



Mini Review

Unzipping haplotypes in diploid and polyploid genomes

Xingtang Zhang^{a,b}, Ruoxi Wu^a, Yibin Wang^a, Jiaxin Yu^a, Haibao Tang^{a,*}

^a Center for Genomics and Biotechnology, Fujian Provincial Key Laboratory of Haixia Applied Plant Systems Biology, Key Laboratory of Genetics, Breeding and Multiple Utilization of Corps, Ministry of Education, Fujian Agriculture and Forestry University, Fuzhou 350002, China

^b State Key Laboratory of Ecological Pest Control for Fujian and Taiwan Crops, College of Plant Protection, Fujian Agriculture and Forestry University, Fuzhou 350002, China

ARTICLE INFO

Article history:

Received 16 September 2019

Received in revised form 25 November 2019

Accepted 26 November 2019

Available online 9 December 2019

Keywords:

Genome assembly

Haplotype phasing

Ploidy

Reference genome

Heterozygosity

ABSTRACT

Diploid genomes consist of two homologous copies of chromosomes with one from each parent while polyploid genomes contain more than two homologous sets of chromosomes. Most of the reference genome assemblies collapsed haplotypes that represent ‘mosaic’ sequences, ignoring allelic variants that may be involved in important cellular and biological functions. Unzipping haplotypes into distinct sets of sequences has been a growing trend in recent genome studies, as it is an essential tool towards resolving important clinical and biological questions, such as compound heterozygotes, heterosis, and evolution. Herein, we review existing methods for alignment-based and assembly-based haplotype phasing for heterozygous diploid and polyploid genomes, as well as recent advances of experimental approaches for improved genome phasing. We anticipate that full haplotype phasing could become a routine procedure in genome studies in the near future.

© 2019 The Authors. Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Contents

1. Introduction	66
2. Alignment-based haplotype phasing	68
3. Assembly-based haplotype phasing with haploid representation	69
4. Assembly-based haplotype phasing with full haplotype resolution	69
5. Chromosomal scale of haplotype reconstruction using Hi-C technology	70
6. Experimental methods targeted at haplotype-resolved genome sequencing	70
7. Summary and outlook	71
Declaration of Competing Interest	71
Acknowledgements	71
Appendix A. Supplementary data	71
References	71

1. Introduction

Assembly of the reference genome is a common route to improve the utilization of the genetic resources for many organisms nowadays. With the availability of single molecule long-read sequencing, such as PacBio SMRT (Single Molecular Real-time) sequencing, Oxford Nanopore Technologies (ONT) and Bio-nano Genomics, as well as high-throughput chromatin conforma-

tion capture (Hi-C) technology, most genome sequencing studies can now be completed to chromosomal level assemblies at a fractional cost compared to just several years ago.

In most reference genome assemblies, typically two homologous copies of every chromosome, one from each parent, were collapsed together and considered as a ‘mosaic’ reference of the two haplotypes in the current diploid genome assembly approaches. Such reference is called a ‘monoploid’ representation which only reflects a single haplotype throughout the genome. For example, the human reference genome (the Genome Reference Consortium, with the most recent release of GRCh38) is derived from 13 volun-

* Corresponding author.

E-mail address: tanghaibao@gmail.com (H. Tang).

teers living in New York, and only contains the O allele for the ABO blood group locus. Similarly, for polyploid genomes, a single haploid was often targeted in the reference genome, thereby ignoring a large amount of genetic diversity within the sequenced organism.

While the ‘monoploid’ references are much simpler to reason about and compare against, they often fail to capture the diploid or polyploid nature of the organisms and ignore allelic variants that can have potentially important functions. Since the initial completion of Human Genome Project (HGP) two decades ago, researchers have been trying to improve the genome assemblies to completely resolve both haplotypes in the same sequencing study. Recent improvements in sequencing technologies and particularly the development of long read sequencing provide opportunities to resolve structural variations between haplotypes that are completely absent in the linear reference genome. In human genome studies, reconstruction of both haplotypes is also clinically relevant, for example, when determining the presence of compound heterozygous mutations, as well as for accurate Human Leukocyte antigen (HLA) typing.

Another important application of genome phasing is to study allele-specific expression (ASE) or allelic imbalance (AI), which has been suggested as an important mechanism for causing heterosis. ASE or AI reveals a pattern of preferential expression of one parental allele over another [1]. This differential expression pattern likely results from allelic variants located in *cis*-regulatory elements, which are able to interact with environmental factors to regulate complex expression networks [2] and eventually lead to large phenotypic variation. For instance, evidence increasingly reveals transcriptionally more active alleles with dominant expression patterns are important contributors to heterosis in hybrid rice

[3,4]. The current approaches to identify ASE relied on phasing RNA-seq reads onto different haplotypes. Typically, RNA-seq reads produced on Illumina short reads sequencing platforms are aligned to a reference genome and variants that belong to each parental genome are phased and the corresponding reads are further used to quantify the gene expression for their respective alleles [3].

Additionally, identification of allelic variants can also facilitate research of polyploid evolution and inform crop breeding. The complex polyploid sweet potato genome was first assembled into a consensus reference genome and a novel haplotype phasing method successfully generated a haplotype-resolved genome assembly [5]. Phylogenetic analysis using phased allelic variants uncovered a total of six haplotypes, thereby tracking the hexaploidization history in the lineage of sweet potato [5]. In the sugarcane project, we recently developed a novel ALLHiC algorithm that is able to build allele-aware and chromosomal-scale assembly of autopolyploid sugarcane genome by incorporating PacBio long reads sequencing and Hi-C technology [6,7]. The phased sugarcane genome with individually assembled haplotypes further revealed the complex chromosome rearrangement and evolutionary history in the *Saccharum* lineage [6].

Haplotype phasing has become a fundamental problem in heterozygous and polyploid genome assemblies. Haplotype phasing is the key for the most accurate representation of genetic composition for a given organism. Herein, we review the existing computational approaches for haplotype phasing – including both alignment-based and assembly-based phasing approaches for heterozygous and polyploid genomes. An overview of the two phasing approaches is illustrated in Fig. 1. We follow up the discussion of computational approaches by a brief summary of experi-

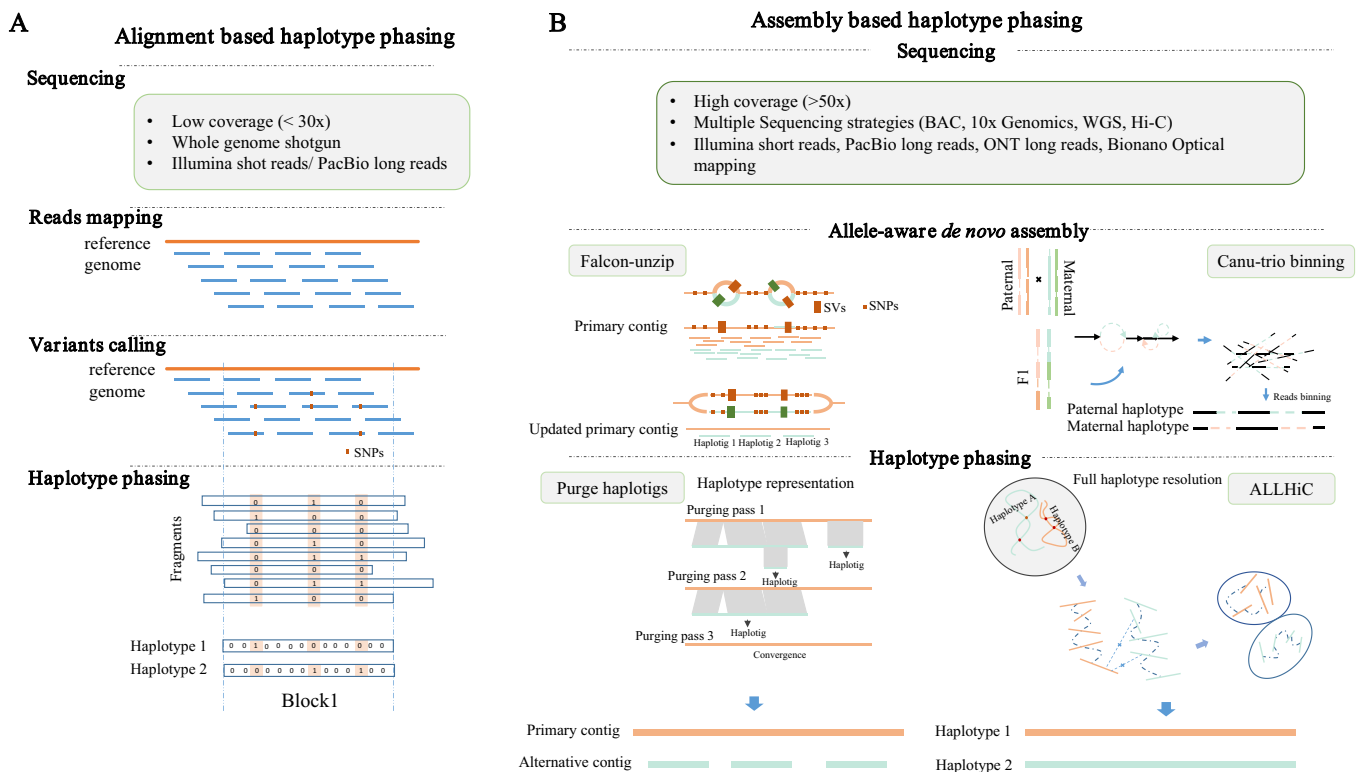


Fig. 1. Overview of the two main classes of haplotype phasing strategies. The left panel (A) is alignment-based and the right one is assembly-based haplotype phasing workflow, respectively. In the alignment-based haplotype phasing, reads are sequenced with relatively low coverage (<30x) and are mapped to a reference genome for variant calling. Afterwards, linked variants are extended into phased blocks each containing a number of neighboring SNPs represented as 0(REF)/1(ALT). In the assembly-based haplotype phasing (B), much deeper sequencing is typically carried out using a variety of sequencing technologies. Allele-aware *de novo* assembly can be achieved using Falcon-unzip or Canu trio-binning methods. When working with multiple haplotypes, primary contigs can be selected as an arbitrary haplotype representation, e.g. using *purge_haplotigs*, for downstream analysis. Alternatively, full set of haplotypes can be resolved through Hi-C technology, e.g. using ALLHiC.

mental advances to assist genome phasing. Finally, the limitation of current approaches as well as future directions are discussed.

2. Alignment-based haplotype phasing

Plummeting cost of Next-generation sequencing (NGS) provided large amounts of data to perform genome-wide scale of haplotype phasing and facilitated the continuous development of phasing algorithms during the past two decades. When the reference sequence is available, the most straightforward method of haplotype phasing is by first aligning the whole genome sequencing (WGS) reads, followed by compiling a set of heterozygous genotypes at polymorphic sites and finally pairing the neighboring haplotypes successively (Fig. 1). The exact pairing information of neighboring sites is provided by the co-occurrence of alleles on the same WGS read or read pair.

Different alignment-based phasing algorithms have been designed based on different optimization criteria or objective functions, including minimum error correction (MEC), Weighted mini-

mum letter flip (WMLF) and Maximum fragment cut (MFC) [8]. These optimization criteria are often NP-hard, and thus often relying on a number of heuristics to speed up the computation. Briefly, MEC aims at reconstruction of the two haplotypes by applying the minimum number of base corrections [9]. Modified from MEC algorithm, WMLF also measures error by the number of base flips [8]. MFC converts the haplotype determination problem to a Max-Cut problem and finds the solution by searching for the maximum distance that can link each SNP edge of a haplotype block. Other algorithms, such as Graph [10], Heuristic dynamic programming [11], Mixture model [12] and Fuzzy conflict graphs [13], are also applicable to genome phasing, each leveraging different statistical models for errors.

A number of haplotype phasing programs based on these algorithms have been reviewed by Je-Keun et al. in 2015 [8]. Therefore, we only focus on the newly developed NGS-based phasing programs since 2015 and highlight their respective strengths as well as limitations (Table 1). WhatsHap provides a dynamic programming algorithm to solve the problem of wMEC, which makes the runtime complexity linear to the number of SNPs [14]. HapCUT2,

Table 1
Overview of softwares listed in this study.

Program	Algorithm	Input data	Highlights	Limitations	Citation
<i>Alignment based haplotype phasing</i>					
WhatsHap	Dynamic programming algorithm	VCF/BAM/reference genome	Good performance in completeness and accuracy; Applicable to both short and long reads	Ignore structural variations; Allow only $15 \times$ the maximum coverage of input data; Only for diploid genome	[14]
HapCUT2	MAX-CUT-based heuristic algorithm	VCF/BAM/reference genome	Capable of handling a wide range of sequencing technologies, including Illumina short, PacBio long, $10 \times$ Linked and Hi-C reads	Ignore structural variations; Only for diploid genome	[15]
SHAPEIT3&4	HMM/MCMC/PBWT	VCF/Genetic map	Excellent performance on accuracy and speed; Able to handle large data size	Ignore structural variations; Not suitable for directly phasing low-coverage sequencing data	[16,17]
<i>Assembly based haplotype phasing - Haplotype representation</i>					
HaploMerger1/2	Whole genome comparison	Draft genome assembly	Suitable for diploid assemblies with high heterozygosity; Implement flexible and sensitive assembly error detection	Not suitable for too fragmented scaffold (e.g. $N50 < 100$ kb)	[26,27]
Redundans	Whole genome comparison	Draft genome assembly	Multiple functionalities, including removing heterozygous sequences, scaffolding, gap closure	May throw away some repetitive and paralogous contigs in reducing step	[23]
Purge_haplotigs	Read depth	BAM/draft genome assembly	It is able to avoid part of repetitive and paralogous contigs to get over-purged; Very fast and scales well to large genome size	Cannot resolve haplotype switching in draft genome due to arbitrary retention of contigs, 'pseudo-haploid' doesn't represent true phasing in polymorphic regions	[28]
FALCON&FALCON-Unzip	Heuristic algorithms to identify 'bubble' structure and greedy algorithm to assist constructing haplotype	PacBio raw reads	FALCON-Unzip is capable of assembling highly accurate, contiguous primary contigs and haplotigs that allows further downstream-analysis on haplotype level	Primary contigs contain haplotype switching error between adjacent phase blocks; The detection of repeated and heterozygous sequences can interfere with each other, resulting in erroneous haplotype assembly.	[19,20]
CANU	Trio binning	Parental short reads/ F1 long reads	Able to generate two sets of haploid genomes for each parent line; Trio binning performing extremely well in continuity and accuracy	Limited application for highly heterozygous genomes with no recorded pedigree information	[21]
<i>Assembly based haplotype phasing - Full haplotype resolution</i>					
FALCON-Phase	A pipeline integrating PacBio reads and Hi-C data to reassign haplotypes for diploid genome	Output for FALCON-unzip assembly/Hi-C reads	Can benefit from allele aware contig assembly by FALCON-Unzip; Integrate PacBio contig assembly and Hi-C reads	Only for diploid genome; Not compatible with other contig assemblers	[36]
ALLHiC	Prune/optimize/Genetic Algorithm	Draft contig assembly/Hi-C mapping BAM/Allele table	It is applicable to a wide range of genomes with different complexity, including simple diploid, heterozygous diploid, allo-polyloid genomes and auto-polyloid genomes; Better performance when contig continuity is low	Sensitive to the accuracy of the starting contig assembly; Requires a closely related reference genome to generate an allelic contig table	[7]

which is an extension of HapCUT program, is able to handle a wide range of sequencing technologies, including NGS short reads, long reads, linked-reads and Hi-C reads. Different from HapCUT, HapCUT2 adopts a likelihood-based model to estimate technology-specific errors and iteratively searches for a subset of variants using max-cut computations in the read-haplotype graph [15, p. 2]. SHAPEIT is a series of software programs that are designed for the estimation of haplotypes based on population-level polymorphism data. The recently developed SHAPEIT3 modified the Markov chain Monte Carlo (MCMC) sampling routine, which speeds up the process and is capable to handle biobank-scale data sets with very low error rates of haplotype switching [16]. As an extension of SHAPEIT3, SHAPEIT4 incorporates a Positional Burrow Wheeler Transform (PBWT) based approach to rapidly select a small set of informative haplotypes from the reference panel [17]. SHAPEIT4 also exhibits sub-linear scaling with sample size and allows for integrating external phasing information such as large reference panels of haplotypes, collections of pre-phased variants and long sequencing reads [17].

3. Assembly-based haplotype phasing with haploid representation

In contrast to the alignment-based phasing approaches that are mostly targeted at small variants, assembly-based approaches are often more accurate and can cover larger types of genomic variations, such as large indels and structural variations (Table 1). However, heterozygous diploid or polyploid genome assembly can be challenging to assemble due to the presence of multiple haplotypes, leading to ambiguities and redundancies in the initial contig-level assemblies.

To deal with these ambiguities and redundancies, a common practice of assembling heterozygous genomes is to simply reconstruct a single haplotype to represent the whole genome (Fig. 1). Assembly of heterozygous genomes likely resulted in sequences with high levels of divergence separated in the contig level assembly, i.e. presence of different alleles at the same loci in homologous chromosomes. Due to these common sequences between the haplotypes, the preliminary contig-level assemblies often contain sequences that are over 90% similar between the parental alleles. Pair-wise genome alignment of input contigs can be used to determine redundant contigs, which represent distinct haplotypes from polymorphic regions by comparing amongst the preliminary contigs. This approach has been implemented in Redundans program [23], and successfully assisted heterozygous genome assemblies, such as *Echinochloa crus-galli* [24] and Colorado potato beetle [25].

Another automated pipeline, “HaploMerger” was proposed to reconstruct the allelic relationships of contigs in the diploid assembly [26]. HaploMerger uses the LASTZ–ChainNet method for whole-genome comparison, and the so-called diploid genome assembly (DGA) diagram is used to describe the relationship between intermediate genes or homologues in the assembly of the diploid genome [26]. HaploMerger showed excellent performance on several polymorphic diploid genomes during its testing and did not introduce new assembly errors, showing its efficacy to analyze and utilize polymorphic genome assembly. HaploMerger2 (HM2) is a major upgrade of the old pipeline, with a re-design of the haploid reconstruction from short and long read diploid assembly [27, p. 2]. The HM2 can handle both low and high heterozygous assemblies, and also provides more flexible assembly error detection and reliable gap closure methods, thereby greatly improving the continuity of the final diploid-level assembly than older pipelines [27, p. 2].

In addition to the whole genome alignment approaches, read depth-based methods can also be used to determine allelic contigs

in heterozygous diploid genome assembly. For a diploid genome assembly with a high level of heterozygosity, a bimodal distribution of read depth is expected which forms the basis to classify the contigs of different origins. The “Purge Haplotigs” program is able to utilize read depth of individual contigs to identify contigs that are suspected to be duplicated sequences, which are assumed to be allelic haplotypes that need to be re-assigned in order to obtain a simple, ‘pseudo-haploid’ reference [28]. While relatively simple to derive, it is important to note that this ‘pseudo-haploid’ does not represent a true phasing of the heterozygous genome since the choice of which contigs to retain can still be arbitrary.

4. Assembly-based haplotype phasing with full haplotype resolution

While a single haploid representation is relatively straightforward to derive in heterozygous assemblies, it also throws away a large amount of sequence information that belongs to the other haplotypes. Modern sequencing approaches, such as Single-molecule, real-time Sequencing (SMRT) developed by Pacific Biosciences and Oxford Nanopore Technology (ONT) offers long read length, promising overwhelming performance of assembly for complex genomes [18] (Fig. 1). In particular, long reads have the potential to recover much longer stretches of haplotypes than short reads.

Currently, Pacific Biosciences technology is often used to assemble many plant genomes, but most of the genome assemblies completed thus far are focused on homozygous individuals or inbred lines, yielding a single representative haplotype. However, for many plant crops that are difficult to be inbred as homozygous individuals, such as many tropical fruit crops, the assembly of heterozygous individuals needs to be handled with additional care. In order to solve this problem, several algorithms are developed to simultaneously reconstruct multiple haplotypes during *de novo* genome assembly. Such algorithms seek to go beyond the overly simplistic monoploid representation of the genome, which are especially valuable for heterozygous diploid and polyploid organisms.

The FALCON and FALCON-Unzip algorithm provide a clean solution that enables sequential assembly of the original sequencing data and ongoing identification of the phased diploid genome [19]. Firstly, FALCON reads sequence alignment from corrected PacBio sequence data, and then builds a string graph based on read overlaps [20]. During this process, the string graph typically contains multiple groups of “haplotype fusion” which are overlapping read groups that show up as ‘bubbles’ in the graph. The ‘bubbles’ represent major structural variations and highly divergent regions between homologous sequences. To resolve these ‘bubbles’ within the assembly graph, FALCON-Unzip analyzes haplotype fusion groups and finds hybrid variants as a basis to ‘unzip’ the otherwise fused haplotypes. One haplotype, often arbitrary, is identified first as a primary path in the graph, or ‘primary’ contigs; while the other haplotype, representing an alternative path, is called ‘associated’ contigs. This method was used to re-assemble the F1 genome from crossing Columbia-0 (col-0) and Cape Verde Island-0 (Cvi-0) ecotypes of *Arabidopsis thaliana*, *Vitis vinifera* CV (Cabernet Sauvignon) and highly heterozygous wild diploid *Clavicornia pyxidate* [19]. These haplotype-resolved assemblies reflect a more realistic representation of their respective genome and allow the study of haplotype structure and heterozygosity in much better accuracy and resolution.

A more recent approach, “trio binning” is proposed to simplify haplotype assembly by addressing allelic variation prior to assembly [21], which has been implemented in the CANU assembler [22]. Compared with existing methods, the effectiveness of this method

increases with the level of heterozygosity. Trio binning starts by partitioning the long reads from the offspring into haplotype specific groups, guided by genome sequencing in each of the two parents. After the partition, each haploid is then assembled independently, resulting in complete diploid reconstruction. To illustrate the utility of “trio binning”, F1 hybrid between cattle subspecies *Bos taurus taurus* and *Bos taurus indicus* were sequenced and fully resolved into two parent haplotypes, with a NG50 of haplotig size of over 20 Mb with 99.998% accuracy, which exceeds the quality of current bovine reference genome [21]. While highly accurate, the setup of “trio binning” requires the parent to be known *a priori* as well as sequenced in order the phase the F1 genome. This requirement not only results in higher cost, but also can be limiting in scenarios such as when the plant species was either collected from the wild or has less developed breeding programs so that the parentage information may not be available.

5. Chromosomal scale of haplotype reconstruction using Hi-C technology

Genome assembly incorporating PacBio long reads sequencing and proximity ligation-based methods is an efficient approach to construct chromosomal level genome assembly. High-throughput chromatin conformation capture (Hi-C) is a technology derived from chromatin conformation capture (3C) technology, which combines chromatin proximity ligation method and high-throughput sequencing to obtain a fine map of chromatin interaction across the whole chromosome [29]. Cross-linked chromatin is cleaved by restriction enzymes and proximity ligated *in situ* to obtain interacting DNA fragments. The ligated DNA fragments are captured by biotin and then sequenced by paired-end sequencing. The interacting DNA fragments, shown as pairs of linked reads, reveal long-distance information about the grouping and linear organization of sequences across the entire chromosomes [30]. The probability of intrachromosomal contacts decays rapidly with linear distance following a power law, but still interacts with a much higher probability with loci on the same chromosome (intra-chromosomal) than loci on different chromosomes (inter-chromosomal), with possible interaction even when separated by over 200 Mb on the same chromosome [31]. Based on the proximity linkage information, Hi-C data can effectively identify linkage between contigs or scaffolds, allowing contigs being linked to nearly whole chromosome-scale.

In the past decade, Hi-C based assembly approaches have become broadly available to generate reliable chromosome-scale *de novo* assemblies of the genome projects in mammals, plants and insects. Hi-C data can also be used to phase genome onto separate haplotypes at chromosome-scale since homologous chromosomes occupy distinct territories in nuclei [32], which could be used to distinguish different haplotypes. Therefore, based on these distinct DNA structural domains, haplotype-aware chromosome-level assemblies using Hi-C based methods have been published and successfully applied on several complex polyploid genomes recently, such as the bread wheat [33], peanut [34] and cotton [35] genomes.

One recently published genome phasing and scaffolding software is FALCON-Phase developed by Phase Genomics [36]. FALCON-Phase is capable of phasing contigs or scaffolds onto high-quality haplotype genome assemblies by integrating long-read sequencing data and Hi-C chromatin contact data of a diploid individual [36]. The pipeline builds upon the aforementioned FALCON-Unzip, which generates phased blocks with reduced haplotype switch errors. FALCON-Phase can address the problem of switches and backfill homozygous regions to produce chromosome-level phased diploid genome assemblies.

However, FALCON-Phase pipeline is designed for phasing diploid genomes and does not yet support the construction of chromosomes for polyploid genomes. To address this limitation, a new pipeline of Hi-C scaffolding named ALLHiC was specifically developed for polyploid genomes [7]. ALLHiC uses a novel pruning step to remove the Hi-C links that are allelic or cross-allelic (linkage between different haplotypes). Such links are typically problematic during phasing since they often prevent the haplotypes from getting partitioned separately. Consequently, ALLHiC enables the accurate phasing of allopolyploid and heterozygous diploid genomes to construct chromosome-scale haplotype assemblies by combining ultra-long-range haplotype information in Hi-C data with a high-quality draft assembly. It has been applied successfully to several chromosome-scale autopolyploid genomes, including an auto-tetraploid [6] and an auto-octoploid sugarcane genomes [7]. In addition to autopolyploid genomes, ALLHiC is applicable to a wide range of genomes including simple diploid, heterozygous and allopolyploid genomes.

6. Experimental methods targeted at haplotype-resolved genome sequencing

The initial Human Genome Project was primarily carried out through the hierarchical sequencing of large-insert clones with long DNA fragments (50–200 kb) inserted in bacterial artificial chromosomes (BACs) [37]. Similarly, fosmids can also be used as DNA vectors as it utilizes F-plasmid origin of replication and partitioning mechanisms to allow cloning of large DNA fragments. Sequencing of fosmid libraries has been successfully applied to highly heterozygous genomes, such as the heterozygous diamond-back moth genome [38]. Fosmid and BAC clones represent a ‘divide-and-conquer’ approach, by breaking the genome into chunks that are much easier to assemble but still retaining the haplotype information since each individual clone represents a single haplotype.

Linked-Reads, a new sequencing technology developed by 10x Genomics, holds the promise to whole genome haplotype phasing at a relatively low cost. This technology leverages microfluidics to partition and barcode high molecular weight DNA molecules, which are then subject to Illumina short-reads sequencing. Paired reads that contain the same barcode are considered to be derived from the same haplotype and linked together during the genome assembly process, which greatly reduces the complexity of genome assembly. Using the Linked-Reads technology, a recent research study on human genomes reveals a comprehensive list of alternative haplotypes in diverse populations [39].

Another related technology, single-tube long fragment read (stLFR) technology developed by Beijing Genomic Institute (BGI), recently showed its potential in haplotyping and *de novo* assembly [40]. Construction of stLFR libraries involves the integration of transposons into long fragments of DNA molecules, which are further mixed with beads containing shared adapters. After the genomic DNA is captured onto the beads, the transposons are ligated to the barcode adapters and then the co-barcoded DNA fragments are subjected to paired end reads sequencing [40].

Clones, linked reads and long fragment reads all provide long-range linkage evidence in the range of hundreds of kilobases to several megabases in length in order to support the phasing of haplotypes – extending well beyond the range that are typically offered by a single WGS read including the latest nanopore reads. Both alignment-based and assembly-based approaches are capable of utilizing these experimental evidence during genome phasing. The popularity of the experimental approaches for phasing would be largely determined by their respective throughput and cost.

7. Summary and outlook

Even though there are a number of existing approaches, both computational as well as experimental, it is still challenging to achieve complete and fully accurate haplotype phasing. Currently, there are two overall strategies for haplotype phasing that we reviewed (Fig. 1). The first class of haplotype phasing methods relies on variant identification from whole genome sequencing reads. The heterozygous variants in a single diploid or polyploid genome are classified into the same block if they are present in the same read or linked read pair. We refer to this class of phasing strategies as alignment-based haplotype phasing, which only reports phased allelic variants that are typically small variants, such as SNPs or small indels. Since reads from larger variants are often missed in the genome alignments, the alignment-based phasing method is not able to recover larger indels or structural variations, such as the human HLA loci or the plant S-locus that are known to be structurally diverse. The second class of haplotype phasing is assembly-based haplotype construction. For this strategy, full-length sequences for individual haplotypes are produced, which are often the preferred phasing method when available due to the higher resolution provided than the alignment-based phasing.

Currently, most state-of-the-art algorithms employ the alignment-based strategy to determine adjacent loci based on Illumina short reads and/or PacBio long reads [8], such as HAPCUT2 [9], WHATSHAP [14] and WinHAP2 [41]. However, phasing with Illumina short reads often produce limited length of phased blocks since the distance of adjacent polymorphic sites can exceed the length of a typical Illumina read or read pair. Although long reads sequencing technology is able to extend haplotype blocks in some cases, it suffers from lots of sequencing errors, leading to high level of switch errors that mix allelic information in the same block.

For the assembly-based strategy, newer generation of genome assemblers such as FALCON-Phase and ALLHiC, are able to separate allelic contigs and produce chromosomal scale of genome assemblies for heterozygous diploid or polyploid genomes [12,7]. However, with the current implementation, we still have a high level of chimeric contig assembly or collapsed sequences that could lead to a large proportion of mis-joined scaffolds or even result in missing chromosomes [7]. At present, neither of the two strategies are perfect and both can cause significant portion of errors during the reconstruction of haplotypes.

Further developments on haplotype phasing and haplotype-resolved genome assembly techniques are needed [42]. Lately, Pacific Biosciences launches the new Sequel II system with much higher throughput than the previous Sequel system. This new system is able to deliver highly accurate individual long reads (HiFi reads), taking advantage of long-read technology with much improved accuracy (over 99.9% accuracy). *De novo* assembly of a human genome using HiFi reads produced a contiguous and accurate genome with a contig N50 of over 15 Mb and consensus accuracy of 99.997% [43]. With the advent of similar new sequencing technologies as well as accompanying advances in genome assembly algorithms, full haplotype resolution of heterozygous genomes and complex polyploid genomes may soon become a routine procedure in genome studies.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (No. 2016YFD0100305 to H.T.), a National Natural Science Foundation of China grant (No. 31701874 to X.Z.), a Natural Science Found Foundation of Fujian grant (No. 2018J01604 to X.Z.) and State Key Laboratory of Ecological Pest Control for Fujian and Taiwan Crops (No. SKL2018001 to X.Z.).

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.csbj.2019.11.011>.

References

- [1] Gaur U, Li K, Mei S, Liu G. Research progress in allele-specific expression and its regulatory mechanisms. *J Appl Genet* 2013;54(3):271–83.
- [2] Knowles DA et al. Allele-specific expression reveals interactions between genetic variation and environment. *Nat Methods* 2017;14(7):699–702.
- [3] Huang X et al. Genomic analysis of hybrid rice varieties reveals numerous superior alleles that contribute to heterosis. *Nat Commun* 2015;6(1):6258.
- [4] Shao L et al. Patterns of genome-wide allele-specific expression in hybrid rice and the implications on the genetic basis of heterosis. *Proc Natl Acad Sci* 2019;116(12):5653–8.
- [5] Yang J et al. Haplotype-resolved sweet potato genome traces back its hexaploidization history. *Nat Plants* 2017;3(9):696–703.
- [6] Zhang J et al. Allele-defined genome of the autopolyploid sugarcane *Saccharum spontaneum* L. *Nat Genet* 2018;50(11):1565–73.
- [7] Zhang X, Zhang S, Zhao Q, Ming R, Tang H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nat Plants* 2019;5(8):833–45.
- [8] Rhee J-K, Li H, Joung J-G, Hwang K-B, Zhang B-T, Shin S-Y. Survey of computational haplotype determination methods for single individual. *Genes Genomics* 2016;38(1):1–12.
- [9] Bonizzoni P, Dondi R, Klau GW, Pirola Y, Pisanti N, Zaccaria S. On the minimum error correction problem for haplotype assembly in diploid and polyploid genomes. *J Comput Biol* 2016;23(9):718–36.
- [10] Duitama J, Huebsch T, McEwen G, Suk E-K, Hoehe MR. ReFHap: a reliable and fast algorithm for single individual haplotyping. In Proceedings of the first ACM international conference on bioinformatics and computational biology - BCB '10, Niagara Falls, New York, 2010, p. 160.
- [11] Xie M, Wang J, Jiang T. A fast and accurate algorithm for single individual haplotyping. *BMC Syst Biol* 2012;6(Suppl 2):S8.
- [12] Matsumoto H, Kiryu H. MixSIH: a mixture model for single individual haplotyping. *BMC Genomics* 2013;14(S2):S5.
- [13] Mazrouee S, Wang W. FastHap: fast and accurate single individual haplotype reconstruction using fuzzy conflict graphs. *Bioinformatics* 2014;30(17):i371–8.
- [14] Patterson M et al. WHATSHAP: weighted haplotype assembly for future-generation sequencing reads. *J Comput Biol* 2015;22(6):498–509.
- [15] Edge P, Bafna V, Bansal V. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Res* 2017;27(5):801–12.
- [16] O'Connell J et al. Haplotype estimation for biobank-scale data sets. *Nat Genet* 2016;48(7):817–20.
- [17] Delaneau O, Zagury J-F, Robinson MR, Marchini J, Dermitzakis E. Integrative haplotype estimation with sub-linear complexity. *Bioinformatics*, preprint, Dec. 2018.
- [18] Rhoads A, Au KF. PacBio sequencing and its applications. *Genomics Proteomics Bioinf* 2015;13(5):278–89.
- [19] Chin C-S et al. Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods* 2016;13(12):1050–4.
- [20] Myers EW. The fragment assembly string graph. *Bioinformatics* 2005;21(Suppl). pp. 2, pp. ii79–ii85.
- [21] Koren S et al. De novo assembly of haplotype-resolved genomes with trio binning. *Nat Biotechnol* 2018;36(12):1174–82.
- [22] Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res* 2017;27(5):722–36.
- [23] Prysacz LP, Gabaldón T. Redundans: an assembly pipeline for highly heterozygous genomes. *Nucleic Acids Res* 2016;44(12). e113–e113.
- [24] Guo L et al. *Echinochloa crus-galli* genome analysis provides insight into its adaptation and invasiveness as a weed. *Nat Commun* 2017;8(1):1031.
- [25] Schoville SD et al. A model species for agricultural pest genomics: the genome of the Colorado potato beetle, *Leptinotarsa decemlineata* (Coleoptera: Chrysomelidae). *Sci Rep* 2018;8(1):1931.
- [26] Huang S et al. HaploMerger: reconstructing allelic relationships for polymorphic diploid genome assemblies. *Genome Res* 2012;22(8):1581–8.

- [27] Huang S, Kang M, Xu A. HaploMerger2: rebuilding both haploid sub-assemblies from high-heterozygosity diploid genome assembly. *Bioinformatics* 2017;33(16):2577–9.
- [28] Roach MJ, Schmidt SA, Borneman AR. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinf* 2018;19(1):460.
- [29] Lieberman-Aiden E et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 2009;326(5950):289–93.
- [30] van Berkum NL, et al. Hi-C: a method to study the three-dimensional architecture of genomes. *J Vis Exp* May 2010;39:1869.
- [31] Schmitt AD, Hu M, Ren B. Genome-wide mapping and analysis of chromosome architecture. *Nat Rev Mol Cell Biol* 2016;17(12):743–55.
- [32] Korb J, Lee C. Genome assembly and haplotyping with Hi-C. *Nat Biotechnol* 2013;31(12):1099–101.
- [33] Maccaferri M et al. Durum wheat genome highlights past domestication signatures and future improvement targets. *Nat Genet* 2019;51(5):885–95.
- [34] Zhuang W et al. The genome of cultivated peanut provides insight into legume karyotypes, polyploid evolution and crop domestication. *Nat Genet* 2019;51(5):865–76.
- [35] Hu Y et al. *Gossypium barbadense* and *Gossypium hirsutum* genomes provide insights into the origin and evolution of allotetraploid cotton. *Nat Genet* 2019;51(4):739–48.
- [36] Kronenberg ZN, et al. Extended haplotype phasing of de novo genome assemblies with FALCON-Phase. *Genomics*, preprint, May 2018.
- [37] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* Feb. 2001;409(6822):860–921.
- [38] You M et al. A heterozygous moth genome provides insights into herbivory and detoxification. *Nat Genet* 2013;45(2):220–5.
- [39] Wong KHY, Levy-Sakin M, Kwok P-Y. De novo human genome assemblies reveal spectrum of alternative haplotypes in diverse populations. *Nat Commun* 2018;9(1):3040.
- [40] Wang O et al. Efficient and unique cobarcoding of second-generation sequencing reads from long DNA molecules enabling cost-effective and accurate sequencing, haplotyping, and de novo assembly. *Genome Res* 2019;29(5):798–808.
- [41] Pan W, Zhao Y, Xu Y, Zhou F. WinHAP2: an extremely fast haplotype phasing program for long genotype sequences. *BMC Bioinf* 2014;15(1):164.
- [42] Schatz MC, Cosgrove A. Graph genomes article collection. *Genome Biol* 2019;20(1).
- [43] Wenger AM et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol* 2019.