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

Brock A. Peters1*, Bahram G. Kermani1*, Andrew B. Sparks1†, Oleg Alferov1, Peter Hong1, Andrei Alexeev1, Yuan Jiang1, Fredrik Dahl1, Y. Tom Tang1, Juergen Haas1, Kimberly Robasky1,3, Alexander Wait Zaranek2, Je-Hyuk Lee1,2, Madeleine Price Ball2, Joseph E. Peterson1, Helena Perazich1, George Yeung1, Jia Liu1, Linsu Chen1, Michael I. Kennemer1, Kalirprasad Pothuraju1, Karel Konvicka1, Mike Tsoukko-Sitnikov1, Krishna P. Pant1, Jessica C. Ebert1, Geoffrey B. Nilsen2, Jonathan Baccash1, Aaron L. Halpern1, George M. Church2 & Radoje Drmanac1

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 (30X or greater base coverage) from different ethnicities and using different technologies2–13 and at a fraction of the cost10 of sequencing the original human reference genome14,15. Although these processes have yet to be used for whole-genome sequencing and haplotyping at a clinically relevant cost, quality and scale.

The LFR approach can generate long-range phased variants because it requires 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 kb23 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 clones13,15. 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 diminishes25. 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...

1Complete Genomics, Inc., 2071 Stierlin Court, Mountain View, California 94043, USA. 2Department of Genetics, Harvard Medical School, Cambridge, Massachusetts 02115, USA. 3Program in Bioinformatics, Boston University, Boston, Massachusetts 02215, USA. 4Wyss Institute for Biologically Inspired Engineering, Harvard Medical School, Cambridge, Massachusetts 02115, USA. 5Present 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.). 6These authors contributed equally to this work.
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\(^{10}\); (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.

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\(^ {1,13,20}\), this resulted in lower coverage read data (<2×) for each fragment in each of the ~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). 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 ~1 Mb and ~500 kb for the NA19240 and European samples, respectively (Table 1 and Supplementary Table 2).
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 N50DNA (kb) | Fragment size for N25DNA (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 (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 replicate 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  | 1,851,032                           | 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±211                        | 528                    |
| NA12886                    | European  | 1,845,300                           | 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.

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 previously12 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 technologies10–12. 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
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 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\textsuperscript{31,32,34,35} and many other applications.

| Sample                           | Ethnicity | Coding only | Coding and TFBS |
|----------------------------------|-----------|-------------|-----------------|
| NA19240 replicate 1              | Yoruban   | 47          | 182             |
| NA19240 replicate 2              | Yoruban   | 55          | 207             |
| NA19240 10-cell pipeline         | Yoruban   | 62          | 175             |
| NA19240 replicate 1 high coverage| Yoruban   | 65          | 235             |
| NA19240 replicates 1+2 combined  | European  | 65          | 241             |
| NA12877 replicate 1              | European  | 45          | 144             |
| NA12877 replicate 2              | European  | 44          | 146             |
| NA20431 replicate 1+2 combined   | European  | 49          | 167             |
| NA12885                           | European  | 34          | 143             |
| NA12886                           | European  | 32          | 168             |
| NA12891                           | European  | 36          | 130             |
| NA12892                           | European  | 37          | 125             |
| NA20431 high coverage            | European  | 36          | 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 (N50 ~ 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\textsuperscript{17,18}. 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 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 verification14. 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 (NS50s 400–1,500 kb in length). Even though 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 haplotype 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 self-assembling DNA nanorarrays) sequence resolution can be improved by the detection of two independent allele variant calling.

METHODS SUMMARY

High molecular mass DNA was purified from cell lines GM12877, GM12878, GM12885, GM12886, GM12891, GM12892 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 ×200 magnification with a micromanipulator (Eppendorf) and deposited into a 1.5-mL microtube with 10 μL of distilled H2O. 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 Polypeh2 (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 Mouse’s database (http://www.life.umd.edu/labs/mount/RNAinfo). JASPAR models36,37 were used to extract potential TFBSs from the reference genome with masst (http://meme.sdsc.edu/meme/mast-intro.html). Variations falling with 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 context. More detailed descriptions of all methods used in this paper can be found in the Supplementary Methods.

Received 24 January; accepted 15 May 2012.

1. Human genome: Genomes by the thousand. Nature 467, 1026–1027 (2010).
2. Wheeler, D. A. et al. The complete genome of an individual by massively parallel DNA sequencing. Nature 452, 872–876 (2008).
3. Wang, J. et al. The diploid genome sequence of an Asian individual. Nature 456, 60–65 (2008).
4. Ley, T. J. et al. DNA sequencing of a cytogenetically normal acute myeloid leukemia genome. Nature 456, 66–72 (2008).
5. Bentley, D. R. et al. Accurate whole genome sequence sequencing using reversible terminator chemistry. Nature 456, 53–59 (2008).
6. Ahn, S. M. et al. The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. Genome Res. 18, 1622–1629 (2008).
7. Kim, J. I. et al. A highly annotated whole-genome sequence of a Korean individual. Nature 460, 1011–1015 (2009).
8. McKernan, K. J. et al. Sequence and structural variation in a human genome uncovers by short-read, massively parallel sequencing using two-base encoding. Genome Res. 19, 1527–1541 (2009).
9. Pushkarev, D., Neff, N. F. & Quake, S. R. Single-molecule sequencing of an individual human genome. Nature Biotechnol. 27, 847–850 (2009).
10. Drmacan, R. et al. Human genome sequencing using unchained base reads on self-assembling DNA nanorarrays. Science 327, 78–81 (2010).
11. Kitzman, J. O. et al. Haplotyp-resolved genome sequencing of a Gujarati Indian individual. Nature Biotechnol. 29, 59–63 (2011).
12. Rothenberg, J. M. et al. An integrated semiconductor device enabling non-optical genome sequencing. Nature 475, 346–352 (2011).
13. Suk, E. K. et al. A comprehensively molecular haplotype-resolved genome of a European individual. Genome Res. 21, 1672–1685 (2011).
14. Venter, J. C. et al. The sequence of the human genome. Science 291, 1304–1351 (2001).
15. Lander, E. S. et al. Initial sequencing and analysis of the human genome. Nature 409, 860–921 (2001).
16. 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).
17. Browning, S. R. & Browning, B. L. Haplotyping: existing methods and new developments. Nature Rev. Genet. 12, 703–714 (2011).
18. Roach, J. C. et al. Chromosomal haplotypes by genetic phasing of human families. Am. J. Hum. Genet. 89, 382–397 (2011).
19. Levy, S. et al. The diploid genome sequence of an individual human, PLoS Biol. 5, e254 (2007).
20. 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).
21. Zhang, K. et al. Long-range polony haplotyping of individual human chromosome molecules. Nature Genet. 38, 382–387 (2006).
22. Ma, L. et al. Direct determination of molecular haplotypes by chromosome microdissection. Nature Methods 7, 299–301 (2010).
23. Fan, H. C., Wang, J., Potanina, A. & Quake, S. R. Whole-genome molecular haplotyping of single cells. Nature Biotechnol. 29, 51–57 (2011).
24. Yang, H., Chen, X. & Wong, W. H. Completely phased genome sequencing through chromosome sorting. Proc. Natl Acad. Sci. USA 108, 12–17 (2011).
25. Drmanac, R. Nucleic acid analysis by random mixtures of non-overlapping fragments. US patent 7,901,891 (2006).
26. Dean, F. B. et al. Comprehensive human genome amplification using multiple displacement amplification. Proc. Natl Acad. Sci. USA 99, 5261–5266 (2002).
27. 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 of 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.

**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. Reprints and permissions information is available at 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. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to B.A.P. (bpeters@completegenomics.com) or R.D. (rdrmanac@completegenomics.com).