Inference of Chromosome-Length Haplotypes Using Genomic Data of Three or a Few More Single Gametes

Ruidong Li,[†,1,2] Han Qu,[†,1] Jinfeng Chen,[1] Shibo Wang,[1] John M. Chater,[1] Le Zhang,[2] Julong Wei,[1,3] Yuan-Ming Zhang,[4] Chenwu Xu,[5] Wei-De Zhong,[6] Jianguo Zhu,[7] Jianming Lu,[1,6] Yuanfa Feng,[1,6] Weiming Chen,[7] Renyuan Ma,[1,8] Sergio Pietro Ferrante,[1] Mikeal L. Roose,*[1,2] and Zhenyu Jia*[1,2]

[1]Department of Botany and Plant Sciences, University of California, Riverside, Riverside, CA

[2]Graduate Program in Genetics, Genomics, and Bioinformatics, University of California, Riverside, Riverside, CA

[3]Center for Molecular Medicine and Genetics, Wayne State University, Detroit, MI

[4]Statistical Genomics Lab, College of Plant Science and Technology, Huazhong Agricultural University, Wuhan, China

[5]Jiangsu Provincial Key Laboratory of Crop Genetics and Physiology, Co-Innovation Center for Modern Production Technology of Grain Crops, Key Laboratory of Plant Functional Genomics of Ministry of Education, Yangzhou University, Yangzhou, China

[6]Department of Urology, Guangdong Key Laboratory of Clinical Molecular Medicine and Diagnostics, Guangzhou First People's Hospital, School of Medicine, South China University of Technology, Guangzhou, China

[7]Department of Urology, Guizhou Provincial People's Hospital, Guizhou, China

[8]Department of Mathematics, Bowdoin College, Brunswick, ME

[†]These authors contributed equally to this work.

***Corresponding authors:** E-mails: arthur.jia@ucr.edu; mikeal.roose@ucr.edu.

Associate editor: Rebekah Rogers

Abstract

Compared with genomic data of individual markers, haplotype data provide higher resolution for DNA variants, advancing our knowledge in genetics and evolution. Although many computational and experimental phasing methods have been developed for analyzing diploid genomes, it remains challenging to reconstruct chromosome-scale haplotypes at low cost, which constrains the utility of this valuable genetic resource. Gamete cells, the natural packaging of haploid complements, are ideal materials for phasing entire chromosomes because the majority of the haplotypic allele combinations has been preserved. Therefore, compared with the current diploid-based phasing methods, using haploid genomic data of single gametes may substantially reduce the complexity in inferring the donor's chromosomal haplotypes. In this study, we developed the first easy-to-use R package, *Hapi*, for inferring chromosome-length haplotypes of individual diploid genomes with only a few gametes. *Hapi* outperformed other phasing methods when analyzing both simulated and real single gamete cell sequencing data sets. The results also suggested that chromosome-scale haplotypes may be inferred by using as few as three gametes, which has pushed the boundary to its possible limit. The single gamete cell sequencing technology allied with the cost-effective *Hapi* method will make large-scale haplotype-based genetic studies feasible and affordable, promoting the use of haplotype data in a wide range of research.

***Key words:* chromosome-length haplotype, gamete, evolutionary genetics, quantitative genetics, recombination.**

Introduction

A haplotype in a diploid individual is a set of DNA variants on a chromosome that are coinherited from a parent. Knowledge of haplotypes is essential in many research areas, including evolutionary genetics and quantitative genetics. For example, haplotype data have been applied to imputation of unobserved low-frequency and rare variants (Huang et al. 2015; McCarthy et al. 2016), determination of parental origins of genetic variants (Kong et al. 2009; Goldmann et al. 2016), characterization of DNA–phenotype associations (Trégouët et al. 2009; Lambert et al. 2013; Xue et al. 2016), identification of recombination hotspots (Coop et al. 2008), detection of selection signatures (Sabeti et al. 2002; International HapMap Consortium 2005; Pendleton et al. 2018), and inference of genetic admixture, introgression, and demographic history in a population (Lohmueller et al. 2009; Palamara et al. 2012). Mounting studies have indicated that using haplotype variants rather than single nucleotide polymorphisms (SNPs) may dramatically improve the power for detection of the signatures of positive selection (Fariello et al. 2013).

Moreover, long-range haplotypes, which provide higher DNA resolution than short-range haplotypes or individual SNPs, have been demonstrated to be very useful for deducing genetic admixture, introgression, and demographic history (Palamara et al. 2012; Harris and Nielsen 2013; Schiffels and Durbin 2014; Snyder et al. 2015; Leitwein et al. 2020). Despite these advantages of using haplotype data, the utility of this genetic resource is still quite constrained due to the lack of a cost-effective method for phasing individual genomes, especially for the inference of high-quality chromosome-length haplotypes.

Phasing, or haplotyping, is the process of inferring haplotype structure based on genotypic data. The most widely used haplotyping strategy is to phase common genetic variants using population data (Stephens et al. 2001; Stephens and Scheet 2005; Scheet and Stephens 2006; Browning and Browning 2007; Howie et al. 2009; Li et al. 2010; Loh et al. 2016; O'Connell et al. 2016); however, this approach is incapable of phasing de novo mutations, rare variants, or structural variants and is limited to inferring short-range haplotype fragments (Glusman et al. 2014). Experimental whole-chromosome phasing approaches usually involve the physical separation of homologous chromosomes in diploid cells using chromosome microdissection, FACS-mediated chromosome sorting, or microfluidics, followed by single-chromosome sequencing (Ma et al. 2010; Fan et al. 2011; Yang et al. 2011). Nevertheless, these approaches usually require specialized and expensive equipment. Numerous sequencing technologies, including fosmid-based dilution pool sequencing, long fragment read technology, PacBio single molecule real-time long-read sequencing, 10X Genomics linked-read sequencing, and proximity ligation (Hi-C) sequencing can also be employed to generate long-range haplotype fragments (Kitzman et al. 2011; Peters et al. 2012; Selvaraj et al. 2013; Edge et al. 2017), but phasing haplotypes that span entire chromosomes is still arduous for these approaches. A recent single-cell DNA template strand-based technique, called *Strand-seq*, sequences either the Watson strand or the Crick strand of a chromosome in a somatic cell and then uses pooled libraries to phase chromosomal haplotypes (Porubský et al. 2016; Porubsky et al. 2017). Generally, phasing complete chromosomes with these sequencing technologies is expensive, making large-scale haplotype-based research infeasible. There is a high demand for innovative methods which can phase entire chromosomes for individual genomes in a cost-effective manner.

Gametes, including pollen grains in plants or sperm and eggs in animals, are the natural packaging of haploid complements that are formed during meiosis. Compared with the current phasing approaches that analyze diploid materials, using haploid genomic data of single gametes substantially reduces the complexity in inferring the donor's chromosomal haplotypes. To infer the chromosome-scale haplotypes with gametes, the objective simply becomes the identification of recombination events which are rare with an average of <3 events affecting most gametic chromosomes (Beye et al. 2006). Current development of gamete-based phasing methodologies is still at an early stage, requiring either a large number of gametes or manual inspection to ensure phasing accuracy (Lu et al. 2012; Hou et al. 2013; Kirkness et al. 2013; Hinch et al. 2019). No easy-to-use software is available for phasing chromosome-length haplotypes with gametes. To fill this void, we developed an innovative methodology, named *Hapi* (haplotyping with imperfect genotype data), for automatic inference of an individual's chromosomal haplotypes using a few gamete cells, given the heterozygous loci on the chromosome are known. Comprehensive comparisons, involving the use of a simulated data set, a maize microspore sequencing data set, and a human sperm sequencing data set, demonstrated that *Hapi* outperformed the only haploid-based algorithm, *PHMM* (pairwise hidden Markov model [HMM]) (Hou et al. 2013), and two commonly applied diploid-based phasing methods, *WhatsHap* (Martin et al. 2016) and *HapCUT2* (Edge et al. 2017) in terms of accuracy, reliability, completeness, and cost-effectiveness. The results also suggested that chromosomal haplotypes may be inferred by using only three gamete cells if the genotype data are of high quality. The rapid advancement of biotechnologies will substantially reduce the experimental costs in isolation, lysis, and whole-genome amplification of single gamete cells, which if allied with the new *Hapi* method will make large-scale haplotype-based studies affordable and feasible. In addition, the crossover analysis module in the *Hapi* R package may be employed to investigate meiotic recombination events on gamete chromosomes to disclose recombination hotspots in a target population.

New Approaches

We developed an innovative *Hapi* methodology to infer chromosomal haplotypes of individual diploid genomes using three or a few more single gametes, which has pushed this boundary to its possible limit. Implementing the *Hapi* algorithm to phase an entire chromosome consists of three steps: 1) data preprocessing, 2) inference of draft haplotypes, and 3) assembly of high-resolution chromosomal haplotypes (fig. 1). In step (1), markers with potential genotyping errors in any gamete cells are filtered out by iteratively analyzing gamete pairs via an HMM. A subset of markers, which have been successfully genotyped in at least three gametes, are selected to form a "precursor" framework. In the framework, missing data in each gamete are iteratively imputed using data available in other gametes. The markers, usually of a small number, with missing data that cannot be fully resolved by imputation are eliminated, resulting in the final framework for building draft haplotypes. In step (2), the draft haplotypes are derived by sequentially analyzing two neighboring markers in the framework with majority voting, through which the phase for any two adjacent framework markers is determined by the majority (or most frequent) link type represented in the gamete cells. The maximum parsimony of recombination (MPR) principle is then adopted to proofread disputed positions of the draft haplotypes. In step (3), each gamete chromosome is compared with the draft haplotypes to identify haplotype-converting points (HCPs) to deduce gamete-specific haplotypes, with the nonframework markers being

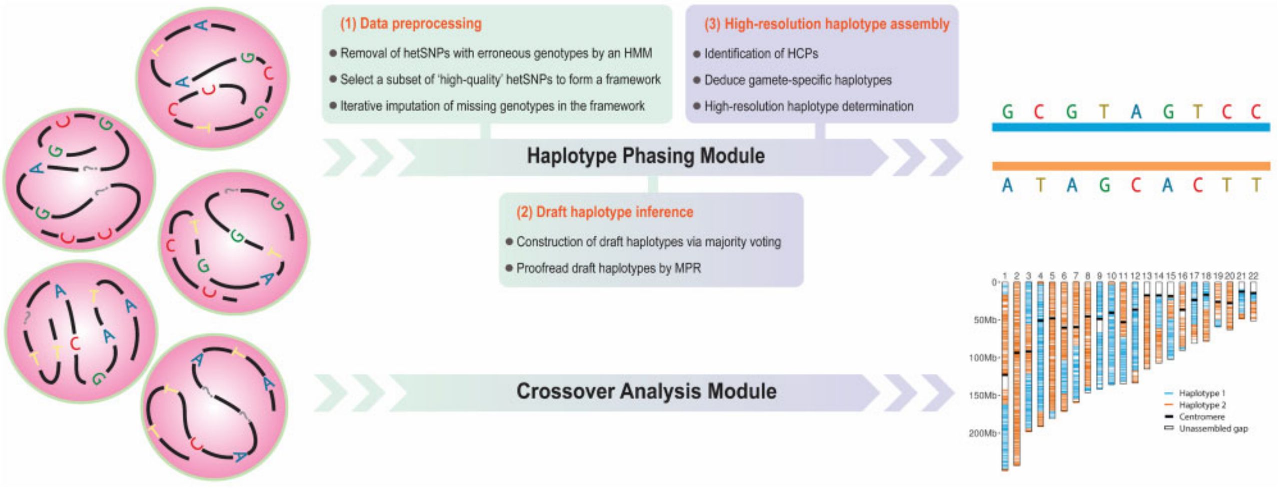

(1) Data preprocessing
● Removal of hetSNPs with erroneous genotypes by an HMM
● Select a subset of 'high-quality' hetSNPs to form a framework
● Iterative imputation of missing genotypes in the framework
(3) High-resolution haplotype assembly
● Identification of HCPs
● Deduce gamete-specific haplotypes
● High-resolution haplotype determination
Haplotype Phasing Module
(2) Draft haplotype inference
● Construction of draft haplotypes via majority voting
● Proofread draft haplotypes by MPR
Crossover Analysis Module
G C G T A G T C C
A T A G C A C T T
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
0
50Mb
100Mb
150Mb
200Mb
Haplotype 1
Haplotype 2
Centromere
Unassembled gap

Fig. 1. Overview of the *Hapi* package. *Hapi* consists of two modules: the Haplotype Phasing Module and the Crossover Analysis Module. In the Haplotype Phasing Module, three main steps are required for haplotype phasing using genomic data of single gamete cells: 1) data preprocessing; 2) draft haplotype inference; 3) high-resolution haplotype assembly. Crossovers in each gamete cell can be identified and recombination-associated analysis can be performed by the Crossover Analysis Module.

phased in this step. Consensus high-resolution haplotypes are eventually determined by these gamete-specific haplotypes through voting. An easy-to-use R package has been developed for implementing the *Hapi* algorithm to infer chromosome-length haplotypes using single gamete cells. The package also includes a crossover analysis module, allowing for downstream analyses and visualization of crossover positions identified in each gamete.

Results

Comparison of Phasing Methods by Analyzing the Human Sperm Data Set

A human sperm sequencing data set consisting of 11 independent sperm cells from the donor of the HuRef diploid genome sequence (Kirkness et al. 2013) was used to compare the phasing performance of the two gamete-based phasing methods (*Hapi* and *PHMM*) and the two read-based phasing methods (*WhatsHap* and *HapCUT2*). Although the true chromosomal haplotypes for this donor were unknown, a "phased" genome consisting of 1.82 million hetSNPs had been suggested based on a joint analysis of these 11 sperm cells sequenced at 1.5–3.7× coverage and 16 additional sperms genotyped using the Illumina HumanOmni-Quad v1.0 BeadChip (array data not publicly available) (Kirkness et al. 2013) which was adopted here as "ground truth" to evaluate the phasing performance of the four competing methods. Variant calling was conducted with the sequencing reads to yield 1.66 million high-quality hetSNPs (out of the total of 1.82 million SNPs), each of which was present in at least one sperm. The number of hetSNPs on 22 autosomes ranged from 15,340 (Chr22) to 141,669 (Chr2), and the rate of missing genotype data ranged from 70.95% to 86.49% (supplementary table S1, Supplementary Material online). The 11 sperm cells were sorted based on the rate of missing hetSNP data in descending order, that is, the first sperm cell has the most missing SNP data and so forth.

Four quality metrics, including completeness (COM), largest haplotype segment (LHS), switch error rate (SER, the fraction of incorrectly inferred phase connections), and hamming error rate (HER, the fraction of incorrectly phased hetSNPs), were used to evaluate the phasing performance for different phasing methods (see Materials and Methods for details).

Various numbers of gametes, 3 through 11 from the sorted sperm list, were successively used for haplotyping by four different methods, respectively, to compare phasing completeness and accuracy at the whole-genome scale (fig. 2; supplementary table S2, Supplementary Material online). The phasing completeness for the two gamete-based methods was steadily and evidently greater than that for the two read-based methods (fig. 2). Chromosome-length haplotypes for the 22 autosomes can be successfully inferred with 99.947% of hetSNPs being phased on the LHS using the gamete-based methods; in comparison, hundreds of thousands of small haplotype segments were deduced with only 3.223% of hetSNPs being phased on the LHS by the read-based methods even when 11 sperms were used (supplementary table S2, Supplementary Material online). These results indicated that the latter two methods only inferred haplotype segments whereas the former two methods were suitable for phasing entire chromosomes. *Hapi* consistently had lower SER and HER at the whole-genome scale than *PHMM* when different numbers of gametes were analyzed (fig. 2). Although the ways we calculate SER and HER were biased toward read-based methods (see Materials and Methods for details), *Hapi* was still superior to *WhatsHap* and *HapCUT2* in terms of accuracy. The running time for each method to phase the 22 autosomes using 3 or 11 sperms has been summarized in supplementary table S3, Supplementary Material online.

The phasing performance of these four methods was further compared at the chromosomal scale. When only the first three sperms were used, the completeness levels of *Hapi* and *PHMM* were much higher than those for *WhatsHap* and

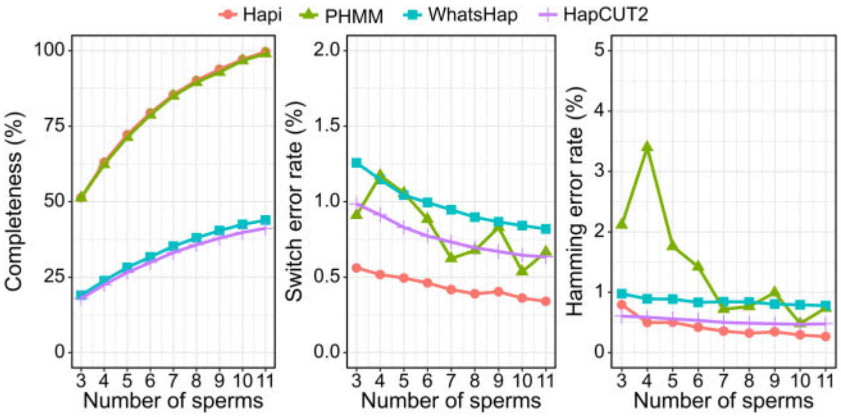


Fig. 2. Comparison of the two gamete-based phasing methods (*Hapi* and *PHMM*), and the two read-based methods (*WhatsHap* and *HapCUT2*) in terms of COM, SER, and HER at the genome scale using the human sperm cell sequencing data set. Various numbers of gametes, 3 through 11 from the sorted sperm list (based on the rate of missing data of hetSNPs in a descending order, that is, the first sperm cell has the most missing SNP data), were successively used for haplotyping analysis by the four different methods, respectively.

HapCUT2 across 22 autosomes (consistent with the results of comparison at the whole-genome scale shown in fig. 2), and the two read-based phasing methods could only generate small haplotype segments (fig. 3, inner circles). Therefore, in the subsequent analysis of phasing accuracy, we mainly focused on the comparison between the two gamete-based phasing methods. As the chromosomal haplotypes suggested in the original article (Kirkness et al. 2013) may be subject to errors, we defined a successful phasing of a chromosome as having over 95% of phased markers on that chromosome were in agreement with the suggested haplotypes. The results showed that *Hapi* can correctly phase all 22 autosomes with three sperm cells, whereas *PHMM* required at least seven sperm cells to achieve the same level of accuracy. When seven or fewer sperm cells were used, *Hapi* performed consistently well but the performance of *PHMM* fluctuated wildly, indicating *Hapi* provided more reliable phasing results with small samples. Interestingly, *PHMM* can correctly infer the haplotypes of Chr1 with 6–10 gametes but failed when all 11 sperms had been used. Out of a total of 198 scenarios (22 chromosomes × 9 numbers of gametes) for the analysis by *Hapi*, 164 scenarios (82%) achieved phasing accuracies of 99% or greater. The majority of scenarios with phasing accuracies between 95% and 99% was for the analyses of Chr15, Chr16, and Chr21, which also appeared to be challenging to *PHMM*, suggesting a complication in the genomic data for these three chromosomes. Overall, among the 1.66 million hetSNPs phased by *Hapi* using all the 11 sperms, 99.73% (1,658,197/1,662,611) of them were concordant with the chromosomal haplotypes suggested in the original paper (Kirkness et al. 2013). An inspection of the nonconcordant hetSNPs showed that 49.1% of them were only supported by one sperm cell and 33.4% of them had discordancy among two or more supporting sperm cells. The disputably phased hetSNPs tended to cluster around the centromere or at either end of the chromosomes (supplementary fig. S1, Supplementary Material online). The hetSNPs that were not in agreement between *Hapi* and the suggested haplotypes on Chr15 were evenly distributed along the chromosome, which might be ascribed to a complication in data of sperm Y47 that was contaminated by DNA from other lysed cells as mentioned in the original article (Kirkness et al. 2013).

Comparison of Phasing Methods in the Maize Microspore Data Set

A maize microspore sequencing data set from F1 hybrid individuals of a cross between two inbred lines (Li et al. 2015) was used to further evaluate the performance of *Hapi* versus *PHMM*. This is an ideal validation data set because the parental haplotypes were known. To avoid using microspores from the same meiosis event, one microspore from each of the 24 tetrads was randomly selected to form a 24-gamete pool. The number of hetSNPs on the maize chromosomes ranged from 42,691 (Chr10) to 82,689 (Chr1). The average rate of missing genotype data for ten chromosomes across the 24 selected gametes was about 50%, with the maximum missing rate equal to 72.46% (supplementary table S4, Supplementary Material online). For each of ten maize chromosomes, the 24 selected gametes were sorted in a similar way as we did for human sperm data. Various numbers (3–15) of gametes from the sorted list were sequentially analyzed with *Hapi* and *PHMM*, to infer the complete haplotypes for that chromosome. This process was repeated to phase all ten chromosomes, yielding a total of 260 scenarios (13 numbers of gametes × 10 chromosomes × 2 methods). In each scenario, the phased chromosome was compared with the known parental haplotypes to calculate phasing accuracy.

At the whole-genome scale, the two methods had the same completeness but *Hapi* generally outperformed *PHMM* in terms of accuracies, especially when few gametes were used in the phasing analysis (fig. 4A; supplementary table S5, Supplementary Material online). The comparison of the phasing results between *Hapi* and *PHMM* at the chromosome scale indicated that *Hapi* consistently had lower HER than *PHMM*. The haplotypes inferred by *Hapi* had HER < 1% in almost all the scenarios, except for Chr2

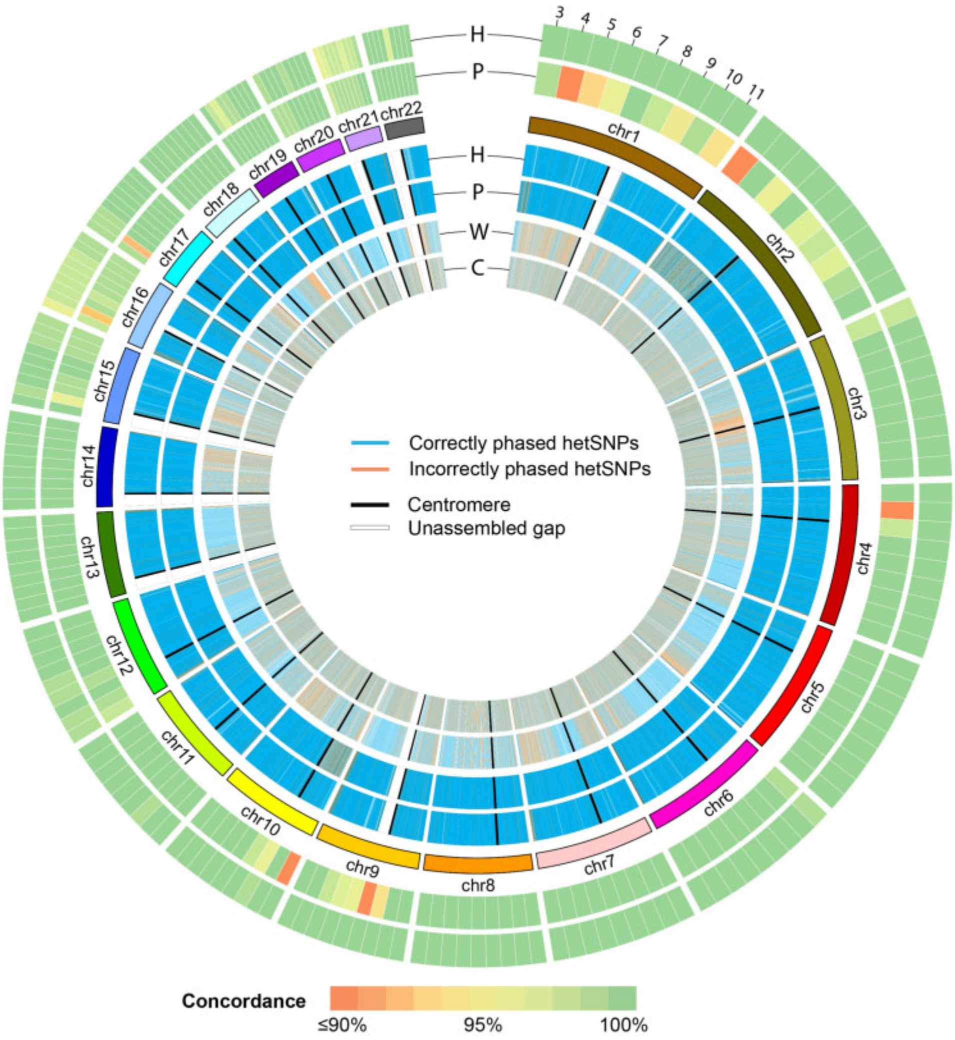


Fig. 3. Circos plot visualizing the comparison of four phasing methods, that is, H (*Hapi*), P (*PHMM*), W (*WhatsHap*), and C (*HapCUT2*) for phasing the 22 autosomes in the human sperm cell sequencing data set. The four inner circles show the phasing results when three sperm were used, with blue/orange representing the correctly/incorrectly phased hetSNPs. Only the phased hetSNPs are shown and density of the hetSNPs indicates the completeness of phasing for each of the four phasing methods. The two outer circles show the phasing accuracies based on HER for two gamete-based methods when 3 through 11 sperms were used for haplotyping.

when three gametes were analyzed (fig. 4*B*). A close look at Chr2 of these three gametes disclosed two crossovers on two gamete chromosomes in a small region (39 hetSNPs in between) near one end of the chromosome. In the default setting of *Hapi*, any small block (<100 hetSNPs) delimited by two crossovers from the draft haplotypes will be excised, prior to implementation of MPR, to construct a reliable draft haplotype; thus, in some cases, the phase of the two merging framework markers may be incorrectly inferred by misinterpreting the link types in between due to the removal of this block. The results showed that Chr2 was also challenging for *PHMM*. Moreover, at least seven gametes were required for *PHMM* to achieve the same phasing accuracies (especially for HER) across all of the ten chromosomes. When a small number of samples (<7 gametes) were analyzed, the phasing performance for *PHMM* fluctuated and did not monotonically increase as the number of gametes increased, suggesting that *PHMM* is not suitable for handling small samples.

Comparison of Phasing Methods in a Simulated Data Set

We carried out a comprehensive simulation study to further benchmark the *Hapi* algorithm for haplotype phasing. Three factors that may affect phasing accuracy and completeness were considered in each scenario, that is, 1) the number of hetSNPs on the chromosome, 2) the number of gametes, and 3) the rate of missing genotype data. As phasing one chromosome is independent of phasing another chromosome, we only considered a single chromosome in the simulated study where a pool of 100 haploid gametes were generated from a diploid donor. The number of hetSNPs on the chromosome ranged from 5,000 (or 5K) to 100,000 (or 100K). Three to 15 gametes, each with one to three crossovers generated on the chromosome, were arbitrarily selected from the 100 haploid gametes without replacement. The majority of the crossovers was randomly positioned, but in some scenarios, we intentionally placed some crossovers approaching the ends of the

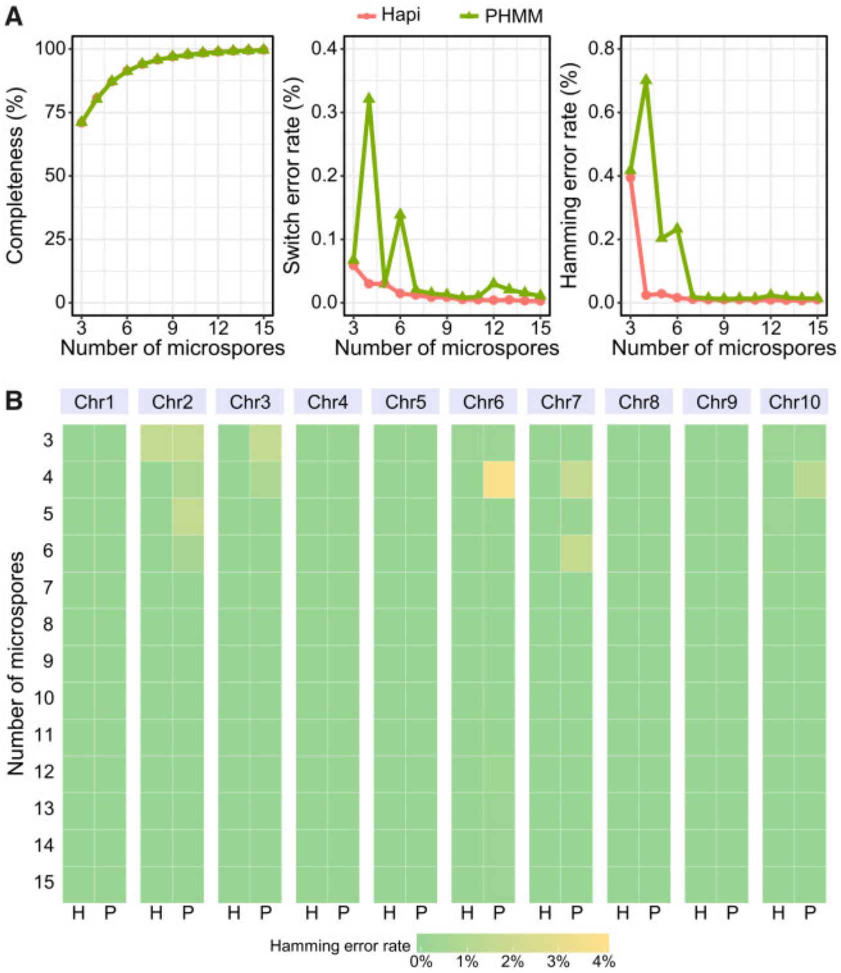


Fig. 4. Comparison of the two gamete-based phasing methods (H: *Hapi* and P: *PHMM*) in the maize microspore sequencing data set. (*A*) Comparison of *Hapi* and *PHMM* in terms of COM, SER, and HER at the genome scale. (*B*) HER of each individual chromosome for haplotype phasing with *Hapi* and *PHMM* methods when 3–15 microspores were used.

chromosomes, which are generally challenging regions to phase. We also generated a few imitative noncrossover (NCO) gene conversions (GCs), each of which possessed two seemingly apparent “crossovers” in a very small region on the gametic chromosomes to increase the complexity. Missing hetSNPs data (i.e., *NA*) ranging from 10% to 70% were randomly introduced to each simulated gamete chromosome. Moreover, 1% genotyping errors were randomly placed on the simulated gamete chromosomes. The 15 gametes were sorted using the same method that was used for the analyses of human sperm data and maize microspore data. We compared the two gamete-based methods under different scenarios with a predetermined number of gametes, number of hetSNPs, and missing genotype rate. Each scenario was repeated 100 times. A high-quality inference in a scenario was defined if more than 99% of the hetSNPs were correctly phased.

The results showed that the average performance (based on 100 replicates of each scenario) of *Hapi* and *PHMM* was similar when nine or more gametes were included in the analysis; however, *Hapi* outcompeted *PHMM* significantly in terms of SER and HER when fewer gametes were used (fig. 5).

We used a heatmap to depict the phasing repeatability or reliability of the two methods based on the 100 replicates for each scenario (fig. 6). The results indicated that the repeatability of *Hapi* steadily increased when 1) the number of hetSNPs increases, 2) the missing genotype rate decreases, or 3) more gametes were used for analysis. In contrast, the repeatability of *PHMM* did not change with the number of hetSNPs or missing genotype rate. Although *PHMM* became more repeatable when more gametes were used for phasing, the trajectory fluctuated rather than increasing monotonically. Asymptotically, *Hapi* can correctly infer chromosomal haplotypes only using three gametes if the hetSNPs were dense enough and the missing genotype rate was not too high, which did not seem to be achievable by *PHMM*.

The same simulation data set was then used to systematically benchmark *Hapi* for crossover detection on the basis of true positive rate (TPR, the proportion of actual crossovers which were correctly identified) and false discovery rate (the proportion of false crossovers). As aforementioned, we intentionally designed the simulation to include some crossovers at the ends of the chromosomes and also introduced a few mimic NCO GCs in the chromosomes. It was not surprising

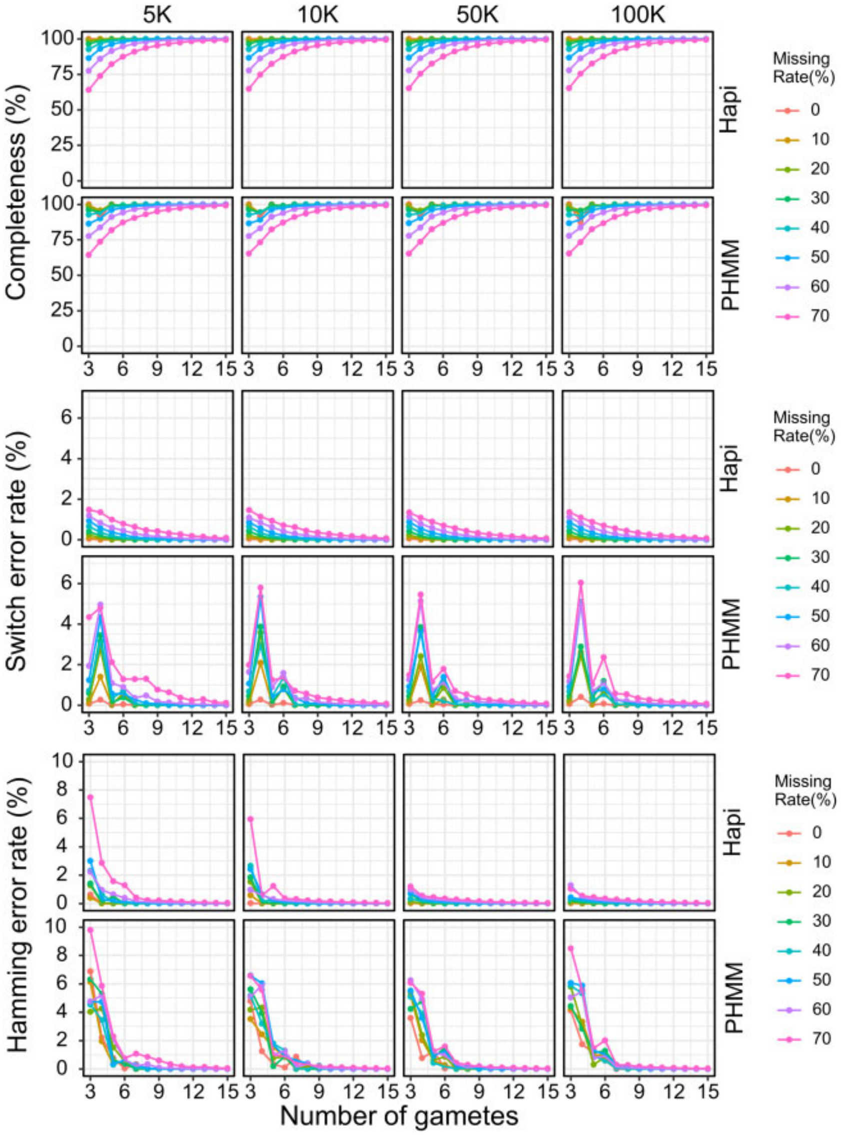
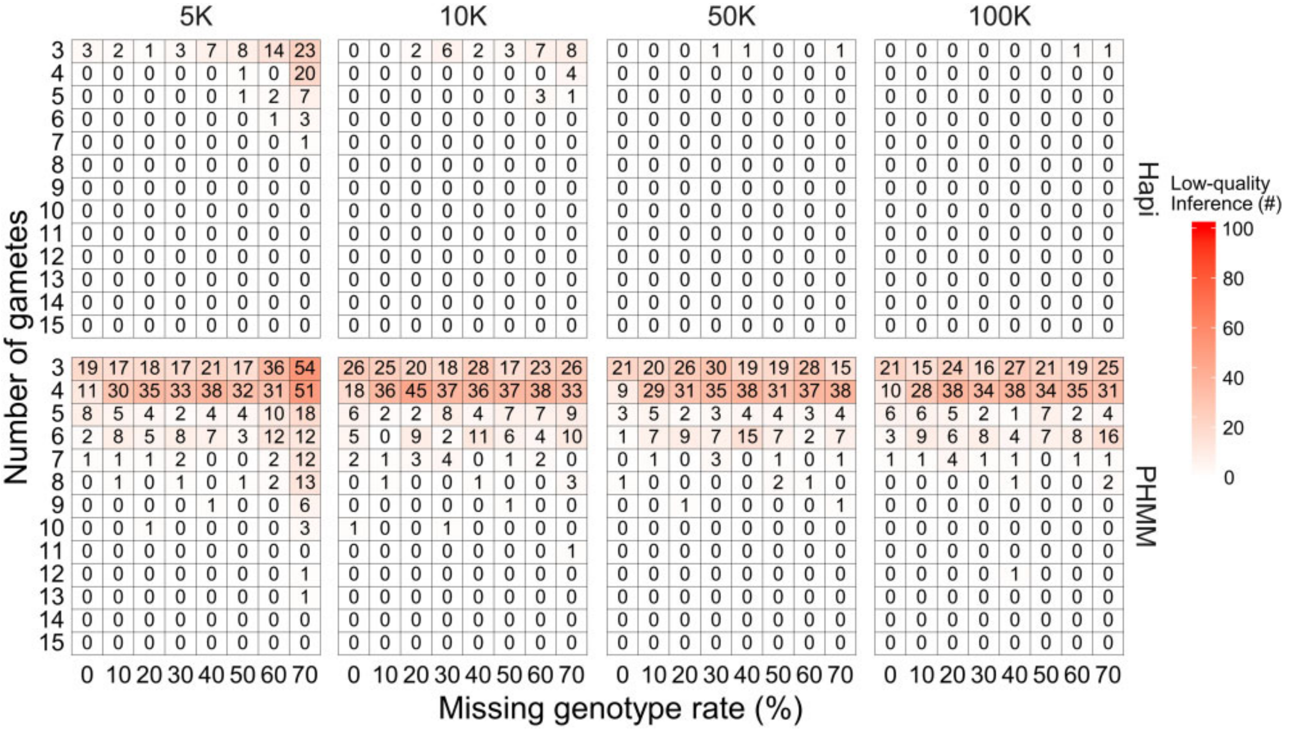


Fig. 5. Comprehensive simulation study comparing the performance of the two gamete-based phasing methods *Hapi* and *PHMM* in terms of COM, SER, and HER. A pool of 100 haploid gametes were simulated where the number of hetSNPs ranged from 5,000 (5K) to 100,000 (100K) and the rate of missing genotype data ranged from 10% to 70%. For the chromosome in each simulated gamete, one to three crossovers and 1% genotyping errors were introduced. In each comparison, 3–15 simulated gametes were randomly selected from the gamete pool for haplotyping and the process was repeated for 100 times to compute the average of COM, SER, and HER.

to see that the performance of *Hapi* and *PHMM* for crossover detection was consistent with that for phasing chromosomes because the identification of crossovers relied on the inferred haplotypes (fig. 7). Thus, imprecise haplotype phasing may lead to inaccurately identified crossovers. For both *Hapi* and *PHMM*, the TPR increased when the number of hetSNPs or the number of gametes rose, whereas TPR decreased if missing genotype rate declined. Over 99.5% of the crossovers can be accurately identified by *Hapi* when four gametes with 5,000 hetSNPs and <50% of missing data were used. With more than 50,000 hetSNPs, all the crossovers can be identified under almost all of the scenarios. The capping strategy designed in the *Hapi* phasing module ensured the accuracy of phasing of hetSNPs at either end of a chromosome and, therefore, led to a successful detection of crossovers in those challenging regions. The HMM adopted in *Hapi* recognized NCO GCs and did not erroneously call them as crossovers. Although *PHMM* also had a satisfactory level of TPR, many false crossovers were identified. The performance of *PHMM* was even worse when more hetSNPs were used. This was likely owing to the fact that a direct inference of crossover positions in the core strategy of *PHMM* is rather sensitive to regions with ambiguous data (i.e., genotyping errors, or complications caused by multiple crossovers in more than one gamete) and dense hetSNPs data would add to the intricacy. Such a problem may be resolved by increasing the number of gametes (i.e., nine or more) in the phasing analysis, which was also the case in the simulation study.



Hapi

Number of gametes	5K								10K								50K								100K							
	0	10	20	30	40	50	60	70	0	10	20	30	40	50	60	70	0	10	20	30	40	50	60	70	0	10	20	30	40	50	60	70
3	3	2	1	3	7	8	14	23	0	0	2	6	2	3	7	8	0	0	0	1	1	0	0	1	0	0	0	0	0	0	1	1
4	0	0	0	0	0	1	0	20	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	1	2	7	0	0	0	0	0	0	3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	1	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

PHMM

Number of gametes	5K								10K								50K								100K							
	0	10	20	30	40	50	60	70	0	10	20	30	40	50	60	70	0	10	20	30	40	50	60	70	0	10	20	30	40	50	60	70
3	19	17	18	17	21	17	36	54	26	25	20	18	28	17	23	26	21	20	26	30	19	19	28	15	21	15	24	16	27	21	19	25
4	11	30	35	33	38	32	31	51	18	36	45	37	36	37	38	33	9	29	31	35	38	31	37	38	10	28	38	34	38	34	35	31
5	8	5	4	2	4	4	10	18	6	2	2	8	4	7	7	9	3	5	2	3	4	4	3	4	6	6	5	2	1	7	2	4
6	2	8	5	8	7	3	12	12	5	0	9	2	11	6	4	10	1	7	9	7	15	7	2	7	3	9	6	8	4	7	8	16
7	1	1	1	2	0	0	2	12	2	1	3	4	0	1	2	0	0	1	0	3	0	1	0	1	1	1	4	1	1	0	1	1
8	0	1	0	1	0	1	2	13	0	1	0	0	1	0	0	3	1	0	0	0	0	2	1	0	0	0	0	0	1	0	0	2
9	0	0	0	0	1	0	0	6	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0
10	0	0	1	0	0	0	0	3	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
13	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Fig. 6. Heatmap visualizing the reliability and repeatability of the two gamete-based phasing methods (*Hapi* and *PHMM*) under the scenarios with different number of hetSNPs (5K–100K), different missing genotype rate (10–70%), as well as different number of gametes (3–15). The number in each cell represents the counts of low-quality phasing (HER > 1%) out of the 100 replicates in each scenario.

Recombination Analysis in the Human Sperm Data Set

With the phased chromosome-length haplotypes, an *HMM* was used to infer crossover positions in the sperm genomes by successively contrasting hetSNPs in each sperm with the inferred chromosomal haplotypes (supplementary fig. S2, Supplementary Material online). A total of 254 crossovers along the 22 autosomes were identified in the 11 sperms with an average of 1.05 per chromosome. Compared with the 260 crossovers identified in the original article (Kirkness et al. 2013), 251 were also identified by the *Hapi* method (supplementary table S6, Supplementary Material online). The 12 inconsistent crossovers were all located at the ends of chromosomes, and such inconsistency may be ascribed to either of the two following reasons: 1) The method in the original article did not accurately infer haplotypes at the chromosome ends, yielding incorrect crossovers in those regions, or 2) the observed double crossovers in a very small region were considered to be either caused by a GC event or consecutive genotyping errors and thus were filtered out by *Hapi*. The number of crossovers was counted in each bin (5 Mb in length) along 22 autosomes and distributions of the 254 crossovers are depicted in figure 8A. The resolution of crossover locations ranged from 79 bp to 788 kb with a median of 89.3 kb, which was roughly the same as the 82.5-kb resolution reported in the original article (Kirkness et al. 2013). Over 75% of the 254 crossovers were located within an interval of < 200 kb (fig. 8B). Distribution of distances between any two chromosomally adjacent crossovers was provided (fig. 8C), which can be used for recombination-relevant research such as interference in the formation of chromosomal crossovers during meiosis. Functions for downstream analysis and visualization were included in the “crossover analysis” module of the *Hapi* package.

Discussion

In the past decades, genetics and evolution studies have benefited from various types of advanced genotyping technologies that can survey genome-wide SNP variants. However, most of the studies were based on the analysis of individual SNPs, yielding limited interpretation of the genomes. Haplotypes, which represent definitively phased neighboring SNPs, provided improved resolution in terms of DNA variants for various genetics analyses. For example, The use of long-range haplotypes or microhaplotypes has been demonstrated to benefit many studies, such as increasing the accuracy for inferring kinship or population structure (Baetscher et al. 2018), and enhancing the analytical power in the detection of genetic stock (McKinney et al. 2017), in the detection of positive selection signatures (Fariello et al. 2013), or in the assessment of admixture, introgression, and demographic history in target populations (Palamara et al. 2012; Schiffels and Durbin 2014; Snyder et al. 2015; Leitwein et al. 2020). Thus, obtaining complete and accurate haplotype data at the chromosome scale will provide tremendous potential to advance various types of genetic research.

The current knowledge of haplotypes is often fragmented or even biased due to the limitations of the existing phasing methods which are mostly based on the analysis of diploid materials. Moreover, most of these diploid-based phasing

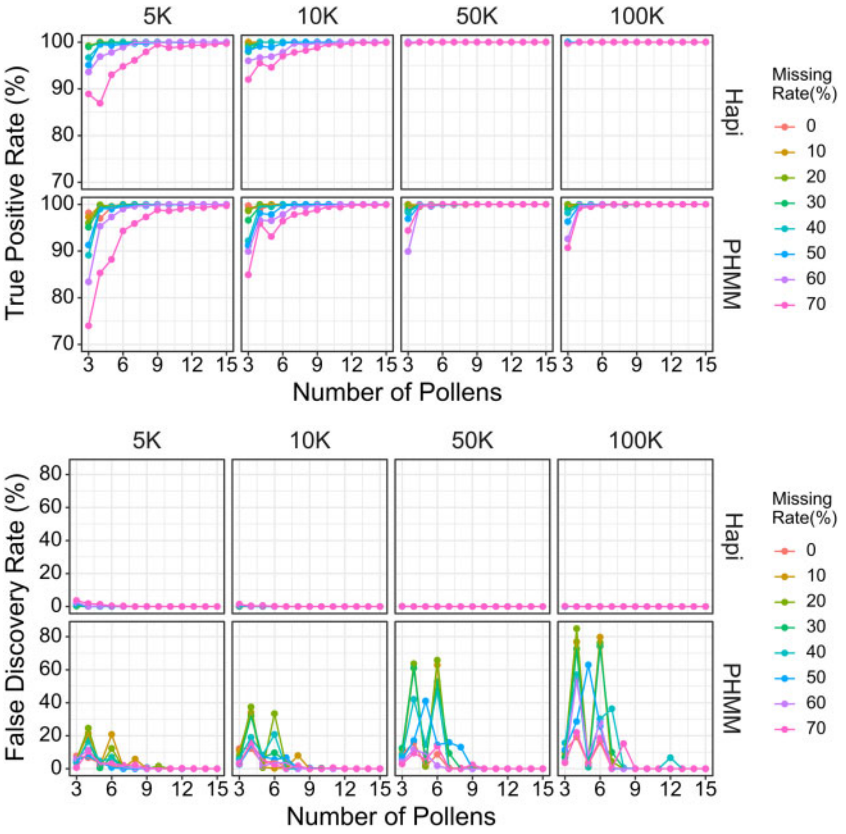


Fig. 7. Comprehensive simulation study comparing the performance of *Hapi* and *PHMM* for crossover detection in terms of TPR and false discovery rate (FDR). The same simulation data set for haplotype phasing analysis was used; that is, a pool of 100 haploid gametes were simulated where the number of hetSNPs ranged from 5,000 (5K) to 100,000 (100K) and the rate of missing genotype data ranged from 10% to 70%. For the chromosome in each simulated gamete, one to three crossovers and 1% genotyping errors were introduced. In each comparison, 3–15 simulated gametes were randomly selected from the gamete pool for haplotyping and the process was repeated for 100 times to compute the TPR and FDR for crossover detection.

methods are costly, limiting a wider application of haplotypes. Such boundaries may be lifted by the alternative phasing methods using individual gametes, where the complexity of phasing will be dramatically reduced because the majority of the haplotypic allele combinations has been preserved between recombination breakpoints. In this study, we developed a cost-effective methodology that only requires a few gametes to correctly reconstruct high-resolution chromosomal haplotypes. We first demonstrated that read-based methods are not suitable for inference of chromosomal haplotypes as they only infer haplotype segments, the phase of which need to be resolved too. It should be noted that *WhatsHap* and *HapCUT2* were not devised to analyze gametes for phasing; therefore, it is not surprising to see unfavorable results when they were applied to haploid genomic data. The comparison between *Hapi* and the other existing gamete-based algorithm, *PHMM*, indicated that *Hapi* outperformed *PHMM* in accuracy, repeatability, and cost-effectiveness based on the human sperm data, maize microspore data, and comprehensively simulated data. To achieve the same level of phasing accuracy, *Hapi* required fewer gametes and can tolerate more missing hetSNPs than *PHMM*. The major deficiency in the phasing algorithm of *PHMM* is due to its core strategy of a direct inference of crossover positions, which is sensitive to regions with ambiguous data due to genotyping errors or GC events, or complications caused by multiple crossovers in more than one gamete. In *Hapi*, such genomic regions harboring complicated multiple CV-links will be detected and excised from the draft haplotypes to reduce the chance of phasing errors. Moreover, novel algorithms for handling imperfect data (missing and erroneous genotypes) are also devised for the *Hapi* method. When different numbers of gametes were used for phasing, the new *Hapi* method performed consistently well but the performance of *PHMM* fluctuated wildly, indicating that the *Hapi* method can sufficiently handle ambiguous data to produce reliable phasing results.

Our study also indicated that three gametes may be enough to reconstruct chromosome-length haplotypes by *Hapi* if the genotype data are of high quality. This is theoretically reasonable because DNA recombination is very rare with an average of <3 events affecting most gametic chromosomes in most studied diploid organisms (Beye et al. 2006). It should be noted that using three gametes may fail in a special scenario when two sampled gametes each have a crossover within a very small region. This is because, in the step of proofreading draft haplotypes, small blocks (i.e., <100 hetSNPs) are excluded from the draft haplotypes by default,

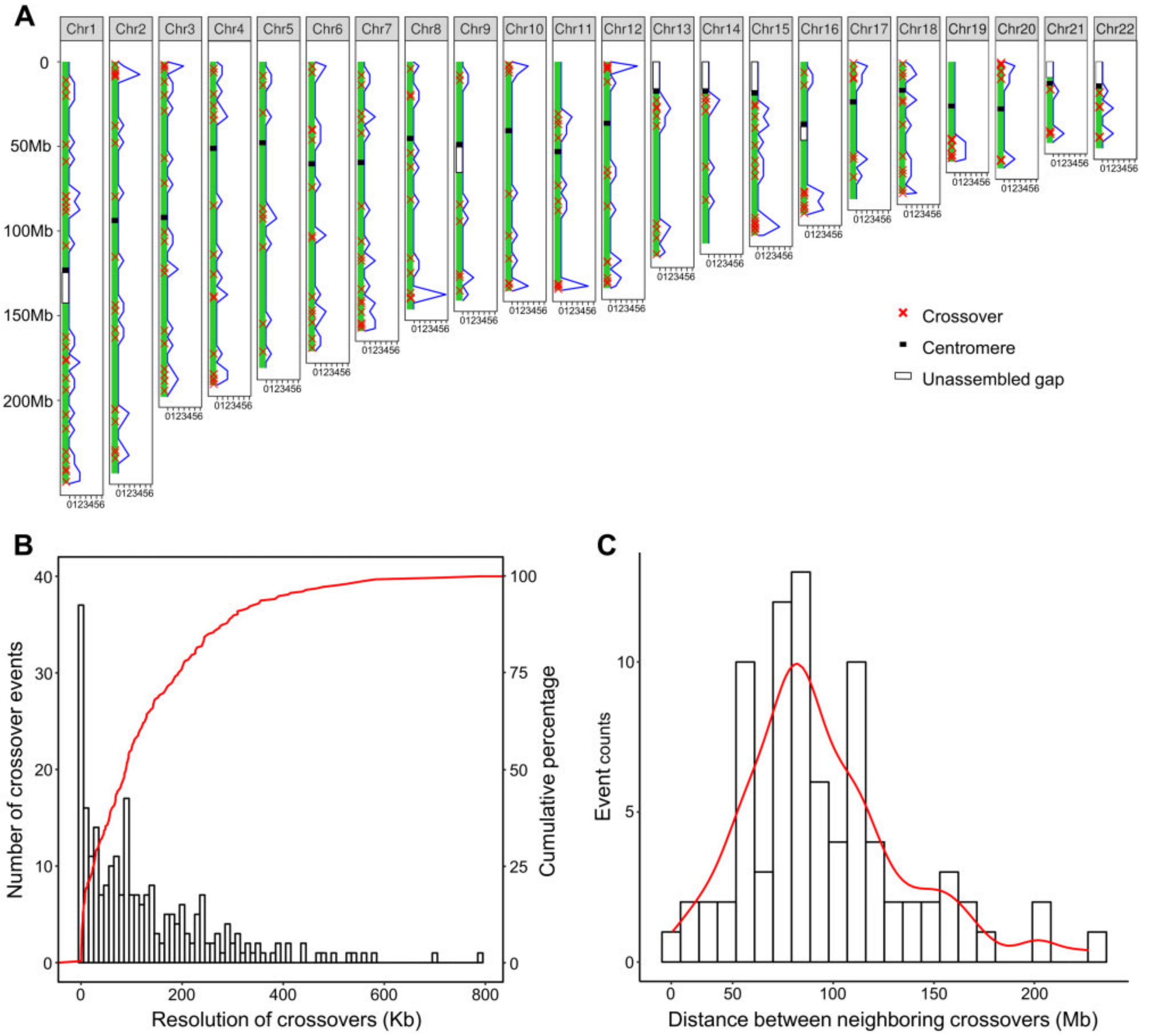


Fig. 8. Crossover analysis in the human sperm sequencing data set. (*A*) The distribution of 254 identified crossovers on the 22 autosomes. (B) The distribution of the crossover resolutions (distance between two adjacent markers that involve a crossover). (*C*) The distribution of distances between two neighboring crossovers.

assuming the probability of having multiple crossovers within these blocks in more than one gamete is low. In this specific but rare scenario, removal of such blocks may lead to the wrong determination of the major link type and thereafter the haplotypes. If only three gametes are available, it is recommended to implement the *Hapi* method with and without removing blocks in constructing draft haplotypes and check the consistency in results from two different settings. Such ambiguity can be easily resolved by slightly increasing the sample size to four or five.

Unlike many other phasing algorithms that demand sequencing long-reads or linked-reads in diploid cells, the *Hapi* method can analyze hetSNP data of single gamete cells generated using any genotyping platform. Advanced technologies, such as 10X Genomics linked-read sequencing, are not necessary for the *Hapi* method but may be used as ancillary approaches to generate designated long-range haplotype fragments for complex and challenging genomic regions, further perfecting the chromosomal haplotypes inferred by the *Hapi* method. Alternatively, the *Hapi* method may be first used to analyze hetSNPs acquired by inexpensive genotyping methods for the inference of a sketch of chromosomal haplotypes; then, these sketch chromosome-length haplotypes can be used to guide haplotype-resolved genome assembly.

Another application of the *Hapi* package is to implement the crossover analysis module to derive maps of recombination in gametes based on the inferred chromosome-length haplotypes. Results from both the comprehensive simulation study and the human sperm sequencing data analysis suggested the feasibility of *Hapi* for crossover detection in the gamete cells. This unique function may be used to investigate recombination hotspots in a population of interest, or monitor the chromosomal crossovers in individual plants to facilitate crop improvement. Recombination events are normally rare, with an overall frequency in a measurable range, which may be used as a gauge for diagnosis of abnormal recombination activity. In our study, we used the available whole-genome sequencing data of 11 sperm cells to benchmark the performance of *Hapi*; whereas, in a clinical setting, it is simple to survey many more sperm cells with little extra cost to produce a high-resolution and reliable recombination map for each male subject. For example, Lu et al. (2012) sequenced 99 sperm cells from a male and observed a decreased crossover frequency in companion with an increase of autosomal aneuploidy in human sperm. This strategy can be equally applied to female subjects too. Hou et al. (2013) identified 2,370 and 2,355 crossovers in the second polar body and female pronucleus of 55 euploid oocytes from eight donors, respectively, and they used these data to create a female personal genetic map and to study crossover interference and chromatid interference. The second study, which focused on human female gametes, also reported lower crossover activity in aneuploid oocytes. Abnormality in recombination frequency during meiosis is one of the primary causes leading to miscarriage and birth defects. These research or clinical practices on the human reproductive system may remarkably benefit from the crossover analysis function, suggesting a translational potential of *Hapi*.

In summary, we have developed the innovative *Hapi* method for an accurate and efficient inference of chromosome-length haplotypes in individual genomes. The crossover detection module may be used to study DNA recombination and its underlying biological mechanisms. This cost-effective tool will promote a large-scale use of haplotype data in many research areas and inspire scientists who have never used single-gamete sequencing technology to design improved experiments for their studies.

Materials and Methods

Key Component Algorithms Employed in *Hapi*

HMM for Detection of Genotyping Errors

An HMM is adopted to linearly scrutinize hetSNP markers along the chromosome in two gametes to identify markers bearing genotyping errors (supplementary fig. S3, Supplementary Material online). In the HMM, there are two observations "s" and "d" indicating the two possible outcomes, either same or different, in terms of the relationship of observed genotype calls at a hetSNP locus between two gametes. Two hidden states, "S" and "D," represent the invisible relationship between the true genotypes of this marker in these two gametes, with "S" and "D" denoting the same and different genotypes, respectively. The initial probabilities of the two states are 0.5. Because the observed genotype outcomes may be different from the hidden states due to the genotyping errors at rate E, the emission probabilities to observe the same genotype calls, that is, s, given the S hidden state, is $1 - 2E \times (1 - E)$, and to observe the different genotype calls, that is, d, is $2E \times (1 - E)$. The emission probabilities given the D state are defined in the same way. A transition is defined as a change in state when scanning two adjacent markers, indicating that a meiotic recombination likely occurs between these two markers on either gamete chromosome. Suppose the recombination frequency is R, the transition probabilities from one state to itself is $1 - 2R \times (1 - R)$, and to the other state is $2R \times (1 - R)$. After defining the HMM, Viterbi's algorithm (Viterbi 1967) can be used to determine the most likely sequence (or path) of the hidden states for the DNA markers along the chromosome. Markers with genotyping errors are determined where there are conflicts between the observed outcomes and the inferred states. The HMM is iteratively applied to all gamete pairs for the detection of disputed SNP loci with potential genotyping errors.

Imputation of Missing Genotypes

We define a *framework* as a set of selected hetSNPs for constructing draft haplotypes for each chromosome. Missing data for the framework markers in the gametes are imputed in an iterative manner (supplementary fig. S4, Supplementary Material online). When a missing region (either a single marker or consecutive markers) of a "target" gamete is to be imputed, the two markers immediately around this region, called comparator markers, are first compared with those in other "support" gametes. The missing region can be imputed with the information from a support gamete cell only if the genotype calls for these two comparator markers in the target gamete are either both identical or both complementary to those in the support gamete. For example, if genotype calls of the two comparator markers in the target gamete are both identical to those in the support gamete, the missing region on the target gamete is simply imputed with genotype calls of markers in the same region in the support gamete. Otherwise, the missing region in the target gamete is imputed with the reciprocal genotypes in the support gamete. Missing genotypes in one gamete can be eventually resolved only if the imputations are supported by more than two support gametes and no imputation conflict is incurred. Once all the gametes are imputed in one iteration, genotypes in the missing regions are updated and the entire process described above will be repeated until no more missing data can be further imputed.

Majority Voting

With the assumption that recombination is generally rare on the chromosome and even rarer between two neighboring framework markers (a small region) in multiple gametes, the haplotypes of these two adjacent framework markers are deduced by analyzing genotype links (genotype patterns for these two markers) across all gametes based on the majority voting principle. There are two types of links between these two neighboring framework markers, that is, type I links include genotype patterns 0-0 and 1-1 and type II links include genotype patterns 0-1 and 1-0, where 1 and 0 represent two complementary genotype calls that are arbitrarily and independently assigned at either locus (supplementary fig. S5, Supplementary Material online). The most frequent link type is determined as HAP-link which represents the likely haplotypes for the two framework markers, whereas the minority link type is considered as CV-link arising from a crossover. The final draft haplotypes can be deduced through walking and voting along the framework of the chromosome.

Maximum Parsimony of Recombination

MPR (Xie et al. 2010), an optimality criterion to search for the haplotype arrangement with minimum number of crossovers in a chromosomal region across all gametes, is adopted by *Hapi* to proofread the equivocal regions (two adjacent framework markers) of draft haplotypes where disputed CV-links have been observed. When five or more gametes are analyzed, we treat any two adjacent markers with two or more CV-links as candidate regions for proofreading (supplementary fig. S6, Supplementary Material online). If very few (e.g., 3 or 4) gametes are in use, every two adjacent markers with any CV-link are subject to proofreading. The draft haplotypes are first segmented into blocks by the equivocal regions. Small blocks (<100 hetSNPs) with little genotypic data are excluded from the construction of the draft haplotypes. To phase two neighboring blocks, raw genotype calls (with possible missing data) of the joining hetSNPs markers, that is, the last 100 consecutive hetSNPs in the first block and the first 100 consecutive hetSNPs in the second block, are retrieved. As haplotypes within each block are unambiguous, there are only two possible combining haplotypes for these two blocks. The total number of crossovers in all gametes is counted given the two combining haplotypes, and the one generating less crossovers is preferred by the MPR algorithm.

Assembly of Consensus Chromosome-Length Haplotypes

One of the inferred draft haplotypes is arbitrarily selected and used as a blueprint to deduce gamete-specific haplotypes and eventually assemble the chromosome-length consensus haplotypes through three steps (supplementary fig. S7, Supplementary Material online). In step 1, genotype calls of framework markers in each gamete chromosome are compared with the blueprint to identify HCPs which are caused by potential recombination. These HCPs partition each gamete chromosome into k haplotype segments, where $k - 1$ is the number of HCPs identified for this gamete chromosome. For the segments 1 through k, genotype calls of hetSNPs in every second segment are flipped to form a gamete-specific haplotype, where “flip” refers to switching the current genotype call to its reciprocal genotype. In step 2, each gamete-specific haplotype is synchronized with the blueprint by either remaining the same or flipping over the genotypes of entire chromosomal hetSNPs. In step 3, the first consensus chromosome-length haplotype is reconstructed via voting for the most frequent allele at each hetSNP locus across all the gamete-specific haplotypes. The second consensus haplotype is obtained by simply flipping genotypes of hetSNPs on the first chromosome-length haplotype.

If a crossover occurs at the end of a gamete chromosome where hetSNPs are not enclosed in the framework, it becomes challenging to correctly infer the haplotypes for this chromosome-tip region. *Hapi* employs an additional capping strategy to polish two ends of chromosomal haplotypes. First, hetSNPs in such a region are combined with the immediately adjacent 200 consecutive hetSNPs at the joining end of the framework to form a capping block, of which the haplotypes can be inferred by treating them as a small chromosome. Then, small-scale draft haplotypes are constructed for the selected framework markers of this capping block by using the most frequently represented genotype calls across the gametes. The same strategy is adopted to generate gamete-specific haplotypes to deduce consensus haplotypes for this small chromosome-tip region. Lastly, the inferred haplotypes for the capping block are integrated into the chromosome-length haplotypes.

Rival Phasing Methods

Pairwise HMM

The *PHMM* approach, the only published gamete-based phasing pipeline, adopted a reference-offspring pairwise-comparison strategy to identify HCPs in each gamete using an HMM to assemble the chromosome-length haplotypes (Hou et al. 2013). For each reference chromosome, a crossover can be directly inferred if, within a 1-Mb sliding window, HCPs can be identified in over 60% of the reference-offspring pairs. Detailed description of the pipeline can be found in the original article (Hou et al. 2013). The source code, which consists of a series of C++ programs and Perl scripts for implementing the *PHMM* pipeline, is publicly available from https://sourceforge.net/projects/phacro/files/ (last accessed July 21, 2020). To facilitate the comparison analysis in this study, we directly applied the C++ programs for crossover identification but rewrote the Perl scripts in R (without changing the original algorithm) for the inference of consensus haplotypes.

WhatsHap

WhatsHap is a read-based method which was initially devised for phasing long-read sequencing data from diploid somatic cells sequenced by third-generation sequencing technologies, such as PacBio and Oxford Nanopore sequencing (Martin et al. 2016). Nevertheless, the method can also be adapted to next-generation sequencing data for inference of haplotypes. *WhatsHap* directly uses mapped sequencing reads spanning at least two heterozygous variants to assemble haplotype segments of an individual. The core algorithm of *WhatsHap* is to compare all potential haplotypes to determine the optimal one, which can assign all reads with the least amount of sequencing errors to be corrected and/or erroneous reads to be removed by solving the weighted minimum error correction problem. In this study, sequencing data of single gametes from the donor were combined and the default settings were applied to infer the haplotypes of the individual.

HapCUT2

The *HapCUT2* approach, similar to *WhatsHap*, is another popular read-based phasing method for data generated using various sequencing platforms (Edge et al. 2017). This method infers longer haplotypes that are most agreeable to the observed mini-haplotypes represented by the sequence of alleles at heterozygous variant sites identified from aligned sequence reads. The default settings were applied to the phasing analysis in the study.

Quality Metrics for Evaluation of Phasing Performance

Four quality metrics, which have been used in previous research (Porubsky et al. 2017), were adopted in the study to assess and compare the phasing performance of the compared methods.

Completeness

A phase connection between two neighboring hetSNPs is defined if these two hetSNPs can be phased by a method. The number of phase connections at the chromosomal scale equals to the number of phased hetSNPs minus the number of inferred haplotype segments on that chromosome. The COM of phasing a chromosome is defined as the ratio of the number of phase connections to the maximum possible number of phase connections, which equals to the total number of hetSNPs on that chromosome minus 1. We calculated the COM of an entire genome as the weighted sum of COMs across chromosomes, with the weights being proportional to the numbers of hetSNPs on each of these chromosomes.

Switch Error Rate

The SER is defined as the number of incorrect phase connections (switch errors) divided by the total number of phase connections within each inferred haplotype segment. It is a commonly used measure of "local" accuracy for a phased haplotype segment or a phased chromosome. The SER for a whole chromosome can be accurately calculated for the gamete-based phasing methods because chromosome-length haplotypes can be inferred. Nevertheless, block-wise SER (cumulative switch errors across all phased segments divided by the total number of phase connections on the chromosome) was calculated for the read-based phasing approaches to represent the chromosomal SER, which may be severely underestimated because the connections between haplotype segments cannot be taken into account. The SER of an entire genome was calculated as the weighted sum of SERs across chromosomes, with the weights being proportional to the numbers of phase connections on each of these chromosomes.

Hamming Error Rate

The fraction of incorrectly phased hetSNPs, called the HER, was proposed to evaluate the phasing accuracy for the LHS (Porubsky et al. 2017). This is because a single switch error in a large phased haplotype segment may result in substantial difference (or Hamming distance) between the inferred and true haplotypes. Similar to SER, the HER for a whole chromosome can be accurately calculated for the gamete-based phasing methods, whereas block-wise HER (cumulative Hamming distance across all phased segments divided by the total number of hetSNPs on the chromosome) was calculated for the read-based phasing approaches to represent the chromosomal HER. Similarly, the HER for a genome can be calculated as a weighted sum of chromosomal HERs, with the weights being proportional to the numbers of the phased hetSNPs on these chromosomes.

Largest Haplotype Segment

As we are interested in ability of inferring the haplotypes that span the whole length of a chromosome, the fraction of phased hetSNPs in the LHS was reported for each chromosome to reflect the contiguity for each phasing method.

Human Sperm Sequencing Data Set

Single sperm cell sequencing data of 11 sperms from the donor of the HuRef diploid genome were downloaded from the NCBI SRA under the accession number SRP017516 (Kirkness et al. 2013). Sequencing reads were aligned to the human GRCh37 reference genome using *BWA-MEM* (Li and Durbin 2009) implemented in the *SpeedSeq* software (Chiang et al. 2015). Duplicate-marked, sorted, and indexed BAM files were produced by the *SpeedSeq* align module, which utilizes *SAMBLASTER* (Faust and Hall 2014) to mark duplicates and uses *Sambamba* (Tarasov et al. 2015) to sort and index BAM files. For each sperm, the genotypes at 1.95 million heterozygous SNP loci in the HuRef genome were determined using the Genome Analysis Toolkit (*GATK*) (DePristo et al. 2011).

Maize Microspore Sequencing Data Set

The raw sequencing data in the maize microspore sequencing data set were available from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) (https://www.ncbi.nlm.nih.gov/sra; last accessed July 21, 2020) under the accession number SRP047362 (Li et al. 2015). In the study, a total of 96 (24 × 4) microspores from 24 tetrads were isolated from F1 hybrid individuals of a cross between two inbred lines (SK and ZHENG58) and were sequenced at ∼1.4× depth coverage. Parents of the F1 hybrid were also sequenced at up to 8× (SK) and 15.7× (ZHENG58) genome coverage depth, respectively. After a stringent filtering process, a total of 599,154 high-quality SNPs were obtained for both parents and the microspores.

Data Availability

Hapi is an R package that is freely available at https://github.com/Jialab-UCR/Hapi.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank Dr Jianbing Yan and Dr Xiang Li at Huazhong Agricultural University for sharing the processed genotype data for the microspores from their maize tetrad study. We also acknowledge the helpful discussions with Dr Hongru Wang in the Department of Integrative Biology at

University of California, Berkeley, and with Dr Li Lei at the US Department of Energy (DOE) Joint Genome Institute (JGI) of Lawrence Berkeley National Laboratory. This work was supported by start-up funding of University of California, Riverside (to Z.J.); UC Academic Senate Regents Faculty Fellowship and Faculty Development Award (to Z.J.); UCR Hellman Fellowship (to Z.J.); the National Institute of Food and Agriculture (Grant No. 2013-67013-21110 to M.L.R.); National Key Basic Research Program of China (Grant No. 2015CB553706 to W.-D.Z.); National Natural Science Foundation of China (Grant No. 81571427 to W.-D.Z.); Guangzhou Municipal Science and Technology Project (Grant Nos. 201803040001 and 201707010291 to W.-D.Z.); National Natural Science Foundation of China (Grant No. 81660426 to J.Z.); and the Science and Technology Project of Guizhou Province (Grant No. 2017-5803 to J.Z.).

References

Baetscher DS, Clemento AJ, Ng TC, Anderson EC, Garza JC. 2018. Microhaplotypes provide increased power from short-read DNA sequences for relationship inference. *Mol Ecol Resour*. 18(2):296–305.

Beye M, Gattermeier I, Hasselmann M, Gempe T, Schioett M, Baines JF, Schlipalius D, Mougel F, Emore C, Rueppell O, et al. 2006. Exceptionally high levels of recombination across the honey bee genome. *Genome Res*. 16(11):1339–1344.

Browning SR, Browning BL. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet*. 81(5):1084–1097.

Chiang C, Layer RM, Faust GG, Lindberg MR, Rose DB, Garrison EP, Marth GT, Quinlan AR, Hall IM. 2015. SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nat Methods*. 12(10):966–968.

Coop G, Wen X, Ober C, Pritchard JK, Przeworski M. 2008. High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. *Science* 319(5868):1395–1398.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*. 43(5):491–498.

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

Fan HC, Wang J, Potanina A, Quake SR. 2011. Whole-genome molecular haplotyping of single cells. *Nat Biotechnol*. 29(1):51–57.

Fariello MI, Boitard S, Naya H, SanCristobal M, Servin B. 2013. Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. *Genetics* 193(3):929–941.

Faust GG, Hall IM. 2014. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics* 30(17):2503–2505.

Glusman G, Cox HC, Roach JC. 2014. Whole-genome haplotyping approaches and genomic medicine. *Genome Med*. 6(9):73.

Goldmann JM, Wong WSW, Pinelli M, Farrah T, Bodian D, Stittrich AB, Glusman G, Vissers LELM, Hoischen A, Roach JC, et al. 2016. Parent-of-origin-specific signatures of de novo mutations. *Nat Genet*. 48(8):935–939.

Harris K, Nielsen R. 2013. Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS Genet*. 9(6):e1003521.

Hinch AG, Zhang G, Becker PW, Moralli D, Hinch R, Davies B, Bowden R, Donnelly P. 2019. Factors influencing meiotic recombination revealed by whole-genome sequencing of single sperm. *Science* 363(6433):eaau8861.

Hou Y, Fan W, Yan L, Li R, Lian Y, Huang J, Li J, Xu L, Tang F, Xie XS, et al. 2013. Genome analyses of single human oocytes. *Cell* 155(7):1492–1506.

Howie BN, Donnelly P, Marchini J. 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet*. 5(6):e1000529.

Huang J, Howie B, McCarthy S, Memari Y, Walter K, Min JL, Danecek P, Malerba G, Trabetti E, Zheng H-F, et al. 2015. Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat Commun*. 6(1):8111.

Consortium International HapMap. 2005. A haplotype map of the human genome. *Nature* 437:1299–1320.

Kirkness EF, Grindberg RV, Yee-Greenbaum J, Marshall CR, Scherer SW, Lasken RS, Venter JC. 2013. Sequencing of isolated sperm cells for direct haplotyping of a human genome. *Genome Res*. 23(5):826–832.

Kitzman JO, MacKenzie AP, Adey A, Hiatt JB, Patwardhan RP, Sudmant PH, Ng SB, Alkan C, Qiu R, Eichler EE, et al. 2011. Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nat Biotechnol*. 29(1):59–63.

Kong A, Steinthorsdottir V, Masson G, Thorleifsson G, Sulem P, Besenbacher S, Jonasdottir A, Sigurdsson A, Kristinsson KT, Jonasdottir A; DIAGRAM Consortium, et al. 2009. Parental origin of sequence variants associated with complex diseases. *Nature* 462(7275):868–874.

Lambert J-C, Grenier-Boley B, Harold D, Zelenika D, Chouraki V, Kamatani Y, Sleegers K, Ikram MA, Hiltunen M, Reitz C, et al. 2013. Genome-wide haplotype association study identifies the FRMD4A gene as a risk locus for Alzheimer's disease. *Mol Psychiatry*. 18(4):461–470.

Leitwein M, Duranton M, Rougemont Q, Gagnaire P-A, Bernatchez L. 2020. Using haplotype information for conservation genomics. *Trends Ecol Evol*. 35(3):245–258.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25(14):1754–1760.

Li X, Li L, Yan J. 2015. Dissecting meiotic recombination based on tetrad analysis by single-microspore sequencing in maize. *Nat Commun*. 6(1):6648.

Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. 2010. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol*. 34(8):816–834.

Loh P-R, Danecek P, Palamara PF, Fuchsberger C, A Reshef Y, K Finucane H, Schoenherr S, Forer L, McCarthy S, Abecasis GR, et al. 2016. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet*. 48(11):1443–1448.

Lohmueller KE, Bustamante CD, Clark AG. 2009. Methods for human demographic inference using haplotype patterns from genomewide single-nucleotide polymorphism data. *Genetics* 182(1):217–231.

Lu S, Zong C, Fan W, Yang M, Li J, Chapman AR, Zhu P, Hu X, Xu L, Yan L, et al. 2012. Probing meiotic recombination and aneuploidy of single sperm cells by whole-genome sequencing. *Science* 338(6114):1627–1630.

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

Martin M, Patterson M, Garg S, Fischer S, Pisanti N, Klau GW, Schöenhuth A, Marschall T. 2016. WhatsHap: fast and accurate read-based phasing. BioRxiv: 085050.

McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, Kang HM, Fuchsberger C, Danecek P, Sharp K, et al. 2016. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet*. 48(10):1279–1283.

McKinney GJ, Seeb JE, Seeb LW. 2017. Managing mixed-stock fisheries: genotyping multi-SNP haplotypes increases power for genetic stock identification. *Can J Fish Aquat Sci*. 74(4):429–434.

O'Connell J, Sharp K, Shrine N, Wain L, Hall I, Tobin M, Zagury J-F, Delaneau O, Marchini J. 2016. Haplotype estimation for biobank-scale data sets. *Nat Genet*. 48(7):817–820.

Palamara PF, Lencz T, Darvasi A, Pe'er I. 2012. Length distributions of identity by descent reveal fine-scale demographic history. *Am J Hum Genet*. 91(5):809–822.

Pendleton AL, Shen F, Taravella AM, Emery S, Veeramah KR, Boyko AR, Kidd JM. 2018. Comparison of village dog and wolf genomes highlights the role of the neural crest in dog domestication. *BMC Biol*. 16(1):64.

Peters BA, Kermani BG, Sparks AB, Alferov O, Hong P, Alexeev A, Jiang Y, Dahl F, Tang YT, Haas J, et al. 2012. Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature* 487(7406):190–195.

Porubsky D, Garg S, Sanders AD, Korbel JO, Guryev V, Lansdorp PM, Marschall T. 2017. Dense and accurate whole-chromosome haplotyping of individual genomes. *Nat Commun*. 8(1):10.

Porubský D, Sanders AD, Van Wietmarschen N, Falconer E, Hills M, Spierings DC, Bevova MR, Guryev V, Lansdorp PM. 2016. Direct chromosome-length haplotyping by single-cell sequencing. *Genome Res*. 26(11):1565–1574.

Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419(6909):832–837.

Scheet P, Stephens M. 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet*. 78(4):629–644.

Schiffels S, Durbin R. 2014. Inferring human population size and separation history from multiple genome sequences. *Nat Genet*. 46(8):919–925.

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

Snyder MW, Adey A, Kitzman JO, Shendure J. 2015. Haplotype-resolved genome sequencing: experimental methods and applications. *Nat Rev Genet*. 16(6):344–358.

Stephens M, Scheet P. 2005. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am J Hum Genet*. 76(3):449–462.

Stephens M, Smith NJ, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet*. 68(4):978–989.

Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. 2015. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* 31(12):2032–2034.

Trégouët D-A, König IR, Erdmann J, Munteanu A, Braund PS, Hall AS, Großhennig A, Linsel-Nitschke P, Perret C, DeSuremain M, et al. 2009. Genome-wide haplotype association study identifies the SLC22A3-LPAL2-LPA gene cluster as a risk locus for coronary artery disease. *Nat Genet*. 41(3):283–285.

Viterbi A. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans Inform Theory*. 13(2):260–269.

Xie W, Feng Q, Yu H, Huang X, Zhao Q, Xing Y, Yu S, Han B, Zhang Q. 2010. Parent-independent genotyping for constructing an ultrahigh-density linkage map based on population sequencing. *Proc Natl Acad Sci U S A*. 107(23):10578–10583.

Xue S, Bradbury PJ, Casstevens T, Holland JB. 2016. Genetic architecture of domestication-related traits in maize. *Genetics* 204(1):99–113.

Yang H, Chen X, Wong WH. 2011. Completely phased genome sequencing through chromosome sorting. *Proc Natl Acad Sci U S A*. 108(1):12–17.