

# Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells

Brock A. Peters<sup>1\*</sup>, Bahram G. Kermani<sup>1\*</sup>, Andrew B. Sparks<sup>1†</sup>, Oleg Alferov<sup>1</sup>, Peter Hong<sup>1</sup>, Andrei Alexeev<sup>1</sup>, Yuan Jiang<sup>1</sup>, Fredrik Dahl<sup>1†</sup>, Y. Tom Tang<sup>1</sup>, Juergen Haas<sup>1</sup>, Kimberly Robasky<sup>2,3</sup>, Alexander Wait Zaranek<sup>2</sup>, Je-Hyuk Lee<sup>2,4</sup>, Madeleine Price Ball<sup>2</sup>, Joseph E. Peterson<sup>1</sup>, Helena Perazich<sup>1</sup>, George Yeung<sup>1</sup>, Jia Liu<sup>1</sup>, Linsu Chen<sup>1</sup>, Michael I. Kennemer<sup>1</sup>, Kaliprasad Pothuraju<sup>1</sup>, Karel Konvicka<sup>1</sup>, Mike Tsoumpko-Sitnikov<sup>1</sup>, Krishna P. Pant<sup>1</sup>, Jessica C. Ebert<sup>1</sup>, Geoffrey B. Nilsen<sup>1</sup>, Jonathan Baccash<sup>1</sup>, Aaron L. Halpern<sup>1</sup>, George M. Church<sup>2</sup> & Radoje Drmanac<sup>1</sup>

Recent advances in whole-genome sequencing have brought the vision of personal genomics and genomic medicine closer to reality. However, current methods lack clinical accuracy and the ability to describe the context (haplotypes) in which genome variants co-occur in a cost-effective manner. Here we describe a low-cost DNA sequencing and haplotyping process, long fragment read (LFR) technology, which is similar to sequencing long single DNA molecules without cloning or separation of metaphase chromosomes. In this study, ten LFR libraries were made using only ~100 picograms of human DNA per sample. Up to 97% of the heterozygous single nucleotide variants were assembled into long haplotype contigs. Removal of false positive single nucleotide variants not phased by multiple LFR haplotypes resulted in a final genome error rate of 1 in 10 megabases. Cost-effective and accurate genome sequencing and haplotyping from 10–20 human cells, as demonstrated here, will enable comprehensive genetic studies and diverse clinical applications.

The extraordinary advancements made in DNA sequencing technologies over the past few years have led to the elucidation of ~10,000 (refs 1–13) individual human genomes (30× or greater base coverage) from different ethnicities and using different technologies<sup>2–13</sup> and at a fraction of the cost<sup>10</sup> of sequencing the original human reference genome<sup>14,15</sup>. Although this is a monumental achievement, the vast majority of these genomes have excluded a very important element of human genetics. Individual human genomes are diploid in nature, with half of the homologous chromosomes being derived from each parent. The context in which variations occur on each individual chromosome can have profound effects on the expression and regulation of genes and other transcribed regions of the genome<sup>16</sup>. Furthermore, determining whether two potentially detrimental mutations occur within one or both alleles of a gene is of paramount clinical importance.

Almost all recent human genome sequencing has been performed on short read length (<200 base pairs (bp)), highly parallelized systems starting with hundreds of nanograms of DNA. These technologies are excellent at generating large volumes of data quickly and economically. Unfortunately, short reads, often paired with small mate-gap sizes (500 bases–10 kilobases (kb)), eliminate most single nucleotide polymorphism (SNP) phase information beyond a few kilobases<sup>8</sup>. Population-based genotype data has been used to successfully assemble short-read data into long haplotype blocks<sup>3</sup>, but these methods suffer from higher error rates and have difficulty phasing rare variants<sup>17</sup>. Although using pedigree information<sup>18</sup> or combining it with population data provides further phasing power, no combination of these methods is able to resolve *de novo* mutations<sup>17</sup>.

At present, four personal genomes—J. Craig Venter<sup>19</sup>, a Gujarati Indian (HapMap sample NA20847)<sup>11</sup>, and two Europeans (Max Planck One<sup>13</sup> and HapMap Sample NA12878 (ref. 20))—have been

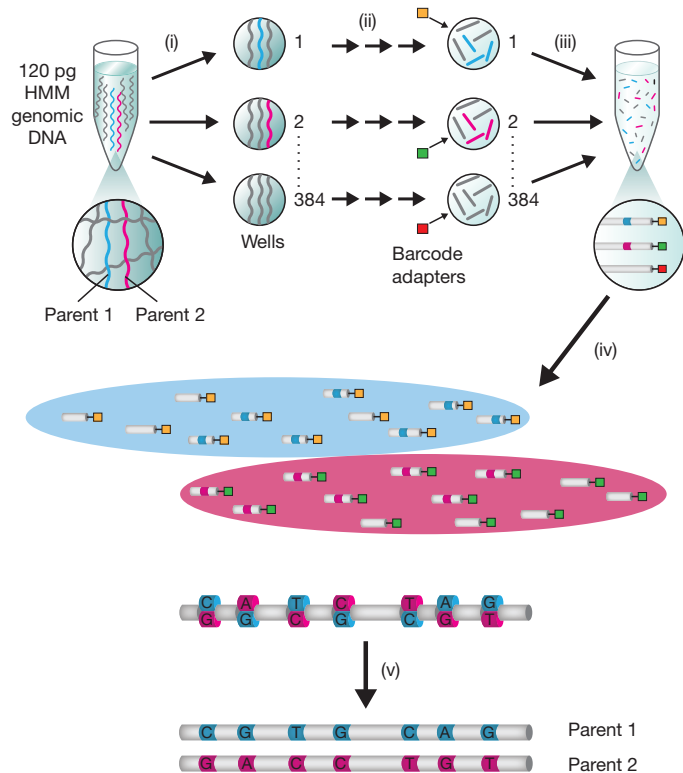
sequenced and assembled as diploid. All have involved cloning long DNA fragments in a process similar to that used for the construction of the human reference genome<sup>14,15</sup>. Although these processes generate long-phased contigs (N50 values (50% of the covered bases are found within contigs longer than this number) of 350 kb<sup>19</sup>, 386 kb<sup>11</sup> and 1 megabase (Mb)<sup>13</sup>, and full-chromosome haplotypes in combination with parental genotypes<sup>20</sup>) they require a large amount of initial DNA, extensive library processing, and are currently too expensive<sup>11</sup> to use in a routine clinical environment. Furthermore, several reports have recently demonstrated whole chromosome haplotyping through direct isolation of metaphase chromosomes<sup>21–24</sup>. These methods have yet to be used for whole-genome sequencing and require preparation and isolation of whole metaphase chromosomes, which can be challenging for some clinical samples. Here we introduce long fragment read (LFR) technology, a process that enables genome sequencing and haplotyping at a clinically relevant cost, quality and scale.

## LFR technology

The LFR approach can generate long-range phased variants because it is conceptually similar to single-molecule sequencing of fragments 10–1,000 kb<sup>25</sup> in length. This is achieved by the stochastic separation of corresponding long parental DNA fragments into physically distinct pools followed by subsequent fragmentation to generate shorter sequencing templates (Fig. 1). The same principles are used in aliquoting fosmid clones<sup>11,13</sup>. As the fraction of the genome in each pool decreases to less than a haploid genome, the statistical likelihood of having a corresponding fragment from both parental chromosomes in the same pool markedly diminishes<sup>25</sup>. For example, 0.1 genome equivalents (300 Mb) per well yields an approximately 10% chance that two fragments will overlap, and a 50% chance that

<sup>1</sup>Complete Genomics, Inc., 2071 Stierlin Court, Mountain View, California 94043, USA. <sup>2</sup>Department of Genetics, Harvard Medical School, Cambridge, Massachusetts 02115, USA. <sup>3</sup>Program in Bioinformatics, Boston University, Boston, Massachusetts 02215, USA. <sup>4</sup>Wyss Institute for Biologically Inspired Engineering, Harvard Medical School, Cambridge, Massachusetts 02115, USA. †Present addresses: Aria Diagnostics, 5945 Optical Court, San Jose, California 95138, USA (A.B.S.); Halo Genomics, Dag Hammarskjolds vag 54A, 751 83 Uppsala, Sweden (F.D.).

\*These authors contributed equally to this work.



**Figure 1 | The LFR technology.** An overview of the LFR technology and controlled random enzymatic fragmenting is shown. (i) First, 100–130 pg of high molecular mass (HMM) DNA is physically separated into 384 distinct wells; (ii) through several steps, all within the same well without intervening purifications, the genomic DNA is amplified, fragmented and ligated to unique barcode adapters; (iii) all 384 wells are combined, purified and introduced into the sequencing platform of Complete Genomics<sup>10</sup>; (iv) mate-paired reads are mapped to the genome using a custom alignment program and barcode sequences are used to group tags into haplotype contigs; and (v) the final result is a diploid genome sequence.

those fragments will be derived from separate parental chromosomes. The end result is a roughly 5% overall chance that a particular well will be uninformative for a given fragment. Likewise, the more individual pools interrogated the greater the number of times a fragment from the maternal and paternal homologues will be analysed in separate pools. The current version of LFR uses a 384-well plate with 10–20% of a haploid genome in each well, yielding a theoretical 19–38 $\times$  physical coverage of both the maternal and paternal alleles of each fragment (see Supplementary Materials and Supplementary Table 1 for an explanation of how this amount of material was selected). This high initial DNA redundancy of 19–38 $\times$  versus recently described strategies using fosmid pools of 3 $\times$  (ref. 11) or 6 $\times$  (ref. 13) ensures complete genome coverage and higher variant calling and phasing accuracy.

To prepare LFR libraries in a high-throughput manner we developed an automated process that performs all LFR-specific steps in the same 384-well plate. First, a highly uniform amplification using a modified,  $\Phi$ 29 polymerase-based, multiple displacement amplification (MDA)<sup>26</sup> is performed to replicate each fragment about 10,000 times. Next, through a process of five enzymatic steps within each well, without intervening purification steps, DNA is fragmented and ligated with barcode adapters. In brief, long DNA molecules are processed to blunt-ended 300–1,500-bp fragments through the new process of controlled random enzymatic fragmenting (Supplementary Methods and Supplementary Figs 2 and 3). Unique 10-base Reed–Solomon<sup>27</sup> error-correcting barcode adapters (Supplementary Fig. 4) are then ligated to fragmented DNA in each well using a high yield, low chimaera formation protocol<sup>10</sup>. Lastly, all 384 wells are combined and

an unsaturated PCR using primers common to the ligated adapters is used to generate sufficient template for massively parallel short-read sequencing platforms (see Supplementary Methods). The addition of the LFR pre-processing steps to the standard library process adds at present about US\$100 to the reagent cost of our genome sequencing (Supplementary Table 2).

### LFR libraries from 10 cells or 100 pg of isolated DNA

As a demonstration of the power of LFR to determine an accurate diploid genome sequence, we generated three libraries of Yoruban female HapMap sample NA19240, six libraries from European HapMap pedigree 1463 (Supplementary Fig. 5), and a single library from Personal Genome Project sample NA20431. Pedigree 1463 and NA19240 have been extensively studied in the HapMap Project<sup>28,29</sup>, the 1,000 Genomes Project<sup>30</sup> and our own efforts (<http://www.completegenomics.com/sequence-data/download-data/>). As a result, highly accurate haplotype information can be generated for these samples based on the redundant sequence data from familial samples. One NA19240 LFR library was made from 10 cells of the corresponding immortalized B-cell line, all other libraries were made from an estimated 100–130 pg (equivalent to 15–20 cells) of denatured high molecular mass genomic DNA (Supplementary Fig. 6 and Supplementary Methods). Libraries were analysed using the sequencing platform of Complete Genomics<sup>10</sup>. Thirty-five-base mate-paired reads were mapped to the reference genome using a custom alignment algorithm<sup>10,31</sup>, yielding on average more than 250 gigabases (Gb) of mapped data with an average genomic coverage of  $>80\times$  (Table 1 and Supplementary Table 3). Analysis of the mapped LFR data shows two distinct characteristics attributable to MDA: slight underrepresentation of GC-rich sequences (Supplementary Fig. 7) and an increase in chimaeric sequences (Supplementary Table 3). In addition, coverage normalized across 100-kb windows was more variable (Supplementary Fig. 8). Nevertheless, almost all genomic regions were covered with sufficient reads (five or more) demonstrating that 10,000-fold MDA amplification by our optimized protocol can be used for comprehensive genome sequencing. Barcodes were used to group mapped reads based on their physical well location within each library, resulting in sparse regions of coverage interspersed between long spans with almost no read coverage (Supplementary Fig. 9). Each of these discrete regions of coverage represents a physical DNA fragment. On average, each well contained 10–20% of a haploid genome (300–600 Mb), in fragments ranging from 10 kb to more than 300 kb in length with N50s of  $\sim 60$  kb (Table 1). Initial fragment coverage was very uniform between chromosomes (Supplementary Fig. 10). As estimated from all detected fragments, the total amounts of DNA used to make the two NA19240 libraries from extracted DNA were  $\sim 62$  pg and 84 pg (equivalent to 9.4 and 12.7 cells, respectively). This is less than the expected 100–130 pg, indicating some lost or undetected DNA or imprecision in DNA quantification. Notably, the 10-cell library seemed to be made from  $\sim 90$  pg (13.6 cells) of DNA, probably due to some of the cells being in S phase during isolation (Table 1).

### LFR haplotyping results

To ensure complete representation of the genome we maximized the input of DNA fragments for a given read coverage and number of aliquots (Supplementary Materials and Supplementary Table 1). Unlike other experimental approaches<sup>11,13,20</sup>, this resulted in low-coverage read data ( $<2\times$ ) for each fragment in each of the  $\sim 40$  wells a fragment is found in. This type of data is not useful for defining haplotypes for each initial fragment and required the development of a new phasing algorithm that statistically combines reads from related fragments found in separate aliquots (Supplementary Methods and Fig. 2). Application of our algorithm to the LFR libraries resulted in the placement of on average 92% of the phasable heterozygous SNPs into long contigs with N50s of  $\sim 1$  Mb and  $\sim 500$  kb for the NA19240 and European samples, respectively (Table 1 and Supplementary

**Table 1 | Comparison of haplotyping performance between different genome assemblies**

Sample	Ethnicity	Number of heterozygous phased SNPs	LFR phasing rate	Haploid fragment coverage (cells)	Fragment size for N50 DNA (kb)	Fragment size for N25 DNA (kb)	DNA bases sequenced (Gb), LFR + STD	N50 contig length (kb)
NA19240 replicate 1	Yoruban	2,386,741	91%	38 (9.4)	64	84	237+176	1,210
NA19240 replicate 2	Yoruban	2,433,621	91%	51 (12.7)	66	96	313+176	1,010
NA19240 10-cell pipeline 2.0	Yoruban	2,369,433	89%	54.3 (13.6)†	80	120	308+176	943
NA19240 replicate 1 high coverage	Yoruban	2,578,903	96%	48 (11.9)	82	116	509+176	1,429
NA19240 replicates 1+2 combined	Yoruban	2,646,352	97%	89 (22.1)	65	90	550+176	1,577
NA19240 replicate 1 LFR only pipeline 2.0	Yoruban	2,031,514	91%	38 (9.4)	64	84	237	1,036
NA19240 replicate 1 high coverage LFR only	Yoruban	2,274,696	95%	48 (11.9)	82	116	509	1,282
NA12877 replicate 1	European	1,831,032	93%	65 (16.3)	74	104	258+218	530
NA12877 replicate 2	European	1,810,540	92%	51 (12.7)	76	106	238+218	535
NA12877 replicates 1+2 combined	European	1,946,089	97%	116 (29)	75	105	496+218	600
NA12885	European	1,850,409	92%	46 (11.6)	72	98	272+221	528
NA12886	European	1,854,360	93%	44 (11)	66	88	293+216	535
NA12891	European	1,825,427	90%*	46 (11.6)	80	112	280+246	545
NA12892	European	1,917,442	93%*	93 (23.3)	94	138	285+213	553
NA12892 LFR only	European	1,720,750	97%*	93 (23.3)	94	138	285	525
NA20431 high coverage	European	1,703,047	84%*	30 (7.4)	94	142	514+189	411

Variant calls for standard (STD) and LFR-assembled libraries were combined and used as loci for phasing except where specified. Two samples were run with the Complete Genomics pipeline 2.0 algorithms, which are expected to call more heterozygous SNPs, the remaining samples were analysed with previous versions (1.7–1.8) of Complete Genomics algorithms. The LFR phasing rate was based on a calculation of parental phased heterozygous SNPs (Supplementary Table 4). N50 calculations are based on the total assembled length of all contigs to the NCBI build 36 (build 37 in the case of NA19240 10-cell and high coverage, and NA20431 high coverage) human reference genome. Haploid fragment coverage is four times greater than the number of cells as a result of all DNA being denatured to single-stranded before being dispersed across a 384-well plate. The insufficient amount of starting DNA explains lower phasing efficiency in the NA20431 genome.

\* For individuals without parental genome data (NA12891, NA12892 and NA20431) the phasing rate was calculated by dividing the number of phased heterozygous SNPs by the number of heterozygous SNPs expected to be real (number of attempted to be phased SNPs – 50,000 expected errors).

† The 10-cell sample was measured by individual well coverage to contain more than 10 cells; this is probably the result of these cells being in various stages of the cell cycle during collection.

Table 4). The large reduction in the N50 contig size for European samples can be explained by many more regions of low heterozygosity (RLHs) found in these genomes (Supplementary Tables 5–7, Supplementary Fig. 11 and Supplementary Materials). Doubling the number of reads to ~160× coverage or combining replicate samples (a total of 768 independent wells), each with ~80× coverage, pushed the phasing rate to ~96% (Table 1). Using only the SNP loci called in the LFR library for phasing resulted mostly in a reduction in the total number of phased SNPs by 5–15% (Table 1 and Supplementary Materials). Importantly, the 1.72 million heterozygous SNPs called and phased by the NA12892 LFR library alone was slightly higher than the number of SNPs phased for a comparable sample using a fosmid approach<sup>13,20</sup> (Table 1). For NA19240, the 10-cell library phased more than 98% of the variants phased by the two libraries made from isolated DNA, demonstrating that LFR can be successful starting from a small number of cells.

### LFR reproducibility and phasing error rate analysis

To test LFR reproducibility we compared haplotype data between the two NA19240 replicate libraries. In general, the libraries were very concordant, with only 64 differences per library in ~2.2 million heterozygous SNPs phased by both libraries (Supplementary Table 8) or 1 of this error type in 44 Mb. LFR was also highly accurate when compared with the conservative but accurate whole-chromosome phasing generated from the parental genomes NA19238 and NA19239 previously sequenced by multiple methods (refs 28, 29 and <http://www.completegenomics.com/sequence-data/download-data/>; Supplementary Table 4). Only ~60 instances in 1.57 million comparable individual loci were found in which LFR phased a variant inconsistent with that of the parental haplotyping (false phasing rate of 0.002% if half of discordances are due to sequencing errors in parental genomes). The LFR data also contained ~135 contigs per library (2.2%), with one or more flipped haplotype blocks (Supplementary Table 8). Extending these analyses to the European replicate libraries of sample NA12877 (Supplementary Table 8) and comparing them with a recent high quality family-based analysis<sup>18</sup> yielded similar results assuming each method contributes half of the observed discordance (Supplementary Table 9). In both NA19240 and NA12877 libraries several contigs had dozens of flipped segments. Most of these contigs were located in RLHs, low read coverage regions, or repetitive regions observed in an unexpectedly large number of wells (for example, subtelomeric or centromeric regions). Most of these

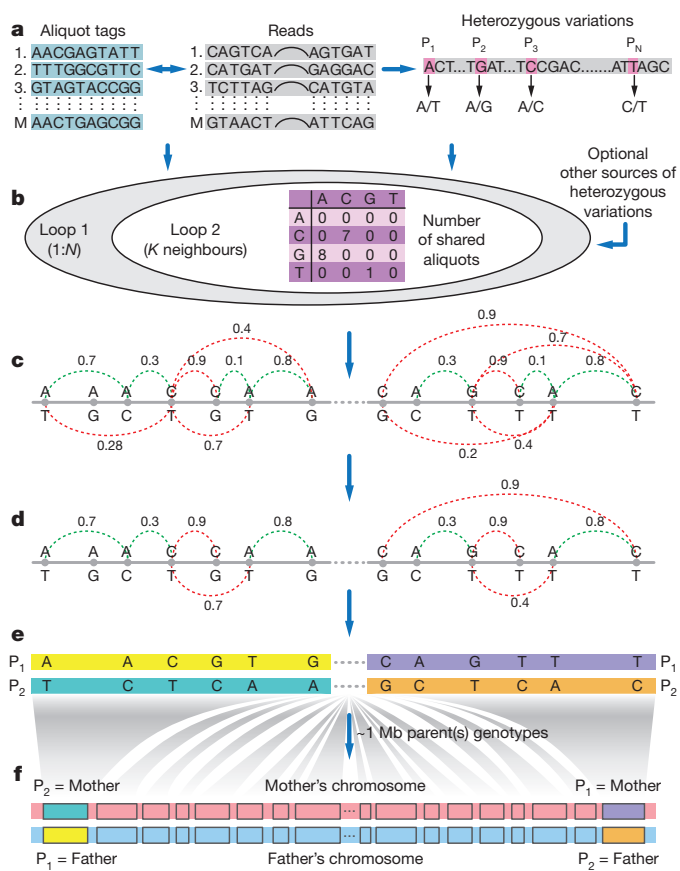
errors can be corrected by forcing the LFR phasing algorithms to end contigs in these regions. Alternatively, these errors can be removed with the simple, low cost addition of standard high density array genotype data (~1 million or greater SNPs) from at least one parent to the LFR assembly. We found that parental genotypes can connect 98% of LFR-phased heterozygous SNPs in full chromosome haplotypes. Furthermore, this data allows haplotypes to be assigned to maternal and paternal lineages; information that is crucial for incorporating parental imprinting in genetic diagnoses in any experimental haplotyping approach. If parental data are unavailable, population genotype data could also be used to connect many of these LFR contigs, although at the cost of increased phasing errors<sup>17</sup>.

### Phasing *de novo* mutations

As a demonstration of the completeness and accuracy of our diploid genome sequencing we assessed phasing of 35 *de novo* mutations recently reported in the genome of NA19240 (ref. 32). Thirty-four of these mutations were called in either the standard genome or one of the LFR libraries. Of those, 32 *de novo* mutations were phased (16 coming from each parent) in at least one of the two replicate LFR libraries (Supplementary Table 10). Not surprisingly, the two non-phased variants reside in RLHs. Of these 32 variants, 21 were phased previously<sup>32</sup> and 18 were consistent with LFR phasing results (M. Hurles, personal communication). The three discordances are probably due to errors in the previous study (M. Hurles, personal communication) confirming LFR accuracy, but not affecting the substantive conclusions of the report.

### Error reduction for accurate sequencing from 10 cells

Substantial error rates (~1 single nucleotide variants (SNV) in 100–1,000 called kilobases) are a common attribute of all current massively parallelized sequencing technologies<sup>2–10,12</sup>. These rates are probably too high for diagnostic use and complicate many studies searching for new mutations. The vast majority of errors are no more likely to occur on the maternal or paternal chromosome. This lack of consistent phasing or presence in only a few aliquots can be exploited by LFR to eliminate these errors from the final assembled haplotypes. To demonstrate this we defined a set of heterozygous SNPs in the NA19240 and NA12877 LFR libraries that were reported with high confidence in each of the individual's parents as matching the human reference genome at both alleles. There were about 44,000 of these heterozygous SNPs in NA19240 and 30,000 in NA12877 that met this



**Figure 2 | LFR haplotyping algorithm.** **a**, Variation extraction. Variations are extracted from the aliquot tagged reads. The 10-base Reed–Solomon codes enable tag recovery by error correction. *M* denotes the number of genomic reads in the set (approximately 8 billion); *N* denotes the number of the candidate heterozygous loci in the genome (~3 million). **b**, Heterozygous SNP pair connectivity evaluation. The matrix of shared aliquots is computed for each heterozygous SNP pair within a certain neighbourhood. Loop 1 is over all the heterozygous SNPs. Loop 2 is over all the heterozygous SNPs on the chromosome that are in the neighbourhood of the heterozygous SNPs in loop 1 (*K*). This neighbourhood is constrained by the expected number of heterozygous SNPs and the expected fragment lengths. **c**, Graph generation. An undirected graph is made, with nodes corresponding to the heterozygous SNPs and the connections corresponding to the orientation and the strength of the best hypothesis for the relationship between those SNPs. The orientation is binary and is shown in the figure with a colour. Red and green depict a flipped and unflipped relationship between heterozygous SNP pairs, respectively. The strength is defined by using fuzzy logic operations on the elements of the shared aliquot matrix. **d**, Graph optimization. The graph is optimized by a minimum spanning tree operation. **e**, Contig generation. Each sub-tree is reduced to a contig by keeping the first heterozygous SNP unchanged, and flipping or not flipping the other heterozygous SNPs on the sub-tree, based on their paths to the first heterozygous SNP. The designation of parent 1 (*P*<sub>1</sub>) and parent 2 (*P*<sub>2</sub>) to each contig is arbitrary. The gaps in the chromosome-wide tree define the boundaries for different sub-trees/contigs on that chromosome. **f**, Optional mapping of LFR contigs to parental chromosomes. Using parental information, a ‘mother’ or ‘father’ label is placed on the *P*<sub>1</sub> and *P*<sub>2</sub> haplotypes of each contig.

criterion (85% sensitivity). By virtue of their nonexistence in the parental genomes these variations are *de novo* mutations, cell-line-specific somatic mutations, or false positive variants. Approximately 1,000–1,500 of these variants were reproducibly phased in each of the two replicate LFR libraries from samples NA19240 and NA12877 (Supplementary Table 11). These numbers are similar to those previously reported for *de novo* and cell-line-specific mutations in NA19240 (ref. 32). The remaining variants are likely to be initial false positives of which only about 500 are phased per library. This represents a 60-fold reduction of the false positive rate in those variations

that are phased. Only ~2,400 of these false variants are present in the standard libraries, of which only ~260 are phased (<1 false positive SNV in 20 Mb; 5,700 haploid megabases per 260 errors). Each LFR library exhibits a 15-fold increase, compared with a genome sequenced by the standard process, in library-specific false positive calls before phasing. Most of these false positive SNVs are likely to have been introduced by MDA; sampling of rare cell-line variants may be responsible for a smaller percentage. Despite making LFR libraries from 100 pg of DNA and introducing a large number of errors through MDA amplification, applying the LFR phasing algorithm described above reduces the overall sequencing error rate to 99.99999% (~600 false heterozygous SNVs per 5.7 Gb), approximately 10-fold lower than the previous published error rates using the same ligation-based sequencing chemistry<sup>18</sup>. These accurate haplotypes allow detection of highly diverged human sequences (Supplementary Materials and Supplementary Table 13) and many other applications.

**Many genes have inactivating variations in both alleles**

To demonstrate how LFR could be used in a diagnostic/prognostic environment we analysed the coding SNP data of all libraries for two or more nonsense, splice site or PolyPhen2 (ref. 33) predicted detrimental missense variations that co-occur in the same gene. Of these, approximately 40 genes were found in each individual that contained at least one detrimental variation in each allele (Table 2). Extending this analysis to variants that disrupt transcription factor-binding sites (TFBS) introduces a further ~100 genes per individual (additional analyses of the effects of TFBS disruption on allele-specific expression can be found in Supplementary Materials and Supplementary Table 12). Owing to the high accuracy of LFR it is unlikely that these variants are a result of sequencing errors and many could have been introduced in the propagation of these cell lines. Furthermore, some of these variants are likely to have little to no effect on the function of these gene products<sup>34</sup> and much more work is required to understand how changes in TFBS affect transcription. A few of these variants were found in unrelated individuals, suggesting that they could be improperly annotated or the result of a systematic mapping or reference error. The genome of NA19240 contained a further ~10 genes predicted to have complete loss of function; this is most likely due to biases introduced by using a European reference genome to annotate an African genome. Nonetheless, these numbers are similar to those found in several recent studies on individual genomes<sup>13,34,35</sup>, and suggest that most generally healthy individuals

**Table 2 | Number of genes with multiple detrimental variations.**

Sample	Ethnicity	Coding only		Coding and TFBS	
		Both alleles	One allele	Both alleles	One allele
NA19240 replicate 1	Yoruban	47	79	182	162
NA19240 replicate 2	Yoruban	55	85	207	174
NA19240 10-cell pipeline 2.0	Yoruban	62	86	197	156
NA19240 replicate 1 high coverage	Yoruban	65	95	235	185
NA19240 replicates 1+2 combined	Yoruban	65	99	241	197
NA12877 replicate 1	European	45	78	144	144
NA12877 replicate 2	European	44	82	146	141
NA12877 replicates 1+2 combined	European	49	96	167	168
NA12885	European	34	79	143	141
NA12886	European	32	101	140	168
NA12891	European	36	69	130	140
NA12892	European	37	65	125	136
NA20431 high coverage	European	36	70	115	127

All phased SNPs were analysed by PolyPhen2 (ref. 33) and a custom splice site detection algorithm (Supplementary Methods) to find variants with a high probability of coding for non-functional proteins. Only variants that were contained within the same contig for each gene were examined. Because LFR contigs are very long (*N*<sub>50</sub> > 500 kb) very few variants were excluded based on this criteria. In each gene 5 kb of the regulatory region upstream of the transcription start site and 1 kb downstream were scanned for SNVs that significantly altered more than 300 TFBS<sup>36,37</sup>. These potentially detrimental variations in TFBS were also phased with coding SNPs to create a more comprehensive list of genes in which the function and/or expression might be altered in these individuals (Supplementary Methods).

probably have a small number of genes, not absolutely required for normal life, which encode ineffective protein products. Further studies are required to understand the meaning of these types of change. Importantly, we have demonstrated that LFR is able to identify genes in which two detrimental variants are found in different alleles without the need for costly verification<sup>34</sup>. This information is crucial for effective clinical interpretation of patient genomes.

## Discussion

In this study we have demonstrated the efficiency of LFR to accurately phase up to 97% of all detected heterozygous SNPs in a genome into long contiguous stretches of DNA (N50s 400–1,500 kb in length). Even LFR libraries phased without candidate heterozygous SNPs from standard libraries, and thus using only 10–20 human cells, are able to phase 91–97% of the available SNPs. In several instances, the LFR libraries used in this paper had less than optimal starting input DNA (NA20431, Table 1). Phasing rate improvements seen by combining two replicate libraries or starting with more DNA (NA12892, Table 1) agree with this conclusion. Furthermore, underrepresentation of GC-rich sequences resulted in less of the genome being called (Supplementary Table 3). Improvements to the MDA process, removal of amplification steps as future single molecule sequencing processes improve, or modifications to how we perform base and variant calling in LFR libraries will help to increase the coverage in these regions (see Supplementary Materials and Supplementary Fig. 12 for a demonstration of how LFR can make calls in low coverage regions). Moreover, as the cost of whole-genome sequencing continues to fall, higher coverage libraries, demonstrated in this paper to markedly improve call rates and phasing, will become more affordable.

A consensus haploid sequence is sufficient for many applications; however, it lacks two very important pieces of data for detecting disease causing variants in personal genomes: phased heterozygous variants and the identification of false positive and negative variant calls. By providing sequence data from both the maternal and paternal chromosomes independently, LFR is able to detect regions in the genome assembly in which only one allele has been covered. Likewise, false positive calls are avoided because LFR independently, in separate aliquots, sequences both the maternal and paternal chromosomes 10–40 times. The result is a statistically low probability that random sequencing or DNA amplification errors would repeatedly occur in several aliquots at the same base position on one parental allele. Thus, LFR allows for the first time, to our knowledge, both accurate and cost-effective sequencing of a genome from a few human cells in spite of the required extensive DNA amplification. Furthermore, by phasing SNPs over hundreds of kilobases (or over entire chromosomes by integrating LFR with routine genotyping of at least one parent), LFR is able to more accurately predict the effects of compound regulatory variants and parental imprinting on allele-specific gene expression and function in various tissue types. Additionally, separation of mate-pair reads by haplotype may also help to detect expanded trinucleotide repeats in diseases such as Huntington's disease, even though LFR does not provide direct length measure of these or similar repeats. Taken together, this provides a highly accurate report about the potential genomic changes that could cause gain or loss of protein function. This kind of information, obtained inexpensively for every patient, will be crucial for clinical use of genomic data. Moreover, successful and affordable diploid sequencing of a human genome starting from ten cells opens the possibility for comprehensive and accurate genetic screening of micro-biopsies from diverse tissue sources such as circulating tumour cells or pre-implantation embryos generated through *in vitro* fertilization.

## METHODS SUMMARY

High molecular mass DNA was purified from cell lines GM12877, GM12878, GM12885, GM12886, GM12891, GM12892 GM19240 and GM20431 (Coriell Institute for Medical Research) using a RecoverEase DNA isolation kit

(Agilent) following the manufacturer's protocol. Individual cells of NA19240 were isolated under  $\times 200$  magnification with a micromanipulator (Eppendorf) and deposited into a 1.5-ml microtube with 10  $\mu$ l of distilled H<sub>2</sub>O. LFR libraries were made as outlined in the text; a more detailed description can be found in the Supplementary Methods. LFR libraries were sequenced, mapped and assembled using the sequencing pipeline of Complete Genomics. Phasing was performed using custom haplotyping algorithms as described in Fig. 2 and in further detail in the Supplementary Methods. Variations adversely affecting protein function or expression were found using several methods. Missense variations were analysed using Polyphen2 (ref. 33). For this study both 'possibly damaging' and 'probably damaging' were considered to be detrimental to protein function, as were all nonsense mutations. Variations determined to adversely affect messenger RNA splicing were found with a custom algorithm based on consensus splice position models from Steve Mount's database (<http://www.life.umd.edu/labs/mount/RNAinfo>). JASPAR models<sup>36,37</sup> were used to extract potential TFBSs from the reference genome with mast (<http://meme.sdsc.edu/meme/mast-intro.html>). Variations falling within these regions were compared with the models to determine what affect they had on transcription factor binding. Genes found to have two or more detrimental mutations were further analysed only if all mutations were found within the same haplotype contig. More detailed descriptions of all methods used in this paper can be found in the Supplementary Methods.

Received 24 January; accepted 15 May 2012.

- Human genome: Genomes by the thousand. *Nature* **467**, 1026–1027 (2010).
- Wheeler, D. A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872–876 (2008).
- Wang, J. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456**, 60–65 (2008).
- Ley, T. J. *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456**, 66–72 (2008).
- Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
- Ahn, S. M. *et al.* The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Res.* **19**, 1622–1629 (2009).
- Kim, J. I. *et al.* A highly annotated whole-genome sequence of a Korean individual. *Nature* **460**, 1011–1015 (2009).
- McKernan, K. J. *et al.* Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res.* **19**, 1527–1541 (2009).
- Pushkarev, D., Neff, N. F. & Quake, S. R. Single-molecule sequencing of an individual human genome. *Nature Biotechnol.* **27**, 847–850 (2009).
- Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78–81 (2010).
- Kitzman, J. O. *et al.* Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nature Biotechnol.* **29**, 59–63 (2011).
- Rothberg, J. M. *et al.* An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**, 348–352 (2011).
- Suk, E. K. *et al.* A comprehensively molecular haplotype-resolved genome of a European individual. *Genome Res.* **21**, 1672–1685 (2011).
- Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
- Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
- Tewhey, R., Bansal, V., Torkamani, A., Topol, E. J. & Schork, N. J. The importance of phase information for human genomics. *Nature Rev. Genet.* **12**, 215–223 (2011).
- Browning, S. R. & Browning, B. L. Haplotype phasing: existing methods and new developments. *Nature Rev. Genet.* **12**, 703–714 (2011).
- Roach, J. C. *et al.* Chromosomal haplotypes by genetic phasing of human families. *Am. J. Hum. Genet.* **89**, 382–397 (2011).
- Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5**, e254 (2007).
- Duitama, J. *et al.* Fosmid-based whole genome haplotyping of a HapMap trio child: evaluation of Single Individual Haplotyping techniques. *Nucleic Acids Res.* **40**, 2041–2053 (2012).
- Zhang, K. *et al.* Long-range polony haplotyping of individual human chromosome molecules. *Nature Genet.* **38**, 382–387 (2006).
- Ma, L. *et al.* Direct determination of molecular haplotypes by chromosome microdissection. *Nature Methods* **7**, 299–301 (2010).
- Fan, H. C., Wang, J., Potanina, A. & Quake, S. R. Whole-genome molecular haplotyping of single cells. *Nature Biotechnol.* **29**, 51–57 (2011).
- Yang, H., Chen, X. & Wong, W. H. Completely phased genome sequencing through chromosome sorting. *Proc. Natl Acad. Sci. USA* **108**, 12–17 (2011).
- Drmanac, R. Nucleic acid analysis by random mixtures of non-overlapping fragments. US patent 7,901 891 (2006).
- Dean, F. B. *et al.* Comprehensive human genome amplification using multiple displacement amplification. *Proc. Natl Acad. Sci. USA* **99**, 5261–5266 (2002).
- Kermani, B. G. & Shannon, K. W. Method and apparatus for quantification of DNA sequencing quality and construction of a characterizable model system using Reed–Solomon codes. US patent PCT/US2010/023083 (2010).

28. The International HapMap Consortium.. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
29. Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
30. The 1000 Genomes Project Consortium.. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
31. Carnevali, P. *et al.* Computational techniques for human genome resequencing using mated gapped reads. *J. Comput. Biol.* **19**, 279–292 (2011).
32. Conrad, D. F. *et al.* Variation in genome-wide mutation rates within and between human families. *Nature Genet.* **43**, 712–714 (2011).
33. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nature Methods* **7**, 248–249 (2010).
34. MacArthur, D. G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823–828 (2012).
35. Lohmueller, K. E. *et al.* Proportionally more deleterious genetic variation in European than in African populations. *Nature* **451**, 994–997 (2008).
36. Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W. W. & Lenhard, B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* **32**, D91–D94 (2004).
37. Bryne, J. C. *et al.* JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.* **36**, D102–D106 (2008).

**Supplementary Information** is linked to the online version of the paper at [www.nature.com/nature](http://www.nature.com/nature).

**Acknowledgements** We would like to acknowledge the continuing contributions and support of all Complete Genomics employees, in particular M. McElwain, D. Bailey,

D. Kruse and J. Turcotte for their help with preparing the manuscript. We also wish to thank W. Chao for his help with Figures 1 and 2. Some of this work was supported by the US Department of Commerce, National Institute of Standards and Technology, Advanced Technology Program, Cooperative Agreement Number 70NANB7H7027 and National Institutes of Health grant P50HG005550. We would like to thank J. Chen for managing the NIST grant.

**Author Contributions** B.A.P., B.G.K., A.B.S. and R.D. conceived the study. B.A.P., B.G.K., R.D., O.A., Y.T.T., J.H., J.C.E., J.B., A.L.H. and G.B.N. performed analyses. B.A.P., A.B.S., P.H., A.A., Y.J., F.D., J.E.P., H.P., G.Y., J.L. and L.C. developed the laboratory processes and generated the LFR libraries. K.K., M.T.-S. and K.P.P. developed the basecaller and parts of the analysis pipeline. M.I.K. formatted, managed and uploaded data to the public archives. K.R., A.W.Z., J.-H.L., M.P.B. and G.M.C. generated and analysed the RNA sequencing data. B.A.P., B.G.K. and R.D. coordinated the study and wrote the paper. All authors contributed to revision and review of the manuscript.

**Author Information** Tagged read data has been deposited with the NCBI short-read archive under accession number SRP012316.1. All sequence data and haplotype information for LFR libraries generated in this study are also available at <http://www.completegenomics.com/LFR>. This paper is distributed under the terms of the Creative Commons Attribution-Non-Commercial-Share Alike licence, and is freely available to all readers at [www.nature.com/nature](http://www.nature.com/nature). Reprints and permissions information is available at [www.nature.com/reprints](http://www.nature.com/reprints). The authors declare competing financial interests: details accompany the full-text HTML version of the paper at [www.nature.com/nature](http://www.nature.com/nature). Readers are welcome to comment on the online version of this article at [www.nature.com/nature](http://www.nature.com/nature). Correspondence and requests for materials should be addressed to B.A.P. ([bpeters@completegenomics.com](mailto:bpeters@completegenomics.com)) or R.D. ([rdmanac@completegenomics.com](mailto:rdmanac@completegenomics.com)).