A Supplementary Methods

A.1 Creating Synthetic Diploid Simulations

To create artificial haplotypes, we used variant positions from the set of common human variants from dbSNP build 150. We restrict to the set of single-nucleotide variants. dbSNP reports the population allele frequencies at each position, \([p_A, p_C, p_G, p_T]\). We generate two haplotypes \(H, \bar{H}\) for each individual such that 
\[H = [h_1, h_2, \ldots, h_n], \quad \bar{H} = [\bar{h}_1, \bar{h}_2, \ldots, \bar{h}_n],\]
where \(h_i \in \{A, C, G, T\}\), and 
\[\Pr(h_i = x) = p_x.\]
Note that this approach does not maintain linkage disequilibrium between adjacent SNPs, but the algorithms tested on this data set do not rely on population level information.

Single cells were downloaded from SRA accession numbers SRP041470 and SRP061939. For the simulation we used 16 cells from patient UMB1465. All cells were downsampled to a target coverage of 3x using samtools. Cells were separated in to pairs \(c, \bar{c}\) based on the total number of positions in chromosome X that were covered by at least one read, where cell \(c\) corresponds to haplotype \(H\) and cell \(\bar{c}\) corresponds to haplotype \(\bar{H}\). Two cells were excluded due to irregular drop-out rates. This yields 7 pairs of cells. From a cell \(c\), we extract the set of reads that cover at least one position \(i\) in the generated haplotypes \(H, \bar{H}\). For each read \(r\) covering position \(i\), we alter the nucleotide \(r_i\) mapped to position \(i\) as
\[r_i = \begin{cases} h_i & \text{with probability } p_{r,i} \\ \{x \in \{A, C, G, T\} | x \neq h_i\} & \text{with probability } (1 - p_{r,i}/3) \end{cases}\]
with \(p_{r,i}\) is the probability of an incorrect based call implied by the reported Phred score 
\[q_{r,i} = 10 \log_{10} p_{r,i}.\]
These from cells \(c, \bar{c}\) were merged into a single BAM file.

B Supplementary Results

B.1 Comparison to Reference-Based Phasing

An alternate approach to determining the haplotype phasing of an individuals genome is statistical or reference-based phasing with a reference panel [3, 2, 4, 1]. These methods take advantage of linkage disequilibrium, the correlation between the occurrence of certain alleles at nearby variants across a population. They take as input genotype data from of one or more individuals as well as a panel of phased genomes from a larger population. Here we compare against the performance of EAGLE2 [3], the current state-of-the-art reference-based phasing approach on the single-cell breast cancer data from Wang et al. [5].

We used the Sanger Imputation Server which integrates the Eagle2 software, to perform phasing, using the 1000 Genomes Phase 3 reference panel. As the service was unable to run on the full set of SNPs from chromosome 17, we created an input file which included all heterozygous variants identified in the sample, in addition to all homozygous variants in
which the called allele had a population frequency of less than 0.95. This yielded a set of
119,990 SNPs.

On this dataset, we found that EAGLE2 had a total switch error of 17.3% over the set
of heterozygous variants. As we do not obtain a phasing over the entire chromosome, we
measured We evaluate the switch error of EAGLE2 on the same set of haplotype blocks we
recover using amplification fragments for a range of phasing scores (see Figure 3 in the main
text). Figure S1 shows the resulting switch errors. Note that the EAGLE2 switch error
decreases with the cutoff used for the phasing score. While EAGLE2 does not have access to
read information used to calculate the phasing score, we expect both the phasing score and
the performance of EAGLE (based on linkage disequilibrium) to correlate with the distance
between SNPs in the block. We see that for scores >= 2, using amplification fragments
with HAPCUT achieves lower switch-error rates than EAGLE2. These results suggest the
potential utility of a hybrid approach, combining the accuracy of phasing from amplification
fragments at short distances, with the ability of reference-based phasing to span pairs of
SNPs beyond amplicon length.

B.2 Effect of Number of Cells on Fragment Accuracy

To evaluate the number of cells needed to obtain accurate results for pairwise phasing,
we ran the model on subsets of cells from the whole-exome dataset [5]. We use the same
experimental setup as in Section 3.2, and randomly subsample n cells for \( n \in [2, 15] \). As a
comparison, we measure the accuracy of the reported phasing for the top 20% of SNP pairs.
Figure 2 shows the results of this analysis.

References

[1] B. L. Browning and Z. Yu. Simultaneous genotype calling and haplotype phasing im-
proves genotype accuracy and reduces false-positive associations for genome-wide asso-


Figure 1: **Comparison to EAGLE2 on whole-exome data.** We evaluate the switch error obtained by EAGLE2 over the set of haplotype blocks recovered using amplification fragments, on the single-cell breast cancer data from Wang et al. [5]. The total switch error rate across all the entire chromosome is shown as the red line.
Figure 2: The number of cells affects the accuracy of the phasing model. We show the accuracy over the high scoring 20% of SNP pairs, for different numbers of input cells.