**Next Generation Genomics:**

**In-course assessment: Quality assessment of a newly sequenced eukaryotic genome**

Li et al (2023): *De novo* assembly of a chromosome-level reference genome of the ornamental butterfly Sericinus montelus based on nanopore sequencing and Hi-C analysis

**Max Falk**

Word Count: \_\_\_\_[[1]](#footnote-1)

# PART 1 – Describing and Critiquing Sequencing & Assembly Methods

## Introduction

The species I have decided to analyse is the *Sericinus* *montelus*, commonly referred to as the ornamental swallowtail butterfly. With a chromosome-level assembly of 581Megabases and an N-50 of 5.74Mb1, the authors claimed this formed the *‘most contiguous reference genome’* for swallow butterflies to date1.

I selected this species as it was the first reported swallow butterfly to be sequenced using Oxford Nanopore long-read technology (ONT). I therefore wanted to analyse the exact sequencing and assembly approach used to determine in which aspects ONT outperformed PacBio and other methods. Additionally, the authors reported a high proportion of repeat elements (48.86%1) in the annotated genome compared to related species and I was interested in how their sequencing approach was curated to tackle this.

## Outlining Overall Technical Approach

To generate a high-quality assembly, the authors used a combination of long & short sequencing methods alongside Hi-C genome scaffolding.

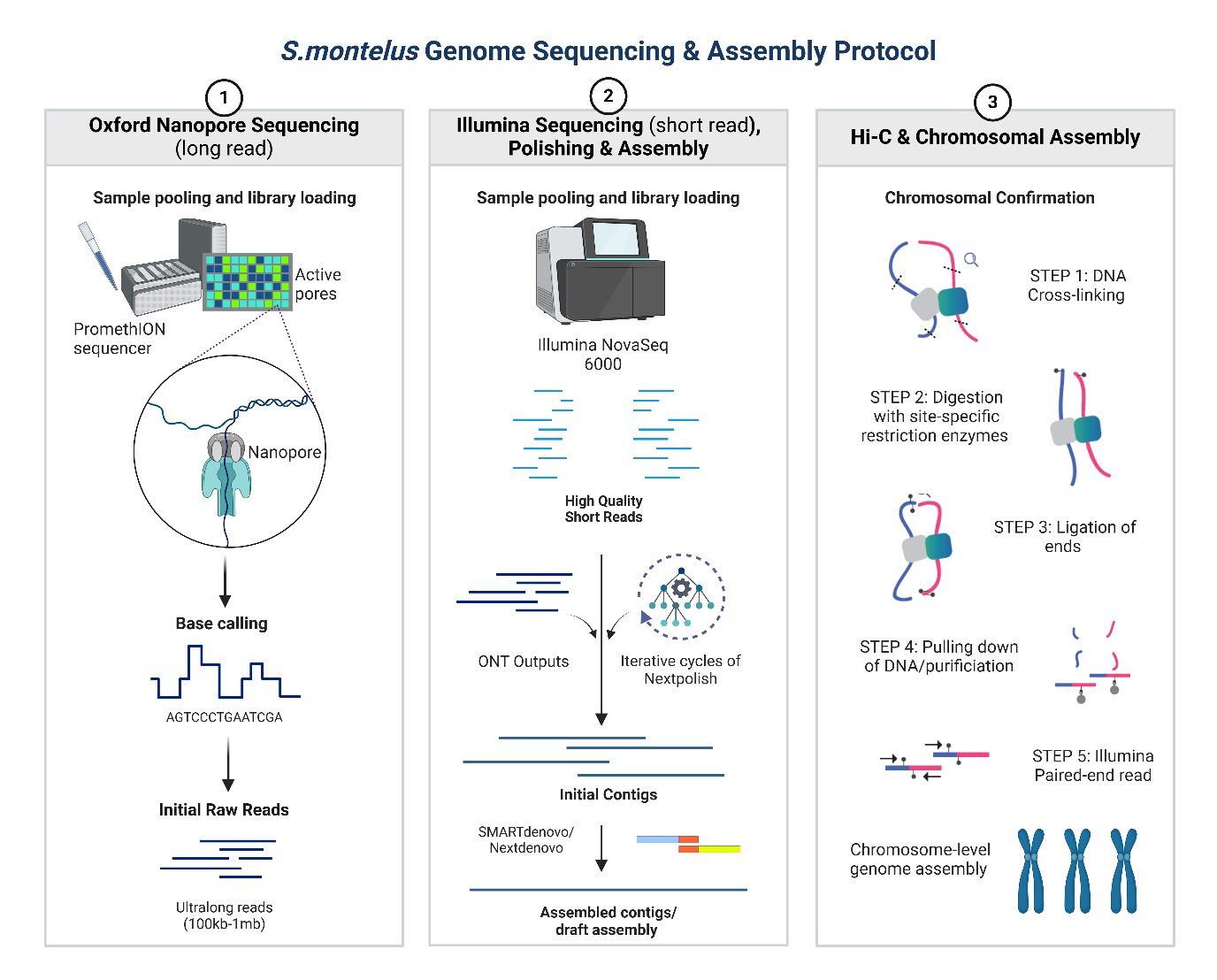
The initial long-read assembly used ONT *PromethION* sequencing, which relies on the traditional ONT technique of threading nucleic acid through a porous, artificial membrane and detecting the precise changes in current that occur when individual bases pass through *(figure 1).* BluePippin was used to ensure only adequately-long fragments (>20kb) were sequenced, and adaptors were ligated using standard kits.

Subsequent base-error correction was done via Illumina short-read sequencing and polishing algorithms. Specifically, they used the Illumina NovaSeq 6000, which deploys bridge PCR amplification of adaptor-bound DNA fragments followed by sequencing by synthesis wherein fluorescently labelled terminator nucleotides are sequentially added to sample clusters and nucleotide-specific emission signals are detected by laser-excitation. Cycles of polishing via Nextpolish algorithms then mapped these short-reads to the long-read signal to generate primary contigs and tidy the inherently erroneous raw ONT signal.

For genome assembly, the authors used a combination of SMARTdenovo and Nextdenovo. SMARTdenovo uses an Overlap-Layout-Consensus (OLC) approach, rapidly building a draft assembly by chaining overlapping reads into contigs. Nextdenovo, however, uses a kmer-based correction method prior to assembly for more accurate, computationally efficient assembly. Additional Illumina short-read data was generated for Hi-C scaffolding (a process that estimates genome proximity by cross-linking 3D interactions) and assembly of contigs into chromosomes.

Assembly quality was assessed with a variety of methods. The authors used a kmer based approach with k=17 for initial genome size estimation, which enabled them to compare genome size to reference genomes from a range of close relatives (such as P.*bianor* 2*).* BUSCO analysis using the insect\_odb10 was performed to assess genome completeness and identify conserved orthologs. Finally, summary metrics for overall quality, including N-50, number of contigs and GC rate were reported.

Overall, this approach combined ultra-long read sequencing with short read and polishing techniques in order to correct inherent error. Hi-C was then applied to scaffold the draft genome into chromosomes for effective assembly prior to gene annotation *(figure 1).*



**Figure 1 - An Outline of the Genome Assembly Protocol for S.montelus**

***1*** *– Generation of Long Reads via ONT Sequencing. I****onic current changes*** *are detected as single-stranded DNA passes through a biological nanopore, generating raw electrical signals. Nucleotide-specific signal changes are used to reconstrue sequences (basecalling), enabling generation of uniquely ultralong reads ranging from 100kb-1mb in length.*

*2 – Three iterative rounds of polishing with NextPolish v1.2.3, incorporating ONT and Illumina short reads to improve base accuracy and continuity. Primary contigs were corrected using NextDenovo v2.5.0 (seed length cutoff: 23 kb) and assembled with SMARTdenovo v1.0.0 (-k 21 -J 3000) to generate a draft assembly.*

*3 – Hi-C sequencing was used to generate chromosome-scale scaffolds. Crosslinked chromatin was digested with restriction enzymes, followed by proximity ligation to join interacting DNA fragments before purification, fragmentation, and sequencing. Valid read pairs were identified and mapped to the assembly, generating an interaction matrix. This was used to calculate interaction frequencies between genomic regions and was processed with LACHESIS to cluster, order, and orient contigs into chromosome-scale scaffolds.*

## Sequencing Methods – A Critical Assessment

ONT is advantageous as it generates *ultra*-long reads and consequently longer contigs and simpler assembly, particularly for more complicated or polypoid genomes. For this reason, it won Nature method of the year, with one authors writing that ONT can ‘*resolve those complex genomic aberrations in cancer that are recalcitrant to Illumina sequencing’* 3*.* ONT is consequently good for resolving repetitive regions and structural variants4, making it suitable here given the reasonably high (48.86%) repeat elements1. Despite this, ONT has an inherently high error rate which may arise from inhomogeneous translocation, faulty segmentation, or issues during sample preparation/signal detection. Larger frame-shift errors may cause significant issues with read-quality, meaning ONT is unsuitable on its own without further polishing. Library prep methods were robust, but further quality control protocols to confirm the distribution of DNA fragments post-size selection would have confirmed the efficacy of these methods.

As mentioned, the authors performed subsequent corrections using Nextpolish algorithms to correct residual base-calling in Nanopore reads. Illumina reads alone are capped at 2 x 150bp due to the issue of phasing, but are renowned for their high accuracy. Combining with polishing algorithms allows them to exploit this feature and *correct* long-reads. Such combination is a tried-and-tested method, such as by Triebel et al who generated a ‘Mycovista’ pipeline for effectively resolving repetitive regions in M.*bovis* 5. The NovaSeq6000 sequencing device used incorporates design elements such as nanowells, exclusion amplification and performing paired end reads, generating the high-quality sequencing data necessary for polishing. Nextpolish is proven to be an efficient algorithmic tool for base-level sequence correction, outperforming older techniques, such as Pilon, in terms of speed and accuracy6,7. Finally, iterative Nextpolish techniques are known to work best with high short-read sequence coverage, which the authors sufficiently generated (~25Gb Illumina short-reads, ~40X coverage)1.

Overall, combining ONT long-reads with short reads is a good method for creating an effective draft assembly. However, there are alternatives the authors could have considered. Modern improvements to ONT protocols have vastly improved its base accuracy, yet there is no evidence in the methods that the authors employed these. For example, Duplex or re-read sequencing, which together are known to reduce long-read sequencing error to Q20+8. Alternatively, PacBio HiFi sequencing could have been used, which reliably produce 99.9% read accuracy due to their adaptor-mediated circular consensus sequence approach. However, reads tend to be much shorter than ONT (~10-25kb) which makes it less suitable for long tandem/highly homologous repeats. Additionally, the faster turnaround time and higher throughput of ONT makes it superior for pan-genome projects with limited budgets or urgent deadlines, which may have been the case here. As this butterfly specific is native to Eastern Asia, it may be difficult to capture and transport samples, which is avoidable with ONT’s portable MinION hardware. Overall, the limited contextual detail in the methods makes it hard to determine if one method would have been preferable, but both have their benefits.

## Assembly Methods – A Critical Assessment

The authors selected SMARTdenovo/Nextdenovo in combination because, despite being overlap consensus (OLC) based methods, they are good for resolving the high error rates coming out of ONT reads. In one study, Liu et al compared SMARTdenovo with an array of other assemblers (Flye, Canu, etc) and compared accuracies. SMARTdenovo and Canu were the most proficient at creating long assemblies with higher coverage, and SMARTdenovo was inferior only to Canu in contig N50 length9. Another study showed that Nextdenovo outperforms in terms of N50 for repetitive genomes, but this was limited only to molluscs10.

If Canu excels in terms of N50, there’s an argument that it, still an OLC-based method, could have been used alone instead, reducing the costs of employing multiple assemblers. However, Canu is computationally inefficient. Additionally, the evidence it provides a longer N50 is weak and is only associated with assembling PacBio reads (not necessarily ONT)9. Canu, despite having in built correction software, also requires additional polishing and would otherwise struggle with noisy ONT reads. Canu might be suitable in combination with PacBio HiFi reads instead (which it’s indeed optimised for), but it depends on genome size, which in this instance is relatively large. Given its computational burden, Canu is best left for smaller genomes (< 1Gb).

The authors could have used a De Bruijn Graph-based (DBG) assembler instead, such as SPAdes, which result in faster assemblies. They struggle, however, with complex repeats and the high error rate of ONT reads. Additionally, they require extremely high-quality short reads in order to produce *reliable* kmers, without which the graphs would be extremely complicated and computationally expensive. Even with the sophisticated polishing steps the authors used, not all errors are fixed (such as Indels), making the choice of an OLC-based approach wise. Even if the authors combined Nextpolish with additional tools (like Medaka or Racon), the costs would be high, and the assemblers would still struggle to resolve repeats. Some modern methods do exist that combine the benefits of DBG and OLC approaches (like MaSuRCA), but these are also resource intensive and likely unnecessary. Other assemblers are more specialised at resolving uncollapsed, diploid genome assemblies, such as Triocanu or HiFiasm. Although HiFiasm is only compatible with PacBio HiFi data, Triocanu might have been beneficial given the reasonably high level of heterozygosity/genome complexity.

The authors finally used Hi-C to scaffold contigs into chromosome-level assemblies. This is an excellent technique for distinguishing repeat regions, given the ability to identify interacting loci in 3D space. Specifically, contigs were split into 50kb interaction matrices and LACHESIS was used for contig orientation. Although this is a standard set up, additional work optimising the matrix bin size for this specific genome would have been beneficial to maximise the Hi-C signal. Additionally, although significant sequence coverage (150X) for Hi-C accuracy was generated, some regions may be underrepresented due to inefficient crosslinking/digestion/PCR amplification bias; a limitation not addressed by the authors. Hi-C alone is also not good at differentiating between maternal and paternal chromosomes and cannot discern structural variations between haplotypes. On top of the many experimental variables could have been optimised, as outlined by Yamaguchi et al11, the authors could have improved their assembly here by selecting more refined scaffolding algorithms, such as ALLHiC or HapHiC12. These are optimised to identify and separate allelic variants and require lower sequence coverage. Finally, the authors could have used optical mapping tools to better resolve large structural variations. Using both technologies in tandem for a ‘hybrid scaffolding’ approach would provide the highest genome-wide scaffolding accuracy, especially for a relatively complex eukaryotic genome.

## Quality Control methods – A Critical Assessment

The kmer-based approach for genome size estimation ensured high quality read signals by comprehensively filtering out of reads with high adaptor content/low Phred scores (using strict thresholds of e.g. >10% Ns) as well as using blastn to detect contamination. The Jellyfish2 and genomescope2 algorithms for counting kmers/analysing kmer distributions respectively are widely respected tools, which, alongside the estimate being similar to sequenced close relatives, adds credibility to this approach. However, the authors decided to remove duplicates from PCR-based amplification which could have accidentally remove true biological replicates in the sample. Specialised algorithms like Picard could have been used here to make this distinction clear. Additionally, the kmer size was modestly selected as 17. Increasing this might overcome the underestimation caused otherwise by repeat regions in the *S.montelus* genome.

A BUSCO score of >95% indicated genome completeness and no major gaps in the assembly. This seems positive, especially as the Nextdenovo + SMARTdenovo assembly methods have been reported to partially compromise BUSCO completeness10. However, the insect\_odb10 database used is optimised for model organisms like Drosophila and many species-specific genes outside the common orthologues may have been missed.

The N-50 score reported is useful but doesn’t say anything about sequencing *quality*. The authors did perform additional coverage assessments, such as via mapping RNA-seq reads back to the assembly, but when making inter-species comparisons, they only evaluated BUSCO/N50. Combining BUSCO reports with RNA-seq outputs as well as other metrics, like Merqury QV50 values/QUAST, would have provided more insights into genome quality, accuracy & misassembly detection. Finally, the authors could have reported NG50 rather than N50 to get a more accurate stat, especially as the genome size estimation protocol was robust and low quality contigs were removed prior to assembly.

# PART 2 – A critical comparison to other published genomes

Altogether, the assembly method assessed above appears robust and thought out, particularly given the complex eukaryotic query genome. Does this approach differ in terms of sequencing approach and assembly method from close relatives, however, particularly those with *lower* repeat content, such as *D.plexippus* (6.21%)?

## A Broad Overview

The percent repeat regions of *S.montelus* was found to be similar to the closely related *K.inachus* and *P.bianor (*49.9% & 55.3% respectively), which had both been sequenced via PacBio. On the other hand, *P.xuthus* and *P.machaon*, which had been sequenced exclusively with Illumina short reads, was reported had much lower repeats (22.4% & 32.3%)1. Interestingly, *P.xuthus* is much closer related to *S.montelus* than *K.inachus* or *P.bianor*, so you would expect a more similar statistic. Instead, this provides more evidence that short-read sequencing is insufficient for effective repeat region characterisation, and additional ONT/PacBio reads are needed to generate longer contigs.

Compared to all previously sequenced de novo butterfly genome assemblies, the quality of *S.montelus* was reported as one of the best in terms of N-50 and BUSCO assessment1. However, I will criticise this statement by referencing other species with different sequencing/assembly methods.

## Species-Specific Comparisons

### *Papilio bianor*

The *P. bianor* (a closely related swallowtail butterfly) genome was sequenced in a similar way, but used PacBio SMRT sequencing alongside Illumina short reads for polishing and Hi-C2. SMRT sequencing is a highly accurate sequencing technique which involves tethering a DNA polymerase to the base of a well (a ‘zero mode waveguide’). Single, fluorophore-conjugated nucleotides freely diffuse into the well and, when incorporated, release a fluorescent pulse upon cleavage of this dye-phosphate group. This allows for individual sequences to be read base-by-base, without the need for error-prone amplification. Interestingly, the authors here used a Wtdbg assembly algorithm, which is a fuzzy DBG approach. This allows for errors, reducing graph complexity, yet it still only produces a single collapsed genome assembly and is therefore poor as identifying haplotypes.

Using this approach resulted in a similar BUSCO completeness to *S.montelus*, but the N-50 was slightly lower (5.5 vs 5.7Mb)2. Additionally, the recovered genome was 421Mb; only 85% of the estimated genome size (496Mb) and a much lower percentage than the S.montelus approach (581Mb assembled, 619Mb estimates, 93.8% recovery) (*Table 1*)1,2.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Organism** | **Sequencing Approach** | **N-50 (Mb)** | **BUSCO (%)** | **Genome Recovery (%)** |
| *S.montelus* | Oxford Nanopore | 5.7 | 99.4 | 93.8 |
| *P.bianor* | PacBio | 5.5 | 96.3 | 84.8 |
| *K.inachus* | PacBio | 3.3 | 97.3 | 99.0 |
| *P.glaucus* | Illumina | 0.2 | 95.5 | 103.1 |

**Table 1 – comparing quality metrics across species and sequencing approach**

N-50 and BUSCO scores were extracted from the relevant publications1,2,13,14.Genome recovery was calculated by taking assembled genome length and divided by estimated genome size.

It could be argued that this PacBio approach is inferior in terms of genome recovery, however the high *K.inachus* score, which also uses SMRT sequencing, contrasts this13. The issue may therefore not be inherent to sequencing method, rather the details of the library prep in this instance which weren’t made available. Alternatively, as the kmer based method is a statistical method based on interpolations from a distribution, these results may be normal fluctuations due to noise. Further statistical assessment would be needed to determine this.

The authors here did do a further circos plot comparison to *P.xuthus* to show that 61 Mb of the *P.bianor* genome could be aligned 1:1 with this reference, suggesting the assembled genome is of high quality. However, the validity of this is uncertain – 61/421Mb does not seem exceptionally high for such a closely related organism and no reference stats are provided.

Overall, it’s difficult to say whether there’s truly any difference in quality statistics when using PacBio vs ONT. However, the N-50 does seem to be lower on average when PacBio is used, particularly in the case of *K.inachus1,13*. It must be said, however, that the *K.inachus* study came 3 years prior to the other species and these differences may just result from more rudimentary versions of the technology.

### *Papilio* *glaucus*

*P.glaucus* was the first of the Papilionoidea (swallowtail butterfly) family to be sequenced, done mainly by paired-end Illumina reads (i.e. 150bp run at both ends of the genome fragments). Given this approach, the resultant N50 value was extremely poor (0.2Mb, *Table 1*) with only 22.2% repeat sequences identified15. Clearly, using short reads on their own is an inferior, albeit cheaper, sequencing method. Further, the authors describe here an extensive inbreeding process prior to sequencing to remove heterozygosity. This ‘genome simplification’ protocol limits the applicability of the process and highlights how limited short reads are at resolving complex genomic problems.

### *Bombyx* *mori*

*B.mori* is a model insect organism, but not a butterfly, making it worth assessing how sequencing approaches might vary from *S.montelus* for insects more broadly. The authors here used a similar combination of PacBio SMRT sequencing and Illumina short reads to assemble reads into primary contigs. Interestingly, they then used Bacterial Artificial Chromosome (BAC) and Fosmid sequences inserted into bacterial cells to resolve gaps and identify repeat regions/misassemblies, rather than Hi-C or optical mapping techniques. This resulted in an enormous N50 of 12.2Mb14. Preparing BAC/Fosmid libraries is very labour intensive however and more useful for genomes with even higher percentages of repeat regions, like plants or mammals. This group were also not performing a de novo assembly so they could use the existing reference genome in tandem with this technique to reduce costs.

# CONCLUSIONS

Long read methods in addition to short read polishing is a tried and tested combination for assembling larger genomes with reasonable heterozygosity3,5. However, for butterflies this had historically been done using PacBio approaches, and the choice in this paper to try ONT was somewhat novel. This provided the authors certain advantages, like higher throughput and lower cost, but came with associated limitations like an inherently higher error rate and limited choice for uncollapsed de novo genome assembly (by excluding HiFiasm, for example).

When comparing to other similar species with similar polishing and assembly protocols, this ONT method appeared slightly superior in terms of N-50 and BUSCO score, with the only exception being *P.xuthus*, for which Illumina sequencing was used yet an exceptionally high N50 of 6.2Mb was reported1. This is the exception rather than the norm, however, as all other methods using 2nd generation NGS tools reported very low N50 scores. One modern paper using PacBio on a more distant butterfly species reported a whopping contig N50 of 23.85Mb16. To make further comparisons to ONT, a paper displaying the most recent ONT technology would need to selected, but it’s clear both technologies are suitable and are evolving at a rapid rate.

Finally, there appears to be no black and white rule for selecting de novo assembly tools. In comparison reports, all appear to do well in terms of genome coverage7,17. In this instance, the tools chosen (Nextdenovo/SMARTdenovo) had been validated by considering both N-50 and coverage depth, a sensible assessment that should be done for each individual sequencing project. Going forward, it would be useful to develop more species-specific comparisons (e.g. Insects) as generally the comparison papers remain broad and deduced applicability is generic7,10.

**Bibliography**

1. Li, J. *et al.* De novo assembly of a chromosome-level reference genome of the ornamental butterfly Sericinus montelus based on nanopore sequencing and Hi-C analysis. *Front. Genet.* **14**, (2023).

2. Lu, S. *et al.* Chromosomal-level reference genome of Chinese peacock butterfly (Papilio bianor) based on third-generation DNA sequencing and Hi-C analysis. *GigaScience* **8**, giz128 (2019).

3. Marx, V. Method of the year: long-read sequencing. *Nat Methods* **20**, 6–11 (2023).

4. Koo, H. *et al.* Two long read-based genome assembly and annotation of polyploidy woody plants, Hibiscus syriacus L. using PacBio and Nanopore platforms. *Sci Data* **10**, 713 (2023).

5. Triebel, S. *et al.* De novo genome assembly resolving repetitive structures enables genomic analysis of 35 European Mycoplasmopsis bovis strains. *BMC Genomics* **24**, 548 (2023).

6. Hu, J., Fan, J., Sun, Z. & Liu, S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* **36**, 2253–2255 (2020).

7. Wang, J. *et al.* Systematic Comparison of the Performances of De Novo Genome Assemblers for Oxford Nanopore Technology Reads From Piroplasm. *Front Cell Infect Microbiol* **11**, 696669 (2021).

8. Wagner, G. E. *et al.* Real-Time Nanopore Q20+ Sequencing Enables Extremely Fast and Accurate Core Genome MLST Typing and Democratizes Access to High-Resolution Bacterial Pathogen Surveillance. *Journal of Clinical Microbiology* **61**, e01631-22 (2023).

9. Liu, H. *et al.* SMARTdenovo: a de novo assembler using long noisy reads. *Gigabyte* **2021**, 1–9 (2021).

10. Sun, J., Li, R., Chen, C., Sigwart, J. D. & Kocot, K. M. Benchmarking Oxford Nanopore read assemblers for high-quality molluscan genomes. *Philos Trans R Soc Lond B Biol Sci* **376**, 20200160.

11. Yamaguchi, K. *et al.* Technical considerations in Hi‐C scaffolding and evaluation of chromosome‐scale genome assemblies. *Mol Ecol* **30**, 5923–5934 (2021).

12. Zeng, X. *et al.* Chromosome-level scaffolding of haplotype-resolved assemblies using Hi-C data without reference genomes. *Nat. Plants* **10**, 1184–1200 (2024).

13. Yang, J. *et al.* Chromosome-level reference genome assembly and gene editing of the dead-leaf butterfly Kallima inachus. *Molecular Ecology Resources* **20**, 1080–1092 (2020).

14. High-quality genome assembly of the silkworm, Bombyx mori. *Insect Biochemistry and Molecular Biology* **107**, 53–62 (2019).

15. Cong, Q., Borek, D., Otwinowski, Z. & Grishin, N. V. Tiger Swallowtail Genome Reveals Mechanisms for Speciation and Caterpillar Chemical Defense. *Cell Rep* **10**, 910–919 (2015).

16. Hu, W. *et al.* Genome assembly of an endemic butterfly (Minois aurata) shed light on the genetic mechanisms underlying ecological adaptation to arid valley habitat. *BMC Genomics* **25**, 1134 (2024).

17. Cosma, B.-M. *et al.* Evaluating long-read de novo assembly tools for eukaryotic genomes: insights and considerations. *GigaScience* **12**, giad100 (2023).

1. Excluding Reference, figure legends & title page [↑](#footnote-ref-1)