International Journal of Computer Science and Information Technology
ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 2, Number 1, Year 2024
DOI: https://doi.org/10.62051/ijcsit.v2n1.49
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Research Status of Haplotype Assembly Technology

Jiayi Wang*

College of School of computer science and technology, Henan Polytechnic University, Jiaozuo, China

ABSTRACT

Haplotype, also called haploid genotype, refers to the combination of a series of variant loci located on the same chromosome. Most of the current assembly algorithms ignore the heterozygous SNP variant sites on the chromosome and assemble a pseudohaplotype sequence. Due to the importance of haplotypes for the treatment of diseases, the mining of pathogenic genes and other medical organisms, haplotype analysis methods have become particularly important. At present, the indirect haplotype inference technology mainly includes two types: the alignment-based haplotyping algorithm and the assembly-based haplotyping method. This paper briefly summarizes the current typical haplotype indirect regression technology, and also looks forward to the future of haplotype typing technology.

KEYWORDS

Haplotypes, Haplotype Indirect Inference Techniques, Alignment-Based Haplotyping Methods, Assembly-Based Haplotyping Methods

1. INTRODUCTION

Haplotypes are combinations of genetic variation loci that reside on the same chromosome[1]. Single Nucleotide Variants (SNVs) refer to the polymorphisms in genes caused by variations of a single nucleotide, and a group of such variants that are linked within a particular region on the same chromosome is known as a haplotype[2].

In diploid and polyploid organisms, homologous chromosomes demonstrate high similarity (with about 99.9% of bases being identical), yet it is the combination of these rare nucleotide variations that contribute to individual differences[3]. These variations may even determine aspects such as life span, physical characteristics, and susceptibility to certain diseases. Haplotypes provide a complete set of genetic information, forming the foundation of individual genomic description and constituting an indispensable part of genomic research. Having information on haplotypes can yield significant insights into various aspects of population genetics, such as the structure, migration, and environmental pressures of populations. At an individual level, haplotype variation data can significantly aid clinical decision-making, which is critical in the study of phenotypes related to individual traits and the specificity of allele expression. The goal of the International HapMap Project is to construct common patterns of polymorphic sites in the human DNA sequence, creating what is known as the HapMap[4]. The HapMap will serve as an important reference for genetic research related to human shape variations, diseases, drug effects, and more.

Contemporary sequencing studies often overlook genomic polymorphism. During the genome assembly process, these single nucleotide variation sites are frequently disregarded as sequencing errors and dealt with through random selection, resulting in a final product that is a pseudo-haplotype sequence containing only half the haplotype information. Although the lost haplotype information

Content from this work may be used under the terms of CC BY-NC 4.0 licence (https://creativecommons.org/licenses/by-nc/4.0/).
Published by Warwick Evans Publishing.

may seem negligible in comparison to the vast expanse of the genome, recent research findings indicate that a thorough understanding of the mysteries of biology's genes necessitates overcoming the current technical barriers to acquire complete haplotype information from the genome[5]. Hence, in the near future, full haplotyping is expected to become a routine procedure in genomic studies.

There are two methods to obtain haplotype information: indirect inference and direct experimental techniques. Due to current technological constraints, the direct experimental method is impractical for individual haplotype research as it demands high-end equipment and is cost-prohibitive. Indirect inference thus becomes the primary means of haplotype acquisition[6]. This article mainly provides an overview of the current state of indirect haplotyping techniques, highlighting the principles and applications of several typical methods. The issues arising from using the indirect inference method to obtain haplotypes, as well as its future prospects, are discussed in the conclusion.

Currently, haplotype assembly methods fall into two categories: alignment-based and assembly-based. Alignment-based methods involve mapping reads to a reference genome to identify single nucleotide variation sites, and then genotyping these variants, essentially addressing the weighted Minimum Error Correction (wMEC) problem[7]. Assembly-based haplotyping, on the other hand, incorporates additional data during the assembly process to genotype the reads containing single nucleotide variations, either by genotyping the bubbles formed in de novo assembly or by directly sorting the reads according to their chromosomal origin before separately assembling them.

Each of these genotyping methods has its advantages and disadvantages, with existing haplotyping technologies making progress in both approaches. The alignment-based haplotyping method does not require additional data and genotypes haplotypes by aligning reads to reference genome sequences. However, it has inherent drawbacks; it can become redundant and less efficient for organisms with high heterozygosity. Assembly-based haplotyping genotypes reads during assembly, using supplemental data to address variations that arise during the process and assigning them to the correct haplotype block. Despite its higher accuracy, faster speed, and independence from a reference genome, this method can be less reliable due to issues like genomic repeats causing incorrect or incomplete genotyping. The need for various supplemental data to genotype haplotypes, which are often difficult or impossible to obtain, poses a significant barrier to assembly-based haplotyping methods.

2. CLASSIFICATION OF HAPLOTYPE ASSEMBLIES

2.1. Alignment-based haplotype typing method

The alignment-based method for haplotyping primarily consists of four steps as illustrated in Figure 1. The initial step involves aligning reads to the reference genome. The second step is the identification of single nucleotide variant (SNV) sites and their locations. The third step involves genotyping these SNVs, grouping SNPs that originate from the same chromosome together. The final step is to either assemble the reads according to their groups or modify the SNPs in the reference genome to produce two haplotype sequences.The process is shown in Figure 1.

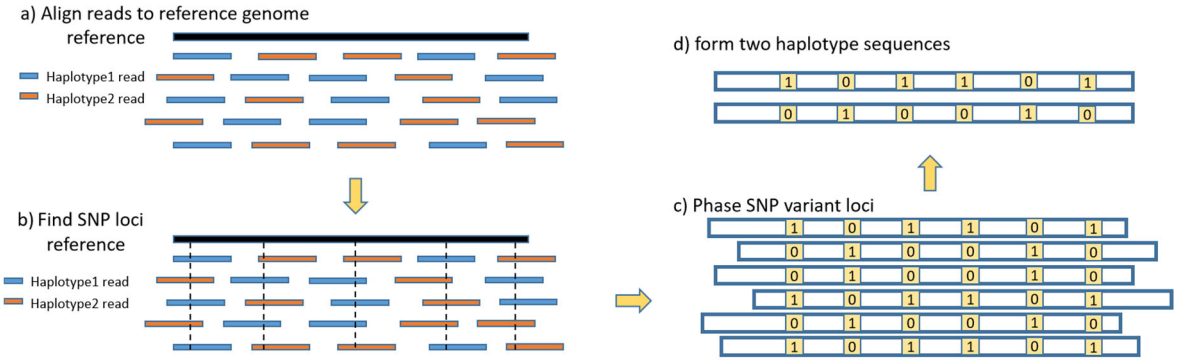

Figure 1 illustrates the alignment-based method for haplotyping. a) Reads are aligned to the reference genome, yielding alignment information against the reference. b) Based on the alignment, single nucleotide variant (SNV) sites are identified, differentiating between sequencing errors and actual SNVs. For instance, with HiFi reads, an SNV is considered genuine if supported by more than three reads; otherwise, it is treated as a sequencing error. c) Phasing is conducted on the SNVs using common phasing techniques such as weighted Minimum Error Correction (wMEC), weighted Minimum Letter Flipping (wMLF), Maximum Fragment Cut (MFC), among other algorithms. d) Utilizing the information from the previous step, the haplotype variants are genotyped, ultimately resulting in two distinct phased haplotype sequences.

Alignment-based haplotyping methods are primarily focused on resolving the phasing of haplotype information within a reference genome. The objective is to accurately classify the haplotype data we obtain into the correct haplotype blocks. Below are some classic alignment-based haplotyping algorithms described.

GenHap is a haplotyping approach that employs genetic algorithms[8]. This method correctly divides reads into two groups using genetic algorithms before assembling them separately in each group. GenHap primarily tackles the Weighted Minimum Error Correction (wMEC) problem. This is an NP-hard problem that involves determining the minimal number of SNV corrections needed to partition sequencing reads into two disjoint haplotype subsets. The Minimum Error Correction (MEC) problem, in the context of haplotype assembly, is regarded as one of the most successful approaches despite being proven NP-hard. Consequently, its weighted variant was developed, the weighted Minimum Error Correction (wMEC) problem, with weights indicating the likelihood of sequencing errors. This consideration uses phred-scaled error rates during the correction process. Here's a brief outline of the steps involved in solving the problem:

Given 'n' variant sites on two homologous chromosomes of a diploid organism and 'm' sequencing reads obtained, each read can be simplified as a sequence 'f' containing elements '0', '1', or '-', where '0' indicates the read matches the reference genome's base at that position, '1' signifies a single nucleotide polymorphism (SNP) against the reference, and '-' indicates the position is not covered by the read. Consequently, all reads can be reduced to an SNV matrix, where each element represents the variation information of the read at that position. Since reads from the same chromosome will have identical haplotyping information, our goal is to categorize reads from the same haplotype into one group. This enables the classification of reads into groups based on identical haplotyping information from each other. The problem ultimately becomes one of read classification, so we can establish an array with elements equal to the number of reads. Each element corresponds to a read, whereby values of 0 and 1 indicate assignments to the first and second partitions, respectively. After partitioning, haplotype synthesis for each partition follows the majority rule approach at each variant site, choosing the haplotype with the most read support.

GenHap utilizes a genetic algorithm to solve this problem. It begins by randomly generating initial populations of read partitions with each individual representing a unique phasing method for the reads. Fitness is determined by the sum of Hamming distances between the reads in a partition and the

haplotype generated from that partition; a larger distance indicates more dispersion and thus a poorer partitioning method. GenHap evolves these partitions through tournament selection and single-point crossover to create offspring, which are then refined through mutation. Iterating this process until the fitness of the best individuals ceases to improve, or when the maximum number of iterations is reached, yields the most effective haplotyping result – the final haplotype sequence.

WhatsHap is another classic alignment-based haplotyping method that leverages the benefits of third-generation sequencing technologies[9]. Due to their length advantage, third-generation sequencing reads can span multiple haplotyping sites, with variations on the same read belonging to the same haplotype, thus aiding in haplotype determination. WhatsHap's uniqueness lies in its combination of read-based phasing and pedigree-based phasing, with the additional provision of multiple related samples data to further enhance its accuracy.

Similar to the GenHap method mentioned earlier, WhatsHap employs the minimum error correction technique by aligning reads to a reference genome to obtain information on single nucleotide variant (SNV) sites, aiming to partition the acquired haplotype variant information such that the error between it and the obtained haplotypes is minimized. For two reads in the haplotype matrix M, they are deemed conflicting if their values in the jth column differ and are not '-', and compatible if they're entirely the same. A haplotype matrix M is considered feasible if all its reads can be divided into two non-intersecting subsets with each subset's reads being compatible. An SNV matrix is feasible when and only when a pair of haplotype sequences can be identified such that any given read in M is compatible with one of the haplotypes, hence deriving the matrix M from these haplotypes. In error-free sequencing, since every row in M invariably comes from one of a pair of chromosomes, there always exists a corresponding pair of haplotypes compatible with the reads. However, as errors are inevitable during sequencing, haplotype assembly becomes a challenge of processing the reads within SNV matrix M by certain rules to find a pair of haplotypes that derive the processed M.

WhatsHap is an exact, fixed-parameter tractable algorithm for the wMEC problem, where the coverage of reads is the sole fixed parameter. This algorithm uses dynamic programming to identify the partitioning of reads into two conflict-free subsets, thereby incurring the minimal cost in terms of weights. The algorithm begins with the variant closest to the 5' end of a chromosome and proceeds towards the 3' end across all variant sites. At each site, it calculates the costs of all 2^k possible bipartitions among k reads covering the site. WhatsHap incorporates Gray code enumeration for bipartitions, allowing the costs to be calculated in constant time, thereby efficiently processing all reads into two conflict-free subsets with minimal correction. The running time of the WhatsHap algorithm is unaffected by the length of reads, making it particularly well-suited for dealing with reads whose lengths continue to increase with future generations of sequencing technologies, unlike other related haplotyping tools.

Beyond GenHap and WhatsHap, there exist numerous highly practical methods for alignment-based haplotyping. Here is a brief introduction to some of them:

1) HapCompass proposes a graph-based algorithm for haplotyping assembly, implementing novel optimization (minimum error correction) techniques[10]. It performs with excellence on datasets such as the 1000 Genomes Project, Pacific Biosciences, and simulated data, using algorithms like the Minimum Error Correction (MEC) and Minimum Weighted Edge Removal (MWER) for local optimization steps.

2) Haptree introduces a probabilistic framework-based haplotyping algorithm and a metric called Relative Likelihood (RL) for assembly quality assessment in polyploid haplotyping, serving as an alternative to minimum error correction[11]. The RL-score outperforms the MEC-score in capturing haplotype assembly quality as polyploidy increases.

3) SDhaP addresses heterozygous issues in data overlooked by most existing haplotyping solutions, proposing a new technique to solve the low-rank semidefinite optimization problems[12]. It offers optimal computational efficiency on both diploid and polyploid levels.

4) H-PoP models the polyploid haplotyping problem as a polyploid balanced optimal partitioning algorithm[13]. For read sequences from a k-ploid genome, this model divides the reads into k groups to minimize intragroup dissimilarities and maximize intergroup distinctions. This algorithm efficiently and accurately processes long reads and high-coverage data, also determining ploidy level in organisms.

5) HapCut2, an extension of the HapCUT program, processes a wide range of sequencing technologies including NGS short reads, long reads, linked reads, and Hi-C reads[14]. Unlike HapCUT, HapCUT2 utilizes a likelihood-based model for specific error estimation and iterative maximal cut computations to search variant subsets in the read haplotyping graph.

6) PolyHarsh introduces two algorithms: a) poly-harsh, based on Gibbs sampling, alternately draws haplotypes and reads to decrease the mismatches between them; b) an effective algorithm to connect haplotype blocks into continuous haplotypes[15].

7) SHAPEIT comprises a series of software aiming to estimate haplotypes based on population-level polymorphism data[16]. The recent SHAPEIT3 speeds up the process with modified Markov chain Monte Carlo (MCMC) sampling algorithms, handling biobank-scale datasets with very low haplotype switch error rates. As an extension, SHAPEIT4 incorporates position-based Burrow-Wheeler Transform (PBWT) to rapidly select informative haplotypes from a reference panel. SHAPEIT4 also exhibits sub-linear scaling with sample size and integrates external phasing information like large reference genomes, a collection of phased variants, and long sequencing reads.

8) Ranbow is designed for short reads, creating haplotype fragments as a reference genome for reads, followed by constructing a graph wherein the haplotype fragments and their overlaps are nodes and edges, used to merge overlapping parts and compute haplotype blocks.

9) Hap10, a haplotyping software package specifically for 10X Genomics linked-reads derived from polyploid genomes[17]. It offers a generic framework based on SDhaP, enabling chromosome-scale haplotyping with a new optimization approach for more accurate haplotypes.

10) Phasebook is a diploid haplotyping method utilizing long-read sequencing data, distinct from traditional alignment-based haplotyping methods[18]. Phasebook first sorts reads by length, extracting the longest reads to merge with overlapping ones to form a reference genome, then repeats this with the next longest reads from the remainder, dividing the reads into clusters. Each cluster undergoes minimum-error correction in parallel, reducing computations and error rates.

The objective of alignment-based haplotyping methods is consistent: to determine haplotype variations from sequencing reads. The variance among existing programs lies in their use of different heuristic algorithms to expedite computational processing. Thus, due to the varied methodologies employed, these programs differ in adaptability to data and requirements for organism ploidy, as well as in their computational efficiency. As an important branch of current haplotyping methodologies, alignment-based haplotyping techniques still have vast potential for improvement, a task awaiting further exploration by researchers in the field.

2.2. Assembly-based haplotyping method

Compared to alignment-based haplotyping methods, which mainly target small variants, assembly-based haplotyping methods are generally more precise and capable of capturing larger and more complex types of genomic variations, such as large insertions, deletions, and structural variations. Current gene assembly methodologies tend to selectively overlook haplotype variations, commonly resolving ambiguities and redundant sites caused by single nucleotide variations by simply choosing

one branch for output. This approach inevitably results in the loss of half the information from heterozygous haplotypes. Modern sequencing technologies that offer long-read capabilities promise overwhelming assembly performance in intricate genomes. Notably, the HiFi reads developed by Pacific Biosciences deliver higher sequencing accuracy, significantly enhancing assembly efficiency and reducing error rates.

Assembly-based haplotyping methods classify haplotype information that arises during the assembly process, assigning it to the correct haplotype groups and ultimately producing two complete haplotype gene sequences. The process is mainly divided into four steps: first, the reads undergo de novo assembly; second, divergences caused by single nucleotide variations, or 'bubbles' in the de novo graph, are identified; third, supplementary data is used to phase these haplotype branches (such data generally consists of reads that can be definitively traced to the same chromosome, which are then aligned to branches within the bubbles for phasing); and fourth, complete haplotype gene sequences are outputted based on this phasing information. Fig. 2 briefly outlines the execution flow of assembly-based haplotyping methods.

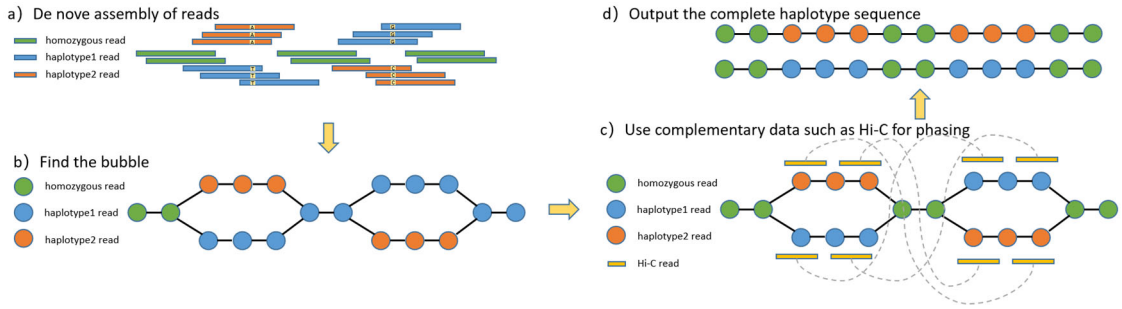

Figure 2 presents an assembly-based haplotyping method as follows. a) Commencing with de novo assembly, one identifies overlapping regions as well as single nucleotide variation (SNV) sites; for instance, in the case of HiFi reads, a position supported by more than three reads is deemed to be an SNV. b) If two reads overlap, an edge is added between them, and bubbles like the ones depicted in the figure emerge at variation sites, representing potential SNVs. c) These bubbles typically have two edges, where the phasing between them is indiscernible, necessitating supplementary data that can phase the information, such as Hi-C reads. Since Hi-C reads are generated based on chromosomal interactions—with both reads coming from the same chromosome—if multiple Hi-C reads' ends match precisely to both sides of a bubble pair, then those edges are considered part of the same haplotype. d) Using the haplotype phasing information obtained, the reads are divided into two separate haplotype gene sequences for output.

The primary challenge that assembly-based haplotyping methods need to address is phasing the bifurcations arising during the assembly process. This is usually solved by aligning reads known to derive from the same chromosome to the two branches of the bifurcation. After phasing, based on this information, reads are divided into two separate haplotypes for output. Below is an introduction to a typical alignment-based haplotyping method.

hifiasm is currently one of the best-performing alignment-based haplotyping methods, utilizing HiFi reads for assembly and accepting three types of supplementary data for phasing[19]. Its main implementation steps are as follows: first, all-versus-all read comparisons are performed to gather overlapping information; second, an overlap graph is constructed, with each read representing a vertex and edges representing overlaps between reads; third, redundant edges are pruned from the graph; fourth, the graph is simplified, and simple paths are merged; fifth, supplementary data, such as Hi-C reads, is aligned to the assembly graph to obtain phasing information between haplotypes; and finally, two separate haplotype gene sequences are outputted based on this information.

Long-read sequencing data have revolutionized genome assembly of simple genomes by resolving repetitive and heterozygous regions, greatly improving assembly accuracy and contiguity. However, assembling complex polyploid genomes with high heterozygosity and repeat content is still

challenging when using CLR or ONT data. HiFi reads, with their high accuracy and long-read lengths, in conjunction with software specifically designed for HiFi data, have a significant advantage in assembling large, complex genomes. For genome assembly of complex polyploid species, employing HiFi reads and hifiasm presents a favorable option.

Besides hifiasm, there are several other assembly-based haplotyping methods, a few of which are briefly described below.

1) FALCON, which sequentially assembles raw sequencing data, continuously identifying diploid genomic information[20]. FALCON constructs an overlap graph from corrected PacBio reads, which often contain many haplotype fusions represented as bubbles in the graph. To resolve these bubbles, FALCON analyzes and decompresses the fused haplotypes, with Hi-C data as supplementary data for phasing, enabling correct allocation of main and alternate contigs to their respective haplotypes, eventually outputting two complete haplotype genome sequences.

2) Redundans aims to assist the assembly of heterozygous genomes, accepting contigs, sequencing reads, and an additional reference genome to ultimately construct a homozygous scaffold assembly[21]. The process involves denovo assembly followed by redundancy removal, contig linking using paired-end reads, and gap filling.

3) HaploMerger2, an upgrade of HaploMerger1, is optimized for highly heterozygous diploids, improving the flexibility of error detection and assembly of haplotypes[22].

4) Trio-binning, which begins by sequencing the parents of the target progeny using second-generation Illumina sequencing and the progeny using third-generation PacBio sequencing, then uses specific k-mers to segregate the reads before assembling separate haplotype gene sequences[23].

5) ALLHiC uses a five-step process for assembling haplotype chromosome genome sequences and tackles the difficulty of assembling highly heterozygous genomes and those of polyploid species with Hi-C data.

6) Strand-seq, a single-cell technique, enables parental-independent phased assembly of the human genome using single-cell strand sequencing and long reads, providing high-accuracy haplotype sequences[24].

7) DipAsm, developed by the same creator as hifiasm, utilizes HiFi and Hi-C data to produce phased contigs, scaffolds, and eventually full haplotype genome sequences by parsing SNV information from HiFi and Hi-C reads[25].

In essence, the goal of assembly-based haplotyping methods is remarkably aligned across various methodologies: to segregate reads based on their origin during assembly, often requiring supplementary data as a basis for phasing. These supplementary datasets, regardless of type, share a common trait of being classifiable by their original chromosomes. Such data allows for "tracing the origins" of at least a pair of reads. With extensive supplementary data, unordered reads are grouped by haplotype, then assembled separately or outputted accordingly, achieving the purpose of haplotyping.

Assembly-based haplotyping is generally unaffected by species heterozygosity because it utilizes supplementary data for phasing, even in highly heterozygous polyploids. However, the downside to this method is the reliance on supplementary data, which can be difficult to obtain, such as parental sequencing data required by Trio binning or single-cell sequencing data needed by Strand-seq. Even when using more mature technologies like Hi-C, the cost can become an unavoidable burden. Hence, in the future, assembly-based haplotyping methods need more streamlined, rapid, affordable, and effective sequencing techniques to facilitate widespread use in phasing reads.

3. CONCLUSIONS

From the foregoing, current haplotyping technologies can be divided into two categories: those based on alignment and those based on assembly. Both aim to obtain haplotype genome sequences, yet they differ in their processes and initial preparations. In practice, the most suitable haplotyping method can be chosen based on specific circumstances. Next, a brief summary of these two methods regarding their preparations and processes follows.

Alignment-based haplotyping primarily utilizes algorithms to segregate the reads we acquire by their haplotype variant information, minimizing conflicts within groups and maximizing conflicts between them. Each group is considered a haplotype, and assembling each independently should yield a complete haplotype genome sequence. This method requires a reference genome to identify single nucleotide variations by alignment but does not need supplementary data. When samples are limited, budgets are tight, parental sequences are difficult to obtain or other sequencing methods are impractical, alignment-based haplotyping without the need for additional data presents the best choice.

The goal of assembly-based haplotyping is to leverage supplementary data, which can confirm the origin of the haplotype, to type the divergences or bifurcations that arise due to single nucleotide variations during assembly. Ambiguities represented as bubbles in the overlap graph emerge from single nucleotide variations during de novo assembly. Standard assembly methods typically choose a random branch to output when encountering these bubbles, resulting in the loss of half the information at heterozygous sites. However, certain sequencing techniques can determine the origin relationship between at least two reads, such as Hi-C data, which is generated based on chromatin interaction mainly within the same chromosome—with a higher interaction frequency indicating closer proximity. Thus, ends of Hi-C reads can be deemed to originate from the same chromosome. By aligning these to de novo assembled bubbles and categorizing the branches by chromosomal origin, separate outputs can be created, forming the sought haplotype genome sequence. This is the essence of assembly-based haplotyping. It typically achieves high accuracy during phasing, but the difficulty in obtaining supplementary data becomes the method's main limitation. However, when funding is sufficient and such data is readily available, assembly-based haplotyping remains the superior choice.

Each method has its pros and cons, and the selection of a particular one in a given experiment should be based on a detailed analysis of the specific situation. Regardless of the chosen method, the ultimate goal is the same—to produce haplotype-level chromosome assemblies, a crucial step in the future of genome research.

4. DISCUSSION

The human genome is composed of approximately 3 billion base pairs, with an average of one variation per thousand base pairs, known as single nucleotide polymorphisms (SNPs). With the successful completion of the Human Genome Project and the International 1000 Genomes Project, research into complex diseases has progressively shifted towards genome-wide association studies (GWAS) based on single nucleotide polymorphisms . Haplotypes play a vital role in modern genetic epidemiological research, particularly in the genetic localization studies of complex diseases, where linkage and association analyses based on haplotypes are more effective than those based on individual SNP sites . Therefore, haplotype inference is an essential step in analyzing many types of genetic variations within the human genome . Due to the limitations of biological gene sequencing technologies, haplotypes are typically inferred from population genotype data using computational methods . Most organisms are polyploids, necessitating multiple genome sequences to allocate the heterozygous SNPs, with the resulting haplotype sequences being crucial for downstream analyses in population genetics .

Existing haplotyping methods, to varying degrees, have preconditions or yield assembly results that may not fully satisfy researchers. However, as the fields of biomedical and other scientific research advance, the analysis of haplotype genomes is becoming increasingly important due to the vast amount of hidden biological information contained within haplotype data. Consequently, the development of haplotyping technology is an inevitable trend. With advancements in gene sequencing technologies and improvements in heuristic algorithms, haplotyping methods regardless of the approach are bound to mature. It is hoped that in the near future, the use of complete haplotype genome sequences for research in biomedical and other scientific fields will become the norm.

REFERENCES

[1] Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A. Finding the missing heritability of complex diseases. Nature 2009, 461(7265): 747-753.

[2] Consortium IH. A second generation human haplotype map of over 3.1 million SNPs. Nature 2007, 449(7164): 851.

[3] Consortium GP. A global reference for human genetic variation. Nature 2015, 526(7571): 68.

[4] Gibbs RA, Belmont JW, Hardenbol P, Willis TD, Yu FL, Yang H, Ch'ang L-Y, Huang W, Liu B, Shen Y. The international HapMap project. 2003.

[5] Delaneau O, Marchini J, Zagury J-F. A linear complexity phasing method for thousands of genomes. Nature methods 2012, 9(2): 179-181.

[6] Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nature Reviews Genetics 2010, 11(7): 499-511.

[7] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England) 2009, 25(14): 1754-1760.

[8] Tangherloni A, Spolaor S, Rundo L, Nobile MS, Liò P. GenHap: A novel computational method based on genetic algorithms for haplotype assembly. 2017.

[9] Martin M, Patterson M, Garg S, Fischer SO, Pisanti N, Klau GW, Schoenhuth A, Marschall T. WhatsHap: fast and accurate read-based phasing. 2016.

[10] Zhang R, Geng Y, Liu J, Zhao Z, Wang J. Abstract 5301: SubHap: An efficient algorithm for reconstructing clonal haplotypes of tumor sample from NGS data. Cancer Research 2018, 78(13 Supplement): 5301-5301.

[11] Berger E, Yorukoglu D, Peng J, Berger B. HapTree: A Novel Bayesian Framework for Single Individual Polyplotyping Using NGS Data. International Conference on Research in Computational Molecular Biology; 2014; 2014. p. e1003502.

[12] Das S, Vikalo H. SDhaP: haplotype assembly for diploids and polyploids via semi-definite programming. BMC genomics 2015, 16: 1-16.

[13] Xie M, Wu Q, Wang J, Jiang T. H-PoP and H-PoPG: heuristic partitioning algorithms for single individual haplotyping of polyploids. Bioinformatics 2016, 32(24): 3735-3744.

[14] Edge P, Bafna V, Bansal V. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. Genome research 2017, 27(5): 801-812.

[15] Schrinner SD, Mari RS, Ebler J, Rautiainen M, Seillier L, Reimer JJ, Usadel B, Marschall T, Klau GW. Haplotype threading: accurate polyploid phasing from long reads. Genome biology 2020, 21: 1-22.

[16] A., Conti, M., H., Bickel. History of Drug Metabolism: Discoveries of the Major Pathways in the 19th Century. Drug Metabolism Reviews 1977.

[17] Majidian S, Kahaei MH, Ridder DD. Hap10: reconstructing accurate and long polyploid haplotypes using linked reads. Cold Spring Harbor Laboratory 2020(1).

[18] Luo X, Kang X, Schoenhuth A. phasebook: haplotype-aware de novo assembly of diploid genomes from long reads. Genome Biology 2021, 22(1).

[19] Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nature Methods 2021, 18(2): 1-6.

[20] Kawamura S, Choe W, Tanaka S, Pandian SR. Development of an ultrahigh speed robot FALCON using wire drive system. IEEE 2010.

[21] Leszek, Pryszcz, Toni, Gabaldón. Redundans: an assembly pipeline for highly heterozygous genomes. Nucleic Acids Research 2016.

[22] Shengfeng, Huang, Zelin, Chen, Guangrui, Huang, Ting, Yu, Ping, Yang. HapIoMerger: Reconstructing allelic relationships for polymorphic diploid genome assemblies. Genome research 2012, 22(8): 1581-1588.

[23] Yen EC, Mccarthy SA, Galarza JA, Generalovic TN, Jiggins CD. A haplotype-resolved, de novo genome assembly for the wood tiger moth (Arctia plantaginis) through trio binning. Cold Spring Harbor Laboratory 2020(8).

[24] Sanders AD, Falconer E, Hills M, Spierings DCJ, Lansdorp PM. Single-cell template strand sequencing by Strand-seq enables the characterization of individual homologs. Nature Protocols 2017, 12(6): 1151-1176.

[25] Garg S, Fungtammasan A, Carroll A, Chou M, Li H. Chromosome-scale, haplotype-resolved assembly of human genomes. Nature Biotechnology 2020: 1-4.