Complete Haplotype Determination and Single-Chromosome Analysis

Richard W. Tourdot[1,2,3] **and Cheng-Zhong Zhang**[1,2,3]

***For correspondence:** richardW_tourdot@dfci.harvard.edu (rick); cheng-zhong_zhang@dfci.harvard.edu (CZ)

[1]Department of Data Sciences, Dana-Farber Cancer Institute, 3 Blackfan Circle, Boston MA 02215; [2]Department of Biomedical Informatics, 10 Shattuck Street, Harvard Medical School, Boston MA 02115; [3]Cancer Program, Broad Institute of MIT and Harvard, 415 Main Street, Cambridge MA 02142

Abstract Haplotype phase represents the collective genetic variation between homologous chromosomes and is an essential feature of non-haploid genomes. Although shotgun sequencing can reveal genotypes at variant sites, it cannot resolve linkage between variant genotypes on each chromosome as most sequencing fragments ($\lesssim$1kb) span at most one variant site (~1 per 1.5 kb in human genomes). Long-range sequencing can generate variant linkage information, but whether such information can determine whole-chromosome haplotype phase remains unclear. Here we describe a general framework for haplotype inference from variant linkage data using a one-dimensional spin model. Building on this model, we design and implement algorithms that can determine complete chromosomal haplotypes of diploid human genomes using 60× linked-reads sequencing in combination with 60× Hi-C sequencing. The computationally inferred haplotypes show high completeness (>98%) and accuracy (>99%) when compared to reference haplotypes determined from sequencing of single chromosomes. We further present examples of single-chromosome analysis enabled by whole chromosome haplotype information.

Introduction

Haplotype phase ("haploid genotype") is the combination of genotypes at sites of genetic variation along a chromosome (***The International HapMap Consortium, 2005***). In population genetics, haplotypes are used to represent strongly correlated genotypes at variant sites in linkage disequilibrium (***Reich et al., 2001***; ***Daly et al., 2001***). Each haplotype reflects a unique segment of an ancestral chromosome and serves as a specific marker of inheritance. In a single diploid genome, the chromosomal haplotype defines a barcode map of sequence variation between homologous chromosomes. Knowledge of chromosomal haplotypes is required both for interrogating molecular differences between homologous chromosomes, including variations both in the DNA sequence and in the epigenetic features, and to relate these variations to allelle-specific gene expression patterns (***Shendure and Aiden, 2012***; ***Adey et al., 2013a***; ***Hansen and van Oudenaarden, 2013***; ***Levesque and Raj, 2013***; ***Deng et al., 2014***; ***Snyder et al., 2015***). Chromosomal haplotype information can also significantly improve the resolution and accuracy of genome analysis, such as diploid genome assembly (***Pendleton et al., 2015***; ***Seo et al., 2016***) and *de novo* mutation detection in mixed cell populations (***Nik-Zainal et al., 2012***; ***Loh et al., 2018***) or single cells (***Zhang and Pellman, 2015***).

Haplotype inference requires knowledge of DNA sequence of each homologous chromosome. One strategy to obtain haplotype information is by analyzing the genotypes of related individuals (***Kong et al., 2008***; ***Browning and Browning, 2011***; ***Loh et al., 2016b***). The haplotypes of an individual's genome can be directly determined from the parents' genotypes, or statistically inferred from the genotypes of related individuals who provide "surrogate parent genotypes." Another strategy to

infer haplotypes is by statistical modeling of recombination based on the frequencies of genotypes in a population (***Scheet and Stephens, 2006***; ***Browning and Browning, 2007***; ***Loh et al., 2016a***). This strategy aims to infer the haplotype of an individual based on the recombination probabilities between adjacent polymorphic sites estimated from a panel of reference haplotypes. Although statistical phasing can be highly accurate for local haplotype inference (>99%), low-frequency switching errors (>0.1%) preclude long-range haplotype (10 Mb or above) inference unless the genotypes of closely related individuals are available. Another limitation of statistical phasing is that it cannot resolve the haplotype phase of *de novo* mutations only present in a single genome.

Haplotype information can be directly obtained by sequencing or genotyping single chromosomes or sub-haploid chromosomal fragments (***Browning and Browning, 2011***; ***Snyder et al., 2015***). This strategy requires experimental separation or tagging of DNA fragments from different chromosomes prior to genotyping or sequencing. Direct sequencing of single chromosomes can produce whole-chromosome haplotypes (***Ma et al., 2010***; ***Fan et al., 2010***; ***Yang et al., 2010***; ***Zhang et al., 2015***; ***Porubsky et al., 2017***), but the experimental procedure of chromosome isolation or tagging is laborious and restricted to dividing cells. Segmental haplotype information can be obtained by long-read sequencing or long-range sequencing technologies, which preserve genetic linkage within long DNA molecules that are either sequenced directly or tagged by unique molecular identifiers (***Kitzman et al., 2010***; ***Suk et al., 2011***; ***Kaper et al., 2013***; ***Adey et al., 2013b***, ***2014***; ***Peters et al., 2014***; ***Zheng et al., 2016***; ***Chu et al., 2017***; ***Marks et al., 2019***). The typical size of DNA molecules in long-range sequencing (∼10-100 kb) is sufficient for inferring haplotypes in regions of normal variant density (∼1 per kb), but inadequate to extend haplotypes in regions of low variant density ($\lesssim 1$ per 10 kb), and impossible to bridge large gaps (> 100 kb) with no identifiable variants, including all centromeres. Therefore, long-range sequencing alone cannot determine complete whole-chromosome haplotypes.

The only technology that can generate very long-range linkage information (but does not require single-chromosome isolation) is Hi-C sequencing (***Rao et al., 2014***). As chromosomes are spatially isolated in separate chromosomal territories in the cell nucleus, proximity-based ligation ensures that Hi-C fragments are predominantly generated between loci on a single chromosome. Genetic linkage from Hi-C sequencing can in principle extend to the whole chromosome (***Selvaraj et al., 2013***). However, because the contact frequency between two genomic loci decays rapidly with their genomic distance, the majority of Hi-C fragments are formed between loci in close proximity (<1 kb). Hi-C reads spanning polymorphic sites separated by more than 1 Mb are extremely rare and insufficient to generate long contiguous haplotypes in most genomes (***Edge et al., 2017***), except in special cases with an extremely high variant density (∼1 per 150 bp) (***Selvaraj et al., 2013***).

Although Hi-C data alone cannot generate long contiguous haplotypes, they can produce a long-range haplotype scaffold with missing haplotype information that can be filled by other technologies. This idea has led us to develop a two-tier approach to determine complete whole-chromosome haplotypes combining long-range sequencing and Hi-C sequencing (Figure 1). This approach first determines local haplotype blocks (∼1 Mb) using long-range sequencing and then merge these blocks into a single haplotype using Hi-C links. We first define a general formalism for inferring binary haplotypes based on linkage data. This formalism converts both local haplotype inference and long-range scaffolding of haplotype blocks to a minimization problem. We describe algorithms for solving the minimization problem by simple steepest-descent methods and a software package that performs both linkage extraction and haplotype inference. We perform haplotype inference on two genomes and benchmark the performance against reference haplotype data determined from single-chromosome sequencing. The computationally inferred haplotypes contain no long-range switching errors and can resolve the haplotype phase of >98% of all heterozygous variants with >99% accuracy. In particular, our computational method simultaneously performs haplotype inference and false variant filtration based on haplotype-specific linkage, thus providing a simple solution for variant detection and phasing. Finally, we describe two applications of whole-chromosome haplotype information for resolving structural alterations of single chromosomes. The

concepts and computational algorithms described here lay the foundation for haplotype-resolved genome analysis that enables the interrogation of genetic and epigenetic alterations specific to each chromosome with superior accuracy and resolution than standard genomic analysis.

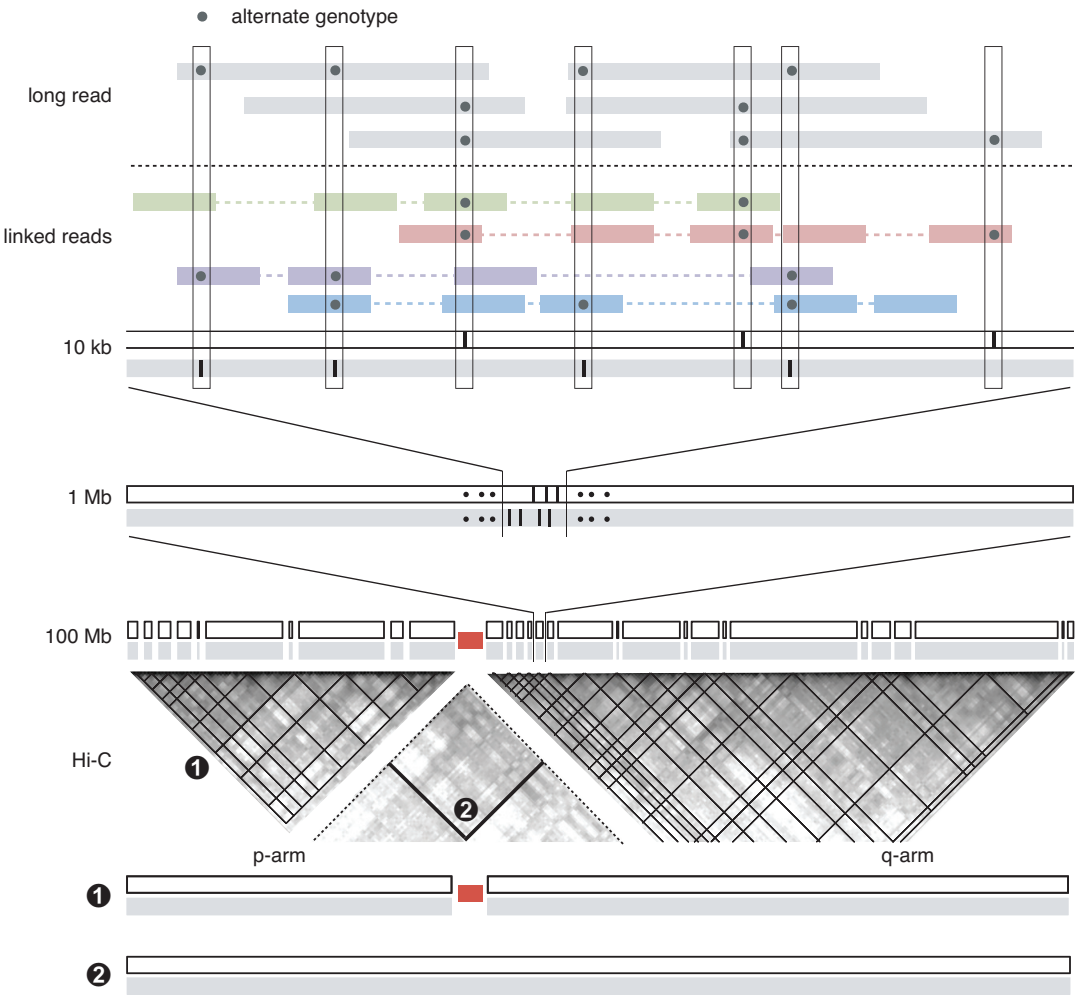

Figure 1. A hierarchical strategy for determining whole-chromosome haplotypes. We first generate local haplotype blocks (∼ Mb) using linkage information from linked-reads or long-read sequencing. We then concatenate these blocks into a single haplotype for each chromosome arm using Hi-C contacts between different haplotype blocks. Finally, we join p- and q-arm haplotypes using Hi-C contacts between the arms.

Results

Generation of linked-reads, Hi-C, and single-chromosome sequencing data

We performed computational haplotype inference and validation on the NA12878 cell line and the retinal pigmental epithelium (RPE-1) cell line. Haplotype inference for the NA12878 genome was done using linked-reads whole-genome sequencing data from 10X genomics (two datasets each with ∼35× mean depth) and Hi-C data from ***Rao et al.*** (***2014***); the inferred haplotypes were benchmarked using haplotypes derived from parental genomes that was released by the Genome-In-A-Bottle (GIAB) consortium (https://www.nist.gov/programs-projects/genome-bottle). Haplotype inference for the RPE-1 genome was done using newly generated linked-reads sequencing data (60× mean depth) and Hi-C data published by ***Darrow et al.*** (***2016***). To benchmark the computationally inferred RPE-1 haplotypes, we determined whole-chromosome RPE-1 haplotypes using a previously described strategy based on single-cell sequencing (***Zhang et al., 2015***). Instead of isolating single

chromosomes, this strategy generates monosomic cells by inducing random chromosome mis-segregation after release from mitotic arrest by nocodazole, a microtubule depolymerization agent. We constructed sequencing libraries of single cells after the nocodazole treatment and identified monosomic cells based on arm-level DNA copy number estimated from low-pass whole-genome sequencing (~0.1×). Monosomic cells were then sent for deeper whole-genome sequencing at 5-10×, resulting in 50% average allelic coverage. Monosomic chromosomes were independently validated by loss-of-heterozygosity using genotype coverage in the deep sequencing data as described before (***Zhang et al., 2015***). We generated a total of 98 monosomic chromosomes from 39 cells, with at least one for each chromosome, and aggregated all the data to generate a single reference haplotype of each chromosome. The sources and metrics of all the sequencing data used for haplotype inference are summarized in Table 1.

RPE-1 cells have a mostly diploid genome that is also very stable. They are frequently used in cell biology and molecular genetic studies (***Zhang et al., 2015***; ***Maciejowski et al., 2015***). The sequencing and haplotype data of RPE-1 cells generated in this study thus not only provide a new reference dataset for bioinformatic tool development and benchmarking, but are also a valuable resource for performing sequencing analysis of samples from *in vitro* genome evolution experiments.

Table 1. Data used for haplotype inference and benchmarking

Sample	Data type	Data source	Read count	Mean depth	Contacts (> 1 Mb)	Application
RPE-1	bulk WGS	***Zhang et al.*** (***2015***)	228,708,769[a]	13×		variant calling
RPE-1	linked reads	to be released	941,518,426[b]	60×[c]		variant calling & local phasing
RPE-1	Hi-C	***Darrow et al.*** (***2016***)	281,285,484[d]		48,124,211	long-range phasing
RPE-1	single cell with monosomies	to be released				Hi-conf variants & reference haplotypes
NA12878	linked reads v.1	10X Genomics[e]	422,179,395[f]	35×[c]		local phasing
NA12878	linked reads v.2	10X Genomics[g]	423,854,243[h]	35×[c]		local phasing
NA12878	Hi-C	***Rao et al.*** (***2014***)	413,230,948[i]		54,348,766	long-range phasing
NA12878	Hi-C	***Rao et al.*** (***2014***)	486,848,169[j]		91,428,507	long-range phasing
NA12878	phased VCF	GIAB[k]				Hi-conf variants & reference haplotypes

[a]SRR1778442: median insert 243; 208,151,992 aligned in pair; 101 × 2; duplication rate 0.024.
[b]Mean molecular length 24.8 kb; median insert 551; 913,660,083 aligned in pair; 150 × 2; duplication rate 0.255.
[c]excluding the GEMcode sequence
[d]SRS1045722: median insert 364; 279,027,892 aligned in pair; 150 × 2; duplication rate 0.067.
[e]https://support.10xgenomics.com/genome-exome/datasets/2.1.0/NA12878_WGS_210
[f]Mean molecular length 68.7 kb; median insert 349; 407,015,530 aligned in pair; 150 × 2; duplication rate 0.062.
[g]https://support.10xgenomics.com/genome-exome/datasets/2.2.1/NA12878_WGS_v2
[h]Mean molecular length 85.6 kb; median insert 370; 418,283,435 aligned in pair; 150 × 2; duplication rate 0.079.
[i]SRR1658571: median insert 459; 405,241,372 aligned in pair; 101 × 2; duplication rate 0.303.
[j]SRR1658572: median insert 377; 484,211,662 aligned in pair; 101 × 2; duplication rate 0.028.
[k]https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/NA12878_HG001/latest/GRCh38/

Quantitative features of variant linkage from linked-reads and Hi-C sequencing

The most important features of variant linkage information are the density and the percentage of *cis*-linkage (*i.e.*, linkage along the same chromosome). Both metrics decay with the genomic distance between variant sites. To optimize haplotype inference, we first analyze the distance dependence of linkage density and linkage accuracy in different types of long-range sequencing data. We focus on linked-reads sequencing and Hi-C sequencing due to their better availability and lower cost relative to long-read sequencing. We expect the density of linkage from long-read sequencing, including PacBio Single-Molecule Real Time (SMRT) sequencing and nanopore sequencing, to have similar distance dependence as that from linked-reads sequencing, both of which determined by

the size distribution of input DNA. The higher rate of sequencing errors in long-read sequencing technologies can result in lower linkage accuracy but the frequency of such errors is independent of the distance between variant sites.

Figure S1 shows the density maps of variant linkage derived from linked-reads (∼60× mean depth) and Hi-C sequencing data (∼50 million contacts above 1Mb) of RPE-1 cells. Although linkage is enriched between variants in close proximity (near the diagonal) in both data, the density maps show qualitative difference. In linked-reads data, linkage is concentrated between variants within 0.1 Mb (Fig. S1A, middle panel); by contrast, variant linkage in Hi-C data is very sparse above 1 kb (Fig. S1B, middle panel) but extends to a much longer range (Fig. S1B, left panel). To assess this difference quantitatively, we have calculated three statistical metrics of variant linkage as a function of the distance between variants: (1) percentage of linked variants; (2) average number of links between linked variants; and (3) accuracy of links between variants. These results are shown in Fig. 2 for both the RPE-1 data and the NA12878 data.

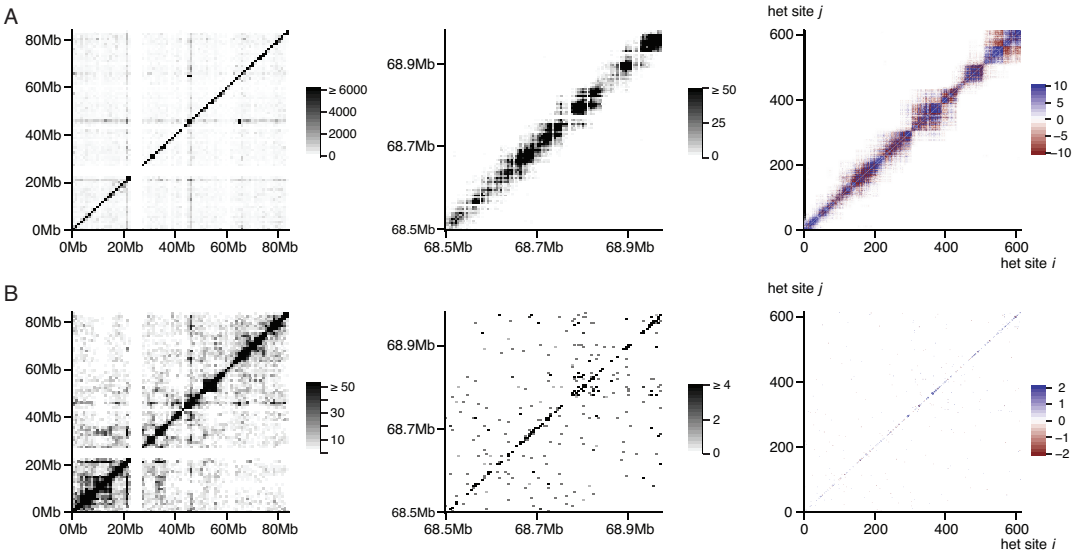

Figure S1. Density maps of variant linkage in linked-reads and Hi-C sequencing data of the RPE-1 genome. A. Heatmap of variant linkage across Chr.17 generated from 10^7 links randomly sampled from 6.1×10^7 total links in the linked-reads data (60× mean depth). Left: variant linkage across Chr.17 at 1Mb resolution; Middle: variant linkage in 68.5-70 Mb at 5 kb resolution; Right: linkage between single variant sites in 68.5-70 Mb. In the left and middle panels, the density map represents the total number of links between two bins (off-diagonal) and within each bin at the diagonal; in the right panel, the color map represents combined haplotype linkage between two variant sites calculated according to Eq. (2). B. Heatmap of variant linkage across Chr.17 generated from 1.6×10^5 long-range Hi-C links (> 1kb).

We observe a sharp drop in linkage density in the linked-reads data when the distance between variants exceeds a threshold. The threshold is ∼100 kb in the RPE-1 data (Fig. 2A,C) and ∼300 kb in the NA12878 data (Fig. 2B,D). These thresholds are consistent with the maximum size of input DNA molecules estimated by the LongRanger pipeline: For the NA12878 data, the mean molecular length was 85 kb (standard deviation: 67 kb); for the RPE-1 data, the mean molecular length was 24 kb (standard deviation: 41 Mb). There is a weak, distance-independent linkage signal above the molecular size that often consisted of 1-2 links (Fig. 2C,D). This signal most likely reflects contamination of a small number of molecular barcodes that tag sequencing fragments derived from different DNA molecules. Consistent with this speculation, 50% of variant linkage above the molecular length threshold is incorrect (Fig. 2E,F) when compared against the reference haplotype, as expected for a random background.

We observe a power-law decay of linkage density against variant distance that is similar to the decay of contact frequency between distal genomic loci (***Lieberman-Aiden et al., 2009***; ***Sanborn***

et al., 2015). The power-law decay against variant distance holds between 10 kb to more than 10 Mb in both Hi-C datasets (Fig. 2A,B). This has two implications. First, it implies that the majority of Hi-C links are intra-chromosomal as contacts between different chromosomes will not follow the power-law dependence. This is further validated using the reference haplotype showing the majority of Hi-C links are in *cis* (Fig. 2E,F). Second, the power-law decay implies that long-range Hi-C links are rare: Less than 1% of all variants separated by 10 kb or more are linked by Hi-C reads (Fig. 2A,B), most of which by one read (Fig. 2C,D). Therefore, linkage from Hi-C data is too sparse to generate contiguous phase blocks and is sensitive to sequencing errors.

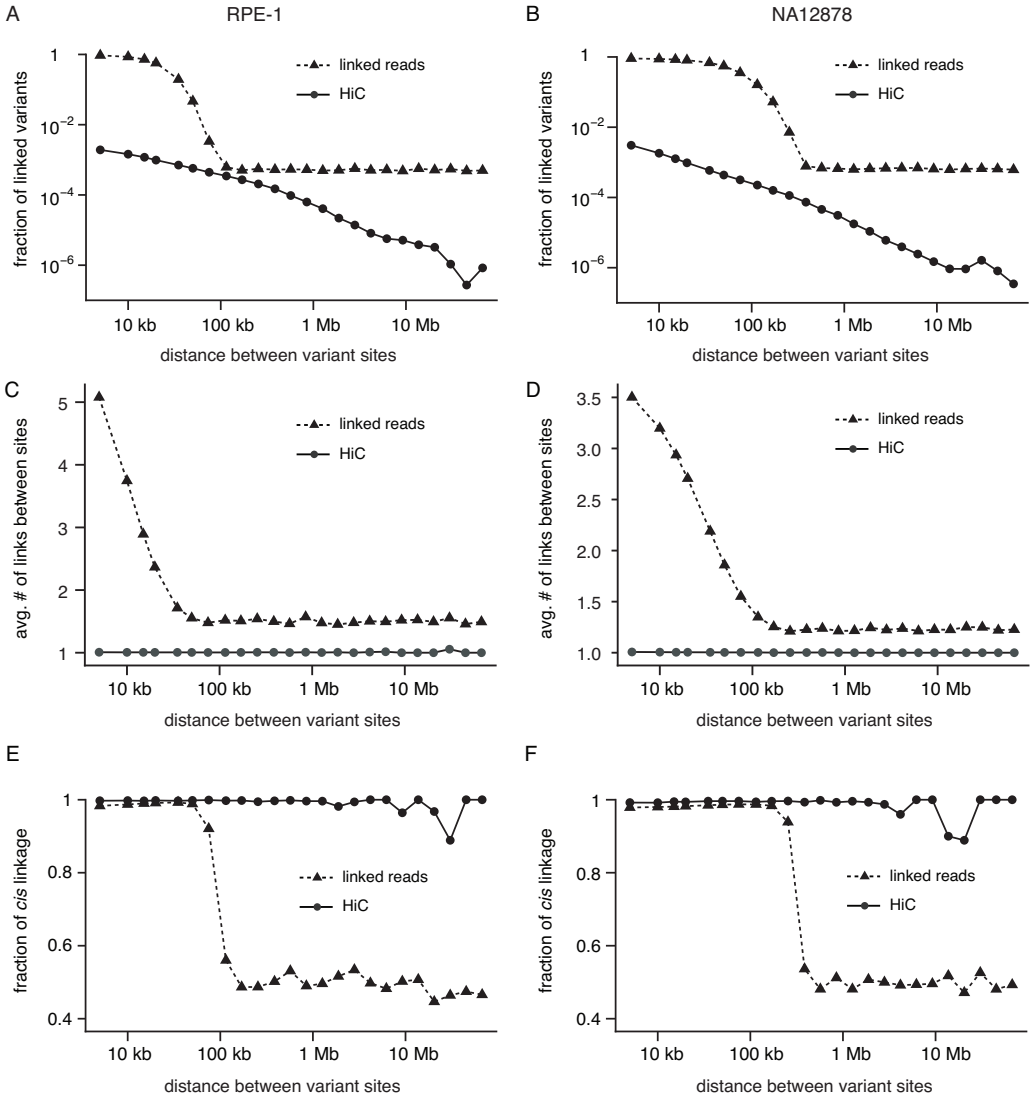

Figure 2. Quantitative metrics of variant linkage calculated from linked-reads and Hi-C sequencing data of RPE-1 (A,C,E) and NA12878 cells (B,D,F). A and B. Percentage of linked variants at different genomic distance. C and D. Average number of links between linked variants. E and F. Percentage of links consistent with linkage on the same homolog (*cis* linkage). Because any random link between a pair of variants has 50% probability of being consistent with *cis*-linkage, the real accuracy of linkage should be $2 \times$ (fraction of correct linkage) $- 1$.

The limitations and advantages of linked-reads and Hi-C data are complementary: Linkage in linked-reads data is dense and accurate but limited to variants within 1 Mb, whereas linkage in Hi-C data extends above 10 Mb but is very sparse. We can overcome both limitations by first generating local contiguous haplotype blocks using linkage in linked-reads data and then concatenating these blocks using long-range Hi-C links between blocks (Fig. 3A). Even though Hi-C links are very sparse,

the signal of Hi-C linkage aggregated from all Hi-C links between large blocks can be significant. For example, the average density of Hi-C links at 5 Mb distance is 10^{-5} in the RPE-1 data (Fig. 2A); given an average variant density of 1 per 1.5 kb in the human genome, there are about 600 variant sites within a 1 Mb block; therefore, there are approximately $10^{-5} \times 600 \times 600 \approx 4$ Hi-C links between two 1 Mb haplotype blocks separated by 5 Mb. This signal is sufficient to determine their linkage.

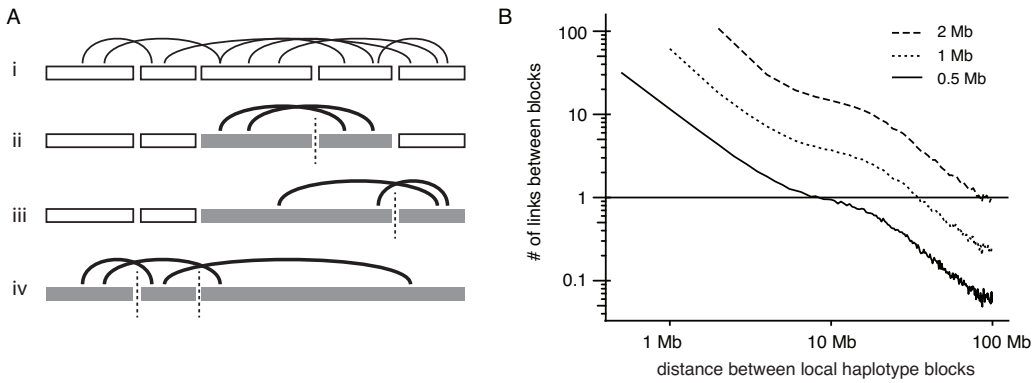

Figure 3. Linking small phase blocks with phased Hi-C reads. A. Schematic illustration of block-wise phasing by Hi-C reads. (i) Local phase blocks (open rectangles) and phased Hi-C links (curves above phase blocks); (ii) and (iii) merging of adjacent phase blocks based on significant linkage signal (thick links) from Hi-C links; the example here assumes that a minimum of two links is required for determining the phase between blocks. (iv) joint haplotype inference using all links between blocks. B. Average number of Hi-C links between two segments of 0.5, 1, and 2 Mb at different genomic distance calculated using the RPE-1 Hi-C data (Table 1).

To demonstrate this quantitatively, we calculated the average number of Hi-C links between two blocks of 0.5, 1, and 2 Mb at different genomic distance contained in the RPE-1 data (Fig. 3B). On average, there should exist ≥4 Hi-C links between two 1 Mb blocks within 10 Mb. The linkage signal between two blocks of 2 Mb is 4× stronger and can extend to 40 Mb, which is sufficient to bridge all gaps in the human genome. We can therefore achieve whole-chromosome haplotype inference in two steps, first generating local haplotype blocks ∼1Mb using linked-reads or long-read data, then concatenating the local haplotype blocks using Hi-C linkage. Both the inference of local haplotypes and their concatenation can be solved using the same strategy that will be discussed next.

Haplotype inference as a minimization problem

We formulate haplotype inference as an energy minimization problem of a one-dimensional spin model. The genotype at each variant site i is represented as $s_i = +1$ for the reference base and $s_i = -1$ for the alternate base. The haplotype phase at sites $i = 1, 2, \cdots N$ is given by

$$\mathbf{S} = \left(s_1, s_2, \cdots, s_N\right).$$

$\mathbf{S}$ and $-\mathbf{S}$ are complementary haplotypes of homologous chromosomes and thus equivalent.

There are four possible genotype combinations at any two variant sites:

$$\begin{aligned}
&\text{reference-reference linkage:} && s_i = 1, s_j = 1;\\
&\text{alternate-alternate linkage:} && s_i = -1, s_j = -1;\\
&\text{reference-alternate linkage:} && s_i = 1, s_j = -1;\\
&\text{alternate-reference linkage:} && s_i = -1, s_j = 1.
\end{aligned}$$

These four combinations can be grouped into two types of haplotype linkage:

$$\begin{aligned}
&\text{reference-reference/alternate-alternate linkage:} && s_i \cdot s_j = 1;\\
&\text{reference-alternate/alternate-reference linkage:} && s_i \cdot s_j = -1.
\end{aligned}$$

Linkage information from either genetic linkage disequilibrium or long-range sequencing can be represented as a coupling μ_{ij} between variant genotypes at site i and site j such that

$$\mu_{ij} > 0 \Rightarrow s_i \cdot s_j = 1;$$
$$\mu_{ij} < 0 \Rightarrow s_i \cdot s_j = -1.$$

Haplotype inference is equivalent to finding the spin chain $\mathbf{S}$ that minimizes the total energy

$$E = -\frac{1}{2}\sum_{i,j,\mu} \mu_{ij} s_i s_j, \tag{1}$$

where the sum runs over all pairs of spins and all links between spins.

We accumulate all linkage information between spin i and spin j as

$$M_{ij} = \begin{cases} \chi_{ij}\left(n_{ij}^{\mathsf{RR}} + n_{ij}^{\mathsf{AA}} - n_{ij}^{\mathsf{AR}} - n_{ij}^{\mathsf{RA}}\right) & i \neq j \\ 0 & i = j \end{cases} \tag{2}$$

where n_{ij}'s are the counts of unique reads (for long-read data) or unique molecules (for linked-reads data) supporting each type of linkage.

$$\chi_{ij} = \frac{\left|n_{ij}^{\mathsf{RR}} + n_{ij}^{\mathsf{AA}} - n_{ij}^{\mathsf{AR}} - n_{ij}^{\mathsf{RA}}\right|}{n_{ij}^{\mathsf{RR}} + n_{ij}^{\mathsf{AA}} + n_{ij}^{\mathsf{AR}} + n_{ij}^{\mathsf{RA}}} \tag{3}$$

measures the fraction of dominant linkage: For perfect linkage data in a diploid genome, either $n^{AR} + n^{RA} = 0$ or $n^{AA} + n^{RR} = 0$, and $\chi = 1$. M_{ij} represents the signal of combined haplotype linkage between variant site i and j ($|M_{ij}| \gg 0$ reflects significant linkage and $M_{ij} \approx 0$ implies little or no linkage) and χ_{ij} serves to penalize the linkage signal when it consists of a large fraction of conflicting evidence. With the introduction of M_{ij}, the correct haplotype $\mathbf{S} = \left(s_1, s_2, \cdots s_N\right)$ can be solved by minimizing the total energy given by

$$E\left(\mathbf{S}\right) = -\frac{1}{2}\sum_{i,j} M_{ij} s_i s_j = -\frac{1}{2}\,\mathbf{S}\mathbf{M}\mathbf{S}^T. \tag{4}$$

Minimization of Eq. (4) is equivalent to finding the haplotype configuration $\mathbf{S}$ with the least amount of conflict against the observed linkage M_{ij}. This formalism of haplotype inference is general and allows straightforward integration of linkage information from different data sources (*e.g.*, genetic linkage disequilibrium or long-range sequencing) or from different data types (long-read, linked-reads, or Hi-C sequencing). Before discussing strategies to solve the minimization problem, we want to highlight several features of this formalism.

First, the intrinsic symmetry between $\mathbf{S}$ and $-\mathbf{S}$ in the definition of $E(\mathbf{S})$ and M_{ij} ensures that haplotype inference is independent of whether linkage information is derived from one or both parental homologs. This is convenient for haplotype inference in aneuploid genomes where linkage information may be dominated by one homolog (*e.g.*, in trisomic regions with 2:1 allelic ratio).

Second, it can be shown that if M_{ij} contains only intra-chromosomal linkage from either parental homolog, then minimization of Eq. (4) is guaranteed to find a global minimum of $E(\mathbf{S})$, with degeneracy arising only where no linkage is present. Moreover, the energy function $E(\mathbf{S})$ is related to the log-likelihood of a haplotype configuration $\mathbf{S}$ given the observed linkage M_{ij}. This interpretation is convenient for assessing phasing confidence. See Appendices A and B for more technical details.

Finally, Eq. (4) is a general formalism for solving binary haplotypes. For local haplotype blocks $\mathbf{B}_s$ $(s = 1, 2, \cdots m)$ (and their complements $\overline{\mathbf{B}}_s$), we can define linkage between blocks similar to Eq. (2):

$$L_{st} = n(\mathbf{B}_s \leftrightarrow \mathbf{B}_t) + n(\overline{\mathbf{B}}_s \leftrightarrow \overline{\mathbf{B}}_t) - n(\overline{\mathbf{B}}_s \leftrightarrow \mathbf{B}_t) - n(\mathbf{B}_s \leftrightarrow \overline{\mathbf{B}}_t). \tag{5}$$

The phase of local haplotype blocks can be solved by minimizing

$$E\left(\mathbf{B}\right) = -\frac{1}{2}\mathbf{B}\mathbf{L}\mathbf{B}^T = -\frac{1}{2}\sum_{s,t} L_{st} b_s b_t, \tag{6}$$

where the chromosomal haplotype is represented using local haplotype blocks as

$$\mathbf{B} = \left(b_1\mathbf{B}_1, \cdots, b_m\mathbf{B}_m\right), \quad b_s = \pm 1.$$

Computational strategies for solving the haplotype

To find the haplotype configuration $\mathbf{S}$ that minimizes $E(\mathbf{S})$ given by Eq. (4), we introduce two types of perturbations:

$$\text{spin flip:} \quad (s_1, \cdots s_{i-1}, s_i, s_{i+1}, \cdots s_N) \rightarrow (s_1, \cdots s_{i-1}, -s_i, s_{i+1}, \cdots s_N) \tag{7}$$

$$\text{block switch:} \quad (s_1, \cdots s_i, s_{i+1} \cdots s_N) \rightarrow (s_1, \cdots s_i, -s_{i+1}, \cdots - s_N). \tag{8}$$

The energy changes associated with these perturbations are given by

$$\Delta E_i = s_i \sum_j M_{ij} s_j \quad \text{(spin flip)} \tag{9}$$

and

$$\Delta E_{k|k+1} = \sum_{i \leq k} \sum_{j > k} M_{ij} s_i s_j. \quad \text{(block switch)} \tag{10}$$

Starting from a random initial state

$$p(s_i = 1) = p(s_i = -1) = 1/2,$$

we iteratively introduce spin flipping or block switching perturbations with a negative energy change ($\Delta E < 0$) until the total energy $E(\mathbf{S})$ reaches a steady state (all $\Delta E > 0$). More details on the implementation of the minimization algorithms and related discussion can be found in the Materials and Method section and in Appendix A.

The energy function $E(\mathbf{S})$ can be interpret as a log-likelihood of a haplotype configuration $\mathbf{S}$ and the energy changes ("penalties") ΔE_i or $\Delta E_{i|i+1}$ are related to the probability of phasing errors:

$$\epsilon = 1 - p(\mathbf{S}|M_{ij}) \approx \frac{1}{1 + e^{\beta \Delta E}}, \tag{11}$$

where β is a heuristic parameter dependent on the background error frequency. A low energy penalty ($\Delta E \approx 0$) implies low phasing confidence ($\epsilon \approx 0.5$), and vice versa ($\Delta E \gg 0 \Rightarrow \epsilon \approx 0$). See Appendix B for more discussion on the relationship between energy penalty and phasing confidence.

We note that the spin-flipping energy penalty (Eq. (9)) can be rewritten as:

$$\begin{aligned}
\Delta E_i &= s_i \sum_{j \neq i} M_{ij} s_j \approx s_i \sum_{j \neq i} (n_{ij}^{RR} + n_{ij}^{AA} - n_{ij}^{AR} - n_{ij}^{RA}) s_j \\
&= s_i \left[\sum_{j \neq i, s_j = 1} (n_{ij}^{RR} + n_{ij}^{AA} - n_{ij}^{AR} - n_{ij}^{RA}) - \sum_{j \neq i, s_j = -1} (n_{ij}^{RR} + n_{ij}^{AA} - n_{ij}^{RA} - n_{ij}^{AR}) \right] \\
&= s_i \underbrace{\left(\sum_{j \neq i, s_j = 1} n_{ij}^{RR} + \sum_{j \neq i, s_j = -1} n_{ij}^{RA} + \sum_{j \neq i, s_j = 1} n_{ij}^{AA} + \sum_{j \neq i, s_j = -1} n_{ij}^{AR} \right)}_{\eta_{R \leftrightarrow \mathbf{S}} = \#(\text{reference} \leftrightarrow \mathbf{S}) + \#(\text{alternate} \leftrightarrow -\mathbf{S})} \\
&\quad - s_i \underbrace{\left(\sum_{j \neq i, s_j = 1} n_{ij}^{AR} + \sum_{j \neq i, s_j = -1} n_{ij}^{AA} + \sum_{j \neq i, s_j = 1} n_{ij}^{RA} + \sum_{j \neq i, s_j = -1} n_{ij}^{RR} \right)}_{\eta_{A \leftrightarrow \mathbf{S}} = \#(\text{alternate} \leftrightarrow \mathbf{S}) + \#(\text{reference} \leftrightarrow -\mathbf{S})}
\end{aligned} \tag{12}$$

The two terms $\eta_{R \leftrightarrow \mathbf{S}}$ and $\eta_{A \leftrightarrow \mathbf{S}}$ in Eq. (12) reflect the total linkage between the two genotypes at site i (R for reference and A for alternate) and other sites ($j \neq i$) in homologous chromosomes with complementary haplotypes $\mathbf{S}$ and $-\mathbf{S}$ ("haplotype linkage"). For true heterozygous variants, the correct linkage has to be either $R \leftrightarrow \mathbf{S}$ (and hence $A \leftrightarrow -\mathbf{S}$), or $A \leftrightarrow \mathbf{S}$ ($R \leftrightarrow -\mathbf{S}$), but not both. Therefore, either $\eta_{R \leftrightarrow \mathbf{S}}$ or $\eta_{A \leftrightarrow \mathbf{S}}$ has to be very small, and the flipping energy ΔE_i is then equivalent to the correct haplotype linkage, $\eta_{R \leftrightarrow \mathbf{S}}$ (when $s_i = 1$) or $\eta_{A \leftrightarrow \mathbf{S}}$ (when $s_i = -1$). A low flipping energy penalty can reflect either weak haplotype linkage $\eta \approx 0$, or a false variant with conflicting haplotype linkage $\eta_{R \leftrightarrow \mathbf{S}} \approx \eta_{A \leftrightarrow \mathbf{S}} > 0$. This feature is useful for excluding false variants during haplotype inference.

Local haplotype inference from linked-reads data

We first performed haplotype inference on the RPE-1 and NA12878 linked-reads data. We restricted the analysis to bi-allelic single-nucleotide variant (SNV) sites as they constitute the majority of genetic variation and have better detection accuracy than more complex variations (insertion/deletion/structural variants). We accumulated all linkage between SNVs within 100 kb into the coupling matrix given by Eq. (2) and solved for haplotypes by minimizing Eq. (4), which produced a single haplotype for each chromosome. (More technical details are provided in Materials and Method). We assessed phasing confidence at each variant site using the energy penalties calculated from the final haplotype solution (higher penalty implies higher confidence): The spin-flipping penalty (Eq. (9)) measures the probability that the phased genotype at a variant site is incorrect; the block-switching penalty (Eq. (10)) measures the likelihood that there is a switching from one haplotype to the complementary haplotype before and after a variant site.

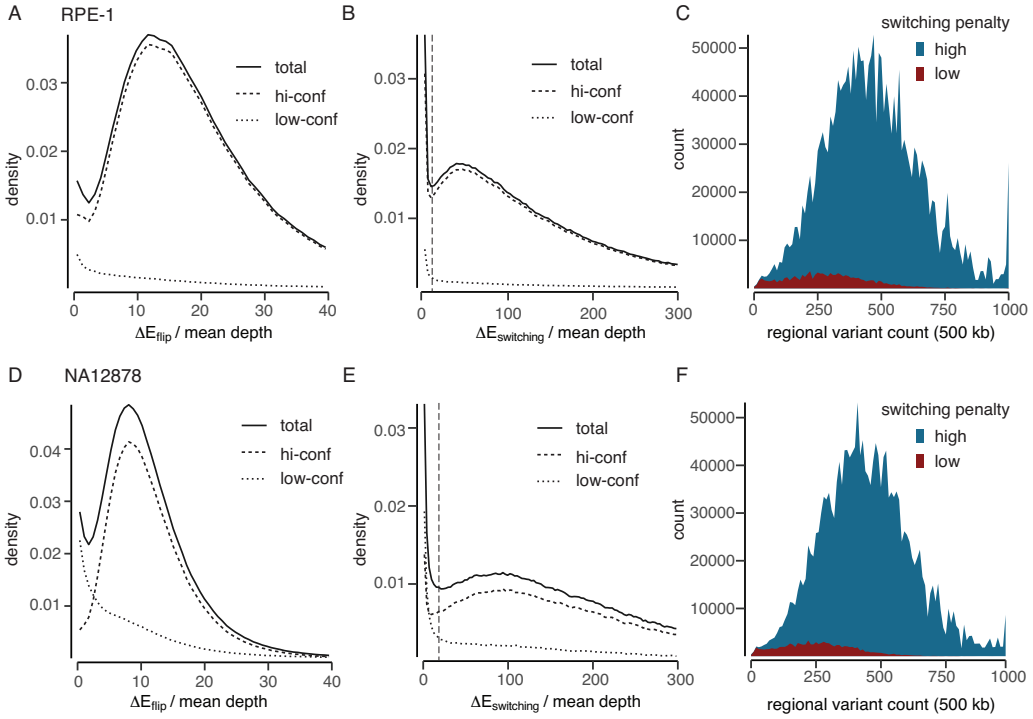

Figure 4. Quality metrics of the haplotype solution inferred from linked-reads data of RPE-1 (A and B) and NA12878 (D and E) genomes. Phasing confidence is measured by the energy penalty of single base (A and D) and block-switching (B and E) errors. The energy penalty is normalized by the mean sequencing depth. For each type of phasing error, the distributions of penalty scores are shown for all variants (solid line) and separately for high-confidence variants (dashed line) and low-confidence variants (dotted line). We further show a histogram of local variant density near each variant site (estimated by the number of high-confidence variants within a 500 kb neighborhood) in the RPE-1 genome (C) and in the NA12878 genome (F). The local variant density near variant sites with low switching penalty scores (low phasing confidence) are shown in dark red; those with high switching penalty scores (high phasing confidence) are shown in dark green.

Figure 4 summarizes the key metrics of spin-flipping and block-switching energy penalties of the haplotype solution from the RPE-1 (Fig. 4A-C) and the NA12878 data (Fig. 4D-F). Both penalty scores show bimodal distributions. The peak at $\Delta E = 0$ consists of sites with low phasing confidence, which can be caused by either sparse linkage in regions of low variant density or conflicting linkage due to false heterozygosity. To distinguish between these two scenarios, we show separate distributions of the phasing quality scores for high-confidence and low-confidence heterozygous variant sites. High- and low-confidence variants were determined using independent data as follows: For the RPE-1 genome, we selected high-confidence sites based on the average allelic fractions in a population of

single-cell genomes. (See Table 5 for the allelic fraction thresholds for high-confidence variants.) For the NA12878 genome, we chose trio-phased variants in the data released by the Genome-In-A-Bottle (GIAB) consortium as high-confidence sites.(Note: The GIAB data do not contain phased variants in several regions in the NA12878 genome. These regions are marked as low density regions in Fig. S4. We put all heterozygous variants in these regions in the low-confidence group even though a large fraction of these variants are real. This does not affect the benchmarking comparison but results in fewer variants in the truth set for a couple of chromosomes.)

We see that low-confidence variants overall have low phasing confidence (low penalty scores). In particular, low-confidence variants comprise a large fraction of sites with low spin-flipping penalty scores (Fig. 4A,D). This prompts us to use the spin-flipping energy score to exclude low-confidence variants from haplotype inference (marked as $s_i = 0$) with a threshold determined by the location of the local minimum in the spin-flipping energy distribution ($\Delta E \approx 3 \times$ mean depth in both samples).

Even after excluding low-confidence variants, we still observe a large number of sites with low confidence with regard to block-switching errors (Fig. 4B,E). The low phasing confidence at these sites cannot be attributed to conflicting linkage as the majority of links between variants within 100 kb are consistent with *cis* linkage (Fig. S1E,F). A more plausible explanation is that these sites are located in regions of low variant density and have sparse linkage to other variants. To test this hypothesis, we plot histograms of local-variant density at variant sites with low or high phasing confidence (determined by the switching penalty score relative to the local minimum in the switching penalty distribution) (Fig. 4C,F). The histograms show that sites with low switching penalty are enriched in regions of low variant density (dark red); in particular, all sites with regional variant density less than 25 variants per 500 kb are in the low-confidence group. We note that low variant density can reflect either less genetic variation (the most obvious example being Chr.X) or limited

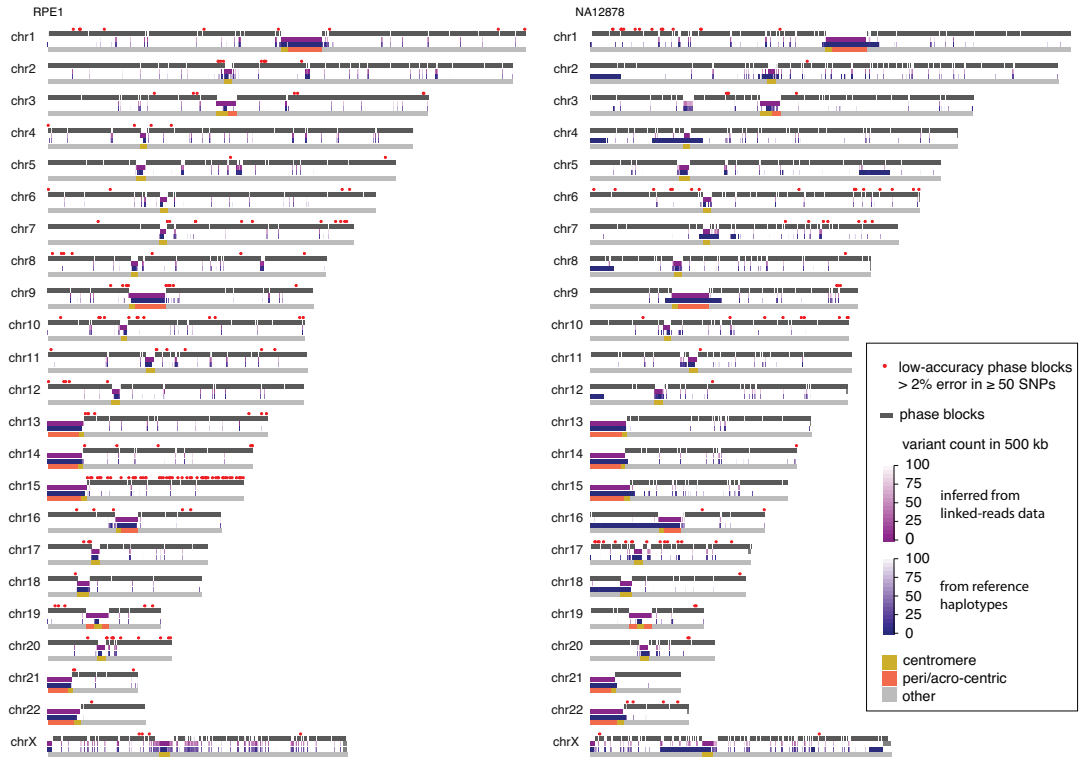

Figure S4. Genome-wide map of haplotype blocks determined from the RPE-1 and NA12878 linked-reads data. For each chromosome, five tracks are shown (bottom to top): centromere bands, regional density (500 kb bins) of high-confidence variants in the reference haplotype data, regional density (500 kb bins) of phased variants in the haplotype solution from linked-reads data, haplotype blocks, low-accuracy haplotype blocks (< 98% accuracy in blocks with at least 50 variants).

variant-detection sensitivity in low-complexity or highly polymorphic sequences (*e.g.*, pericentric and subtelomeric regions). To better resolve haplotype linkage in these regions requires different strategies such as long-read sequencing or *de novo* assembly.

We generate high-confidence haplotype blocks by disassembling the haplotype solution at sites with low block-switching penalty scores. The haplotype blocks are shown in Fig. S4 with their cumulative length distributions shown in Fig. S5A. The threshold switching penalty is determined by the location of local minimum in the switching penalty distribution (dash lines in Fig. 4B and E). For the RPE-1 data, the threshold of $\Delta E = 700$ ($\approx$ 12× mean depth) results in haplotype blocks with an N50 length of 0.3 Mb. For the NA12878 data, the threshold of $\Delta E \approx$ 20× mean depth results in haplotype blocks with N50 lengths of 0.5 Mb using a single linked-reads library and 0.8 Mb using both libraries (each 35×). The longer N50 length of haplotype blocks in the NA12878 data is consistent with the larger size of input DNA (Fig. S1).

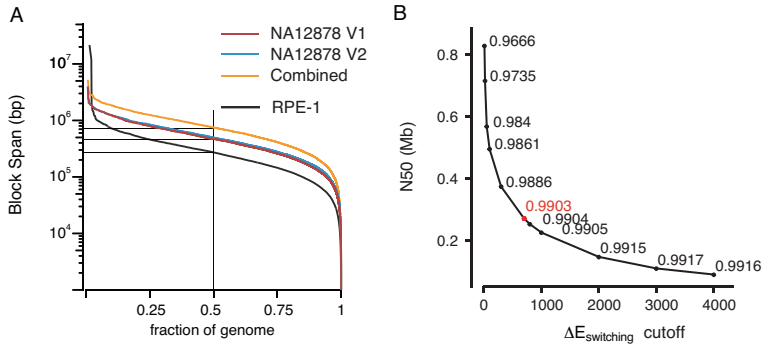

Figure S5. N50 lengths of haplotype blocks determined from linked-reads data of RPE-1 and NA12878 genomes. A. Cumulative distribution of the fraction of phased genome against the length of haplotype blocks determined using the stringent cutoff (dash lines in Fig. 4B and E). The N50 length of haplotype blocks determined from the RPE-1 data is 0.3 Mb. The N50 length of haplotypes determined from a single NA12878 library is 0.5 Mb and 0.8 Mb when both NA12878 libraries are used. B. Dependence of the N50 length of haplotype blocks on the switching-energy cutoff in the RPE-1 haplotype solution. The accuracy (shown above the curve) is calculated as the fraction of sites correctly phased within each haplotype block.

The N50 lengths of haplotype blocks determined from the NA12878 data in our analysis (0.5-0.8 Mb) are much shorter than what was reported (5-8 Mb) in a previous study (***Marks et al., 2019***). We attribute this difference to be primarily due to our choice of more conservative cutoffs for controlling haplotype-switching errors. More permissive cutoffs result in significantly longer haplotype blocks. For example, the threshold of $\Delta E = 700$ chosen for the RPE-1 data gives an overall phasing accuracy of 99% (red dot) between all adjacent variants; the overall phasing accuracy reduces to 98.4% at a lower threshold $\Delta E = 50$ and 97.3% at $\Delta E = 20$ (Fig. S5B). Meanwhile, the N50 length of haplotype blocks increases from 0.25 Mb ($\Delta E = 700$) to 0.6 Mb ($\Delta E = 50$) and 0.7 Mb ($\Delta E = 20$). As the main purpose of local haplotype inference is to generate high confidence haplotype blocks that can be concatenated by long-range Hi-C links, our choice of conservative cutoffs reflects the priority of phasing accuracy over block size.

Concatenating local haplotype blocks using Hi-C data

The local haplotype blocks determined from linked-reads data were assembled using Hi-C links (Hi-C reads spanning variant sites) between blocks. We first concatenated haplotype blocks within each chromosome arm using Hi-C linkage by the same minimization strategy as local haplotype inference. (More technical details can be found in Materials and Method.) Only Hi-C links between variants separated by ≤10 Mb were included to reduce the fraction of *trans* linkage (Fig. 2E,F). We then joined p- and q-arm haplotypes using all Hi-C links between the arms. To expedite convergence of the haplotype solution, we excluded short haplotype blocks with ≤ 5 total Hi-C links to other blocks from the calculation. The haplotype solution excluding short blocks is later referred to as the "scaffold haplotype." Most sites in the scaffold haplotype should have high phasing confidence.

We further calculated the linkage between variant genotypes and the scaffold haplotype according to Eq. (12) ("haplotype linkage"). Haplotype linkage was used to refine the scaffold haplotype and phase additional variant sites, producing the final haplotype solution.

The consistency between the haplotype solution and Hi-C linkage can be verified from the ratio between *cis* and *trans* Hi-C links between blocks. As an example, Fig. S6A and Fig. S6C show that the majority of Hi-C linkage between haplotype blocks on Chr.X is consistent with their haplotype phase in the final solution. A more convenient strategy is to compare *cis* and *trans* Hi-C links between chromosome arms. If the p- and q-arm haplotypes are given by $\mathbf{H}_p$ and $\mathbf{H}_q$, then the chromosomal haplotype $\mathbf{H}$ is either $(\mathbf{H}^p, \mathbf{H}^q)$ or $(\mathbf{H}^p, -\mathbf{H}^q)$: Hi-C links consistent with these two possibilities reflect either *cis* (major) or *trans* (minor) contacts. The dominance of *cis*-linkage implies a large disparity in the number of Hi-C links in each group, as is shown in Fig. S6B (RPE-1) and Fig. S6D (NA12878). The presence of long-range switching errors in the haplotype of either arm will bring the ratio between *cis* and *trans* links closer to 1.

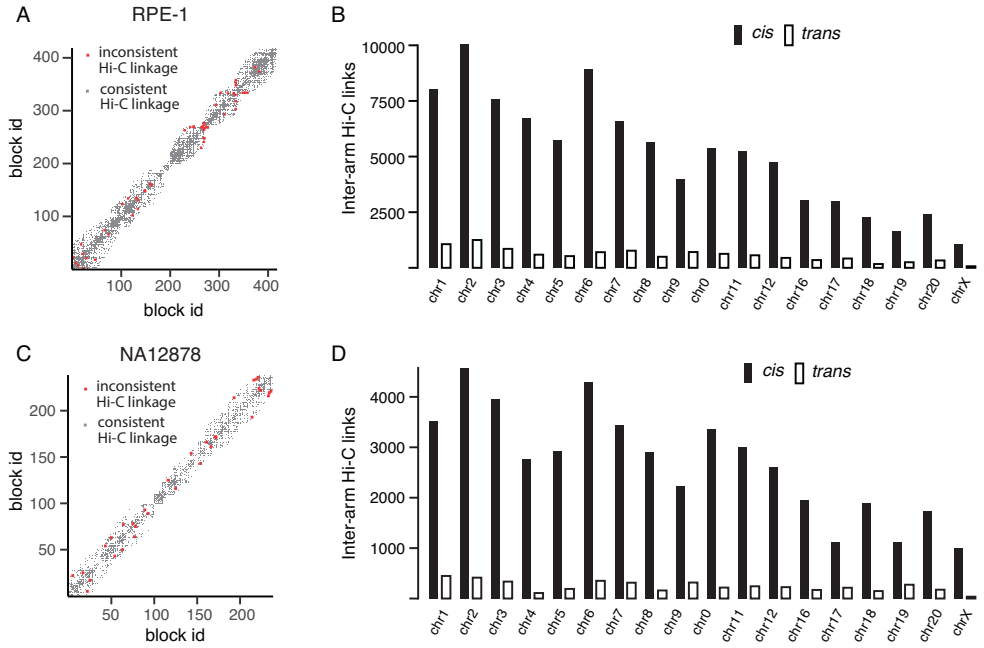

Figure S6. Consistency between Hi-C linkage and the haplotype solution. A and C. Comparison of Hi-C linkage between haplotype blocks on Chr.X to their haplotype phase in the RPE-1 (A) and in the NA12878 data (C). B and D. Total number of *cis* and *trans* Hi-C links between p- and q-arm haplotypes in the RPE-1 data (B) and in the NA12878 data (D).

Evaluation of the accuracy and completeness of haplotype inference

We evaluated the accuracy and completeness of the computationally inferred haplotypes using independent reference haplotype data (Table 1). For the NA12878 genome, we used the trio-phased haplotypes as the truth data. We first benchmarked the scaffold haplotype solution constructed from large haplotype blocks (Table 2). The scaffold haplotype solution includes 94% of all variants in the truth data with 99.6% accuracy. (Chr. 19 has the lowest accuracy of 98.4%.) We further evaluated the accuracy of local haplotype blocks determined only using the linked-reads data. Haplotype blocks with ≥50 variants but <98% accuracy are highlighted with red dots in Fig. S4: These blocks are rare and scattered across the genome, indicating consistent accuracy across each chromosome. The final haplotype solution determined from variant linkage to the scaffold haplotype contains 129,930 additional phased variants and shows 99.6% accuracy and 98.2% completeness in comparison to the reference haplotype (Table 3).

Table 2. Comparison between the scaffold haplotype solution and reference haplotypes of NA12878

ref_id	Total variant sites[a]	Phased from parents[b]	Phased from bulk data[c]	Comparable sites	Fraction of agreement	Fraction of completion[d]
chr1	193,281	145,417	161,679	137,088	0.992	0.936
chr2	208,196	147,402	175,237	139,069	0.999	0.943
chr3	175,998	133,236	145,418	126,411	0.999	0.948
chr4	179,803	115,676	153,510	110,215	0.999	0.952
chr5	164.837	120,290	141,897	114,664	0.999	0.952
chr6	174,820	132,441	153,729	126,867	0.986	0.945
chr7	147,495	108,363	126,641	103,199	0.996	0.948
chr8	136,150	94,908	117,927	89,921	0.999	0.946
chr9	119,635	84,264	101,213	80,633	0.996	0.953
chr10	133,655	96,743	112,378	91,878	0.992	0.942
chr11	125,208	92,565	103,672	87,544	0.999	0.945
chr12	120,828	85,634	102,808	81,066	0.999	0.946
chr13	89,785	71,352	79,064	68,771	0.999	0.963
chr14	81,921	63,767	70,690	60,446	0.999	0.947
chr15	72,770	51,991	58,857	49,057	0.999	0.943
chr16	80,402	36,561	67,847	34,835	0.994	0.947
chr17	68,924	45,363	55,341	42,288	0.989	0.922
chr18	73,239	43,961	62,585	41,558	0.999	0.945
chr19	63,856	33,739	43,452	31,699	0.984	0.924
chr20	68,144	42,642	50,240	40,348	0.998	0.944
chr21	44,623	26,893	31,226	25,749	0.999	0.957
chr22	44,187	24,229	31,282	23,092	0.990	0.943
chrX	84,624	51,414	61,548	45,017	0.999	0.875
Total	2,652,381	1,848,821	2,208,241	1,751,415	0.996	0.943

[a]heterozygous single-nucleotide variants detected by HaplotypeCaller with no filtering
[b]phased variants released by the Genome-In-A-Bottle consortium, excluding sites in centromeric regions
[c]scaffold haplotype solution
[d]percentage correctly phased sites in the reference data

Table 3. Comparison between the final haplotype solution and reference haplotypes of NA12878

Total variant sites	Phased from bulk data[a]	Phased from parental data	Comparable sites	Agreed	Accuracy	Fraction of completion
2,652,381	2,338,171	1,848,821	1,824,112	1,816,453	0.996	0.982

[a]Including both the scaffold haplotype and variants phased by their linkage to the scaffold haplotype

For the RPE-1 genome we used chromosomal haplotypes determined directly from the sequencing data of monosomic RPE-1 cells as an independent reference. The scaffold haplotype solution includes 89% of all phased variants in the reference haplotypes and shows 98% agreement. (Chromosome 9 has the lowest percentage of agreement of 96.2%.) Although there are more haplotype blocks with less than 98% agreement on Chr.15 and Chr.20 than other chromosomes (Fig. S4), the overall percentage of agreement for these two chromosomes is still above 96%. The final haplotype solution determined using haplotype linkage contains 139,412 additional phased variants but shows a modest increase in the percentage disagreement from 1.8% to 3.6% (Table 5).

Although the RPE-1 haplotype data show more disagreement than the NA12878 data, this does not necessarily reflect lower accuracy of the computational inference because the reference RPE-1 haplotypes derived from single-cell sequencing also contain errors (*e.g.*, accumulated during single-cell genome amplification). In particular, neither dataset has an explicit filter for false heterozygous variants, whose presence will contribute more discordance. To exclude false variants, we introduced a filter based on the variant allele fraction observed in all the RPE-1 single cell data (>100 samples). This filter removed variants whose allele fractions deviate significantly from the expectation based on the DNA copy number. Among the 2,172,689 variants passing this filter, 2,131,012 (98%) are phased by computational inference and 2,109,589 (97%) are phased by monosomy sequencing,

Table 4. Comparison between the scaffold haplotype and reference haplotypes of RPE-1

ref_id	Total variant sites	Sites for phasing[a]	Phased from bulk data[b]	Phased from monosomies[c]	Comparable sites	Agreed	Fraction of agreement
chr1	185,997	175,806	163,062	179,216	160,291	157,619	0.983
chr2	194,641	186,340	174,640	188,297	171,784	168,516	0.981
chr3	169,194	153,681	144,579	157,887	137,380	134,878	0.982
chr4	168,626	162,344	153,751	161,562	149,932	148,714	0.992
chr5	149,212	143,375	134,514	146,447	133,049	131,996	0.992
chr6	156,723	149,646	142,482	155,379	141,808	140,555	0.991
chr7	143,405	137,658	129,741	139,365	127,635	123,992	0.971
chr8	124,489	121,157	113,688	120,140	110,678	109,590	0.990
chr9	115,189	109,208	102,110	110,993	100,042	96,279	0.962
chr10	119,524	111,765	105,803	112,758	102,832	100,604	0.978
chr11	116,321	108,386	101,840	113,375	100,707	99,384	0.987
chr12	111,225	105,643	99,591	106,331	96,569	94,553	0.979
chr13	84,586	82,366	78,232	83,092	77,149	76,496	0.992
chr14	74,463	72,190	67,351	71,195	65,027	64,172	0.987
chr15	70,936	66,347	61,721	62,662	55,448	53,586	0.966
chr16	78,049	73,164	68,373	70,375	62,875	61,448	0.977
chr17	75,407	69,479	64,875	60,418	52,916	51,624	0.976
chr18	64,006	61,019	57,706	61,622	57,072	56,230	0.985
chr19	61,461	44,384	41,468	26,906[d]	18,480	18,080	0.978
chr20	61,029	49,817	46,970	55,063	44,265	42,769	0.966
chr21	44,178	33,227	31,892	36,240	28,219	27,857	0.987
chr22	40,635	30,275	28,131	36,154	26,104	25,785	0.988
chrX	66,015	62,985	52,486	64,676	51,855	51,035	0.984
Total	2,475,311	2,310,262	2,165,006	2,320,153	2,072,117	2,035,762	0.982

[a]Excluding sites in centromeric regions or not having linkage to both reference and alternate genotypes
[b]Scaffold haplotype constructed from linked reads and Hi-C sequencing
[c]Whole-chromosome haplotype determined from single-cell sequencing of monosomies
[d]This is generated from one monosomic sample only, resulting in a lower percentage of phased variants.

Table 5. Comparison between the final haplotype solution and reference haplotypes of RPE-1 using different variant filters

Filter	Total variant sites	Phased from bulk data	Phased from monosomies	Agreed	Discordant	Fraction of discordance
None	2,475,311	2,304,418	2,320,153	2,122,574	78,307	0.036
Allele fraction[a]	2,172,689	2,131,012	2,109,589	2,031,208	43,018	0.021
Allele fraction[b]	2,153,678	2,148,402	2,069,123	2,017,695	46,625	0.023
Linkage[c]	2,116,557	2,116,536	2,033,513	2,013,413	20,085	0.010
Combined[d]	2,017,897	2,017,897	1,943,901	1,933,260	10,641	0.006

[a]From single-cell data:
[b]From haplotype linkage from bulk data:
minor allele fraction ≥ 0.3 in disomic regions and $\in [0.18, 0.48]$ in the trisomic region in Chr.10q.
[c]≥ 1 link connecting ref, alt, HapA, and HapB & minor linkage ≤ 2 or minor linkage/total linkage ≤ 0.1
[d]With both the allele fraction (b) and the linkage (c) filter from bulk data

with an overall agreement of 98% between the two data (Table5). In particular, 2,018,878 out of 2,165,006 (93.2%) phased variants in the scaffold haplotype solution pass the allele fraction filter, confirming that our phasing algorithm can effectively exclude false variants from haplotype inference by implicitly penalizing conflicting linkage evidence (Eq. (3)). We next describe a more specific variant-filtering strategy solely based on haplotype linkage.

Application I: Variant filtration using haplotype linkage

We expect variant linkage from true heterozygous sites to have the following features: First, links from the reference genotype and links from the alternate genotype should map to complementary

haplotypes ("linkage filter"). Second, the number of links to each haplotype should be proportional to DNA copy number of that haplotype; in disomic regions, the allelic ratio should be roughly 1:1 ("allele fraction filter"). These filters are implemented using haplotype linkage

$$\eta_{\mathrm{rA}} = \#(\text{reference-haplotype A}) + \#(\text{alternate-haplotype B})$$
$$\eta_{\mathrm{rB}} = \#(\text{reference-haplotype B}) + \#(\text{alternate-haplotype A}),$$

and allelic linkage

$$\rho_{\mathrm{A}} = \#(\text{reference-haplotype A}) + \#(\text{alternate-haplotype A})$$
$$\rho_{\mathrm{B}} = \#(\text{reference-haplotype B}) + \#(\text{alternate-haplotype B}).$$

The linkage filter implies

$$\max(\eta_{\mathrm{rA}}, \eta_{\mathrm{rB}}) \gg \min(\eta_{\mathrm{rA}}, \eta_{\mathrm{rB}}).$$

For the RPE-1 data, we impose the following criteria:

$$\min(\eta_{\mathrm{rA}}, \eta_{\mathrm{rB}}) \leq \max[2, 0.1 \times (\eta_{\mathrm{rA}} + \eta_{\mathrm{rB}})]$$

and further require at least one link connecting each of the following: reference genotype, alternate genotype, haplotype A, and haplotype B. The allele fraction filter implies

$$\rho_{\mathrm{A}}/\rho_{\mathrm{B}} \approx 1,$$

and we use the following thresholds for the allele fractions:

$$0.3 \leq \frac{\rho_{\mathrm{A,B}}}{\rho_{\mathrm{A}} + \rho_{\mathrm{B}}} \leq 0.7$$

in disomic regions, and

$$\left| \frac{\min(\rho_{\mathrm{A}}, \rho_{\mathrm{B}})}{\rho_{\mathrm{A}} + \rho_{\mathrm{B}}} - 0.33 \right| \leq 0.15$$

in the trisomic region in Chr.10q.

Each filter removes about 13% of variants (Table 5). Although the two filters utilize different features of molecular linkage, most variants either pass or fail both criteria, indicating good consistency between these independent filters. Among variants passing both filters, the percentage of agreement between the haplotypes determined computationally and those determined by monosomy sequencing is 99.4%. This indicates similar accuracy of the RPE-1 haplotype solution as the NA12878 haplotype solution.

Although we demonstrate the utility of haplotype-linkage based variant filtering using single-nucleotide variants, the same filtering strategy can be directly applied to variants that are more sensitive to alignment errors, including indels, structural variants, and any type of variants in repetitive sequences or in sequences that show significant divergence from the reference genome.

Application II: Digital karyotyping using phased Hi-C reads

The determination of complete haplotype information enables the analysis of genetic or epigenetic features that are specific to each homologous chromosome. For example, comparison of Hi-C contacts phased to active and inactive X chromosomes has provided important insight into the organization of the inactive X chromosome and its regulation (***Darrow et al., 2016***). Here we describe an application using Hi-C data to infer the 2D structure of chromosomes ("digital karyotyping").

Hi-C contact map has previously been used for detecting long-range chromosomal rearrangements (***Rao et al., 2014***; ***Dixon et al., 2018***; ***Zhou et al., 2019***). The formation of new junctions between distal loci (separated by > 1 Mb genomic distance or located on different chromosomes) results in a significant increase in the contact density between these loci, creating a specific signal for detecting such events. With haplotype information, we can extend this idea further and detect chromosomal rearrangements from the density map of phased Hi-C reads. As most translocations

involve one homolog only, the increase in contact density should only be seen in one out of two (for intra-chromosomal rearrangements) or one out of four (for inter-chromosomal rearrangements) phased Hi-C contact maps between different homologs. For example, a translocation between the A homolog of Chr.1 and the A homolog of Chr.2 will only result in an increase in Hi-C contacts between the A homologs:

$$\underline{\#\text{A-A}} \gg \#\text{A-B}, \#\text{B-A}, \#\text{B-B}.$$

For "semiphased" Hi-C reads (*i.e.*, with only one end phased), we expect

$$\#\text{A-unphased} = \underline{\#\text{A-A}} + \#\text{A-B} \gg \#\text{B-unphased} = \#\text{B-A} + \#\text{B-B};$$
$$\#\text{unphased-A} = \underline{\#\text{A-A}} + \#\text{B-A} \gg \#\text{unphased-B} = \#\text{A-B} + \#\text{B-B}.$$

Using non-specific Hi-C contacts between intact homologs as control further increases the sensitivity of the signal of *cis* contacts on the rearranged chromosome. Here we demonstrate this using examples from the RPE-1 genome and the K-562 genome (Fig. 5).

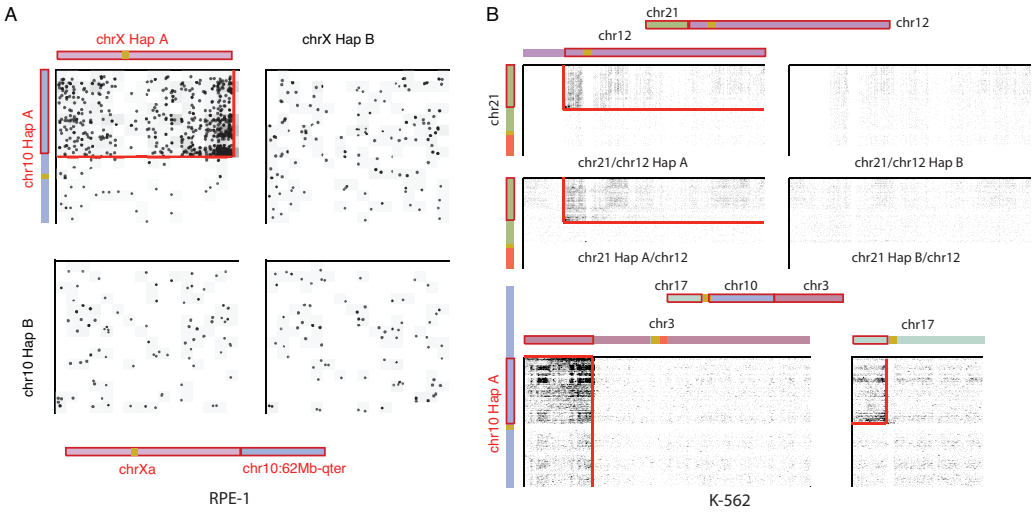

Figure 5. Digital karyotyping using phased Hi-C reads. A. Phased Hi-C links (dots) and Hi-C link density (10 Mb bins) in four different haplotype combinations between Chr.10 and Chr.X of RPE-1 cells. A significant enrichment of Hi-C links is seen in only one haplotype combination, reflecting the translocation between a single homolog of Chr.10 and Chr.X. The enrichment of Hi-C links is seen throughout the whole X chromosome on the translocated homolog, reflecting *cis*-linkage between Chr.X and 10q that is consistent with the derivative chromosome shown below. There is also a modest increase of *trans* Hi-C links between haplotype A of Chr.10q and haplotype B of Chr.X due to elevated DNA copy number of Chr.10q. B. Semi-phased Hi-C links (dots) in the K-562 genome supporting two rearranged chromosomes, der(12)t(12;21) (top) and der(10;17)t(3;10;17) (bottom). Top: The density maps of semi-phased Hi-C links between Chr.12 and Chr.21 are shown in four panels and the signal of *cis* linkage on the derivative chromosome der(12)t(12;21) (whose structure is shown above the density maps) is seen in only one homolog of each chromosome. (The p-terminus of Chr.12 has undergone loss-of-heterozygosity and is omitted in the haplotype-specific plots.) Bottom: The density maps of semi-phased Hi-C links show two translocations between a single Chr.10 homolog and both Chr.3 and Chr.17. A modest increase in the density of Hi-C contacts between Chr.3 and Chr.17 (data not shown) verifies that these two translocations occur on a single derivative chromosome. The structure of the derivative chromosome der(10;17)t(3;10;17) is shown above the density maps.

The RPE-1 genome contains an extra copy of a segment of Chr.10q (~62 Mb-qter). Sequence information near the breakpoint on Chr.10q cannot resolve the partner locus, but cytogenetic analysis indicates that the 10q segment is translocated to the q-terminus of Chr.X. On the phased Hi-C contact maps, this translocation creates an increase in Hi-C contacts only in one haplotype combination (Fig. 5A). The translocation junction can be identified from the site with the highest density of Hi-C links (~62 Mb on Chr.10 and the q-terminus of Chr.X). Remarkably, the enrichment

of Hi-C contacts persists through Chr.X and reaches the p-terminus. This indicates that the 10q segment is fused to a complete X chromosome. Therefore, examination of the phased Hi-C contact map exactly recapitulates the cytogenetic map.

The K-562 genome is highly aneuploid with multiple marker chromosomes (***Gribble et al., 2000***; ***Naumann et al., 2001***) and copy-number alterations (***Zhou et al., 2019***). In particular, many regions of the genome undergo loss-of-heterozygosity and do not contain informative phasing information. We determined the haplotype phase in heterozygous regions using published linked-reads data (https://www.encodeproject.org/experiments/ENCSR053AXS/) and Hi-C data (***Rao et al., 2014***) and used semi-phased Hi-C reads to infer the structure of two derivative chromosomes (Fig. 5B).

The upper panel in Fig. 5B shows the density maps of semi-phased Hi-C reads between Chr.12 and Chr.21. The p-terminal segment of Chr.12 is hemizygous and omitted in the haplotype-specific plots. Semi-phased Hi-C reads are enriched in only one haplotype of Chr.12 and Chr.21 and consistent with a single rearranged chromosome schematically shown above the Hi-C maps. The lower panel shows the density maps of semi-phased Hi-C links between Chr.10 and Chr.3, and between Chr.10 and Chr.17. The hemizygous region on Chr.10q is omitted in the plots. Enrichment of semi-phased Hi-C links suggests two translocations involving Chr.10: the first joins the p-terminal segment of Chr.3 to Chr.10q; the second joins the 17p arm to the Chr.10 centromere. We further observe a modest increase in the density of Hi-C contacts between Chr.3 and Chr.17 near the translocation sites to Chr.10 (not shown). Together, these data indicate the presence of a derivative chromosome whose structure is shown above the Hi-C maps. Although the large-scale structure (~ 5 Mb) of these derivative chromosomes were previously resolved by cytogenetic analyses (***Gribble et al., 2000***; ***Naumann et al., 2001***), Hi-C maps significantly improve the resolution of rearrangement junctions and further reveal the long-range linkage between these junctions.

Application III: Resolving genetic heterogeneity by haplotype copy number

Haplotype information enables separation of the read depth signals from homologous chromosomes, which can significantly improve DNA copy number analysis (***Nik-Zainal et al., 2012***; ***Zhang and Pellman, 2015***; ***Loh et al., 2018***). In comparison to the total read depth signal, the ratio between haplotype-specific read coverage is insensitive to non-uniformity of read coverage due to sequence content; the haplotype ratio can also be calculated as an average over over multiple variant sites to suppress statistical noise at individual variant site.

To demonstrate the utility of haplotype-resolved copy-number analysis, we re-analyzed whole-genome sequencing data from a previous study that sought to elucidate the mutational consequences of telomere crisis (***Maciejowski et al., 2015***). The authors performed whole-genome sequencing on the progeny of single cells that underwent telomere crisis. We downloaded and processed the sequencing data in this study and calculated haplotype coverage in 250 kb bins as

$$C^{(i)}_{A,B} = D^{(i)} \cdot \overline{R^{(i)}_{A,B}},$$

where $D^{(i)}$ is the normalized average sequence coverage in the ith bin ($\overline{D^{(i)}} = 1$) and $\overline{R^{(i)}_{A,B}}$ is the mean haplotype fraction (A or B) across all variants within the ith bin. We then calculated haplotype DNA copy number by normalizing $C_{A,B}$ by its median across the whole genome (from both haplotypes). The median corresponds to the expected (normalized) coverage of a single chromosome (haplotype) assuming the genome is mostly diploid.

Figure 6 shows the haplotype-specific copy number of three chromosomes from three different samples that each underwent telomere crisis and displayed complex genomic alterations. Figure 6A shows an example of chromothripsis with rearrangements restricted to the 8q arm accompanied by copy-number alterations affecting the red haplotype only; the non-integer copy-number states of the red haplotype between 1 and 2 on the 8q arm implies that alterations are present in a subclonal population (~45%) given that the karyotype of this sample is close to diploid; by the same argument, loss of the red haplotype on 8p is also subclonal (~50%). Whether the rearrangement

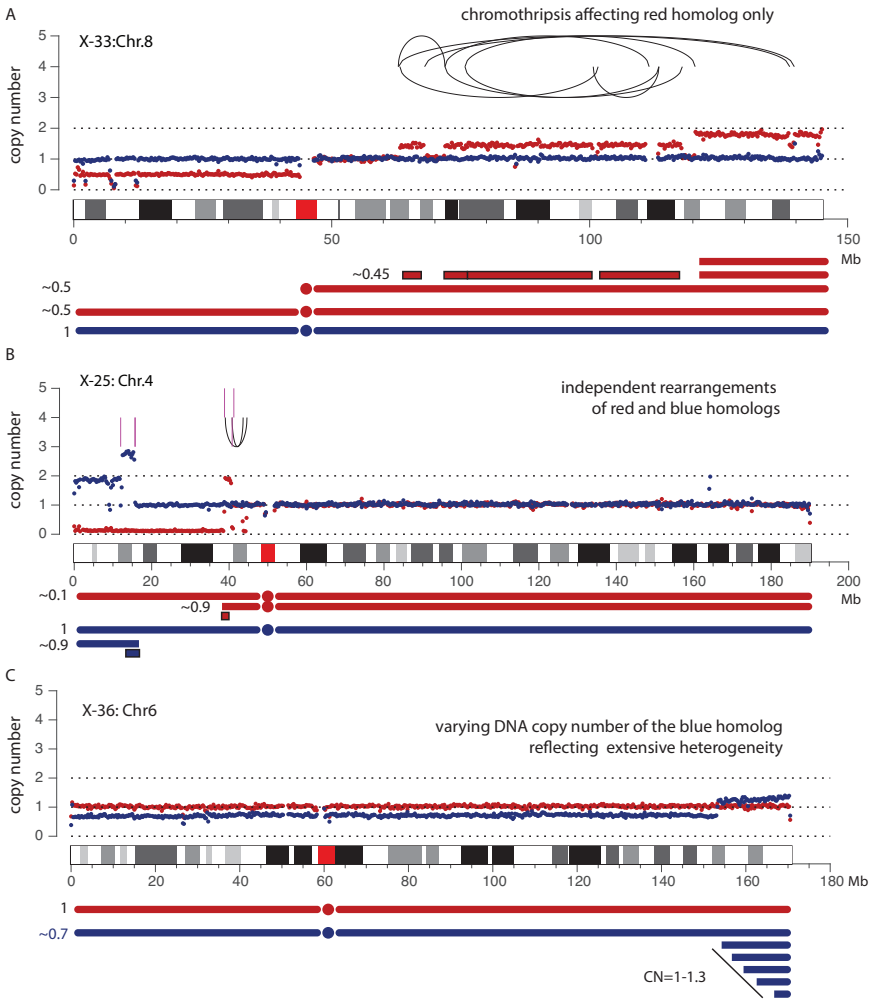

Figure 6. Haplotype-resolved DNA copy number and rearrangement analysis of bulk RPE-1 Sequencing data generated by ***Maciejowski et al.*** (***2015***). Schematic diagrams of alterations to each homolog and their clonal fractions are shown below the copy-number plots. Chromosomal fragments involved in complex rearrangements are outlined. A. Haplotype-specific DNA copy number (blue and red dots for two parental haplotypes) and rearrangement (black arcs: intra-chromosomal; magenta vertical lines: inter-chromosomal) of Chr.8 in the X-33 sample. Rearrangements on the 8q arm are accompanied by copy-number changes of the red homolog; the non-integer changes between different copy-number states of the red homolog indicate that the rearrangements are only present in a subclonal population (∼45%). B. Haplotype-specific DNA copy number and rearrangement of Chr.4 in the X-25 sample. Rearrangements involve both homologs as inferred from haplotype-specific copy number. Moreover, the terminal loss on the p-arm of the red homolog is subclonal, with ∼10% of cells retaining the whole p-arm. C. Haplotype-specific copy number of Chr.6 in the X-36 sample. The q-terminus of the blue homolog shows non-constant copy number (1-1.3) that contrasts with the stable copy number of the red homolog or the rest of the blue homolog (∼0.7). We interpret this pattern as reflecting extensive genetic heterogeneity in the population.

on 8q was present in the ancestor cell but subsequently lost in a sub-population, or it was not present in the ancestor cell but arose in a sub-population during clonal outgrowth cannot be determined. Figure 6B shows an example that the authors classified as chromothripsis. The haplotype copy number suggests that the rearrangements on Chr.4 involve both homologs: the alterations between 10 and 20 Mb affect the blue homolog and those near 40 Mb affect the red homolog; these alterations are unlikely to be the direct outcome of a single dicentric chromosome since they are not joined together but connected Chr.13 (magenta lines). The large terminal loss on the red homolog between 0 and 40 Mb most likely have resulted from breakage of a dicentric

chromosome, but this must have been a secondary event (as opposed to being present in the ancestor cell) as the loss is subclonal (in ∼90% of all cells). Finally, Figure 6C shows an example of extensive copy-number heterogeneity in the cell population. The blue homolog shows simultaneous subclonal loss (in∼30% of all cells) and segmental gain at the q-terminus. Interestingly, the terminal gain does not show a constant copy number state when compared to either the red haplotype (clonal single copy) or the rest of the blue haplotype (∼0.7). That the red haplotype shows a constant copy number excludes the possibility that the varying copy number of the blue haplotype (1-1.3) is due to technical artifacts. We therefore conclude that the varying copy number reflects extensive heterogeneity between cells that retain different portions of the terminal segment. Together, the results from haplotype copy-number analysis show that "clonal" populations derived from single cells passing through telomere crisis are not clonal after all: they frequently show subclonal heterogeneity and ongoing genome evolution. The genetic heterogeneity can be either caused by the formation of unstable chromosomes (dicentric or acentric), or attributed to the disruption of Rb and p53 pathways used to facilitate clonal outgrowth. The feature of non-constant haplotype copy number is particularly interesting and may be used to infer genome instability and assess genetic heterogeneity in a cell population directly from bulk DNA sequencing.

Discussion

Here we describe a computational method that can accurately determine complete chromosomal haplotypes from a combination of linked-reads sequencing (30-60× mean depth) and Hi-C sequencing (≥50 million long-range contacts). Our method achieves high accuracy (>99%) and completeness (>98%) in the inferred haplotypes in comparison to reference haplotypes determined by single-chromosome sequencing (RPE-1) or from parental genomes (NA12878) (Table 1).

Our method offers several advantages over previous methods. First, it uses data generated by two widely used bulk sequencing technologies and does not involve special experimental techniques required for single-chromosome isolation (***Ma et al., 2010***; ***Fan et al., 2010***; ***Yang et al., 2010***), single-cell sequencing (***Zhang et al., 2015***), or similar techniques such as "Strand-Seq" (***Falconer et al., 2012***; ***Porubsky et al., 2017***). Both linked-reads and Hi-C sequencing libraries can be constructed from clinical samples in addition to cultured cells. Second, the phasing algorithm automatically filters out false variants based on the specificity of haplotype linkage. This contrasts with previous methods (***Edge et al., 2017***) that require high-quality variants as input. The haplotype linkage can further be used for validating *de novo* variants that are prone to false detection to alignment inaccuracies, such as small insertion/deletions or structural variants.

Our formalism of haplotype inference as a minimization problem also has several unique features. The representation of binary haplotypes as 1D spin chains integrates linkage information from both homologous chromosomes and simplifies the inference of complementary haplotypes into one minimization problem. The haplotype inference algorithm is not affected by allelic imbalance, including loss-of-heterozygosity, and therefore can be applied to diploid genomes as well as aneuploid tumor genomes. Although we introduce a simple iteration strategy that can efficiently solve the haplotype in both examples and likely in most cases, incorporation of more sophisticated techniques (*e.g.*, Monte-Carlo methods) may further improve the efficiency and/or the accuracy of haplotype inference, especially if the linkage data contain more errors. One useful extension of our method is to perform joint haplotype inference using population genotypes and Hi-C data. Population-based statistical phasing (***Loh et al., 2016a***) can produce long haplotype blocks (>1 Mb) that contain random but rare switching errors. It should be possible to correct these errors using Hi-C data and extend the statistical haplotype phase to the entire chromosome.

With whole-chromosome haplotype information, one can jointly analyze variations in DNA sequence, epigenetic marks, chromatin structure, and allelic gene expression on each chromosome. This is especially relevant for the analysis of cancer genomes where homologous chromosomes often acquire independent alterations in the DNA sequence that can cause differential changes in chromatin folding and gene expression between the homologs (***Spielmann et al., 2018***). The

presence of large segmental or chromosome gains creates a new problem as the haplotype phase is no longer binary. This problem will be addressed in a forthcoming study.

Acknowledgments

Richardo Rengifo Pinto contributed to the C++ code of energy minimization. The single-cell sequencing data were generated in collaboration with David Pellman's laboratory by Hauke Cornils, Alexander Spektor and Neil Umbreit with help from Joshua Francis, Yingying Zhang, and Lili Sun at the single-cell genome center at the Dana-Farber Cancer Institute. The linked-reads data of RPE-1 cells were generated at the Yale Center for Genome Analysis (https://medicine.yale.edu/keck/ycga/) with help from Jinyu Wang and Lili Sun. We want to thank David Pellman, Heng Li, Po-Ru Loh, Matthew Meyerson, and members of Peter Park's laboratory for critical reading and feedbacks on the manuscript. The software package is maintained at (https://github.com/rwtourdot/linker). All the sequencing data will be deposited to the Short Read Archive along with a companion manuscript under preparation.

Materials and Methods

Linked-reads data of RPE-1 cells

High-molecular weight DNA from RPE-1 cells was extracted using the RevoluGen PuriSpin Fire Monkey kit following the protocol provided by the vendor with the following modifications: cells were lysed at 56°C for 2 hours, followed by addition of ∼100 ng RNase A and additional incubation for 15 minutes at 56°C. A single linked-reads library was constructed using the Chromium Genome Library Kit v2 from 10X Genomics following the standard protocol. The library was then sequenced on the Illumina NovaSeq platform to generate 941,518,426 read pairs with 60× mean depth of coverage. The mean size of input DNA was estimated to be 25 kb by the LongRanger software. See Table 1 for additional metrics of the sequencing data.

Sequencing of monosomic RPE-1 cells

Monosomic RPE-1 cells were generated using three different strategies: (1) Nocodazole block and release (***Zhang et al., 2015***); (2) Induction of dicentric chromosome bridges; (3) Treatment with Taxol. The experimental protocols for bridge chromosome induction is described in detail in a separate manuscript (https://www.biorxiv.org/content/10.1101/835058v1).

For the current study, we selected 39 cells: 32 from nocodazole release, 5 from bridge induction, and 2 from Taxol treatment, containing 98 monosomic chromosomes (some cells contain multiple monosomies) in total. The monosomic chromosomes were identified based on the observed heterozygosity normalized by the mean allelic coverage (***Zhang et al., 2015***):

$$\frac{\text{observed heterozygosity}}{(\text{observed allelic coverage})^2}.$$

The observed heterozygosity is defined as the fraction of heterozygous sites in the RPE-1 genome that show heterozygous coverage in a single cell p_{het}; the observed allelic coverage is determined from the average fraction of heterozygous sites that show either reference or alternate coverage $(p_{\text{ref}} + p_{\text{alt}})/2$. To eliminate "false heterozygosity" due to sequencing or amplification errors, we considered a variant site to show reference or alternate coverage only when the number of sequencing reads showing either genotype to exceed a threshold set as $d^* = \max(2, 1 + 0.1 \times$ mean sequencing depth of chromosome): $d^* = 2$ if the mean sequencing depth is $\leq 10\times$ (most samples) and $d^* = 4$ if the mean sequencing depth is 30×. Monosomic chromosomes were selected based on the criteria that the expected heterozygosity is less than $0.1\times$ the median from all the cells (the median is usually ≈ 1). Monosomic cells were sequenced to 5-30× on either the Illumina HiSeq 2500 or the Illumina NovaSeq platforms at the Broad Institute of MIT and Harvard.

Sequence data alignment

We used the LongRanger software from 10X Genomics to extract the molecular barcode of each sequencing fragment in the linked-reads data, which is preserved in the "BX" tag in the BAM record.

Instead of using the Lariat aligner (https://github.com/10XGenomics/lariat) in the LongRanger pipeline, we chose to align all sequencing data (both linked reads and Hi-C) using a standard short-read aligner (https://github.com/lh3/bwa) with default parameters ("`bwa mem`"). Using a barcode-agnostic aligner ensures better specificity of linkage information and therefore achieves better phasing accuracy than using a barcode-aware aligner. The rationale is explained below.

Linked-reads data contain two sources of linkage information. Direct physical linkage between sequence fragments derived from the same DNA molecule is embedded in the shared molecular barcode:

$$\text{physical linkage} \Rightarrow \text{same molecular barcode.}$$

Additional evidence of linkage is contained in the alignment positions: sequence fragments derived from the same DNA molecule should be aligned to proximal loci in the reference genome:

$$\text{physical linkage} + \text{correct alignment} \Rightarrow \text{proximity of alignment positions.}$$

A barcode-aware aligner (*e.g.*, Lariat) uses physical linkage information to help determine the best alignment positions of sequence fragments (***Marks et al., 2019***):

$$\left.\begin{array}{r}\text{identical molecular barcode} \dashrightarrow \text{physical linkage} \\ \text{sequence information}\end{array}\right] \rightsquigarrow \text{optimal alignment positions.}$$

This strategy assumes that **the molecular barcode information provides stronger linkage evidence than the sequence content**. This assumption is valid for sequence fragments derived from repeat regions, whose alignment positions cannot be uniquely determined from their sequences. However, the majority of 100-150 bp fragments derived from the human genome are unique and the alignment positions provide specific evidence about the adjacency of these fragments. By contrast, sharing the same molecular barcode does not guarantee singular molecular origin of the sequencing fragments since multiple long DNA molecules are encapsulated in each droplet and tagged by the same molecular barcode. Therefore, to **ensure the best specificity of linkage between sequence fragments**, we want to use the alignment positions as independent support of linkage:

$$\left.\begin{array}{r}\text{identical molecular barcode} \\ \text{proximity of alignment position (< 100 kb)}\end{array}\right] \Rightarrow \textit{cis}\text{ linkage between fragments.}$$

In the current study, we considered sequence reads to be linked in *cis* only when they both share the same molecular barcode and are also mapped to positions within 100 kb from each other independent of the molecular barcodes. The 100 kb threshold was determined from the distance dependence of linkage density and linkage accuracy in linked-reads data, which roughly corresponds to the maximum length of input DNA.

Using a barcode-aware aligner can improve alignment accuracy of sequence fragments sharing the same molecular barcode if (1) at least some sequence fragments in the group are aligned correctly; and (2) the aligner can output the confidence of alignment for each sequence fragment with reasonably accuracy. For sequence fragments derived from short interspersed repeats (typically <10 kb), they are often linked to fragments derived from the flanking non-repeat regions that can be aligned accurately; therefore, the barcode-aware alignment strategy will improve their alignment accuracy. However, for sequence fragments derived from regions of segmental duplications, or from regions not represented in the human genome reference (*e.g.*, centromeres), their alignment positions cannot be uniquely determined or are not included in the human genome reference after all. In these scenarios, a barcode-aware aligner may place linked sequence fragments near

each other but at an incorrect location, thus creating incorrect linkage information. By contrast, barcode-agnostic aligners usually assign a randomly picked position to a fragment with multiple possible hits, or place a fragment not contained in the reference to a locus with the most sequence similarity. The incorrect alignments of these fragments are often scattered across the genome and thus not producing usable linkage information that is restricted to fragments aligned within 100 kb. For haplotype phasing, the accuracy of linkage information is more important than the quantity and it is more approriate to terminate haplotype blocks at regions with low linkage accuracy than to aggressively extend them based on inaccurate linkage. We therefore choose the more conservative barcode-agnostic alignment approach and use the barcode-agnostic alignment positions as independent evidence to ensure the specificity of *cis*-linkage between sequence reads. We note that the choice of barcode-agnostic alignment may not be ideal for applications such as **variant discovery**, especially for **structural variant detection**, for which **the benefit of improved alignment accuracy is more significant**. It will therefore be convenient if a newly developed aligner can output both the (original) barcode-agnostic alignment positions and the barcode-aware alignment positions (if different) that are suited for different applications.

Post-alignment processing

We inferred the insert size distribution of sequence fragments from 2,000,000 unambiguously aligned sequence fragments (both mates having mapping quality 60, inferred insert size <2000, and forward-reverse orientation). We then chose the primary alignment positions for each aligned read pair based on the inferred insert size range (99.9% percentile). We inferred read pairs that were PCR duplicates based on the primary alignment positions using `MarkDuplicates` from Picard (https://broadinstitute.github.io/picard/). We then adjusted the duplicate tag of non-primary alignments based on the duplicate status of the corresponding primary alignments (with the same readname).

Variant calling and filtering

We ran `HaplotypeCaller` from GATK (v4.0.12.0-6-gfef36e3-SNAPSHOT) on the discovery mode ("`--genotyping-mode DISCOVERY`") to detect genetic variants in the RPE-1 genome jointly from the linked-reads data and standard bulk whole-genome sequencing data ($\sim 13\times$) published previously (***Zhang et al., 2015***). We imposed the following read filters in addition to the standard parameters and read filters used by `HaplotypeCaller` to exclude reads with improper or inaccurate alignment:

```
 --read-filter PairedReadFilter \
 --read-filter MateOnSameContigOrNoMappedMateReadFilter \
 --read-filter FragmentLengthReadFilter --max-fragment-length 1000 \
 --read-filter MateDifferentStrandReadFilter \
 --read-filter MappingQualityReadFilter --minimum-mapping-quality 30 \
 --read-filter OverclippedReadFilter --filter-too-short 25 \
 --read-filter GoodCigarReadFilter --read-filter AmbiguousBaseReadFilter
```

We selected bi-allelic single-nucleotide variant sites (one reference plus one alternate) as the input for haplotype inference, excluding sites in peri-/acro-/centromeric regions based on the standard chromosome banding annotation ("acen","gvar","stalk") provided by the UCSC genome browser. No other quality filter (such as variant quality score recalibration) was applied.

Extraction of variant linkage information from long-range sequencing

Our haplotype solver "`linker`" (https://github.com/rwtourdot/linker) performs multiple tasks related to haplotype inference. It takes long-range sequencing data and a set of heterozygous variants as input and runs on either a whole chromosome or a specified region of interest.

"`linker extract`" generates a hash map from variant genotypes to sequencing reads through iteration over all sequencing reads overlapping each variant site. For example,

```
198801_A_C -> {read_identifier1, read_identifier2, read_identifier3};
```

represents the following

```
198801: variant position;
A: reference base;
C: genotype observed in the supporting reads (read_identifier1, etc.).
```

For linked-reads data, `linker` uses the 16-bp molecular barcodes as read identifiers. For Hi-C data, `linker` uses the read name as read identifiers. For long-read data, the default read identifier is set to be the read name concatenated with the start and end positions of alignment; this definition ensures that only linkage within a single contiguous alignment is preserved, but not between non-contiguous (split) alignments of the same read. `linker` applies the following read filters when extracting linkage from reads (these will be customizable in a future release):

Table 6. Read filters for linkage extraction from different sequencing data

Data type	Mapping quality	Base Quality	Bam Flag
linked-reads	< 20	< 20	duplicate, non-primary
long-read	< 20	< 8[a]	duplicate, non-primary
Hi-C	< 20	< 20	duplicate

[a]This threshold applies to uncorrected PacBio reads but not to corrected PacBio reads.

As `linker` iterates over all variants to generate the variant-to-read map, it simultaneously creates the inverse map from reads to linked variants. (In this process, the program stores all unique read-identifiers in memory.) For example,

```
read_identifier1 -> {198801_A_C, 198322_T_T, 196990_C_G};
```

indicates that the genotypes `C`, `T`, and `G` at positions 198801, 198322, and 196990 are linked by a single read or linked-read molecule with identifier `read_identifier1`. The inverse map contains all linkage information between variant genotypes.

Calculation of haplotype-specific linkage between variants

The linkage signal between two variant genotypes (*e.g.*, `198801_A_C` and `198322_T_G`) is measured by the number of unique molecules linking them. To calculate this, `linker` intersects the list of reads associated with each variant genotype using the variant-to-read map to get the list of unique reads linking the two genotypes. To calculate haplotype linkage between all pairs of variant genotypes (*e.g.*,`198801_A_C`,`198322_T_G`, etc.), it is more efficient to iterate over all read identifiers and add a link between all variant genotypes associated with each read. The overall complexity of this algorithm is $O(RL)$ where R is the total number of unique read identifiers and L is the average number of variants covered by each read. This strategy was used to calculate n_{ij} and M_{ij} in Eqs. (2) and (3).

The linkage signal between one variant genotype (*e.g.*, `198801_A_C`) and two parental haplotypes (H_A and H_B) can be measured in two ways. One can simply add the linkage between this variant genotype (`198801_A_C`) and all different variant genotypes in each haplotype. Because haplotype linkage has a finite range (< 100 kb), the sum of linkage only needs to be carried out among variants within this range. One caveat of this strategy is that it will count a read more than once if it covers more than two variant sites. Alternatively, one can intersect the list of reads associated to the variant genotype of interest (`198801_A_C`) with union of reads associated to all variant genotypes in each haplotype. The haplotype coverage thus calculated reflects the number of unique molecules (unique read identifiers) linking a variant genotype (`198801_A_C`) to a specific haplotype. The latter was used to compute haplotype linkage in "`linker recover`".

Haplotype inference by energy minimization

The minimization approach for haplotype inference can be implemented in different ways. We have implemented a simple strategy that alternately performs spin flipping or block switching on the haplotype configuration to lower the total energy ("`linker solve`"). During each round of minimization, the program first performs spin flips associated with negative energy penalties given by Eq. (9) and then performs block switches with negative energy penalties given by Eq. (10).

In calculating the block switching energy $\Delta E_{k|k+1}$, we haven taken advantage of the following recursive relationship:

$$\begin{aligned}\Delta E_{k|k+1} - \Delta E_{k-1|k} &= \sum_{i\le k}\sum_{j>k} M_{ij}s_i s_j - \sum_{i\le k-1}\sum_{j\ge k} M_{ij}s_i s_j \\ &= s_k\left(\sum_{j>k} M_{kj}s_j - \sum_{i<k} M_{ik}s_i\right) \\ &= s_k\left(\sum_{i>k} M_{ki}s_i - \sum_{i<k} M_{ki}s_i\right). \end{aligned} \tag{13}$$

The last step uses the symmetric property of $M_{ij} = M_{ji}$. Introducing

$$h_i^+ = \sum_{j>i} M_{ij}s_j \quad \text{and} \quad h_i^- = \sum_{j<i} M_{ij}s_j, \tag{14}$$

we can calculate the spin flipping energy ΔE_i and the block switching energy $\Delta E_{k|k+1}$ using

$$\Delta E_i = s_i\left(h_i^+ + h_i^-\right) \quad \text{and} \quad \Delta E_{i|i+1} = \Delta E_{i-1|i} + s_i\left(h_i^+ - h_i^-\right). \tag{15}$$

Therefore, the calculation of both energy penalties over all sites has the same complexity as the calculation of h_i^+ and h_i^-, which is proportional to the total number of non-zero M_{ij}'s.

The energy penalties are calculated iteratively from the first spin to the last using Eqs. (14) and (15) while the spin configuration is continually updated as spin flips or block switches are introduced. The asynchronous iteration scheme achieves faster convergence to an energy minimum than synchronous iteration in which all spins are updated at once. This is illustrated by a toy example shown in Fig. S2.

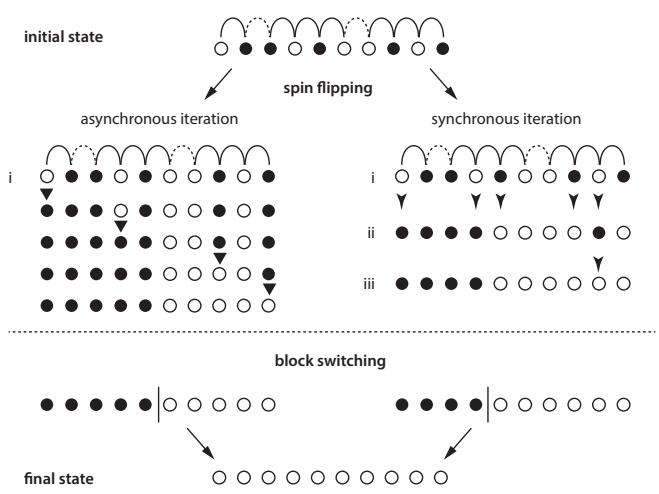

Figure S2. A toy example of haplotype solution by iteration of spin flips and block switches.

In this example, we start from a random initial configuration ($s_i^{(0)} = \pm 1$) with open and filled circles representing genotypes in complementary haplotypes. The interactions between adjacent genotypes are shown as arcs above the haplotype configuration. The interaction is positive (dashed arc) between genotypes in the same haplotype (*cis* linkage) and negative (solid arc) between genotypes from opposite haplotypes (minus *cis* linkage). Spin flips are introduced if the net interaction with adjacent spins is negative.

In asynchronous iteration (left), the energy penalty is calculated as spin flips are introduced: In one round of iteration, the 1st, 4th, 8th, and 10th spins are flipped due to negative interactions with the neighbor spins. In synchronous iteration (right), the 5th and the 9th spins are also flipped due

to negative interactions with their neighbors as given in the initial state (ii); it will take another round of iteration to reverse the 9th spin to the optimal configuration (iii). At the end of spin flipping, there are only long-range switching errors, which are then resolved through block switching moves to get the final state corresponding to a single haplotype.

For both the RPE-1 and the NA12878 linked-reads data, the minimization converges to a stable solution within 10 rounds of iterations for every chromosome (Fig. S3). The fast convergence is due to several factors. First, because the majority of coupling between spins are *cis*-links between genotypes in the two underlying (complementary) haplotypes, local minima consisting of mixed genotypes from both haplotypes are rare (as they require strong *trans*-links). The smooth energy landscape (*i.e.*, without too many local minima) makes the minimization relatively straightforward. Second, the high density of *cis*-links between variant sites up to 100 kb in linked-reads data creates strong local correlation between spins: the strong correlation results in fast convergence to correct local haplotypes by spin flipping (see Appendix A for a more in-depth discussion). Third, although spin flipping cannot efficiently remove long-range switching errors, the generation of correct local haplotype blocks improves the accuracy of block-switching energy calculation and therefore simplifies the correction of block-switching errors. The alternation between spin flipping and block switching operations in the optimization is a distinct feature of our algorithm in comparison to previous methods (***Edge et al., 2017***; ***Marks et al., 2019***). Finally, the dynamic iteration scheme improves the efficiency of both spin-flipping and block-switching moves due to continual improvement of phasing accuracy.

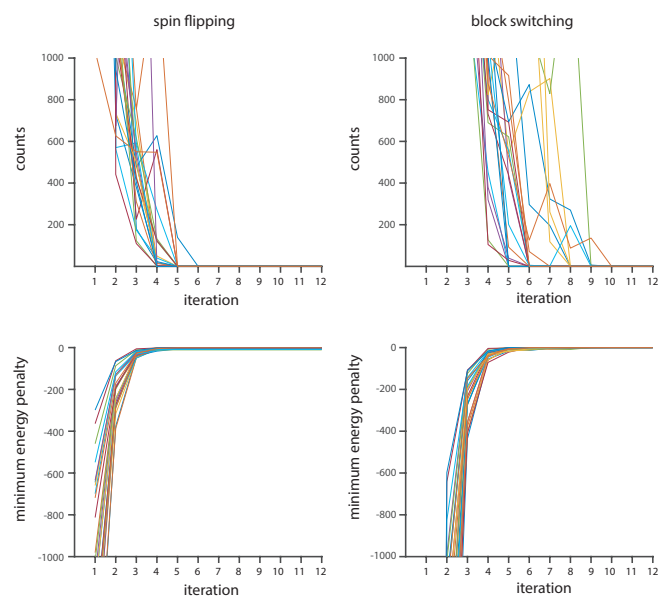

Figure S3. Rapid convergence of the haplotype solution through energy minimization. Top: Total number of spin flips (left) and block switches (right) during each round of iteration through all variant sites on each chromosome (lines of different colors). Bottom: Lowest energy penalty of spin flips (left) and block switches (right) during each round of iteration.

Our minimization strategy can efficiently determine the phase of binary genotypes or haplotype blocks (more to be discussed below) as long as the signal of *cis*-linkage is much stronger than the (random) background noise. If the linkage data are more sparse or contain more errors, the energy landscape may contain more local minima. This problem can be addressed with more sophisticated minimization algorithms. For example, one can use the Metropolis algorithm for spin flipping and block switching that allows more randomness and can sample a larger number of states. It is also straightforward to run multiple simulations starting from different initial states and construct an optimal solution from the ensemble of final states. For example, we can calculate a "thermodynamic" average of the phase between genotypes at sites i and j as

$$\overline{s_i s_j} = \mathrm{sgn}\left(\sum_r s_{i,r} s_{j,r} \cdot \frac{1}{1 + e^{\beta \min(\Delta E_{i,r}, \Delta E_{j,r})}} \right).$$

Here the sum is over the output of $s_{i,r} s_{i+1,r}$ from multiple simulations weighed by the likelihood of phasing errors at either s_i or s_j given by Eq. (11); sgn is the sign function that ensures $s_i s_{i+1} = +1$ or -1.

The haplotype solution can be determined iteratively from $\overline{s_i s_{i+1}}$.

Assessment of phasing accuracy

The heuristic relationship between phasing accuracy and energy penalty given by Eq. (11) provides a convenient strategy to assess phasing accuracy of the final haplotype solution using the energy penalty scores. Low phasing accuracy primarily arises in two scenarios. The first is near regions of low variant density, including those near large gaps in the reference with no variant or linkage information. These regions have weak or no linkage signal. The second scenario is in regions with a high fraction of false variants. These regions can be highly divergent from the reference or contain sequences that are duplicated multiple times in both homologous chromosomes. Alignment inaccuracy is the main cause of false variant detection in these regions and the false variants cannot be completely filtered by haplotype linkage.

As the energy penalty score measures the signal of true *cis* linkage relative to the background, it can reveal regions of low phasing accuracy due to either low variant density or high fraction of false variants. We can therefore use a single cutoff on the energy penalty to ensure phasing accuracy. To generate haplotype blocks with high intra-block phasing accuracy that is critical for deriving accurate inter-block linkage from Hi-C reads, we have chosen a conservative cutoff for block-switching errors determined by the minimum in the distribution of energy penalty scores. The haplotype solution is then disassembled to short haplotype blocks flanked by sites with lower block-switching penalty scores than the cutoff. Importantly, the switching energy penalty measures the probability of long-range switching errors and is insensitive to sporadic false variants or phasing errors at single variant sites.

Concatenation of haplotype blocks using Hi-C links

Local haplotype blocks are concatenated using Hi-C links ("`linker scaffold`"). The program first generates maps between variants and haplotype blocks:

```
198801 -> haplotype_block_1;
198322 -> haplotype_block_1;
haplotype_block_1 -> {198801, 198322, ...};
haplotype_block_2 -> {199005, 200142, ...}.
```

Phased genotypes within each haplotype block are represented as (1 for reference and -1 for alternate):

```
haplotype[haplotype_block_1]=[1, -1, -1, ...].
```

The program then iterates over all Hi-C reads spanning two or more phased variants (using the read-to-variant map of Hi-C data generated by `linker extract`) to calculate phased linkage between haplotype blocks. For example, the number of links supporting *cis*-linkage between two haplotype blocks $\mathbf{B}_s$ and $\mathbf{B}_t$ is given by

$$n(\mathbf{B}_s \leftrightarrow \mathbf{B}_t) = \sum_m \mathbb{1}[C_m(x_m) = \mathbf{B}_s(x_m)] \cdot \mathbb{1}[C_m(y_m) = \mathbf{B}_t(y_m)],$$

where the sum is over all Hi-C links $C_m(x_m, y_m)$ with variant positions x_m and y_m in the two haplotype blocks; $C_m(x_m)$ and $C_m(y_m)$ are the genotypes observed at the variant sites in the Hi-C read; $\mathbf{B}_s(x_m)$ and $\mathbf{B}_t(y_m)$ are the genotypes at the variant sites in the two haplotype blocks $\mathbf{B}_s$ and $\mathbf{B}_t$. Similarly,

$$n(\overline{\mathbf{B}}_s \leftrightarrow \mathbf{B}_t) = \sum_m \mathbb{1}[C_m(x_m) = -\mathbf{B}_s(x_m)] \cdot \mathbb{1}[C_m(y_m) = \mathbf{B}_t(y_m)],$$

gives the signal of *trans*-linkage. The combined linkage signal is calculated using Eq. (5).

As the linkage signal is generated by accumulating phased Hi-C links between local haplotype blocks, the specificity of linkage is very sensitive to long-range switching errors within each block.

Consider two blocks with fractions of haplotype A given by f_1 and f_2, and assume all Hi-C links to be *cis* links, then the observed fraction of *cis*-linkage between the two blocks is approximately

$$\underbrace{f_1 \cdot f_2}_{A \leftrightarrow A} + \underbrace{(1 - f_1) \cdot (1 - f_2)}_{B \leftrightarrow B},$$

and the fraction of *trans*-linkage is

$$f_1 + f_2 - 2f_1 f_2.$$

If there is no switching within either block, $f_{1,2} = 0$ or 1, then either the fraction of *cis*-linkage is equal to 1 ($f_1 = f_2$) or the fraction of *trans*-linkage is equal to 1 ($f_1 = 1, f_2 = 0$ or $f_1 = 0, f_2 = 1$). When $0 < f_1 < 1$ or $0 < f_2 < 1$, both *cis* and *trans* linkage fractions are above zero; when f_1 or $f_2 \approx 0.5$ (which can be caused by a single switching error), there is roughly equal *cis* and *trans* linkage with no net signal. Ensuring the accuracy of local haplotype blocks is therefore critical for determining the phase between these blocks accurately and efficiently.

To concatenate haplotype blocks within each chromosome arm, it is sufficient to use Hi-C links between variants within 10 Mb. The concatenation of p- and q-arm haplotypes uses all phased Hi-C links between the two arms, which always shows a strong bias (>10:1) dominated by intra-chromosomal linkage (Fig. S6). The dominance of *cis*-linkage between p- and q-arm haplotypes can serve as a good indicator of phasing accuracy even when the ground truth is unknown, because any significant long-range switching error in either arm will result in a substantial drop in the ratio between *cis*- and *trans*-linkage as discussed in the previous paragraph.

Appendix A: The minimization approach of haplotype inference

To better understand the connection between haplotype inference and energy minimization, we switch to a different representation of a haplotype solution using the true parental haplotypes $\mathbf{H}_A$ and $\mathbf{H}_B = \overline{\mathbf{H}_A}$. We represent haplotype A as follows:

$$\begin{aligned} h_{A,i} &= 1 \quad \text{if haplotype A has the reference genotype at site } i\,, \\ h_{A,i} &= -1 \quad \text{if haplotype A has the alternate genotype at site } i\,. \end{aligned}$$

A haplotype solution $\mathbf{S}$ can be written as

$$\mathbf{S} = \mathbf{S}' \cdot \mathbf{H}_A, \tag{16}$$

with

$$\begin{aligned} s'_i &= 1 \quad \text{if the genotype at site } i \text{ agrees with haplotype A,} \\ s'_i &= -1 \quad \text{if the genotype at site } i \text{ agrees with haplotype B.} \end{aligned}$$

The coupling term in Eq. (4) can be rewritten as

$$M_{ij} s_i s_j = M_{ij} s'_i h_{A,i} s'_j h_{A,j} = M'_{ij} s'_i s'_j. \tag{17}$$

with

$$M'_{ij} = M_{ij} h_{A,i} h_{A,j}. \tag{18}$$

As $h_{B,i} = -h_{A,i}$, we also have

$$M'_{ij} = M_{ij} h_{B,i} h_{B,j}. \tag{19}$$

It is straightforward to verify that M'_{ij} is proportional to the difference between *cis* (A-A/B-B) linkage μ_{ij} and *trans* (A-B/B-A) linkage δ_{ij}:

$$M'_{ij} = \chi_{ij} \left(\mu_{ij} - \delta_{ij} \right),$$

with

$$\chi_{ij} = |\mu_{ij} - \delta_{ij}|/(\mu_{ij} + \delta_{ij})$$

measures the fraction of dominant linkage. The energy function (Eq. (4)) can be rewritten as

$$E(\mathbf{S}') = -\frac{1}{2} \sum_{i,j} \chi_{ij} \left(\mu_{ij} - \delta_{ij}\right) s'_i s'_j. \tag{20}$$

Finding $\mathbf{S}$ that minimizes $E(\mathbf{S})$ is equivalent to finding $\mathbf{S}'$ that minimizes Eq. (20).

If $\mu_{ij} > \delta_{ij}$ $(\forall i, j)$, Eq. (20) describes a finite-range 1D Ising model that has two global minima $s'_i = 1$ or $s'_i = -1$. Block switching and spin flipping are the two elementary fluctuations with energy changes given by

$$\Delta E_{i|i+1} = \sum_{j \le i} \sum_{k > i} \chi_{jk} \left(\mu_{jk} - \delta_{jk}\right) s'_j s'_k \quad \text{(block switch)} \tag{21}$$

$$\Delta E_i = \sum_j \chi_{jk} \left(\mu_{ij} - \delta_{ij}\right) s'_i s'_j = s'_i h'_i. \quad \text{(spin flip)} \tag{22}$$

Starting from a random configuration:

$$p(s'_{k,0} = 1) = p(s'_{k,0} = -1) = 1/2,$$
$$p(s'_{k,0} \cdot s'_{l,0} = 1) = p(s'_{k,0} \cdot s'_{l,0} = -1) = 1/2,$$

if we introduce spin flips according to Eq. (22):

$$h'_{i,0} > 0 \to s_{i,1} = -1;$$
$$h'_{i,0} < 0 \to s_{i,1} = 1,$$

the probability that two sites are phased correctly relative to each other after one round of spin flipping is given by

$$p(s'_{i,1} \cdot s'_{j,1} = 1) = p(h'_{i,0} \cdot h'_{j,0} > 0).$$

We have

$$\begin{aligned} h'_{i,0} \cdot h'_{j,0} &= \sum_{k,l} M'_{ik} M'_{jl} s'_{k,0} s'_{l,0} \\ &= \sum_k M'_{ik} M'_{jk} + \sum_{k \ne l} M'_{ik} M'_{jl} s'_{k,0} s'_{l,0}. \end{aligned} \tag{23}$$

For a random configuration,

$$p(s'_{k,0} \cdot s'_{l,0} = 1) = p(s'_{k,0} \cdot s'_{l,0} = -1) = 1/2,$$

we have

$$\begin{aligned} E(h'_{i,0} \cdot h'_{j,0}) &= \sum_k M'_{ik} M'_{jk} + \sum_{k \ne l} M'_{ik} M'_{jl} E(s'_{k,0} \cdot s'_{l,0}) \\ &\approx \sum_k M'_{ik} M'_{jk} > 0. \end{aligned} \tag{24}$$

Thus the first term in Eq. (23) is always positive when $M'_{ij} \ge 0$. This implies that

$$p(s'_{i,1} \cdot s'_{j,1} = 1) = p(h'_{i,0} \cdot h'_{j,0} > 0) > p(h'_{i,0} \cdot h'_{j,0} < 0) = p(s'_{i,1} \cdot s'_{j,1} = -1).$$

As more sites are phased correctly, $p(s'_i \cdot s'_j = 1) \uparrow$ and $E(h'_i \cdot h'_j) \uparrow$ according to Eq. (24). This implies that spin flipping will continue to improve phasing accuracy as long as *cis* linkage outweighs *trans* errors ($\mu_{ij} > \delta_{ij}$).

Spin flipping will not be able to correct long-range haplotype switching:

$$s'_k = 1,\ i - m \le k \le i; s'_k = -1,\ i + 1 \le k \le i + n.$$

The block-switching energy penalty near position i

$$\Delta E_{i|i+1} = -\sum_{j\leq i}\sum_{k>i} M_{jk}$$

will be negative if $M_{jk} \geq 0$. Therefore, by iterating spin flipping and block switching, we are guaranteed to get either $s'_i = 1$ or $s'_i = -1$. Exceptions arise when $\Delta E = 0$, which implies the existence of sites j, k with $\mu_{jk} - \delta_{jk} \leq 0$, *i.e.*, more *trans*-linkage than *cis*-linkage.

Appendix B: The log-likelihood function of haplotype inference

The energy function $E(\mathbf{S})$ as defined in Eq. (4) can be related to the log-likelihood of a haplotype solution given observed linkage. To derive this relationship, we first note that $E(\mathbf{S})$ is equivalent to

$$E(\mathbf{S}) \approx \underbrace{(\text{\# of observed links inconsistent with } \mathbf{S})}_{T(\mathbf{S})} - \underbrace{(\text{\# of observed links consistent with } \mathbf{S})}_{C(\mathbf{S})},$$

when conflicting linkage is rare ($\chi_{ij} \approx 1$).

Assume *trans* errors (linkage between paternal and maternal genotypes) due to sequencing or other technical artifacts to be random and have a combined error frequency δ, the probability of seeing $T(\mathbf{S})$ inconsistent links and $C(\mathbf{S})$ consistent links is given by

$$p\left(M_{ij}|\mathbf{S}\right) = \binom{T+C}{T}\delta^T(1-\delta)^C.$$

The likelihood $p\left(\mathbf{S}|M_{ij}\right)$ of a haplotype solution $\mathbf{S}$ given M_{ij} is given by

$$p\left(\mathbf{S}|M_{ij}\right) = \frac{p\left(M_{ij}|\mathbf{S}\right)p(\mathbf{S})}{\sum_{\hat{\mathbf{S}}} p\left(M_{ij}|\hat{\mathbf{S}}\right)p(\hat{\mathbf{S}})}.$$

For two haplotype solutions $\mathbf{S}$ and $\mathbf{S}'$, we have

$$\begin{aligned}\ln p\left(\mathbf{S}|M_{ij}\right) - \ln p\left(\mathbf{S}'|M_{ij}\right) &= \ln p\left(M_{ij}|\mathbf{S}\right) - \ln\left(M_{ij}|\mathbf{S}'\right) + \ln p(\mathbf{S}) - \ln p\left(\mathbf{S}'\right)\\ &\approx \ln p\left(M_{ij}|\mathbf{S}\right) - \ln\left(M_{ij}|\mathbf{S}'\right)\end{aligned} \tag{25}$$

assuming a uniform prior of $p(\hat{\mathbf{S}})$.

As the total number of links $M = T(\mathbf{S}) + C(\mathbf{S})$ is constant, we have

$$\Delta E = E(\mathbf{S}') - E(\mathbf{S}) = T(\mathbf{S}') - T(\mathbf{S}) + C(\mathbf{S}) - C(\mathbf{S}') = 2(T(\mathbf{S}') - T(\mathbf{S})).$$

Equation (25) becomes

$$\begin{aligned}\ln p\left(\mathbf{S}|M_{ij}\right) - \ln p\left(\mathbf{S}'|M_{ij}\right) &= \ln p\left(M_{ij}|\mathbf{S}\right) - \ln\left(M_{ij}|\mathbf{S}'\right)\\ &= \ln\frac{T(\mathbf{S}')!C(\mathbf{S}')!}{T(\mathbf{S})!C(\mathbf{S})!} + \underbrace{\left(T(\mathbf{S}) - T(\mathbf{S}')\right)}_{-\Delta E/2}\ln\delta + \underbrace{\left(C(\mathbf{S}) - C(\mathbf{S}')\right)}_{\Delta E/2}\ln(1-\delta)\end{aligned}$$

For the first term in the second row (entropic term), we have

$$\begin{aligned}&\ln T(\mathbf{S}')! + \ln C(\mathbf{S}')! - \ln T(\mathbf{S})! - \ln C(\mathbf{S})!\\ &\approx T(\mathbf{S}')\ln T(\mathbf{S}') + (M - T(\mathbf{S}'))\ln(M - T(\mathbf{S}')) - T(\mathbf{S})\ln T(\mathbf{S}) - (M - T(\mathbf{S}))\ln(M - T(\mathbf{S}))\\ &= (T(\mathbf{S}') - T(\mathbf{S}))\cdot\ln\frac{T(\mathbf{S})}{M - T(\mathbf{S})} + \frac{(T(\mathbf{S}') - T(\mathbf{S}))^2}{2T(\mathbf{S})(1 - T(\mathbf{S})/M)} + O(T(\mathbf{S}') - T(\mathbf{S})/M)^2\\ &= \frac{\Delta E}{2}\left[\ln\frac{T(\mathbf{S})/M}{1 - T(\mathbf{S})/M} + \frac{\Delta E}{M}\cdot\frac{1}{2T(\mathbf{S})/M\cdot(1 - T(\mathbf{S})/M)} + O\left(\frac{\Delta E}{M}\right)^2\right].\end{aligned}$$

Putting these together and introducing

$$\tau = T(\mathbf{S})/M,$$

which measures the fraction of observed links that are inconsistent with $\mathbf{S}$, we have

$$\ln p\left(\mathbf{S}|M_{ij}\right) - \ln p\left(\mathbf{S}'|M_{ij}\right) = \frac{\Delta E}{2}\left[-\ln\frac{\delta}{1-\delta} + \ln\frac{\tau}{1-\tau} + \frac{\Delta E}{M}\cdot\frac{1}{2\tau(1-\tau)} + o\left(\frac{\Delta E}{M}\right)^2\right]. \tag{26}$$

As $\Delta E \ll M$, this simplifies to

$$\ln p\left(\mathbf{S}|M_{ij}\right) - \ln p\left(\mathbf{S}'|M_{ij}\right) \approx \frac{\Delta E}{2}\left[-\ln\frac{\delta}{1-\delta} + \ln\frac{\tau}{1-\tau}\right]. \tag{27}$$

In general, we expect the fraction of inconsistent links $\tau \gtrsim \delta$, the frequency of *trans* errors. Therefore,

$$\ln p\left(\mathbf{S}|M_{ij}\right) - \ln p\left(\mathbf{S}'|M_{ij}\right) = \beta\left[E(\mathbf{S}') - E(\mathbf{S})\right],$$

and minimizing $E(\mathbf{S})$ is equivalent to maximizing the likelihood $p\left(\mathbf{S}|M_{ij}\right)$.

$$\beta = \frac{1}{2}\left[-\ln\frac{\delta}{1-\delta} + \ln\frac{\tau}{1-\tau}\right]$$

is a heuristic parameter that is related to linkage accuracy.

Moreover, the energy changes associated with spin flipping or block switching as given by Eq. (9) and (10) directly reflect local phasing accuracy. As the haplotype at any site i is either $s_i = +1$ or $s_i = -1$, we have

$$p(s_i = 1) + p(s_i = -1) = 1.$$

Together with

$$\ln p(s_i|M) - \ln p(s_i \to -s_i|M) \approx \beta\Delta E_i,$$

the probability that s_i is phased correctly is given by

$$p(s_i|M) \approx \frac{e^{\beta\Delta E_i}}{1+e^{\beta\Delta E_i}}. \tag{28}$$

Similarly, the probability that there is haplotype switching between site i and $i+1$ is given by

$$p(s_i s_{i+1}|M) \approx \frac{e^{\beta\Delta E_{i|i+1}}}{1+e^{\beta\Delta E_{i|i+1}}}. \tag{29}$$

References

Adey, A., Burton, J. N., Kitzman, J. O., Hiatt, J. B., Lewis, A. P., Martin, B. K., Qiu, R., Lee, C., and Shendure, J. (2013a). The haplotype-resolved genome and epigenome of the aneuploid hela cancer cell line. *Nature*, 500(7461):207–211.

Adey, A., Kitzman, J. O., Burton, J. N., Daza, R., Kumar, A., Christiansen, L., Ronaghi, M., Amini, S., L Gunderson, K., Steemers, F. J., and Shendure, J. (2014). In vitro, long-range sequence information for de novo genome assembly via transposase contiguity. *Genome Research*, 24(12):2041–2049.

Adey, A., Patwardhan, R. P., Qiu, R., Kitzman, J. O., Burton, J. N., and Shendure, J. (2013b). Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nature Biotechnology*, 31(12):1119–1125.

Browning, S. R. and Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics*, 81(5):1084–1097.

Browning, S. R. and Browning, B. L. (2011). Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics*, 12(10):703–714.

Chu, W. K., Edge, P., Lee, H. S., Bansal, V., Bafna, V., Huang, X., and Zhang, K. (2017). Ultraaccurate genome sequencing and haplotyping of single human cells. *Proceedings of the National Academy of Sciences*, 114(47):12512–12517.

Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J., and Lander, E. S. (2001). High-resolution haplotype structure in the human genome. *Nature Genetics*, 29(2):229–232.

Darrow, E. M., Huntley, M. H., Dudchenko, O., Stamenova, E. K., Durand, N. C., Sun, Z., Huang, S.-C., Sanborn, A. L., Machol, I., Shamim, M., Seberg, A. P., Lander, E. S., Chadwick, B. P., and Aiden, E. L. (2016). Deletion of DXZ4on the human inactive X chromosome alters higher-order genome architecture. *Proceedings of the National Academy of Sciences*, 113(31):E4504–E4512.

Deng, Q., Ramskold, D., Reinius, B., and Sandberg, R. (2014). Single-cell rna-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science*, 343(6167):193–196.

Dixon, J. R., Xu, J., Dileep, V., Zhan, Y., Song, F., Le, V. T., Yardımcı, G. G., Chakraborty, A., Bann, D. V., Wang, Y., Clark, R., Zhang, L., Yang, H., Liu, T., Iyyanki, S., An, L., Pool, C., Sasaki, T., Rivera-Mulia, J. C., Ozadam, H., Lajoie, B. R., Kaul, R., Buckley, M., Lee, K., Diegel, M., Pezic, D., Ernst, C., Hadjur, S., Odom, D. T., Stamatoyannopoulos, J. A., Broach, J. R., Hardison, R. C., Ay, F., Noble, W. S., Dekker, J., Gilbert, D. M., and Yue, F. (2018). Integrative detection and analysis of structural variation in cancer genomes. *Nature Genetics*, 50(10):1388–1398.

Edge, P., Bafna, V., and Bansal, V. (2017). HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Research*, 27(5):801–812.

Falconer, E., Hills, M., Naumann, U., Poon, S. S. S., Chavez, E. A., Sanders, A. D., Zhao, Y., Hirst, M., and Lansdorp, P. M. (2012). DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. *Nature Methods*, 9(11):1107–1112.

Fan, H. C., Wang, J., Potanina, A., and Quake, S. R. (2010). Whole-genome molecular haplotyping of single cells. *Nature Biotechnology*, 29(1):51–57.

Gribble, S. M., Roberts, I., Grace, C., Andrews, K. M., Green, A. R., and Nacheva, E. P. (2000). Cytogenetics of the chronic myeloid leukemia-derived cell line k562: karyotype clarification by multicolor fluorescence in situ hybridization, comparative genomic hybridization, and locus-specific fluorescence in situ hybridization. *Cancer Genet Cytogenet*, 118(1):1–8.

Hansen, C. H. and van Oudenaarden, A. (2013). Allele-specific detection of single mrna molecules in situ. *Nature Methods*, 10(9):869–871.

Kaper, F., Swamy, S., Klotzle, B., Munchel, S., Cottrell, J., Bibikova, M., Chuang, H.-Y., Kruglyak, S., Ronaghi, M., Eberle, M. A., and Fan, J.-B. (2013). Whole-genome haplotyping by dilution, amplification, and sequencing. *Proc. Natl. Acad. Sci. U.S.A.*, 110(14):5552–5557.

Kitzman, J. O., MacKenzie, A. P., Adey, A., Hiatt, J. B., Patwardhan, R. P., Sudmant, P. H., Ng, S. B., Alkan, C., Qiu, R., Eichler, E. E., and Shendure, J. (2010). Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nature Biotechnology*, 29(1):59–63.

Kong, A., Masson, G., Frigge, M. L., Gylfason, A., Zusmanovich, P., Thorleifsson, G., Olason, P. I., Ingason, A., Steinberg, S., Rafnar, T., Sulem, P., Mouy, M., Jonsson, F., Thorsteinsdottir, U., Gudbjartsson, D. F., Stefansson, H., and Stefansson, K. (2008). Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature Genetics*, 40(9):1068–1075.

Levesque, M. J. and Raj, A. (2013). Single-chromosome transcriptional profiling reveals chromosomal gene expression regulation. *Nature Methods*, 10(3):246–248.

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., and et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293.

Loh, P.-R., Danecek, P., Palamara, P. F., Fuchsberger, C., A Reshef, Y., K Finucane, H., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G. R., Durbin, R., and L Price, A. (2016a). Reference-based phasing using the Haplotype Reference Consortium panel. *Nature Genetics*, 48(11):1443–1448.

Loh, P.-R., Genovese, G., Handsaker, R. E., Finucane, H. K., Reshef, Y. A., Palamara, P. F., Birmann, B. M., Talkowski, M. E., Bakhoum, S. F., McCarroll, S. A., and Price, A. L. (2018). Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. *Nature Publishing Group*, pages 1–22.

Loh, P.-R., Palamara, P. F., and Price, A. L. (2016b). Fast and accurate long-range phasing in a UK Biobank cohort. *Nature Genetics*, 48(7):811–816.

Ma, L., Xiao, Y., Huang, H., Wang, Q., Rao, W., Feng, Y., Zhang, K., and Song, Q. (2010). Direct determination of molecular haplotypes by chromosome microdissection. *Nature Publishing Group*, 7(4):299–301.

Maciejowski, J., Li, Y., Bosco, N., Campbell, P. J., and de Lange, T. (2015). Chromothripsis and kataegis induced by telomere crisis. *Cell*, 163(7):1641 – 1654.

Marks, P., Garcia, S., Barrio, A. M., Belhocine, K., Bernate, J., Bharadwaj, R., Bjornson, K., Catalanotti, C., Delaney, J., Fehr, A., Fiddes, I. T., Galvin, B., Heaton, H., Herschleb, J., Hindson, C., Holt, E., Jabara, C. B., Jett, S., Keivanfar, N., Kyriazopoulou-Panagiotopoulou, S., Lek, M., Lin, B., Lowe, A., Mahamdallie, S., Maheshwari, S., Makarewicz, T., Marshall, J., Meschi, F., O'Keefe, C. J., Ordonez, H., Patel, P., Price, A., Royall, A., Ruark, E., Seal, S., Schnall-Levin, M., Shah, P., Stafford, D., Williams, S., Wu, I., Xu, A. W., Rahman, N., MacArthur, D., and Church, D. M. (2019). Resolving the full spectrum of human genome variation using Linked-Reads. *Genome Research*, 29(4):635–645.

Naumann, S., Reutzel, D., Speicher, M., and Decker, H. J. (2001). Complete karyotype characterization of the k562 cell line by combined application of g-banding, multiplex-fluorescence in situ hybridization, fluorescence in situ hybridization, and comparative genomic hybridization. *Leuk Res*, 25(4):313–22.

Nik-Zainal, S., Van Loo, P., Wedge, D. C., Alexandrov, L. B., Greenman, C. D., Lau, K. W., Raine, K., Jones, D., Marshall, J., Ramakrishna, M., and et al. (2012). The life history of 21 breast cancers. *Cell*, 149(5):994–1007.

Pendleton, M., Sebra, R., Pang, A. W. C., Ummat, A., Franzen, O., Rausch, T., Stütz, A. M., Stedman, W., Ananthara-man, T., Hastie, A., Dai, H., Fritz, M. H.-Y., Cao, H., Cohain, A., Deikus, G., Durrett, R. E., Blanchard, S. C., Altman, R., Chin, C.-S., Guo, Y., Paxinos, E. E., Korbel, J. O., Darnell, R. B., McCombie, W. R., Kwok, P.-Y., Mason, C. E., Schadt, E. E., and Bashir, A. (2015). Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nature Methods*, 12(8):780–786.

Peters, B. A., Kermani, B. G., Sparks, A. B., Alferov, O., Hong, P., Alexeev, A., Jiang, Y., Dahl, F., Tang, Y. T., Haas, J., Robasky, K., Zaranek, A. W., Lee, J.-H., Ball, M. P., Peterson, J. E., Perazich, H., Yeung, G., Liu, J., Chen, L., Kennemer, M. I., Pothuraju, K., Konvicka, K., Tsoupko-Sitnikov, M., Pant, K. P., Ebert, J. C., Nilsen, G. B., Baccash, J., Halpern, A. L., Church, G. M., and Drmanac, R. (2014). Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature*, 487(7406):190–195.

Porubsky, D., Garg, S., Sanders, A. D., Korbel, J. O., Guryev, V., Lansdorp, P. M., and Marschall, T. (2017). Dense and accurate whole-chromosome haplotyping of individual genomes. *Nature Communications*, 8(1).

Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S., and Aiden, E. L. (2014). A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–80.

Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., Lavery, T., Kouyoumjian, R., Farhadian, S. F., Ward, R., and Lander, E. S. (2001). Linkage disequilibrium in the human genome. *Nature*, 411(6834):199–204.

Sanborn, A. L., Rao, S. S. P., Huang, S.-C., Durand, N. C., Huntley, M. H., Jewett, A. I., Bochkov, I. D., Chinnappan, D., Cutkosky, A., Li, J., and et al. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proceedings of the National Academy of Sciences*, 112(47):E6456–E6465.

Scheet, P. and Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *The American Journal of Human Genetics*, 78(4):629–644.

Selvaraj, S., Dixon, J. R., Bansal, V., and Ren, B. (2013). Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nature Biotechnology*, 31(12):1111–1118.

Seo, J.-S., Rhie, A., Kim, J., Lee, S., Sohn, M.-H., Kim, C.-U., Hastie, A., Cao, H., Yun, J.-Y., Kim, J., Kuk, J., Park, G. H., Kim, J., Ryu, H., Kim, J., Roh, M., Baek, J., Hunkapiller, M. W., Korlach, J., Shin, J.-Y., and Kim, C. (2016). De novo assembly and phasing of a Korean human genome. *Nature Publishing Group*, pages 1–18.

Shendure, J. and Aiden, E. L. (2012). The expanding scope of DNA sequencing. *Nature Publishing Group*, 30(11):1084–1094.

Snyder, M. W., Adey, A., Kitzman, J. O., and Shendure, J. (2015). Haplotype-resolved genome sequencing: experimental methods and applications. *Nature Reviews Genetics*, 16(6):344–358.

Spielmann, M., ez, D. x. o. G. L. x. a. x., and Mundlos, S. (2018). Structural variation in the 3D genome. *Nature Reviews Genetics*, pages 1–15.

Suk, E. K., McEwen, G. K., Duitama, J., Nowick, K., Schulz, S., Palczewski, S., Schreiber, S., Holloway, D. T., McLaughlin, S., Peckham, H., Lee, C., Huebsch, T., and Hoehe, M. R. (2011). A comprehensively molecular haplotype-resolved genome of a European individual. *Genome Research*, 21(10):1672–1685.

The International HapMap Consortium (2005). A haplotype map of the human genome. *Nature*, 437(7063):1299–1320.

Yang, H., Chen, X., and Wong, W. H. (2010). Completely phased genome sequencing through chromosome sorting. *Proceedings of the National Academy of Sciences*, 108(1):12–17.

Zhang, C.-Z. and Pellman, D. (2015). From mutational mechanisms in single cells to mutational patterns in cancer genomes. *Cold Spring Harbor Symposia on Quantitative Biology*, 80:117–137.

Zhang, C.-Z., Spektor, A., Cornils, H., Francis, J. M., Jackson, E. K., Liu, S., Meyerson, M., and Pellman, D. (2015). Chromothripsis from dna damage in micronuclei. *Nature*, 522(7555):179–184.

Zheng, G. X. Y., Lau, B. T., Schnall-Levin, M., Jarosz, M., Bell, J. M., Hindson, C. M., Kyriazopoulou-Panagiotopoulou, S., Masquelier, D. A., Merrill, L., Terry, J. M., Mudivarti, P. A., Wyatt, P. W., Bharadwaj, R., Makarewicz, A. J., Li, Y., Belgrader, P., Price, A. D., Lowe, A. J., Marks, P., Vurens, G. M., Hardenbol, P., Montesclaros, L., Luo, M., Greenfield, L., Wong, A., Birch, D. E., Short, S. W., Bjornson, K. P., Patel, P., Hopmans, E. S., Wood, C., Kaur, S., Lockwood, G. K., Stafford, D., Delaney, J. P., Wu, I., Ordonez, H. S., Grimes, S. M., Greer, S., Lee, J. Y., Belhocine, K., Giorda, K. M., Heaton, W. H., McDermott, G. P., Bent, Z. W., Meschi, F., Kondov, N. O., Wilson, R., Bernate, J. A., Gauby, S., Kindwall, A., Bermejo, C., Fehr, A. N., Chan, A., Saxonov, S., Ness, K. D., Hindson, B. J., and Ji, H. P. (2016). Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nature Biotechnology*, 34(3):303–311.

Zhou, B., Ho, S. S., Greer, S. U., Zhu, X., Bell, J. M., Arthur, J. G., Spies, N., Zhang, X., Byeon, S., Pattni, R., Ben-Efraim, N., Haney, M. S., Haraksingh, R. R., Song, G., Ji, H. P., Perrin, D., Wong, W. H., Abyzov, A., and Urban, A. E. (2019). Comprehensive, integrated, and phased whole-genome analysis of the primary ENCODE cell line K562. *Genome Research*, 29(3):472–484.