LETTERS

The complete genome of an individual by massively parallel DNA sequencing

David A. Wheeler[1]*, Maithreyan Srinivasan[2]*, Michael Egholm[2]*, Yufeng Shen[1]*, Lei Chen[1], Amy McGuire[3], Wen He[2], Yi-Ju Chen[2], Vinod Makhijani[2], G. Thomas Roth[2], Xavier Gomes[2], Karrie Tartaro[2]†, Faheem Niazi[2], Cynthia L. Turcotte[2], Gerard P. Irzyk[2], James R. Lupski[4,5,6], Craig Chinault[4], Xing-zhi Song[1], Yue Liu[1], Ye Yuan[1], Lynne Nazareth[1], Xiang Qin[1], Donna M. Muzny[1], Marcel Margulies[2], George M. Weinstock[1,4], Richard A. Gibbs[1,4] & Jonathan M. Rothberg[2]†

The association of genetic variation with disease and drug response, and improvements in nucleic acid technologies, have given great optimism for the impact of 'genomic medicine'. However, the formidable size of the diploid human genome[1], approximately 6 gigabases, has prevented the routine application of sequencing methods to deciphering complete individual human genomes. To realize the full potential of genomics for human health, this limitation must be overcome. Here we report the DNA sequence of a diploid genome of a single individual, James D. Watson, sequenced to 7.4-fold redundancy in two months using massively parallel sequencing in picolitre-size reaction vessels. This sequence was completed in two months at approximately one-hundredth of the cost of traditional capillary electrophoresis methods. Comparison of the sequence to the reference genome led to the identification of 3.3 million single nucleotide polymorphisms, of which 10,654 cause amino-acid substitution within the coding sequence. In addition, we accurately identified small-scale (2–40,000 base pair (bp)) insertion and deletion polymorphism as well as copy number variation resulting in the large-scale gain and loss of chromosomal segments ranging from 26,000 to 1.5 million base pairs. Overall, these results agree well with recent results of sequencing of a single individual[2] by traditional methods. However, in addition to being faster and significantly less expensive, this sequencing technology avoids the arbitrary loss of genomic sequences inherent in random shotgun sequencing by bacterial cloning because it amplifies DNA in a cell-free system. As a result, we further demonstrate the acquisition of novel human sequence, including novel genes not previously identified by traditional genomic sequencing. This is the first genome sequenced by next-generation technologies. Therefore it is a pilot for the future challenges of 'personalized genome sequencing'.

To catalogue the genomic diversity within a single individual, a total of 106.5 million high-quality reads were generated by 454-sequencing[3], representing approximately 24.5 billion DNA bases. Reads that aligned to the genome were further filtered using stringent criteria to ensure the accuracy of mapping, resulting in 93.2 million reads aligned to reference genome sequence. The reference genome sequence was thus covered to an average depth of 7.4-fold (Fig. 1a). The alignments between the uniquely mapped reads and the reference genome were used to catalogue genetic variation in the subject's DNA, including single nucleotide polymorphisms (SNPs), small insertions and deletions (indels), and copy number variation (CNV).

The 454 base-calling software provides error estimates (*Q* values) for each base. We developed a three-step filtering process using the patterns of error and associated *Q* values from the 454 base-calling software to improve the accuracy of SNP discovery. An initial 14 million variant positions were filtered to 3.32 million putative SNPs (Table 1).

Comparison of these putative SNPs in the subject's genome with those in the dbSNP (dbSNP: http://www.ncbi.nlm.nih.gov/projects/SNP/) revealed 2.72 million in common ('known SNPs'). Approximately 99% of SNPs in dbSNP are bi-allelic. At only 10,425 positions did the subject's variant not match the variant found in dbSNP. Although some of these could represent a third allele in the population, or an error in the dbSNP polymorphism record, we conservatively estimated the false discovery rate in the known SNPs to be approximately 0.38% based on the mismatches with dbSNP.

The remaining 0.61 million SNPs were at positions not previously identified as polymorphic in dbSNP ('novel SNPs'). The known SNPs were divided almost equally between homozygous (50.2%) and heterozygous (49.8%) SNPs, whereas within the novel SNPs heterozygotes predominate (83.3%) compared with homozygotes (16.7%). Because most common alleles in human populations are already captured in dbSNP, novel variants are expected to be rare, and therefore much more likely to be found as heterozygotes.

We assessed the accuracy of the known SNPs derived from DNA sequencing by comparison with the experimental genotyping of the subject's DNA using an Affymetrix 500K microarray. Compared with a haploid reference sequence, there are four possible outcomes of SNP array genotyping: homozygous for the reference allele; homozygous for the variant (non-reference) allele; heterozygous; and assay failure. Table 2 shows the results for 494,713 markers that were successfully genotyped. The subject's DNA sequence exhibited only the reference allele at 99.4% of the markers homozygous for the reference and at 95.1% of markers homozygous for the variant. Genotyping identified 135,413 heterozygous markers of which 75.8% exhibited two alleles in the 454-reads. The lower sensitivity of detection of heterozygotes is predicted by a Poisson process of sampling DNA fragments modelled on a diploid genome (Methods). Consistent

[1]Human Genome Sequencing Center, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA. [2]454 Life Sciences, Roche Diagnostics, 20 Commercial Street, Bradford, Connecticut 06405, USA. [3]Center for Ethics and Health Policy, Baylor College of Medicine, One Baylor Plaza, Houston Texas 77030, USA. [4]Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, Houston Texas 77030, USA. [5]Department of Pediatrics, Baylor College of Medicine, One Baylor Plaza, Houston Texas 77030, USA. [6]Texas Children's Hospital, Texas Medical Center, Houston, Texas 77030, USA. †Present addresses: Molecular Imaging Systems, Carestream Health, Inc., 4 Science Park, New Haven, Connecticut 06511, USA (K.T.); Rothberg Institute for Childhood Diseases, 530 Whitfield Street, Guilford, Connecticut 06437, USA (J.M.R.).
*These authors contributed equally to this work.

with this model, the coverage was lower at the 24.2% of heterozygous positions where DNA sequencing represented only one of the correct alleles (Fig. 1b and Supplementary Table 2). The Poisson model further shows that 13-fold average coverage would be required to detect 99% of all heterozygous SNPs (Supplementary Fig. 4).

The DNA sequencing genotypes disagree with the SNP array genotyping in 4,948 cases, or 1.0% of the time; another 3,499 markers (0.30%) had no coverage, consistent with the genome-wide redundancy of the sequence. Assuming the sensitivity and specificity of the markers on the microarray is representative of those found throughout the human genome, we estimate the total number of SNPs in the subject's genome to be approximately 3.7 million (see 'Sensitivity and specificity of SNP discovery' in Supplementary Information).

We identified 222,718 indels ranging from 2 to 38,896 bp, with 113,539 in common with indels in dbSNP; 85,418 indels are found as homozygotes, and 137,300 were heterozygotes. Insertions account for 65,677 events, and deletions 157,041. A portion of the deletion size distribution from 2 to 1000 bp is shown in Fig. 1c. The distribution of deletions shows the relative enrichment for events in the range 300–350 bases in length, as expected from the known polymorphism of *Alu* SINE elements[4,5]. The size range over which insertions are detected is limited by the length of the reads; in the analysis of our 250 base-pair 454-sequencing reads, the largest observed was 208 bp.

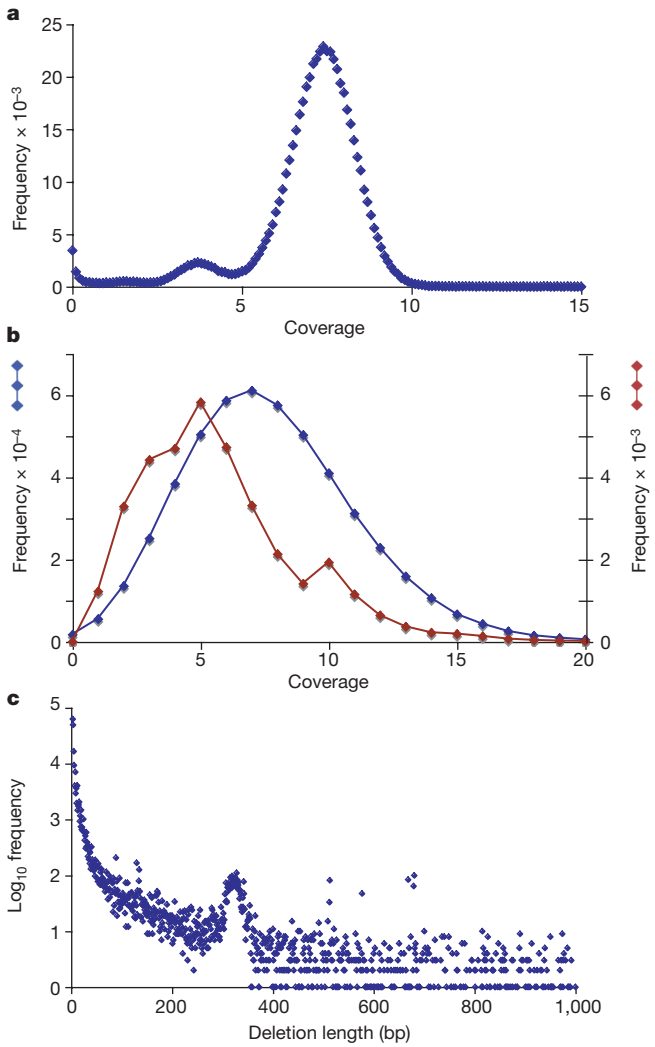

Figure 1 | 454-Sequencing of individual genome generated even coverage enabling genome-wide detection of variation. a, Distribution of sequence coverage of reference genome by 454-reads is random. Coverage, calculated in consecutive 5-kb windows, exhibited a Poisson distribution with a mean of 7.4-fold across all chromosomes except the X. Shoulder at 3.7X represents coverage of the X chromosome. **b**, Coverage is a key factor in detection of both alleles at heterozygous positions. For 31,709 markers heterozygous by microarray, but which exhibited only a single allele by DNA sequencing, the coverage was lower (red line, mean 5.7X) than the overall coverage for all SNPs (blue line, mean 7.8X). **c**, Size distribution of deletions. Deletions were readily observed in alignments of 454-reads to the reference genome. Note the peak in the size range at 300–350 bases owing to polymorphic Alu transposon insertion sites.

Table 1 | Single nucleotide variation in 454 reads

Subject	Filter*	Total variation	Known†	Novel
Watson	Raw	14,829,087	3,283,273	11,545,814
	1	4,427,488	2,815,322	1,612,166
	2	3,971,513	2,752,991	1,218,522
	3	3,322,093	2,715,296	606,797
Venter‡	4	3,470,669	2,822,902	647,767

* Filters: raw, all base substitution from cross_match alignments; 1, $S_v > 28$ (see Methods); 2, filter 1 plus ratio of variant to total coverage >0.2; 3, filter 2 plus eliminate SNPs close to homopolymer runs >5 bp; 4, (Venter) Phred (ref. 20) Q > 15, ratio of variant to total coverage >0.2.
† Variants found in build 126 of dbSNP (http://www.ncbi.nlm.nih.gov/SNP/).
‡ SNPs found in genome of Venter: see ref. 2 and supplementary material therein.

A total of 345 indels were observed to overlap coding sequence and had the potential to alter protein function. We designed primers to amplify and validate by Sanger sequencing 111 of these events with a size range of 3–50 bases. A total of 78 indels were successfully validated of which 66 were observed to be in length multiples of 3, ranging from 3 to 33 bp, and hence not expected to cause protein translation frame shifts. Sixty-five of these indels were found as heterozygotes. Surprisingly, a 4-bp deletion in exon 11 of *SGEF* was found to be homozygous; however, this gene is highly conserved in vertebrate species from rhesus macaque to stickleback, and all manifest the same 4-bp deletion in their genome (Supplementary Fig. 5). Furthermore, two other independent human messenger RNAs (mRNAs) harbour the deletion as well, suggesting the subject's allele was the wild type and that the reference harbours a rare insertion.

CNVs are local gains or losses of regions in the genome owing to duplication or deletion that can be associated with genetic disease[6] and which should be detectable by variation in the average DNA sequence coverage of the region. A comparative genomic hybridization (CGH) microarray analysis of the subject's DNA revealed 23 apparent CNV regions ranging in size from 26 kb to 1.6 Mb: 9 with DNA gains and 14 with a loss. The sequence coverage data exhibited a gain or loss congruent with the CGH result at 18 of the 23 regions (Supplementary Table 4). Regions of CNV are polymorphic in populations, segregating as alleles with varying frequency[7,8]. Consequently, the interpretation of a CGH microarray depends on the reference genome with which the subject is compared. This difference in reference standard is unavoidable when comparing CGH with DNA sequencing results using National Center for Biotechnology Information (NCBI) build 36 as the reference, which is not based on a single individual and for which no physical DNA sample exists. We experimentally demonstrated variation in CGH results by repeating the CGH array using a second reference genome and two different array platforms, demonstrating the effect the reference DNA has on the outcome of a CGH experiment (Supplementary Table 4).

An individual region of homozygous loss was characterized further using CGH results and DNA sequencing (Supplementary Fig. 6c). Sequence alignment of the subject's reads spanning the breakpoint of

Table 2 | Microarray validation of 454 SNPs

Affymetrix genotype*	Affymetrix SNP array	454 Sequence†	Agreement (%)
Homo ref.	254,753	253,348	99.4
Homo var.	104,547	99,387	95.1
Hetero	135,413	102,702	75.8

* Homo ref., homozygous for reference allele; homo var., homozygous for the variant allele; hetero, heterozygous.
† The genotype based on the alleles observed in 454 reads at each position of an Affymetrix marker.

a homozygous deletion region reveals a 2-bp addition at the breakpoint junction, suggesting non-homologous end joining[6] was the mechanism involved in generating the deletion, and demonstrating the feasibility of using 454 sequence reads for identifying CNV breakpoints (Supplementary Fig. 6c). Several other CNV regions were flanked by repeats and segmental duplications, and likely occur by non-allelic homologous recombination, as was reported recently for the CNV loss at 22q13.1 (ref. 9) (see CNV 23 in Supplementary Table 4).

None of the CNV regions we defined are currently known to be involved in a recognizable phenotype; however, either trait or disease susceptibility correlations could occur in the future[10]. Thirty-four genes are predicted to be affected by these gains and losses, including two separate olfactory receptor groups, several genes with possible roles in cancers of the prostate, breast and colon, a gene from the HLA-D locus, and two proteins thought to be involved in RNA editing (Supplementary Table 4).

Among the 3.3 million SNPs found in the subject's genome were 8,996 non-synonymous changes in known SNPs and 1,573 in novel SNPs. We compared the non-synonymous known SNPs with the Human Gene Mutation Database (HGMD), the largest current compendium of human disease alleles[11]. Thirty-two alleles exactly matched mutations reported in the HGMD whereas an additional 310 of these were in HGMD genes but were either alleles or amino-acid positions not previously characterized as disease-causing. In 12 cases the specific alterations and loci consisted of genes where homozygous recessive alleles can give rise to disease or other recognizable phenotype (Table 3); and 20 cases are reported to be associations with increased disease risk (Supplementary Table 5).

Ten of the 12 alleles in Table 3 are thought to be highly penetrant, Mendelian recessive disease-causing alleles. Seven of ten were heterozygous in the subject's genome sequence; the other three only exhibited one allele but have an average sequence coverage less than fivefold. Because the subject does not have these three diseases, and we expect not to recover the second allele for 24% of heterozygous positions, it is likely that he is not homozygous for these disease allele positions. We note that there are not yet any systematic studies of the population frequencies of these alleles. Nevertheless, the subject is a carrier for ten highly penetrant genetic disease loci found in the HGMD data set consisting of 900 genes. It has been estimated there are fewer than ten lethal equivalents in each person[12–14]. Because we have drawn ten from an HGMD subset of the genome, we would predict the subject harbours a much greater number of deleterious Mendelian mutant alleles than is commonly estimated.

In addition, a sampling of 3,898 of the non-synonymous SNPs were tested for their possible functional impact on the protein sequence using the software Polyphen[15]. Polyphen classified 7.3% as 'probably damaging', suggesting these changes will be of functional consequence to the protein. The remainder were classified as either 'possibly damaging' (13%) or 'benign' (74%).

The genome sequence of another individual (C. Venter) was recently reported[2]. That study reported a 7.5-fold genome coverage using Sanger reads. The Venter genome harboured approximately 2.8 million known SNPs and about 0.74 million novel SNPs, in close agreement with the results from 454-reads of the Watson genome at a similar fold coverage. The two individuals shared 1.68 million of the SNPs, of which 5,230 were non-synonymous, accounting for 58% of the subject's non-synonymous SNPs. Watson and Venter are each distinguished from the reference by 3,766 and 3,882 non-synonymous SNPs, respectively, and therefore are different from each other by 7,648 protein coding changes. This is the most comprehensive comparison of the non-synonymous difference between two diploid genomes yet undertaken.

The subject's data also contained 1.5 million reads of novel sequence that did not map to build-36, corresponding to about 1.4% of the total sequence data. Approximately 65% of the unmapped sequences matched to known human repeats enriched for satellite DNA and other repeat elements characteristic of heterochromatin (Supplementary Fig. 7). The novel reads were assembled into approximately 170,000 contigs spanning 48 Mb. After removing contigs with fewer than 100 bp of contiguous unique sequence, 110,000 contigs spanning 29 Mb remained, which is close to the 25 Mb of euchromatic sequence predicted to be absent from the reference genome[1]. These non-repeat contigs closely match to 33 human complementary DNA (cDNA) sequences from a variety of tissues and predicted functions (Supplementary Table 6) having no known map location on the human reference genome (see Methods). To assess further the gene-coding potential of these novel DNA sequences, we compared conceptual translations of the contigs greater than 1,000 bp (1,279 in all) with the GenBank non-redundant (NR) protein database. This search yielded 60 significant, but not identical, matches to 49 different proteins in humans and other vertebrates (Supplementary Table 7). The annotations of several of these transcripts are consistent with transcription factor or signalling molecules. Therefore it is possible this diploid genome sequence will contribute important new genes to the human genome.

The sequencing method used in this study has many advantages over traditional capillary sequencing. It is inherently scalable, which means that sequencing costs in the miniature continue to decrease and throughput increases as the density of the sequencing reactions on the chip increases, and read lengths get longer[3]. In this study we sequenced the genome of Dr Watson for less than US$1 million,

Table 3 | SNPs matching HGMD mutations causing disease or other phenotypes

HGMD accession	Chromosome	Coordinate	HUGO symbol	Gene name	Cytogenetic	Phenotype	Zygosity
CM003589	1	97937679	DPYD	Dihydropyrimidine dehydrogenase	1q22	Dihydropyrimidine dehydrogenase deficiency	Heterozygous
CM950484	1	157441978	FY	Duffy blood-group antigen	1q	Duffy blood group antigen, absence	Homozygous*
CM942034	4	619702	PDE6B	Phosphodiesterase 6B, cGMP-specific, rod, beta	4p16.3	Retinitis pigmentosa 40	Heterozygous
CM021718	9	36208221	GNE	UDP-N-acetylglucosamine 2-epimerase	9p	Myopathy, distal, with rimmed vacuoles	Heterozygous
CM980633	10	50348375	ERCC6	Excision repair cross-complementing rodent repair deficiency, complementation group 6 protein (CSB)	10q	Cockayne syndrome	Homozygous†
CM050716	11	76531431	MYO7A	Myosin VIIA	11q13.5	Usher syndrome 1b	Homozygous†
CM950928	12	46812979	PFKM	Phosphofructokinase, muscle	12q13.3	Glycogen storage disease 7	Homozygous*
CM032029	14	20859880	RPGRIP1	Retinitis pigmentosa GTPase regulator interacting protein 1	14q11	Cone–rod dystrophy	Heterozygous
CM984025	19	18047618	IL12RB1	Interleukin-12 receptor, beta 1	19p13.1	Mycobacterial infection	Heterozygous
CM024138	19	41014441	NPHS1	Nephrosis-1, congenital, Finnish type	19q	Congenital nephrotic syndrome, Finnish type	Heterozygous
CM910052	22	49410905	ARSA	Arylsulphatase A	22q	Metachromatic leukodystrophy	Heterozygous

* Coverage at these SNP positions is less than 5. However, both produce benign phenotypes.
† Coverage at these SNP positions is greater than 5. Both would produce severe phenotypes if they were truly homozygous.

Box 1

Protection of human subjects

Is institutional review board approval required for this project?

Considerations. Approval by an institutional review board (IRB) is required for all research involving human subjects. Federal regulation defines research as 'a systematic investigation, including research, development, testing and evaluation, designed to develop or contribute to generalizable knowledge.' A human subject is defined according to the regulations as 'a living individual about whom an investigator ... conducting research obtains (1) data through intervention or interaction with the individual, or (2) identifiable private information.' (45 CFR 46.102). Baylor College of Medicine, Houston, Texas, requires that all proposed activities at the college be reviewed to determine if they meet the regulatory definitions for research involving human subjects (Baylor College of Medicine, IRB Procedures, November 2006). The research team and the Baylor College of Medicine IRB agreed that the activities associated with this project constitute research involving human subjects. IRB review helps to ensure ethical research conduct and appropriate subject protection. It also sets an important standard for future research in the field of personalized genomics.

Management. The research protocol was written in consultation with an ethicist and reviewed by the Baylor College of Medicine IRB. The research participant's identity was not revealed to the IRB, to ensure objectivity. Although the practical management of many of the ethical issues depended on the unique expertise of the research participant, this did not affect review or approval of the research protocol.

Returning research results to research participants

Should the research participant be able to receive information about their individual genome sequence?

Considerations. Dr Watson requested that he receive information about all data generated from this research project. Generally, patients have a right to receive medical information, but this right does not typically extend to individual research results. Concerns include validity of genetic tests not performed in a Clinical Laboratory Improvement Amendments (CLIA)-approved laboratory, the unknown clinical significance of much of the generated data, and protecting subjects from potentially harmful information, such as information about genetic risk of uncertain penetrance. These considerations have to be weighed against the ethical principles of respect for autonomy and the right to receive relevant information about oneself, as well as the principle of reciprocity, which suggests a right to receive information in exchange for research participation.

Management. The research team felt that because of Dr Watson's unique expertise he would be able to understand adequately the significance and limitations of these data. Therefore, out of respect for his autonomous decision to receive the information, Dr Watson was given his entire genome sequence on a miniature hard drive. Genetic counselling was provided to help with interpretation and to ensure adequate understanding of the information given and the limitations of its clinical significance. It remains controversial whether other research participants who do not share Dr Watson's expertise ought to be informed of individual results of genetic research. Certainly for many, whole-genome data will be meaningless. A more analysed form of the data may therefore be required.

Should the research participant be able to request that certain information be redacted before individual and/or public disclosure?

Considerations. Although patients generally have a right to receive certain medical information, they can waive that right and request that information be withheld from them. The right not to know of a genetic risk is legally sanctioned but remains ethically controversial. Even if this right is recognized, however, decisions about redactions must be made *a priori* to preserve the right not to know a particular finding. In the context of whole-genome sequencing research, because genomic information is stable over time but our understanding of the clinical significance of that information continues to grow at a very fast pace, we cannot anticipate future findings that may reveal genetic risk information that the participant may not have wanted to know. We also cannot entirely eliminate the risk that inferences could be made about the missing information from downstream and/or upstream data.

Management. Because Dr Watson is knowledgeable about and familiar enough with the current literature in genetics to assess research findings and to make an informed decision about what risk information he does and does not want to receive, his right to redact information was respected. Decisions about redactions were made *a priori* and the problems associated with future findings, as well as general concerns about receiving specific genetic information, were discussed with a genetic counsellor. Dr Watson requested that all gene information about apolipoprotein E be redacted, citing concerns about the association that has been shown with Alzheimer's disease. These data were redacted and were not analysed by the research team. Again, this approach is not generalizable and may not be appropriate for other research participants.

Data release and data flow

Should the participant's genome sequence be publicly released?

Considerations. There is great scientific interest in accessing and studying the data generated from this project. To maximize scientific and clinical use, public data release is strongly encouraged in genomic research. Dr Watson is personally committed to a policy of open access to DNA data. However, because DNA is a unique identifier, there are privacy risks associated with data sharing. Because this project was publicly announced and Dr Watson was individually identified, there was concern about his privacy interests and the potential harm that could result from the misuse of his genetic information.

Management. An individual can waive their right to privacy and share personal information with others. Dr Watson decided to share his personal genome by releasing it into a publicly accessible scientific database. The privacy risks associated with public data broadcast were explained.

What, if any, obligations are owed to third-party relatives?

Considerations. Because genetic information is familial by nature, Dr Watson's participation in this research raised concerns about what obligations, if any, were owed to his close biological relatives. Dr Watson's autonomy-based rights had to be weighed against the rights and welfare of his biological relatives. There are three research-related activities that raise concerns about obligations to third-party relatives: consent for research participation; returning research-related results; and data release.

1. Third-party relatives are not typically considered research participants and their consent is not generally required for research participation. The participant's autonomous decision to participation in research typically outweighs any objections raised by third-party relatives.
2. In the clinical context, the physician's duty to warn biological relatives of genetic risk remains controversial. In the research context, because the data are not validated in an approved laboratory, the obligation to warn at-risk relatives is even more tenuous. As in the clinic, all research participants should be informed of the risks to relatives and encouraged to discuss research results with them. In this project, genetic counselling was offered for family members, free of charge.
3. There are privacy risks associated with the public release of genomic data for the individual research participant. There are also risks, though of a more uncertain nature, of public release for close biological relatives, especially when the participant is clearly identified and the release of his data is publicized. What obligation, if any, does an investigator have to protect third-party relatives from these privacy risks?

Management. Risks to relatives were disclosed and thoroughly discussed with the research participant. He was strongly encouraged to discuss these issues with biological relatives and to make a family decision about research participation and data release. The issue of whether investigators should publicly release Dr Watson's data without familial consent was avoided by presenting the edited genome sequence to Dr Watson who then released it himself directly into publicly accessible databases. This did not alleviate any moral obligation Dr Watson may have to protect his biological relatives from the uncertain privacy risks associated with public data release. It is also unlikely that future research participants will be able to facilitate their own data release, making the question of obligations to third-party relatives a policy priority.

whereas the genome of Venter by Sanger sequence reportedly cost approximately US$100 million. Although not used in this study, this sequencing technology allows the production of mate-paired reads. The use of mate-pair reads will enable assessment of a wider range of indels and other structural rearrangements, and facilitate the incorporation of new sequence into the reference genome[16]. The principal weakness of the method is that it currently does not allow efficient detection of single-base indels in homopolymers. Future developments in chemistry and software will improve the ability to identify single-base indels.

A key aim of personal genome sequencing is to identify genome sequences that may be associated with disease, or are predictive of response to medication. The need to make genotype–phenotype correlations before having predictive value is at the heart of both the excitement and the dilemma of the new era of genomic medicine[17]. Thus the ability to sequence individuals readily using high-throughput, scaleable, low-cost, completely *in vitro* technology, as demonstrated here, is an important milestone in our ability to connect 'personalized genomes' to 'personalized medicine' and enable these critical correlations to be made.

METHODS SUMMARY

Mapping and alignment to the genome. DNA sequencing on the Genome Sequencer FLX instrument (454, Inc.) is described in detail in Methods. Reads, averaging approximately 250 bases, were mapped on to the human reference genome, NCBI build 36, by sequence alignment using Basic Local Alignment Search Tool (BLAST)-like alignment tool (BLAT). Reads were removed from subsequent analysis that failed to meet minimum criteria (see Methods). Reads that passed the alignment quality criteria were realigned to local reference genome fragments using Cross_match software; the refined alignments were parsed for sequence variation between the subject and the reference. An error model was developed to separate sequencing error from true genomic variation, and the location and type of each putative true variant was tabulated (see Methods).

Assembly of non-matching reads. One and a half million reads that failed to find a match in the reference genome sequence ('no-hit' reads) and another 2.2 million reads with low-quality alignments to the reference were pooled for sequence assembly. All reads were trimmed to remove the last 50 bases in which most of the sequencing error lies. Reads less than 50 bases after trimming were discarded. The assembly of the remaining 2.6 million trimmed reads followed the standard ATLAS-WGS procedure[18,19].

Laboratory analysis of genomic DNA. A DNA sample from the subject was labelled and annealed to the Affymetrix 500K GeneChip array to provide independent laboratory analysis of the subject's SNPs and conformation of the DNA sequencing coverage. To compare local fluctuation in DNA sequencing coverage with copy number variation in the subject's genome, three additional DNA samples were labelled and mixed each in a 1:1 ratio with separately labelled control DNAs. The mixtures were annealed to each of two Agilent 244K array CGH chips and one Nimblegen HD2 chip.

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

Received 3 December 2007; accepted 4 March 2008.

1. The International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431,** 931–945 (2004).
2. Levy, S. *et al.* The diploid genome sequence of a single individual. *PLoS Biol.* **5,** e254–e286 (2007).
3. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437,** 376–380 (2005).
4. Bennett, E. A., Coleman, L. E., Tsui, C., Pittard, W. S. & Devine, S. E. Natural genetic variation caused by transposable elements in humans. *Genetics* **168,** 933–951 (2004).
5. The Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437,** 69–87 (2005).
6. Lupski, J. R. & Stankiewicz, P. Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. *PLoS Genetics* **1,** 627–633 (2005).
7. Weber, J. L. *et al.* Human diallelic insertion/deletion polymorphisms. *Am. J. Hum. Genet.* **71,** 854–862 (2002).
8. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444,** 444–454 (2006).
9. Kidd, J. M., Newman, T. L., Tuzun, E., Kaul, R. & Eichler, E. E. Population stratification of a common APOBEC gene deletion polymorphism. *PLOS Genetics* **3,** 584–592 (2007).
10. Lupski, J. R. Structural variation in the human genome. *N. Engl. J. Med.* **356,** 1169–1171 (2007).
11. Stenson, P. D. *et al.* Human Gene Mutation Database (HGMD): 2003 update. *Hum. Mutat.* **21,** 577–581 (2003).
12. Halligan, D. L. & Keightley, P. D. How many lethal alleles? *Trends Genet.* **19,** 57–59 (2003).
13. Bittles, A. H. & Neel, J. V. The costs of human inbreeding and their implications for variations at the DNA level. *Nature Genet.* **8,** 117–121 (1994).
14. Vogel, F. & Motulsky, A. G. *Human Genetics: Problems and Approaches* 2nd edn 487–502 (Springer-Verlag, New York, 1986).
15. Ramensky, V., Bork, P. & Sunyaev, S. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.* **30,** 3894–3900 (2002).
16. Korbel, J. O. *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318,** 420–426 (2007).
17. McGuire, A. L., Caulfield, T. & Cho, M. K. Research ethics and the challenge of whole genome sequencing. *Nature Rev. Genet.* **9,** 152–156 (2008).
18. Havlak, P. *et al.* The Atlas genome assembly system. *Genome Res.* **14,** 721–732 (2004).
19. Richards, S. *et al.* Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal, gene, and cis-element evolution. *Genome Res.* **15,** 1–18 (2005).
20. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8,** 186–194 (1998).

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

Acknowledgements We thank J. D. Watson for his participation, encouragement and engagement. We thank J. Belmont for discussion of HGMD data. We also thank the following individuals for their technical support: S. Attiya, M. Braverman, J. Brunelle, C. Celone, Z. Chen, A. Sancher, W. Song, and personnel from 454 Sequencing Center and R&D.

Author Information The CEL files for the Affymetrix 500K Genome array and GPR files for the Agilent_1 and Agilent_2 CGH arrays are deposited in Gene Expression Omnibus under series accession number GSE10668. All SNPs and insertion/deletions from 454-reads are deposited in dbSNP under handle bcmhgsc_jdw; contigs from the assembly of 454-reads not matching the reference genome are deposited in Genbank under accession number ABKV01000000. All 454-reads are deposited in the Trace Archive of the National Center for Biotechnology Information under Center_name = 'CSHL' and Center_project = 'Project JIM'. Reprints and permissions information is available at www.nature.com/reprints.

The authors declare competing financial interests: details accompany the paper at www.nature.com/nature. M.E., W.E., Y.-J.C., V.M., G.T.R., X.G., F.N., C.L.T., G.P.I. and M.M. are current employees of 454 Life Sciences, which is owned by Roche Diagnostics, receive a salary for their work and are eligible for bonuses administered by 454 Life Sciences and Roche Diagnostics. Correspondence and requests for materials should be addressed to J.M.R. (jonathan.rothberg@gmail.com) or R.A.G. (agibbs@bcm.tmc.edu).

METHODS

DNA sequencing. Genomic DNA was purified from white blood cells from Dr Watson by using the Flexigene DNA kit (Qiagen). Five micrograms of DNA were sheared by nebulization and fractionated on agarose gel to isolate 450–550 base fragments. These were used to construct a single-stranded library that was used as template for single-molecule PCR on 28-μm diameter beads in emulsions[3]. The amplified template beads were recovered after emulsion breaking and selective enrichment. Sequencing primer was annealed to the template and the beads were incubated with *Bst* DNA polymerase, apyrase and single-stranded binding protein. A slurry of the template beads, enzyme beads (required for signal transduction) and packing beads (for *Bst* DNA polymerase retention) was loaded into the wells of a 70 mm × 75 mm picotiter plate. The picotiter plate was inserted in the flow cell and subjected to pyro-sequencing on the Genome Sequencer FLX instrument (454, Inc.).

The Genome Sequencer FLX flows 100 cycles of four solutions containing either dTTP, αSdATP, dCTP and dGTP reagents, in that order, over the cell. For each dNTP flow, a single 38-s image was captured by a CCD (charge-coupled device) camera on the sequencer. The images were processed in real time to identify template-containing wells and to compute associated signal intensities. The images were further processed for chemical and optical cross-talk, phase errors and read quality before base calling was performed for each template bead.

Mapping 454-reads to reference genome. We generated sequence with 234 runs on Genome Sequencer FLX instruments (454 Inc.), which produced over 105 million bases per run. Reads were aligned to the human reference genome, NCBI build 36 (http://hgdownload.cse.ucsc.edu/goldenPath/hg18/bigZips/), using BLAT. All but 1.5 million reads found one or more match locations in the reference genome. The best match in the genome was used as the location for the reads with multiple matches. Poor-quality alignments were defined as those reads aligning over less than 90% of their length, or with more than four substitutions or insertions/deletions with respect to the reference, or reads that matched two locations with nearly equal match score. Ninety-three million reads comprising 7.4-fold coverage (Fig. 1a, see also Supplementary Fig. 2) passed the filtering criteria and were realigned to local genome segments (41 kb each) using Cross_match software. The limitation of a 41-kb genomic fragment placed an upper limit on the sizes of indels that could be detected.

Filtering criteria for SNPs. Mismatch base positions found among the 454-reads were scored using a scaling of the associated error probabilities (*Q*, see 'Scoring system for mismatch base positions' in Supplementary Information). The variant score, S_v, was the sum of scaled 454 *Q* values. Only variant positions with scores $S_v \geq 28$ were considered. In addition, the ratio of variant bases to total coverage was required to be not less than 0.2.

Known SNPs were associated with homopolymer runs of more than 5 bp less than 3% of the time, whereas novel SNPs were associated 33% of the time. Therefore novel variants were removed if they were associated with a homopolymer run of more than 5 bp within 13 bases of the SNP; a novel variant was also removed if it was associated with a 5-bp homopolymer run and $S_v \leq 54$ (see 'Filtering, criteria for SNPs' in Supplementary Information).

Genotyping with sequence data. The sensitivity of detecting heterozygous SNPs by whole-genome shotgun sequencing is limited by the depth-coverage. If the average depth-coverage is *C*, the coverage *k* for each base of the genome approximately follows the Poisson distribution[21]:

$$f(k|C) = \frac{C^k \bullet e^{-C}}{k!}$$

For each heterozygous SNP site of the diploid genome, covered by *K* reads, the number of reads *i* representing one of the two alleles follows the binomial distribution:

$$f(i|K,0.5) = \frac{K!}{i! \bullet (K-i)!} \bullet 0.5^K$$

Given that a heterozygous SNP call required observation of at least two reads from both alleles at the SNP site, the sensitivity of detecting heterozygous SNPs was computed based on the two statistical distributions described above (Supplementary Fig. 4).

Filtering criteria for insertions and deletions. Indels were often associated with short tandem repeats sequences, which caused ambiguity in sequence alignments. Among separately aligned reads, indels in close proximity to one another may represent the same event. Based on our analysis of the spacing between indels, a given deletion or insertion is grouped with the previous one if: (1) the two events are the same type (insertion or deletion); (2) the ratio of the smaller to the larger is greater than or equal to 0.8; (3) the distance between the start coordinates is less than 15 bp for indels larger than 6 bp, or seven times the indel size for indels less than or equal to 6 bp. Furthermore, we required valid indels to be supported by at least two reads and have a ratio of variant to reference greater than 0.25. Errors in measurement of homopolymer length were a source of systematic error for indels one base in length, so they were ignored for this study.

Analysis of 'no-hit' reads. A total of 169,643 contigs were assembled with a total size of 48 Mb and an N50 size (the size at which 50% of the genome is contained within contiguous sequences of this size or greater) of 296 bp. Human repeats in the contigs were masked using RepeatMasker, and contigs containing fewer than 100 contiguous bases of unique sequence were set aside. The remaining110,353 contigs spanned 29 Mb, with an N50 size of 267 bp; 1,294 contigs were longer than 1,000 bp, and the longest was 10,724 bp.

We used Mega BLAST (http://www.ncbi.nlm.nih.gov/blast/megablast.shtml) with the 'expect' parameter set to 10^{-30} to compare these contigs to a human mRNA sequence database of over 40,000 sequences (http://hgdownload.cse.ucsc.edu/goldenPath/hg18/bigZips/mrna.fa.gz). Significant matches were found to 417 mRNA sequences, which had no map coordinates on build 36. We selected from this set 104 mRNA that matched 886 contigs with greater than 96% identity, and which were covered across more than 40% of the mRNA length. The contigs matching these mRNA sequences ranged in size from 296 to 5121 bp. The 104 mRNA sequences were compared with the reference genome using BLAT to confirm they did not have a matching gene on the reference genome. Forty-four were eliminated by this test; 27 had a partial hit, but the genomic match and the contig match did not overlap on the mRNA sequence; 33 of the mRNA sequences had no hit.

Comparative genome hybridization. For the Agilent 244K array, the Human Genome CGH 244K Array (Agilent Technologies, Inc.) contains 238,459 formatted 60-base oligonucleotides, representing a compiled view of the human genome at an average resolution of 9 kb. DNA digestion, labelling and hybridization were performed according to the manufacturer's instructions, with minor modifications. Two separate experiments differed in the reference DNA used for comparison: 'Agilent_1' used a standard reference caucasian male (Kleberg Cytogenetics Laboratory, Baylor College of Medicine); 'Agilent_2' used a caucasian male, NA10851, Coriell Institute for Medical Research. For the Nimblegen HD2 array, we tested a second sample of experimental DNA co-annealed with reference DNA, Coriell Institute for Medical Research number NA10851, to a NimbleChip HD2 Array (NimbleGen Systems, Inc.) in collaboration with NimbleGen Systems. The HD2 array has 2.1 million oligonucleotides, each between 50 and 75 bases, with a reported resolution of about 5 kb. Log_2 ratios were analysed for variation in copy number using NimbleGen SignalMap software.

Affymetrix Gene Chip 500. Duplicate genomic DNA samples from lymphocytes, 250 ng each, were annealed to the Affymetrix 250K *Nsp*I and 250K StyI arrays according to the manufacturer's protocol (see 'Affymetrix Gene Chip 500' in Supplementary Information for further details).

21. Lander, E. S. & Waterman, M. S. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2**, 231–239 (1988).