**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 who’s de novo genome assembly 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), as opposed to PacBio. Given the authors claims, I therefore wanted to analyse the exact sequencing and assembly approach used to see *for what measured* and *under which conditions* ONT indeed proved superior. Additionally, the authors reported a high proportion of repeat elements (48.86%1) in the annotated genome relative to related species and I was interested in assessing 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 main long-read assembly used the ONT *PromethION* sequencing technique, 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 before using this to reconstrue sequences. This technique enables generation of uniquely ultralong reads ranging from 100kb-1mb in length.

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 were then used to map these short-reads to the long-read signal and tidy up the inherently erroneous ONT draft assembly.

For genome assembly, the authors used a combination of SMARTdenovo and Nextdenovo tools. 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. Together, they enable the rapid composition of a draft which can then be updated and corrected. Additional Illumina short-read data was then used for Hi-C scaffolding (a process that estimates genome proximity by cross-linking 3D interactions and assessing these using an enzyme-based pull-down method) 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 also 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.

## 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. In fact, it one Nature method of the year for this exact reason, with one authors writing that ONT can ‘*resolve those complex genomic aberrations in cancer that are recalcitrant to Illumina sequencing’* 3*.* ONT is consequently also good for resolving repetitive regions and structural variants4, making it suitable here given the S.*montelus* genome consists of 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 error-correction and polishing protocols.

As mentioned, the authors here perform subsequent corrections using Illumina short-reads and Nextpolish algorithms. Although Illumina reads alone are capped at 2 x 250bp due to the issue of phasing, combining with long-read approaches reassigns its role to *correction* and removes this issue. Combining short-reads with ONT 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 modern NovaSeq6000 sequencing device used incorporates design elements such as nanowells, exclusion amplification and performing paired end reads, generating the high-quality sequencing data which 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 and removing the need for traditional two-stage polishing7. Finally, iterative Nextpolish techniques are known to work best with high short-read sequence coverage, which the authors sufficiently generated (~25Gb Illumina short-reads, equalling ~40X coverage)1.

Overall, combining ONT long-reads with short reads is a good method for creating an effective draft genome assembly, despite some of the individual techniques’ shortcomings. However, there are some 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 could have been used, 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. 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 long-read sequencing method would have been preferable. It is worth noting that all previously sequenced close relatives used a PacBio approach, a matter which will be discussed in more detail later.

## Assembly Methods – A Critical Assessment

The authors likely 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, however 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 instead, particularly as using it alone reduces the costs of employing multiple assemblers. However, Canu is infamously computationally inefficient, counteracting the increased speed afforded by the ONT sequencing choice. 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 would perhaps be suitable in combination with PacBio HiFi reads instead (which it’s indeed optimised for), but it really depends on genome size, in this instance which is relatively large. Given its computational burden, Canu is probably best left for smaller genomes (< 1Gb).

The authors could have used a De Bruijn Graph-based (DBG) assembler instead, such as SPAdes, which tend to result in faster assemblies. However, they tend to struggle 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), meaning the choice of an OLC-based approach was wise in this instance. Even if the authors combined Nextpolish with additional tools (like Medaka or Racon), the costs would be extremely high and the assemblers would still struggle to resolve the repeats. Some modern methods do exist that combine the benefits of DBG and OLC approaches (like MaSuRCA), but again these are 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 complex genome and reasonably high level of heterozygosity.

After generating a draft assembly, the authors 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. Hi-C accuracy inherently depends on draft genome assembly quality and sequence coverage. Although significant sequence coverage (150X) was generated here, 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, saving resources. 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 involved comprehensive read filtering out of reads with high adaptor content/low Phred scores as well as blastn to detect contamination. The Jellyfish2 and genomescope2 algorithms for counting kmers/analysing kmer distribitions respectively are widely used and 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. Additionally, the kmer size was selected as 17 (relatively small). Increasing this might provide higher accuracy in size estimations, especially given the high % repeat regions in the S.*montelus* genome.

A BUSCO score of >95% was reported, indicating genome completeness and no major gaps in the assembly. This seems positive, especially as the Nextdenovo + SMARTdenovo assembly methods have been reported to somewhat compromise BUSCO completeness10. However, the insect\_odb10 database used is better 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 RNA-seq, 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

## A Broad Comparison (*rename*)

## Species-Specific Comparisons

### Papilio *bianor*

### Papilio *glaucus*

### Bombyx *mori*

# CONCLUSIONS

**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).

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