SCIENTIFIC REPORTS
nature research

OPEN

High throughput barcoding method for genome-scale phasing

David Redin[1], Tobias Frick[1], Hooman Aghelpasand[1], Max Käller[1], Erik Borgström[1], Remi-Andre Olsen[2] & Afshin Ahmadian[1*]

The future of human genomics is one that seeks to resolve the entirety of genetic variation through sequencing. The prospect of utilizing genomics for medical purposes require cost-efficient and accurate base calling, long-range haplotyping capability, and reliable calling of structural variants. Short-read sequencing has lead the development towards such a future but has struggled to meet the latter two of these needs. To address this limitation, we developed a technology that preserves the molecular origin of short sequencing reads, with an insignificant increase to sequencing costs. We demonstrate a novel library preparation method for high throughput barcoding of short reads where millions of random barcodes can be used to reconstruct megabase-scale phase blocks.

Elucidating the true impact of genetic variation and its potential contribution to healthcare requires a complete characterization of the human genome. While massive efforts and technological developments have been made towards this goal, the vast majority of whole genome sequencing data produced has been limited to profiles of unphased nucleotide variations. High throughput short read sequencing has become the backbone of genomics as a field, yet generating a haploid consensus rather than a haplotype-resolved genome has limited our ability to associate genetic variation with health and disease[1]. Deriving phenotypes from more than just profiles of SNVs (single nucleotide variations), by investigating structural variants, gene fusion events, and the cumulative effects of mutations across long distances is likely to be greatly beneficial. It is estimated that more than half of human genomic variation is constituted by structural variants[2,3] in the form of deletions, insertions, inversions, duplications and translocations, and studies have shown such events to have a larger effect on gene expression than SNVs[4]. Identifying such variants all but equates to a need for long-range haplotype information so differences between maternal and paternal alleles can be resolved[2,3]. The importance of long-range phasing information is exemplified by the characterization of compound heterozygosity being essential for diagnosis of recessive Mendelian diseases. Furthermore, delving deeper into the genetic basis of elaborate phenotypes have shown structural variants to be drivers of cancers and complex diseases[5,6].

Despite the promise of new-found insights for medical genomics, the adoption of whole genome haplotyping technologies have been limited by high costs due to specialized instruments and platform-dependent reagents. Assays based on dilution of genomic fragments in discrete compartments, combined with compartment-specific barcoding, have been proven effective for obtaining long-range phasing information by linking reads that share a common barcode[7–10]. Such technologies are able to utilize the high throughput and accuracy of short read sequencing platforms while maintaining the long-range information crucial for haplotyping. One technology in particular[7], has in recent years established itself as the foremost alternative for whole genome haplotyping, but its reliance on microfluidic equipment and barcoded beads have limited its scalability and flexibility. Furthermore, the high cost of preparing libraries for sequencing has notably limited widespread adoption of this technology. Sequencing based on 'contiguity preserving transposition on beads'[11], offers an alternative solution for genome-wide haplotyping with a single tube reaction setup. Performing tagmentation of genomic fragments on uniquely barcoded beads, rather than in discrete compartments, opens up the potential for automated library preparation in the future. However, an evident bottleneck of this technology is the laborious generation of a transposase-linked bead library with barcodes of sufficient complexity to resolve a human genome. Furthermore, a product that enables phasing of DNA molecules has not yet been made commercially available. Alternative platforms based on long read sequencing of single molecules[12,13] provide long-range phasing information, albeit with a lower sequencing accuracy and throughput than short read sequencing, rendering them unfit for the scale

[1]Royal Institute of Technology (KTH), School of Engineering Sciences in Chemistry, Biotechnology and Health, Department of Gene Technology, Science for Life Laboratory, SE-171 65, Solna, Sweden. [2]Stockholm University, Department of Biochemistry and Biophysics, Science for Life Laboratory, Box 1031, 171 21, Solna, Sweden. *email: afshin.ahmadian@scilifelab.se

required to haplotype human-sized genomes[14–16]. Ultimately, these platforms have mostly been used to provide longer scaffolding information for genome assembly, whilst still relying on short read sequencing for coverage[17,18].

A number of droplet-based barcoding strategies have been developed to enable high-throughput analysis of single molecules[10] or single cells[19–21], all utilizing microfluidic devices for droplet generation and/or uniquely barcoded beads to distinguish the contents in each droplet. In-house manufacturing of microfluidic systems has been the solution for many academic groups to avoid commercial devices for droplet generation which are expensive and typically constricted to particular assays. Although the hardware components for microfluidic systems are easy to obtain, the manufacturing of single-use microfluidic chips requires expertise and equipment that goes beyond what is available in most laboratories[10]. Likewise, the production of barcoded beads is not a trivial matter as highly complex libraries of unique barcodes are required to perform high-throughput analyses. The process of barcoding beads can for instance be done by clonal amplification in droplets[22], by combinatorial extension cycling[20] or by split-pool cycles of phosphoramidite synthesis[21]; but regardless of strategy it remains a costly and laborious pre-requisite for the intended assay reaction.

Here we describe a new method for whole genome DNA phasing based on previous work of barcoding DNA fragments in emulsion droplets formed by simple shaking[23]. The assay does not require microfluidic devices or complex libraries of barcoded beads, making it more scalable and affordable than alternative methods. Being free from these requirements also means libraries can be prepared in any laboratory setting, using readily available low-cost reagents (Supplementary Table S1) and without investments in platform-specific laboratory equipment. To demonstrate the integrity of the data produced we haplotype a complete human genome and investigate the potential for reference-free assembly.

First, a tagmentation reaction introduces a universal DNA sequence at arbitrary yet evenly distributed positions throughout the genome. Bead-linked transposases preserve the proximity of tagmented constituents from long DNA fragments, and by linking the template molecule(s) from each bead to a mutually exclusive barcode it enables the information of proximity to be conserved through DNA sequencing (Fig. 1). This is achieved by separating the beads into millions of discrete compartments (emulsion droplets) together with single copies of a barcoding oligonucleotide. The barcoding oligonucleotide features a semi-randomized sequence with an unrestricted complexity, ensuring the barcode present in each compartment is unique. Within each droplet, PCR amplification is used to first generate clonal populations of single stranded barcoding oligos, and then to couple the barcode to template molecules (Supplementary Fig. S1). As a consequence of limited dilution, droplets without either the barcoding or template molecules will be formed, but neither will yield coupled amplicons. Such products are removed from the library through a target-specific enrichment following breakage of the emulsion reaction. Libraries consisting of barcode-linked molecules are then sequenced using the standard Illumina short read platform and reads are grouped according to the barcode to reconstruct long-range haplotype information of the original fragment(s) (Supplementary Fig. S1).

Results

A library from the human 'genome in a bottle' (GIAB) reference individual GM24385 was generated to enable benchmarking of results against external haplotyping technologies and sequencing data. A total of 451 M sequencing read pairs were processed (Supplementary Table S2), leaving 321 M read pairs assigned to 2,204,497 barcoded read groups that were used as input for phasing analysis. With our barcode sequences translated to 10x barcodes (Methods), the Long Ranger pipeline yielded an N50 phase block length of 1,832,815 bp. The haplotype was resolved for 3,620,251 SNVs (97.9% of identified SNVs) with a mean sequencing depth of 19.1X and 0.7% of the reference genome not covered (Table 1). These figures are comparable to that of previously published transposition-based phasing data[11], in which figures of 98% SNVs phased and an N50 phase block length of 1.14 Mb are presented for a comparable coverage of 19.2X. Figure 2a shows that we obtain an even coverage across the whole genome and for both haplotypes, with discrepancies contained to heterochromatic chromosome regions. The software calculated that 74.8% and 18.6% of input molecules were over 20 Kb and 100 Kb, respectively (Table 1). Variant calling of constructed phase blocks yielded 29 large structural variant (LSV) calls (Fig. 2a, Supplementary Table S3) and 4,008 short deletion calls. Based on sequenced bases and mapping coordinates, we estimate the coupling efficiency of our assay to be 74.4% (Supplementary Note). We also ran a duplicate experiment, showcasing these statistics are reproducible (Supplementary Table S4).

Comparing SNVs to the GIAB 'ground truth' callset for GM24385[24,25] yielded an accuracy of 95.4% and accounted for 76.1% of the SNVs detected with less than 10-fold of the sequencing depth (Table 1). To exemplify the clinical relevance of phasing, our data shows all HLA family genes have their haplotypes resolved. Out of 29 large structural variant calls, 28 (96.6%) were validated through manual review of correspondence to a GIAB (10x Genomics, 42X) dataset (Supplementary Table S3). Called structural variants consist of inversions, deletions and duplication events across the genome, varying from 40 kb to 1.2 Mb, three of which are visualized in Fig. 2b,c.

In addition, we complemented the GM24385 library with an additional library and increased the coverage to 34.7X to ensure variant calling and phasing scaled with higher sequencing depths. With 540 M sequencing read pairs as input, the analysis resulted in 98.8% of SNVs phased and phase blocks up to 11.9 megabases (N50 phase block length of 2,812,019 bp). An additional 310,000 more SNVs were discovered, leveling the sensitivity difference to the benchmarking GIAB (10x Genomics, 42X) dataset while maintaining higher accuracy (Table 1). Structural variant calling was in agreement with the dataset of lower sequencing depth, resulting in 29 LSVs and 4,047 short deletions. The added reads also enabled evaluation of the method's potential for reference-free assembly wherein the data was run through the Supernova pipeline. This generated an assembly with a total length of 3.20 Gb (ungapped length of 2.68 Gb) where scaffolds covered 2.65 Gb (85.4%) of the GRCh38 reference genome (heterochromatic regions not excluded, Supplementary Fig. S2). For context, scaffolding based on barcode linkages of reads increased the N50 from 24.4 kb to 5.60 Mb whereof the longest contig spanned 47.9 Mb.

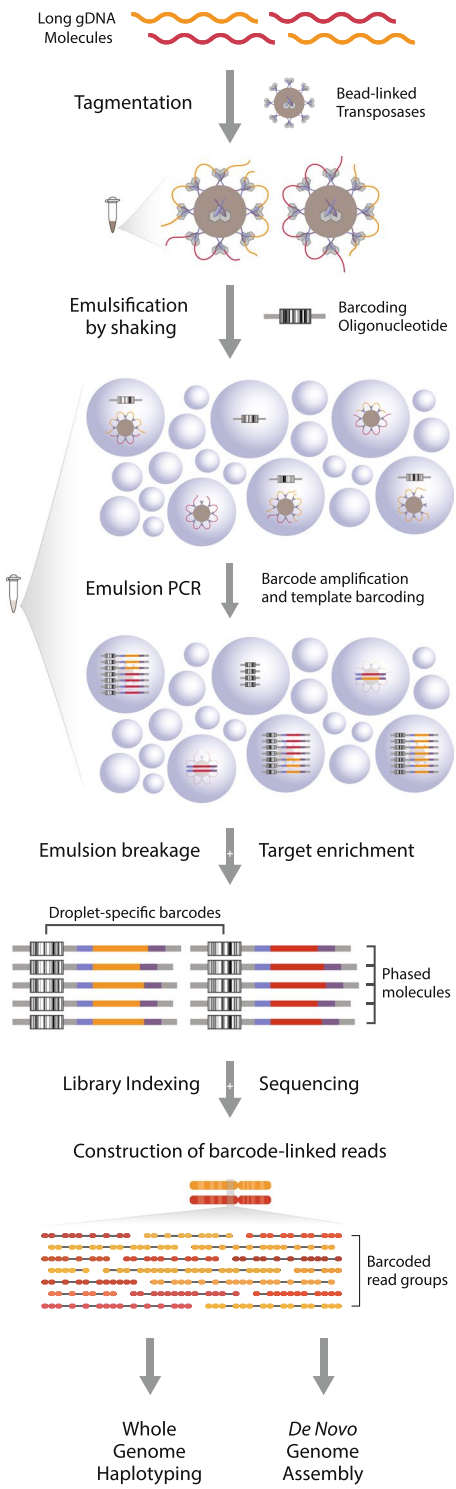

Figure 1. Overview of the phasing technology. High molecular weight DNA fragments are diluted and tagmented with bead-linked transposases. DNA-loaded beads are put into emulsion droplets with barcoding oligonucleotides and primers for amplification, and the constituents of each original molecule is coupled to a unique barcode sequence through emulsion PCR. Following the removal of uncoupled molecules, the library undergoes standard short read sequencing and subsequent grouping of reads according to the barcode sequence. The resulting barcode-linked reads are utilized for long-range DNA phasing, genome-wide haplotyping or reference-free genome assembly.

Library	GM24385 (19X)	GM24385 (35X)	GIAB (10x Genomics)
Sequencing reads	641,457,522	1,080,294,792	976,557,530
Mean depth	19.1 X	34.7 X	41.7 X
SNPs Phased	97.9%	98.8%	98.50%
N50 Phase Block	1,832,815 bp	2,812,019 bp	9,657,460 bp
Longest Phase Block	7,771,012 bp	11,919,151 bp	35,805,844 bp
Mean Molecule Length	25,946 bp	26,780 bp	104,745 bp
Molecules >20 kb	74.8%	74.8%	92.2%
Molecules >100 kb	18.6%	18.6%	44.7%
LSV Calls*	35	35	35
Short Deletion Calls	4,008	4,047	4,383
Median Insert Size	231 bp	234 bp	308 bp
Mapped Reads	82.7%	89.6%	96.2%
Zero Coverage	0.735%	0.507%	0.178%
Q30 bases, Read 1	82.4%	86.2%	100%**
Q30 bases, Read 2	64.2%	69.3%	100%**
SNV Calls (Q > 60)	3,445,072	3,804,691	4,054,372
SNV Detection Sensitivity	76.1%	83.3%	85.1%
SNV Detection Accuracy	95.4%	94.6%	90.7%

Table 1. Phasing analysis and variant calling for internal 19X and 35X datasets, as well as for the 42X dataset from 10x Genomics. *Large structural variant (LSV) calls featured multiple heterozygous deletions calls in chromosome X, which following correspondence with 10x Genomics were confirmed as erroneous (Supplementary Table S3). **Figures indicate an undisclosed pre-filtering of sequencing reads Q < 30 for the dataset from 10x Genomics. For all datasets, SNV detection sensitivity and accuracy was calculated by comparing to the GIAB 'ground truth' callset for GM24385 for SNVs with phred score >60. The reference dataset contained 4,756,689, of which 4,319,399 SNVs had a minimum phred score (Q) of 60. See Supplementary Table S5 for raw SNV counts.

Discussion

Geneticists are recognizing that short read sequencing by itself is not sufficient to resolve the connection between genetic variation and common aspects of health and disease[26]. Expanding on the capability of short read sequencing platforms to generate vast amounts of high quality data, by resolving haplotypes and calling structural variants, presents a pivotal change in genomics that enables more accurate reconstruction of genomes[27,28]. It is clear that preserving the contiguity of short sequences is currently the most affordable strategy for obtaining haplotyping information on a human genome-wide scale. We have described a novel library preparation method for barcode-linked reads that enables whole genome haplotyping with insignificant increases to sequencing costs, substantially less than the most established alternative[7]. Furthermore, as the assay does not require an investment in a platform-specific instrument it can be implemented in any laboratory with a thermal cycler and performed in a single day (Supplementary Fig. S3). Featuring amplification from single barcode molecules gives the method significant advantages since barcode degeneracy is essentially unlimited. This, in combination with high throughput and a low per-reaction cost makes it possible to run several parallel reactions where the number of input molecules can be tuned to yield libraries with single molecule barcoding resolution. Such a technology provides the next stepping stone towards a future based on reference-free genomics studies[29].

Our results show that we can phase up to 99% of called SNVs in the human genome with an accuracy of 95% and generate phase blocks with an N50 of 2.8 Mb. Analysis of data from a single lane of sequencing identified 29 large structural variants in the human genome, of which 28 could be independently verified (96.6%). Additionally, we showcase a *de novo* application where scaffolds cover at total of 2.65 Gb of a state-of-the-art 3.10 Gb human genome reference (GRCh38) from a modest sequencing depth of 35X.

The presented method offers laboratories all over the world the benefit of adding long-range phasing information to short read sequencing, through a simple protocol independent of specialized equipment and expensive reagents. We recognize that droplets formed by shaking results in a non-uniform, albeit controllable[23], size distribution compared to microfluidic chips. Though if microfluidic devices are readily available the proposed chemistry would be fully compatible. Regardless, the results in this study show that microfluidic devices are not necessary for producing high quality phased data. As the field seeks to draw upon the advantages of long-range haplotyping information, standard practices for extracting DNA will need to shift towards more meticulous protocols aimed at maintaining the integrity of large genomic fragments for phasing. Unlike long read sequencing platforms, an advantage of our assay is that there is no inherent bias in the length of fragments that can be phased.

We have developed a flexible and scalable solution for barcoding long molecules for short read sequencing and present it applied in whole genome haplotyping and *de novo* sequencing. The use of single barcode molecules in an efficient high throughput setting means the assay can be tailored according to the size or complexity of the genome and the resolution to which a biological study would require. These key elements makes future prospects of this technology include reference-free assembly of complex metagenomic samples and haplotype-resolved genomes from single cells in simple workflows with cost-efficient library preparation. For the purpose of

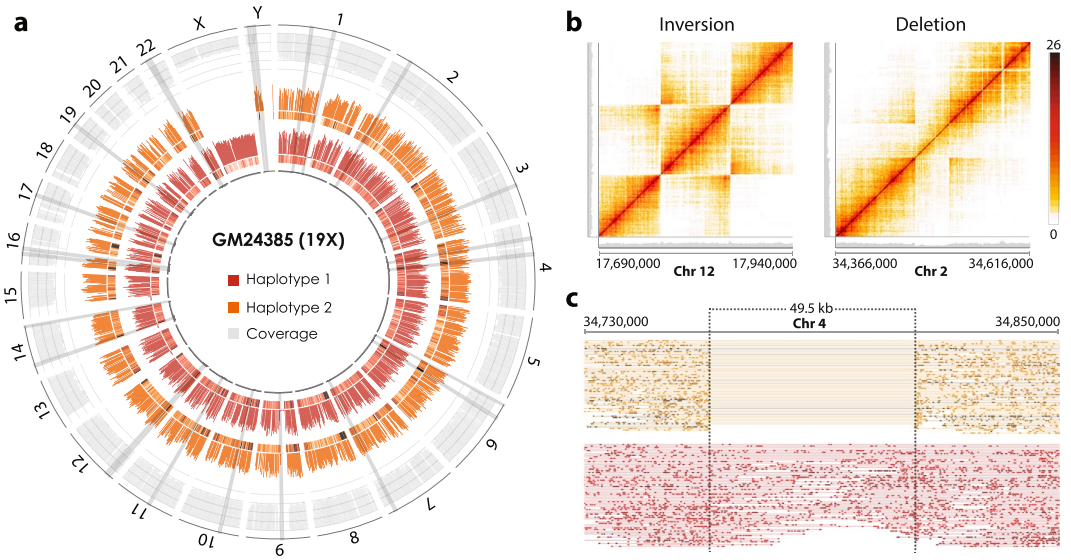

Figure 2. Whole genome haplotyping results. (**a**) Sequence data of haplotype-resolved human genome, GM24385 (19X). From center, phased SNV density and relative read coverage for haplotype 1 (red), phased SNV density and relative read coverage for haplotype 2 (orange), total read coverage (light grey) on a scale from 0 to 25X. The localization of large structural variants is visualized by bands in grey (not drawn to scale). (**b**) Heatmap of barcode overlap for reads spanning called structural variants with a window size of 250 kb, an 86.0 kb inversion in chromosome 12 (left) and a 40.8 kb heterozygous deletion in chromosome 2 (right). Relative read coverage collapsed for the two haplotypes shown in grey, x and y axis are identical. (**c**) Barcode-linked reads of a heterozygous deletion identified in chromosome 4, with reads assigned to either haplotype, and where the reads on each line share a mutually exclusive barcode. Reads in the top haplotype (shown in orange) are linked across the deletion spanning 49.5 Kb.

expanding our understanding of population diversity and individual variance, the next frontier for large-scale genomics ought to be *de novo* and haplotype-resolved genome analyses[30]. The rise of long read sequencing and linked-read platforms show that more and more researchers are realizing the benefits of long-range phasing information. The method proposed within offers a unique opportunity for researchers to tackle the hurdles of *de novo* sequencing and genome-wide haplotyping without being limited by a lack of resources. Combined with the continued reduction in short read sequencing costs, the need for an affordable library preparation that maximizes the yield of medically relevant information is evident. In the near future, there will simply be no room for library preparation assays that cost hundreds of dollars per sample when the cost of sequencing the human genome will be reduced to a fraction of that.

Methods

Sample preparation. A sample of human gDNA from GIAB (Genome in a Bottle) individual GM24385 was obtained from the 10x Chromium Genome Kit (Control gDNA). On-bead tagmentation of DNA was performed using Nextera DNA Flex Library Prep (Illumina) according to the manufacturer's reference guide for tagmentation; except that each reagent was scaled down to 15% of the specified volumes to reduce the amount of BLT (bead-linked transposases) used per reaction. For both libraries, a total of 1 ng HMW gDNA was added to the on-bead tagmenentation reaction. Rather than amplification of tagmented DNA, the polymerase amplification was prepared without the addition of Nextera Flex indexes, and beads were subjected to incubation at 68 degrees for 10 min instead of the specified PCR cycling protocol. The beads were subsequently washed twice with the supplied TWB reagent and then resuspended in 5 ml Elution Buffer (Qiagen) prior to emulsification.

Reaction emulsification and retrieval. Assay reactions consist of 50 µl PCR reagents that are mixed and added on top of emulsion oil before shaking for emulsification. The PCR volumes consist of 5 µl beads with tagmented DNA (see section above), 1x Phusion Flex Master Mix (New England Biolabs), 1 M Betaine (Sigma Aldrich), 3%vol DMSO (Thermo Scientific), 2%wt PEG-6000 (Sigma Aldrich), 2%vol Tween-20 (Sigma Aldrich), 400 nM Enrichment Oligo, 80 nM Coupling Oligo and 330 fM Barcoding Oligo (see Supplementary Table S6 for oligonucleotide sequences; purchased from Integrated DNA Technologies). Emulsification is carried out by pipetting the 50 µl PCR volume on top of 100 µl HFE-7500 oil with 5%(w/V) 008-Fluorosurfactant (Ran Biotechnologies) in a Qubit tube (Life Technologies) and shaking at 14.0 Hz for 8 min using a Tissuelyser instrument (Qiagen). Emulsion reactions were then left to stand upright for 15 min to settle, the excess oil was removed from the bottom and the remaining emulsion phase was transferred to a PCR tube with 60 µl FC-40 oil with 5%(w/V) 008-Fluorosurfactant (Ran Biotechnologies) already added to it. 85 µl of mineral oil (Sigma Aldrich) was then added on top of the emulsion reactions, before placing the reaction tubes in a Mastercycler Pro S (Eppendorf) instrument for reaction cycling with the following protocol: 5 min at 95 °C, 30 cycles of [95 °C for

30 s–55 °C for 30 s–72 °C for 30 s], followed by 8 cycles of [95 °C for 1 min–40 °C for 2 min–72 °C for 5 min] and ending the protocol with 10 min at 72 °C and holding at 12 °C. A ramp speed of 3% was used to ramp between temperatures of 40 °C and 72 °C. Following emulsion PCR, the mineral oil was discarded with a pipette and 4 µl EDTA (100 mM) was added. The entire emulsion reaction and excess emulsion oil was transferred to a 0.5 ml tube (Eppendorf) and 100 µl 1 H,1 H,2 H,2H-Perfluoro-1-octanol (Sigma Aldrich) was added before vortexing at maximum speed. After centrifugation for 1 min at 20,000 g, the aqueous phase was collected from the top and a magnetic rack was used to discard the beads.

Sample enrichment and sequencing library preparation. Following retrieval of aqueous phases from emulsion reactions the library preparation was continued by a bead-based purification to remove short and uncoupled barcoding amplicons below 200 bp using sample purification beads included in the Nextera Flex kit. Biotinylated and barcode-coupled molecules were enriched for by washing and incubating the sample with 20 µl DynaBeads MyOne Streptavidin T1 beads (Life Technologies) in B&W buffer (1 M NaCl, 5 mM Tris-HCl, 500 mM EDTA) for 30 min under rotation at room temperature. The supernatant was then discarded and the beads washed twice with Elution Buffer, four times with NaOH (0.125 N), and finally two more times with Elution Buffer. An indexing PCR was then performed on the washed enrichment beads in 1x Phusion Flex Master Mix with 400 nM Indexing Oligo; using a protocol starting with 2 min at 95 °C, 5 min at 55 °C (with 10% ramp speed), 10 min at 72 °C (with 3% ramp speed), and 1 min at 95 °C. At this point, the PCR reaction was paused and placed on a magnetic rack (heated to 80 °C), and the supernatant was quickly transferred to a fresh PCR tube. The i5 Adapter Oligo (Supplementary Table S6) was then added to a final concentration of 400 nM to the reaction and the PCR indexing protocol was continued by running 4 cycles of [95 °C for 30 s–55 °C for 30 s–72 °C for 1 min], followed by 2 min at 72 °C. Reactions were then purified and samples were quantified by Qubit (Life Technologies). Samples were diluted to 2 nM and sequenced using the HiSeq X platform (Illumina) with 150 bp paired-end sequencing and an 8 bp single-end index.

Data analysis overview. The analysis pipeline combines custom written python scripts, commonly used bioinformatics tools and three pipelines for sequencing reads linked by barcodes, used depending on the sample type and purpose as detailed below and outlined in Supplementary Fig. S4. Samtools (v1.7) and pysam (v0.14) were used extensively in in-house developed scripts[31]. All parts of the pipeline are available through GitHub (https://github.com/FrickTobias/BLR).

Read trimming and barcode deconvolution. For all samples, sequencing reads were initially trimmed with cutadapt (v1.16)[32] using a similarity threshold of 20%, to remove universal handle sequences upstream of the barcode sequence, downstream of the barcode sequence, and downstream of genomic inserts (when applicable). Reads not following the structure and sequence of the universal handles were omitted from downstream analysis steps (Supplementary Table S2). The barcode sequence, matching a predefined design (Supplementary Fig. S1), were extracted from reads and divided into 8 subsets based on the first two bases, using bc_extract.py. These files were then individually clustered using CD-HIT-454 (v4.6)[33] with k-mer cutoff of 0.9, to deconvolute the barcoded read groups.

Human genome haplotyping and *de novo* assembly. Trimmed reads were first mapped using Bowtie2 (v2.3.4.1)[34] with GRCh38 as reference, thereafter read duplicates with the same barcode were called and removed using picard tools (v2.5.0) (http://broadinstitute.github.io/picard/). To identify and collapse barcode sequences originating from the same droplet, barcode-linked reads sharing at least two proximal (<100 Kb) read pairs were merged using cluster_rmdup.py. To exclude potentially erroneous phasing information from abnormally large droplets, the barcode sequence was stripped from all reads originating from droplets with >260 molecules using filter_Clusters.py. For this purpose, molecules were defined as barcode-linked reads whereby each read mapped no further than 30 kb from its closest neighbor. Reads along with corresponding barcodes were then converted in accordance to the input of the 10x Genomics analysis pipelines (www.10xgenomics.com) with wfa2tenx.py. To enable use of these pipelines, which feature a limit of 4.8 M barcodes, the complexity of our barcode population was reduced by stripping the barcode information from all barcoded read groups with less than 4 read pairs. For whole genome haplotyping analysis, phasing metrics were generated using the Long Ranger analysis pipeline (v2.2.2), run with GATK (v3.8) for SNV calling and a 10x GRCh38 blacklist to reduce false positive variant calls. For the *de novo* assembly, the Supernova[35] pipeline was run with the no-preflight option to allow for shorter insert sizes than the typical Chromium data and pseudohap as the output style.

Data validation and visualization. For the *de novo* assembly, the final scaffolds zero coverage and mismatch percentages were calculated based on metrics supplied from QUAST (4.6.4)[36], aligned using Minimap2 (v2.4)[37]. SNV calls were evaluated by comparing output variant files from the Long Ranger analysis to variant calls from the GIAB reference callset for GM24385[24,25] using VCFtools[38] and large structural variation calls (>30 kb) were compared manually (Supplementary Table S3, Supplementary Figs. S5 and S6) to variants identified in the resource dataset GIAB (10x Genomics, 42X)[24]. The GIAB resource dataset was downloaded and run through Long Ranger with the same version and parameters as previously described. In addition to the previously mentioned software, MultiQC (v1.6.dev0)[39], bedtools (v2.27.1) (Quinlan and Hall, 2010), Circos (v0.69.6)[40] and dotPlotly (https://github.com/tpoorten/dotPlotly) were used for calculations and data visualization.

Data availability

All sequencing data is available on the Sequence Read Archive (SRA) under BioProject PRJNA590332, BioSample SAMN13324933. External GIAB data used can be downloaded through GIABs homepage, for both the 10x Genomics dataset (42X) and the GIAB ground truth callset. The external 10x Genomics assembly (58X) can be found on 10x Genomics official homepage.

Received: 20 May 2019; Accepted: 13 November 2019;
Published online: 02 December 2019

References

1. Sudmant, P. H. *et al*. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75 (2015).
2. Huddleston, J. *et al*. Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome research* **27**, 677–685 (2017).
3. Huddleston, J. & Eichler, E. E. An incomplete understanding of human genetic variation. *Genetics* **202**, 1251–1254 (2016).
4. Chiang, C. *et al*. The impact of structural variation on human gene expression. *Nat. genetics* **49**, 692 (2017).
5. Feuk, L., Carson, A. R. & Scherer, S. W. Structural variation in the human genome. *Nat. Rev. Genet.* **7**, 85 (2006).
6. Moncunill, V. *et al*. Comprehensive characterization of complex structural variations in cancer by directly comparing genome sequence reads. *Nat. biotechnology* **32**, 1106 (2014).
7. Zheng, G. X. *et al*. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat. biotechnology* **34**, 303 (2016).
8. Amini, S. *et al*. Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. *Nat. genetics* **46**, 1343 (2014).
9. Peters, B. A. *et al*. Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature* **487**, 190 (2012).
10. Lan, F., Haliburton, J. R., Yuan, A. & Abate, A. R. Droplet barcoding for massively parallel single-molecule deep sequencing. *Nat. communications* **7**, 11784 (2016).
11. Zhang, F. *et al*. Haplotype phasing of whole human genomes using bead-based barcode partitioning in a single tube. *Nat. biotechnology* **35**, 852 (2017).
12. Clarke, J. *et al*. Continuous base identification for single-molecule nanopore dna sequencing. *Nat. nanotechnology* **4**, 265 (2009).
13. Eid, J. *et al*. Real-time dna sequencing from single polymerase molecules. *Science* **323**, 133–138 (2009).
14. Laver, T. *et al*. Assessing the performance of the oxford nanopore technologies minion. *Biomol. Detection quantification* **3**, 1–8 (2015).
15. Quail, M. A. *et al*. A tale of three next generation sequencing platforms: comparison of ion torrent, pacific biosciences and illumina miseq sequencers. *BMC genomics* **13**, 341 (2012).
16. Loman, N. J., Quick, J. & Simpson, J. T. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. methods* **12**, 733 (2015).
17. Koren, S. *et al*. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat. biotechnology* **30**, 693 (2012).
18. Pendleton, M. *et al*. Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat. methods* **12**, 780 (2015).
19. Zheng, G. X. *et al*. Massively parallel digital transcriptional profiling of single cells. *Nat. communications* **8**, 14049 (2017).
20. Klein, A. M. *et al*. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
21. Macosko, E. Z. *et al*. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
22. Borgström, E. *et al*. Phasing of single dna molecules by massively parallel barcoding. *Nat. communications* **6**, 7173 (2015).
23. Redin, D. *et al*. Droplet barcode sequencing for targeted linked-read haplotyping of single dna molecules. *Nucleic acids research* **45**, e125–e125 (2017).
24. Zook, J. M. *et al*. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci. data* **3**, 160025 (2016).
25. Zook, J. M. *et al*. Integrating human sequence data sets provides a resource of benchmark snp and indel genotype calls. *Nat. biotechnology* **32**, 246 (2014).
26. Chaisson, M. J. *et al*. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. communications* **10** (2019).
27. Church, D. M. *et al*. Extending reference assembly models. *Genome biology* **16**, 13 (2015).
28. Schneider, V. A. *et al*. Evaluation of grch38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome research* **27**, 849–864 (2017).
29. Bishara, A. *et al*. High-quality genome sequences of uncultured microbes by assembly of read clouds. *Nat. biotechnology* (2018).
30. Aleman, F. The necessity of diploid genome sequencing to unravel the genetic component of complex phenotypes. *Front. Genet.* **8**, 148 (2017).
31. Li, H. & Durbin, R. Fast and accurate short read alignment with burrows–wheeler transform. *bioinformatics* **25**, 1754–1760 (2009).
32. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* **17**, 10–12 (2011).
33. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. Cd-hit: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
34. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with bowtie 2. *Nat. methods* **9**, 357 (2012).
35. Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M. & Jaffe, D. B. Direct determination of diploid genome sequences. *Genome research* **27**, 757–767 (2017).
36. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. Quast: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
37. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
38. Danecek, P. *et al*. The variant call format and vcftools. *Bioinformatics* **27**, 2156–2158 (2011).
39. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. Multiqc: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
40. Krzywinski, M. *et al*. Circos: an information aesthetic for comparative genomics. *Genome research* **19**, 1639–1645 (2009).

Acknowledgements

The authors would like to thank the National Genomics Infrastructure (NGI) and UPPMAX for providing sequencing and computational support and infrastructure. This work was supported by the Erling Persson Family Foundation, and Olle Engkvist Foundation [2015/347]. Open access funding provided by Royal Institute of Technology.

Author contributions

A.A. and D.R. conceived the technology. D.R., T.F. and A.A. designed the experiments. T.F. developed the analysis pipeline, with support from E.B., R.O and M.K. D.R. performed all experiments, with support from H.A. R.O. produced the human *de novo* assembly. All authors contributed in writing the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-019-54446-x.

Correspondence and requests for materials should be addressed to A.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.