Whole Chromosome Haplotype Phasing from Long-Range Sequencing

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, 44 Binney Street, Boston MA 02215; [2]Department of Biomedical Informatics, Harvard Medical School; [3]Cancer Program, Broad Institute of MIT and Harvard

Abstract Haplotype phase represents the collective genetic variation between homologous chromosomes and is an essential feature of polyploid genomes. Determining the haplotype phase requires knowledge of both the genotypes at variant sites and their linkage across each homologous chromosome. Although short-read sequencing can produce accurate genotype information, it cannot resolve linkage between genotypes due to the short size ($\lesssim$1kb) of sequencing fragments. Long-read and long-range sequencing technologies can reveal linkage information across a wide range of genomic lengths (10kb-100 Mb), but such information is often sparse and contaminated with different sources of errors. To what extent can long-range sequencing produce accurate long-range haplotype information remains unknown. Here we describe a general computational framework for inferring haplotype phase and assessing phasing accuracy from long-range sequencing data using a one-dimensional spin model. Building on this model, we demonstrate a two-tier phasing strategy that enables complete whole-chromosome phasing of diploid genomes combining 60× linked-reads sequencing and 60× Hi-C sequencing. The computationally inferred haplotypes from long-range sequencing show high completeness (>95%) and accuracy (∼99%) when compared to haplotypes directly determined from sequencing of single chromosomes. Our results provide a scalable solution to generating completely phased genomes from bulk sequencing and enable haplotype-resolved genome analysis at large.

Introduction

Haplotype phase ("haploid genotype") refers to a combination of genotypes that are linked along a chromosome (***The International HapMap Consortium, 2005***). In population genetics, haplotypes are used to represent strongly correlated genotypes at adjacent sites in linkage disequilibrium (***Reich et al., 2001***; ***Daly et al., 2001***). Each haplotype corresponds to a unique segment of an ancestral chromosome and serves as a specific marker for ancestral and demographic inference. In a single polyploid genome, haplotypes denote the combinations of genetic variants along each homologous chromosome. Chromosomal haplotypes are required for resolving molecular differences between homologous chromosomes (in both DNA sequence and chromatin organization) and relating allele-specific genetic and epigenetic variations to changes in gene expression (***Shendure and Aiden, 2012***; ***Adey et al., 2013a***; ***Snyder et al., 2015***). Knowledge of chromosomal haplotype can also significantly improve the accuracy of genomic analysis, including diploid genome assembly (***Pendleton et al., 2015***; ***Seo et al., 2016***) and somatic mutation detection in both heterogenous cell populations (***Nik-Zainal et al., 2012***; ***Loh et al., 2018***) and single-cell genomes (***Zhang and Pellman, 2015***).

Haplotype inference requires knowledge of DNA sequence on each homologous chromosome. The most common strategy to get haplotype information is by analyzing the genotypes of related individuals (***Kong et al., 2008***; ***Browning and Browning, 2011***; ***Loh et al., 2016b***) The haplotype information of an individual can be straightforwardly determined from the genotypes of the parents, or inferred from the genotypes of a number of related individuals who provide "surrogate parent genotypes." Another strategy to infer haplotypes is by statistical modeling of recombination based on genotype frequencies in a population (***Loh et al., 2016a***). This strategy aims to estimate the probability of recombination between adjacent polymorphic sites from a panel of reference haplotypes and produces the haplotype of an individual as the most likely configuration based on the estimated recombination probabilities. Although statistical phasing can produce highly accurate local haplotypes, generating long-range haplotype is only possible for very closely related individuals or with very large sample cohorts. It has limited power in identifying low-frequency haplotypes and cannot phase *de novo* mutations.

Haplotype information can also be directly obtained from sequencing or genotyping of single chromosomes or chromosomal fragments (***Browning and Browning, 2011***; ***Snyder et al., 2015***). This strategy requires experimental separation or tagging of DNA sequences from homologous chromosomes prior to genotyping or sequencing. Experimental phasing can be done at the single chromosome level (***Ma et al., 2010***; ***Fan et al., 2010***; ***Yang et al., 2010***; ***Zhang et al., 2015***; ***Porubsky et al., 2017***), which can produce whole-chromosome haplotype phase with no switching errors. But the haplotype phase is often incomplete due to uneven sequence coverage caused by amplification bias in the generation of sequencing libraries from a single chromosome. Moreover, experimental isolation of single chromosomes is extremely laborious and not scalable. Segmental haplotype information is available from long-read sequencing or long-range sequencing (***Tlemann-Boege and Betancourt, 2017***), which preserve linkage between genetic variants up to the size of a sequencing fragment or within a long DNA molecule that is tagged with a unique molecular identifier (***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 sequencing reads or fragments is around 10-100 kb, which is sufficient for building long haplotype blocks in regions with normal variant density (1 per 1.5 kb). But even 100 kb-range of variant linkage is insufficient to build haplotypes across regions of low variant density (< 1 per 10 kb) or gaps in the reference assembly, which limit the average size of phase blocks determined by long-range sequencing to $\sim$ 1Mb.

Besides single-chromosome sequencing, the only other technology that can generate extremely long-range linkage information is Hi-C sequencing based on proximity-based chromatin ligation (***Selvaraj et al., 2013***; ***Rao et al., 2014***). Spatial isolation of chromosomes in separate chromosomal territories ensures that Hi-C fragments are predominantly formed between loci on a single chromosome; Hi-C reads can in principle reveal linkage information across the whole chromosome. However, because the contact frequency between two genomic loci decays rapidly with their distance, the majority of Hi-C fragments are formed between loci in close proximity (< 1 kb), and very few span distal loci. Among the very few long-range Hi-C contacts, an even smaller fraction span polymorphic sites at both ends. Due to such sparsity, Hi-C data alone are insufficient to generate long contiguous haplotypes (***Edge et al., 2017***) except for special genomes with an extremely high density of variants ($\sim$ 1 per 150 bp) (***Selvaraj et al., 2013***).

Although sparse Hi-C data alone cannot generate long contiguous haplotypes, they provide a scaffold that can be used to connect short phase blocks into long haplotypes (***Edge et al., 2017***). This idea has led us to pursue a hybrid approach of haplotype inference combining long-range sequencing and Hi-C sequencing, using long-range sequencing to generate megabase-scale phase blocks and then connecting these blocks by sparse yet specific long-range Hi-C reads (Figure 1). Here we implement this idea using a simple spin-glass model formalism. In this formalism, both local phasing and scaffolding of local phase blocks are solved as a minimization problem. This formalism further enables quantitative assessment of phasing accuracy at every polymorphic site, which is critical for ensuring accuracy of local phase blocks and specificity of long-range phase

scaffolds. We demonstrate that the hybrid approach can generate complete whole-chromosome haplotypes of diploid human genomes using only bulk linked-reads and Hi-C sequencing of human retinal pigmental epithelium (RPE-1) cells. Comparison of the computationally inferred haplotype with that determined from single-chromosome sequencing confirms the high accuracy of the inferred haplotype, with no global switching errors and very few phasing errors (<2%). We further discuss two applications of haplotype linkage in variant recalibration and digital karyotyping. The concepts and computational algorithms described here lay the foundation for haplotype-resolved genome analysis that offers superior accuracy than standard whole-genome sequencing analysis and significantly expands the scope of genome sequencing.

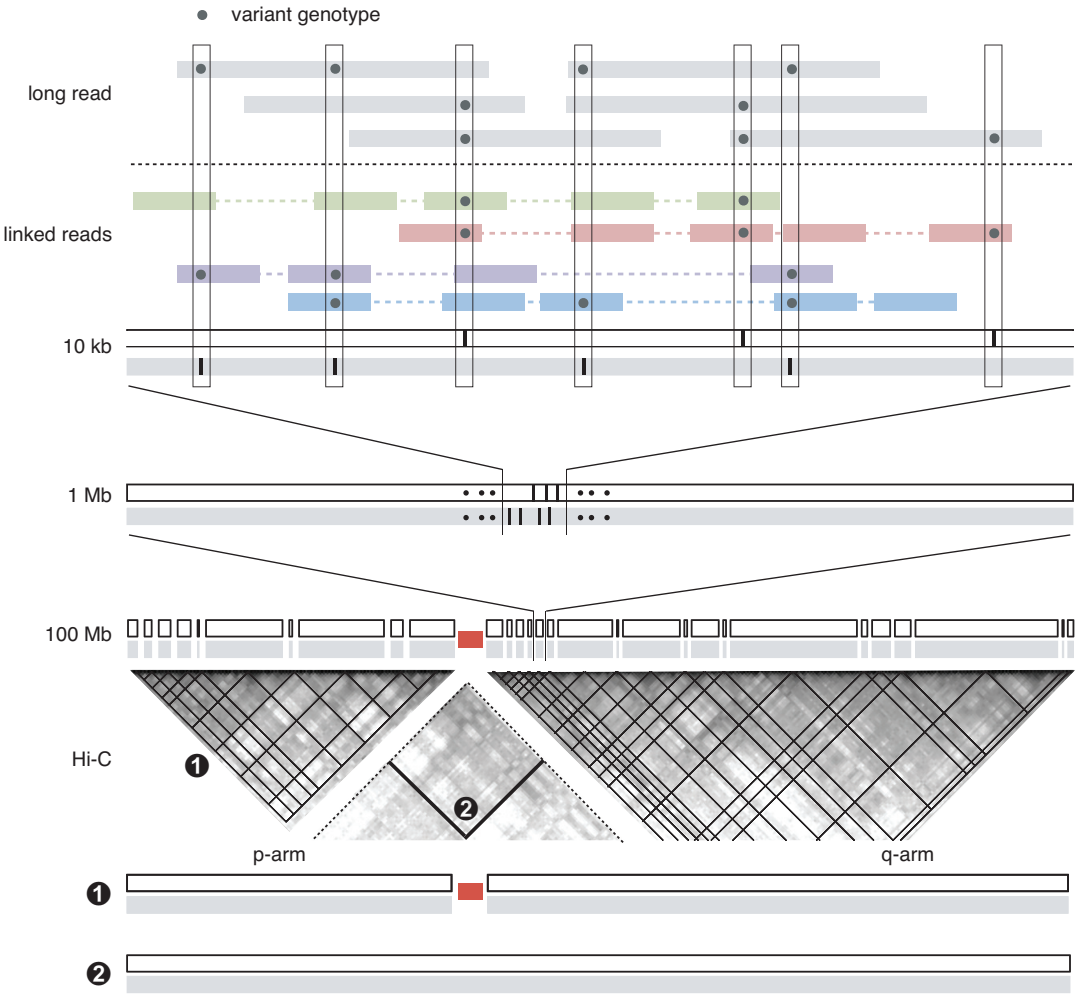

Figure 1. Whole-chromosome haplotype phasing by long-range and Hi-C sequencing. We first generate megabase-scale haplotype blocks using linkage information from long-read or linked-reads sequencing. We then connect these blocks within each chromosome arm combining all phased Hi-C links between blocks. Finally, we bridge p- and q-arm haplotypes using phased Hi-C links between the arms. To ensure accuracy of linkage between blocks, it is critical that each block is phased accurately with no switching errors.

Results

Generation of long-range, Hi-C, and single-chromosome sequencing data

All the analyses in this paper were done on retinal pigmental epithelium cells (RPE-1), an epithelial cell line immortalized with hTERT. RPE-1 cells have a mostly diploid genome that is also very stable. They are frequently used in cell biology and molecular genetic studies. Deep bulk Hi-C data of RPE-1

cells were available from a previously published study (***Darrow et al., 2016***). We used sequencing data of one bulk Hi-C library of wild-type RPE-1 cells (SRS1045726) for this analysis. A single linked-reads sequencing library of RPE-1 cells was generated using the droplet-based technology from 10X Genomics and sequenced to $60\times$ on the Illumina NovaSeq platform, yielding a total of 941,518,426 read pairs (150×2, PCR duplication rate 25%).

We further generated whole-chromosome haplotypes by single-cell sequencing using a previously described experimental strategy (***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. Single-cell DNA libraries were generated from cells after the nocodazole treatment. Monosomic cells were identified based on the arm-level DNA copy number estimated from low-pass whole-genome sequencing ($\sim 0.1\times$) and sent for deeper whole-genome sequencing at 5-10×, which on average covers 50% of each allele. Monosomy was further validated by loss-of-heterozygosity determined from genotype coverage in the deep sequencing data as described before (***Zhang et al., 2015***). We generated a total of 98 monosomic chromosomes from 40 single cells, with at least one for each chromosome, and aggregated all the data to generate a single haplotype of each chromosome. More details are described in Methods.

Haplotype inference as a minimization problem

We formulate the problem of haplotype inference as an energy minimization problem of a one-dimensional Spin Glass (SG) model. This formalism is based on the binary nature of genotypes at each heterozygous site, Reference (R) or Alternate (A). The genotype at site i is represented as $s_i = +1$ (reference) or $s_i = -1$ (alternate). The haplotype phase of a chromosome at sites $i = 1, 2, \cdots N$ is represented as a vector

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

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

Current phasing methods, whether designed for statistical phasing from a population or for haplotype inference from long-range sequencing, focus on determining the linkage between adjacent sites from one of the following scenarios:

$$\begin{aligned}
&\text{reference-reference linkage:} && s_i = 1, s_{i+1} = 1\\
&\text{alternate-alternate linkage:} && s_i = -1, s_{i+1} = -1\\
&\text{reference-alternate linkage:} && s_i = 1, s_{i+1} = -1\\
&\text{alternate-reference linkage:} && s_i = -1, s_{i+1} = 1.
\end{aligned}$$

These can be simplified to two cases:

$$\begin{aligned}
&\text{reference-reference/alternate-alternate linkage:} && s_i \cdot s_{i+1} = 1\\
&\text{reference-alternate/alternate-reference linkage:} && s_i \cdot s_{i+1} = -1,
\end{aligned}$$

which can be further generalized to describe the linkage between any two sites i and j:

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

In short-read sequencing, linkage between variants is only available when they are covered by a single sequencing fragment. As the maximum fragment length ($\lesssim$ 1 kb) is significantly shorter than the average spacing between variant sites in a human genome ($3 \times 10^9/2 \times 10^6 \approx$ 1.5 kb), linkage information is mostly limited to between adjacent sites. In population-based phasing, local haplotype is inferred from linkage disequilibrium between variants that also has the strongest signal between adjacent sites. By contrast, long-range (including long-read) and Hi-C sequencing can reveal linkage between variants separated by 100 kb or across whole chromosomes. For long-range linkage data, direct inference of the global haplotype is more efficient than sequential inference of linkage between adjacent variants.

To have a simple formalism for global haplotype inference, we introduce a matrix that accumulates all linkage information between variant sites:

$$M_{ij} = \begin{cases} \chi_{ij}\left(n_{ij}^{\mathsf{RR}} + n_{ij}^{\mathsf{AA}} - n_{ij}^{\mathsf{AR}} - n_{ij}^{\mathsf{RA}}\right) & i \neq j \\ 0 & i = j \end{cases} \tag{1}$$

where n_{ij}'s represent read counts supporting each type of linkage; χ_{ij} is defined as

$$\chi_{ij} = \frac{\left|n_{ij}^{\mathsf{RR}} + n_{ij}^{\mathsf{AA}} - n_{ij}^{\mathsf{AR}} - n_{ij}^{\mathsf{RA}}\right|}{n_{ij}^{\mathsf{RR}} + n_{ij}^{\mathsf{AA}} + n_{ij}^{\mathsf{AR}} + n_{ij}^{\mathsf{RA}}} \tag{2}$$

and measures the percentage of discrepancy between support for either type of linkage. For perfect data, we expect to see only one type of linkage (either R-R/A-A or R-A/A-R), and $\chi_{ij} = 1$. χ_{ij} serves to penalize the read count signal when there is substantial evidence supporting both types of linkage: such ambiguity arises mostly from sequence duplications or other forms of structural variation that affect either or both loci (i and j). With the introduction of M_{ij}, it is straightforward to see that the correct haplotype $\mathbf{S} = \left(s_1, s_2, \cdots s_N\right)$ will minimize the quadratic "energy" function defined as

$$E\left(\mathbf{S}\right) = -\frac{1}{2}\sum_{i,j} M_{ij} s_i s_j = -\frac{1}{2}\mathbf{S}\mathbf{M}\mathbf{S}^T. \tag{3}$$

Before presenting a computational method to find $\mathbf{S}$ in the next subsection, we want to highlight several features of this formalism. First, the symmetry between $\mathbf{S}$ and $-\mathbf{S}$ is obvious in the definition of $E(\mathbf{S})$ and M_{ij}. In general, M_{ij} can incorporate linkage information across different ranges from multiple technologies, providing a general framework for joint haplotype inference. Second, $E(\mathbf{S})$ can be interpret as the log-likelihood for a haplotype configuration $\mathbf{S}$ given the observation M_{ij}. Assume that conflicting linkage information due to genotype/trans errors are rare, then $\chi_{ij} \approx 1$ and we have

$$E\left(\mathbf{S}\right) \approx (\text{\# of linkage consistent with } \mathbf{S}) - (\text{\# of linkage inconsistent with } \mathbf{S}).$$

If we further assume sequencing errors or technical artifacts to be random and have a combined error frequency δ, then the likelihood $p\left(\mathbf{S}|M_{ij}\right)$ of haplotype $\mathbf{S}$ given M_{ij} is given by

$$\begin{aligned} p\left(\mathbf{S}|M_{ij}\right) &\approx \delta^{\text{\# of inconsistent linkage}} \\ &= e^{\ln\delta\cdot\left(\text{\# of inconsistent linkage}\right)} \\ &= \exp\left[-2\ln\delta\cdot E(\mathbf{S}) + \ln\delta\cdot\left(\text{total \# of linkage}\right)\right]. \end{aligned}$$

The second term on the last line is a constant, therefore

$$E(\mathbf{S}) \approx -\frac{1}{2\ln\delta}\, p\left(\mathbf{S}|M_{ij}\right) + \text{Const}. \tag{4}$$

This relationship is convenient for the assessment of phase errors as described below.

Finally, Eq. (3) is a general formalism for solving binary haplotypes. Assume we have knowledge of local haplotype blocks $\mathbf{B}_s$ $(s = 1, 2, \cdots m)$ (and their complements $\overline{\mathbf{B}}_s$) and want to determine how they are combined on each chromosome, we can present the global haplotype as a vector of spins

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

We can then compute the linkage matrix between blocks similar to Eq. (1) as

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

and then solve b_s by minimizing

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

Computational strategies for solving the haplotype

Determining the haplotype from the linkage matrix given by Eq. (1) is equivalent to minimizing $E(\mathbf{S})$ given by Eq. (3). We solve this problem by alternately introducing two types of perturbations:

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

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

These two perturbations are equivalent to correcting single-base phasing errors and block-switching errors. 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

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

Starting from a random initial state

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

we alternately consider spin flipping at every site and block flipping between every pair of adjacent sites to improve haplotype inference, until the energy reaches a stable minimum against both perturbations.

As we iterate over every spin, it is straightforward to see that the overall complexity of calculating the flipping energy ΔE_i is proportional to the total number of non-zero M_{ij}'s. When computing $\Delta E_{k|k+1}$ between every pair of adjacent variant sites, we can take advantage of the following recursive relationship (The last step uses the symmetric property of $M_{ij} = M_{ji}$.):

$$\begin{aligned} \Delta E_{k|k+1} - \Delta E_{k-1|k} &= \sum_{i \leq k} \sum_{j > k} M_{ij} s_i s_j - \sum_{i \leq k-1} \sum_{j \geq k} M_{ij} s_i s_j \\ &= s_k \left(\sum_{j > k} M_{kj} s_j - \sum_{i < k} M_{ik} s_i \right) \\ &= s_k \left(\sum_{i > k} M_{ki} s_i - \sum_{i < k} M_{ki} s_i \right). \end{aligned} \tag{11}$$

Introducing

$$h_i^+ = \sum_{j > i} M_{ij} s_j \quad \text{and} \quad h_i^- = \sum_{j < i} M_{ij} s_j, \tag{12}$$

we can calculate both ΔE_i and $\Delta E_{k|k+1}$ as

$$\Delta E_i = s_i \left(h_i^+ + h_i^- \right) \quad \text{and} \quad \Delta E_{i|i+1} = \Delta E_{i-1|i} + s_i \left(h_j^+ - h_j^- \right). \tag{13}$$

Thus, the total complexity for calculating both ΔE_i and $\Delta E_{i|i+1}$ over all sites is proportional to the total number of non-zero M_{ij}'s.

After the total energy $E(\mathbf{S})$ reaches a steady state, we can assess switching or single base phasing errors by ΔE. The interpretation of $E(\mathbf{S})$ as a log-likelihood of a haplotype configuration $\mathbf{S}$ (Eq. (4)) suggests that the energy change ΔE_i or $\Delta E_{i|i+1}$ is related to the probability of single-base or block-switching phasing errors as

$$\epsilon = \frac{e^{-\beta \Delta E}}{1 + e^{-\beta \Delta E}}. \tag{14}$$

$\Delta E \sim 0$ indicates high probability of phasing or switching errors.

Features of linkage information derived from long-range and Hi-C sequencing

The energy landscape of $E(\mathbf{S})$ given by Eq. (3) is dependent on non-zero M_{ij} values. Contributions to M_{ij} come from both true molecular linkage and spurious linkage generated by technical artifacts. True molecular linkage connects variant genotypes on each homologous chromosome and should always be consistent with the chromosomal haplotype, whereas spurious linkage produces random connections between genotypes on both homologs that is 50% correct. It can be shown that if the signal of true molecular linkage outweighs that due to spurious linkage, then iteration by spin and block flipping is guaranteed to find one of the global minima, with the only degeneracy arising where no linkage exists between adjacent phase blocks (See Supplementary Discussion). It is therefore important to know the signal-to-noise characteristics of M_{ij} in different sequencing technologies.

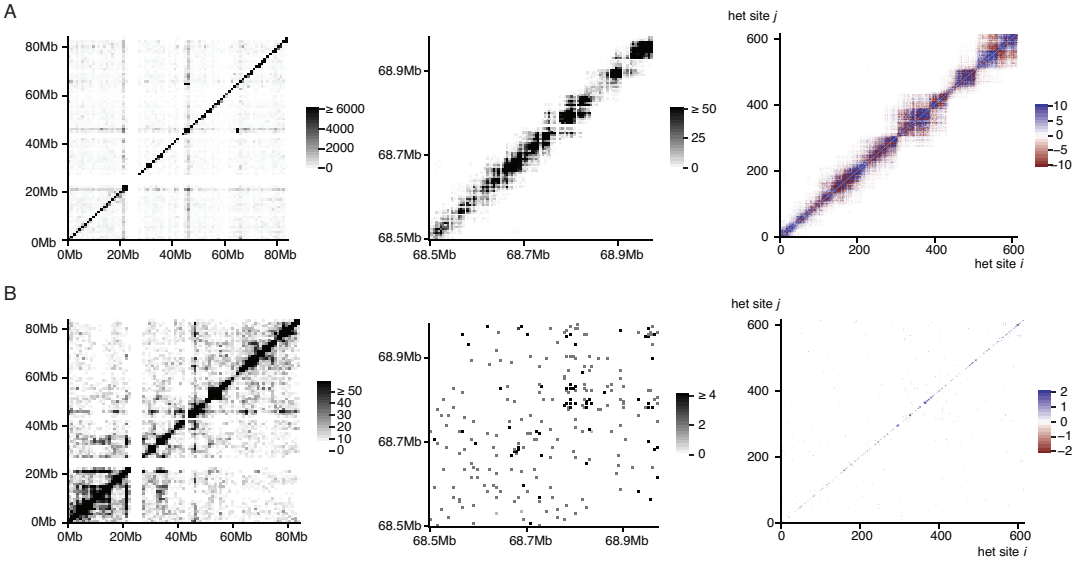

Figure 2. Density map of variant linkage from linked-reads and Hi-C sequencing. A. Density map of linkage between variants on Chr. 17 of RPE-1 cells from linked-reads sequencing generated from 10^7 links randomly sampled from a total of 6.1×10^7 links. Left: total variant linkage at 1Mb resolution; Middle: variant linkage at 5 kb resolution in a 500 kb region; Right: graphic representation of the linkage matrix defined in Eq. (1) calculated for the region shown in the middle panel. B. Density map of variant linkage generated from 160,000 long-range Hi-C links in $60\times$ Hi-C sequencing of RPE-1 cells.

Figure 2 shows the density of variant linkage derived from linked-reads and Hi-C sequencing data of RPE-1 cells (both at 60× average depth) at different resolutions. Although variant linkage derived from both technologies is enriched near the diagonal, there are several important differences. It is obvious from Fig. 2 that linkage generated by linked-reads data is mostly between variants less than 0.1 Mb, whereas linkage from Hi-C data is sparse but present at a much longer range. To quantitatively assess the distance dependence of linkage information from both technologies, we calculates three metrics of linkage between variants at different distances. (1) total linkage between variants (Fig. 3A); (2) percentage of linked variants (Fig. 3B); and (3) average number of links between linked variants (Suppl. Fig. X).

Comparing linkage from Hi-C data and linked-reads data (both at $60\times$ average depth), we observed that Hi-C data generate much less linkage information ($\sim 10^5$) than linked-reads data ($\sim 10^8$) and covers a tiny fraction of pairs of variant sites. The sparsity of Hi-C data implies that they are insufficient to generate contiguous phase blocks. By contrast, linkage information from linked-reads spans almost all variants within 10 kb and covers a significant fraction of variants separated by as much as 100 kb, which corresponds to the length of DNA molecules selected in linked-reads library construction. As any two variants on a long DNA molecule are linked, the density of variant linkage should be roughly constant within the size range of input DNA molecules but

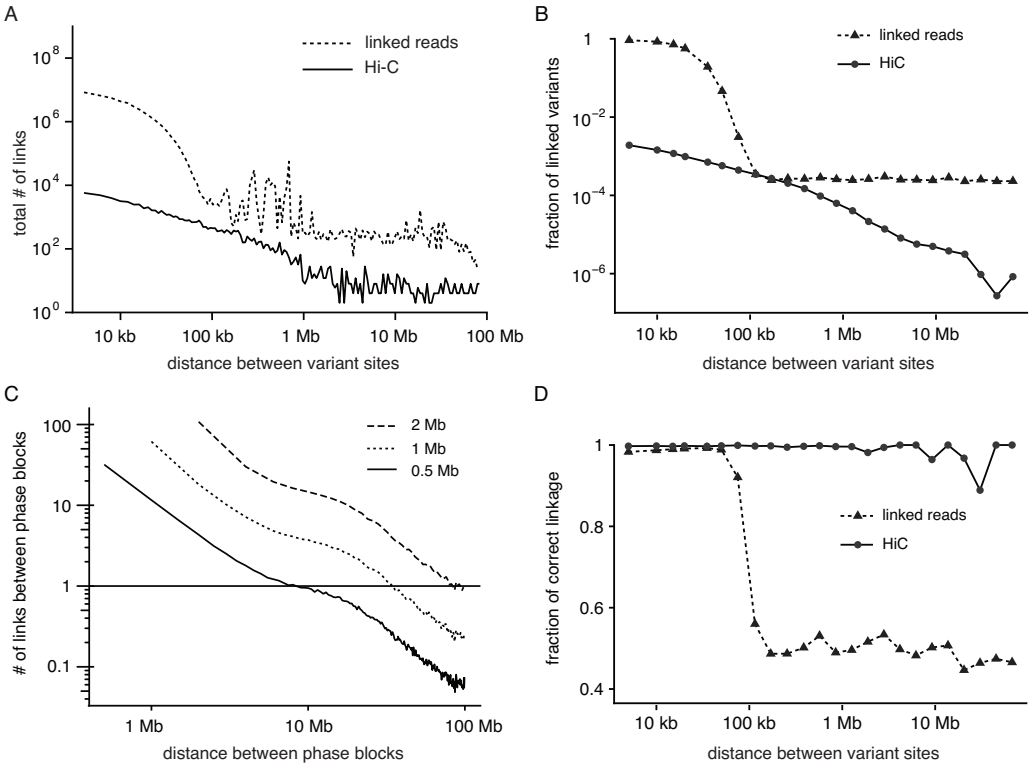

Figure 3. Distance dependence of variant linkage derived from linked-reads and Hi-C sequencing. A. Total number of links between variants at different distance. B. Fraction of linked variants at different distance. C. Average number of phased Hi-C reads connecting phasing blocks of 0.5, 1, and 2 Mb at different distance. D. Average percentage of *cis*-links at different distance.

quickly decay to zero above this range. In reality, there is a low-level background signal of linkage above 100 kb that is distance-independent and generated by technical artifacts. This is confirmed by comparing linkage from linked-reads data against complete whole-chromosome haplotypes (Fig. 3D) derived independently from sequencing of monosomic cells.

The capping of linkage information at 100 kb implies that haplotype inference from linked-reads data alone cannot bridge phase blocks across large gaps in the reference assembly (including all centromeres) and may perform poorly in variant-poor regions. The long-range linkage information missing in linked-reads data is however, available in Hi-C data. Variant linkage in Hi-C data is generated from proximity-based ligation between intra-chromosomal segments. The density of molecular linkage is proportional to the contact frequency between the variant sites, which shows a power-law decay as a function of their genomic distance. As shown in Fig. 3B, the power-law dependence of variant linkage extends above 10 Mb, suggesting that the majority of Hi-C linkage is intramolecular even at this range. This is again confirmed by comparing linkage from Hi-C reads against whole-chromosome haplotypes (Fig. 3D).

Although Hi-C data are too sparse to directly link individual variants, aggregation of sparse but accurate Hi-C links can produce a significant linkage signal between large phasing blocks even when they are far apart. This is demonstrated in Fig. 3C, which shows that phased Hi-C links between megabase-scale phase blocks can generate linkage signal at 10 Mb or above. For a human genome with an average variant density of 1 per 1.5 kb, there are approximately 600 variants in a 1Mb phase block. Given an average number of Hi-C links at 10 Mb distance of 10^{-5} (Fig. 3B), the total linkage between two 1Mb blocks separated by 1Mb is approximately

$$10^{-5} \times 600 \times 600 \approx 4.$$

Because Hi-C links are highly accurate, this signal is sufficient to connect phase blocks that are far

apart even when intermediate haplotype information is not available, *e.g.*, in centromeres or other long heterochromatin regions.

The accuracy of long-range Hi-C contacts further suggests a gradual merging strategy to construct long haplotypes. Given that the density of Hi-C contacts $\rho(l)$ between two loci separated by distance l scales approximately as l^{-1}, the total linkage between two blocks of size d scales as

$$\rho(l)d^2 \sim \frac{d^2}{l}.$$

If we require a minimum threshold of linkage signal to be ρ^*, the range of Hi-C linkage scales as

$$l^* \sim \rho^* d^2. \tag{15}$$

This linkage signal should allow us to grow phase blocks to l^* by merging, which will generate a linkage signal at even longer range. For example, if Hi-C data can produce linkage between 1 Mb phased blocks up to 10 Mb, this linkage signal will allow us to grow phase blocks to ~ 10 Mb or so; aggregation of Hi-C linkages between 10 Mb blocks will generate a specific linkage signal up till 10 Mb $\times 10^2$ =1000 Mb that exceeds even the longest chromosome in a human cell and enables whole-chromosome haplotype phasing. By contrast, starting with 0.1 Mb blocks, Hi-C linkage only extends to 10 Mb $\times 0.1^2$ =0.1 Mb and cannot extend the size of phase blocks. These considerations suggest that to effectively use Hi-C linkage, it is critical to first generate accurate phase blocks above 0.1 Mb.

The above results have led us to implement a two-tiered phasing strategy to generate whole-chromosome haplotypes combining linked-reads sequencing and Hi-C sequencing. First, we search for a haplotype configuration $\mathbf{S}$ that minimizes Eq. (3) using only linkage between variants separated by 100 kb or less from linked-reads data. We identify sites with a high probability of phasing errors using Eqs. (9) and Eq. (10) and break $\mathbf{S}$ at sites with potential switching errors to generate high-confidence local phase blocks. We then determine the linkage between these phase blocks using phased Hi-C reads spanning different blocks through minimization of Eq. (6) by the same iteration approach as in minimizing Eq. (3). The accuracy of local phase blocks ensures specificity of inter-block linkage from Hi-C data calculated using Eq. (5). The long-range linkage signal from Hi-C reads enables us to join all phased blocks within a single chromosome into one haplotype.

Local haplotype inference from linked-reads sequencing

We first determined local haplotypes of RPE-1 cells from linked-reads data (60× average depth) using the minimization approach. We restricted the analysis to bi-allelic single-nucleotide variant (SNV) sites detected by HaplotypeCaller (https://software.broadinstitute.org/gatk/) with minimum variant filtration. The binary spin model formalism cannot incorporate multi-allelic variants, but can include all bi-allelic variants. We excluded insertion/deletion (indel) events in haplotype inference due to their higher false detection rate than SNV. As will be discussed below, the density of SNV variants is sufficient to build complete parental haplotypes; indel variants can then be added to parental haplotypes based on their linkage to each haplotype.

As a measure of local phasing accuracy, we computed the energy penalties of site flipping and block switching given by Eqs. (9) and (10)); their distributions are shown in Fig. 4A and 4B. Both penalties show a bimodal distribution with an enrichment of sites with low energy penalty (*i.e.*, high error frequency). We further assessed the quantitative relationship between phasing accuracy and energy penalty and focused on switching errors as they will destroy long-range haplotype accuracy but flipping errors will not. We calculated the percentage of correct linkage between adjacent sites as a function of the switching error penalty (Fig. 4C) using the haplotype data derived from single-chromosome sequencing as the ground truth. As expected, phasing accuracy is correlated with switching error penalty. We set a cutoff for switching errors at the local minimum in Fig. 4B (dotted line) and cut the haplotype at sites with a lower switching penalty. With this cutoff, intra-block phasing accuracy was estimated to be $> 98\%$ (Fig. 4C, dotted line) and the N50 length of phase

blocks is 0.4 Mb (Fig. 4F). The length distributions of contiguously phased haplotype blocks and gaps between blocks are shown in Fig. 4D and 4E. The average gap between phased blocks is about 10 kb; most of these gaps correspond to regions with low variant density. Most phase blocks are in the range of 0.1-1Mb, with a few longer than 1Mb. The block length distribution indicates that there exists at least one region with low variant density (less than one variant per 10 kb) every 100 kb to 1Mb.

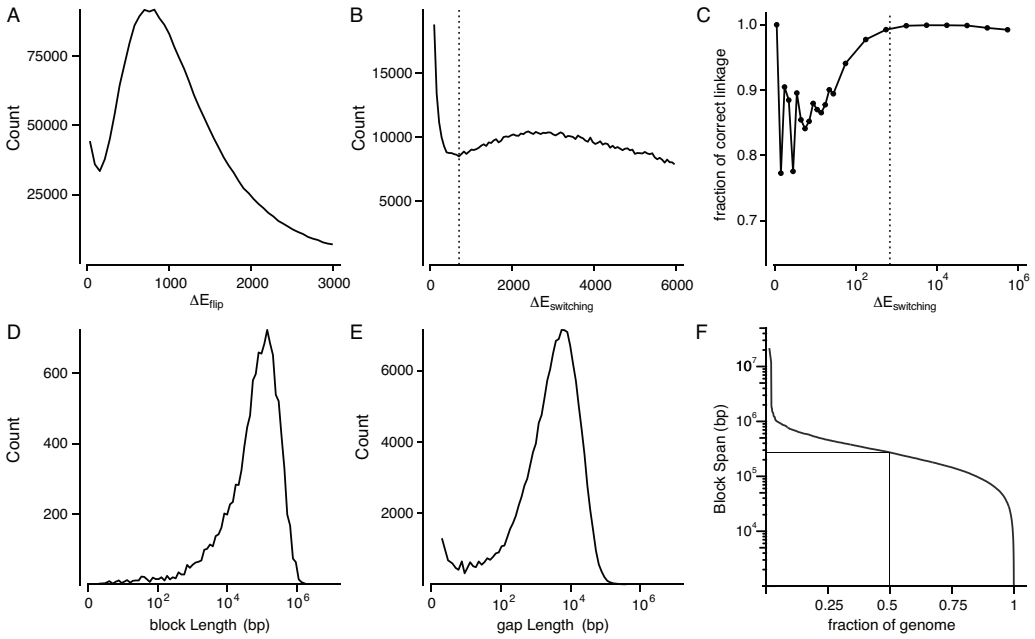

Figure 4. Accuracy and length distribution of haplotype phase blocks inferred from linked-reads sequencing. A and B. Distributions of the energy penalty as a quality score of single base (A) and block-switching (B) errors. Both quality scores show a bimodal distribution. C. Fraction of switching errors as a function of the energy penalty. Dotted line represents $\Delta E \approx 700$ corresponding to the local minimum (~ 700) in B, with > 98% accuracy. D and E. Length distribution of contiguously phased haplotype blocks (D) and gaps between blocks (E) with the switching energy score cutoff $\Delta E = 700$ F. Cumulative distribution of the fraction of phased genome against the the length of phased blocks with $\Delta E = 700$. The N50 phased block length is about 0.4 Mb.

We note that the choice of $\Delta E = 700$ corresponding to the minimum in Fig. 4B is conservative: Figure 4C indicates that even with $\Delta E = 100$, phasing accuracy is above 90%. A permissive cutoff can generate longer blocks, but at the cost of more frequent intra-block switching errors. The choice of a conservative cutoff ensures intra-block phasing accuracy, which is crucial for deriving accurate inter-block linkage from phased HI-C reads spanning different blocks. The average length of phase blocks at this conservative cutoff still exceeds 0.1 Mb and is sufficient to accumulate significant Hi-C linkage to build long haplotypes.

Merging of local phase blocks using linkage from Hi-C reads

After determining local phase blocks from linked-reads data, we then computed linkage between these blocks by aggregating all phased Hi-C reads spanning different blocks using Eq. (5). In this calculation we only included Hi-C links between sites no more than 10 Mb apart to suppress errors due to *trans* contacts. We further excluded short phase blocks with fewer than 5 total links to other blocks to make the minimization converge more quickly. Minimization of Eq. (6) produced completely phased haplotypes of each chromosome arm and the p- and q-arm haplotypes were connected into a single haplotype based on the number of phased Hi-C links between the arms (Figure 5).

The high energy penalty of switching (Fig. 5A) indicates that Hi-C reads generate very specific linkage signal between blocks. Indeed, the vast majority of phase blocks are linked correctly by Hi-C

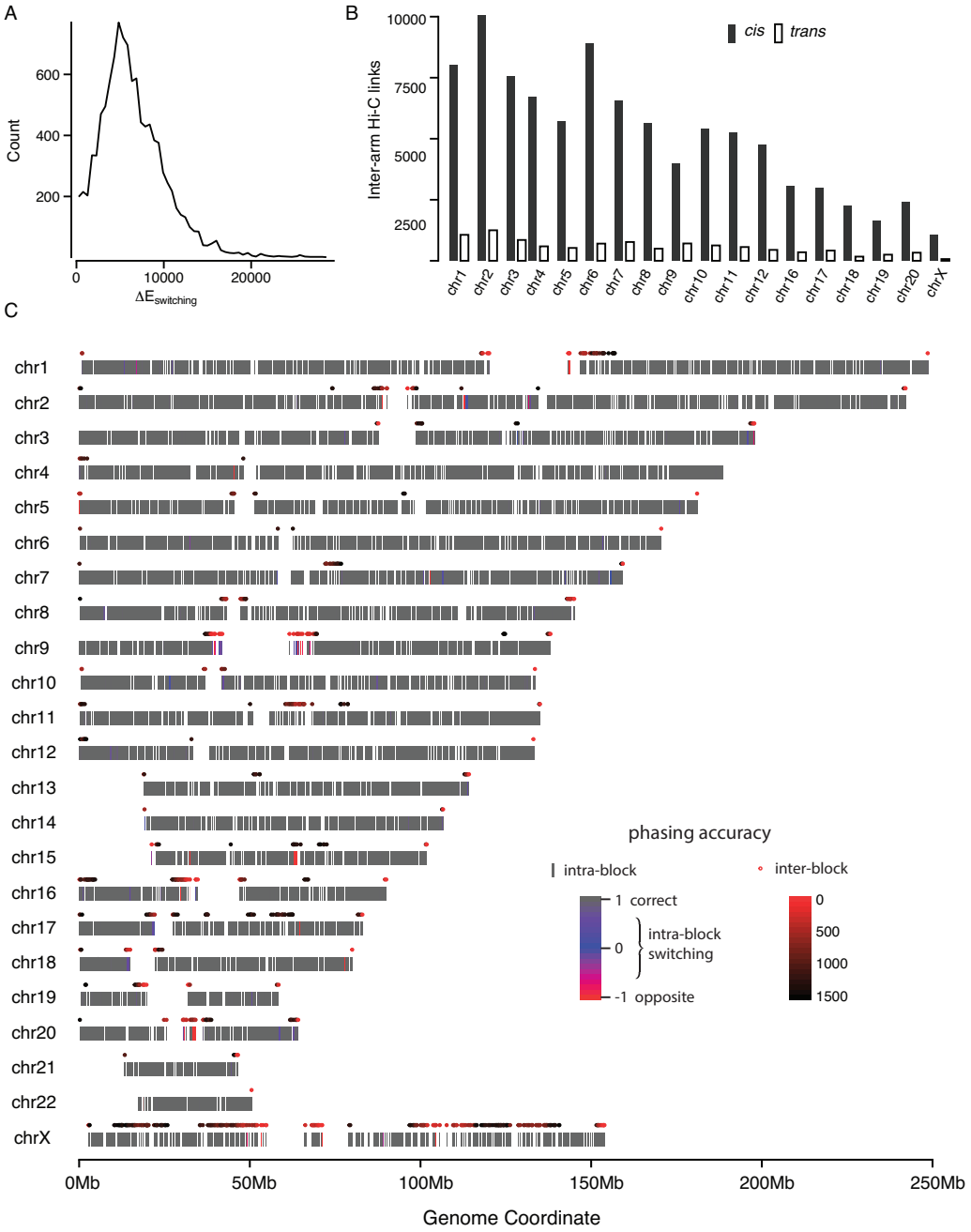

Figure 5. Whole-chromosome haplotype inference from linked-reads and Hi-C sequencing. A. Distribution of the energy penalty of switching between blocks. B. Total number of *cis* and *trans* Hi-C links between p- and q-arm haplotypes. C. Regions of low phasing accuracy and potential phasing errors. Each rectangle represents a phase block derived from linked-reads data; the face color of phase blocks reflects phasing accuracy measured by the percentage of agreement with whole-chromosome haplotypes determined from single-cell sequencing of monosomic chromosomes. Phase blocks in red show completely opposite haplotypes and correspond to block flipping errors; phase blocks in blue contained mixed haplotypes and reflect intra-block switching errors; both types of errors are rare and mostly localized to pericentric or subtelomeric regions. Regions with potential switching errors are indicated by red dots corresponding to low energy penalty of switching, most of which are located near the telomere or centromere. (Regions with high switching error penalty (>1,500) are not shown.) Switching errors in pericentric regions are not an issue as the p- and q-arm haplotypes are determined later by aggregating all Hi-C links between the arms. Chromosome X shows overall lower phasing accuracy (indicated by lower switching error penalty) than other chromosomes due to the low variant density.

reads (gray rectangles in Fig. 5C) when compared to reference haplotypes derived from sequencing of monosomic RPE-1 cells. We observed a small number of incorrectly phased blocks that are often adjacent to sites of potential switching errors (low switching energy penalty). Phasing errors include whole block flipping, when the haplotype phase is completely opposite (red rectangles), and intra-block switching, when a fraction of the haplotype is phased incorrectly. Both types of phasing errors are almost exclusively found in subtelomeric and pericentric regions with a high frequency of false variants. Chromosome X also has slightly lower phasing accuracy due to its low variant density.

The overall phasing accuracy of the computationally inferred haplotype is 98% (Table 1). We note that only 93.6% of input variants were phased. As will be discussed below, the majority of unphased variants are false variants with ambiguous or no linkage to flanking variants.

Table 1. Comparison of computationally inferred RPE-1 haplotype phase with single-cell data

ref_id	Total variant sites[a]	Sites for phasing[b]	Phased from bulk data[c]	Phased from monosomies[d]	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[e]	61,461	44,384	41,468	26,906	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 within a short deletion
[b]excluding sites in centromeric regions or not having linkage from both ref and alt genotypes
[c]whole-chromosome haplotype inferred from linked-reads and Hi-C sequencing
[d]whole-chromosome haplotype determined from single-cell sequencing of monosomies
[e]from one monosomic sample only

Application I: Variant filtration using phased linkage

A major source of false variant detection is due to incorrect alignment of short sequencing reads. Incorrectly aligned sequence reads (*e.g.*, due to low sequence complexity or incompleteness of the reference assembly) can be mistaken as variant-supporting reads in a locus that is homozygous. But mis-alignment also abolishes molecular linkage. The lack of linkage to flanking variants can be a specific filter against false variants that is robust against false alignment.

To test this idea, we computed four types of haplotype linkage for each heterozygous variant: reference-haplotype A, reference-haplotype B, alternate-haplotype A, alternate-haplotype B. Haplotype linkage can be calculated from the number of phased links, the number of linked variant sites, or the number of unique molecules. We found the number of unique molecules to be the most informative measure. We expect linkage from true heterozygous variants to have the following

features: First, both reference and alternate genotypes are linked to flanking sites but to opposite haplotypes. Second, the number of molecules showing reference or alternate genotypes should be proportional to the allelic DNA copy number; in disomic regions, the allelic ratio should be roughly 1:1.

We first determined the haplotype phase at every variant site based on linkage to either parental haplotypes. This increased the total number of phased sites but with a slightly lower phasing accuracy (Table 2). We further introduced two simple filters based on allelic coverage and haplotype linkage. Phasing accuracy is $\sim 98\%$ with the allelic ratio filter and $\sim 99\%$ with the linkage filter; combining both filters improves phasing accuracy to $> 99.5\%$. The significant overlap between sites preserved after either filter suggests that both filters work very effectively. Moreover, as filtered variants tend to fail both criteria, we expect most of the filtered variants to be false variants. With these filters, we estimated the total number of true heterozygous variants to be approximately 2.1 million and more than 98% of these are phased correctly using bulk linked-reads and Hi-C data.

We note that the majority of true variants after imposing either the allele ratio or the linkage filters were contained in the original haplotype phase (Table). Therefore, haplotype phasing by itself can also effectively remove false variants. With the parental haplotype information, we can calculate and use haplotype linkage to improve the accuracy of somatic variant detection.

Table 2. Comparison of computationally inferred RPE-1 haplotype phase with single-cell data

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 ratio[a]	2,153,678	2,148,402	2,069,123	2,017,695	46,625	0.023
Linkage[b]	2,116,557	2,116,536	2,033,513	2,013,413	20,085	0.010
Combined	2,017,897	2,017,897	1,943,901	1,933,260	10,641	0.006

[a]Minor allele ratio $\in [0.3, 0.7]$ in disomic regions, $\in [0.18, 0.48]$ in the trisomic region in Chr. 10q.
[b]At least one link connecting ref, alt, HapA, and HapB; plus minor linkage ≤ 2 or minor linkage/Total linkage ≤ 0.1

Application II: Digital karyotyping using phased Hi-C links

The capability to connect all genetic variants and chromatin alterations on each homologous chromosome is extremely powerful for investigating long-range chromatin interactions and chromosomal alterations. For example, comparison of Hi-C contacts phased to active and inactive X chromosomes has provided new insight into the folding of the inactive X chromosome and its regulation (***Darrow et al., 2016***). Because the majority of Hi-C reads are intra-chromosomal, it is possible to phase Hi-C contacts even if only one side of the Hi-C contact is phased.

The specificity of Hi-C reads arising from intra-chromosomal contacts implies that we can further use phased Hi-C links infer the structure of rearranged chromosomes ("digital karyotyping"). Because inter-chromosomal (including between homologous chromosomes) interactions are much more rare than intra-chromosomal, Hi-C contacts between different chromosomal segments in a rearranged chromosome result in a very specific signal at the sites of translocation (***Rao et al., 2014***; ***Dixon et al., 2018***; ***Zhou et al., 2019***). Moreover, as Hi-C links can extend very far along the chromosome (> 10 Mb), one can use this signal to assemble a scaffold of multiple rearranged segments. An example is shown in Fig. 6. RPE-1 cells contain an extra copy of a large segment from Chr. 10q (Fig. 6A). Sequence information at the breakpoint on Chr. 10q cannot resolve the partner locus, but cytogenetic analysis showed that the 10q segment is translocated to the end of Chr. Xq. Phased Hi-C contacts between Chr. 10 and Chr. X show a clear enrichment near the end of Chr. Xq and at the site of translocation on Chr. 10. Importantly, this enrichment is only present in one haplotype combination, thus not only confirming the cytogenetic analysis but further pinpointing the translocation to the active X chromosome. Remarkably, the phased linkage signal

persists through Chr. X and reaches the p-terminus. This indicates that the 10q segment is fused with a complete X chromosome and generates a digital karyotype of the 10q:X translocation.

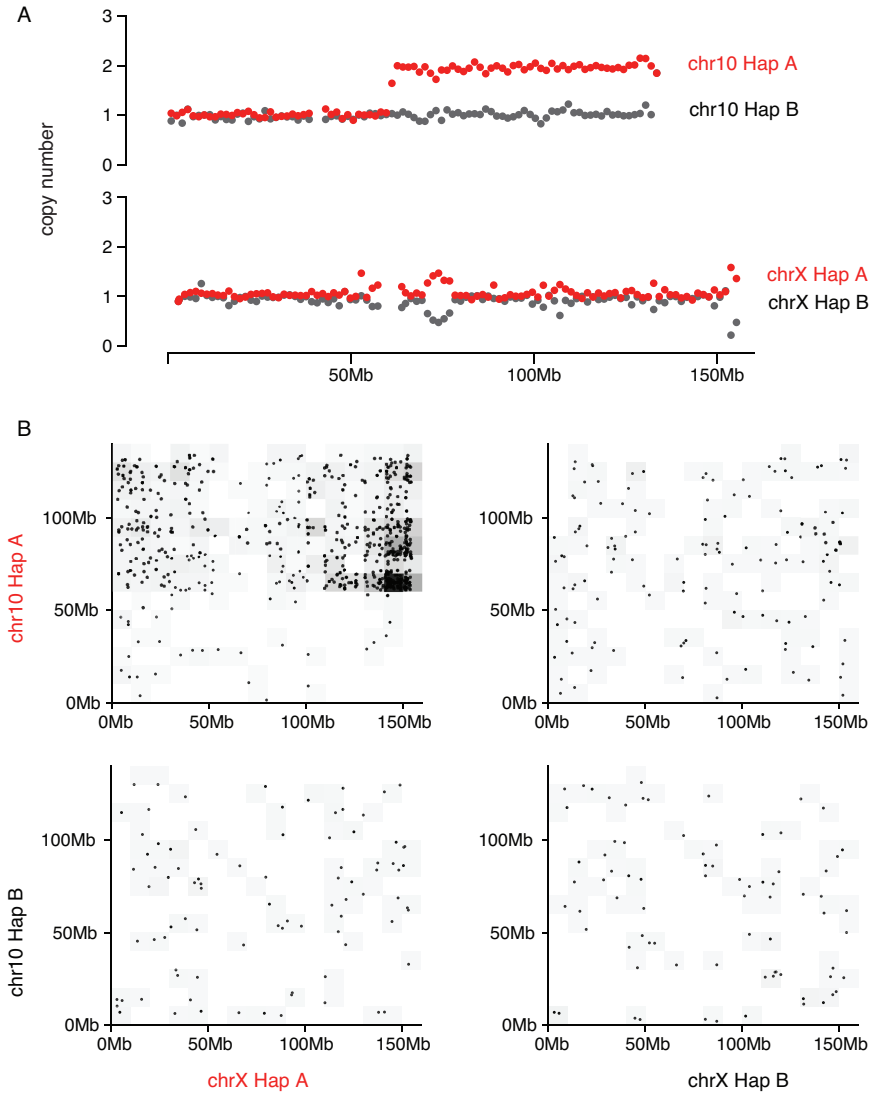

Figure 6. Digital karyotyping using phased Hi-C links. A. Segmental DNA copy number of both parental haplotypes of Chr. 10 and Chr. X, showing an extra copy of a segment from one Chr. 10q homolog. B. Phased Hi-C links (aggregated in 10 Mb bins) in four different haplotype combinations between Chr. 10 and Chr. X. A significant enrichment of Hi-C links is seen in only one specific haplotype combination, reflecting the translocation between specific homologs of Chr. 10 and Chr. X. The enrichment of Hi-C links is throughout the whole X chromosome of haplotype A, suggesting *cis*-linkage between Chr. X and 10q. There is also a slight increase of Hi-C contacts between haplotype A of Chr. 10q and haplotype B of Chr. X; this is due to elevated DNA copy number and much weaker than the signal of *cis*-linkage.

Discussion

Chromosomal haplotypes are essential to a complete description of polyploid genomes. Although standard shotgun sequencing cannot resolve haplotype information directly, various experimental strategies now enable preservation and retrieval of haplotype information through molecular barcoding of shotgun sequencing fragments that are derived from a single long DNA molecule. Haplotype information can also be inferred computationally using population panels or determined experimentally from sequencing of single chromosomes. Despite all these efforts, it remains

unclear whether complete whole-chromosome haplotypes can be accurately determined using some combinations of these approaches, or require new experimental technologies. Here we have demonstrated the feasibility to generate complete whole-chromosome haplotypes combining long-range sequencing and Hi-C sequencing using a two-tiered computational phasing approach. Comparison of the computationally inferred haplotypes with those determined experimentally from sequencing of single chromosomes validated that the computational inference shows no global switching errors across any chromosome and has an overall phasing accuracy above 98%.

The determination of complete haplotypes further enabled us to systematically analyze linkage information from linked-reads and Hi-C sequencing data. Our analysis showed that linked-reads sequencing produces dense linkage information that is highly accurate till the 100 kb range corresponding to the targeted size of high molecular DNA in linked-reads library preparation. By contrast, linkage information from Hi-C data is highly accurate throughout whole chromosomes but extremely sparse due to 3D chromatin organization. The generation of megabase-scale phase blocks using linked-reads data allows us to overcome the sparsity of linkage information from Hi-C data and take advantage of highly accurate long-range Hi-C linkage information to build whole-chromosome haplotypes. Our statistical analysis of linkage information from linked-reads and Hi-C data indicates that the combination of these two technologies (each at 60× depth) is sufficient to generate completely phased genomes at high accuracy.

Our computational method is built on a simple spin-glass model for binary haplotype configurations and converts haplotype inference to a minimization problem. This framework has several advantages. First, the accuracy of linkage information from linked-reads or Hi-C sequencing ensures fast convergence to an optimal haplotype configuration. This makes the overall computational complexity proportional to the total number of links and is a significant improvement over existing algorithms (***Edge et al., 2017***). Second, sites of potential phasing errors are automatically detected from the energy function during minimization with no additional computation. This enables quantitative assessment of phasing errors at every site and the generation of phase blocks of maximal size at high accuracy, which is critical for building long haplotypes from short phase blocks using sparse linkage data. Finally, the spin formalism allows integration of other data with phasing information, such as linkage from somatic variants, copy-number information, or known haplotypes from population data. In particular, the spin glass model formalism is invariant when there are allelic copy-number changes, including loss-of-heterozygosity.

We note that our phasing algorithm is robust against false variants (10-20%). This is because specific molecular linkage information is absent or ambiguous between false variant sites and their flanking sequences, resulting in poor haplotype inference at false variant sites that can be readily identified and excluded by the phasing algorithm. We further design a filter that can effectively distinguish between true and false heterozygous variants based on linkage between variant genotypes and parental haplotypes. The linkage filter is largely independent of the sequence context of variants and can be very useful for validating *de novo* variants that are difficult to detect accurately due to alignment inaccuracies, such as small insertion/deletions or structural variants.

A special application of haplotype phasing is digital karyotyping. The ability to phase Hi-C data using whole-chromosome haplotype information suggests a strategy to simultaneously interrogate spatial co-localization between thousands of loci in an allele-specific manner. The significant enrichment of Hi-C reads between loci in close proximity relative to those in distal regions provides a specific signal to detect long-range chromosomal rearrangements. In particular, phased Hi-C linkage information can be used to assemble the structure of rearranged chromosomes. As an example, we showed that combining Hi-C contacts and haplotype phase not only enables reconstruction of a translocated chromosome, but further reveals the parental alleles involved in the translocation. We are currently developing methods to analyze and construct a digital karyotype of aneuploid cancer genomes using this strategy.

Besides high-resolution genome analysis, whole-chromosome haplotype provides a channel to relate local genetic/epigenetic alterations to variations in gene expression, and to global molecular

changes on the chromosome. The combination of haplotype-resolved genomic analysis and single-cell sequencing can serve as a high-throughput assay to probe molecular changes of the chromatin (both genetic and epigenetic alterations) and relate these changes to transcriptional changes or other molecular/cellular phenotypes. As molecular changes usually affect one homologous chromosome at a time, sequencing readout from the other homolog provides as intrinsic control that is not available in traditional molecular genetic assays.

In summary, we have developed a computational approach to generate whole-chromosome haplotypes with high accuracy and completeness from two widely used bulk sequencing methods. This approach offers a convenient and scalable solution to generating completely phased genomes of individuals as well as experimental cell lines, setting the stage for precise 2D and 3D genome analysis with haplotype resolution.

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. (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.

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*, pages 1–16.

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

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.

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.

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.

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.

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.

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.

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.

Tlemann-Boege, I. and Betancourt, A., editors (2017). *Haplotyping*. Methods in Molecular Biology. Humana Press.

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.