Vol. 12 (2026): Proceedings of the 2026 International Conference on Life Sciences and Health Engineering
(IC-LSHE 2026)
E-ISSN: 3105-0522 P-ISSN：3105-0514
DOI: https://doi.org/10.70267/lshe.20260103

Hi-C Scaffolding Challenges in Non-Model Lepidoptera

Xueting Wang[*]

College of Life Sciences, Tianjin Normal University, Tianjin, China

**Corresponding author: Xueting Wang*

Abstract

The limited availability of high-quality genomic resources for non-model Lepidoptera severely hampers research on their evolutionary adaptations and restricts their application in biodiversity conservation and agroforestry. This study systematically addresses a critical bottleneck-quality degradation in Hi-C scaffolding-that undermines the structural integrity of Lepidoptera genome assemblies. The findings show that intrinsic genomic features, such as high heterozygosity and abundant repetitive sequences, coupled with technical noise in Hi-C data and algorithmic limitations, frequently lead to large-scale structural errors (including misjoins, inversions, and translocations) that are not easily detected by conventional assembly metrics. To address these challenges, the paper proposes an integrated framework that combines haplotype-resolved assembly, manual curation, and multi-dimensional evaluation. This framework integrates optimized experimental and computational workflows, stringent quality control procedures, and standardized evaluation criteria to establish a comprehensive quality assurance system. Additionally, this paper emphasizes the importance of community collaboration and data transparency in fostering reproducible and scalable genomic research. Together, these advances are expected to significantly enhance the reliability and applicability of genomic resources for non-model Lepidoptera, providing a robust genetic foundation for evolutionary biology, comparative genomics and integrated pest management.

Keywords

non-model lepidoptera, Hi-C scaffolding, genome assembly, quality assessment, haplotype phasing

1. Introduction

Non-model lepidopteran species constitute vital components of global biodiversity and provide substantial research value across diverse fields, including agriculture, biomedicine, biomimetic materials, and ecological conservation. Nevertheless, genome assembly for these organisms has historically been impeded by their distinctive genomic characteristics-notably high heterozygosity and abundant repetitive sequences-along with persistent technical barriers. Current genomic research on non-model Lepidoptera is rapidly advancing, with technological innovations shifting the focus from foundational genome sequencing toward comprehensive functional and mechanistic interpretation. For instance, phylogenomic demonstrated that even analyses incorporating hundreds of genes remain constrained by early transcriptomes and draft genomes-characterized by fragmented contigs and compositional bias-hindering resolution of key evolutionary relationships within lineages such as Gelechioidea[1]. These findings underscore the critical necessity for high-quality, chromosome-level genome assemblies. Consequently, the research paradigm is evolving from basic sequence assembly toward an integrated framework combining precise genomic data with multi-omics approaches to elucidate underlying biological mechanisms.

Long-read sequencing technologies, notably PacBio (HiFi and CLR) and Oxford Nanopore Technologies (ONT), coupled with chromatin conformation capture techniques such as Hi-C/Omni-C, have substantially enhanced the efficiency and cost-effectiveness of generating high-quality genome assemblies. These advancements have facilitated pivotal shifts in genomic research: from single-genome analyses to pangenomic investigations elucidating adaptive evolution across populations; from static sequence assembly to dynamic regulatory mapping via integration with transcriptomic and other multi-omics data; and from correlation-based inference toward causal validation utilizing gene-editing technologies like CRISPR/Cas9. Recent systematic evaluations provide comprehensive comparisons of these technologies and their associated bioinformatic workflows [2, 3].

This study addresses critical limitations in genome assemblies for non-model Lepidoptera, focusing specifically on challenges encountered during Hi-C scaffolding. Key issues include interference from tandem repeat arrays, optimization of proximity ligation efficiency parameters and the requirement for manual curation of chromatin interaction maps. To resolve these challenges, this paper proposes an integrated strategy combining long-read sequencing (PacBio HiFi or ONT) with Hi-C scaffolding using established pipelines such as chromap+YaHS or BWA+HapHiC, enhanced by manual refinement using Juicebox. This approach demonstrably improves the completeness and accuracy of genome assemblies for non-model lepidopteran species [4].

2. Challenges and Limitations in Hi-C Scaffolding for Non-Model Lepidoptera Genomes

While long-read sequencing and Hi-C scaffolding have become the “gold standard” for high-quality genome assembly, yielding advanced evaluation metrics such as Scaffold N50 and BUSCO scores, these continuity indicators often obscure substantial structural deficiencies, undermining the reliability of downstream biological analyses [5]. A major concern is that high continuity metrics may coexist with high structural error rates, posing a significant limitation to current assembly evaluation frameworks. For instance, in the benchmarking study, the genome assembled from ONT data achieved an impressive Scaffold N50 of 17.48 Mb and a BUSCO completeness of 94.8%, yet contained 852 structural errors. This discrepancy underscores the inadequacy of conventional metrics in detecting misassemblies related to sequence order and orientation at the chromosomal level.

The persistence of errors-such as misjoins, inversions, and translocations-compromises the integrity of genome-based studies, particularly in comparative genomics and gene family analyses. These issues are especially pronounced in non-model Lepidoptera, due to their high heterozygosity and repetitive content. Manual curation, using tools like Juicebox, and advanced evaluation methods, such as EagleC, are essential for ensuring assembly accuracy, even when standard metrics suggest high continuity and completeness [6].

Achieving accurate chromosome-scale assemblies is primarily hindered by inherent limitations in the Hi-C scaffolding process. First, tandem repetitive regions constitute a major impediment to correct scaffold construction. For instance, Zhang et al.’s effort to assemble the first telomere-to-telomere genome of the silkworm demonstrated that the complex “TTAGG” telomeric repeats prevented the generation of single contigs spanning three chromosomes, necessitating gap filling using external reference sequences. This underscores that highly repetitive regions-such as telomeres and centromeres-disrupt Hi-C interaction signals and frequently lead to chimeric scaffolds.

Second, parameterizing proximity ligation thresholds in scaffolding algorithms entails a fundamental compromise. Walden et al. explicitly illustrated this dilemma: their Omni-C contact maps revealed pervasive off-diagonal signal blocks and anomalous interaction patterns within scaffolds. These artifacts highlight the challenge for algorithms in balancing sensitivity and specificity-excessively lenient parameters induce misjoins, while overly stringent thresholds result in premature chromosomal fragmentation.

These challenges necessitate substantial reliance on manual curation within contemporary genome assembly workflows, highlighting a fundamental limitation in achieving fully automated processing at critical stages. As noted by Zhang et al., “manual correction using Juicebox, while effective for error resolution, lacks standardized protocols." Similarly, Walden et al. emphasized that "most assemblies would benefit from additional manual curation." These observations collectively confirm that human intervention remains indispensable in current scaffolding practices.

A particularly illustrative example is Zhang's comparative analysis, which identified a megabase-scale inversion on chromosome 24 of the previously published P50T-SilkBase reference genome-an error unequivocally revealed by Hi-C interaction maps. This case underscores a critical concern: even publicly accessible reference genomes may harbor undetected large-scale structural errors. Rectifying such inaccuracies still relies on researchers' empirical judgment and laborious manual curation, which not only introduces subjectivity but also impedes the scalable, reproducible utilization of genomic data.

3. Analysis of the Root Causes of Quality Defects and Technical Route Optimization

The assembly of high-quality genomes presents significant challenges for non-model Lepidoptera species, which frequently exhibit extreme genomic characteristics including large genome sizes, high repetitive content, and substantial heterozygosity. Although the integrated approach of PacBio HiFi sequencing with Hi-C scaffolding represents the current "gold standard," the resulting assemblies often contain undetected structural errors that compromise their biological utility. A comprehensive investigation into the origins of these deficiencies is imperative, both for assessing the reliability of extant genomic resources and for informing the design of future sequencing endeavors targeting this ecologically and economically important insect group.

The structural errors in Hi-C scaffolding mentioned above do not exist in isolation, and their root causes can be systematically analyzed from three aspects: intrinsic genomic features, technical data attributes, and algorithm limitations. The primary challenges in Hi-C scaffolding arise from both the intrinsic characteristics of Hi-C data and the limitations inherent in current computational algorithms. Hi-C technology infers linear genomic organization from three-dimensional chromatin proximity data; however, experimental artifacts-including uneven cross-linking efficiency and restriction enzyme bias-significantly compromise the signal-to-noise ratio, obscuring the distinction between genuine chromatin interactions and random collisions.

These data limitations are further amplified by variations in algorithmic approaches. Global partitioning methods, such as 3D-DNA, effectively correct large-scale interchromosomal translocations but frequently fail in highly repetitive regions, exemplified by centromeres in lepidopteran genomes. Signal depletion in these regions often results in assembly breaks or chimeric joins. Conversely, graph-based approaches like SALSA offer enhanced flexibility but exhibit high sensitivity to connection thresholds. Suboptimal parameterization in complex genomes can consequently introduce small-scale inversions or translocations.

Benchmark studies consistently indicate that no single algorithm resolves all structural errors, underscoring the necessity for specialized post-assembly validation tools such as EagleC to correct such errors. This highlights the inherent limitations of fully automated scaffolding pipelines in addressing genomic complexity, particularly for non-model organisms possessing intricate genomic architectures.

Hi-C scaffolding challenges frequently arise from upstream sequencing quality. PacBio HiFi sequencing delivers high accuracy, producing reliable contigs and minimizing structural errors. Conversely, Oxford Nanopore Technologies (ONT) offers longer reads and higher throughput but exhibits higher raw error rates, which can result in uncorrected sequence errors within initial assemblies. When these erroneous contigs are utilized for Hi-C scaffolding, algorithms may inadvertently incorporate these errors into chromosomal structures. Consequently, minor contig errors can escalate into significant chromosomal defects, as demonstrated in case studies.

A key challenge in genome assembly resides in the context-dependent performance of diverse software tools. Systematic benchmarking studies on model species, such as the silkworm, have revealed that even when processing identical datasets, assemblers like HiFiASM and NextDenovo generate contigs exhibiting divergent error profiles. This observation underscores a fundamental challenge: no universally optimal assembler exists, as tool efficacy is heavily contingent upon data type and intrinsic genomic characteristics. The current lack of comprehensive benchmarking across the field compels most non-model genome projects to rely on empirical tool selection. This approach invariably introduces software-specific artifacts, thereby compromising data reliability and reusability.

Current technological approaches present several trade-offs. While the “PacBio HiFi + Hi-C” combination remains the gold standard, its high cost limits large-scale species surveys. Oxford Nanopore technology provides a scalable alternative, but its effectiveness hinges on careful evaluation of the error rate-to-cost ratio. Algorithmically, the context-dependence of tools and the lack of systematic benchmarking present significant bottlenecks in enhancing data quality. Therefore, advancing comprehensive benchmarking initiatives and developing next-generation algorithms capable of integrating multi-source data and automating error correction are crucial steps to overcome current limitations and ensure that genomic resources provide reliable biological insights--This issue suggests that optimizing the technical route requires personalized selection of assembly tools based on the genomic characteristics of the target species, such as heterozygosity and repeat sequence proportion, rather than adopting a one size fits all universal process.

4. An Integrated Approach to Constructing High-Quality Genomes of Non-Model Lepidoptera

The challenges in Hi-C scaffolding of non-model Lepidoptera genomes arise from high heterozygosity, abundant repetitive sequences, experimental noise, and algorithmic limitations. These factors amplify noise levels within chromatin proximity signals, resulting in large-scale structural errors in assemblies. Such inaccuracies are often obscured by high contiguity metrics yet compromise biological validity. Addressing this necessitates a systematic, integrated approach optimizing technological, evaluative, and community standards, thereby advancing Lepidoptera genomics from qualitative assessment towards quantitative precision.

The transition from conventional hybrid genome assembly to haplotype-resolved assembly constitutes a significant methodological advancement in genomic reconstruction. This paradigm shift necessitates a strategic reorientation of assembly strategies, prioritizing high-fidelity long-read sequencing technologies such as PacBio HiFi, in conjunction with advanced phasing-enabled assemblers like hifiasm [7]. These tools enable the precise resolution of highly heterozygous genomes into two complete haplotypes during the initial assembly phase, thereby providing purified input data for subsequent Hi-C scaffolding.

Implementing independent Hi-C scaffolding for each haplotype utilizing specialized workflows-such as chromap coupled with yahs-effectively mitigates allelic interference and substantially reduces the risk of large-scale misassemblies. This approach has been empirically validated in pioneering initiatives, including the Human Pangenome Reference Consortium, establishing novel benchmarks for the production of high-quality, allele-aware reference genomes [8].

The systematic application of this methodology to non-model Lepidoptera species addresses assembly fragmentation and structural inaccuracies induced by high heterozygosity, transitioning genomic data quality from merely “usable” to scientifically “reliable.” Furthermore, the targeted incorporation of characterized telomeric sequences, such as the “TTAGG” repeat motif, during the assembly process enhances chromosomal end completeness, laying a robust foundation for achieving authentic telomere-to-telomere genome assemblies.

Developing a multidimensional quality assessment framework that surpasses conventional metrics is essential for ensuring the credibility of genomic assemblies and is a crucial step toward standardizing the field. While traditional indicators such as Scaffold N50 and BUSCO scores assess assembly contiguity and gene completeness, they exhibit systematic limitations in detecting chromosomal-scale structural errors. Therefore, establishing a more rigorous and comprehensive quality evaluation protocol is critical, incorporating mandatory validation steps, such as Hi-C contact map-based automated assessments using deep learning tools like EagleC, combined with long-read reconciliation analysis. At the same time, chromosome mounting verification can be supplemented with LACHESIS software to further reduce misjudgment rates through multi-tool cross-validation.

Hi-C-based evaluation enables the quantitative and objective identification of cryptic structural variations, such as inversions and translocations, which manifest as characteristic anomalous patterns in the interaction matrix. Simultaneously, long-read reconciliation verifies assembly authenticity at the sequence level by assessing alignment concordance between raw long reads and the assembled genome, providing a complementary evaluation of both accuracy and completeness. This integrated “multi-layered validation” framework should become the new standard for defining “reference-grade” genomes, effectively addressing the current overemphasis on contiguity at the expense of structural accuracy in genomic quality assessment.

To promote sustainable progress across the field, it is essential to establish a more open and collaborative community ecosystem. We propose launching a "Benchmark Dataset" initiative specifically targeting non-model Lepidoptera genomes. This program will systematically collect and curate comprehensive datasets from species representing varying levels of genomic complexity, including HiFi/ONT long-read sequences and Hi-C interaction data. These meticulously curated benchmark resources will provide a transparent and equitable platform for evaluating diverse assembly algorithms and workflows.

Standardized datasets will facilitate the establishment of standardized methodologies and the iterative refinement of computational tools, enabling researchers to evaluate and select approaches based on consistent criteria. Concurrently, we advocate for implementing transparent data release practices, mandating the submission of essential raw data-including Hi-C interaction matrices-alongside publications. This requirement enhances the rigor of peer review and maximizes the long-term utility of datasets,as demonstrated in recent high-quality chromosome-level assemblies such as that of the Atlas moth [9], thereby reducing redundant research efforts and resource expenditure arising from data inaccessibility. This commitment to open collaboration will steer Lepidoptera genomics toward a more reproducible and cumulative research paradigm, establishing a foundation for sustained progress within the field [10].Collectively, the systematic implementation of haplotype-resolved assembly strategies, the establishment of rigorous multidimensional quality assessment protocols, and the promotion of open community collaboration will directly address the core challenges in Hi-C scaffolding for non-model Lepidoptera genomes. This integrated approach will establish a comprehensive genomic framework essential for deciphering the exceptional diversity and evolutionary history of this extensive insect group, ushering in a transformative phase for Lepidoptera genomics.

Future research should prioritize three areas: first, developing more advanced algorithms capable of autonomously detecting and correcting structural errors; second, establishing comprehensive evaluation standards that account for both sequence accuracy and biological validity; and third, fostering broader data sharing through standardized formats and repositories. Collectively, these efforts will contribute to a more efficient and sustainable genomics research ecosystem for Lepidoptera, advancing insights into insect evolution and biodiversity conservation.

5. Conclusion

The assembly of high-quality genomes for non-model Lepidoptera species remains a substantial challenge, with the primary bottleneck arising from quality deficiencies in Hi-C scaffolding. This study rigorously examines the underlying causes of these limitations: the interplay between genomic features-notably elevated heterozygosity and abundant repetitive elements-and technical artifacts inherent in Hi-C data, exacerbated by algorithmic constraints. Collectively, these factors induce pervasive structural errors undetectable by conventional assembly metrics.

To address this multifaceted challenge, we propose an integrated solution comprising three core components: optimized technical workflows incorporating haplotype-resolved assembly strategies, expert-guided manual curation, and multidimensional evaluation frameworks utilizing three-dimensional genomic evidence. This approach establishes a robust quality assurance system for genome assembly projects.

Prospectively, fostering community collaboration through standardized benchmark datasets and promoting data transparency will be imperative for the sustainable advancement of this field. These initiatives will enhance the reliability of genomic resources, accelerate understanding of Lepidoptera evolution and biology, and ultimately advance biodiversity conservation and pest management efforts.

This research delineates a clear technical roadmap and quality standards for non-model Lepidoptera genome studies, propelling the field from merely "obtaining genomes" toward systematically "acquiring high-quality genomic resources." Addressing Hi-C scaffolding quality issues not only bolsters the reliability of comparative genomics and evolutionary biology research but also furnishes a more precise genetic foundation for practical applications, including pest management and biodiversity conservation.

It is crucial to acknowledge current solution limitations, particularly the reliance on expert knowledge during manual curation and the substantial computational demands of haplotype-resolved assembly strategies-the reliance on expert experience for manual calibration directly limits the efficiency of this approach in large-scale genome assembly of non model Lepidoptera species; However, high computing demands have raised the technical threshold for small and medium-sized laboratories, which is not conducive to the popularization of technology. Future advancements in sequencing technologies and algorithmic innovations-especially deeper integration of artificial intelligence into genome assembly pipelines-are anticipated to reduce dependence on manual intervention, enabling more efficient and precise automated assembly processes.

Furthermore, enhancing international collaborations and establishing rigorous quality control frameworks will propel non-model Lepidoptera genomic research into a new phase. Such concerted initiatives will provide enhanced technical support for elucidating the evolutionary mechanisms underpinning insect diversity, thereby enabling the development of more effective conservation and management strategies for this ecologically pivotal group.

References

[1] Kawahara, A. Y., Storer, C., Carvalho, A. P. S., Plotkin, D. M., Condamine, F. L., Braga, M. P., Ellis, E. A., St Laurent, R. A., Li, X., Barve, V., et al. Evolutionary genomics of the Lepidoptera: From silkmoths to pest species. Annual Review of Entomology. 2019, 64, pp. 257-276. https://doi.org/10.1146/annurev-ento-011118-112424.

[2] Zhang, T., Xing, W., Wang, A., Zhang, N., Jia, L., Ma, S. and Xia, Q. Comparison of long-read methods for sequencing and assembly of lepidopteran pest genomes. International Journal of Molecular Sciences. 2023, 24(1), p. 649. https://doi.org/10.3390/ijms24010649.

[3] Walden, K. K. O., Cao, Y., Fields, C. J., Hernandez, A. G., Rendon, G. A., Robinson, G. E., Skinner, R. K., Stein, J. A. and Dietrich, C. H. High-quality genome assemblies for nine non-model North American insect species representing six orders (Insecta: Coleoptera, Diptera, Hemiptera, Hymenoptera, Lepidoptera, Neuroptera). Molecular Ecology Resources. 2024, 24(8), p. e14010. https://doi.org/https://doi.org/10.1111/1755-0998.14010.

[4] Kim, S.-R., Kwak, W., Kim, H., Caetano-Anolles, K., Kim, K.-Y., Kim, S.-B., Choi, K.-H., Kim, S.-W., Hwang, J.-S., Kim, M., et al. Genome sequence of the Japanese oak silk moth, Antheraea yamamai: the first draft genome in the family Saturniidae. GigaScience. 2018, 7(1), p. gix113. https://doi.org/10.1093/gigascience/gix113.

[5] Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., Shamim, M. S., Machol, I., Lander, E. S., Aiden, A. P., et al. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science. 2017, 356(6333), pp. 92-95. https://doi.org/10.1126/science.aal3327.

[6] Wang, X., Luan, Y. and Yue, F. EagleC: A deep-learning framework for detecting a full range of structural variations from bulk and single-cell contact maps. Science Advances. 2022, 8(24), p. eabn9215. https://doi.org/10.1126/sciadv.abn9215.

[7] Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. and Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nature methods. 2021, 18(2), pp. 170-175. https://doi.org/10.1038/s41592-020-01056-5.

[8] Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Fungtammasan, A., Kim, J., et al. Towards complete and error-free genome assemblies of all vertebrate species. Nature. 2021, 592(7856), pp. 737-746. https://doi.org/10.1038/s41586-021-03451-0.

[9] Li, H., Liu, C., Zhang, Y., Wang, X., Xiang, Z., Xia, Q., Miao, X. and Dai, F. A chromosome-level genome assembly of the Atlas moth (*Attacus atlas*) provides insights into its gigantism and conservation. Nature Ecology & Evolution. 2023, 7(9), pp. 1452-1464. https://doi.org/10.1038/s41559-023-02132-7.

[10] Van Oosterhout, C., Breden, F., Hohenlohe, P. A., Garner, B., Hand, B. K., Harrisson, K. A. and Funk, W. C. Genomic erosion in threatened species: A practical guide for conservation planners. Trends in Genetics. 2023, 39(4), pp. 281-297. https://doi.org/10.1016/j.tig.2022.12.007.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Open Access