Bioinformatics, 34, 2018, i211–i217
doi: 10.1093/bioinformatics/bty286
ISMB 2018

OXFORD

Haplotype phasing in single-cell DNA-sequencing data

Gryte Satas[1,2] and Benjamin J. Raphael[1,*]

[1]Department of Computer Science, Princeton University, Princeton, NJ 08540, USA and [2]Department of Computer Science, Brown University, Providence, RI 02912, USA

*To whom correspondence should be addressed.

Abstract

Motivation: Current technologies for single-cell DNA sequencing require whole-genome amplification (WGA), as a single cell contains too little DNA for direct sequencing. Unfortunately, WGA introduces biases in the resulting sequencing data, including non-uniformity in genome coverage and high rates of allele dropout. These biases complicate many downstream analyses, including the detection of genomic variants.

Results: We show that amplification biases have a potential upside: long-range correlations in rates of allele dropout provide a signal for phasing haplotypes at the lengths of amplicons from WGA, lengths which are generally longer than than individual sequence reads. We describe a statistical test to measure concurrent allele dropout between single-nucleotide polymorphisms (SNPs) across multiple sequenced single cells. We use results of this test to perform haplotype assembly across a collection of single cells. We demonstrate that the algorithm predicts phasing between pairs of SNPs with higher accuracy than phasing from reads alone. Using whole-genome sequencing data from only seven neural cells, we obtain haplotype blocks that are orders of magnitude longer than with sequence reads alone (median length 10.2 kb versus 312 bp), with error rates $<$2%. We demonstrate similar advantages on whole-exome data from 16 cells, where we obtain haplotype blocks with median length 9.2 kb—comparable to typical gene lengths—compared with median lengths of 41 bp with sequence reads alone, with error rates $<$4%. Our algorithm will be useful for haplotyping of rare alleles and studies of allele-specific somatic aberrations.

Availability and implementation: Source code is available at https://www.github.com/raphael-group.

Contact: braphael@cs.princeton.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

In recent years, single-cell DNA-sequencing technologies have enabled the measurement of the genomic changes in individual cells (Gawad *et al.*, 2016). This technology has been used to measure somatic mutations in normal tissue (Lodato *et al.*, 2015 ; McConnell *et al.*, 2013), to quantify somatic evolution in cancer (Navin, 2015; Wang *et al.*, 2014), to investigate the genomes of unculturable microorganisms (Marcy *et al.*, 2007), and for other applications.

Unfortunately, it is not yet possible to directly sequence the DNA molecule(s) present in a single cell. Rather, current single-cell DNA-sequencing technologies first perform whole-genome amplification (WGA), in order to obtain sufficient DNA to sequence. Several WGA methods have been introduced, with the three most common being degenerate oligonucleotide primed PCR (DOP-PCR), multiple displacement amplification (MDA), and multiple annealing and looping-based amplification cycles (MALBAC) (Gawad *et al.*, 2016). The lengths of the amplified genomic fragments, or amplicons, range from 200 to 300 bp for DOP-PCR, to 1–5 kb for MALBAC, and up to 10–100 kb for MDA (Sherman *et al.*, 2017). As WGA uses repeated cycles of amplification, any errors or non-uniformity in coverage obtained during early cycles of amplification are amplified in later cycles. Thus, WGA results in highly non-uniform coverage of the genome Figure 1a, which is apparent in the observed strong correlations between read depth at genomic loci within the distance of an amplicon (Zhang *et al.*, 2015).

The amplification bias resulting from WGA leads to challenges in identifying genomic variants in single cells, and thus is a negative of single-cell sequencing technology. Considerable efforts have been made to develop analysis algorithms that overcome this bias

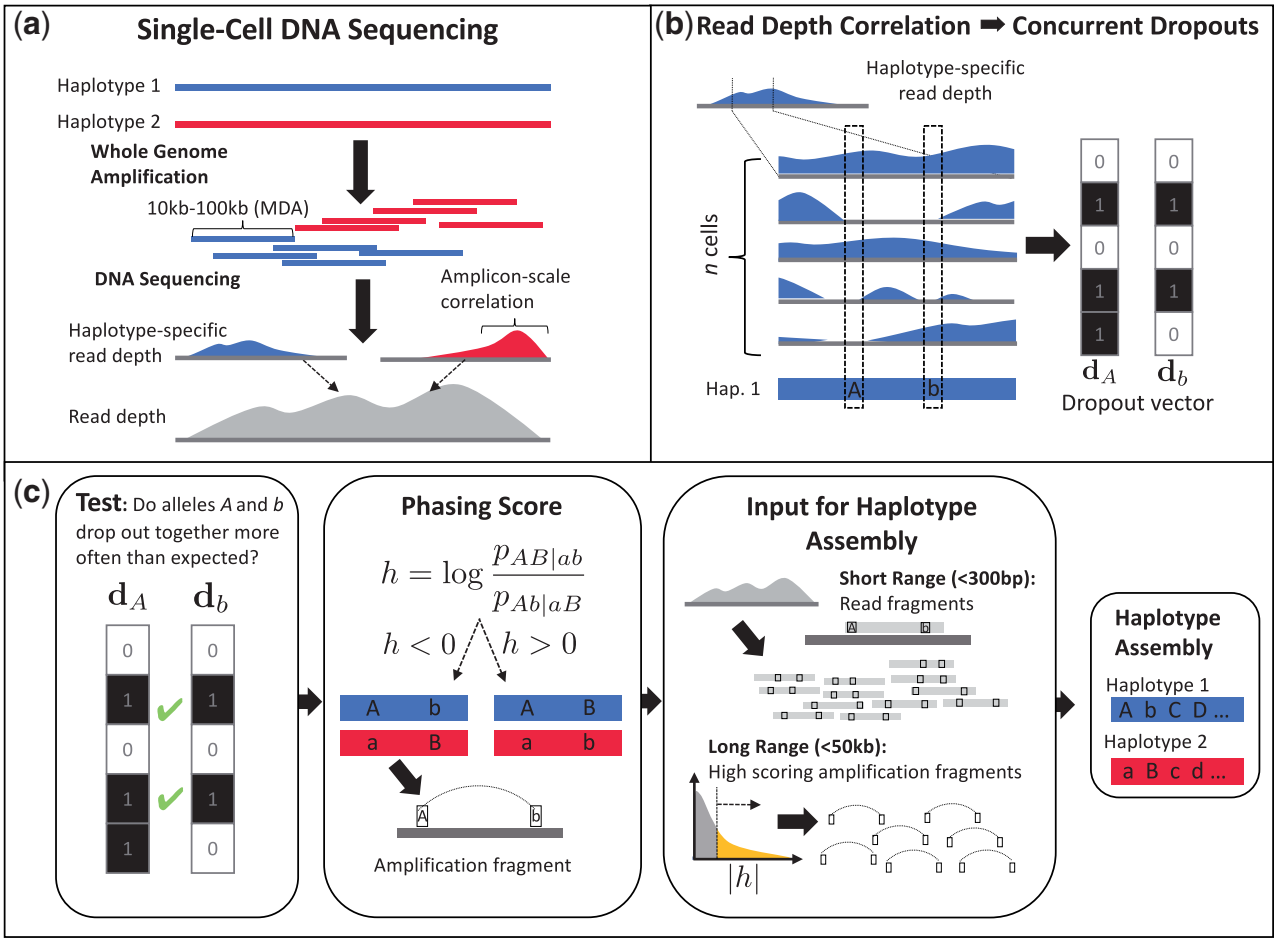

Fig. 1. (a) Single-cell DNA sequencing typically requires WGA to obtain sufficient quantities of DNA, which results in non-uniform read depth with correlation at scale of amplicons. Since the two homologous chromosomes are amplified independently, read-depth correlations are strongest between sequence reads originating from the same chromosome/haplotype. **(b)** Amplicon-scale read-depth correlations, combined with high rates of allelic dropout result in increased rates of concurrent allelic dropout for pairs of alleles originating from the same haplotype, where entries of the dropout vectors d_A and d_b indicate whether alleles *A* and *b*, respectively, are measured in each cell. **(c)** We derive a phasing score for pairs of nearby SNPs based on the *P*-values of concurrent dropout for different phasings of alleles. High or low values of the phasing score correspond to amplification fragments containing pairs of alleles that are likely to be on the same haplotype. These amplification fragments are used as input to haplotype assembly algorithms, augmenting phasing information from read fragments containing alleles found on the same read

(Bakker *et al.*, 2016; Garvin *et al.*, 2015) and to design WGA methods with less bias (Chen *et al.*, 2017; Picher *et al.*, 2016).

In this article, we demonstrate a positive aspect of amplification bias: since neighboring genomic loci are often co-amplified, the correlation in sequence coverage between neighboring alleles on the same chromosome can be used to phase haplotypes in diploid genomes (Fig. 1b). Diploid genomes, such as the human genome, consist of pairs of homologous chromosomes, distinguished by single-nucleotide polymorphisms (SNPs), and other small genomic differences. Current DNA-sequencing technologies yield sequence reads that originate from a mixture of both homologous chromosomes, losing information about the chromosomal origin of each read. Thus, for any pair of heterozygous SNPs that are further apart than a read length, the alleles that are present on the same chromosome are unknown.

Haplotype assembly is the process of reconstructing the haplotypes of an individual—i.e. assigning the alleles of heterozygous SNPs to the corresponding chromosome of origin—from sequence reads obtained from an individual. Since each read generally derives from a single chromosome, if a read spans multiple SNPs, then the observed alleles are presumed from a single haplotype. Haplotype and SNP phase information has applications in population genetics (Tewhey *et al.*, 2011) and clinical and medical genomics (Glusman *et al.*, 2014; Roach *et al.*, 2010; van de Ven *et al.*, 2012) as well as being used to improve other analyses such as SNP imputation and genotyping (Browning and Yu, 2009; Marchini *et al.*, 2007) and somatic variant calling (Bohrson *et al.*, 2018). Obtaining long continuous haplotype blocks is a challenge as the distance between adjacent SNPs is longer than reads and read fragments in most sequencing technologies. Long reads and other technologies that provide long-range phase information, such as linked-read and Hi-C data, have been used to improve the length of haplotype assemblies (Edge *et al.*, 2017; Patterson *et al.*, 2015; Pirola *et al.*, 2016; Zheng *et al.*, 2016), and there are microfluidic techniques designed to recover haplotypes of single-cell data (Chu *et al.*, 2017; Fan *et al.*, 2011). However, none of these techniques can be applied to existing short-read single-cell data, and may be prohibitively expensive for new experiments.

We describe an algorithm that exploits amplification bias across a collection of sequenced single cells to assemble haplotypes

(Fig. 1c). This algorithm is based on the observation in (Zhang *et al.*, 2015) that homologous chromosomes are amplified almost independently during WGA, and thus show similar rates of amplification bias. Thus, amplicon-scale correlations in read depth provide a signal to phase heterozygous SNPs across amplicon lengths. Specifically, our model leverages *allele dropout*, where one of the two alleles of a heterozygous SNP is not measured in a cell, a common feature of single-cell sequencing data (Gawad *et al.*, 2016). Alleles of two nearby SNPs on the same chromosome are likely to drop out (not be covered by an amplicon) concurrently. We derive a statistical test of concurrent dropout of alleles of heterozygous SNPs. We validate our approach using both whole-genome and whole-exome DNA single-cell sequencing data, and show that our approach predicts haplotype phase with high accuracy, achieving >90% accuracy on top-ranked 22% of pairs of SNPs within amplicon-length distances. We use pairs of SNPs exhibiting high rates of concurrent dropout to define amplification fragments that we input into an existing haplotype assembler (Edge *et al.*, 2017). We obtain haplotype blocks that are three to four orders of magnitude larger (10.2 kb versus 312 bp on whole-genome data, 9.2 kb versus 41 bp on whole-exome data) than obtained using read information alone, with low increase in assembly errors.

2 Materials and methods

WGA methods used in single-cell sequencing result in datasets with high rates of *allele dropout*, where alleles that are present in the genome are not observed in a sequenced cell. The primary cause of allele dropout is the failure of amplification of a genomic region from one of the two homologous chromosomes during WGA. Thus, one expects to observe correlations in rates of allele dropout between alleles of SNPs whose distances are within the length of an amplicon (up to 10–100 kb, depending on the WGA method). More specifically, since an amplicon contains DNA sequence from one homologous chromosome, one expects to observe a higher rate of concurrent dropout (dropout of both alleles in one cell) for a pair of alleles on the same haplotype and within the length of an amplicon, than for two alleles on different haplotypes. We describe a statistical test to evaluate such concurrent dropout between alleles for a pair of SNPs. We then use pairs of alleles which show strong evidence of significant concurrent dropout as input to a haplotype assembly algorithm.

2.1 Quantifying concurrent dropout

We obtain DNA-sequencing data from n single cells from the same individual, and assume that these cells share m heterozygous SNPs. Consider a pair of heterozygous SNPs with alleles A, a for the one SNP and alleles B, b for the other SNP. For an allele A, we define the *dropout vector* $\mathbf{d}_A$ to be a binary vector of length n, where $d_{A,s} = 1$ if we do not observe any reads containing allele A in cell s, and $d_{A,s} = 0$ otherwise. Thus, d_A indicates the dropout for allele A across cells. For an allele A, let $n_A = \sum_s d_{A,s}$ be the number of *dropouts*, or cells where A is not measured. Similarly, for a pair of alleles A and B, let $n_{AB} = \mathbf{d}_A \cdot \mathbf{d}_B = \sum_{s=1}^{n} d_{A,s} \times d_{B,s}$ be the number of *concurrent dropouts* of alleles, i.e. the number of cells where both A and B are not observed. The key idea of our model is that if the distance between SNPs is less than the length of an amplicon, and if alleles A and B are on the same haplotype, then concurrent dropout of A and B is more frequent than expected by chance. Conversely, if SNPs are far apart or A and B are present on different haplotypes, then we expect that amplification of these alleles is independent, and concurrent dropouts are random events.

Let N_{AB} be a random variable indicating this number of concurrent dropouts between alleles A and B across n cells. Under the null model, dropouts between allele A and allele B are independent. To compute the distribution of N_{AB} under the null, we need to compute the probability $w_{X,s} = \Pr(d_{X,s} = 1)$ that allele X drops out in cell s for each allele and cell. However, $w_{X,s}$ varies by locus, allele X, and cell s. Locus-specific and allele-specific variability in dropout rates results from context-specific amplification, sequencing or alignment biases. Cell-specific variability in dropout rates results from differences in sequencing depth and uniformity of coverage across cells. Since it is difficult to model each of these effects directly, we instead compute a weighted exact distribution $\Pr(N_{AB} = n_{AB} | n_A, n_B, \mathbf{w})$, where $\mathbf{w}$ are cell-specific weights obtained from the observed number of dropouts across all loci in each cell. See Leiserson *et al.* (2016) for details of similar weighted tests used in other biological applications.

Specifically, let $\mathbf{D}$ be the $2m \times n$ matrix whose rows correspond to the 2 m dropout vectors for the set of alleles of m heterozygous SNPs. We compute the P-value $\Pr(N_{AB} \geq n_{AB} | \mathbf{D}_{1\cdot}, \ldots \mathbf{D}_{2m\cdot}, \mathbf{D}_{\cdot 1} \cdots \mathbf{D}_{\cdot n})$ of observing n_{ab} or more concurrent dropouts, conditioned on the observed row sums and column sums of the matrix $\mathbf{D}$. Computing this P-value is non-trivial. Leiserson *et al.* (2016) introduced the WExT algorithm to compute a saddlepoint approximation of a P-value for the related problem of mutually exclusive events. We use a recent release of the WExT software that computes a saddlepoint approximation for the co-occurrence test statistic N_{AB}.

2.2 Augmenting haplotype assembly with amplification fragments

Haplotype assembly is the reconstruction of haplotypes from local information about groups of alleles that are present on the same chromosome. The input for haplotype assembly is a set of fragments $\mathcal{F}$, where a fragment $f \in \mathcal{F}$ defines a phasing over a set of SNPs—e.g. $f = ABc|abC$, where | delineates the two chromosomes, indicates alleles A, B, c are on one chromosome and alleles a, B, C are on the other. Haplotype assembly algorithms aim to find the most likely haplotypes from the set $\mathcal{F}$. Typically, the set $\mathcal{F}$ of fragments is equal to $\mathcal{F}_r$, the set of sequenced reads (or paired-end reads in the case of mate pair libraries). This is because alleles of two SNPs measured on the same read (or paired-read) are highly likely to reside on the same haplotype.

We extend the fragment set $\mathcal{F}$ using a set of *amplification fragments* defined from pairs of alleles from neighboring SNPs that demonstrate concurrent dropout, using the statistical test defined in the previous section. A pair of SNPs, having alleles A, a for one SNP and alleles B, b for the second SNP, can be phased in two ways: $AB|ab$ or $Ab|aB$. As noted by Zhang *et al.* (2015), during WGA, homologous chromosomes are amplified independently. Thus, we assume the random variables N_{AB} and N_{ab} are independent under the null hypothesis, and combine the P-values P_{AB} and P_{ab} using Fisher's method to obtain a single P-value $P_{AB|ab} = \Pr(T \geq t)$ where $t = -2; (\log P_{AB} + \log P_{ab})$ and T follows the χ^2_4 distribution, since two P-values are combined. For each pair of SNPs, this procedures yields two P-values, $P_{AB|ab}$ and $P_{Ab|aB}$, corresponding to the strength of evidence against the null model of independence for each phasing. Under phasing $AB|ab$, we expect high dropout concurrence for allele pair $\{A, B\}$ and allele pair $\{a, b\}$ and independent dropout for allele

pairs $\{A, b\}$ and $\{a, B\}$. Thus, we expect to see a low P-value $P_{AB|ab}$ and a high P-value $P_{Ab|aB}$.

We summarize the evidence in support of each phasing using a *phasing score* defined as

$$h = -\log\left(\frac{P_{AB|ab}}{P_{Ab|aB}}\right). \quad (1)$$

Here, $h > 0$ indicates stronger evidence for phasing $AB|ab$, while $h < 0$ indicates stronger evidence for phasing $Ab|aB$. Large values of $|h|$ suggest that the allele pairs for one phasing have more concurrent dropout than expected by chance compared with the allele pairs for the other phasing, and thus are more likely to have come from same haplotypes. We define a set of amplification fragments $\mathcal{F}_c$ containing allele pairs whose absolute values of the phasing scoring $|h| \geq c$, for a non-negative threshold c. We then use a haplotype assembly algorithm, HapCut2 (Edge *et al.*, 2017) to assemble haplotypes using the combined set of sequence and amplification fragments $\mathcal{F} = \mathcal{F}_r \cup \mathcal{F}_c$ as input.

3 Results

We tested our model on two single-cell DNA-sequencing datasets: whole-genome sequencing of neurons (Lodato *et al.*, 2015) and whole-exome sequencing of breast cancer cells (Wang *et al.*, 2014).

3.1 Whole-genome single cell sequencing of neurons

We first evaluated the performance of our model on synthetic diploid cells constructed from sequenced X chromosomes in single-cell DNA sequencing of neurons from a male individual, UMB1465 (Lodato *et al.*, 2015). This dataset includes whole-genome sequencing from $n = 16$ single cells and two bulk samples. The single cells were amplified using MDA, and a recent analysis (Sherman *et al.*, 2017) estimates that these samples had MDA amplicons lengths up to 200 kb, with a median of 19 kb and 95th percentile of 103 kb. To create synthetic diploid cells, we extracted the haploid X chromosomes from each sequenced cell (Fig. 2a), excluding the pseudo-autosomal regions PAR1 and PAR2 on chromosome X as defined in human reference genome assembly GRCh37.p13. To account for variability in dropout rates between cells, we downsampled all cells to a fixed number of reads, and formed pairs of cells based on the total number of covered positions in the samples. We generated simulated haplotypes using population-level allele frequencies acquired from dbSNP, and spiked these alleles into the sequencing data. We introduced sequencing error into the spiked-in alleles based on the Phred quality scores of individual reads. Further details on the simulation can be found in Supplementary Material S1.

To infer haplotypes, we applied the statistical test described in Section 2 to all pairs of SNPs within 50 kb of each other. Next, we constructed a set $\mathcal{F}_c$ of amplification fragments for a phasing score value $|h| >= c$ and input the amplification fragments $\mathcal{F}_c$ and read fragments $\mathcal{F}_r$ into HapCut2 to assemble haplotypes.

We evaluated the ability of the phasing score to accurately phase pairs of SNPs at varying distances and over a range of values $|h|$ (Fig. 2b). We found that the proportion of fragments whose phase was predicted correctly, increases with the absolute value $|h|$ of the phasing score, indicating that $|h|$ is correlated with the the accuracy of phasing. With larger amplicon lengths (50 kb–1 Mb and 1–5 Mb) the fragment accuracy decreases rapidly as the phasing score decreases. In addition, relatively few fragments have high phasing scores $|h|$, demonstrating that few SNPs are accurately phased when the distance between SNPs exceed the length of WGA amplicons. For haplotype assembly, we use pairs of SNPs whose distance is 50 kb or less. Over this set of SNP pairs, we are able to correctly phase fragments with 77% accuracy overall. However, if we restrict to the set of fragments with scores $|h| \geq 1$, the top 22% of fragments, we obtain an accuracy of 91%.

We assemble haplotypes using amplification fragments $\mathcal{F}_c$ and read fragments $\mathcal{F}_r$ as input to HapCut2 (Edge *et al.*, 2017) as described in Section 2. We ran HapCut2 with a range of thresholds c for the phasing scores, where for each c, we supply HapCut2 with $\mathcal{F}_r \cup \mathcal{F}_c$, the set of sequence fragments and amplification fragments. We also run HapCut2 using only sequence reads $\mathcal{F}_r$ to get a baseline measure of results using only short reads. In each case, we ran HapCut2 with default parameters.

We evaluate the resulting haplotype assemblies using the following metrics (Fig. 2c).

1. The *N50* is the length of a haplotype block such that half of all phased variants are in a block at least as long.
2. *Switch error* is the proportion of phase connections between adjacent SNPs that are incorrect.

Not surprisingly, we observe a trade-off between switch error rate and length of resulting haplotype blocks, Without any amplification fragments ($\mathcal{F}_c = \varnothing$), HapCut2 obtains haplotype blocks with a median (N50) length of 312 bp. As the phasing score threshold c decreases, the block lengths increase by several orders of magnitude with only relatively small corresponding increases in switch error rate With phasing score threshold $c = 2.25$, we obtain a block length of $N50 = 10.2$ kb, with a switch error rate of 0.02. At lower values of phasing score threshold, we see larger increases in error, which corresponds to the observed decrease fragment accuracy at the same values (Fig. 2b). Depending on the downstream analysis that utilizes the resulting haplotypes, these low thresholds may prove useful. For example, at a threshold of $c = 0$, we acquire blocks with an $N50 =$ 9.96 Mb, with 92.6% of blocks containing no switch errors.

3.2 Whole-exome single-cell sequencing of breast cancer

Whole-exome sequencing comprises a significant proportion of available single-cell sequencing datasets (Navin, 2015), due to both the lower sequencing costs compared with whole-genome sequencing and because of interest in measuring variation in coding regions. Assembling haplotypes from short-read whole-exome sequencing data is challenging as reads (or paired reads) are shorter than the lengths of most introns. Introns are estimated to have a median length of ~1334 bp in the human genome (Hong *et al.*, 2006), while the median exon size is ~122 bp (International Human Genome Sequencing Consortium, 2001). Thus haplotype phase is difficult to determine across exons from short read data.

Since WGA amplicons are typically longer than an intron, we hypothesized that we could use our model to obtain haplotype blocks from single-cell whole-exome data that were substantially longer than blocks obtained with short-read sequencing data. To test this hypothesis, we evaluated our model on single-cell whole-exome data from a triple-negative breast cancer patient (Wang *et al.*, 2014). We observed that one copy of Chromosome 17 was lost in eight cancer cells in this datasets that were indicated to be hypodiploid. Using these eight cells we obtained the true haplotypes for Chromosome 17 (Fig. 3a). We then applied our model to 14 normal (non-cancerous) diploid cells from the same individual. We ran the model and constructed amplification fragments as in previous

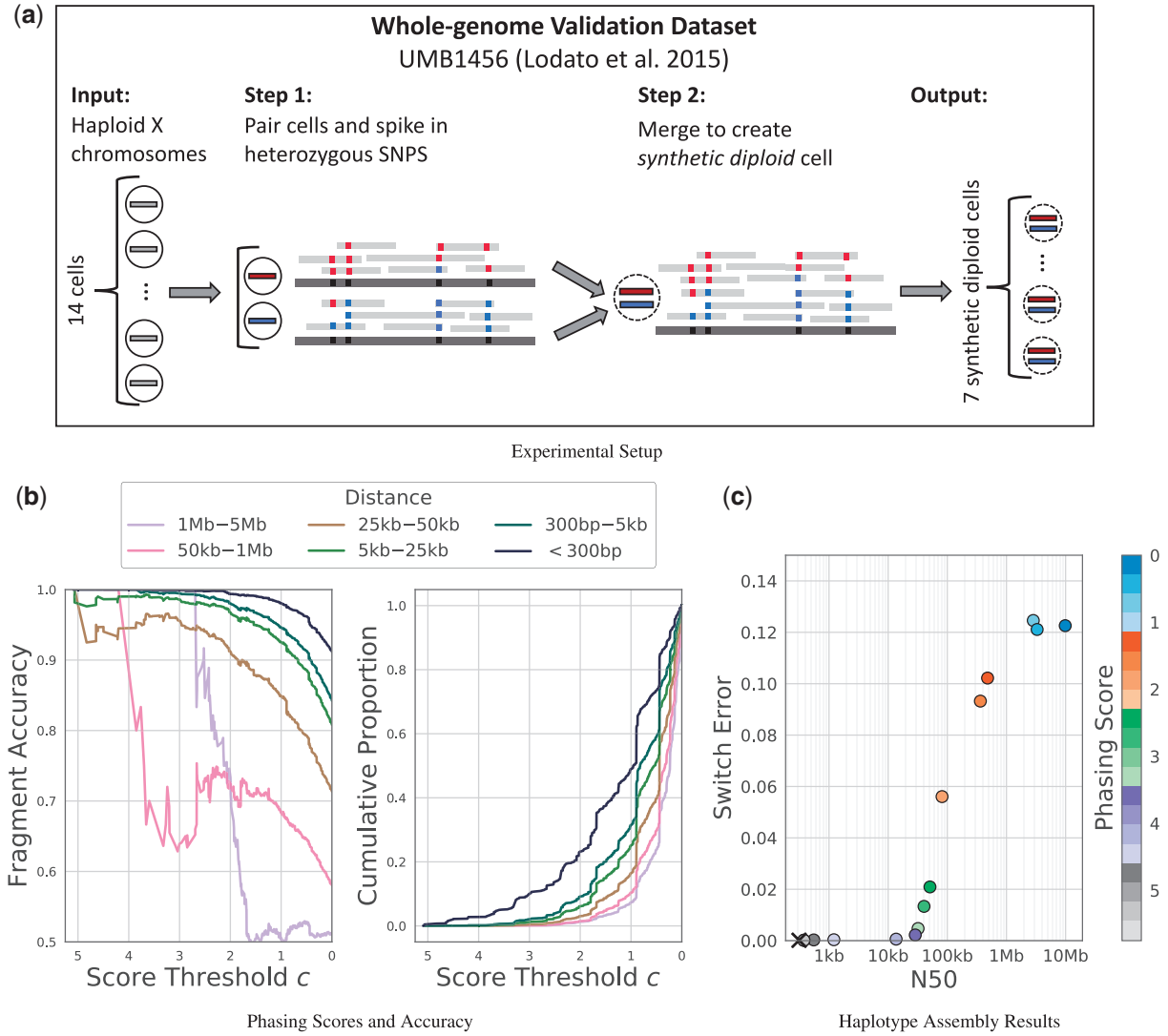

Fig. 2. Haplotype assembly on whole-genome DNA-sequencing data. **(a)** We form a validation dataset of seven synthetic diploid cells with known haplotypes from X chromosomes in whole-genome DNA-sequencing data of single neuron cells from a male (Lodato *et al.*, 2015). **(b)** (Left) The accuracy of the predicted phase for the set $\mathcal{F}_c$ of amplification fragments with the absolute value of the phasing score $|h| > c$. We observe highly accurate prediction of phase for pairs of SNPs whose distance is less than the length of amplicons (here 95th percentile of amplicon length is 103 kb). (Right) The proportion of SNP pairs included in the set of amplification fragments $\mathcal{F}$. **(c)** The *N*50 and switch error for haplotype assembly as we vary the phasing score threshold *c*. The *N*50 and switch error for the haplotype assembly with no amplification fragments is marked with an '×'

sections. Figure 3b shows the features of the resulting haplotype assemblies.

Using only read fragments $\mathcal{F}_r$, the haplotype assemblies have a median block length of $N50 = 41$ bp, which, as expected, is shorter than the length of a single exon. Using amplification fragments $\mathcal{F}_c$ we are able to increase the block length by several orders of magnitude, with only small increases in switch error (Fig. 3b). For example, when the phasing score threshold $c = 2.75$, we obtain a median block length $N50 = 9.3$ kb, with a corresponding switch error rate of 0.04. This indicates that we are able to phase across multiple exons. Indeed this haplotype block length is of the same order of magnitude of the typical gene length.

Although read-based phasing is limited by the length of fragments, reference-based phasing algorithm is generally able to phase over longer genomic distances. We compared the haplotype assemblies we obtain to a reference-based phasing algorithm, EAGLE2 (Loh *et al.*, 2016) (Supplementary Fig. S1). We show that while EAGLE2 was able to provide a phasing over the whole chromosome, our amplification-fragment based phasing provides lower error rates within the blocks that we obtain.

To investigate the number of single-cells required to obtain accurate amplification fragments, we ran the model on subsets of cells of size 2–15 (Fig. 3c). We find that on this data set, the fragment accuracy levels off after 8–10 cells. Thus, we can obtain accurate amplification fragments even with relatively few cells.

4 Discussion

Single-cell DNA sequencing is increasingly being used to explore the genomic content of individual cells, but requires analysis algorithms that are robust to the errors and biases in this data. In this article, we exploit one bias in single-cell-sequencing data, amplification bias, and show how we can leverage this local information to assemble haplotypes.

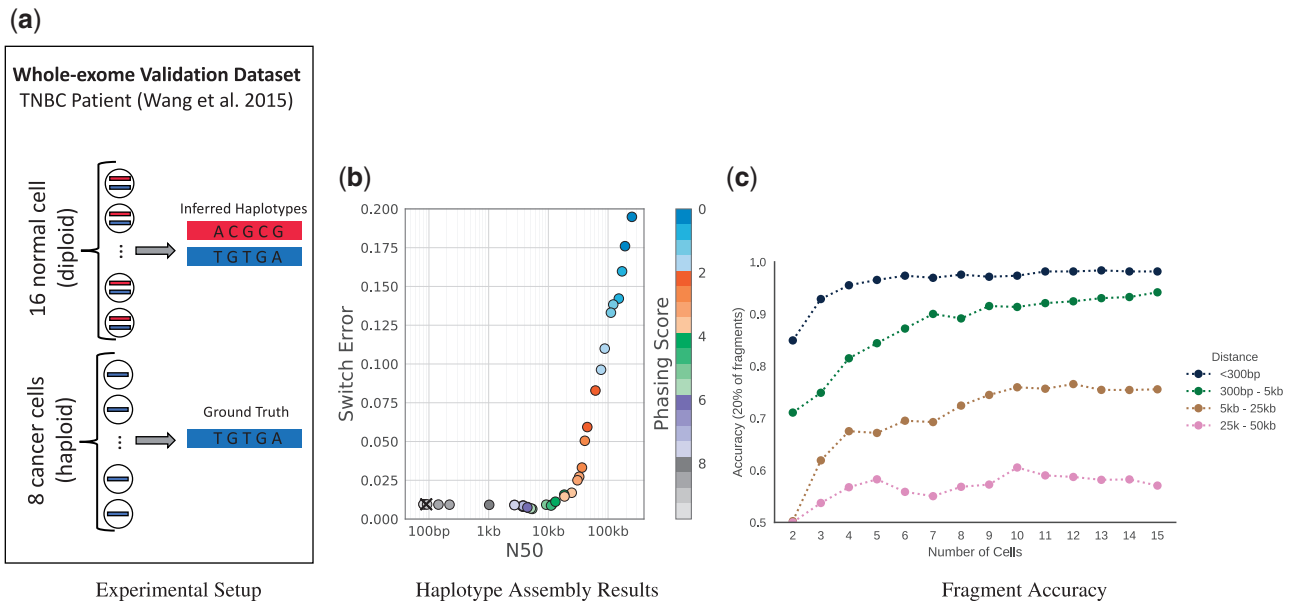

Fig. 3. Assembling haplotypes on whole-exome DNA-sequencing data. **(a)** We validate haplotype assemblies on whole-exome DNA-sequencing data from an individual breast cancer patient (Wang *et al.*, 2014), by comparing to the haplotype of Chromosome 17, whose haplotype we can determine from the eight cancer cells that have lost one homolog of this chromosome. **(b)** Haplotype block length (*N*50) as a function of haplotype switch error for varying threshold of phasing score. The *N*50 and switch error for the assembly with no amplification fragments is marked with an '×'. Amplification fragments increase the length of haplotype assemblies by orders of magnitude with small increase in switch error. **(c)** The accuracy of the phasing for the highest scoring 20% of amplification fragments for varying numbers of cells

Our results demonstrate that concurrent dropout between nearby alleles can provide amplicon-scale correlations that lead to better haplotype assemblies than using only correlations between alleles on the same sequence read. However, there are several limitations and avenues for further improvement of our approach. First, many recent haplotype assembly algorithms, including the HapCut2 (Edge *et al.*, 2017) used here, employ more sophisticated probabilistic models for error in fragments. Extending our model to estimate error rates for amplification fragments could provide better integration with the error models in haplotype assemblers, and yield more accurate haplotype predictions. Although our current model does not calculate likelihood for each phasing, we may be able to estimate the error rates based on the empirical distribution of phasing scores. Alternatively, an improved probability model that accounts for several of the features of sequencing data—including sequencing error and observed read depth—could be developed and would likely outperform the straightforward model of concurrent dropout introduced here. Additionally, extending the model to consider groups of SNPs instead of just pairs of SNPs may be useful for identifying phase when pairwise relationships are weak. This can be particularly useful with the lower sequence coverages that are common in single-cell sequencing datasets.

If one's goal is to obtain a high-quality phased diploid genomes, there are a number of good approaches. These include: the application of high-quality reference-based phasing algorithms (Browning and Yu, 2009; Delaneau *et al.*, 2013; Loh *et al.*, 2016; Stephens *et al.*, 2001) that exploit large populations on genotyped individuals; long-read (Glusman *et al.*, 2014) or linked-read (Zheng *et al.*, 2016) sequencing; or specialized techniques such as Strand-Seq (Porubský *et al.*, 2016). Phasing algorithms also exist for both bulk (Castel *et al.*, 2016) and single-cell RNA-seq data (Castel *et al.*, 2016). We do not expect that researchers will perform single-cell sequencing if their only goal is to obtain a phased, diploid genome.

Rather, we anticipate that the approach described here will be a useful complement for specific analyses of single-cell sequencing data. For example, we expect that this model can be broadly useful in variant calling in single cells, which is typically significantly confounded by amplification bias. Although we validated the model on diploid genomes, the model is readily adaptable to copy-number aberrations, and thus can be applied to cancer genomes which often demonstrate high levels of aneuploidy. The information derived from our model may be useful for allele-specific copy number calling in single cells, which to our knowledge is not currently done in any existing single-cell copy-number caller. Information on haplotype phase will also be useful for calling retrotranspon insertions (Evrony *et al.*, 2012) or single-nucleotide variants (SNVs). Recently, (Bohrson *et al.*, 2018) showed how they were able to reduce false positive rates in SNV calling in single-cell DNA sequencing by phasing SNVs to nearby SNPs. However, many SNVs cannot be phased to SNPs with short reads, especially in whole-exome data. Our method is able to phase across much larger distances and across exons. An additional application is phasing of structural variants in single-cells. SNPs on either side of the breakpoints of a structural variant will also show more dropout concurrence than expected by chance. An extension of our model might allow for improved phasing of structural variants, which could be useful in reconstructing highly rearranged cancer genomes. Finally, additional extensions of dropout concurrence might be applied to single-cell RNA-seq data, e.g. by exploiting correlations in allele-specific expression or allele-specific alternative splicing.

Funding

This work was supported by a US National Science Foundation (NSF) CAREER Award [CCF-1053753] and US National Institutes of Health (NIH) grants [R01HG007069 and R01CA180776 to B.J.R.].

Conflict of Interest: B.J.R. is a co-founder and consultant at Medley Genomics.

References

Bakker,B. *et al*. (2016) Single-cell sequencing reveals karyotype heterogeneity in murine and human malignancies. *Genome Biol*., **17**, 115.

Bohrson,C.L. *et al*. (2018) Linked-read analysis identifies mutations in single-cell dna sequencing data. *Nucleic Acids Res*., **46**, e20.

Browning,B.L. and Yu,Z. (2009) Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. *Am. J. Hum. Genet*., **85**, 847–861.

Castel,S.E. *et al*. (2016) Rare variant phasing and haplotypic expression from rna sequencing with phaser. *Nat. Commun*., **7**, 12817.

Chen,C. *et al*. (2017) Single-cell whole-genome analyses by linear amplification via transposon insertion (lianti). *Science*, **356**, 189–194.

Chu,W.K. *et al*. (2017) Ultraaccurate genome sequencing and haplotyping of single human cells. *Proc. Natl. Acad. Sci. USA*, **114**, 12512–12517.

Delaneau,O. *et al*. (2013) Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods*, **10**, 5.

Edge,P. *et al*. (2017) Hapcut2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Res*., **27**, 801–812.

Evrony,G.D. *et al*. (2012) Single-neuron sequencing analysis of l1 retrotransposition and somatic mutation in the human brain. *Cell*, **151**, 483–496.

Fan,H.C. *et al*. (2011) Whole-genome molecular haplotyping of single cells. *Nat. Biotechnol*., **29**, 51–57.

Garvin,T. *et al*. (2015) Interactive analysis and assessment of single-cell copy-number variations. *Nature Methods*, **12**, 1058.

Gawad,C. *et al*. (2016) Single-cell genome sequencing: current state of the science. *Nat. Rev. Genet*., **17**, 175.

Gawad,C. *et al*. (2016) Single-cell genome sequencing: current state of the science. *Nat. Rev. Genet*., **17**, 175–188.

Glusman,G. *et al*. (2014) Whole-genome haplotyping approaches and genomic medicine. *Genome Med*., **6**, 73.

Hong,X. *et al*. (2006) Intron size, abundance, and distribution within untranslated regions of genes. *Mol. Biol. Evol*., **23**, 2392–2404.

International Human Genome Sequencing Consortium. *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860.

Leiserson,M.D. *et al*. (2016) A weighted exact test for mutually exclusive mutations in cancer. *Bioinformatics*, **32**, i736–i745.

Lodato,M.A. *et al*. (2015) Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science*, **350**, 94–98.

Loh,P.-R. *et al*. (2016) Reference-based phasing using the haplotype reference consortium panel. *Nat. Genet*., **48**, 1443.

Marchini,J. *et al*. (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet*., **39**, 906.

Marcy,Y. *et al*. (2007) Dissecting biological dark matter with single-cell genetic analysis of rare and uncultivated tm7 microbes from the human mouth. *Proc. Natl. Acad. Sci. USA*, **104**, 11889–11894.

McConnell,M.J. *et al*. (2013) Mosaic copy number variation in human neurons. *Science*, **342**, 632–637.

Navin,N.E. (2015) The first five years of single-cell cancer genomics and beyond. *Genome Res*., **25**, 1499–1507.

Patterson,M. *et al*. (2015) Whatshap: weighted haplotype assembly for future-generation sequencing reads. *J. Comput. Biol*., **22**, 498–509.

Picher,A.J. *et al*. (2016) Trueprime is a novel method for whole-genome amplification from single cells based on tthprimpol. *Nat. Commun*., **7**, 13296.

Pirola,Y. *et al*. (2016) Hapcol: accurate and memory-efficient haplotype assembly from long reads. *Bioinformatics*, **32**, 1610–1617.

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

Roach,J.C. *et al*. (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science*, **328**, 636–639.

Sherman,M.A. *et al*. (2017) Pasd-qc: quality control for single cell whole-genome sequencing data using power spectral density estimation. *Nucleic Acids Res*.,

Stephens,M. *et al*. (2001) A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet*., **68**, 978–989.

Tewhey,R. *et al*. (2011) The importance of phase information for human genomics. *Nat. Rev. Genet*., **12**, 215.

van de Ven,M. *et al*. (2012) Effects of compound heterozygosity at the xpd locus on cancer and ageing in mouse models. *DNA Repair*, **11**, 874–883.

Wang,Y. *et al*. (2014) Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature*, **512**, 155–160.

Zhang,C.-Z. *et al*. (2015) Calibrating genomic and allelic coverage bias in single-cell sequencing. *Nat. Commun*., **6**, 6822.

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