De novo assembly of a haplotype-resolved human genome

Hongzhi Cao[1–3,18], Honglong Wu[1,2,18], Ruibang Luo[1,4,18], Shujia Huang[1,5,18], Yuhui Sun[1,5,18], Xin Tong[1], Yinlong Xie[1,4,5], Binghang Liu[1,4], Hailong Yang[1], Hancheng Zheng[1,3], Jian Li[1,3], Bo Li[1], Yu Wang[1,5], Fang Yang[1], Peng Sun[1], Siyang Liu[1,3], Peng Gao[1], Haodong Huang[1,5], Jing Sun[1], Dan Chen[1], Guangzhu He[1], Weihua Huang[1], Zheng Huang[1], Yue Li[1], Laurent C A M Tellier[1,3], Xiao Liu[1,3], Qiang Feng[1,3], Xun Xu[1], Xiuqing Zhang[1], Lars Bolund[1,6,7], Anders Krogh[3], Karsten Kristiansen[1,3], Radoje Drmanac[8], Snezana Drmanac[8], Rasmus Nielsen[1,9,10], Songgang Li[1], Jian Wang[1,11], Huanming Yang[1,11,12], Yingrui Li[1,13], Gane Ka-Shu Wong[1,14,15] & Jun Wang[1,3,12,16,17]

The human genome is diploid, and knowledge of the variants on each chromosome is important for the interpretation of genomic information. Here we report the assembly of a haplotype-resolved diploid genome without using a reference genome. Our pipeline relies on fosmid pooling together with whole-genome shotgun strategies, based solely on next-generation sequencing and hierarchical assembly methods. We applied our sequencing method to the genome of an Asian individual and generated a 5.15-Gb assembled genome with a haplotype N50 of 484 kb. Our analysis identified previously undetected indels and 7.49 Mb of novel coding sequences that could not be aligned to the human reference genome, which include at least six predicted genes. This haplotype-resolved genome represents the most complete *de novo* human genome assembly to date. Application of our approach to identify individual haplotype differences should aid in translating genotypes to phenotypes for the development of personalized medicine.

Haplotype-resolved information for the human genome is essential for understanding the relationship between genotype and phenotype, as different configurations of exactly the same set of variants can sometimes result in different outcomes with regard to phenotype and disease susceptibility[1]. This information has typically been obtained by mapping sequence reads back to the human genome reference[2], and such methods cannot be applied to species for which a reference genome is not available. In addition, the human genome reference is chimeric, having been constructed from several anonymous Caucasian donors, and therefore does not represent the haplotype-resolved genome of any particular individual. Advanced next-generation sequencing (NGS) technology and numerous bioinformatics techniques[3–5] have been developed and applied to the production and analysis of large-scale human sequence data in many individual[6–9] and international projects[10–12]. However, NGS technology provides a short-read format of mixed DNA fragments that are derived from a pair of diploid chromosomes, posing challenges for determining haplotype information.

Several computational and experimental methodologies have been developed to obtain haplotype information, including (i) population-based statistical phasing by integration of unrelated individual data[10,13]; (ii) trio-based phasing applying Mendel's law of inheritance[14]; (iii) phasing by direct usage of sequence reads information[6]; (iv) experimental phasing that includes the use of various forms of PCR, atomic force microscopy with carbon nanotubes[15] and hybridization of probes to single DNA molecules[16,17]; (v) physical methods involving the initial preparation of haploid genomic material, for which the haplotype origin is distinguishable after sequencing[18–20]. Although statistical phasing is a well-established and widely reviewed method[21], it cannot phase rare variants when not enough samples are available from the general population and when additional family-based data are not avaliable[10,11,22]. Family trio–based phasing provides high accuracy for 80% of the known heterozygous single-nucleotide polymorphisms (SNPs) in the human genome, but the remaining 20% have to be resolved by alternative approaches, and collecting family samples for trio analyses remains difficult. Direct

phasing using read pairs spanning two heterozygous SNPs requires a high density of heterozygous markers and therefore results in short-range haplotypes. Experimental phasing cannot phase variants at long range and/or is inefficient at the genome scale, owing to the limited number of markers that can be phased in a single assay.

Recently, three methods for physical DNA preparation that allow long-range haplotyping have been developed. Microfluidics-based[18] chromosomal isolation techniques can physically separate entire chromosomes (from a single cell in metaphase) by haplotype, at least theoretically. Combined with NGS, this can directly produce the haplotype-resolved genome, but the laborious experimental process and need for sophisticated equipment prohibits the routine application of this technology. Methods based on fosmid clone pools[19,20] and long-read fragments[23] can generate long-range haplotypes, covering most of the functional elements in the human genome, although it remains difficult to phase an entire chromosome. Because these methods are based on short-read alignments using a good but incomplete reference[2], they have limited power for identifying intermediate-sized (51~200 bp) indels, long and complex structural variations, and novel sequences—all of which are rare, but likely have a large impact on human phenotypes[24].

Here we report a haplotype-resolved *de novo* assembly of a human genome generated entirely from NGS data without using an existing reference sequence. The total assembled genome size was 5.15 Gb with a haplotype N50 (the length for which the collection of all contigs of that length or longer contains at least half of the sum of the lengths of all contigs) of 484 kb, covering 97% of the autosomes and 96% of the gene regions (at least one haplotype covered), with 84% of the genes completely phased by at least one haplotype. Comparing the assembled diploid genome to the current reference genome (hg19; https://genome.ucsc.edu/cgi-bin/hgGateway?db=hg19), we identified ~4.05 M variation events, including 3.27 M SNPs, 777 k insertions and deletions ranging from 1 bp–38 kb, 111 inversions (1~87 kb), and 167 translocations (1~18 kb). We found 12 k heterozygous indels of 51~200 bp affecting 97 genes in the genome, which were not detected in the diploid sequence [7]. We also detected 7.49 Mb of sequence that could not be aligned to the human reference genome. These novel sequences include at least six predicted genes, which we validated using transcriptome data[25], and 21 kb that could be aligned to mammalian conserved sequence regions.

RESULTS
Data generation
For this study, we used an Asian individual (YH) whose genome, transcriptome and methylome have been well studied[7,25,26]. We generated ~605,566 fosmid clones, ranging in size from 3–50 kb, with a mean of 36.8 kb (**Supplementary Fig. 1**). We put ~33 fosmid clones into each pool and sequenced them on the Illumina HiSeq 2000 (**Fig. 1**). On average, each pool contained 0.04% of the haploid genome (~1.2 Mb). The DNA from each pool represented a small fraction of the genome, hence, the probability that any two fosmids in a given pool came from the same region was only 0.0033. In total, 1,712 Gb of raw sequence data was generated from the pooled fosmid libraries (**Supplementary Table 1**). The inferred fosmids coverage averaged ~8× for the haploid genome, with a median depth of 7× (**Supplementary Fig. 2**). For the initial *de novo* assembly of the YH genome, we used previously constructed whole-genome shotgun (WGS) libraries with both short and large insert sizes, sequenced on the Illumina HiSeq 2000 and amounting to another 203 Gb of raw data[27]. We also generated ~331 Gb of new WGS data on the Complete Genomics platform (**Supplementary Table 1**). In summary, we used ~2,246 Gb of raw data from three different libraries, as well as 67 Gb of transcriptome and ~105 Gb of methylation data from previous work[25,26].

Haplotype-resolved genome assembly
We developed a pipeline that combined fosmid pool-based NGS sequencing with *de novo* assembly[28] and computational haplotyping techniques[14] (**Fig. 1**).

First, we assembled the fosmid data, pool-by-pool, using SOAPdenovo2 (ref. 27). Over 88.5% of the assembled pools contained at least 80% of the corresponding fosmid sequence, and 54.7% of individual fosmids had a fragment equal to or greater than half of the fosmid length (**Supplementary Figs. 3** and **4**). In total, we obtained 23 Gb of fosmid-assembled haploid (FAH) sequences, consistent with the estimated total length of the inferred fosmid clones. The N50 of the FAH scaffolds was 17.9 kb (**Table 1**).

Second, we performed a *de novo* assembly for the YH genome using previously generated WGS data[29,30]. FAH sequences were then treated as long reads to increase the completeness and continuity of

Figure 1 *De novo* assembly of a haplotype-resolved diploid genome. Genetic material is derived from YH. DNA was fragmented and processed according to following steps: (1) WGS library construction. Libraries with insert sizes of 200~20,000 bp were used. (2) Fosmid pooling and library construction. Approximately 30 fosmid clones were combined into each pool. Libraries were constructed using a two-barcode system to distinguish the sequence reads from each pool. (3) DNA sequencing of WGS and fosmid libraries. (4) *De novo* assembly for each of the two kinds of sequence data. WGS reads with different insert sizes were combined to assemble an initial genome sequence. Fosmid reads were sorted by pool and assembled into FAH sequences. (5) Treat FAH sequences as long reads to fill in the gaps between and within the WGS assembly, obtaining an improved YH reference genome (6). (7) Heterozygous marker detection by aligning WGS reads and fosmid reads to the YHref. (8) Localization of the fosmids by mapping each fosmid pool's reads to YHref. (9) Heterozygous marker phasing, haploid sequence extension and (10) construction of the haplotype-resolved diploid genome sequence.
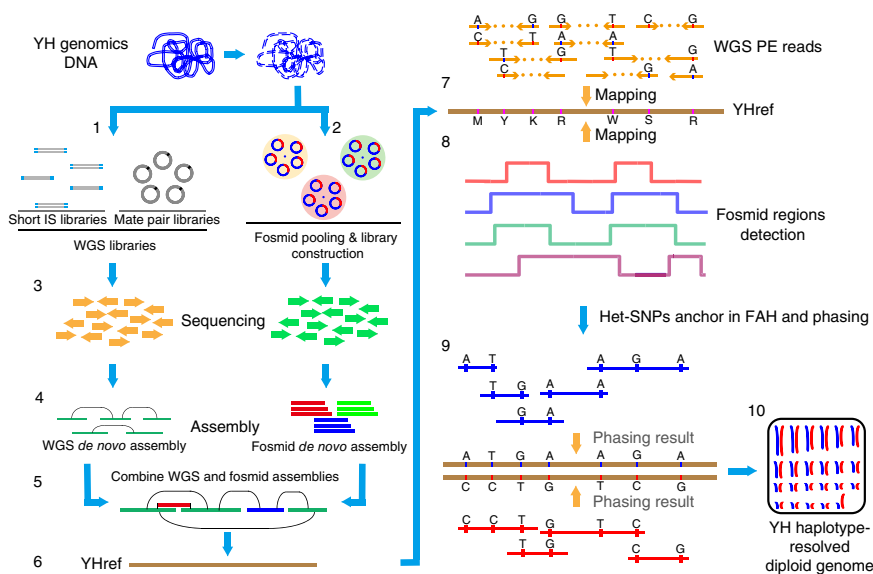
Table 1 Summary of assemblies

	Total length (bp)	Gap length (bp)	N10 (bp)	N50 (bp)	N90 (bp)
Fosmid scaffold	22,653,528,043	696,242,430	37,835	17,946	1,416
WGS scaffold	2,729,656,304	312,365,195	73,830,311	15,700,666	1,975,721
YH reference	2,883,329,361	72,898,567	78,377,698	23,192,260	3,739,289
HDG	5,152,727,603	78,921,926	1,390,629	484,222	99,527

this draft assembly, creating an improved reference genome that we called YHref. The total length of YHref was 2.9 Gb with a scaffold N50 length of 23.2 Mb (**Table 1**).

Third, we phased the FAH sequences by generating high-confidence heterozygous SNP markers and using ReFHap[14] to phase these heterozygous markers into haplotypes. Finally, the FAH sequences were separated into two groups, according to these phased heterozygous markers, and merged with the overlapping sequences from YHref to construct a haploid-resolved diploid genome (HDG) sequence (**Supplementary Fig. 5**).

The HDG sequence was classified into confident and nonconfident bases (**Supplementary Fig. 5**). Confident bases accounted for 92.4% of the HDG sequence, and on this basis, the error rate was estimated to be 8×10^{-5}. In total, we obtained 5.15 Gb of sequence with a haplotype N50 of 484 kb, approaching the expected haplotype N50 (510 kb) for this particular individual (**Supplementary Fig. 6a**). The phasing accuracy was 0.66 at that N50 length and better than 0.99 for phased gene regions. The unphased segments were mostly due to long regions of homozygosity[31] that exceeded the lengths of the fosmids used (**Supplementary Fig. 6b**). We compared our HDG sequence to hg19 using Lastz[32], together with a chaining and netting process, and found that 99% of the sequences could be aligned after excluding ambiguous, low-quality alignments. The assembled HDG covered 97% of hg19 and 96% of the RefSeq gene regions (coverage > 0), with slightly better coverage for the autosomes than for the sex chromosomes (**Supplementary Fig. 7**).

Variation detection and annotation

We modified a previously established[33] method to identify SNPs and structural variants in a comparison of our HDG sequence to hg19. There were a total of 3.27 M SNPs, 745 k short (≤50 bp) indels, 18 k intermediate (51~200 bp) indels, 13 k long (>200 bp; longest, 37 kb) indels, 111 inversions (longest, 87 kb) and 167 translocations (longest, 18 kb) (**Table 2** and **Supplementary Fig. 8**). Approximately 7% and 59% of the detected SNP and short indels, respectively, were novel and not present in either the 1000 Genomes Project or dbsnp137. We found that 75% of the novel SNPs and 72% of the novel indels were located in repetitive regions that were not analyzed in the previous study[7]. By comparing our assembled haploid-resolved sequence variants (ASV) with the variants from two resequencing platforms, Complete Genomics and Illumina HiSeq 2000, we found that 89% of the SNPs and 64% of the small indels from ASV could be validated by at least one resequencing-based method (**Supplementary Figs. 9** and **10**). Experimental and manual validation of randomly selected mutations that detected only in ASV, similar to analyses performed by the Genome in a Bottle Consortium[34], revealed that 30 of 50 selected SNPs and 12 of 20 selected indels were likely to be real. This demonstrates that our *de novo* assemblies were able to detect variants that other methods[35] could not (**Supplementary Figs. 11** and **12**).

Nevertheless, variants that were detected only in ASV still had a high false-positive rate (~40%). A detailed analysis of the false-positive sites showed that, in addition to biases in the fosmid coverage and systematic sequencing errors, repetitive sequences are still the biggest

challenge for *de novo* assembly of short-read data. We also analyzed missing variants that were detected by Complete Genomics and Illumina, but not part of the ASV. It turned out that 47% were in regions of low physical coverage by our fosmid pools, and 45% were removed by our quality controls.

The divergence between the YH genome and the hg19 reference, calculated on a window size of 100 kb, showed a unimodal distribution with a single peak at 0.14%. In contrast, the heterozygosity for YH—regions where the two haplotypes of YH are dissimilar—showed a bimodal distribution with a major peak at 0.1% and a minor peak at 0.02% (**Supplementary Fig. 13**). The minor peak is consistent with the observation of many long regions of homozygosity. In addition, we identified 24.0 Mb of hypervariable regions, defined not only by the levels of variation in comparison to hg19 but also by the heterozygosity between the two haplotypes. These variants are extremely difficult to analyze by traditional resequencing or consensus assembly methods. Some of these hypervariable regions were in functional gene regions that are known to have been subjects of adaptive evolution, including the major histocompatibility complex (MHC) and cytochrome P450 (*CYP*) gene regions. In total, we identified 219 genes in highly divergent regions. GO and KEGG analyses showed a significant enrichment in gene ontologies or pathways related to stimulus ($P =1.2 \times 10^{-13}$) and immune response ($P =2.8 \times 10^{-11}$) (**Supplementary Table 2**).

The HDG sequence allowed us to identify intermediate sized (51~200 bp) indels, in an unbiased manner, and long and complex structural variants, including novel gene sequences. We detected 32 k structural variants (>50 bp) with no obvious bias in terms of size and type (**Supplementary Fig. 8**). Among these, 12 k were intermediate-sized (51–200 bp) indels affecting 97 genes (defined here as having at least 1-bp overlap with an exon). We also detected 2,756 nonredundant novel sequences >100 bp in length, totaling 7.49 Mb in length and with 1.18 Mb shared between the two haplotypes. There were 193 kb of novel insertions and 5.2 Mb of novel haplotypes based on our classification system (**Supplementary Fig. 14a**); 5.6 Mb of the novel sequences were found within the gaps in the hg19 reference, fully covering 12 of these gaps. Not surprisingly, most of them were in repetitive regions, but 0.8% were located in known exons, possibly representing novel or alternative exons of known genes (**Supplementary Fig. 14c**).

As indicated in a previous study[36], genes that have not been previously described and are specific to a sequenced individual may be

Table 2 Summary of variations

|		Total	Heterozygous/homozygous	Rate found in repeats	Maximum length (bp)
SNP	All	3,271,996	1.63	0.54	1
	dbSNP137	3,048,013	1.46	0.53	1
	Novel	223,983	10.51	0.75	1
	Coding	18,735	1.71	0.01	1
	Non-synonymous	8,900	1.75	0.01	1
Small indel	All	745,266	2.55	0.68	50
	dbSNP137	316,638	2.52	0.63	50
	Novel	428,628	3.63	0.72	50
	Coding	661	4.20	0.14	50
	Frameshift	337	5.14	0.02	50
SV	All	32,012	5.85	0.65	87,178
	Intermediate indel	18,480	6.45	0.65	200
	Long indel	13,254	3.99	0.64	37,783
	Inversion	111	22.00	0.50	87,178
	Translocation	167	77.00	0.69	17,796
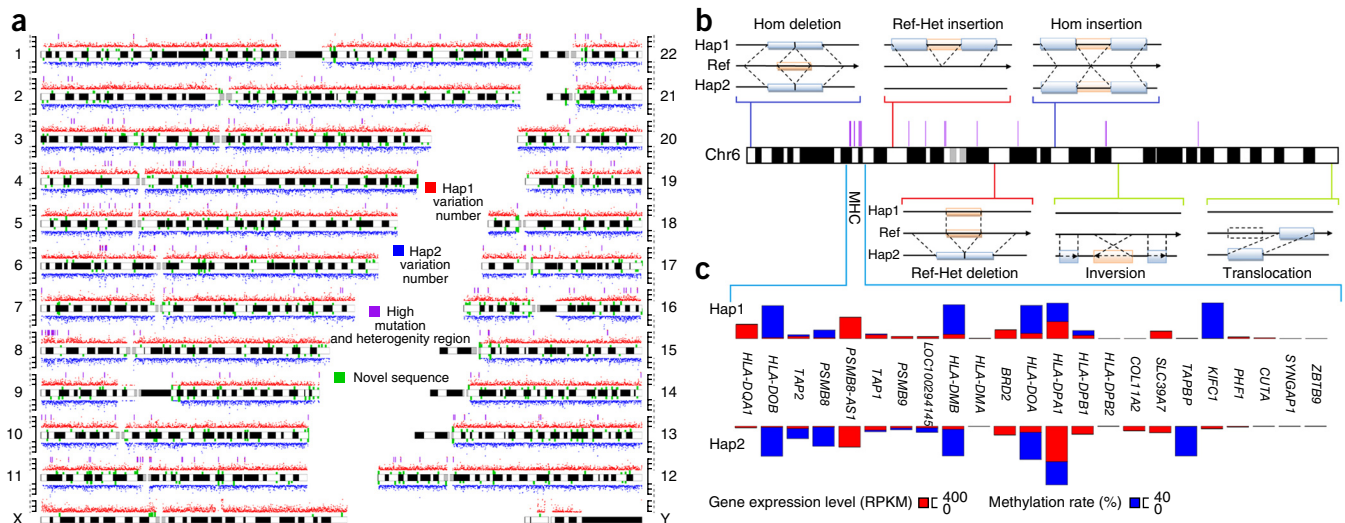
Figure 2 YH diploid genome landscape, generated by comparing our two haplotypes to the reference genome. (**a**) Chromosome overview showing distribution of variation. (**b**) Illustration of insertions, deletions, inversions and translocations of chromosome 6, which includes the MHC complex. (**c**) Methylation and expression patterns for partial MHC class II genes. The methylation rate (percentage) and expression level (RPKM) are calculated using sequencing data.

found among these sequences. We used transcriptome data generated from YH to determine which of these sequences represent transcribed (likely functional) genes, identifying 6 genes and 11 transcripts in cufflink[37] with high confidence (**Supplementary Table 3**). No stronger support could be found by searching the proteomics data generated from the same cell line, suggesting that these may be noncoding genes. However, 21 kb of the novel sequences could be aligned to mammalian conserved sequence regions with ≥95% identity, suggesting that some are functional.

Genome-wide distributions are depicted in **Figure 2**, including a magnification of the MHC class II region, which has one of the highest densities of genes and polymorphisms in the human genome.

Interpretation of the HDG sequence

Based on these data, we can conduct a disease-risk prediction. There were 18,735 SNPs, 661 indels, 21 inversions and 20 translocations in the exon regions, about half of which altered the amino acid sequence. Nonsynonymous SNPs and frame-shifting indels affected 5,054 genes; 1,330 of these changes (representing 1,112 genes) were predicted to be deleterious by both SIFT[38] and Polyphen2 (ref. 39) and were related to 154 known diseases. Among genes with deleterious variants, 33 were homozygously deleted. We identified 225 genes with more than two heterozygous variants that were predicted to be deleterious (**Supplementary Table 4**). Given that deleterious alleles can be on the same chromosome (*cis*-acting) or on opposite chromosomes (*trans*-acting), different haplotypes can have dramatically different phenotypes[40]. Of these 225 genes, ~62% (139) may have little impact because all of the damaging variants are in *cis*-configuration, hence, one copy of the gene on the homologous chromosome remains intact. Unless the deleterious variants have a dominant effect, the nonmutated copy will provide the necessary function (**Supplementary Fig. 15a**). Of the remaining 38% (86) that have damaging variants in the *trans*-configuration, so that both copies of the gene on the two homologous chromosomes have a deleterious variant (**Supplementary Fig. 15b**), 18 are known to play a role in the pathogenesis of complex diseases.

The HDG sequence provides a good opportunity to explore the relationship between allele-specific methylation (ASM) and expression (ASE). Although we had previously[34] described ASM and ASE using

phase information obtained through paired-end reads, the majority of the SNPs were not phased due to technical limitations. We therefore re-analyzed these data with our new improved assembly. In total we detected 1,108 genes with ASM in their regulatory or protein-coding regions (**Supplementary Fig. 16**). Epigenetic alterations within the promoter regions of genes are associated with regulation of gene expression, and can be used as biomarkers for cancer[41]. We found 332 genes with ASM in the promoter regions, including *ADA*, which plays important roles in purine metabolism and adenosine homeostasis[35]. Most of these genes had not been identified in our previous study[26]. We also found 1,621 genes with ASE, including *CKLF*[33], which plays important roles in inflammation and regeneration of skeletal muscle. Additionally, 122 genes had both ASE and ASM (**Supplementary Table 5**). GO enrichment was performed with DAVID[29,42] and 15 GO categories that are related to the basic metabolic and catalytic function, such as GO:0008152 ($P = 2.79 \times 10^{-03}$) and GO:0003824 ($P = 5.45 \times 10^{-03}$) were found significantly enriched (**Supplementary Table 5**). For the remaining genes (92%) with ASE but not ASM, these may be the direct result of haplotype differences, indicating the importance of having a HDG.

DISCUSSION

We have presented here the first HDG sequence based on *de novo* assembly with NGS technology. Four previous studies used a similar library construction and sequencing strategy[19,20,23,43], but they identified variants by mapping fosmid-pooled sequence reads to a pre-existing reference genome, and therefore were limited in their ability to detect complex and long structural variants. Furthermore, our fosmid pools contained a much smaller number of clones (30 in this study versus 5,000 in a previous study[20]), and we were therefore able to get more accurate haploid information.

Notably, we detected a large number of novel intermediate-sized heterozygous indels (51~200 bp), which are difficult or impossible to detect by traditional paired-end sequence mapping as these variant sizes are within the insert size distribution of a typical library (500~800 bp). These variants are also difficult to detect using split-read or local assembly, as indels of this nature often produce incorrect signals or no signal when doing the alignments (**Supplementary Figs. 11** and **12**). We also identified sequences that,

although not necessarily unique to this individual, had not been previously reported. The average lengths (2,717 bp) were much longer than in our previous study[36] (710 bp). Some of these sequences affected known genes, and others contained novel coding or regulatory sequences that had not been reported before.

Given that our HDG assembly is currently the most complete representation of an individual's genome, and that complementary and comprehensive resources are available from this individual, including genome[7] and transcriptome data[25], methylation data[26] and established cell lines, the extant work and resources can serve as a standard for developing new sequencing and assembly techniques, and for functional studies involving RNA or protein analysis. In this sense, the YH resources mirror the work by The National Institute of Standards and Technology and the Genome in a Bottle Consortium[34] to provide a highly accurate set of genotypes across a genome that others can use as a benchmark.

Further, the pipeline developed here lays the foundation for *de novo* assembly of genomes with high levels of heterozygosity and/or of polyploid origins, for which current assembly methods typically produce incomplete and fragmented genomes[7,28]. More efficient implementations of this pipeline should be possible in the future, by adapting the algorithms developed here to deal with emerging technologies, like long fragment reads[23], that do not require the preparation of fosmid libraries, do not have 40-kb size limits and integrate barcodes with mate-paired reads.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Raw WGS reads from the YH genome, generated on the Illumina Hiseq2000 platform, are available from the BGI website: http://yh.genomics.org.cn/download.jsp, the EBI Short Read Archive (SRA) (EMBL: ERP001652), and the GigaScience database (http://gigadb.org/dataset/100015). WGS data that were generated on the Complete Genomics (CG) platform are available in EBI European Nucleotide Archive (ERR792927). Fosmid reads are available in the NCBI Sequence Read Archive (PRJEB7886) as well as in the GigaScience database (http://gigadb.org/dataset/view/id/100096/token/hgxjAoybUwd6pREG). The fosmid assembled haploid (FAH) and haploid-resolved diploid genome (HDG) sequences are available at the YH database: http://yh.genomics.org.cn/download.jsp. Transcriptome data are available from the NCBI Sequence Read Archive (SRA043767.1) and methylation data can be downloaded from the NCBI Gene Expression Omnibus (GSE17972). All variants detected by comparing our HDG sequence to hg19 are available at the YH database: http://yh.genomics.org.cn/download.jsp. The YH cell line or the DNA from that cell line can be obtained by sending a request to the corresponding authors.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

1. Tewhey, R., Bansal, V., Torkamani, A., Topol, E.J. & Schork, N.J. The importance of phase information for human genomics. *Nat. Rev. Genet.* **12**, 215–223 (2011).
2. Lander, E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
3. Li, R. *et al.* SNP detection for massively parallel whole-genome resequencing. *Genome Res.* **19**, 1124–1132 (2009).
4. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
5. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
6. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5**, e254 (2007).
7. Wang, J. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456**, 60–65 (2008).
8. 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).
9. Tong, P. *et al.* Sequencing and analysis of an Irish human genome. *Genome Biol.* **11**, R91 (2010).
10. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
11. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
12. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
13. Browning, S.R. & Browning, B.L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007).
14. 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).
15. Woolley, A.T., Guillemette, C., Li Cheung, C., Housman, D.E. & Lieber, C.M. Direct haplotyping of kilobase-size DNA using carbon nanotube probes. *Nat. Biotechnol.* **18**, 760–763 (2000).
16. Cao, H. *et al.* A short-read multiplex sequencing method for reliable, cost-effective and high-throughput genotyping in large-scale studies. *Hum. Mutat.* **34**, 1715–1720 (2013).
17. Dapprich, J., Ferriola, D., Magira, E.E., Kunkel, M. & Monos, D. SNP-specific extraction of haplotype-resolved targeted genomic regions. *Nucleic Acids Res.* **36**, e94 (2008).
18. Fan, H.C., Wang, J., Potanina, A. & Quake, S.R. Whole-genome molecular haplotyping of single cells. *Nat. Biotechnol.* **29**, 51–57 (2011).
19. Kitzman, J.O. *et al.* Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nat. Biotechnol.* **29**, 59–63 (2011).
20. Suk, E.-K. *et al.* A comprehensively molecular haplotype-resolved genome of a European individual. *Genome Res.* **21**, 1672–1685 (2011).
21. Browning, S.R. & Browning, B.L. Haplotype phasing: existing methods and new developments. *Nat. Rev. Genet.* **12**, 703–714 (2011).
22. Altshuler, D.M. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
23. Peters, B.A. *et al.* Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature* **487**, 190–195 (2012).
24. Alkan, C., Coe, B.P. & Eichler, E.E. Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* **12**, 363–376 (2011).
25. Peng, Z. *et al.* Comprehensive analysis of RNA-seq data reveals extensive RNA editing in a human transcriptome. *Nat. Biotechnol.* **30**, 253–260 (2012).
26. Li, Y. *et al.* The DNA methylome of human peripheral blood mononuclear cells. *PLoS Biol.* **8**, e1000533 (2010).
27. Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1**, 18 (2012).
28. Zhang, G. *et al.* The oyster genome reveals stress adaptation and complexity of shell formation. *Nature* **490**, 49–54 (2012).

29. Huang, W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2009).
30. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
31. Kirin, M. *et al.* Genomic runs of homozygosity record population history and consanguinity. *PLoS ONE* **5**, e13996 (2010).
32. Schwartz, S. *et al.* Human-mouse alignments with BLASTZ. *Genome Res.* **13**, 103–107 (2003).
33. Han, W. *et al.* Molecular cloning and characterization of chemokine-like factor 1 (CKLF1), a novel human cytokine with unique structure and potential chemotactic activity. *Biochem. J.* **357**, 127–135 (2001).
34. Zook, J.M. *et al.* Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat. Biotechnol.* **32**, 246–251 (2014).
35. Aghaei, M., Karami-Tehrani, F., Salami, S. & Atri, M. Adenosine deaminase activity in the serum and malignant tumors of breast cancer: the assessment of isoenzyme ADA1 and ADA2 activities. *Clin. Biochem.* **38**, 887–891 (2005).
36. Li, R. *et al.* Building the sequence map of the human pan-genome. *Nat. Biotechnol.* **28**, 57–63 (2010).
37. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
38. Ng, P.C. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**, 3812–3814 (2003).
39. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
40. Rusk, N. Causal mutations in a haploid landscape. *Nat. Methods* **9**, 36 (2012).
41. Yang, M. & Park, J.Y. DNA methylation in promoter region as biomarkers in prostate cancer. *Methods Mol. Biol.* **863**, 67–109 (2012).
42. Huang, W., Sherman, B.T. & Lempicki, R.a. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37**, 1–13 (2009).
43. Adey, A., Burton, J., Kitzman, J. & Hiatt, J. The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. *Nature* **500**, 207–211 (2013).

ONLINE METHODS

Public data used and new data produced. WGS data for the YH genome, generated on the Illumina HiSeq 2000 (HS) platform, was downloaded from the YH database (yh.genomics.org.cn). Transcriptome and methylation data for the same individual were downloaded from the NCBI Sequence Read Archive (accession number SRA043767.1) and the Gene Expression Omnibus (accession number GSE17972), respectively. In addition to these public data, 1,712 Gb of pooled-fosmid data was generated on the Illumina platform (see below for details), and ~331 Gb of new WGS data were generated on the Complete Genomics (CG) platform (**Supplementary Table 1**). Lastly, genotyping data were also generated using the Illumina HumanOmni2.5M array.

Construction and sequencing of fosmid libraries. Fosmid libraries were constructed using protocols developed for the oyster genome[28]. 100 μg of purified genomic DNA from the YH cell line was randomly sheared and electrophoresed on a 1% Ultrapure agarose gel (Invitrogen) using a Bio-Rad CHEF-DR II Pulsed Field Gel Electrophoresis apparatus. DNA was purified and recovered by the GELase enzyme method with gel-excised fragments of ~35–45 kb. The fragments were blunt-end repaired and ligated to a prelinearized pCC2FOS vector. Fosmids were packaged using the MaxPlax Lambda Packing Extract (Epicentre), used to infect EPI300-T1R cells and diluted to a suitable concentration. The infected bacterium solution was added to an LB medium containing 12.5 μg/ml chloramphenicol and cultured in a 37 °C incubator overnight. On average, 30 fosmid clones were cultured together to form a single fosmid pool. Then, 3 μg of DNA from each pool was digested, and fragments of insert size ranging from 200 to 800 bp were selected. Adapters containing the 11-bp barcode were ligated to these selected fragments to form a single pooled-fosmid library. Barcoded fragments from, on average, ~300 single pooled-fosmid libraries were then pooled again (evenly) to create ~76 stage I barcode libraries. DNA fragments of 180–650 bp (lengths exclude barcode) from each stage I barcode library were used to construct two independent libraries (one with small insert sizes and one with intermediate insert sizes). Each library was PCR amplified with index primers, each of which contained an 8-bp barcode, to form a Stage II barcode library. Some libraries failed, but in total, 139 Stage II barcode libraries comprising ~600 k fosmid clones were constructed and sequenced on the Illumina platform (**Supplementary Fig. 17** and **Supplementary Table 1**). Notice that each fosmid pool can be distinguished by two sets of barcodes (11 bp and 8 bp).

Hierarchical assembly of the HDG sequence. We used the 'hierarchical assembly' strategy previously developed for the highly heterogeneous oyster genome[28], with a few modifications. The overall strategy is represented by the following steps:

1. Assembly of single fosmid clones in each pool. For the 139 Stage II barcode libraries, sequence reads were separated according to their original pool (consisting of ~30 fosmids) by deconstructing the two barcodes. SOAPdenovo2 (ref. 27) was employed to assemble the fosmid clone sequences. A range of *k*-mers from 29 to 63 was tested, and the optimal *k*-mer was chosen to produce the longest contig N50. Missing sequences were filled with the gap-closure module from SOAPdenovo2. A post-assembly correction procedure was applied, using paired-read support and local re-assembly. The fosmid scaffolds so generated are called the FAH sequence, which totaled ~23 Gb.

2. Construction of YH reference genome. WGS reads from YH were assembled by SOAPdenovo2 with a *k*-mer size of 45. The resultant assembly contained 2.73 Gb of sequence with a contig N50 of 2,310 bp and a scaffold N50 of 15.7 Mb. We separated the FAH sequences at their indeterminate (N) bases and treated the resultant fragments as long sequence reads to fill the gaps between and within WGS-assembled scaffolds, to break any erroneous connections, and to connect the nearby scaffolds. This method increased the scaffold N50 of YH consensus assembly from 15.7 Mb to 23.2 Mb and increased the genome size from 2.73 Gb to 2.88 Gb (**Table 1**), indicating better overall completeness, continuity and accuracy. This improved YH assembly was used as the reference (YHref) for subsequent heterozygous marker detection, variant phasing and haploid sequence assembly.

3. Heterozygous marker detection and correction. Phasing is the foundation of haploid-resolved assembly, and heterozygous marker detection is the key challenge for phasing. To increase the accuracy of phasing, we only used heterozygous SNPs as markers because: (i) most of the variations are SNPs, (ii) phasing accuracy is more sensitive to quality than quantity and (iii) accuracy of SNP detection is much higher than other classes of variants. Raw WGS reads and pooled-fosmid reads were mapped to YHref using BWA[4] (0.6.2). All reads were sorted and duplications were removed. The location of each fosmid on YHref was inferred from the mapping results (see below for detail). We calculated the coverage for WGS reads and the physical coverage for fosmids, separately, and then defined a 'golden region' on YHref as a region that meets the following three criteria: (i) WGS read coverage is within 3 s.d. of the WGS mean. (ii) WGS reads with multiple alignments to YHref comprise less than 10% of the total reads mapped. (iii) Inferred fosmid physical coverage ranges from 1 to 20. The golden regions so defined totaled 2.6 Gb (90% of the effective sites). For the WGS alignments, heterozygous SNPs with genotype quality ≥20 were detected using SOAPsnp. For the fosmid alignments, all overlapping fosmids derived from the same haplotype were combined. Heterozygous SNPs were selected from this combined sequence. Both sets of heterozygous SNPs were used as markers for the subsequent processes. Potential false positives were detected through the conflicting linkage information from different fosmids covering the same genome region, for example, for neighboring heterozygous sites A/T and C/G, if all of the fosmids support A-C and T-C linkage, in contrast to the expectation of A-C/T-G or A-G/T-C linkage, we can conclude that the C/G site is a false positive and set it to C/C. In the end, 2.06 million high-confidence heterozygous SNPs were obtained.

4. Construction of the haplotype-resolved diploid genome. ReFHap[14] was used to phase heterozygous SNPs into long haplotype blocks based on the linkage information provided by each fosmid. Different FAH sequences from the same haplotype were merged together to generate a consensus haploid sequence. Bases with the highest support were retained whenever inconsistencies appeared. Sequences from YHref were employed when no FAH sequences were available (**Supplementary Fig. 5**). Each site in the HDG was classified as: (i) Confident, if there is FAH support and no conflicts among the different FAH sequences from the same haplotype. (ii) Un-confident, if there is no FAH support or there are conflicts between different FAH sequences from the same haplotype (**Supplementary Fig. 5**).

Completeness and accuracy of the FAH sequence. To evaluate the completeness of the FAH sequences, we first estimated the fosmid number and length for each pool, by mapping the raw fosmid sequence data to hg19 and YHref. A potential fosmid region would have an average read depth ≥5×, a length of 3–50 kb, and at least 50% coverage. To exclude spurious fosmids caused by repetitive sequences, and fosmids with poor sequence data, we also aligned the FAH sequences to hg19. Only those regions covered by raw reads and FAH sequences were accepted as true fosmids. This procedure gave us 605,566 fosmids, with an average count of 33 per pool, and an average size of 36.8 kb (**Supplementary Fig. 1**) per fosmid. We then assessed completeness from three viewpoints.

1. Completeness of assembled sequence for each pool. The assembled sequence length is divided by the inferred length of each pool. Over 88.5% of the pools contained at least 80% of the expected sequence, and over 53.2% of the pools contained at least 95% of the expected sequence (**Supplementary Fig. 3**).

2. Continuity of the assembled sequence for each fosmid. Even when all of a fosmid is assembled, it may fragment into several pieces. In particular, the N50 size was only 17.9 kb, or half the length of a typical fosmid. We found that 54.7% of our fosmids had a longest assembled fragment sequence equal or longer than half of the inferred fosmid length, and about 18% of the fosmids were completely assembled. (**Supplementary Fig. 4**).

3. Base-pair accuracy of the fosmid-pool assemblies. To evaluate accuracy, we randomly selected 100 fosmid pools, each of which had a fosmid

count and sequence depth near the average for this study. Taking the FAH sequence as a reference, we mapped the raw reads back to the assembled sequence with BWA, and used samtools to detect potential 'mutations'. In principle, homozygous mutations would indicate assembly errors, but in practice, sequencing errors and repetitive sequences make the situation more complicated. We therefore selected both homozygous and heterozygous mutations and defined the assembly error rate as the total length of the detected variants divided by the total assembled sequence length in each pool. The resultant error rate (for these 100 pools) was 0.03%; counting only homozygous mutations the error rate was 0.006%.

Phasing accuracy of HDG sequence. A previous[20] study evaluated phasing accuracy by comparing the results provided by a *de novo* assembly with that of a statistical method based on population data. The overall discordance was 6.4%. However, there are many problems with this method: (i) It requires a reference population data set matched to the studied sample. Statistical methods, such as MACH[44], would then impute the phasing for each individual. (ii) It has little power to evaluate rare variants. (iii) Being probabilistic in nature, it does not count as a gold standard. With these limitations in mind, we noticed that as each pool had only about 30 fosmids, the probability that any two fosmids would come from the same genome region and have a different haplotype is extremely unlikely, hence, the markers in each fosmid region could serve as a gold standard for assessing haplotype accuracy. By mapping each fosmid pool's raw reads to YHref, we inferred the location of and obtained the haplotype information for each fosmid. Then, we randomly selected fosmids that represent ~1× physical coverage of YHref and calculated the consistency of the marker's linkage information between the selected fosmids and the HDG phasing. This showed that the phasing accuracy of two adjacent fosmids is 96.2%. Taking an average fosmid length of 36.8 kb into consideration and assuming the overlap region between two adjacent fosmids is 3/4 of a fosmid length, the phasing accuracy at the phased N50 length (484 kb) is 0.66 (calculated as $0.962^{484/(36.8+36.8/4)}$). For the fully phased genes where both alleles are 100% covered by our HDG sequence, the average gene length was 30,743 bp, shorter than the average fosmid length and haplotype phasing accuracy exceeds 0.995.

Effectiveness of the phasing algorithm. In addition to phasing accuracy, one wants the phased regions to be as long as possible. We therefore compared our HDG result with a previous study on MP1 (ref. 20). Three factors influence phasing length: (i) Long regions of homozygosity (ROH, defined here as homozygous ≥ 20 kb) make it impossible to differentiate haplotypes. The more long ROHs there are, the more fragmented the results will be. (ii) Number of fosmid clones sequenced. (iii) Completeness of reference genome. We found 18,361 and 15,062 long ROHs in YH and MP1, respectively. YH had ~22% more long ROHs then MP1. We wanted to know if this difference was specific to the individuals studied or to the underlying populations. So we calculated the ROH numbers for the 1000 Genomes Project[45] (**Supplementary Fig. 6a**). Populations with African ancestry (LWK, ASW, YRI) had the lowest ROH numbers, as expected. Next were populations with American ancestry (MXL, PUR, CLM) and European ancestry (GBR, TSI, IBS, FIN, CEU). Populations with East Asian ancestry (CHS, CHB, JPT) had the highest ROH numbers, in agreement with one previous study[46], but not with another study[31]. The total number of fosmids used for YH was 605,566, representing 8× physical coverage of the haploid genome (**Supplementary Fig. 2**), in comparison to the MP1 study, the number of fosmids used was 66% larger (or 1,005,000 more) than in the current study. Moreover, in the YH study we used a *de novo*–generated reference with 1,752 scaffolds and an N50 of 23 Mb, in comparison to MP1, they used the 24 chromosomes of hg19 as a reference and had an N50 of 38.4 Mb. To compute the expected haplotype N50 for YH and other previously analyzed genomes (MP1 (ref. 20), Irish[9], NA18507 (ref. 47), we used hg19 as a reference and took the long ROHs into account. Based on the heterozygous markers detected in each study, we calculated expectations for different fosmid coverages (**Supplementary Fig. 6b**). At a fosmid coverage of 4× per haplotype (or 8× assuming a 3-Gb genome), we would expect a haplotype N50 of ~510 kb for YH, in fact, we achieved an N50 of 484 kb, despite having to compute the reference genome *de novo*. In comparison, for MP1 at a fosmid coverage of 6× per haplotype the expected N50 is 964 kb, similar to what was reported in the corresponding study (959 kb at 6.38×).

Completeness and accuracy of the HDG sequence. The assembled HDG contained 5.15 Gb of sequence with an N50 length of 486 kb. To evaluate completeness, we aligned our sequences to hg19 using Lastz, and applied a chaining and netting procedure using certain algorithms from the UCSC Kent tools. After removing duplicates and ambiguous alignments, 97.1% of the autosomes were covered by at least one haplotype and 87.6% were covered by both. Only 96.6% and 60.6% of the X and Y chromosomes were covered, respectively (**Supplementary Fig. 7a**). For the gene regions defined by RefSeq, 97.3% were at least partially covered by one haplotype, and 86.2% were covered by both (**Supplementary Fig. 7b**). If we require complete coverage of the gene regions, 84.8% were covered by at least one haplotype, and 56.9% were covered by both. In total, 91.2% of the human effective diploid genome was assembled, consistent with the assembled length of 5.15 Gb.

We generated additional WGS data on the CG platform with an average depth of 115×. Together with the data from the Illumina platform, we did a resequencing analysis against hg19 for YH on each platform (details given in next two sections). From this we generated a "gold standard" data set of genotypes for YH against hg19, using the intersection of the data from these two platforms. The criteria were: (i) For CG data, sites with score ≥50 and depth ≥30 were considered. (ii) For Illumina data, sites with consensus quality ≥50 and depth ≥30 were considered. (iii) Only sites with consistent genotypes between platforms were retained. The final "gold standard" genotype data set covered 2.17 Gb of the YH genome. In comparing long HDG sequences to hg19, there is always the possibility of alignment errors to complicate the evaluation. Hence, we chopped HDG into 150-bp pieces, requiring 50 bp of overlap between adjacent pieces, and aligned these pieces to hg19 by BWA. After excluding low-quality (map quality <30), simple repeats and low-complexity sequences, the alignments covered 92% of the "gold standard" genotype sites with a consistency of 99.992%, indicating that the error rate for the HDG assembly was <1 per 10 kb.

Variation detection for the assembled HDG sequence. Following the pipeline developed in a previous study[48], Lastz was chosen to align the long HDG sequences to hg19. We chopped the HDG scaffolds into 100-bp segments and pre-aligned them to the reference genome using BWA to determine possible chromosomal locations. These segments were then grouped together and aligned to the corresponding chromosome using Lastz. As different parameter settings can lead to different alignments, we were particularly cautious about biases, for example, for SNPs, small indels, inversion and translocations. Conversely, more fragmented alignments make detection of larger variations difficult, as gap-free alignment is essential for detecting large indels or structural variants. **Supplementary Table 6** shows the different parameters used for our Lastz alignments. A dynamic-programming algorithm "axtBest" confirmed the best hits. Based on this processed alignment, the exact locations of variants (SNPs, indels, inversions and translocations) and breakpoints were determined. For the loci with both haplotypes covered, full genotyping was performed. Otherwise, it was annotated as "half."

Resequencing analysis on CG and Illumina platforms. We performed a resequencing analysis against hg19 using WGS data from CG and Illumina. For CG, where the sequence depth was 115×, genotypes were called using the standard CG pipeline; 3,411,305 SNPs and 510,944 small indels were generated from the high-quality mutations in the masterVar file. For Illumina, only reads with insert size <1 kb were used, and these represented a ~50× coverage of the genome. Reads were mapped to hg19 using BWA, sorted and duplications were removed by samtools. After merging, GATK was used to detect variants. The resultant variants included 3,365,182 SNPs and 633,679 small indels (**Supplementary Table 7**).

Evaluation of detected variation (YH versus hg19). For the YH genome, using the HDG sequence compared against hg19, we detected 3.27 M SNPs, 745 k small indels (1~50 bp), 32 k intermediate length indels (51~200 bp) and 13 k long indels (>200 bp) (**Table 2**). About 7% and 59% of the detected SNPs and short indels were novel —not present in either the 1000 Genomes Project or dbsnp137. We found that 75% of these novel SNPs and 72% of these novel indels were in repetitive regions (**Table 2**). A length distribution analysis of the detected indels showed no obvious bias, in terms of size and

type (**Supplementary Fig. 8**). To further evaluate the accuracy and completeness of the assembled sequence variants (ASV), we compared ASV against YH data generated on an Illumina HumanOmni2.5M genotyping array, as well as variants detected by performing a resequencing analysis on the two sequencing platforms (CG and Illumina). For the genotyping array, concordance (i.e., sites with same genotypes) with ASV was 99.6%. The sensitivity (i.e., fraction of genotyping array SNPs detected by ASV) was 85.6%. Comparing CG and Illumina resequencing data with ASV, as shown in **Supplementary Fig. 9**, a majority (68.2%) of the 4.0 M overall SNPs were concordant between all three data sets. However, there were still tens of thousands of method- and platform-specific calls, which is similar to a previous study[49]. If we use the overlap between CG and Illumina as a gold standard, the concordance with ASV was 99.5%, and the sensitivity was 87.3%, similar to the results from the genotyping array, indicating a FP rate for ASV-detected SNPs of 0.5~0.6% and a FN rate of 12.7~14.4%, as determined by a cross-validation with the genotyping array data and the overlap between the two resequencing analyses on the two different platforms. The concordance was higher when comparing resequencing platforms (CG vs. Illumina, 78.4%) than when comparing assembly-based and resequencing analyses (ASV vs. CG, 71.1%; ASV vs. Illumina, 69.6%), indicating a bias for the method of variant detection. ASV also had the highest fraction of method-specific SNPs (9.2%), compared to CG (3.8%) and Illumina (4.2%). Of these ASV-specific SNPs, 58.4% were novel. We randomly selected 50 (20 known and 30 novel) for Sanger sequencing and manual validation, finding that 60% (15 known and 15 novel) were true variants. We also evaluated the concordance for indels. As resequencing-based analyses are limited to a size of ~50 bp and cannot detect het-het (i.e., both alleles different from reference and each other) indel sites, we compared only heterozygous and homozygous indels with size 1~50 bp. As is well known, positional shifts (mostly caused by indels in tandem repeat regions) can also result from the use of different methods of detection. Our first step was to determine the optimum flank size, starting at zero and excluding indels at the same position, then plotting the concordance between ASV and the two resequencing analyses as a function of increasing flank size. The plot stabilizes above 50 bp (**Supplementary Fig. 18**). Using that as our cutoff, the result is in general agreement with a previous study[11]. Of the ~1 million indels detected, only 27.6% were concordant between three methods/platforms (**Supplementary Fig. 10**). The ASV method seemed to find many more long indels than the two resequencing-based methods (**Supplementary Fig. 19**). To confirm that these were mostly real variants, we used Sanger sequencing and manual validation on 20 ASV-specific indels. Of these, 12 displayed strong support from the mapped fosmid sequence (**Supplementary Figs. 11,12** and **20**).

The validation results indicate that, although HDG assemblies increased our ability to detect new variants that other resequencing analyses could not, these method-specific variants contained a high number of false positives (~40%). To better understand the sources of these errors, we performed a detailed analysis on each site validated to be a false-positive mutation. 60% of them proved to be in repetitive regions. 15% of them proved to be in regions of low coverage by fosmid pools (of these, 66.7% overlapped with repetitive sequences). For the remaining 25% of sites, influenced by neither repetitive sequence nor low fosmid coverage, 80% are likely due to assembly errors—based on analysis of the repeat boundaries in the *de Bruijn* graph data structures which may place two similar sequences incorrectly, or cause mapping errors in Lastz, as we could not detect abnormalities in the assembled bases in mapping the raw sequence reads back to the assembled sequence. Finally, 20% of the sites were due to systematic sequencing errors, as we could detect abnormalities after mapping. In total, 10.2% of the SNPs and 8.5% of the indels could be detected by CG and Illumina, but not by ASV (**Supplementary Figs. 9** and **10**). So we selected 1,072 random sites and analyzed the cause for this peculiarity. Our result showed that 47% were not assembled by ASV due to the low read coverage at those sites, and 45% were located in low-confidence regions, which were removed during variant detection in ASV.

Detection of hypervariable regions in YH. To identify hypervariable regions, we used 100kb windows and shifted these windows in steps of 20 kb. Heterozygosity was defined as the difference (SNPs and indels combined) between the two haplotypes of YH. A similar computation was also done for YH vs. hg19, revealing that the most common rate of variation between YH

and hg19 was 0.14% (**Supplementary Fig. 13**). However, the heterozygosity rate for the difference between the two haplotypes of YH was bimodal, with a major peak at 0.1% and a minor peak at 0.02%. The minor peak is consistent with our long-ROH result (**Supplementary Fig. 6a**), revealing that there are many regions in the YH genome with very low heterozygosity. We defined a hypervariable region as one with a high variation rate (≥0.0040) in comparison to hg19 and a high heterozygosity rate (≥0.0029) within YH, with the cutoffs based on the 99% line in the two distributions (**Supplementary Fig. 13**). We excluded regions that may have been distorted by copy number variations, by examining the depth information in the WGS data. Ultimately, 24.0 Mb of the HDG sequence, involving 255 genes (**Supplementary Table 2**), was declared to be hypervariable.

Novel sequence and novel gene detection. A previous study, also based on *de novo* assembly, showed that an individual's genome can have novel sequences that do not exist in an arbitrary reference genome[36]. We therefore examined all contiguous regions of the HDG sequence that were >100 bp and had no alignment to hg19. These candidate novel sequences were then further aligned to the NCBI nucleotide database (nt) using BLAST. Sequences that had less than 90% identity to hg19, and to the microbe and plant sequences in nt, were declared to be novel (**Supplementary Fig. 14**). We identified 7.49 Mb of these sequences (**Supplementary Table 8**). To identify novel genes that are also transcribed, and hence might be functional, we aligned RNA sequences generated for YH to these novel genome sequences using cufflink (**Supplementary Table 3**).

Annotation of the detected gene variants. All of the detected variants were analyzed by ANNOVAR[30], SIFT[38] and PolyPhen2 (ref. 39) to predict the likelihood that they might have damaging effects. *Cis*- and *trans*-configurations were defined as described in the previous study of MP1 (ref. 5). For genes with more than two heterozygous variants, configurations of the heterozygous variants were determined (**Supplementary Fig. 15**). We only kept those mutations detected in ASV that also had support from CG or Illumina resequencing. A total of 141 *cis*- and 89 *trans*-configuration genes were found (**Supplementary Table 4**).

Allelic-specific methylation and expression. Bisulfite sequencing reads, as well as RNA sequencing reads previously generated for YH, were aligned back to YHref. We restricted this analysis to the most reliable variants that were used as markers in the haplotype resolution process. Bisulfite and RNA sequencing reads were separated into two groups based on which of the two haplotype they were best matched to. For the allele-specific methylation (ASM) study, we calculated read counts for methylcytosines and cytosines on each haplotype block and used Fisher's exact test to define significantly different methylation patterns. ASM regions had to satisfy the following criteria: (i) Fisher's exact test $P < 0.01$; (ii) twofold change in methylation levels; (iii) at least one haplotype with methylation level >20%; (iv) at least five CpG cytosine codons.

In the allele-specific expression (ASE) work, we followed a previous study[50] that used a Poisson distribution to define significantly different levels of expression. More specifically, genes with a twofold change in RPKM, $P < 0.01$ and FDR < 0.01 were said to exhibit ASE. In total there were 1,108 genes with ASM and 1,621 genes with ASE. Of these, 122 had both ASE and ASM (**Supplementary Table 5** and **Supplementary Fig. 16**).

Programs used in the study: SOAPdenovo2: http://soap.genomics.org.cn/soapdenovo.html; BWA: http://bio-bwa.sourceforge.net; SOAPsnp: http://soap.genomics.org.cn/soapsnp.html; samtools: http://samtools.sourceforge.net; GATK: http://www.broadinstitute.org/gatk; ReFHap: http://owww.molgen.mpg.de/~genetic-variation/SIH/Data/algorithms; Lastz: http://www.bx.psu.edu/~rsharris/lastz; Knet: http://compbio.med.wayne.edu/software/KNet/knet.zip; ANNOVAR: http://annovar.openbioinformatics.org/en/latest/; SIFT: http://sift.jcvi.org/; Polyphen: http://genetics.bwh.harvard.edu/pph2 and **Supplementary Code**.

44. Li, Y., Willer, C.J., Ding, J., Scheet, P. & Abecasis, G.R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* **34**, 816–834 (2010).

45. Khurana, E. *et al.* Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* **342**, 1235587 (2013).

46. Pemberton, T.J. *et al.* Genomic patterns of homozygosity in worldwide human populations. *Am. J. Hum. Genet.* **91**, 275–292 (2012).

47. Bentley, D.R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).

48. Li, Y. *et al.* Structural variation in two human genomes mapped at single-nucleotide resolution by whole genome de novo assembly. *Nat. Biotechnol.* **29**, 723–730 (2011).

49. Lam, H.Y.K. *et al.* Performance comparison of whole-genome sequencing platforms. *Nat. Biotechnol.* **30**, 78–82 (2012).

50. Audic, S. & Claverie, J.M. The significance of digital gene expression profiles. *Genome Res.* **7**, 986–995 (1997).