Journal of Computer Science and Artificial Intelligence
ISSN: 3078-8242 (Print), ISSN: 3078-2309 (Online) | Vol. 7, No. 1, 2026

A Review of Haplotype Assembly: Methods and Challenges

Muniu Niu *

School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, Henan, 454000, China

Abstract: Haplotype assembly aims to partition sequencing reads according to their chromosomal origin based on variant information and to reconstruct the allelic configurations along individual chromosomes. Its core significance lies in revealing the true combinations of variants on the same chromosome, thereby enhancing the biological interpretability of genetic information. This paper presents a systematic review of haplotype assembly methods, with a focus on their underlying models and algorithmic strategies. Existing approaches are broadly categorized into reference-guided and de novo methods, and their characteristics are comparatively analyzed. Finally, future research directions are discussed, including the integration of multi-source sequencing data and the development of more robust and efficient models. Overall, this review provides a comprehensive overview of recent advances in haplotype assembly and offers valuable insights for future studies.

Keywords: Haplotype assembly, Reference-guided methods, De novo assembly.

1. Introduction

Haplotype assembly aims to partition sequencing reads according to their chromosomal origin based on variant information and to reconstruct the allelic configurations along individual chromosomes [1]. With the rapid development of High-Throughput Sequencing (HTS) technologies, the cost of genome sequencing has continuously decreased, while sequencing throughput and data quality have significantly improved, enabling genome-scale analyses at the individual level to become a fundamental component of modern life sciences and biomedical research [2]. In diploid and polyploid organisms, genetic variants are not inherited independently but are transmitted in linked combinations along homologous chromosomes [3]. A haplotype refers to the arrangement of alleles at multiple variant loci on a single chromosome, providing a more informative representation of an individual's genetic structure than unphased genotype data. Therefore, haplotype information plays an essential role in population genetics, complex disease studies, evolutionary analysis, and precision medicine.

From a computational perspective, haplotype assembly is typically formulated as a read partitioning and phasing problem, which aims to reconstruct haplotypes from sequencing reads. Depending on whether a reference genome is available, existing methods are generally classified into de novo and reference-guided approaches [4]. De novo methods infer haplotypes solely from overlap relationships and variant information among reads, making them suitable for organisms lacking high-quality reference genomes [5]. In contrast, reference-guided methods first align reads to a reference genome and then leverage variant information to perform read clustering and haplotype reconstruction. Regardless of the strategy, the objective is to obtain long, accurate, and contiguous haplotype sequences while minimizing phasing errors and fragmentation.

Methodologically, existing approaches can be broadly characterized by different computational paradigms, including graph-based models, combinatorial optimization formulations, and clustering-based strategies. These methods typically construct read–variant relationships and transform haplotype assembly into an optimization problem, often evaluated using the Minimum Error Correction (MEC) criterion or related objective functions [6].

With advances in sequencing technologies, data characteristics have become a key factor influencing assembly performance. Second-generation sequencing technologies provide high throughput and accuracy but are limited by short read lengths, which restrict their ability to capture long-range variant associations [7]. Third-generation sequencing technologies generate long reads that improve phasing continuity but exhibit higher error rates [8]. HiFi sequencing offers a favorable balance between read length and accuracy, making it particularly suitable for high-precision haplotype assembly [9]. In addition, Hi-C data provide long-range chromatin contact information that can be integrated with sequencing reads to improve haplotype resolution [10]. Overall, different sequencing technologies exhibit complementary advantages, and integrating multi-source data has become an important direction for improving assembly accuracy and continuity.

Despite significant progress, haplotype assembly still faces several challenges. Sequencing errors are ubiquitous and may introduce false variant signals, leading to incorrect read partitioning. Moreover, read length and coverage distribution significantly affect assembly performance: short reads provide high accuracy but limited connectivity across distant variants, whereas long reads offer improved continuity but introduce higher noise levels. Therefore, balancing read accuracy, length, and coverage remains a key challenge in method design.

From a computational standpoint, haplotype assembly is often modeled as a Minimum Error Correction (MEC) problem, which has been proven to be NP-hard. As sequencing datasets continue to grow in scale, both the number of reads and variant sites impose increasing demands on computational efficiency in terms of time and memory complexity. In addition, in the absence of high-quality reference haplotypes, evaluating the accuracy and completeness of reconstructed haplotypes remains a challenging open problem.

In summary, haplotype assembly methods can be broadly

categorized into reference-guided and de novo approaches, which correspond to different application scenarios and data availability conditions. Although substantial progress has been achieved in both methodological design and theoretical formulation, haplotype assembly still involves trade-offs among accuracy, contiguity, and computational efficiency under complex sequencing conditions. Its NP-hard nature further increases algorithmic difficulty. Therefore, systematic investigation of existing methods is necessary to clarify their strengths, limitations, and applicability. Based on this, this paper provides a comprehensive review of haplotype assembly methods, focusing on model design and optimization strategies, and further discusses key challenges and future research directions.

2. De novo Haplotype Assembly Methods

De novo haplotype assembly methods reconstruct haplotypes without relying on a reference genome, instead utilizing overlap relationships among sequencing reads, variant information, and auxiliary sequencing data, as illustrated in Fig. 1. Such methods are particularly suitable for newly studied species, genomes with abundant structural variations, or scenarios where the reference genome is of low quality. With the rapid development of third-generation sequencing technologies—especially high-accuracy long reads and chromosome conformation capture data—de novo haplotype assembly has gradually become one of the major research directions in this field.

Typically, de novo haplotype assembly methods are based on read overlap relationships, constructing overlap graphs or assembly graphs to represent the connectivity among reads. On this basis, haplotype-specific branching paths are identified within the graph structure. A central challenge for these approaches lies in accurately distinguishing paths originating from different homologous chromosomes within complex graph topologies.

hifiasm is one of the most influential de novo haplotype assembly tools in recent years [11]. This method takes HiFi long reads as input and first constructs a read overlap graph. It then generates an initial assembly graph through operations such as graph simplification, redundant edge pruning, and simple path merging. Based on this graph, hifiasm exploits heterozygous variant information in the reads, treating branching structures in the assembly graph as potential haplotype divergence points, and performs haplotype separation through phasing inference. When combined with Hi-C data or parental information, hifiasm can further improve phasing continuity and produce chromosome-scale haplotype assemblies, achieving excellent performance in genome assembly tasks across multiple species.

DipAsm is a de novo haplotype assembly method designed for diploid genomes. By integrating HiFi reads and Hi-C data, DipAsm first performs an initial de novo assembly to generate contigs. It then identifies SNP sites on the contigs and leverages the long-range linkage information provided by Hi-C reads to assign global phasing across distant variant loci [12].

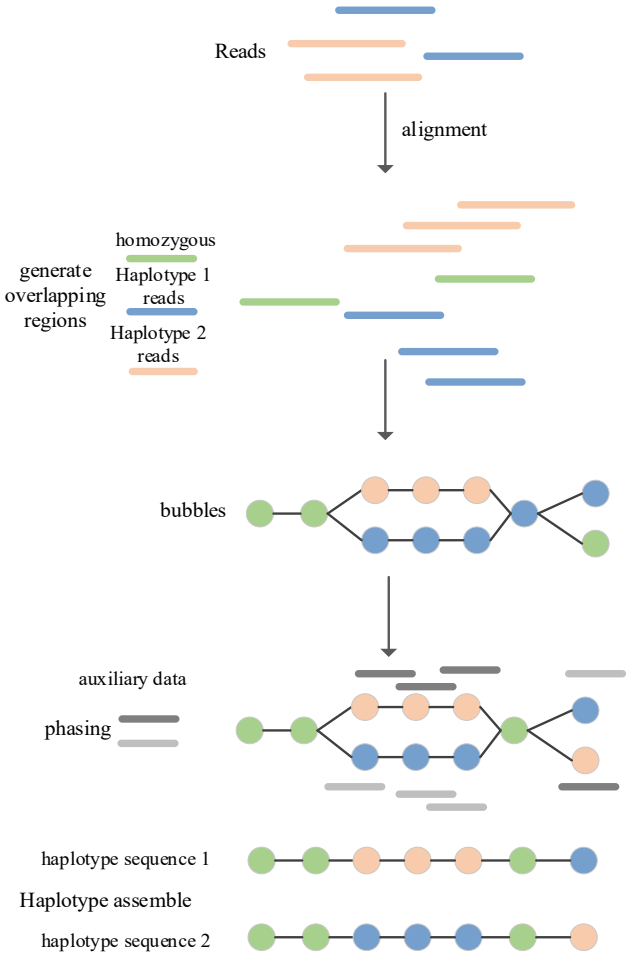

Figure 1. De novo haplotype assembly workflow

Based on the inferred phasing information, DipAsm partitions the contigs into haplotype-specific sequences. Subsequently, through scaffold construction and gap filling, it produces complete haplotype-resolved genome sequences. This method demonstrates strong advantages in terms of phasing continuity and structural consistency.

Another class of de novo haplotype assembly methods focuses on identifying and resolving heterozygous structures within assembly graphs, with FALCON-Unzip being a representative example. This approach first utilizes long reads to generate an initial assembly, producing a set of primary contigs. Subsequently, by analyzing read alignment information, it identifies potential heterozygous regions within the primary contigs and represents them as "bubble" structures in the assembly graph. For each bubble, FALCON-Unzip performs phasing-based read partitioning and resolves it into multiple haplotigs, thereby explicitly representing distinct haplotype sequences [13]. The strength of FALCON-Unzip lies in its ability to capture heterozygous regions while maintaining overall assembly continuity, which has led to its widespread use in early long-read haplotype assembly studies. However, this method typically produces a combination of primary contigs and local haplotigs, and thus still exhibits limitations in terms of haplotype continuity and genome-wide phasing completeness. To balance local phasing accuracy and global assembly continuity, some studies have proposed strategies that integrate local phasing with global assembly.

Phasebook is one of the representative methods in this direction. It first obtains overlap relationships among reads through pairwise alignment and partitions the reads into multiple local groups. Within each group, Phasebook performs local phasing analysis to generate high-quality,

phase-consistent “super-reads.” These super-reads are then treated as new input sequences, and global assembly is carried out using the Overlap-Layout-Consensus (OLC) strategy, ultimately producing two independent haplotype sequences. This “local phasing followed by global assembly” strategy partially mitigates the effects of read noise and uneven coverage on haplotype reconstruction, thereby improving result stability. However, it is relatively computationally expensive and still requires further optimization in terms of scalability to large-scale genomic datasets [14].

ALLHiC proposes a de novo haplotype assembly strategy based on Hi-C data. This method first aligns Hi-C reads to preliminarily assembled contigs and uses inter-chromosomal interaction frequency information to cluster contigs, enabling their assignment to different haplotypes or subgenomes. Subsequently, ALLHiC orders and orients contigs within each cluster to generate haplotype-specific chromosome-scale scaffolds [15]. ALLHiC demonstrates strong applicability in polyploid genome studies; however, it is highly sensitive to the quality and sequencing depth of Hi-C data, and its performance may degrade when high-quality auxiliary data are unavailable.

AsmMix is a hybrid assembly method designed for diploid genomes. It integrates the advantages of third-generation sequencing (TGS) long reads and synthetic long reads (SLR). The method first generates a haplotype-collapsed assembly from TGS data and a haplotype-resolved assembly from SLR data independently [16]. The TGS assembly is then aligned to the SLR assembly to correct errors in the TGS-based contigs. Finally, the SLR-derived haplotypes are combined with the corrected TGS assembly to produce the final haplotype-resolved genome sequences.

Overall, de novo haplotype assembly methods have made significant progress with the support of third-generation sequencing technologies and auxiliary data. However, in scenarios such as polyploid genomes or the coexistence of highly similar haplotypes, challenges such as the lack of direct overlaps between reads and weakly informative variant sites still severely limit the accuracy of read partitioning. Moreover, de novo approaches typically involve complex graph structures and large-scale computations, imposing substantial demands on both time complexity and memory usage. Therefore, while preserving the advantages of de novo assembly strategies, developing more intelligent feature learning mechanisms to effectively capture latent relationships among non-overlapping reads has become an important research direction in this field. This also provides broad opportunities for the application of deep learning-based methods in haplotype assembly.

3. Reference-Guided Haplotype Assembly Methods

Reference-guided haplotype assembly methods represent the earliest developed and most widely applied class of approaches in haplotype assembly research. These methods rely on an existing reference genome and align sequencing reads to the reference sequence to identify the variant loci covered by each read. Haplotype phase relationships are then inferred based on allelic information at these loci, as illustrated in Fig. 2.

Since the reference sequence provides a unified coordinate system for reads, these methods are relatively straightforward in algorithmic design and have been widely applied to organisms with high-quality reference genomes, such as the human genome. Early studies of reference-guided haplotype assembly primarily focused on formulating the problem as probabilistic inference or combinatorial optimization. In this stage, it is typically assumed that variant sites have already been identified through variant calling pipelines, and the main objective is to reconstruct consistent haplotype sequences under sequencing errors and uneven coverage.

HapTree is one representative method in this direction. It constructs a probabilistic graphical model to represent the support relationships between reads and variant sites, and then applies maximum likelihood estimation or Bayesian inference to search for the most probable haplotype configuration [17]. Although HapTree can theoretically identify sequencing errors and produce globally consistent haplotypes, its computational complexity is relatively high, limiting its applicability to large-scale or high-coverage datasets.

H-POP approaches haplotype assembly from a combinatorial optimization perspective by reformulating the problem as a read clustering task [18]. It groups reads that cover the same variant sites and iteratively resolves conflicts among reads, thereby inferring the underlying haplotype structure. H-POP performs well on small- to medium-scale datasets; however, its clustering stability is still challenged in high-noise or highly complex variant regions. With further research progress, graph-based models have gradually become the dominant paradigm in reference-guided haplotype assembly.

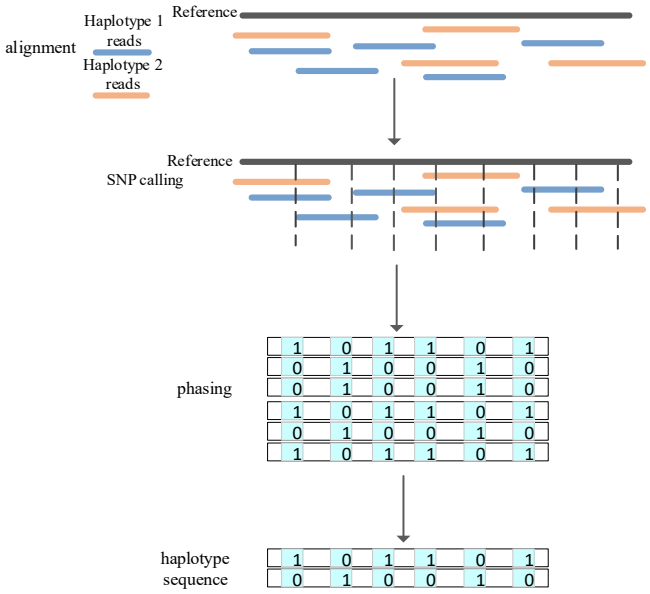

Figure 2. Reference-based haplotype assembly workflow

HapCUT and its improved version HapCUT2 are among the most influential representative tools in this direction. These methods model variant loci as nodes in a graph and exploit reads that simultaneously cover multiple variant sites to construct weighted edges between nodes. The haplotype assembly problem is then formulated as a Maximum Cut (Max-Cut) problem on a weighted graph. By solving this optimization problem, variant sites are partitioned into two disjoint sets, corresponding to the two haplotypes [19].

HapCUT2 extends HapCUT by incorporating support for multiple sequencing platforms and introducing optimizations for long-read and high-error-rate scenarios, leading to improvements in both computational efficiency and phasing accuracy. Owing to its strong performance and general

applicability, HapCUT2 has become a widely used benchmark method in many subsequent studies and practical applications [20].

WhatsHap is a haplotype phasing method based on the weighted Minimum Error Correction (wMEC) model. It exploits read coverage across multiple variant sites and efficiently computes the optimal phasing solution using dynamic programming under bounded coverage constraints. By fully leveraging long-range information from long reads spanning multiple variant sites, WhatsHap achieves higher phasing accuracy and longer haplotype blocks on long-read sequencing data. In addition, the method supports joint analysis of multiple sequencing data types as well as pedigree information, making it highly flexible in practical applications [21].

nPhase is a representative ploidy-agnostic tool in the field of polyploid haplotype assembly [22]. This method integrates short-read and long-read sequencing data, first performing SNP detection using short reads. Long reads are then simplified into feature representations containing only heterozygous SNP information, together with coverage information, which are subsequently used to construct similarity relationships among reads. The core of the method is an iterative clustering algorithm for haplotype partitioning. Specifically, the algorithm is based on read similarity computation. Reads with the highest similarity are first merged into clusters, followed by cluster identity consistency checks to ensure the stability of merged clusters. Consensus sequences are then generated for each cluster to enable efficient inter-cluster comparison. These steps are iteratively repeated until no further valid merging can be performed, ultimately producing clusters corresponding to different haplotypes.

SHAPEIT4 is a haplotype inference tool designed for large-scale genotype datasets and high-coverage sequencing data [23]. It addresses key challenges in large-cohort haplotype analysis, such as low computational efficiency and difficulties in integrating multi-source data, and has become an important benchmark method in subsequent studies and applications in this field. Built upon classical population genetic models, SHAPEIT4 primarily leverages the positional Burrows–Wheeler Transform (PBWT), an efficient data structure that enables rapid haplotype matching and filtering. It can quickly identify long-range haplotype similarity among individuals and dynamically select the most informative haplotypes as references for inference, replacing the traditional strategy of using a fixed number of reference haplotypes. As a result, its computational complexity scales sub-linearly with sample size, meaning that per-sample processing efficiency improves as cohort size increases, making it well-suited for datasets comprising millions of individuals. Another important advantage of SHAPEIT4 is its ability to integrate multi-source phasing information. It supports the joint incorporation of large-scale reference haplotype panels, phasing information from long-read sequencing data, and pre-phased variant sites. For long-read data, local phase blocks are first obtained using tools such as WhatsHap, and then incorporated as probabilistic constraints into the SHAPEIT4 inference framework. This design preserves the long-range contiguity benefits of long reads while correcting sequencing-induced phasing errors using population-level information. Moreover, the integration of reference panels further improves the accuracy of phasing rare variants.

pHapCompass is a haplotype assembly tool applicable to both diploid and polyploid genomes. It addresses core challenges in polyploid phasing, such as ambiguous read assignment and the exponential expansion of the partitioning space, through an innovative graph-based model and probabilistic inference strategy. The method designs two complementary models tailored to different sequencing data types—pHapCompass-short and pHapCompass-long—both of which are reference-based approaches [24]. First, sequencing reads are aligned to a reference genome to identify heterozygous SNP sites, and reads covering at least two heterozygous loci are extracted. Based on these reads and their alignment information to the reference genome, a grouping model is constructed, and finally haplotype sequences consistent with genotype constraints are produced as output.

With the widespread application of deep learning techniques in bioinformatics, some studies have begun to explore the integration of deep neural networks into reference-based haplotype assembly tasks. These methods aim to improve the accuracy of haplotype partitioning by learning latent relationships among sequencing reads, as illustrated in Fig. 3.

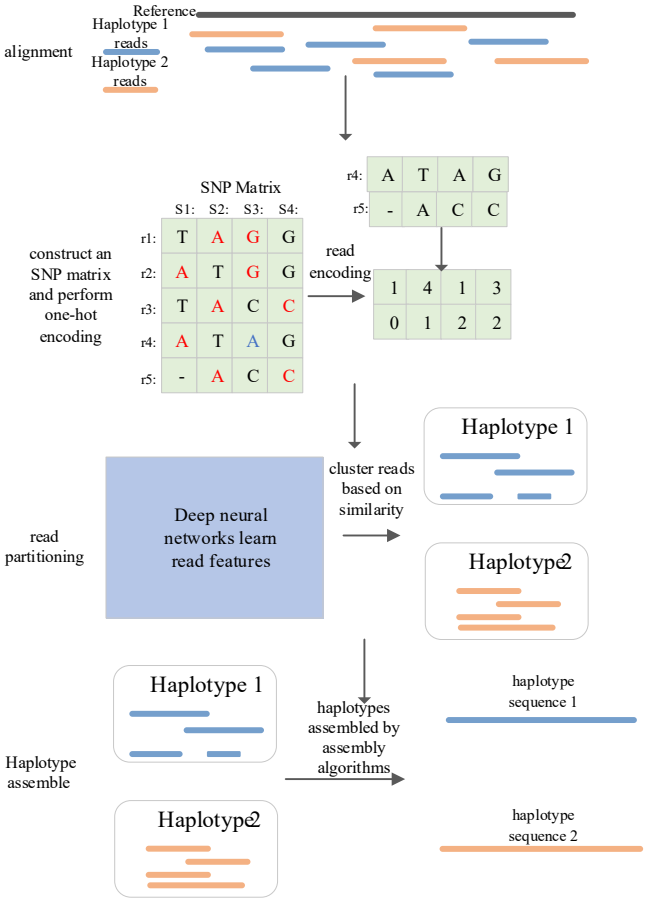

Figure 3. Deep learning-based haplotype assembly workflow

CAECSeq is one of the early representative methods that incorporate deep learning techniques. This method employs a convolutional autoencoder to learn feature representations of reads at variant sites, mapping high-dimensional and sparse read information into a low-dimensional feature space, and then performs read clustering and haplotype reconstruction based on these learned representations. By automatically learning feature embeddings, CAECSeq improves the robustness of the model to noisy data to some extent [25].

XHap introduces a Transformer-based attention

mechanism, which leverages global self-attention to learn long-range dependencies among reads, enabling the model to capture non-local relationships between variant sites [26]. This approach demonstrates good scalability in polyploid and highly complex genomic scenarios, indicating the potential advantages of deep learning models in handling complex haplotype structures.

NeurHap approaches haplotype assembly from a graph learning perspective by formulating the problem as a graph coloring task. It first constructs a read overlap graph and then applies graph neural networks to learn node and edge representations, combined with combinatorial optimization strategies to perform haplotype inference. By incorporating graph structural information, NeurHap achieves relatively stable performance in polyploid haplotype assembly tasks [27].

Ralphi is the first deep reinforcement learning–based tool for diploid haplotype assembly [28]. It innovatively reformulates the haplotype assembly problem as a Maximum Fragment Cut (MFC) optimization task on a fragment graph, providing a novel learning-driven solution for read-based haplotype assembly.

Overall, reference-based haplotype assembly methods rely on the information provided by high-quality reference genomes and infer phase relationships between different haplotypes by analyzing allelic information of sequencing reads at variant sites. These methods are relatively mature in terms of algorithm design and can effectively exploit linkage information across multiple variant loci spanned by reads. The introduction of deep learning techniques enables models to learn latent representations from complex read data, thereby partially mitigating the effects of sequencing noise and data sparsity on haplotype inference.

However, these approaches still have certain limitations in practical applications. Their performance largely depends on the quality of the reference genome; when the reference sequence contains gaps or structural biases, haplotype reconstruction accuracy may be affected. In addition, in complex genomic structures or polyploid scenarios, the conflict relationships among reads become more intricate, and traditional methods still face challenges in terms of phasing continuity and computational efficiency. Therefore, how to fully exploit reference sequence information while incorporating more efficient feature learning strategies to improve haplotype assembly under complex genomic conditions remains an important research direction in this field.

4. Conclusions

This review systematically summarizes current haplotype assembly methods from three main perspectives: de novo assembly methods, reference-guided assembly methods, and emerging deep learning-based approaches. De novo methods reconstruct haplotypes without relying on a reference genome, primarily leveraging read overlap structures, assembly graphs, and auxiliary data. These approaches benefit from advances in long-read and Hi-C sequencing technologies, but still face challenges such as complex graph structures, error propagation, and limited scalability, particularly in polyploid or highly repetitive genomes.

In contrast, reference-guided methods infer haplotypes by aligning sequencing reads to a high-quality reference genome and performing phasing based on variant sites. These methods are relatively mature in algorithmic design and computationally efficient; however, their performance heavily depends on the quality of the reference genome and may degrade in regions with substantial structural variation or complex genomic architectures.

In recent years, advances in deep learning have further expanded the methodological landscape, enabling models to learn data-driven feature representations directly from sequencing reads. These approaches demonstrate advantages in mitigating noise effects, capturing long-range dependencies, and modeling complex interaction patterns. They show promising performance in challenging scenarios such as polyploid genomes and high-error-rate sequencing data. However, they also introduce new challenges in terms of interpretability, generalization capability, and computational cost.

Overall, despite significant progress in this field, haplotype assembly still faces a range of fundamental challenges, including sequencing errors, heterogeneous read lengths, complex genomic structures, and its inherently NP-hard optimization nature. Future research should focus on developing deep learning-based multimodal integration methods for multi-source sequencing data, so as to fully exploit the complementary information across different types of sequencing signals. In addition, it is necessary to design model architectures with higher computational efficiency and stronger representational capacity, in order to improve modeling performance on large-scale genomic datasets and more accurately characterize the latent structure of genomic variation.

5. Discussion

The application of deep learning in haplotype assembly has been increasingly explored, particularly in polyploid genome scenarios, reflecting a transition in this field from rule-based inference methods toward data-driven representation learning paradigms. Although existing studies have made progress in read clustering, phasing accuracy, and robustness to sequencing noise, the application of deep learning to polyploid haplotype assembly remains in an early stage of development. Current approaches typically rely on deep neural networks with a large number of parameters, and their computational and memory requirements still pose significant challenges when applied to ultra-large-scale genomic datasets.

Therefore, there is a need to further develop scalable frameworks capable of efficiently modeling long-range dependencies in large-scale genomic graphs. Existing methods often suffer from quadratic or even higher-order computational complexity when modeling pairwise relationships among reads or genomic fragments. Graph neural networks, sparse attention mechanisms, and hierarchical modeling strategies provide promising directions for capturing global haplotype structures while substantially reducing computational overhead.

In addition, polyploid haplotype resolution typically requires integrating multiple types of sequencing signals, including long reads, Hi-C, optical mapping data, and short reads. Future deep learning frameworks should move toward unified multimodal learning paradigms that jointly encode heterogeneous data sources within a single model, and adaptively weight different modalities according to local genomic context, thereby enabling more accurate and robust haplotype reconstruction.

References

[1] Browning S R, Browning B L. Haplotype phasing: existing methods and new developments [J]. Nature Reviews Genetics, 2011, 12(10): 703-714.

[2] Reuter J A, Spacek D V, Snyder M P. High-throughput sequencing technologies [J]. Molecular cell, 2015, 58(4): 586-597.

[3] Zhang X, Wu R, Wang Y, et al. Unzipping haplotypes in diploid and polyploid genomes [J]. Computational and structural biotechnology journal, 2020, 18: 66-72.

[4] Church D M, Schneider V A, Graves T, et al. Modernizing reference genome assemblies [J]. PLoS biology, 2011, 9(7): e1001091.

[5] Paszkiewicz K, Studholme D J. De novo assembly of short sequence reads [J]. Briefings in bioinformatics, 2010, 11(5): 457-472.

[6] Wang R S, Wu L Y, Li Z P, et al. Haplotype reconstruction from SNP fragments by minimum error correction [J]. Bioinformatics, 2005, 21(10): 2456-2462.

[7] Behjati S, Tarpey P S. What is next generation sequencing? [J]. Archives of disease in childhood-Education & practice edition, 2013, 98(6): 236-238.

[8] Schadt E E, Turner S, Kasarskis A. A window into third-generation sequencing [J]. Human molecular genetics, 2010, 19(R2): R227-R240.

[9] Han Y, He J, Li M, et al. Unlocking the potential of metagenomics with the PacBio high-Fidelity sequencing technology [J]. Microorganisms, 2024, 12(12): 2482.

[10] Forcato M, Nicoletti C, Pal K, et al. Comparison of computational methods for Hi-C data analysis [J]. Nature methods, 2017, 14(7): 679-685.

[11] Cheng H, Concepcion G T, Feng X, et al. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm [J]. Nature methods, 2021, 18(2): 170-175.

[12] Garg S, Fungtammasan A, Carroll A, et al. Chromosome-scale, haplotype-resolved assembly of human genomes [J]. Nature biotechnology, 2021, 39(3): 309-312.

[13] Chin C S, Peluso P, Sedlazeck F J, et al. Phased diploid genome assembly with single-molecule real-time sequencing [J]. Nature methods, 2016, 13(12): 1050-1054.

[14] Luo X, Kang X, Schönhuth A. Phasebook: haplotype-aware de novo assembly of diploid genomes from long reads [J]. Genome biology, 2021, 22(1): 299.

[15] Zhang X, Zhang S, Zhao Q, et al. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data [J]. Nature plants, 2019, 5(8): 833-845.

[16] Wu P, Liu C, Wang O, et al. AsmMix: A pipeline for high quality diploid de novo assembly [J]. bioRxiv, 2021: 2021.01. 15.426893.

[17] Berger E, Yorukoglu D, Peng J, et al. HapTree: a novel Bayesian framework for single individual polyplotyping using NGS data [J]. PLoS computational biology, 2014, 10(3): e1003502.

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

[19] Bansal V, Bafna V. HapCUT: an efficient and accurate algorithm for the haplotype assembly problem [J]. Bioinformatics, 2008, 24(16): i153-i159.

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

[21] Martin M, Patterson M, Garg S, et al. WhatsHap: fast and accurate read-based phasing [J]. BioRxiv, 2016: 085050.

[22] Abou Saada O, Tsouris A, Eberlein C, et al. nPhase: an accurate and contiguous phasing method for polyploids [J]. Genome biology, 2021, 22(1): 126.

[23] Delaneau O, Zagury J F, Robinson M R, et al. Accurate, scalable and integrative haplotype estimation [J]. Nature communications, 2019, 10(1): 5436.

[24] Hosseini M, Veiner E, Bergendahl T, et al. pHapCompass: Probabilistic Assembly and Uncertainty Quantification of Polyploid Haplotype Phase [J]. arXiv preprint arXiv:2512.04393, 2025.

[25] Xue H, Rajan V, Lin Y. Graph coloring via neural networks for haplotype assembly and viral quasispecies reconstruction [J]. Advances in Neural Information Processing Systems, 2022, 35: 30898-30910.

[26] Consul S, Ke Z, Vikalo H. XHap: haplotype assembly using long-distance read correlations learned by transformers [J]. Bioinformatics Advances, 2023, 3(1): vbad169.

[27] Xue H, Rajan V, Lin Y. Graph coloring via neural networks for haplotype assembly and viral quasispecies reconstruction [J]. Advances in Neural Information Processing Systems, 2022, 35: 30898-30910.

[28] Battistella E, Maheshwari A, Ekim B, et al. ralphi: a deep reinforcement learning framework for haplotype assembly [C]//International Conference on Research in Computational Molecular Biology. Cham: Springer Nature Switzerland, 2025: 349-353.