nature biotechnology

Haplotype-resolved genome sequencing of a Gujarati Indian individual

Jacob O Kitzman[1], Alexandra P MacKenzie[1], Andrew Adey[1], Joseph B Hiatt[1], Rupali P Patwardhan[1], Peter H Sudmant[1], Sarah B Ng[1], Can Alkan[1,2], Ruolan Qiu[1], Evan E Eichler[1,2] & Jay Shendure[1]

Haplotype information is essential to the complete description and interpretation of genomes[1], genetic diversity[2] and genetic ancestry[3]. Although individual human genome sequencing is increasingly routine[4], nearly all such genomes are unresolved with respect to haplotype. Here we combine the throughput of massively parallel sequencing[5] with the contiguity information provided by large-insert cloning[6] to experimentally determine the haplotype-resolved genome of a South Asian individual. A single fosmid library was split into a modest number of pools, each providing ~3% physical coverage of the diploid genome. Sequencing of each pool yielded reads overwhelmingly derived from only one homologous chromosome at any given location. These data were combined with whole-genome shotgun sequence to directly phase 94% of ascertained heterozygous single nucleotide polymorphisms (SNPs) into long haplotype blocks (N50 of 386 kilobases (kbp)). This method also facilitates the analysis of structural variation, for example, to anchor novel insertions[7,8] to specific locations and haplotypes.

The high quality of the human reference genome derives from the hierarchical sequencing of large-insert clones, such that the assembly corresponding to each clone represents a single haplotype[9]. One of the first 'personal genomes' exploited clone-based mate pairing and long, accurate Sanger reads to resolve variants into haplotype blocks (N50 of 350 kbp; that is, 50% of resolved sequence is within blocks of at least 350 kbp)[1]. Although new technologies[5] have subsequently enabled >1,000-fold reduction in genome sequencing costs, the short read-lengths and paucity of contiguity information are such that it remains challenging to determine haplotypes at a genome-wide scale. Genomic phase, the assignment of alleles to homologous chromosomes, was determined for SNPs using mate-paired reads on the SOLiD (sequencing by oligonucleotide ligation and detection) platform[10] for an individual genome, but only 43% of heterozygous variants were phased, and nearly all in blocks no greater than the insert size, that is, <3.5 kbp[10]. Experimental limitations on the size and complexity of mate-pair libraries based on *in vitro* circularization[11] make it difficult to improve upon this approach.

An alternative is to infer haplotypes from population-based linkage disequilibrium data or from pedigree analysis. For example, haplotypes were successfully inferred in the YH (YanHuang) genome for variants at which phased CHB/JPT HapMap data were available (CHB, Han Chinese from Beijing, China; JPT, Japanese from Tokyo, Japan)[12]. The genomes of a family of four have been sequenced and these relationships used to infer inheritance blocks[13]. Although they can be successful, inferential methods have limitations. Statistical phasing, whether based on genotyping[2] or sequencing[14], performs poorly when linkage disequilibrium is not high, and for rare variants. Phasing by pedigree analysis requires genome sequencing of many related individuals, increasing costs and limiting practical application.

We describe a cost-effective method for determining long-range haplotypes at a genome-wide scale by massively parallel sequencing of complex, haploid subsets of an individual genome (**Fig. 1**). We apply this method to the first reported whole-genome sequencing of a human of South Asian ancestry. The Indian subcontinent is home to myriad culturally and genetically diverse groups with distinct population histories[15]. We selected a female from the HapMap panel of 'Gujarati Indians in Houston' (GIH; NA20847) for sequencing. Notably, the imputation of genotypes for GIH was the least effective of all non-African populations in HapMap[2].

Genomic DNA from NA20847 was used to construct a single, complex fosmid library, containing clones packaged in phage for infecting *Escherichia coli* cells (>2×10^6 clones with ~37 kbp inserts) (**Fig. 1a** and **Supplementary Methods**). We then split a portion of this library to 115 pools, at a density such that each pool contained ~5,000 independent clones. Each pool was expanded by either scraping a single plate of infected cells and inoculating outgrowth culture, or by direct liquid outgrowth after infection. However, at no point does this method require the isolation of individual colonies. We next constructed 115 barcoded, shotgun sequencing libraries from fosmid DNA isolated from each of the 115 pools[16]. Libraries indexed with barcodes were combined and sequenced (Illumina GAIIx; PE76 or PE101 reads) to a mean 2.4× depth per haploid clone (**Fig. 1b**).

Because each pool captures an essentially random ~3% of the 6-gigabase (Gb) diploid genome (that is, ~5,000 fosmids × ~37 kbp inserts) sequence reads from each pool are overwhelmingly (99.1%) derived from only one homologous chromosome or the other at any single location. Upon mapping reads from each pool to the reference assembly, the approximate boundaries of 538,009 individual clones (37.2 ± 4.7 kbp) were identified by read depth (4,678 ± 1,229 clones per pool). Coverage was uniform across the genome (98.6% covered

[1]Department of Genome Sciences, University of Washington School of Medicine, Seattle, Washington, USA. [2]Howard Hughes Medical Institute, Seattle, Washington, USA. Correspondence should be addressed to J.S. (shendure@uw.edu) or J.O.K. (kitz@uw.edu).

Received 26 October; accepted 29 November; published online 19 December 2010; doi:10.1038/nbt.1740

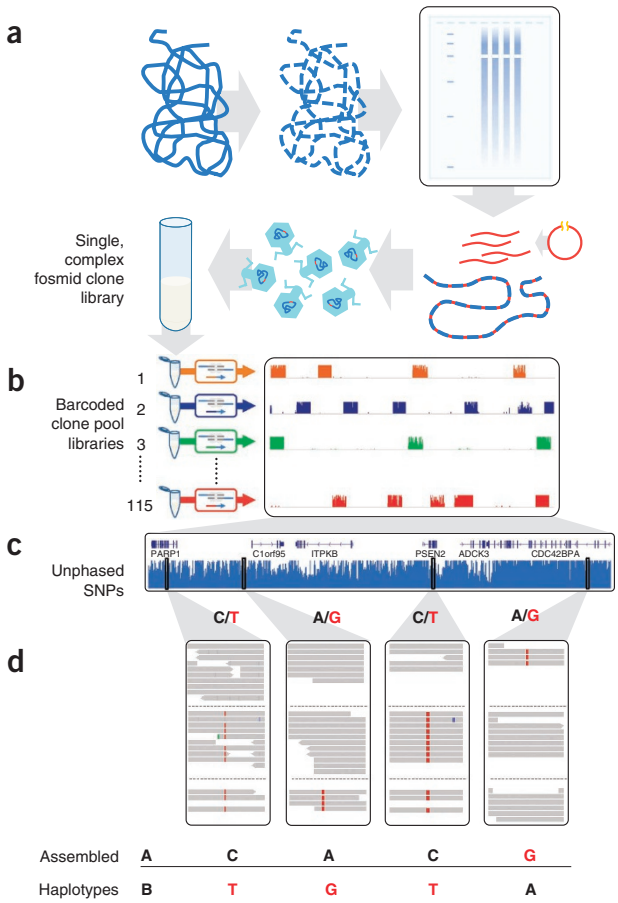

Figure 1 Haplotype-resolved genome sequencing. (**a**,**b**) A single, highly complex fosmid library was constructed (**a**) and split into 115 pools (**b**), each representing ~3% physical coverage of the diploid human genome. Barcoded shotgun libraries from each pool were constructed, then combined and sequenced. As expected, reads from each library map to ~5,000 × ~37 kbp blocks, minimally redundant within each library. (**c**) Whole-genome shotgun sequencing of the same individual generated unphased variant calls. (**d**) Unphased variant calls were combined with haploid genotype calls to assemble haplotype blocks using a maximum parsimony approach[19] (reference allele in black, nonreference allele in red).

by one or more clones) and within each pool (82% of clones with mean read depth within a tenfold range) (**Supplementary Fig. 1**).

For unphased variation discovery, we performed conventional whole-genome resequencing to 15× depth (Illumina HiSeq; PE50) (**Supplementary Table 1** and **Supplementary Fig. 2**). After alignment to the reference, we called 3.3×10^6 SNPs and 3.4×10^5 short indels[17,18] (**Fig. 1c**). Nonreference sensitivity for SNPs was 91%, that is, HapMap variant genotypes at positions also called in our data, and genotype concordance to high-quality HapMap 3 genotypes[2] at called positions was 99.2% (n = 1,436,495). Other bulk statistics, including the heterozygous-to-homozygous call ratio, the fraction of called variants previously ascertained in the NCBI SNP database (dbSNP), the transition-to-transversion ratio, and the numbers and classes of coding variants, were consistent with expectations based on previously sequenced non-African genomes (**Supplementary Table 2**).

Several methods have been described for assembling haplotypes from sequence data[1,19–21]. We adopted a maximum parsimony approach[19] to combine the unphased variants from shotgun whole-genome sequencing with haploid genotype calls from sequencing of the 115 pools (**Fig. 1d**). The resulting assembly incorporated 94% of ascertained heterozygous SNPs into haplotype-resolved blocks, with an N90 of 89 kbp, an N50 of 386 kbp and an N10 of 1 megabase (Mbp) (**Fig. 2a**). Sixty-two percent of genes were fully encompassed by single blocks, and 73% were covered for over half their length.

To evaluate accuracy, we compared our haplotype assembly with HapMap phase predictions for NA20847 (**Fig. 2b**)[2]. For pairs of SNPs in exceptionally high-linkage disequilibrium ($D' > 0.90$ among GIH), we observed nearly perfect concordance (>99.7%). Because NA20847 was not part of a trio, HapMap predictions rely upon linkage disequilibrium between alleles to predict phase from genotypes. Correspondingly, concordance was reduced to ~71% when $D' < 0.10$, which is the case for most (66%) pairwise SNP combinations. Concordance is also reduced when one or both alleles in the pair is rare in GIH (**Fig. 2c**). Note that our haplotype assembly is experimental and specific to an individual, and therefore completely independent of population-based phenomena such as linkage disequilibrium and allele frequency. Consequently, these trends likely reflect errors in HapMap phasing[1].

South Asian history includes admixture between two ancestral groups, one genetically close to Europeans (ANI) and another more highly diverged from well-ascertained populations (ASI)[15]. Furthermore, principal components analysis revealed a distinct subgroup of Indian populations in general and GIH in particular, including NA20847, that may harbor substantial genetic ancestry from a third population distinct from ANI and ASI[15]. We compared haplotype blocks for this individual to HapMap allele frequencies in the GIH and CEPH European (CEU) populations to distinguish 'GIH-like' from 'CEU-like' haplotypes. Notably, novel SNPs are markedly enriched on the most GIH-like haplotypes (**Fig. 3**). We also scored haplotype blocks against allele frequencies from the 1000 Genomes Project[14] (**Supplementary Fig. 3**). Haplotypes that least resembled all three populations in that study (CEU, CHB/JPT and Yoruba) were also markedly enriched for novel SNPs. We propose that GIH-like blocks and other well-differentiated haplotypes may be derived from more poorly ascertained ancestral

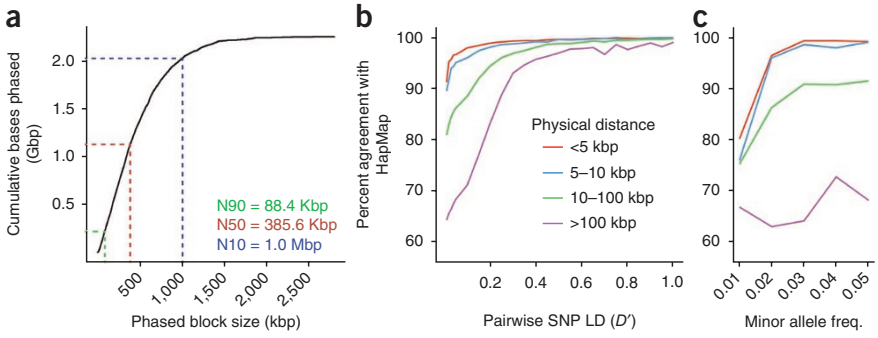

Figure 2 Haplotype assembly results. (**a**) Size distribution of blocks within the haplotype assembly up to a maximum block size of 2.79 Mbp. Half of the assembly comprised blocks longer than 386 kbp (N50). (**b**) Comparison of experimental phasing with HapMap population-based inference[2] for NA20847, with agreement of pairwise haplotype predictions as a function of physical distance and linkage disequilibrium. (**c**) Agreement of pairwise haplotype predictions as a function of physical distance and minor allele frequency (defined as the lower allele frequency of the pair in GIH). Key is the same as for **b**.

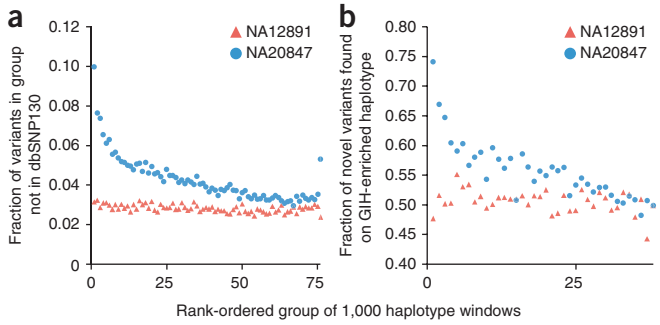

Figure 3 Enrichment of novel variants on 'GIH-like' haplotypes. (**a**) Haplotypes were scored and rank ordered within sliding windows of 20 HapMap variants[2] for greater similarity to GIH or CEU on the basis of population allele frequencies (left on *x* axis: more similar to GIH). Plotted is the fraction of novel variants (not in dbSNP v130) in rank-ordered groups of haplotype windows, demonstrating that the most 'GIH-like' haplotype windows are enriched for novel variants. Values from trio-phased[14] CEU individual NA12891 are shown for comparison (red). (**b**) Scores calculated in **a** for haplotype windows were compared between homologous chromosomes, and haplotypes were ranked based on the extent to which they scored as 'GIH-like' relative to their homolog. Plotted is the fraction of novel variants found on the more 'GIH-like' haplotype in rank-ordered groups of homologous haplotype windows. As above, the analysis was also performed for individual NA12891 using the rank ordering from individual NA20847. Haplotype blocks that are most differentiated relative to their homolog (higher ranked) with respect to GIH versus CEU similarity are enriched for novel variants relative to their homolog, consistent with the pattern observed in **a**.

populations, and therefore enriched for novel variants. Such haplotypes may represent a valuable source of information about human history on the South Asian subcontinent.

A substantial fraction of the human genome consists of gene-rich segmental duplications and otherwise structurally complex regions that continue to defy accurate diploid consensus assembly within individual genomes. We sought to evaluate whether haplotype-resolved sequencing is useful for the fine-mapping and haplotype-assignment of deletions, inversions and novel contigs.

We used shotgun read depth[22], discordant pairing in shotgun data[23] and array-based SNP calls[2] to estimate copy number and detect 58 deletions (>8 kbp), 15 of which were flanked by segmental duplications. Of these, 48 deletions (83%) were unambiguously confirmed by sequenced fosmid clones spanning the breakpoints, providing fine-scale resolution and confirming 30 as hemizygous (**Fig. 4a** and **Supplementary Table 3**). Heterozygous variants in flanking clones allowed for unambiguous incorporation of these deletions into haplotype-resolved blocks.

Inversions are challenging to detect because they are copy-number neutral and frequently mediated by repetitive sequences. As even fosmid end-sequencing tends to overcall inversions[6], the added information from interrogating full ~37-kbp inserts may be useful for discriminating true inversions from false positives (**Supplementary Fig. 4**). Indeed, we observed a number of unambiguous inversions by means of breakpoint-spanning clones (**Supplementary Fig. 5**). However, larger clones (>100 kbp) may be required to span the large duplication blocks where inversion breakpoints typically map[6]. NA20847 is heterozygous for the inversion-containing H2 haplotype at the *MAPT* locus (17q21) (**Supplementary Fig. 6**). Of note, we properly phased all 287 SNPs that tag the H2 haplotype across a 588-kbp span[24].

We also detected common human sequences unrepresented in the reference, that is, the 'pan-genome' (**Supplementary Table 4**)[7,8]. Of 16,904 contigs (total 12.8 Mbp) reported by two recent studies[7,8], we identified 8,993 in NA20847. We exploited the contiguity of fosmids to anchor ~30% of these (**Fig. 4b**), with 73% agreement (±50 kbp) with a previously anchored subset[8]. *De novo* assembly of remaining unmapped reads yielded 2,242 additional contigs after filtering, of which we anchored 396. To validate anchoring accuracy, we simulated novel insertions by deleting 600 intervals (250 bp–10 kbp) *in silico* from the reference and remapping reads to the modified reference. Unmapped reads were *de novo* assembled into 5,435 contigs that covered ~61% of simulated insertions. Of these, we predicted anchoring locations for 2,184 with an accuracy of 87%, with the remaining contigs unassigned because of limited clone coverage. The sensitivity and specificity with which novel contigs can be anchored by this approach is likely to improve with increased clone and shotgun coverage.

We recently demonstrated exome sequencing as a strategy for identifying causal variants in Mendelian disorders[25], for example, implicating compound heterozygote variants in *DHODH* in Miller syndrome[26]. In such studies, phasing reduces the number of candidate genes consistent with a recessive, compound heterozygous model[13]. For example, in this Gujarati Indian individual, unphased variant data included 44 genes consistent with compound heterozygosity (that is, two or more heterozygous, novel, nonsynonymous or splice-site variants that altered the same gene). But after phase was

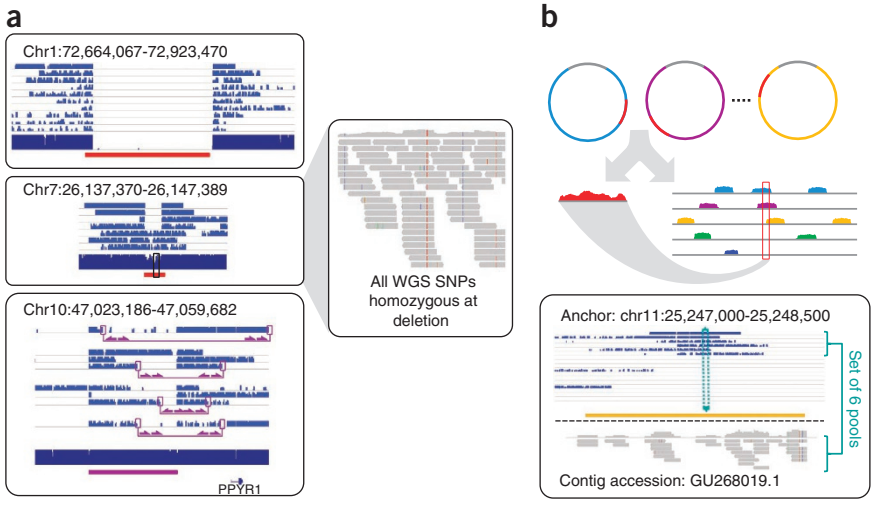

Figure 4 Insertion anchoring and structural variation detection. (**a**) Homozygous deletion (top), hemizygous deletion (middle) and inversion (bottom) with fosmid clone support. Deletion calls were made using read depth and paired-read discordance. Inversions were called by paired-read discordance. SNPs within hemizygous deletions appear as stretches of hemizygosity by whole-genome shotgun sequencing. Purple connections indicate the additional support of strand discordance of read pairs spanning genomic DNA and the vector backbone. (**b**) Novel contigs not present in the reference assembly (red) but detected among clone pool–derived reads (light blue, purple, yellow) are anchored by searching for positions in the reference common to those pools but missing from most or all other pools. This approach anchors 1,733 recently reported insertion sequences[7,8] including contig GU268019.

taken into account, only ten were validated as *trans* heterozygous, with the remainder having both variants on the same haplotype.

This method requires significantly greater expertise and sample preparation than the haplotype-blind shotgun sequencing of an individual genome—specifically, the construction of a single fosmid library and >100 *in vitro* shotgun libraries, as compared with constructing one or a few *in vitro* shotgun libraries. A detailed consideration of the added effort and cost are provided in **Supplementary Table 5**. In summary, sample preparation can be completed in <2 weeks by a single technician at a cost (~$4,000) that is much greater than that of preparing a single shotgun library, but low relative to the overall cost of whole-genome sequencing. We use an unconventional method based on *in vitro* transposition[16] to significantly reduce the time and effort for producing >100 shotgun libraries. Current costs are primarily driven by commercial reagents for fosmid and shotgun library construction, and may therefore be amenable to optimization[16]. Furthermore, most steps are compatible with manual scaling and/or automation.

We also note that the total bases sequenced here (~87 Gb shotgun, ~110 Gb clone-based) is only modestly higher than for other individual human genomes sequenced to date. To estimate the minimal amount of clone sequencing required, we subsampled our data for either the number of independent clones or the depth of clone library sequencing (**Supplementary Fig. 7**). The primary effect was a reduction in the length of assembled haplotype blocks, rather than any decay in accuracy. For example, at 80% of clones and 60% of sequencing depth (which is 48% as much clone-based sequencing), the N50 dropped from 386 kbp to 238 kbp. However, most ascertained heterozygous variants remained phased (85.4%), and phasing remained highly concordant with HapMap (>99% at $D' > 0.9$). Other optimizations, for example, switching from plate-scraping to direct liquid outgrowth to improve clone uniformity (**Supplementary Fig. 1**), may further reduce sequencing requirements.

Haplotypes are essential to the information content that defines a diploid human genome, but have heretofore been intractable to genome-wide, experimental determination in the context of massively parallel sequencing. We anticipate that haplotype-resolved genome sequencing will be valuable in a broad range of scenarios, including the following. (i) Population genetics. Haplotype-resolved genome sequencing eliminates the need for population or pedigree-based haplotype inference. This will be most useful in populations that are poorly ascertained (e.g., South Asians) or have low linkage disequilibrium (e.g., Africans), and more generally for rare variants. (ii) Genetic anthropology. For example, the availability of the haplotype-resolved reference and Venter genomes was critical to the observation of a Neanderthal contribution to some modern humans[3]. (iii) Medical genetics of rare and common phenotypes. Haplotype information can facilitate the analysis of recessive Mendelian disorders[13], the determination of the parent of origin for *de novo* mutations, and the study of complex interactions among multiple SNPs[27]. (iv) Structural variation in both germline and cancer genomes. Our approach is more comprehensive than long-insert mate-pairing (whether by fosmids[6] or *in vitro* circularization[28]), as these methods determine the ends of large molecules but are blind to their internal contents. Also, the intermediate level of partitioning provided by fosmids may be more useful than whole chromosome amplification[29], as many germline and somatic structural events are intrachromosomal. (v) Allele-specific phenomena. Haplotype information may be essential for understanding the genetic basis of phenomena such as allele-specific expression and methylation[30]. (vi) *De novo* genome assembly. Massively parallel sequencing of highly complex pools of minimally redundant haploid clones may facilitate the high-quality *de novo* assembly of new genomes, an area that continues to be a major challenge for the genomics field despite the falling costs of DNA sequencing[11].

METHODS

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

Accession codes. Short read sequence data have been deposited at the NCBI Sequence Read Archive (SRA) under accession no. SRA026360. Assembled haplotype blocks and novel contigs are available from: http://krishna.gs.washington.edu/indianGenome/.

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

ACKNOWLEDGMENTS

We thank C. Lee and M. Malig for technical assistance, J. Akey, T. O'Connor and P. Green for helpful discussions, D. Reich for ancestry information on NA20847, the U.W. Genome Sciences Genomics Resource Center (GS-GRC) for sequencing and the 1000 Genomes Project for early data release. This work was supported by National Institutes of Health grants AG039173 (J.B.H.) and HG002385 (E.E.E.), a National Science Foundation Graduate Research Fellowship (J.O.K.), a Natural Sciences and Engineering Research Council of Canada Fellowship (P.H.S.) and a fellowship from the Achievement Rewards for College Scientists Foundation (J.B.H.). E.E.E. is an investigator of the Howard Hughes Medical Institute.

AUTHOR CONTRIBUTIONS

The project was conceived and experiments planned by J.O.K., E.E.E. and J.S. J.O.K., A.P.M. and R.Q. carried out all experiments. J.O.K., A.A., J.B.H., R.P.P., P.H.S., S.B.N. and C.A. performed data analysis. J.O.K., A.P.M., A.A., J.B.H., R.P.P. and J.S. wrote the manuscript, and all authors reviewed it. All aspects of the study were supervised by J.S.

COMPETING FINANCIAL INTERESTS

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

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

1. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5**, e254 (2007).
2. International HapMap Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
3. Green, R.E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).
4. Anonymous. Human genome: Genomes by the thousand. *Nature* **467**, 1026–1027 (2010).
5. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat. Biotechnol.* **26**, 1135–1145 (2008).
6. Kidd, J.M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64 (2008).
7. Li, R. *et al.* Building the sequence map of the human pan-genome. *Nat. Biotechnol.* **28**, 57–63 (2010).
8. Kidd, J.M. *et al.* Characterization of missing human genome sequences and copy-number polymorphic insertions. *Nat. Methods* **7**, 365–371 (2010).
9. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
10. 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).
11. Schatz, M.C., Delcher, A.L. & Salzberg, S.L. Assembly of large genomes using second-generation sequencing. *Genome Res.* **20**, 1165–1173 (2010).
12. Wang, J. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456**, 60–65 (2008).
13. Roach, J.C. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**, 636–639 (2010).
14. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
15. Reich, D., Thangaraj, K., Patterson, N., Price, A.L. & Singh, L. Reconstructing Indian population history. *Nature* **461**, 489–494 (2009).
16. Adey, A. *et al.* Rapid, low-input, low-bias construction of shotgun fragment libraries by high density *in vitro* transposition. *Genome Biol.* **11**, R119 (2010).
17. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

18. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
19. Bansal, V. & Bafna, V. HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. *Bioinformatics* **24**, i153–i159 (2008).
20. Kim, J.H., Waterman, M.S. & Li, L.M. Diploid genome reconstruction of Ciona intestinalis and comparative analysis with Ciona savignyi. *Genome Res.* **17**, 1101–1110 (2007).
21. Bansal, V., Halpern, A.L., Axelrod, N. & Bafna, V. An MCMC algorithm for haplotype assembly from whole-genome sequence data. *Genome Res.* **18**, 1336–1346 (2008).
22. Alkan, C. *et al.* Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat. Genet.* **41**, 1061–1067 (2009).
23. Hormozdiari, F. *et al.* Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics* **26**, i350–i357 (2010).
24. Zody, M.C. *et al.* Evolutionary toggling of the MAPT 17q21.31 inversion region. *Nat. Genet.* **40**, 1076–1083 (2008).
25. Ng, S.B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272–276 (2009).
26. Ng, S.B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.* **42**, 30–35 (2010).
27. Drysdale, C.M. *et al.* Complex promoter and coding region beta 2-adrenergic receptor haplotypes alter receptor expression and predict in vivo responsiveness. *Proc. Natl. Acad. Sci. USA* **97**, 10483–10488 (2000).
28. Korbel, J.O. *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**, 420–426 (2007).
29. Ma, L. *et al.* Direct determination of molecular haplotypes by chromosome microdissection. *Nat. Methods* **7**, 299–301 (2010).
30. Tycko, B. Allele-specific DNA methylation: beyond imprinting. *Hum. Mol. Genet.* **19**, R210–R220 (2010).

ONLINE METHODS

Fosmid library pool construction. High molecular weight genomic DNA (HMW gDNA) was extracted from HapMap lymphoblastoid cell line GM20847 (Coriell) using the Gentra Puregene kit (Qiagen). A single, complex fosmid library (>2 × 10^6 clones) was created using the CopyControl pCC1Fos Fosmid Library Construction kit (Epicentre), as previously described[31]. After bulk infection, the library was split into 115 pools of ~5,000 clones each. Each pool was then individually expanded, either by scraping plates of infected cells and inoculating outgrowth culture, or by direct liquid outgrowth after infection. Clone DNA was extracted from each pool by alkaline lysis miniprep.

Massively parallel sequencing. Illumina-compatible shotgun sequencing libraries were prepared from each fosmid clone pool DNA and HMW gDNA using the Nextera DNA Sample Prep Kit (Epicentre), as described[16]. For each fosmid pool library, a 9-bp barcoded adaptor was added during PCR amplification[16]. Pool-derived libraries were combined before sequencing (PE76 or PE101 reads, plus index read, on an Illumina GA2x), and the index read was used to deconvolve the original clone pools from the combined reads. For unphased variant discovery, a single whole-genome shotgun library was sequenced across seven lanes (PE50 reads on an Illumina HiSeq).

Read mapping and variant discovery. Basecalling was performed with Illumina RTA v1.8 software. The resulting reads were aligned to the reference assembly (NCBI release GRCh37, UCSC release hg19) using BWA v0.5.8a[17]. The Genome Analysis Toolkit (GATK)[18] was used to recalibrate base quality scores, realign reads surrounding putative and known indels, and call single-nucleotide and indel variants from the whole-genome shotgun data. Quality filters were applied based on coverage, base and mapping quality score, and allelic and strand bias. Copy number genotypes were estimated genomewide by (G+C)-corrected read depth, as previously described[32]. Deletions >8 kbp were identified by intersecting regions of predicted copy less than 2 with split-read calls[23] and published SNP array-based calls[2] and requiring calls by two of the three methods.

Haplotype assembly. Clone coordinates were identified within each pool by searching for intervals of length 25–45 kbp with coverage significantly above background. Heterozygous SNP positions ascertained during whole-genome shotgun sequencing were regenotyped within each haploid clone pool. Clones with an excess of heterozygous positions, likely representing overlapping clones drawn from different haplotypes, were discarded. Haplotype blocks were created from overlapping clones using a custom reimplementation of HAPCUT[19], a parsimony maximization-based haplotype assembly algorithm. The effects of lower sequence coverage upon haplotype assembly accuracy and block length were simulated by leaving out a random subset of clones and/or reads.

Haplotype ancestry analysis. Phased blocks were divided into sliding windows of variants from HapMap[2] (20 SNPs/window) or the 1000 Genomes Project[14] (200 SNPs/window). For the HapMap-based comparison, similarity to GIH and CEU populations was scored for both haplotypes of NA20847 at every window based on the frequencies of alleles in NA20847 among GIH and CEU. Haplotype windows were then rank-ordered by the difference in similarity scores, such that haplotypes with high-frequency alleles among GIH but not CEU were more highly ranked. The fraction of all detected novel variants (not in dbSNP release 130) was then counted for each haplotype window for NA20847, and for comparison in the same rank-ordered windows, the trio-resolved CEU individual NA12871 (ref. 14). Pairs of homologous haplotype windows were rank ordered by differential similarity to GIH, and the fraction of novel variants on the GIH-enriched homolog was computed. For the 1000 Genomes-based comparison, haplotype windows were rank-ordered by divergence from CEU, YRI, and CHB+JPT populations and the fraction of novel variants per haplotype window computed for both NA20847 and NA12871 as before.

Pan-genome and novel contig mapping and anchoring. Whole-genome and clone pool-derived reads that did not align to the human genome reference (GRCh37/hg19) were mapped to novel contigs not present in the human reference genome assembly[7,8] to find contigs covered with ≥50 bp (phred-scaled mapping score ≥Q20). A subset of contigs were anchored by ≥2 reads with mates mapping to the reference. As further evidence of anchoring, intervals were identified in the reference assembly having read depth from clone pools also hitting a given contig but depleted among those pools not hitting that contig. Further novel sequences from NA20847 were assembled *de novo* from remaining unmapped reads using Velvet[33]. Contigs aligning to existing pan-genome sequences and contaminating sequences (*E. coli*, vector backbone, Epstein-Barr virus) were removed and remaining contigs were anchored as above. Sensitivity to detect and accurately anchor novel sequence was simulated by introducing *in silico* deletions into the reference, *de novo* assembling corresponding insertion contigs, anchoring as before, and measuring agreement between predicted anchoring location and the known site of simulated deletion.

31. Raymond, C.K. *et al.* Targeted, haplotype-resolved resequencing of long segments of the human genome. *Genomics* **86**, 759–766 (2005).
32. Sudmant, P.H. *et al.* Diversity of human copy number variation and multicopy genes. *Science* **330**, 641–646 (2010).
33. Zerbino, D.R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008).

doi:10.1038/nbt.1740