Dense and accurate whole-chromosome haplotyping of individual genomes

David Porubsky[1,#]**, Shilpa Garg**[2,3,4,#], **Ashley D. Sanders**[5,6,#], Jan O. Korbel[5], Victor Guryev[1], Peter M. Lansdorp[1,6,7], Tobias Marschall[2,3,*]

1. *European Research Institute for the Biology of Ageing, University Medical Center Groningen, University of Groningen, 9713 AV Groningen, The Netherlands*
2. *Center for Bioinformatics, Saarland University, Saarbrücken, Germany*
3. *Max Planck Institute for Informatics, Saarbrücken, Germany*
4. *Graduate School of Computer Science, Saarland University, Saarbrücken, Germany*
5. *European Molecular Biology Laboratory (EMBL), Heidelberg, Germany*
6. *Terry Fox Laboratory, BC Cancer Agency, Vancouver, BC V5Z 1L3, Canada*
7. *Department of Medical Genetics, University of British Columbia, Vancouver, BC, V6T 1Z4, Canada*

* Corresponding author: Tobias Marschall <t.marschall@mpi-inf.mpg.de>

\# These authors contributed equally to this work.

ABSTRACT

The diploid nature of the genome is neglected in many analyses done today, where a genome is perceived as a set of unphased variants with respect to a reference genome. Many important biological phenomena such as compound heterozygosity and epistatic effects between enhancers and target genes, however, can only be studied when haplotype-resolved genomes are available. This lack of haplotype-level analyses can be explained by a dearth of methods to produce dense and accurate chromosome-length haplotypes at reasonable costs. Here we introduce an integrative phasing strategy that combines global, but sparse haplotypes obtained from strand-specific single cell sequencing (Strand-seq) with dense, yet local, haplotype information available through long-read or linked-read sequencing. Our experiments provide comprehensive guidance on favorable combinations of Strand-seq libraries and sequencing coverages to obtain complete and genome-wide haplotypes of a single individual genome (NA12878) at manageable costs. We were able to reliably assign > 95% of alleles to their parental haplotypes using as few as 10 Strand-seq libraries in combination with 10-fold coverage PacBio data or, alternatively, 10X Genomics linked-read sequencing data. We conclude that the combination of Strand-seq with different sequencing technologies represents an attractive solution to chart the unique genetic variation of diploid genomes.

Human genomes are diploid and possess two copies of each chromosome – one paternal and one maternal copy. At the DNA sequence level, these two homologous copies differ at a number of loci along each chromosome. Such heterozygous variants include single nucleotide variants (SNVs), short indels, as well as larger structural variants like deletions, duplications or inversions that change the copy number or orientation of segments of the genome. Discriminating and phasing alleles to their respective parental homologue is valuable in many areas of human genetics. For instance, resolving haplotype structure is required to track inheritance in human pedigrees and populations[1], map regions of meiotic recombination[2,3], identify variant-disease associations[4], detect instances of compound heterozygosity, and study allele-specific events like DNA methylation or gene expression[5]. In particular, long-range haplotype information is needed to systematically study interactions between enhancers and their target genes. This is critical as many variants that have been linked to traits in genome-wide association studies reside in enhancers and super enhancers[6]. Enhancer-specific variants can show epistatic effects among one another[7], as well as with their target genes that are beyond the reach of linkage disequilibrium[8]. To better understand these interactions, we must move beyond merely locating variant alleles and additionally study their functional relationships over long distances. Constructing genome-wide chromosome-length haplotypes is therefore the clear next step to build a more complete picture of genome architecture and function.

Currently, methods used to chart unique variation of individual genomes rely largely on 2nd and 3rd generation DNA sequencing and can include specialized experimental protocols[9–13]. Sequencing technologies sample the human genome in the form of relatively short molecules (reads) and every read that spans at least two heterozygous variants can essentially be considered as a 'mini haplotype' that can be assembled into longer haplotype segments by partially overlapping reads spanning the same variable locus[4]. To this end, haplotype-informative reads need to be partitioned into two disjoint sets that represent the two haplotypes. This process, however, is complicated by errors in sequencing as well as genotyping. For these reasons assembling haplotypes directly from sequencing data is computationally challenging, and the resulting optimization problems are provenly hard[14,15]. Notwithstanding, a number of computational approaches for read-based phasing have recently been developed[16] and, particularly, progress on fixed-parameter tractable (FPT) algorithms has enabled solving read-based phasing in practice[17–19], for instance through the implementations available in the software package WhatsHap[20]. Beyond phasing reads aligned to a reference genome, various approaches for haplotype-resolved *de novo* assembly have been explored[21–25].

However, all approaches to reconstruct haplotypes from sequencing reads, be it reference-based or reference-free, come with the intrinsic limitation that the distance between subsequent heterozygous markers can be larger than the read length itself. While long-read sequencing (such as PacBio SMRT[26] and Oxford NanoPore MinION[27]), or linked read data (such as those provided by 10X Genomics[28]) help to mitigate this issue, these technologies still fail to phase over long stretches of homozygosity or centromeres. Instead, specialized techniques that enable homologous chromosomes to be discriminated are required to physically connect alleles across whole chromosomes[3,29,30]. As an alternative to whole chromosome separation, chromatin capture (Hi-C) methods[31] can be employed to infer long-range haplotype information, based on the assumption that a chromosome will be cross-linked to itself more often than to its homologue[13]. Recently, Hi-C data have been used in combination with other sequencing methods for long-range phasing[32,33]. However to generate a reliable long-range haplotype scaffold high sequence coverage (~100-fold) is required to reduce bias caused by crosslinks between non-homologues chromosomes[32]. In particular, because these haplotypes need to be inferred statistically, the probability that two heterozygous variants are correctly phased relative to each other deteriorates with increasing chromosomal distances.

For the first time, we introduce a strategy to obtain dense and global haplotypes that span centromeres, homozygosity regions and genome assembly gaps, while keeping error rates, costs and labor at minimum. To this end, we harness the long-range phasing information provided by single cell template strand sequencing (Strand-seq)[34]. Strand-seq is an effective method to assemble highly accurate chromosome-length haplotypes, albeit with lower density of phased alleles in comparison to read-based phasing[9]. Unlike other haplotyping methods, Strand-seq by design distinguishes parental homologues based on the directionality of single-stranded DNA. We emphasize that Strand-seq is therefore able to deliver global haplotypes: its ability to correctly phase two variants with respect to each other does not depend on their distance. To fully exploit this advantage, while at the same time generating dense haplotypes that contain virtually all heterozygous SNVs, we designed a novel unified statistical framework to combine Strand-seq data with short-read, long-read, or linked-read sequencing data. We demonstrate that the combination of long-range Strand-seq haplotypes successfully bridges partial Illumina, PacBio and 10X Genomics phased segments into contiguous and global haplotypes that span whole chromosomes. We offer extensive experimental guidance on favorable combinations of the number of used Strand-seq libraries and the depth of PacBio or Illumina coverage, and thus enable considerable reductions in costs and labor – yielding a novel, affordable and scalable approach for reconstruction of haplotype-resolved individual genomes.
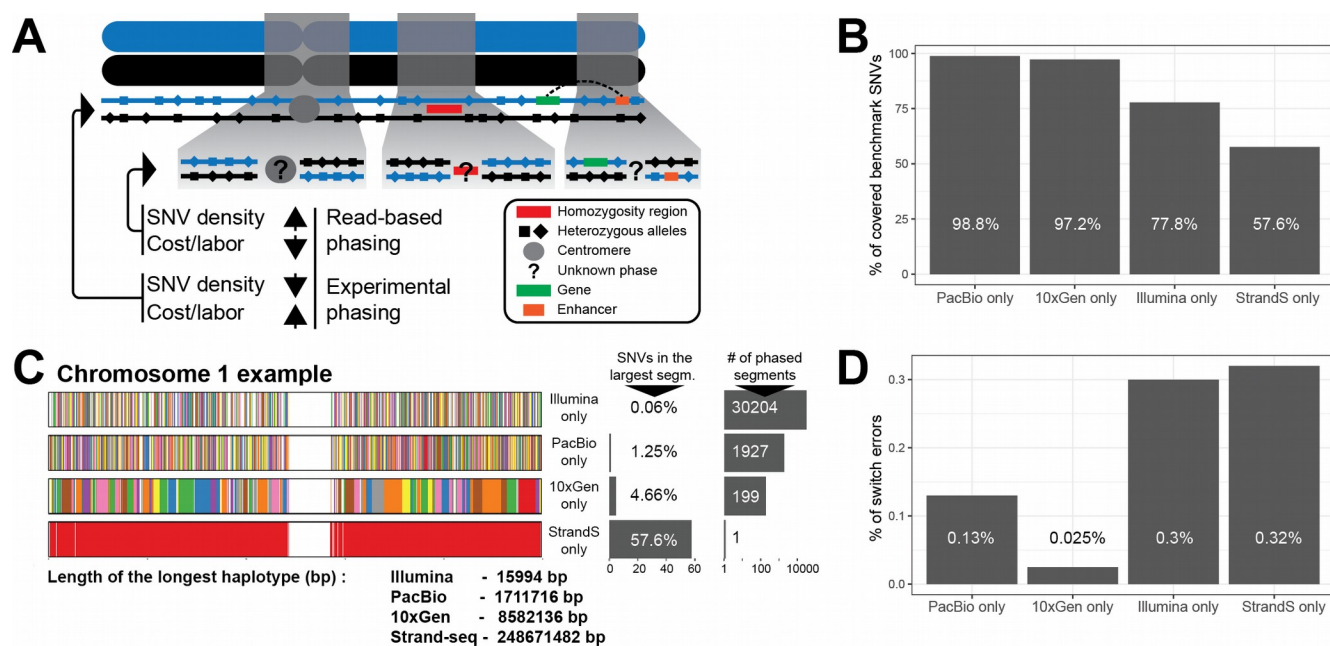
Figure 1: Phasing efficacy of read-based and experimental phasing approaches (example Chromosome 1).

A) Two homologous chromosomes are shown (blue and black). Experimental phasing approaches like Strand-seq can connect heterozygous alleles along whole chromosomes, however, at higher costs (time and labor) and lower density of captured alleles. In contrast, read-based phasing can deliver high-density haplotypes, but only short haplotype segments are assembled with an unknown phase between them. **B**) Barplot showing the percentage of phased variants, for each sequencing technology, from the total number of reference SNVs (Illumina platinum haplotypes). **C**) Graphical summary of phased haplotype segments for Illumina, PacBio, 10X Genomics and Strand-seq phasing shown for chromosome 1. Each haplotype segment is colored in a different color with the longest haplotype colored in red. Side bargraph reports the percentage of SNVs phased in the longest haplotype segment. **D**) Accuracy of each independent phasing approach measured as percentage short switch errors in comparison to benchmark haplotypes.

RESULTS

To explore a new integrative phasing strategy, with the aim of obtaining dense and accurate chromosome-length haplotypes, we used sequencing data available for a well-studied individual (NA12878). The NA12878 genome has been extensively sequenced using multiple technologies, providing high-coverage public sources of sequence information (see **Data Access**). In this study, we focused on read-based phasing data generated from Illumina and PacBio technologies, as they represent current standards for short- and long-read sequencing, respectively. The Illumina dataset was sequenced to an average depth of 49.5x coverage with a median insert size of 433bp, and the PacBio dataset was sequenced to 39.6x coverage with an average read length of ~15kb (**Supplemental Tab. S1**). In addition, we evaluated the performance of 10X Genomics, an emergent linked-read technology. Since none of these technologies alone provides chromosome-length haplotype information, we additionally incorporated single cell Strand-seq data[9], which has the capacity to scaffold haplotype information obtained from other data types (**Fig. 1A**). Here we used 134 single cell libraries sequenced to an average depth of 0.037x coverage per library using a paired-end sequencing protocol (see **Data Access** and **Supplemental Tab. S2**). To evaluate the phasing accuracy of haplotypes reported in this study, we used the publicly available Illumina platinum haplotypes generated for the same individual (NA12878) as a 'reference' standard (see **Data Access**). NA12878 'reference haplotypes' were completed by genetic haplotyping using highly accurate genotypes from seventeen individuals of a three-generation pedigree[35] , and can therefore serve as a gold-standard to assess the phasing accuracy throughout this study.

Phasing Performance of Individual Technologies

To independently assess the phasing performance of each technology we assembled haplotypes directly from sequencing reads (Illumina or PacBio) using WhatsHap (see **Methods**). The main advantage of this algorithm is that it solves the Minimum Error Correction (MEC) problem optimally with a runtime that scales linearly in the number of variants (alleles) and is independent of the read length. Therefore, it performs well with short-read technologies (Illumina) and is especially suited for use with long reads (PacBio, Oxford NanoPore). 10X Genomics haplotype segments were assembled by the vendor using the 10X LongRanger pipeline. To phase the multiple Strand-seq libraries we developed a new phasing algorithm, implemented in the R package StrandPhaseR (see **Methods, and Supplemental Fig. S1**). In comparison to our previously published phasing algorithm[9], the current version implements a more robust sorting-based phasing approach of single cell haplotypes into

consensus haplotypes, such that conflicts of alleles within both consensus haplotypes are reduced. The haplotypes generated by each technology (i.e. Illumina, PacBio, 10X Genomics and Strand-seq) were compared to the Illumina platinum reference haplotypes, to establish the density, completeness and accuracy of the phase blocks delivered by each platform independently. For a more streamlined exposition, we focus on results obtained for Chromosome 1 in the following analysis and present numbers aggregated across all chromosomes in a concluding discussion.

We found both PacBio and 10X Genomics technologies capable to phase nearly the complete set of variants listed in the reference haplotypes (98.8% and 97.2%, respectively), whereas Illumina alone phased only 77.8% and Strand-seq only 57.6% of the reference SNVs (**Fig. 1B**). The comparatively low percentage for Strand-seq can be explained by the relatively low sequencing coverage employed, combined with a slight unevenness in genomic coverage (**Supplemental Fig. S2**). For all technologies except Strand-seq, only short-range haplotypes were assembled using the read-based phasing, with a limited number of alleles phased per haplotype segment (**Fig. 1C**). For instance, we found >30,000 unconnected haplotype segments assembled from Illumina data, with the largest segment of 16kb (median ~500bp) harbouring only 0.06% of the phased variants. This is because heterozygous variants that are further apart than the length of the sequenced DNA fragments cannot be connected, resulting in multiple disjoint haplotype segments with an unknown phase between them. Improvements were achieved using longer sequencing reads from PacBio technology, which effectively decreased the number of phased haplotype segments (1,927) and increased their size; the largest segment of 1.7Mb (median ~21kb) containing 1.25% of all SNVs on Chromosome 1 (**Fig. 1C**). 10X Genomics produced even longer haplotype segments than both Illumina and PacBio data (**Fig. 1C**). The largest haplotype segment contained almost 5% of the heterozygous SNVs and spanned more than 8.5Mb (median ~241kb). Still, the haplotypes of Chromosome 1 came in 199 disconnected segments and, hence, an end-to-end phasing was not achieved. That is, the linked reads from 10X Genomics were not able to connect distant neighboring heterozygous sites, for instance at centromeres, genome assembly gaps or regions of low heterozygosity (**Fig. 1A**). This is in contrast to the global, albeit sparse, haplotypes produced by Strand-seq. Although the completeness of Strand-seq haplotypes was lower compared to the other technologies, all phased variants were placed into a single haplotype segment spanning the entire length of Chromosome 1 (**Fig. 1B,** and **C**).

Finally, we assessed the accuracy of each technology by calculating the extent of switch errors in comparison to the reference haplotypes. High phasing accuracy of each technology was exemplified by the low percentage (<0.4%) of switch errors (**Fig. 1D**) with PacBio and 10X Genomics being the most accurate. Since no single phasing technology was sufficient to generate both global and dense

haplotypes, we explored integrative phasing approaches that combine global, sparse haplotyping as afforded by Strand-seq technology with local high-density haplotypes from read-based phasing.
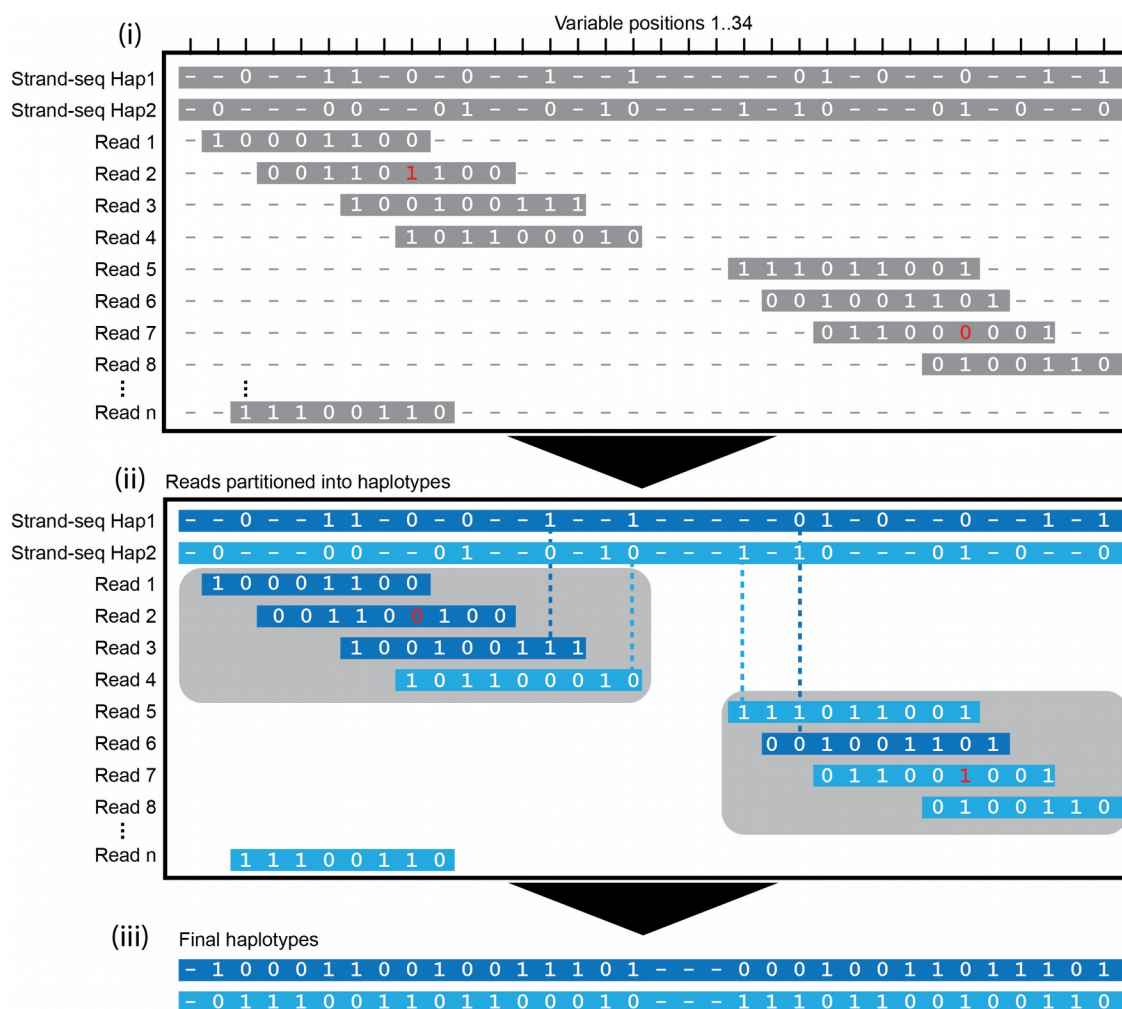

Figure 2: Integration of global and local haplotypes by WhatsHap algorithm.

An example solution of the weighted minimal error correction problem (wMEC) using WhatsHap algorithm is shown. For simplicity base qualities used as weights are omitted from the picture (for details on wMEC see Patterson et al. 2015). (i) The columns of the matrix represent 34 heterozygous variants (SNVs). Continuous stretches of zeros and ones indicate alleles supported by respective reads (0 – reference allele, 1 – alternative allele). First two rows of the wMEC matrix are represented by Strand-seq haplotypes, illustrated as one 'super read' connecting alleles along the whole length of the chromosome. (1st row haplotype 1 alleles, 2nd row haplotype 2 alleles). Subsequent rows of the matrix are represented by reads that map to the reference assembly in short overlapping segments. Sequencing errors (shown in red in read 2 and 7) are corrected when the cost for flipping the alleles is minimized. (ii) Reads are then partitioned into two haplotype groups (Haplotpye 1 – dark blue, Haplotype 2 – light blue) such that a minimal number of alleles are corrected (in red). As an illustration of long haplotype contiguity facilitated by Strand-seq 'super reads', we depict two non-overlapping groups of reads (gray rectangles) that can be stitched together by Strand-seq (dashed lines). (iii) Final haplotypes are exported for both groups of optimally partitioned reads.

Integrative global phasing strategy

To generate more complete and dense haplotypes, we sought to establish a novel and integrative phasing approach using a combination of Strand-seq data with the other data types. That is, we aim to enrich the sparse yet global phasing from Stand-seq using the dense haplotype information provided by Illumina, PacBio or 10X Genomics. However, integrating phase information across platforms poses a non-trivial statistical and algorithmic challenge, which we resolved by treating the sparse Strand-seq haplotypes generated by StrandPhaseR as one row in the fragment matrix processed by WhatsHap (see **Methods**). The other rows correspond to sequencing reads (PacBio, Illumina) or pre-assembled haplotype segments (10X Genomics) (see **Methods**). This allows, for the first time, for integrative phasing by solving the corresponding optimization problem (weighted MEC) provably optimal (**Fig. 2**). We performed extensive experiments to demonstrate that this approach enables excellent results in practice, as we describe in the following section.

To discover the most beneficial combinations of Strand-seq with Illumina or PacBio data, we explored combinations of variable numbers of Strand-seq libraries together with increasing depths of sequencing reads. To this end, we downsampled the number of Strand-seq libraries used in the analysis by randomly selecting subsets of libraries (5, 10, 20, 40, 60, 80, 100, or 120) from the original (N = 134) dataset. Similarly, we randomly downsampled the sequencing reads from the Illumina and PacBio datasets to a lower genomic coverage (2, 3, 4, 5, 10, 15, 25, and 30-fold). We applied our integrative phasing strategy to all pairs of downsampled Strand-seq libraries and the downsampled PacBio/Illumina datasets to assess the completeness (i.e. % of phased SNVs), contiguity (length of the largest haplotype segment) and accuracy (agreement with the 'reference' standard) of each assembled haplotype.

We found that the combination of Strand-seq haplotypes with any of the other data types markedly increased the number of variants that were phased in the largest haplotype segment, albeit to differing degrees (**Fig. 3A)**. Specifically, for the Illumina data we observed the completeness of each haplotype increased gradually with the number of Strand-seq libraries used in the experiment, whereas the depth of coverage of Illumina data had only a minor but noticeable effect (**Fig. 3A, i**). In contrast, the PacBio data showed a significant improvement in haplotype completeness at 10-fold genomic coverage, regardless of the number of Strand-seq libraries used (**Fig. 3A i, gray rectangle**). Similar results were seen when we combined Strand-seq with the 10X Genomics haplotypes (**Fig. 3A, ii)**. In all cases, integration of Strand-seq phasing drastically improved the contiguity of the haplotype spanning Chromosome 1 (**Fig. 3B**). When combining Illumina data with 40 Strand-seq libraries >65% of the reference variants could be phased accurately (**Fig. 3B i, black asterisk**); 5497 haplotype segments

(collectively representing 19.7% of the phased SNVs), however, remained disconnected, even when integrating the complete (N=134) Strand-seq dataset. These results confirm that Illumina data are of limited utility for haplotype phasing.

In contrast, as few as 10 Strand-seq cells combined with 10-fold PacBio coverage were sufficient to phase more than 95% of all heterozygous SNVs into a single haplotype segment (**Fig. 3B ii, black asterisk**), and merely 5 Strand-seq single cell libraries were required to connect all 10X Genomics haplotypes. However, we recommend at least 10 Strand-seq libraries (**Fig. 3B iii, black asterisk**) to ensure that at least one haplotype-informative (i.e. Watson-Crick-type) cell exists for every chromosome with high probability (p=0.978). This global haplotyping was unique to Strand-seq, as the combination of 10X Genomics with PacBio reads proved inefficient to join locally phased segments (**Fig. 3B iv**). That is, the added value of combining these two technologies is limited as the haplotype segments tend to break at similar locations.
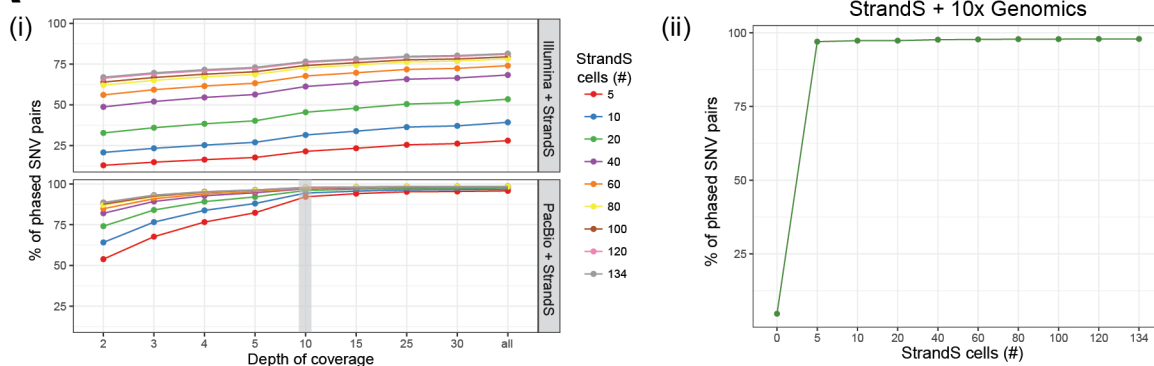
Finally, we assessed the phasing accuracy of the assembled haplotypes (the longest phased segment only) (**Fig. 3C**). Similar to the completeness of the haplotype, the accuracy of Illumina phasing gradually increased with sequencing depth and Strand-seq library number, indicating that Illumina coverage of 30-fold and higher is advisable (**Fig. 3C, i**). We further observed slightly elevated switch error rates at lower PacBio depths, which plateaued at 10-fold coverage. This is likely caused by allele uncertainty resulting from error-prone PacBio reads, especially at lower sequencing depths (**Fig. 3C, i**). The lowest switch error rate (< 0.2%) was achieved by the combination of Strand-seq with 10X Genomics data (**Fig. 3C, ii**).

Switch error rates reflect local inaccuracies expressed by the number of pairs of consecutive heterozygous variants that are wrongly phased with respect to each other. These error rates are not necessarily informative about global haplotype accuracy, which largely depend on how switch errors are spatially distributed (**Methods, Supplemental Fig. S4A**). Note that one single switch error implies that all following alleles (up to the next switch error) are assigned to the wrong haplotype. Since our goal is to generate dense and global haplotypes, we additionally report the Hamming error rate of the largest haplotype segment in comparison to the reference haplotypes (**Methods, Supplemental Fig. S4B**). Illumina reads are highly accurate and therefore we observed lower impact of sequencing depth on the global accuracy of the largest phased haplotypes (**Fig. 3C, iii**). In contrast, PacBio reads exhibited higher sequencing error rates, which translated into higher switch error rates at low sequencing depths. Using 10-fold PacBio coverage combined with at least 10 Strand-seq cells yielded highly accurate global haplotypes (**Fig. 3C, iii gray rectangle**), while lower coverages led to markedly worse results. Furthermore, the combination of Strand-seq with 10X Genomics haplotypes yielded
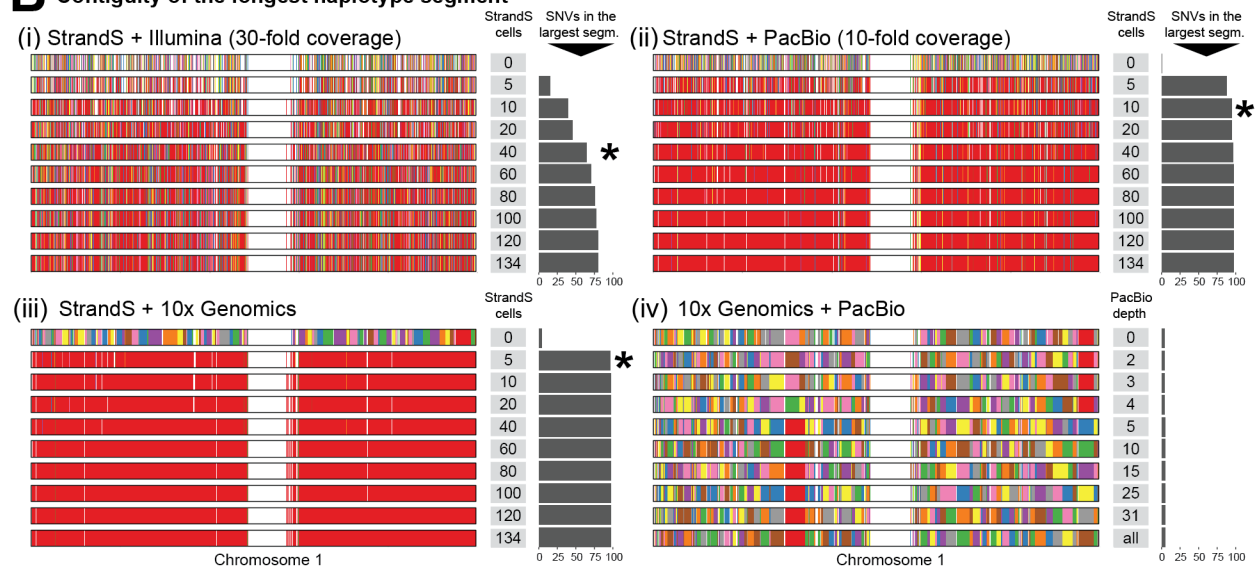
highly accurate global haplotypes, already at the minimal amount of Strand-seq libraries (**Fig. 3C, iv**).

Taken together, these results illustrate that Strand-seq can be used to phase existing sequence data and build dense, global and highly accurate haplotypes. Indeed, we found our approach highly efficient for genome-wide phasing (**Fig. 4A**). Using a combination of 40 Strand-seq libraries with 30-fold Illumina coverage, or 10 Strand-seq libraries with either 10-fold PacBio coverage or the 10X Genomics haplotypes we successfully scaffolded chromosome-length haplotypes for every autosome of NA12878. The completeness of the genome-wide haplotypes measured for the largest haplotype block reached 95.7% and 69.1% using PacBio and Illumina reads, respectively (**Fig. 4A, i**). We further demonstrated the high accuracy of these haplotypes on the local and global scales, which showed low switch (<0.45%) and Hamming error (<0.99%) rates for both the PacBio and Illumina combination (**Fig. 4A, i,ii**). Whereas scaffolding the 10X Genomic haplotypes produced the most accurate local haplotypes (switch error rate of 0.05%), global performance suffered, and the highest Hamming error rate (2.18%) was calculated for this combination. Nevertheless, using Strand-seq to scaffold any of the datasets remarkably improved the completeness, contiguity and accuracy of phasing for each chromosome, highlighting our integrative phasing strategy as a robust method for building dense and accurate whole genome haplotypes.

A Completeness of the longest haplotype segment

(i) — Illumina + StrandS / PacBio + StrandS; x-axis: Depth of coverage; y-axis: % of phased SNV pairs

(ii) StrandS + 10x Genomics; x-axis: StrandS cells (#); y-axis: % of phased SNV pairs

StrandS cells (#): 5, 10, 20, 40, 60, 80, 100, 120, 134

B Contiguity of the longest haplotype segment

(i) StrandS + Illumina (30-fold coverage)

(ii) StrandS + PacBio (10-fold coverage)

(iii) StrandS + 10x Genomics

(iv) 10x Genomics + PacBio

C Accuracy of the longest haplotype segment

(i) Switch error rate — Illumina + StrandS / PacBio + StrandS

(ii) StrandS + 10x Genomics — Switch error rate

(iii) Hamming error rate — Illumina + StrandS / PacBio + StrandS
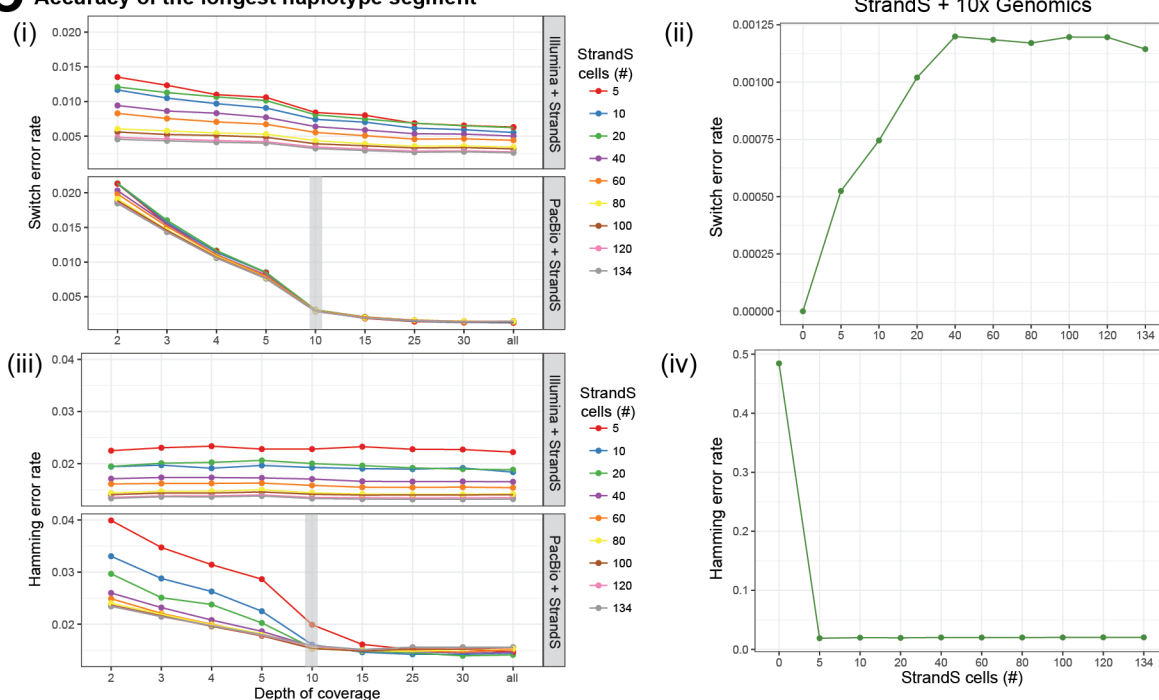
(iv) Hamming error rate

Figure 3: Various combinations of Strand-seq and read-based phasing (Illumina, PacBio, 10X Genomics) - example Chromosome 1.

Plots show haplotype quality measures for various combinations of Strand-seq cells (5, 10, 20, 40, 60, 80, 100, 120, 134) with selected coverage depths of Illumina or PacBio sequencing data (2, 3, 4, 5, 10, 15, 25, 30, >30-fold), or in combination with 10X Genomics haplotypes. **A**) Assessment of the completeness of the largest haplotype segment as the % of phased SNVs. Grey bars highlight PacBio sequencing depth where completeness and accuracy of final haplotypes do not dramatically improve. **B**) Assessment of the contiguity of the largest haplotype segment as the length of the largest haplotype segment. Every phased haplotype segment is depicted as a different color, with the largest segment colored in red. **C**) Assessment of the accuracy of the largest haplotype segment as the level of agreement with the 'reference' standard. Gray bars highlight Illumina and PacBio sequencing depth where accuracy of final haplotypes do not dramatically improve. In case of Illumina sequencing such improvement is more gradual.
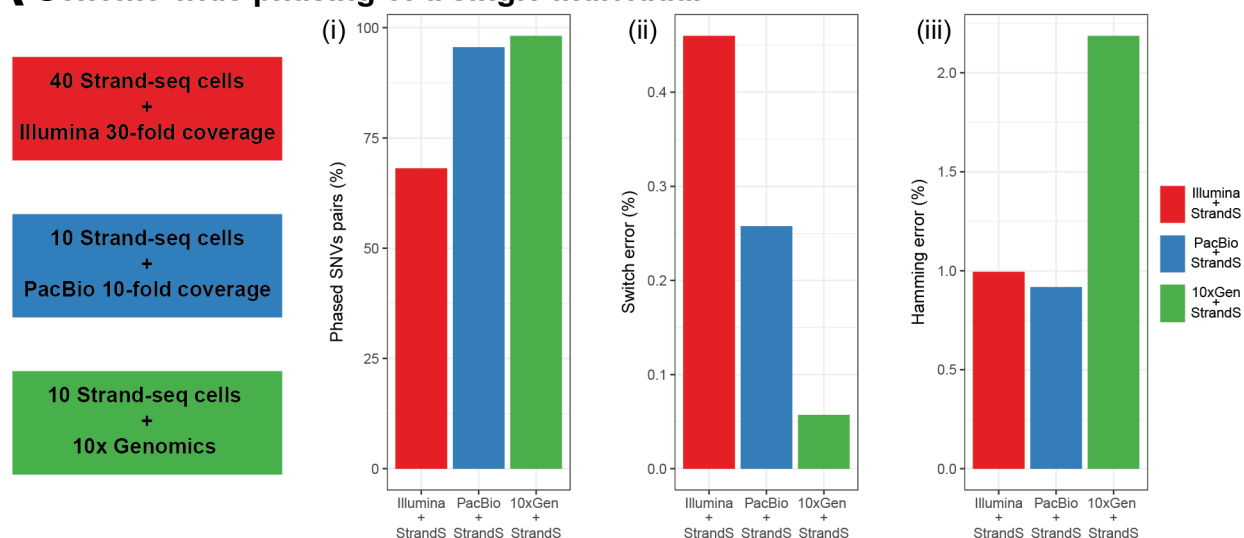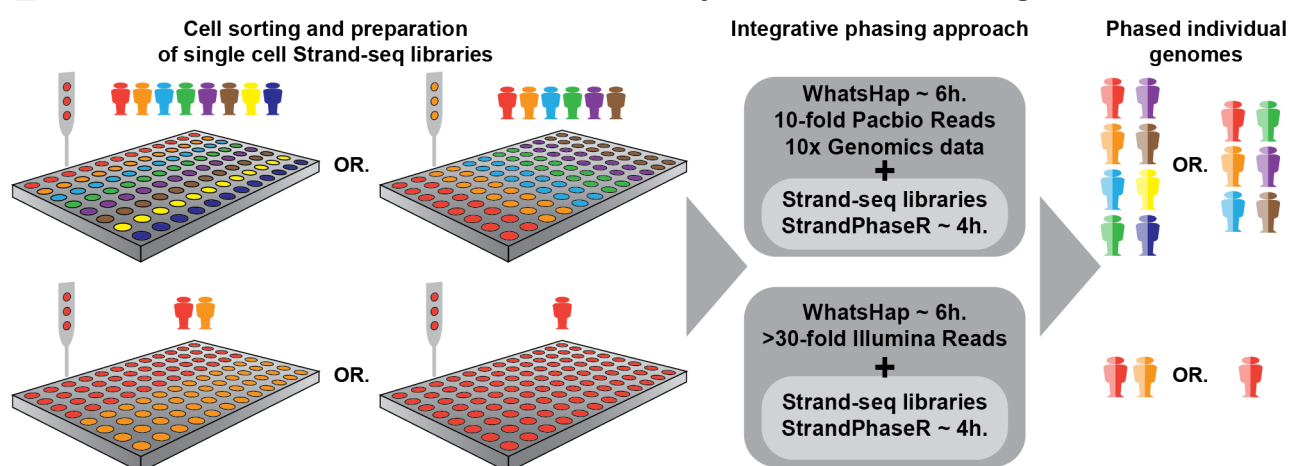
Figure 4: Recommended settings to phase certain amounts of individuals.

A) Genome-wide phasing of NA12878 using combination of 40 Strand-seq libraries with 30x short Illumina reads, 10 Strand-seq libraries with 10-fold long PacBio reads, or 10 Strand-seq libraries with 10X Genomics data. (i) The percentage of phased SNV pairs in the largest haplotype segment for each combination (ii)The cumulative accuracy the genome-wide haplotype was calculated by summing the switch error rates and (iii) the Hamming error rates found for each autosomes. **B**) A diagram providing the recommendations for the required number of Strand-seq libraries to be combined with recommended minimum of 10-fold PacBio and 30x Illumina coverage in order to reach global and accurate haplotypes for a depicted number of individual diploid genomes.

DISCUSSION

Strand-seq has been successfully prepared from a wide range of cell types taken from various organisms[9,34,36] and is currently being adopted by an increasing number of researchers. The integrative phasing strategy we introduce here paves the way to leveraging Strand-seq to obtain chromosome-length dense and accurate haplotypes at a manageable cost and labor investment. Based on the comprehensive evaluation presented above, we recommend three different combinations of Strand-seq with a complementary technology (**Fig. 4**).

As one option, one can combine Strand-seq with standard Illumina sequencing. Although the power of Illumina data for phasing is limited, mainly due to short insert sizes and read lengths, it still has some merit for adding additional variants to Strand-seq haplotypes. This might be of interest to many researchers since Illumina sequencing still constitutes the most common technology and there is an abundance of Illumina sequence data currently available for many sample genomes. To completely phase these preexisting data, we recommend generating 40 Strand-seq libraries for the sample genome, which is sufficient to phase >68% of all heterozygous variants genome-wide with good accuracy (switch error 0.45%, Hamming error 0.99%), see **Figure 4A** and **Supplemental Tab. S3**.

To build more complete haplotypes, we recommend combining Strand-seq with either PacBio or 10X Genomic technologies. A minimum of 10-fold PacBio coverage coupled with 10 Strand-seq libraries will phase >95% of heterozygous variants genome-wide with excellent accuracy (switch error 0.25%, Hamming error 0.91%). PacBio has been demonstrated to be particularly powerful for resolving structural variation[37,38] and, although not explored here, might hence be the best choice when the resolution of haplotypes, structural variation and repetitive regions is desired. However, the cost of this platform is still comparatively high. Therefore, until long-read technologies have become standard practice, we recommend combining 10 Strand-seq libraries with 10X Genomics technology. We found this combination yielded the most complete (>98% heterozygous variants genome-wide) haplotypes with the lowest switch error rate (0.05%). We did observe a slightly increased Hamming error rate (2.18%), however, which indicates that some genomic intervals are placed on the wrong haplotype, most likely due to switch errors in the pre-phased haplotype segments (produced by 10X Genomics) used as input. Overall, combining Strand-seq with 10X Genomics is the most cost-effective (in terms of time and money) strategy to phase an individual genome at extraordinary accuracy.

Our results demonstrate that dense and accurate chromosome-length haplotypes can be generated at manageable costs. This development brings haplotype-level analyses closer to a routine practice, which can be key for understanding disease phenotypes. We emphasize that the strategy we present here works for single individuals without relying on other family members or statistical inference from haplotype reference panels. In contrast to such population-based phasing approaches, the method we advocate here allows insights into rare and *de novo* variants and long-range epistatic effects.

Our future efforts will focus on *de novo* assembly of haplotype resolved-genomes without the alignment to a reference genome. This will provide us with true diploid representations of individual genomes, which will have profound implications to study variability of personal genomes in health and disease.

METHODS

1. StrandPhaseR pipeline

To build whole genome haplotypes from Strand-seq data we developed a new sorting-based pipeline, called StrandPhaseR. StrandPhaseR implements an improved phasing algorithm based on a binary sorting strategy of two parallel matrices, storing haplotype information obtained from single cell Strand-seq libraries. The analysis pipeline takes as input aligned BAM (binary alignment map) files from single cells, which were initially filtered for duplicate and low mapping quality reads (mapq < 10). Haplotype informative WC regions were localized in every Strand-seq library as described in Porubsky et al. (2016). Alleles at variable positions (supplied as set of SNVs obtained from Illumina platinum haplotypes) were identified separately for W and C reads in every informative region to generate low density single cell haplotypes that are then sorted by the phasing algorithm. The partial single cell haplotypes are used to fill two matrices, where rows represent cells and columns represent covered variable positions (SNVs) in any given cell (**Supplemental Fig. S1**). Initially, one matrix stores all variable positions found within the Watson templates, and a second matrix stores all variable positions found within the Crick templates. Cells in the matrices are sorted in decreasing order based on the number of covered variants (i.e. depth of coverage). Initially, a score of each column is calculated as the sum of all covered variants minus the most abundant variant. This represents the level of disagreement across all cells for the given SNV in the column. The sum of scores for each column represents the overall score of the matrix, and a lower matrix score represents a higher level of concordance across all SNV positions. Once the score of both matrices is determined, all SNVs in the first row (i.e. those belonging to the first cell) are swapped between the two matrices. In essence, this exchanges the Watson and Crick template strands of the cell within the matrix, to test whether there is a higher level of agreement across the phased SNVs found for all the cells. To determine this, the matrix scores are recalculated and if the scores are lower than the previous scores the change is kept, otherwise the change is reversed. The algorithm continues with the second row. Again, the covered variants of the second cell are swapped between matrices, the matrix score are recalculated and the decision to preserve or reverse the change is made. This is repeated through all rows (cells) of the matrix, sorting the single cell haplotypes within both matrices to reduce the number of conflicting alleles within each column. We repeated sorting process twice, after which we did not observe any further changes. The resulting haplotypes are reported as the consensus allele found across all the cells for each column of the matrices. Ideally, there is only one allele present for every variable site in each matrix, however sporadic sequencing errors or cell-specific artefacts can introduce discrepancies. Lastly, any missing

alleles at heterozygous sites are rescued by searching within the 'uninformative' reads (i.e. those from WW and CC regions) present in Strand-seq libraries and filled in. The final consensus haplotypes are exported in standardized VCF format, with each variable position has an assigned Phred quality score and entropy value reflecting the confidence in the given allele. All phasing steps of StrandPhaseR have been implemented into a single open-source R package (see **Code availability**).

2. Downsampling of Strand-seq libraries and read data (PacBio or Illumina)

To assess difference combinations of Strand-seq libraries (w.r.t. number of single cell libraries) with read data (w.r.t. depth of coverage), we performed a systematic analysis of the phasing performance for various subsets of each dataset. To achieve this, we downsampled the original publicly available (see Data Access) datasets consisting of: 134 single cell Strand-seq libraries (Porubsky et al. 2016), 39.6x coverage long-read PacBio data[39], and 49.6x coverage short-read Illumina data[40,41]. To simulate Strand-seq datasets consisting of reduced numbers of single cells, we randomly selected subsets of either 5, 10, 20, 40, 60, 80, 100, or 120 libraries from the original number of 134 libraries in the dataset. Read data from the PacBio and Illumina datasets were downsampled using Picard (picard-tools-1.130) to meet a defined depth of coverage of either 2, 3, 5, 10, 15, 25, or 30-fold. The downsampling was performed for 5 independent trials to account for variability in downsampled datasets, and the average phasing performance across all trials was reported (as described below).

3. Integrative phasing using WhatsHap

As an input for integrative phasing, Strand-seq haplotypes were phased using StrandPhaseR (exported in VCF format) and combined with either PacBio or Illumina alignments (both stored in BAM format) or 10X Genomics pre-phased haplotype segments (stored in the VCF produced by LongRanger) to phase heterozygous variants obtained from Illumina platinum genomes (see **Data Access**). We achieved this integrative phasing across platforms by solving the weighted minimum error correction (wMEC) problem using WhatsHap[19,20].

Mathematically, aligned reads from Illumina or PacBio (or pre-phased 10X Genomics haplotype segments) and sparse Strand-seq haplotypes are jointly represented in the form of a fragment matrix, where each row represent either one reads (in case of Illumina and PacBio), one pre-phased haplotype segment (in case of 10X Genomics) or one sparse global haplotype (in case of StrandSeq data) and columns represent the variant sites (**Fig. 2**). The matrix is filled with 0, 1 and '-' entries, where 0 and 1 indicate that the corresponding read supports the reference or alternative allele, respectively, and '-' means the information is missing (e.g. because a read does not cover this variant

site). WhatsHap selects a subset of rows and solves the wMEC problem optimally on these rows, as described earlier[20]. The result is a maximum likelihood bipartition of rows, which corresponds to the two sought haplotypes.

For all analyses, whatshap was provided with a reference genome (option --reference) to enable re-alignment-based allele detection when constructing the fragment matrix from sequencing reads. This has been shown to significantly improve performance for PacBio reads[20].

4. Quality metrics of assembled haplotypes

To assess the quality of assembled haplotypes in this study, we calculated different metrics described in the following.

Completeness: The process of haplotyping establishes phase relations between pairs of consecutive heterozygous variants. We call each such pair a 'phase connection'. For each haplotype segment produced by a (combination of) technologies, we therefore count the number of phase connections, which is equal to the number of heterozygous markers that make part of such a haplotype segment minus one. To measure the completeness of a phasing, we sum the number of phase connections across all haplotype segments and divide by the maximum possible number of phase connections, which is equal to the number of heterozygous variants on a chromosome minus one.

Switch error rate: The switch error rate is the fraction of phase connections for which the phasing between the two involved heterozyous variants is wrong (**Supplemental Fig. S3A)**.

Largest haplotype segment: In this study we are interested in haplotypes that span the whole length of all chromosomes. To measure the completeness of phasing, we report the fraction of heterozygous variants that are part of the largest haplotype segment.

Largest haplotype segment Hamming rate: To assess whether haplotypes are correct also over long genomic distances, we only consider the largest haplotype segment and compute the Hamming distance between true and predicted haplotypes (**Supplemental Fig. S3B)**, divided by the total number of heterozygous variants in this haplotype segment. That is,the Hamming error rate is equal to the fraction of wrongly phased heterozygous variants. Note that, only one switch error (e.g. in the middle of a chromosome) can result into a very high Hamming distance and hence the Hamming distance is a much more stringent quality measure. While the switch error rate assesses whether haplotypes are correct locally, i.e. between pairs of neighboring heterozygous variants, the Hamming distance assesses whether haplotypes are correct globally.

Data Access

Illumina reads[40,41]: Obtained from 1000 Genome Project Consortium (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/NA12878/high_coverage_alignment/).

PacBio reads[39]: Obtained from Genome in a Bottle Consortium (GIAB) (ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/NA12878_PacBio_MtSinai/sorted_final_merged.bam).

10X Genomics haplotypes: pre-assembled 10X Genomics haplotypes were downloaded from 10X Genomics website: https://support.10Xgenomics.com/genomeexome/datasets/NA12878_WGS_210 (we have filtered for only heterozygous and PASS filter SNVs)

Strand-seq libraries[9]: For this study have been downloaded from the European Nucleotide Archive (http://www.ebi.ac.uk/ena), accession number: PRJEB14185.

Reference haplotypes[35]*:* In this study we use as a reference trio based haplotypes of NA12878 obtained from Illumina platinum genomes (http://www.illumina.com/platinumgenomes/)

Code availability

The **StrandPhaseR** software is publicly available through GitHub.

(https://github.com/daewoooo/StrandPhaseR)

The **WhatsHap** software is publicly available through bitBucket.

(https://bitbucket.org/whatshap/whatshap)

Author's contribution

D.P and T.M. designed the study. D.P. and S.G. implemented the pipeline and performed experiments. D.P. prepared the figures. D.P., A.D.S., T.M. and S.G. wrote the manuscript, with the help of all authors. A.D.S., V.G., P.M.L. and J.O.K helped with data interpretation. All authors read and approved final version of the manuscript.

REFERENCES

1.   Tewhey, R., Bansal, V., Torkamani, A., Topol, E. J. & Schork, N. J. The importance of phase information for human genomics. *Nat. Rev. Genet*. **12**, 215–223 (2011).

2.   Wang, J., Fan, H. C., Behr, B. & Quake, S. R. Genome-wide Single-Cell Analysis of Recombination Activity and De Novo Mutation Rates in Human Sperm. *Cell* **150**, 402–412 (2012).

3.   Fan, H. C., Wang, J., Potanina, A. & Quake, S. R. Whole-genome molecular haplotyping of single cells. *Nat. Biotechnol.* **29**, 51–57 (2011).

4.   Glusman, G., Cox, H. C. & Roach, J. C. Whole-genome haplotyping approaches and genomic medicine. *Genome Med*. **6**, 73 (2014).

5.   Leung, D. et al. Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature* **518**, 350–354 (2015).

6.   Hnisz, D. et al. Resource Super-Enhancers in the Control of Cell Identity and Disease. *Cell* **155**: 934–947

7.   Corradin, O. et al. Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome Research* **24**:1–13 (2014).

8.   Corradin, O. et al. Modeling disease risk through analysis of physical interactions between genetic variants within chromatin regulatory circuitry. *Nat. Genet.* **48**(11): 1313–1320 (2016).

9.   Porubský, D. et al. Direct chromosome-length haplotyping by single-cell sequencing. *Genome Research* **26**(11): 1565–1574 (2016).

10.   Snyder, M. W., Adey, A., Kitzman, J. O. & Shendure, J. Haplotype-resolved genome sequencing: experimental methods and applications. *Nat. Publ. Gr*. **16**, 344–358 (2015).

11.   de Vree, P. J. P. et al. Targeted sequencing by proximity ligation for comprehensive variant detection and local haplotyping. *Nat. Biotechnol*. **32**, 1019–1025 (2014).

12.   Amini, S. et al. Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. *Nat. Genet. 46*, 1343–1349 (2014).

13.   Selvaraj, S., R Dixon, J., Bansal, V. & Ren, B. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nat. Biotechnol*. **31**, 1111–1118 (2013).

14.   Cilibrasi, R., Iersel, L. Van, Kelk, S., Tromp, J. & Sep, G. N. On the Complexity of the Single Individual SNP Haplotyping Problem. *Algorithmica.* **49**(1): 13–36 (2007).

15.   Lancia, G., Bafna, V., Istrail, S., Lippert, R. & Schwartz, R. SNPs Problems, Complexity, and Algorithms. 182–183 (2001).

16.   Honglan, J. R. & Joung, L. J. Survey of computational haplotype determination methods for single individual. *Genes Genomics*. **38**, 1–12 (2016).

17.   Deng, F., Cui, W. & Wang, L. A highly accurate heuristic algorithm for the haplotype assembly problem. *BMC Genomics*. **14**(2): 1–10, (2013).

18.   Kuleshov, V. et al. Whole-genome Haplotyping Using Long Reads and Statistical Methods. *Nat. Biotechnol*. **32**, 261–266 (2014).

19.   Patterson, M. et al. WhatsHap: Weighted haplotype assembly for future-generation sequencing reads. *Journal of Computational Biology*. **22**(6): 498–509 (2014).

20.   Martin, M. et al. WhatsHap : fast and accurate read-based phasing. *bioRxiv*. 1–18 (2016).

21.   Weisenfeld, N. I. et al. Direct determination of diploid genome sequences. *bioRxiv*. 1–21 (2016).

22.   Chin, C. et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods*. **13**(12): 1–7 (2016).

23.   Pendleton, M. et al. Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat. Methods* **12**, 780–786 (2015).

24.   Mostovoy, Y. et al. A hybrid approach for de novo human genome sequence assembly and phasing. *Nat. Methods* **13**, 12–17 (2016).

25.   Seo, J. et al. De novo assembly and phasing of a Korean human genome. *Nature* **538**, 243–247 (2016).

26.   Steinberg, K. M. et al. Single haplotype assembly of the human genome from a hydatidiform mole. *Genome Res*. **24**: 1–12 (2014).

27.   Ammar, R., Paton, T. A., Torti, D., Shlien, A. & Bader, G. D. Long read nanopore sequencing for detection of HLA and CYP2D6 variants and haplotypes. *F1000Research* **4**, 17 (2015).

28.   Zheng, G. X. Y. et al. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat. Biotechnol*. **34**: 303–311 (2016).

29.   Ma, L. et al. Direct determination of molecular haplotypes by chromosome microdissection. *Nat. Methods*. *7*, 299–301 (2010).

30.   Brown, P. J. B. et al. Correction for Yang et al., Completely phased genome sequencing through chromosome sorting. *Proc. Natl. Acad. Sci*. **109**, 3190–3190 (2012).

31.   Lieberman-aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*. **326**: 289–293 (2009).

32.   Edge, P., Bafna, V. & Bansal, V. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Res*. 1–23 (2016).

33.   Ben-elazar, S. et al. Extending partial haplotypes to full genome haplotypes using chromosome conformation capture data. *Bioinformatics* **32**: 559–566 (2016).

34.   Falconer, E. et al. DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. *Nat. Methods* **9**, 1107–12 (2012).

35.   Eberle, M. A. et al. A reference dataset of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *bioRxiv* (2016). doi:10.1101/055541

36.   Sanders, A. D. et al. Characterizing polymorphic inversions in human genomes by single cell sequencing. *Genome Res.* **26**(11):1575–1587 (2016).

37.   Chaisson, M. J. P., Wilson, R. K. & Eichler, E. E. Genetic variation and the de novo assembly of human genomes. *Nat. Publ. Gr.* **16,** 627–640 (2015).

38.   Huddleston, J. *et al.* Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res.* (2016). doi:10.1101/gr.214007.116

39.   Zook, J. M. *et al.* Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat. Biotechnol.* **32,** 246–251 (2014).

40.   Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526,** 75–81 (2015).

41.   Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).