Published in final edited form as:

Nat Methods. 2016 July ; 13(7): 587–590. doi:10.1038/nmeth.3865.

A Hybrid Approach for de novo Human Genome Sequence Assembly and Phasing

Yulia Mostovoy[1], **Michal Levy-Sakin**[1], **Jessica Lam**[1], **Ernest T Lam**[2], **Alex R Hastie**[2], **Patrick Marks**[3], **Joyce Lee**[2], **Catherine Chu**[1], **Chin Lin**[1], **Željko Džakula**[2], **Han Cao**[2], **Stephen A. Schlebusch**[4], **Kristina Giorda**[3], **Michael Schnall-Levin**[3], **Jeffrey D. Wall**[5], and **Pui-Yan Kwok**[1,5,6]

[1]Cardiovascular Research Institute, University of California, San Francisco, San Francisco, CA, USA

[2]BioNano Genomics, Inc., San Diego, CA, USA

[3]10X Genomics, Inc., Pleasanton, CA, USA

[4]Department of Molecular and Cell Biology, University of Cape Town, Cape Town, South Africa

[5]Institute for Human Genetics, University of California, San Francisco, San Francisco, CA, USA

[6]Department of Dermatology, University of California, San Francisco, San Francisco, CA, USA

Abstract

Despite tremendous progress in genome sequencing, the basic goal of producing 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 a new approach to perform *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

Correspondence should be addressed to P.Y.K. (Pui.Kwok@ucsf.edu).

Accession codes. Sequencing and assembly data are available under BioProject PRJNA315896 with Sequence Read Archive accession numbers: SRX1675529, SRX1675530 and SRX1675531.

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. DŽ., J. Lee, and H.C. built initial genome maps and performed BNG alignment and SV 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 methods sections.

Competing Financial Interests Statement

E.T.L., A.R.H., J. Lee, Ž. DŽ., H.C. are employees of BioNano Genomics. P.M., K.G., M.S.L. are employees of 10X Genomics, and P.Y.K. is on the scientific advisory board of BioNano Genomics.

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, the basic goal of producing phased (haplotype-resolved) genome sequence with end-to-end contiguity for each chromosome at reasonable cost and effort is still unrealized. Consequently, researchers who engage in human "whole-genome sequencing" have 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 chromosomes[3]. Structural variants, especially those larger than a few thousand bases or embedded in repetitive elements, are almost impossible to identify with short-read sequencing[4]. 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., the HLA region), but the elucidation of the causal mutations are 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) long repetitive sequences of close to 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.

To meet these three challenges, the original plan of the Human Genome Project was to sequence a set of tiling large-insert clones (yeast artificial chromosomes (YACs) initially and bacterial artificial chromosomes (BACs) subsequently) to separate the parental chromosomes and provide DNA fragments of manageable size for sequence assembly. If a BAC clone harbors repetitive elements that cannot be bridged by short-read sequences, it can be analyzed by generating medium size insert clones (such as fosmids, etc.). However, this labor-intensive stepwise approach was later superseded by shotgun sequencing of BAC clones without further subcloning to resolve complex, repetitive regions. As a result, numerous gaps and unresolved repetitive regions were present in the original reference human genome sequence assembly, though most have since been resolved with additional efforts. Because of the complex genomic features present, there are no shortcuts to producing high-quality eukaryotic genome sequence assemblies. One must separate the parental chromosomes and overcome the long repetitive sequences present in many regions[5]. In other words, the original approach of the Human Genome Project is still sound, but it is too costly and requires too much effort.

Recent advances provide several potential paths forward in this regard. Single-molecule long-read sequencing methods from Oxford Nanopore Technologies and Pacific Biosciences (PacBio) continue to improve their read lengths, throughput, and assembly capabilities[6–10], but with per basepair sequencing costs and error rates still much higher than those of standard short-read sequencing. Illumina's synthetic long-read technology (formerly Moleculo) showed promising results[11,12] but has not been compared much with other methods. Putnam and colleagues[13] described a method for generating mate-pairs by 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 were proposed for both phasing and improving the connectivity of assemblies but are not yet commercially available, including fosmid paired-end sequencing[14], pooled fosmid sequencing[15–18], and contiguity preserving transposase sequencing (CPT-seq)[19,20].

So far, the most successful *de novo* assembly project using commercially available methods 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 26 Mb and an N50 contig size greater than 880 kb[21]. The motivation for this approach was that genome mapping with long DNA fragments of hundreds of kilobases would replace the laborious cloning, mapping and tiling of BACs, while PacBio long-read sequences would replace the need for medium size insert clones for resolving the structure of local repetitive elements, and short-read sequences would provide accurate base-calling for single-base resolution of the final genome assembly. One major drawback to their method is the high cost and low throughput of PacBio sequencing. While this cost is coming down and throughput is increasing with PacBio's new Sequel instrument, it is likely to remain considerably more expensive and to take substantially longer to perform the sequencing experiments than routine short-read sequencing for the foreseeable future.

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 colleagues[21] but with 10X Genomics (10XG) "Linked-Read" data rather than PacBio data used for the medium-length contiguity information. Briefly, the 10XG platform uses microfluidics to generate several hundred-thousand emulsified droplets each containing a small fraction (~0.1%) of the human genome in molecules ranging from tens to hundreds of kb in size, a gel bead bearing barcoded oligos ending with a random k-mer, and biochemistry reagents for creating a barcoded sequencing library from within each droplet. The libraries produced within each droplet are combined for sequencing library preparation in the usual manner before they are sequenced on an Illumina sequencer. Because each partition contains a very small fraction of the genome, the likelihood that homologous regions on the two parental chromosomes are found in the same droplet is extremely low. Therefore, sequences from any nearby region bearing the same barcode are almost certainly from the same DNA molecule. Tabulating short read sequences bearing the same barcode produces "Linked-Reads" of tens to hundreds of kb. Our approach of integrating the data from Illumina short-read sequencing, 10XG Linked-Reads, and BNG genome mapping yields high-quality, phased, *de novo* assemblies using commercially available products at a fraction of the cost of comparable approaches. Here we demonstrate the feasibility of this approach by performing *de novo* phased genome assembly and phasing of a human HapMap sample (NA12878).

RESULTS

Sequence assembly

An overview of our assembly strategy is shown in Fig. 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). Next, to order and orient these scaffolds into longer blocks, we obtained sequence data from libraries generated using the 10XG GemCode platform. In total, 97X barcoded sequence coverage of NA12878 was obtained and, after filtering as described in Methods, 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 can identify and orient linked scaffolds. We used the program fragScaff[20] to achieve this scaffolding, which increased the scaffold N50 of our assembly to 7.0 Mb (Table 1), representing 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[24], 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 the 10XG-scaffolded short-read *de novo* assembly (Fig. 2), the final hybrid assembly contained just 170 scaffolds (Fig. 3a, 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, the 10XG-scaffolded assembly was 2.81 Gb, and 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 Long Ranger software developed by 10XG (and freely available to any researcher) where 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 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 assembly[22] (Table 2). Assembly accuracy, as measured by the position and orientation of sequences separated by 100 kb in our assembly compared to the reference genome, was 95.2%, comparable with the Pendleton *et al.* assembly[21] and more accurate than the ALLPATHS-LG assembly[22]. 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 histocompatibility complex (MHC) and 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 polymorphism[23], as well as the other structural variants previously identified in this sample (Supplementary information) [24].

DISCUSSION

We have demonstrated how high-quality *de novo* sequence scaffolds can be assembled using a combination of commercially available technologies from Illumina, 10X Genomics and BioNano Genomics. The quality of our assembly is comparable with the results of previous approaches[21,22], but the sequencing and mapping data can be produced in one week and at a much lower cost.

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 produced by our approach 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 “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, the 10XG “Linked-Reads” are produced by random k-mer amplification of the 50–100 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 length is defined based either on the proportion of shared barcodes 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 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 allow one to identify mostly single nucleotide variants and small

insertion/deletions. Missing 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 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. Based on our experience, data generation on these commercially available platforms can be done in less than one week and at reasonable cost.

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 58X coverage, while the second had a median insert size of 193 bp and was sequenced to a depth of 39X coverage. BNG physical mapping data was obtained from a previous study[24].

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 SOAPdenovo[26] with a 39X short insert library and a 24X 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-MEM[27] 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 fragScaff[20] because shorter scaffolds are difficult to correctly assemble with the barcode-based approach. This alignment file was used as input for scaffolding using fragScaff[20], 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 Lastz[28] 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 was 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 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 NA12878[24] (see Supplementary information) using BNG's Hybrid Scaffold tool as performed previously[24] with the following adaptations: Merging p-value threshold was more stringent, $1e^{-13}$. Where there were divergent structures, suggesting chimeric assemblies, genome map chimera 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 63X coverage of NA12878, similar to a previously published NA12878 trio phasing analysis[29].

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 Lastz[28] 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 Samtools[30] and then to BED format using Bedtools[31]. 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 Browser[32].

To visualize scaffold-to-reference concordance (Fig. 2f), scaffold 52 was aligned to reference chromosome X using Lastz[28] with the parameters listed above, and output in rdotplot format. The resulting file was plotted with matlotlib[33]. 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 Lastz[28] (v 1.03.73). A subset of chromosomes (3, 4, 5, 6, 7, 9, 11, 12, 13, 14, 15 and 16) were used to estimate the number of assembled exons and genes using the Gencode database (v22). 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 Genncode database (v22) using Bedtools[31].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

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.

References

1. Wheeler DA, Wang L. From human genome to cancer genome: the first decade. Genome Res. 2013; 23:1054–1062. [PubMed: 23817046]
2. Duncan E, Brown M, Shore EM. The revolution in human monogenic disease mapping. Genes. 2014; 5:792–803. [PubMed: 25198531]
3. Li R, et al. The sequence and de novo assembly of the giant panda genome. Nature. 2010; 463:311–317. [PubMed: 20010809]
4. Tattini L, D'Aurizio R, Magi A. Detection of genomic structural variants from next-generation sequencing data. Front Bioeng Biotechnol. 2015; 3:92. [PubMed: 26161383]
5. Cao H, et al. De novo assembly of a haplotype-resolved human genome. Nat Biotechnol. 2015; 33:617–622. [PubMed: 26006006]
6. Quick J, Quinlan AR, Loman NJ. A reference bacterial genome dataset generated on the MinION portable single-molecule nanopore sequencer. GigaScience. 2014; 3:22. [PubMed: 25386338]
7. Goodwin, S., et al. Oxford Nanopore sequencing and de novo assembly of a eukaryotic genome. 2015. bioRxivhttp://dx.doi.org/10.1101/013490
8. 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, CA, USA. 2014.
9. Huddleston J, et al. Reconstructing complex regions of genomes using long-read sequencing technology. Genome Res. 24:688–696. [PubMed: 24418700]
10. Chaisson MJP, et al. Resolving the complexity of the human genome using single-molecule sequencing. Nature. 2015; 517:608–611. [PubMed: 25383537]
11. Voskoboynik A, et al. The genome sequence of the colonial chordate, Botryllus schlosseri. eLife. 2013; 2:e00569. [PubMed: 23840927]
12. McCoy RC, et al. Illumina TruSeq synthetic long-reads empower de novo assembly and resolve complex, highly-repetitive transposable elements. PLoSOne. 2014; 9:e106689.

13. Putnam NH, et al. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. Genome Res. 2016; 26:342–50. [PubMed: 26848124]
14. Williams LJS, et al. Paired-end sequencing of Fosmid libraries by Illumina. Genome Res. 2012; 22:2241–2249. [PubMed: 22800726]
15. Kitzman JO, et al. Haplotype resolved genome sequencing of a Gujarati Indian individual. Nat Biotechnol. 2011; 29:59–63. [PubMed: 21170042]
16. Suk E, et al. A comprehensively molecular haplotype-resolved genome of a European individual. Genome Res. 2011; 21:1672–1685. [PubMed: 21813624]
17. Duitama J, et al. Fosmid-based whole genome haplotyping of a HapMap trio child: evaluation of single individual haplotyping techniques. Nucl Acids Res. 40:2041–2053. [PubMed: 22102577]
18. Lo C, et al. On the design of clone-based haplotyping. Genome Biol. 2013; 14:R100. [PubMed: 24028704]
19. Amini S, et al. Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. Nat Genet. 46:1343–1349. [PubMed: 25326703]
20. Adey A, et al. In vitro, long-range sequence information for de novo genome assembly via transposase contiguity. Genome Res. 2014; 24:2041–2049. [PubMed: 25327137]
21. Pendleton M, et al. Assembly and diploid architecture of an individual human genome via single-molecule technologies. Nat Methods. 2015; 12:780–786. [PubMed: 26121404]
22. Gnerre S, et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci USA. 2011; 108:1513–1518. [PubMed: 21187386]
23. Steinberg KM, et al. Structural Diversity and African Origin of the 17q21.31 Inversion Polymorphism. Nat Genet. 2012; 44:872–880. [PubMed: 22751100]
24. Mak AC, et al. Genome-wide structural variation detection by genome mapping on nanochannel arrays. Genetics. 2016; 202:351–362. [PubMed: 26510793]
25. Zook, JM., et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. 2015. bioRxivhttp://dx.doi.org/10.1101/026468
26. Luo R, et al. SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. Gigascience. 2012; 1:18. [PubMed: 23587118]
27. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013 arXiv 1303.3997v2.
28. Harris, RS. ProQuest. 2007. Improved pairwise alignment of genomic DNA.
29. Zheng GXY, et al. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. Nat Biotechnol. 2016
30. Li H, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009; 25:2078–2079. [PubMed: 19505943]
31. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26:841–842. [PubMed: 20110278]
32. Kent WJ, et al. The human genome browser at UCSC. Genome Res. 2002; 12:996–1006. [PubMed: 12045153]
33. Hunter JD. Matplotlib: A 2D graphics environment. Computing in Science & Engineering. 2007; 9:90–95.

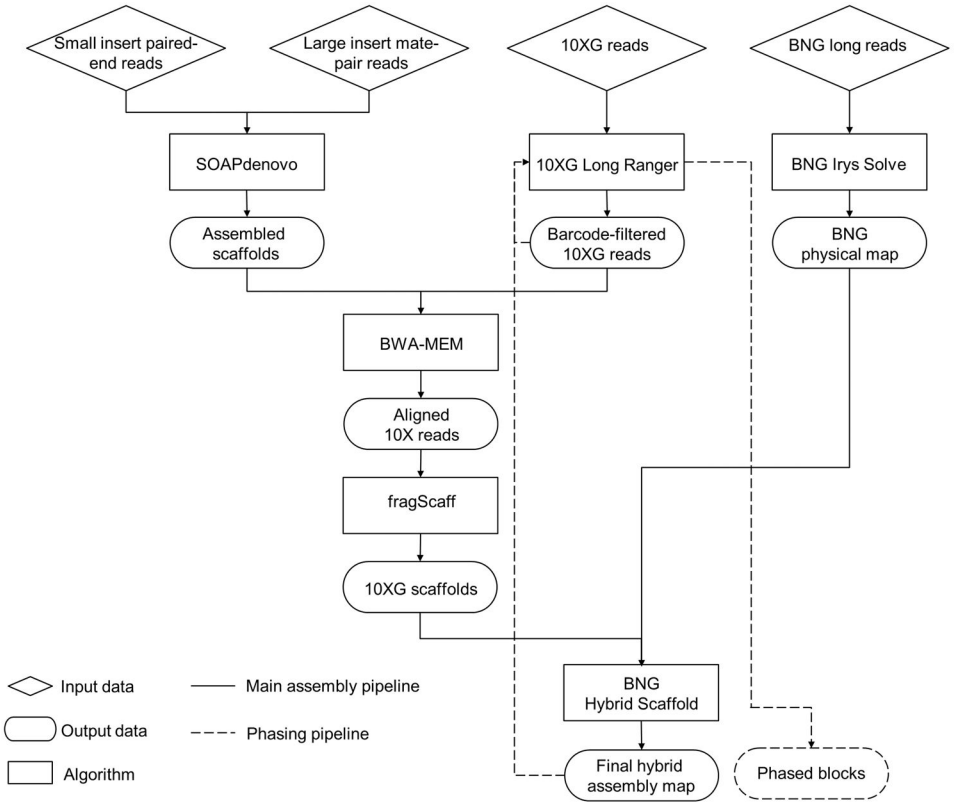

Figure 1.
Flowchart depicting genome sequence assembly strategy.

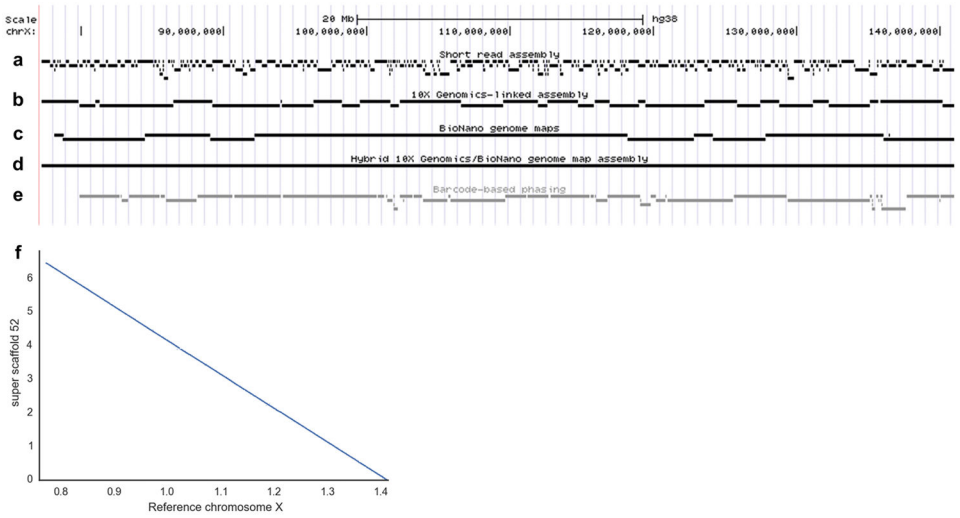

Figure 2. Schematic from the UCSC Genome Browser showing the relative sizes of scaffolds produced during each step of the assembly process, as well as haplotype blocks, for the hybrid scaffold (64 Mb) aligned to the q arm of reference chromosome X

(**a**) Assembly based on short-read Illumina ata filtered for scaffolds longer than 3 kb; (**b**) the short-read assembly scaffolded together 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.

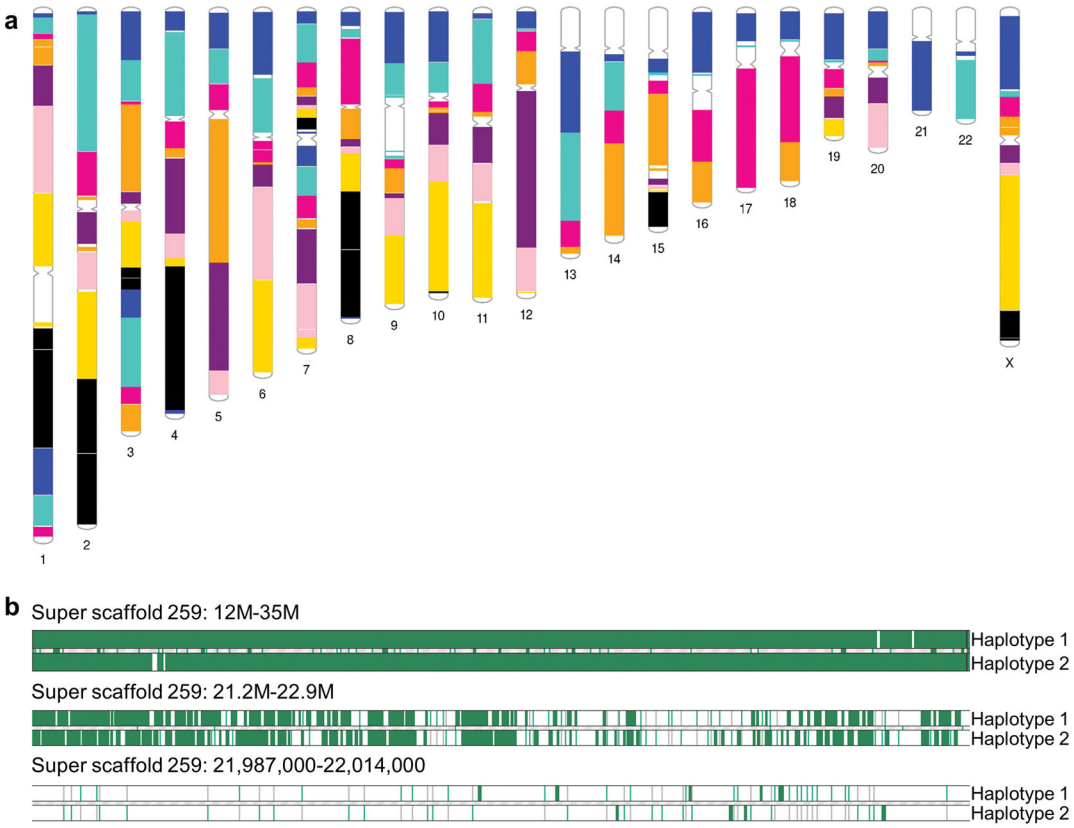

Figure 3. Alignment and phasing of the hybrid assembly

(**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 Chr 3 50 Mb-73 Mb) at increasing resolution showing the alleles on the two haplotypes (green vertical line: assembly allele; grey vertical line, alternate allele). Where a green or grey vertical line is not matched with a corresponding mark, the allele is indeterminate on that haplotype.

Table 1

Summary of assembly statistics for human sample NA12878

The different rows 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 to the next step of the assembly.

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

Table 2

Comparison with other NA12878 assemblies.

	This study	Ref 21	ALLPATHS-LG[22]
Input data	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 100kb (%)	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