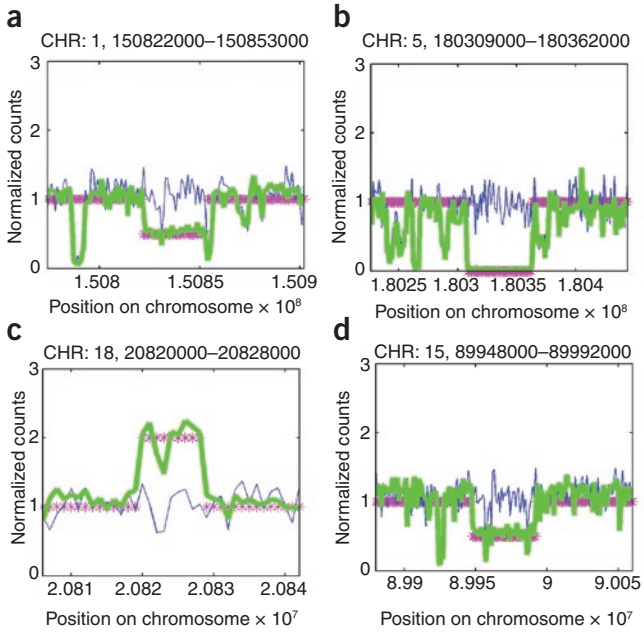

Figure 3 Copy number variation in the P0 human genome. Blue, signal from simulated dataset (simulated reads per 1 kb bin). Magenta, CNV estimate. Green, raw signal (actual reads mapped per 1 kb bin). (**a**) Heterozygous deletion. (**b**) Homozygous deletion. (**c**) Homozygous duplication. (**d**) Heterozygous deletion.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturebiotechnology/.

Accession codes. National Institutes of Health Short Read Archive: sequence data have been deposited with accession code SRA009216.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

We are grateful to V. Natu and J. Coller of the Stanford Functional Genomics Facility for performing the Illumina SNP analysis, R.A. White for assistance with dPCR assays, and A. Sidow for the use of the Covaris sonicator. We acknowledge National Science Foundation award CNS-0619926 for computer resources funding the Bio-X2 cluster, and the National Institutes of Health Pioneer Award (to S.R.Q.).

AUTHOR CONTRIBUTIONS

N.F.N. prepared the libraries, performed the sequencing and wrote the manuscripts. D.P. developed the data analysis algorithms, performed the computations and wrote the manuscript. S.R.Q. designed the research and wrote the manuscript.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturebiotechnology/.

Published online at http://www.nature.com/naturebiotechnology/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Venter, J.C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
2. Lander, E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
3. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5**, e254 (2007).
4. Wheeler, D.A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872–876 (2008).
5. Ley, T.J. *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456**, 66–72 (2008).
6. Wang, J. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456**, 60–65 (2008).
7. Bentley, D.R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
8. Kim, J.I. *et al.* A highly annotated whole-genome sequence of a Korean individual. *Nature* advance online publication doi:10.1038/nature08211 (8 July 2009).
9. Ahn, S.-M. *et al.* The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Res.* published online, doi:10.1101/gr.092197.109 (26 May 2009).
10. Braslavsky, I., Hebert, B., Kartalov, E. & Quake, S.R. Sequence information can be obtained from single DNA molecules. *Proc. Natl. Acad. Sci. USA* **100**, 3960–3964 (2003).
11. Greenleaf, W.J. & Block, S.M. Single-molecule, motion-based DNA sequencing using RNA polymerase. *Science* **313**, 801 (2006).
12. Harris, T.D. *et al.* Single-molecule DNA sequencing of a viral genome. *Science* **320**, 106–109 (2008).
13. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138 (2009).
14. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18**, 1851–1858 (2008).
15. Rumble, S.M. *et al.* SHRiMP: Accurate Mapping of Short Color-space Reads. *PLOS Comput. Biol.* **5**, 1000386 (2009).
16. Daines, B. *et al.* High-throughput multiplex sequencing to discover copy number variants in *Drosophila*. *Genetics* published online, doi:10.1534/genetics.109.103218 (15 June 2009).
17. Herman, D.S. *et al.* Filter-based hybridization capture of subgenomes enables resequencing and copy-number detection. *Nat. Methods* **6**, 507–510 (2009).
18. Xie, C. & Tammi, M.T. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics* **10**, 80 (2009).
19. Iafrate, A.J. *et al.* Detection of large-scale variation in the human genome. *Nat. Genet.* **36**, 949–951 (2004).
20. Qin, J., Jones, R.C. & Ramakrishnan, R. Studying copy number variations using a nanofluidic platform. *Nucleic Acids Res.* **36**, e116 (2008).
21. Vijg, J., Busuttil, R.A., Bahar, R. & Dollé, M.E. Aging and genome maintenance. *Ann. NY Acad. Sci.* **1055**, 35–47 (2005).
22. Janeway, C. *Immunobiology* (Garland Science, New York, 2004).

ONLINE METHODS

Genomic library preparation and Illumina BeadArrays. Genomic DNA was purified from 2 mls of whole blood from P0 using the DNeasy Kit (Qiagen). DNA fragments of ~30 kb were sheared using a Covaris S2 acoustic sonicator using conditions recommended by Helicos. The intensity and time settings were varied to maximize the fragment yield in the 100- to 500-bp range. Sheared DNA fragments were further processed using a Microcon 30 (Millipore) column to remove small fragments and treated with T4 polynucleotide kinase to maximize 3′-OH groups at the ends of the fragments.

Terminal transferase tailing with dATP, 3′-terminal blocking, concentration and quantification used protocols provided by Helicos. Approximately 200 pmoles of dA-tailed molecules were loaded onto each lane of two flow-cells. This represents approximately ~25 pg of DNA per lane. Control oligonucleotides provided by Helicos were loaded into one lane on each flow-cell for sequencing run quality control. Minor modifications to the DNA sample preparation using size selected fragments of 100–300 bp or using Dynal (Invitrogen) oligo-dT magnetic beads to enrich for poly dA–containing molecules did not improve the sequence output.

Two 250 ng samples of P0 genomic DNA were processed and analyzed on Illumina Human610-Quad SNP Beadchips by the Stanford Functional Genomics Facility according to established procedure. *Staphylococcus aureus* genomic DNA, strain USA 300 (size 2.8 Mb, GC 37%) was obtained from ATCC and fragmented to an average size of 220 bp using a Covaris instrument prior to sequencing. The library and its characterization were provided by Helicos and were sequenced at Stanford using two channels of the Heliscope. This yielded ~8.7 M reads per channel aligned to *S. aureus* USA300 reference genome. The entire genome was covered, with a minimum coverage of 7× and median coverage of 180×. Positions were called using majority vote, and three SNPs were identified. Error rates were computed based on actual alignments by dividing number of errors of a given type by the total number of aligned bases.

P0 human genome sequencing. Data was obtained from four machine runs (two full runs and two runs shared with other libraries), yielding 148 GB of raw sequence in 172 channels and an average read length of 33 bp (**Supplementary Table 5**).

Raw reads were filtered with the Helicos filtersms program[14], which removes reads with highly repetitive sequences resembling the base addition order (CTAG) and applies length filters removing reads shorter than 24 bp and longer than 70. Base addition order filter results in a larger fraction of long reads being discarded (**Fig. 1a**). The average length of aligned reads turned out to be 32.

The resulting FASTA files were aligned to human reference build 36 by the open-source aligner IndexDP developed by Helicos. Alignment parameters were chosen to guarantee 100% sensitivity in alignment for reads that share at least one 18 mer with no more than 1 mismatch (substitution, insertion, deletion) with the reference. A low complexity seed filter is applied by the aligner; only the first 65,000 positions for each seed are stored in the index; seeds of a given read that point to >50 positions in the genome are discarded.

The two best alignments for each read were used. In cases where the read had multiple high-quality alignments <200 bp apart, only the best alignment from that region was considered

UMKA scoring. The error rate was estimated based on reads that have the second-best alignment at least 2 errors away from the best alignment. A substitution matrix SM was constructed, the elements of which show the probability of reading a nucleotide given another nucleotide in the reference sequence (**Supplementary Table 6**).

$$SM\,[i,j] = P(B = i \mid R = j),\ ex.SM[2.1] = p(C \mid A) \tag{1}$$

For each hit, the probability of that hit coming from the position on reference it is pointing to was estimated:

$$P(SEQ \mid REF) = \prod_{i=1}^{length} P(SEQ[i] \mid REF[i]) = \prod_{i=1}^{length} SM[SEQ[i], REF[i]] \tag{2}$$

For a given read, for all hits the probability of read coming from position denoted by hit j is:

$$P(HIT[j] \mid SEQ) = P(SEQ \mid HIT[j]) * \frac{P(HIT[j])}{P(SEQ)} = P(SEQ \mid HIT[j]) * \frac{P(HIT[j])}{\sum_{i=1}^{hit} P(SEQ \mid HIT[i]) * P(HIT[i])} \tag{3}$$

By observing that

$$P(HIT[i]) = 1 / L:$$

$$P(HIT[j] \mid SEQ) = \frac{P(SEQ \mid HIT[j])}{\sum_{i=1}^{hit} P(SEQ \mid HIT[i])} \tag{4}$$

UMKA variant calling. Reads that uniquely (second-best alignment is at least 2 errors away from best) align to genome and have no more than 3 errors were used for variant calling.

Five-dimensional (5D) integer vectors were constructed for each position; the first four dimensions of the vector represents the number of reads that call a given base (A,C,T,G) at the position, with the fifth dimension representing the number of reads with gaps at that position. The magnitude of this vector was limited so that the sum of the first four values was normalized at 20.

At the next stage of variant calling we estimate the probability that a given 5D base vector V was obtained as a result of sequencing of a given allelic combination $P(V|XY)$ (XY=[AA,AC,AG,AT,CC,CG,CT,GG,GT,TT]).

In the absence of alignment errors the 5D vector can be thought of as the result of a random walk in 5D sequence space (ACGT-) where the starting point is the true allelic combination and step directions and probability of going in the wrong direction are defined by the substitution matrix defined above.

For example, consider a vector V=[5,7,0,1,1] and base combination XY=AA. If READ(X) is a stochastic process that is equivalent to sequencing base X, then

$$V = \sum_{i=1}^{14} READ(A) \tag{5}$$

Since these read operations are independent, we can obtain the probability distribution resulting from sequencing 14 A bases by taking the convolution of 14 matrices defining probability distribution of single READ() operation.

$$P(V \mid AA) = \bigotimes_{i=1}^{len} \{PD[A]\} \tag{6}$$

where PD[A] is the distribution of outcomes from reading base A as defined by the first column of substitution matrix SM.

After this distribution is computed, $P(V|AA)$ is obtained by simply taking point V in that distribution.

Calling heterozygotes is slightly different, because one should compute the corresponding distribution for each possible combination of heterozygotes (in the case above from 14A-0C to 14C-0A), and these combinations should be weighted by prior probability of each combination, which is derived from the multinomial distribution

$$P(nA, nC) = \frac{(nA + nC)!}{nA! * nC!} * 2^{-nA-nC} \tag{7}$$

After this process is done for all ten allelic combinations (AA,AC,AG,AT,CC, CG,CT,GG,GT,TT), the one with highest probability (MX) is chosen and error rate is estimated to be

$$Perr = 1 - \frac{P(V \mid MX) * P(MX)}{\sum_{XY} P(V \mid XY) * P(XY)} \tag{8}$$

The base-caller outputs a quality score QS = –log10(Perr) that has the meaning of logarithm base 10 of the probability that variant was mistakenly called.

doi:10.1038/nbt.1561

This assumes that sequencing errors are random and are defined by substitution matrix SM.

The main benefit of this approach is the ability to get correct probabilistic estimates of the correctness of the call that take into account instrument-specific substitution and/or deletion rates as well as coverage at the position in question. A simulation was performed to validate that within errors of simulation probabilities reported by the variant caller are correct. This approach allows us to select subsets of data with predefined average accuracy and simplifies the process of planning new sequencing experiments based on coverage and/or accuracy required by the type of subsequent analysis.

Iterated reference. After the initial alignment and variant calling we revised the reference genome and performed a second alignment step to further reduce reference bias and alignment error. The first alignment step yielded 3 million SNPs with a quality score >2.2; for each of these SNPs we constructed a 101-bp sequence that consists of the location that differs from reference (with the variant call as the new reference) and 50 bp of flanking reference sequence. We then aligned all genomic reads to these 3 million sequences (the 'iterated reference') and performed variant calling with UMKA.

Validation. Analysis was performed on the 2,805,471 highest quality SNPs (QS > 2.8).

The P0 genomic sample was also analyzed on the Illumina Human610-Quad SNP BeadArray, and the 279,000 highest quality positions (Illumina concordance score >90, as reported by BeadArray analysis software, and agreement with a replicate BeadArray measurement) were selected as the independent reference set. A false-negative rate of 4.3% was estimated from the fact that only 95.7% of SNPs reported by BeadArray reference set were found in the 2.8 M Helicos high-quality SNP pool.

One can also estimate the false-positive rate by looking at positions of the BeadArray which were called as being identical with the reference. One in 10,000 of these appeared in the 2.8 M Helicos high-quality SNP pool, which yields a 10% false-positive rate estimate under the assumption that the natural SNP occurrence rate is 1 per 1,000 bp. This is most likely an overestimate as it also assumes that the BeadArray internal error rate is <0.01%. A more reliable estimate of the false-positive rate was obtained by independent Sanger sequencing of 100 randomly chosen SNPs (Genewiz). All of the locations sequenced by Sanger sequencing agreed with UMKA variant calls, establishing the false-positive rate as <1%.

SNP annotation. dbSNP build 129 was used for SNP annotation.

Copy number variation. In this study, we binned uniquely aligning reads into 1 kb bins.

We observed variation exceeding the theoretical prediction, which was caused by both alignment errors and is also a natural outcome of substantial nonrandomness in the human genome. To compensate for that error and calibrate the aligner, we simulated 2× of human genome coverage using the simulator described below and repeated the same procedure for genomic reads in terms of CNV analysis.

We have noted a very high correlation (Pearson score 0.72) between the number of hits in corresponding bins for genomic reads and simulated reads; thus, the majority of variation is caused by alignment artifacts, including regions with no sequence (N-regions) and repetitive sequence that gets filtered by the P(REF|HIT) > 0.99 requirement.

The following probabilistic approach was used to perform CNV analysis:

Consider a chromosome, and denote $H[i]$ to be the number of genomic reads aligning to i^{th} 1 kb bin, and $S[i]$ to be the number of simulated reads aligning to the same bin. Thus one would expect $H[i]$ to have close to a Poisson distribution with mean being $S[i]/\text{median}(S)*\text{median}(H)$.

We use that form of the distribution to construct the likelihood estimate of a given copy number at $C[i]$, and penalty J for opening region with copy number different than current.

$$L = \prod_i poisspdf\left(H[i] \mid C[i] * S[i] * \text{median}(H) / \text{median}(S)\right) * J^{1-(C[i]==C[i+1])} \quad (9)$$

Optimal copy number assignments can be obtained by maximizing this function with respect to $C[i]$, which is a well-behaved problem giving optimal solution in linear time.

The Database of Genomic Variants was obtained from the UCSC web site, and studies represented within it were partitioned into three major groups: SNP arrays, CGH and direct sequencing. A CNV region was considered to be validated if it was detected by at least two different types of experiment (5.8% of human genome fall into this category).

We experimentally validated 27 regions of predicted CNV using Digital PCR (BioMark, Fluidigm) (**Supplementary Table 4**). A highly conserved region of chromosome 1 (gene EIF2C1) was used as the experimental reference channel. Out of 27 regions, 25 were called correctly (quantitative agreement of the number of copies with the prediction) and 2 were called incorrectly; however, these two were in qualitative agreement with the prediction as both approaches showed increased copy number relative to the reference gene (ID: 3-60 and 18-20 (**Fig. 3c**)).

Simulator. Simulated reads were constructed by taking 1 million random uniquely aligned reads and creating a set of templates from them, each template recording length, type and position of sequencing errors. Then 2× of human coverage was obtained by applying these templates to random positions of the human genome, such that these reads would capture length distribution and error profile as close as possible.