

A hybrid approach for *de novo* human genome sequence assembly and phasing

Yulia Mostovoy¹, Michal Levy-Sakin¹, Jessica Lam¹, Ernest T Lam², Alex R Hastie², Patrick Marks³, Joyce Lee², Catherine Chu¹, Chin Lin¹, Željko Džakula², Han Cao², Stephen A Schlebusch⁴, Kristina Giorda³, Michael Schnall-Levin³, Jeffrey D Wall⁵ & Pui-Yan Kwok^{1,5,6}

Despite tremendous progress in genome sequencing, the basic goal of producing a phased (haplotype-resolved) genome sequence with end-to-end contiguity for each chromosome at reasonable cost and effort is still unrealized. In this study, we describe an approach to performing *de novo* genome assembly and experimental phasing by integrating the data from Illumina short-read sequencing, 10X Genomics linked-read sequencing, and BioNano Genomics genome mapping to yield a high-quality, phased, *de novo* assembled human genome.

The completion of the human genome reference assembly in 2003 marked a major milestone in genome research. The reference human genome sequence (and the genome sequences of numerous other organisms) and the sequencing technologies developed for the Human Genome Project revolutionized biological research and hastened the discovery of causal mutations for many diseases^{1,2}. Despite tremendous progress, a haplotype-resolved genome sequence with end-to-end contiguity for each chromosome that can be produced at reasonable cost and effort is still needed. Whole-genome sequencing has produced tens of thousands of genomes that are collections of short-read sequences aligned to the composite reference human genome sequence produced from several donors of various ethnic backgrounds. Similarly, *de novo* assemblies of other species generally consist of a set of scaffolds which may (or may not) have been mapped onto chromosomes³. Structural variants, especially those larger than a few thousand bases or those embedded in repetitive elements, are almost impossible to identify with short-read sequencing⁴. In addition, the parental chromosomes for these genomes are not resolved, so one cannot determine if variants that may affect gene function are on the same haplotype. A number of structurally complex regions of the genome are involved in disease syndromes (e.g., DiGeorge syndrome and Williams syndrome) or common disorders (e.g., HLA-region-associated disorders), but elucidation of the causal mutations is hampered by the difficulties in characterizing the sequence and structure of these regions. To produce high-quality genome sequence assemblies, one has to overcome

three challenges: (1) the confounding effects of the long, repetitive sequences of nearly 100% sequence identity that are present in most higher eukaryotic genomes, (2) the diploid nature of the DNA source, and (3) the lack of low-cost sequencing platforms that produce accurate, long DNA sequences.

Recent advances have addressed these challenges in part⁵. Single-molecule long-read sequencing methods from Oxford Nanopore Technologies and Pacific Biosciences (PacBio) are continually improved in terms of their read lengths, throughput, and assembly capabilities^{6–10}, but their per-base-pair sequencing costs and error rates are still much higher than those of standard short-read sequencing. Illumina's synthetic long-read technology (formerly known as Moleculo) showed promising results^{11,12} but has not been adequately compared with other methods. Putnam and colleagues¹³ described a method for generating mate pairs through proximity ligation of *in vitro* reconstituted chromatin that dramatically increased N50 scaffold sizes of *de novo* assemblies. Their approach is available as a service through Dovetail Genomics (<http://dovetailgenomics.com>). Several methods have been proposed for both phasing and improving the connectivity of assemblies. These methods, which are not yet commercially available, include fosmid paired-end sequencing¹⁴, pooled fosmid sequencing^{15–18}, and contiguity-preserving transposase sequencing (CPT-seq)^{19,20}.

A successful *de novo* assembly project combined Illumina short-read sequencing, PacBio sequencing, and BioNano Genomics (BNG) genome mapping to produce a phased assembly of HapMap sample NA12878 with an N50 scaffold size of 31.1 Mb and an N50 contig size greater than 880 kb²¹. One major drawback of this method is the relatively high cost and low throughput of PacBio sequencing.

In this study, we describe a new approach to perform *de novo* genome assembly and experimental phasing that is similar to the method of Pendleton and colleagues²¹ but with 10X Genomics (10XG) "Linked-Read" data rather than PacBio data used for the medium-length contiguity information. We demonstrate the feasibility of this approach by performing *de novo* genome assembly and phasing of a human HapMap sample (NA12878).

¹Cardiovascular Research Institute, University of California, San Francisco, San Francisco, California, USA. ²BioNano Genomics, Inc., San Diego, California, USA.

³10X Genomics, Inc., Pleasanton, California, USA. ⁴Department of Molecular and Cell Biology, University of Cape Town, Cape Town, South Africa. ⁵Institute for Human Genetics, University of California, San Francisco, San Francisco, California, USA. ⁶Department of Dermatology, University of California, San Francisco, San Francisco, California, USA. Correspondence should be addressed to P.-Y.K. (pui.kwok@ucsf.edu).

RESULTS

Sequence assembly

An overview of our assembly strategy is shown in **Figure 1**. We started with a SOAPdenovo-based *de novo* assembly of Illumina sequence data from human HapMap sample NA12878, which had a contig N50 of 11.1 kb and a scaffold N50 of 590 kb after filtering for scaffolds that were at least 3 kb in length (**Table 1**). To order and orient these scaffolds into longer blocks, we obtained sequence data from libraries generated using the 10XG GemCode platform. In total, 97× barcoded sequence coverage of NA12878 was obtained and, after filtering as described in the Online Methods, the result consisted of ~480,000 barcoded pools, each of which contained an average of ~3 Mb of target DNA. Qualitatively, scaffolds that are physically near each other will co-occur in the same barcoded pools more often than expected by chance. By looking at the patterns of co-occurrence of reads from the same pool mapped onto the ends of scaffolds, we could identify and orient linked scaffolds. We used the program fragScaff²⁰ to achieve this scaffolding, which increased the scaffold N50 of our assembly to 7.0 Mb (**Table 1**), a 12-fold improvement.

Genome mapping and hybrid assembly

In parallel, we obtained an assembled sequence motif physical map of the NA12878 genome that was generated using the Irys System from BNG with map assembly N50 of 4.59 Mb. Scaffolding of the Illumina short-read assembly and BNG maps directly yielded a hybrid scaffold N50 length of 7.76 Mb. Because most of the small contigs were not incorporated into the scaffold, it contained 2.39 Gb of the Illumina assembly and 17% *N*-base gaps. However, when we combined the BNG physical map with

Table 1 | Summary of assembly statistics for human sample NA12878

Assembly	Total map length (Gb)	Number of scaffolds	Scaffold N50 (Mb)	Longest scaffold (Mb)
Illumina	2.79	14,047	0.59	5.57
10XG	2.81	5,697	7.03	37.9
BNG	2.93	1,079	4.59	26.6
Hybrid	2.86	170	33.5	99.96

Rows 1–4 correspond to the results from the initial *de novo* short-read-based assembly, the 10XG-scaffolded assembly, the BNG map assembly, and the final hybrid assembly, respectively. Statistics for the Illumina assembly were calculated after filtering for scaffolds that were at least 3 kb in length, since those served as input for the next step of the assembly.

the 10XG-scaffolded short-read *de novo* assembly (**Fig. 2**), the final hybrid assembly contained just 170 scaffolds (**Fig. 3a** and **Supplementary Fig. 1**), with an N50 size of 33.5 Mb (**Table 1**), representing an additional 4.8-fold improvement (and an overall 57-fold improvement relative to the initial Illumina assembly). The total length of the Illumina short-read-based assembly was 2.79 Gb, that of the 10XG-scaffolded assembly was 2.81 Gb, and that of the BNG-assembled mapping data was 2.93 Gb; the final hybrid assembly was 2.86 Gb in length.

Phasing of assembled scaffolds

Phasing of the hybrid assembled scaffolds was performed using the freely available Long Ranger software developed by 10XG, in which single base variants in linked reads are strung together into haplotype blocks. Where copy-number variations across repetitive regions could not be resolved with linked reads, the BNG maps were used to resolve the haplotypes based on long single molecules spanning the regions. Phasing was done with respect to the *de novo* assembly of this study, yielding phase blocks of up to 23 Mb in size, with a median phase block size of 4.7 Mb and 2.8 million single-nucleotide variations (SNVs) phased (97.2%, **Table 2**, **Fig. 3b**, and **Supplementary Table 1**).

Assessment of assembled scaffold

The contiguity and accuracy of the final assembly was assessed and compared with that published by Pendleton *et al.*²¹ and the ALLPATHS-LG assembly²² (**Table 2**). Assembly accuracy, as measured by the position and orientation of sequences separated by 100 kb in our assembly compared with the reference genome, was 95.2%, comparable with the Pendleton *et al.* assembly²¹ and more accurate than the ALLPATHS-LG assembly²². To further assess the accuracy of the final assembly, we compared exon content to hg38 and found that 95.7% of all exons were fully present in the new assembly. 14.3 Mb of sequence in the current assembly was not found in the hg38 reference genome sequence, representing the difference between NA12878 and the reference.

The final phased assembly was further assessed in two ways. First, two complex regions of the genome, the major

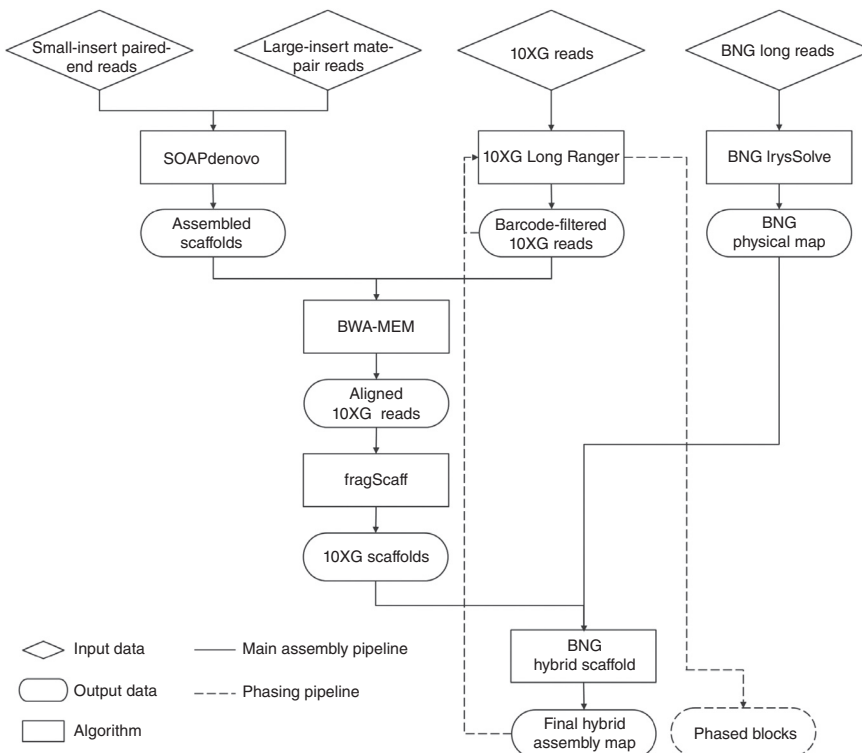


Figure 1 | Flowchart depicting genome sequence assembly strategy.

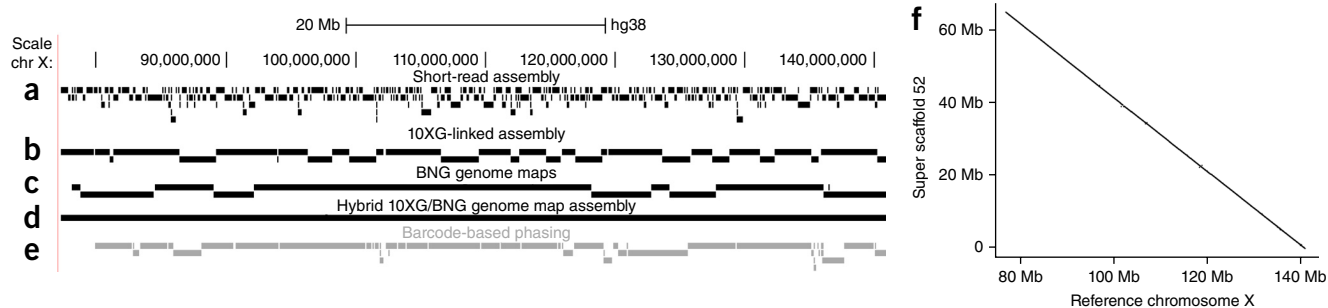


Figure 2 | Schematic of the relative sizes of scaffolds produced during the assembly process. A schematic from the UCSC Genome Browser showing the relative sizes of the scaffolds obtained during each assembly step and haplotype blocks for an example hybrid scaffold (super scaffold 52, covering 64 Mb of chromosome (chr) Xq) is shown: (a) assembly based on short-read Illumina data filtered for scaffolds longer than 3 kb; (b) short-read assembly scaffolded using barcode information from 10XG data; (c) assembled BNG genome maps; (d) hybrid scaffold produced by merging b and c; (e) barcode-based haplotype blocks for this region; (f) dot plot of the region against reference genome hg38.

histocompatibility complex (MHC) and the amylase regions, were analyzed and the phased assemblies with the two haplotypes resolved (Supplementary Fig. 2). Second, we confirmed that our assembly detected the 10p inversion polymorphism (Supplementary Fig. 3) and the large 17q21.31 inversion polymorphism²³, as well as the other structural variants previously identified in this sample (Supplementary Data 2)²⁴.

DISCUSSION

In this proof-of-principle study, we did not attempt to close all the *N*-base gaps in the scaffolds, so the number of sequence contigs is based on the *de novo* short-read assembly and remains large. With two minor optimizations, the *de novo* genome assemblies will likely be even better than the pilot results we have presented. First, the contig and scaffold N50 lengths can be improved by several-fold if a larger range of insert sizes is used (e.g., 250–800 bp for paired-end libraries and 2–15 kb for mate-pair libraries).

Second, the sequence data from 10XG libraries can be used to extend contigs and/or fill in the gaps between neighboring contigs. A purpose-built local assembler is being developed to use uniquely placed barcoded reads to recruit neighboring contigs that share reads of the same barcode, thereby joining the contigs and filling in the gaps.

While our hybrid assembly approach is efficient and cost effective, there are three limitations. First, the genome maps and 10XG Linked-Read sequencing require the use of long DNA molecules, so high molecular DNA preparation (usually from cells) is needed. Most archival DNA samples prepared with commercial kits consist of relatively short DNA fragments (~50 kb) and are therefore not useful for genome mapping or Linked-Read sequencing. Second, Linked-Reads are produced by random *k*-mer amplification of the 50- to 100-kb molecules present in the small partitions. As such, there are times when these molecules are not completely amplified, and so the barcoded sequencing reads will not cover the entire molecule.

When the gaps fall in repetitive regions not uniquely covered by the routine short-read contigs, they result in *N*-base gaps whose lengths are defined based either on the proportion of barcodes shared between flanking regions (if scaffolded by 10XG data) or the distance between nicks (if scaffolded by BNG data). Third, to minimize the number of *N*-base gaps, multiple sequencing libraries of various insert sizes will have to be prepared and sequenced, adding to the work involved. Fortunately, these libraries can be prepared in parallel

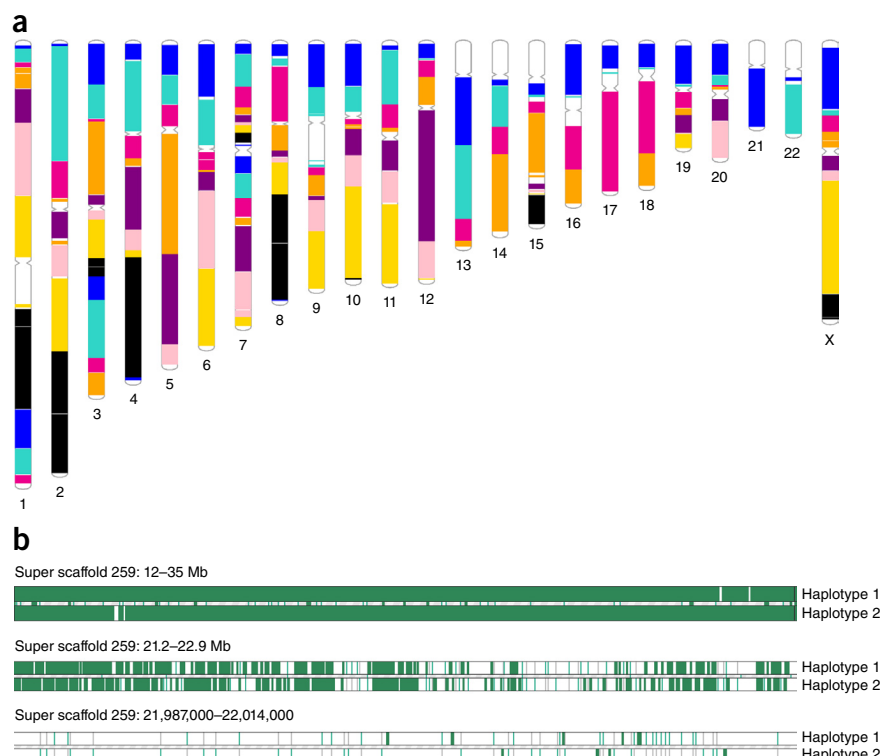


Figure 3 | Alignment and phasing of the hybrid assembly. (a) Ideograms of the hybrid scaffold assembly aligned to the reference genome hg38, with each colored block representing an assembled scaffold. (b) A 23-Mb phase block (super scaffold 259, aligned to chromosome 3 region 50–73 Mb) at increasing resolution showing the alleles on the two haplotypes (green vertical line, assembly allele; gray vertical line, alternate allele). Where a green or gray vertical line is not matched with a corresponding mark, the allele is indeterminate on that haplotype.

Table 2 | Comparison with other NA12878 assemblies

Input data	This study	Ref. 21	ALLPATHS-LG ²²
	Illumina paired-end and mate-pair reads; 10XG reads; BNG genome maps	PacBio reads; BNG genome maps	Illumina paired-end, mate-pair, and fosmid-based short reads
Scaffold N50 (Mb)	33.5	31.1	11.5
Number of scaffolds	170	202	23,634
Assembly length (Gb)	2.86	2.76	2.78
Validity at 100 kb (%)	95.2	97.5	93.5
N content (%)	10.2	4.61	5.90
Phase block N50	4.7 Mb	145 kb	N/A
Phased SNVs	2,783,119	2,421,740	N/A

and sequenced together in multiplex, resulting in just a small increase in effort, time, and cost.

Whole-genome short-read sequencing has been done on tens of thousands of individuals with various diseases, but the genome information obtained thus far is incomplete because short-read sequences only allow researchers to most accurately identify single-nucleotide variants and small insertion-deletions. More challenging are the structural variations that can disrupt genes and/or their regulatory elements as well as haplotype information. With this proof-of-principle study, we have shown that these limitations can be overcome by using three complementary sets of mapping-sequencing data that can be generated in parallel in a short time by an average laboratory at reasonable cost.

METHODS

Methods and any associated references are available in the [online version of the paper](#).

Accession codes. NCBI Sequence Read Archive, BioProject PRJNA315896: SRX1675529, SRX1675530 and SRX1675531.

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

ACKNOWLEDGMENTS

This work was supported in part by R01 HG005946 (P.-Y.K.). The DNA sample was obtained from the Coriell Institute for Medical Research, and the Illumina sequence data were obtained from the US National Institute of Standards and Technology (NIST). We thank the expert sequencing staff at the Institute for Human Genetics at UCSF for generating some of the sequencing data.

AUTHOR CONTRIBUTIONS

P.-Y.K., J.D.W., and Y.M. conceived the project and provided resources and oversight for sequencing and algorithmic analysis. K.G. prepared long libraries for 10XG GemCode sequencing. C.C. and C.L. performed long DNA preparation and BNG genome mapping experiments. E.T.L., A.R.H., Ž.D., J.Lee, and H.C. built initial genome maps and performed BNG alignment and structural variant calling. Y.M. and J.Lam performed scaffold analysis. E.T.L., A.R.H., and J.Lee performed hybrid genome assembly. P.M., K.G., and M.S.-L. performed scaffold phasing. Y.M., M.L.-S., E.T.L., J.Lam, J.Lee, and S.A.S. performed validation and quality measure analyses of the assembled data. Y.M., E.T.L., M.L.-S., and P.-Y.K. primarily wrote the manuscript and revisions, though many coauthors provided edits and Online Methods sections.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the [online version of the paper](#).

Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>.

- Wheeler, D.A. & Wang, L. From human genome to cancer genome: the first decade. *Genome Res.* **23**, 1054–1062 (2013).
- Duncan, E., Brown, M. & Shore, E.M. The revolution in human monogenic disease mapping. *Genes* **5**, 792–803 (2014).
- Li, R. *et al.* The sequence and *de novo* assembly of the giant panda genome. *Nature* **463**, 311–317 (2010).
- Tattini, L., D'Aurizio, R. & Magi, A. Detection of genomic structural variants from next-generation sequencing data. *Front. Bioeng. Biotechnol.* **3**, 92 (2015).
- Cao, H. *et al.* *De novo* assembly of a haplotype-resolved human genome. *Nat. Biotechnol.* **33**, 617–622 (2015).
- Quick, J., Quinlan, A.R. & Loman, N.J. A reference bacterial genome dataset generated on the MinION™ portable single-molecule nanopore sequencer. *GigaScience* **3**, 22 (2014).
- Goodwin, S. *et al.* Oxford Nanopore sequencing, hybrid error correction, and *de novo* assembly of a eukaryotic genome. *Genome Res.* **25**, 1750–1756 (2015).
- Landolin, J. *et al.* Initial *de novo* assemblies of the *D. melanogaster* genome using long-read PacBio sequencing. *55th Annual Drosophila Research Conference, San Diego* (2014).
- Huddleston, J. *et al.* Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Res.* **24**, 688–696 (2014).
- Chaisson, M.J.P. *et al.* Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608–611 (2015).
- Voskoboinik, A. *et al.* The genome sequence of the colonial chordate, *Botryllus schlosseri*. *eLife* **2**, e00569 (2013).
- McCoy, R.C. *et al.* Illumina TruSeq synthetic long-reads empower *de novo* assembly and resolve complex, highly-repetitive transposable elements. *PLoS One* **9**, e106689 (2014).
- Putnam, N.H. *et al.* Chromosome-scale shotgun assembly using an *in vitro* method for long-range linkage. *Genome Res.* **26**, 342–350 (2016).
- Williams, L.J.S. *et al.* Paired-end sequencing of Fosmid libraries by Illumina. *Genome Res.* **22**, 2241–2249 (2012).
- Kitzman, J.O. *et al.* Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nat. Biotechnol.* **29**, 59–63 (2011).
- Suk, E. *et al.* A comprehensively molecular haplotype-resolved genome of a European individual. *Genome Res.* **21**, 1672–1685 (2011).
- 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).
- Lo, C. *et al.* On the design of clone-based haplotyping. *Genome Biol.* **14**, R100 (2013).
- Amini, S. *et al.* Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. *Nat. Genet.* **46**, 1343–1349 (2014).
- Adey, A. *et al.* *In vitro*, long-range sequence information for *de novo* genome assembly via transposase contiguity. *Genome Res.* **24**, 2041–2049 (2014).
- Pendleton, M. *et al.* Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat. Methods* **12**, 780–786 (2015).
- Gnerre, S. *et al.* High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci. USA* **108**, 1513–1518 (2011).
- Steinberg, K.M. *et al.* Structural diversity and African origin of the 17q21.31 inversion polymorphism. *Nat. Genet.* **44**, 872–880 (2012).
- Mak, A.C. *et al.* Genome-wide structural variation detection by genome mapping on nanochannel arrays. *Genetics* **202**, 351–362 (2016).

ONLINE METHODS

Sequence data for NA12878. Barcoded reads were obtained from two different 10XG libraries: library 1 was prepared by 10XG and sequenced in two batches, one at 10XG and the other at the UCSF core sequencing facility; library 2 was fully prepared and sequenced by 10XG. Samples were processed as described in ref. 25. The first library had a median insert size of 176 bp and was sequenced to 58× coverage, while the second had a median insert size of 193 bp and was sequenced to a depth of 39× coverage. BNG physical mapping data were obtained from a previous study²⁴.

Initial *de novo* genome assembly. An Illumina short-read assembly for NA12878 was obtained from <http://sjackman.github.io/abyss-scaffold-paper/>; this assembly was generated using SOAPdenovo²⁶ with a 39× short-insert-size library and a 24× large-insert-size library with inserts of 2.5–3.5 kb.

10X Genomics-based scaffolding. The 10XG libraries were processed by trimming the first 10 bp of the first mate of each pair, as recommended by the company. Reads with barcodes that did not match the company's barcode whitelist were filtered out, as were those barcodes that were seen below a given threshold frequency (22 for library 1 and 101 for library 2, based on the lowest frequency among the number of barcodes that were detected in these libraries by 10XG's Long Ranger software), resulting in 231,022 barcodes retained for library 1 and 247,781 barcodes retained for library 2. Barcodes from the two different libraries were distinguished by flags appended to the barcodes.

The libraries were mapped to the short-read-based assembly using BWA-MEM²⁷ with default settings. The resulting alignments for each library were merged into one BAM file and filtered for reads that aligned to scaffolds that were at least 3 kb in size, as recommended for fragScaff²⁰ because shorter scaffolds are difficult to correctly assemble with the barcode-based approach. This alignment file was used as input for scaffolding using fragScaff²⁰ along with an *N*-base bed file and a repeat bed file produced by self-against-self blastn according to the fragScaff recommendations and processed using scripts distributed with fragScaff. For fragScaff processing stages 1 and 2, default parameters were used with the exception that -C 10 was set for stage 2. For the third stage, various combinations of fragScaff parameters *j*, *u*, and/or *p* were evaluated based on the resulting assembly contiguity and accuracy relative to the reference genome as well as the quality of the resulting hybrid assembly (**Supplementary Table 2**). To check the fragScaff assembly accuracy, we mapped the initial short-read-based scaffolds to each chromosome of the hg38 reference genome using Lastz²⁸ with the following settings: --nogapped --notransition --exact=200 --identity=95 --seed=match15 --twins=1.100 --ambiguous=n --match=1,5 --masking=3. The first alignment for a given scaffold was selected as its position on that chromosome. To resolve cases where a given scaffold mapped to more than one chromosome, the chromosome with the longest combined alignment for that scaffold was selected; scaffolds with ties among chromosomes were discarded, as were scaffolds where the longest single alignment was to a different chromosome than its longest combined alignment. This filtered position and orientation data were input into the fragScaff_checkOrdering.pl script that is distributed with fragScaff to evaluate the accuracy of different assemblies with respect to the reference. Using this information

as well as the contiguity of the hybrid assemblies using each set of parameters, we settled on the parameters of *j* = 1 and *u* = 3 to achieve a relatively conservative assembly and to minimize the introduction of scaffolding errors.

BioNano physical maps and hybrid scaffolding. The assembly produced by fragScaff was *in silico* digested with the nicking enzyme Nt.BspQI. This *in silico* map was scaffolded together with the BNG assembly of NA12878²⁴ (see **Supplementary Data 1**) using BNG's Hybrid Scaffold tool as performed previously²⁴ with the following adaptations. Merging *p*-value threshold was more stringent, $1e^{-13}$. Where there were divergent structures, suggesting chimeric assemblies, genome-map chimeric score, which is based on single-molecule support across the junction, was used to choose the better path, and the unsupported path was cut. Where there were smaller inconsistencies, genome map structure was maintained. This and all other software tools used in this study are freely available to all researchers (see **Supplementary Table 3**).

Phasing. Phasing was performed by 10XG with respect to the hybrid assembly with the Long Ranger software using a third library with 63× coverage of NA12878, similar to a previously published NA12878 trio phasing analysis²⁹.

Final assembly visualization and validation. To visualize the stages of the assembly for hybrid scaffold 52 (**Fig. 2a–d**), the constituent scaffolds from each stage (short-read based, 10XG scaffolded, and BNG scaffolded) were aligned to reference chromosome X using Lastz²⁸ with RepeatMasker (<http://www.repeatmasker.org>) softmasking and the following additional parameters: --format=sam --nogapped --notransition --exact=500 --maxwordcount=90% --identity=95 --seed=match15 --twins=1.100 --ambiguous=n --match=1,5. The resulting SAM file was converted to BAM format using Samtools³⁰ and then to BED format using Bedtools³¹. Multiple BED entries for a single scaffold were merged into a single entry. BNG genome maps were aligned to the reference using BNG's RefAligner tool, and the resulting alignments were converted to BED format. The final BED files were visualized in the UCSC Genome Browser³².

To visualize scaffold-to-reference concordance (**Fig. 2f**), scaffold 52 was aligned to reference chromosome X using Lastz²⁸ with the parameters listed above, and the output was plotted in rdotplot format. The resulting file was plotted with matplotlib³³. To visualize assembled scaffold alignments on chromosomes, IrysSolve RefAligner was used to find the best alignment matches between the final assembly and hg38 genome maps. Genome maps were built by *in silico* nicking with the enzyme Nt.BspQI. Alignment coordinates were plotted on chromosomes using PhenoGram (<http://ritchielab.psu.edu/software/phenogram-downloads>). Haplotype phasing and SV calls were visualized using 10XG Loupe software.

Exon assembly accuracy validation. Scaffolds were aligned to the hg38 chromosomes using Lastz²⁸ (version 1.03.73). A subset of chromosomes (3, 4, 5, 6, 7, 9, 11, 12, 13, 14, 15, and 16) was used to estimate the number of assembled exons and genes using the Gencode database (version 22). From these alignments, we determined the proportion of each chromosome that was

assembled. These assembled portions were then compared to known gene and exon features from the Gencode database (version 22) using Bedtools³¹.

25. Zook, J.M. *et al.* Extensive sequencing of seven human genomes to characterize benchmark reference materials. Preprint at <http://biorxiv.org/content/early/2015/12/23/026468> (2015).
26. Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *GigaScience* **1**, 18 (2012).
27. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* [1303.3997v2](https://arxiv.org/abs/1303.3997v2) (2013).
28. Harris, R.S. *Improved Pairwise Alignment of Genomic DNA* PhD thesis, Pennsylvania State Univ. (2007).
29. Zheng, G.X.Y. *et al.* Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat. Biotech.* **34**, 303–311 (2016).
30. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
31. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
32. Kent, W.J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
33. Hunter, J.D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.