School of Biological Sciences

Next Generation Genomics
ICA Essay

Author: B236494

Monday 18th March, 2024

Introduction

This essay will be focusing on the long read sequenced genome assembly of the beautiful Io moth as published in the journal, "G3: Genes, Genomes, Genetics". The paper is "Long read genome assembly of *Automeris io* (Lepidoptera: Saturniidae) an emerging model for the evolution of deimatic displays". [1] These moths belong to the Lepidoptera order, Saturniidae family and *Automeris io* genus according to its taxonomic lineage. Io moths are easily recognisable from the distinct eyespots on their hindwings which are employed as a deimatic defence against predators. This genome assembly could provide insights into deimatic displays wherein the Io moth can be considered as a model organism. Adults in this species, which show sexual dimorphism, have life spans that are no longer than a few weeks and during adulthood they focus on oviposition (egg-laying) and reproduction. The polyphagous Io moth caterpillars are also quite distinguishable by their urticating spines which cause discomfort to human skin and striking green colouration.

Skojec *et al.* were also interested in the annotation of the final curated assembly so that genes linked to patterns, shapes and colours will have a framework to build up on. Future applications of the genome would be related to the study of the evolution of gene functions that relate to antipredatory phenotypes.

Figure 1: *Automeris io* caterpillar

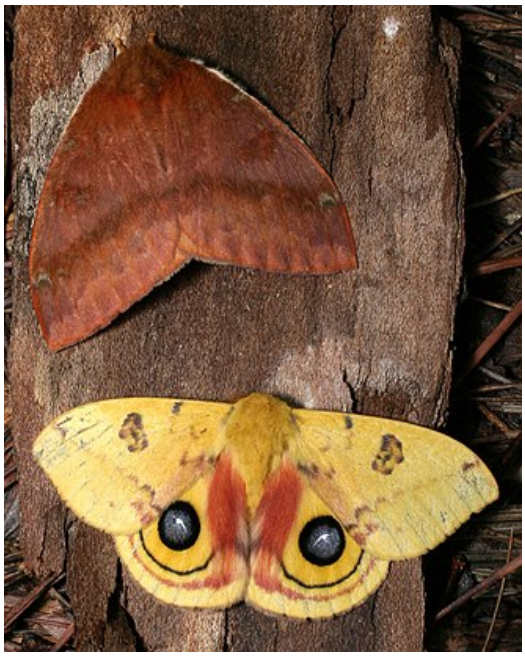

Figure 2: *Automeris io* Adult Moths, Female (above) Male (below)

The first section of the essay will be directed at analysing the technological approaches employed by Skojec *et al.* [1] in assembling this genome. Additionally, the rationale behind the sample collection, library preparation and crucially, the assembly will be discussed. Where applicable, suggestions will also be put forward which could potentially improve the overall assembly.

DNA Isolation and Sequencing

To obtain the DNA that would be used for sequencing, an adult male Io moth was collected and stored at -80 °C. Moths are species that show female heterogamety (female WZ vs male ZZ). The W chromosome is supposedly devoid of protein-coding genes so it is logical to choose a male moth for sequencing considering the future applications as a model organism. [2] The tissue from the thoracic region was extracted using the Qiagen DNeasy Blood and Tissue Kit. There was a slight modification in the isolation protocol to aid in the recovery of high molecular-weight fragments. The samples were gently scraped to avoid clogging the spin column and post homogenisation and digestion, the upper fraction of the spun tissue was used for the DNA isolation steps. The tubes were only inverted to mix the homogenate buffer. Shearing could also have been further prevented using wide bore pipette tips but these were not available in this isolation. I did find a protocol that could improve the quality of the DNA isolated as shown in another species from the Lepidoptera order developed by the International Center for Tropical Agriculture (CIAT). [3]

Skojec *et al.* had decided to sequence the specimen using the Pacific Biosciences (PacBio) Sequel IIe system which is powered by Single Molecule Real Time (SMRT) sequencing technology. They have not described why they decided to go ahead with this particular sequencing technology but coincidentally, it is the same sequencer also available at Edinburgh Genomics and for good reason. This instrument bridges the gap between the highly accurate short reads and comparatively less accurate long reads. It outputs long reads of lengths 10-25 kb and with accuracies greater than 99.5%. These reads are known as HiFi (high fidelity) reads and can be used for a variety of tasks such as the detection of structural and single nucleotide variants, highly repetitive genomic regions and of course, genome assembly. [4] HiFi sequencing has been the provider of data to important projects such as the Telomere-to-Telomere Consortium, Human Pangenome Reference Consortium and the Darwin Tree of Life. This would lead me to believe that they have chosen one of the most recent and advanced sequencing strategies.

Skojec *et al.* estimated the genome size using K-mer counter v.3.2.1 from the PacBio consensus reads. They also visualised a histogram of the k-mer counts for a k-mer size of 23 using the '-m 23' parameter in GenomeScope2.0. [5] I tried to find superior alternatives to these but could not do so.

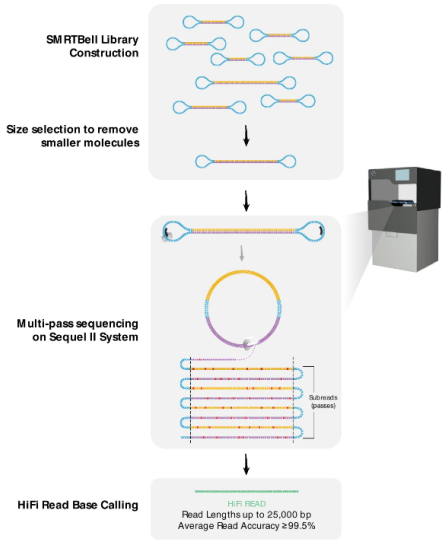

Figure 3: PacBio Highly accurate long-read HiFi read generation [4]

Analysis of Genome Assembly

The assembler of choice to assemble this genome was the high quality de novo assembler, Hifiasm v.0.16.1 r307. [6] The choice of assembler is the most crucial step in a de novo genome assembly. There are several factors to be considered during assembler selection. This includes time, money, computational power, disk space, genome size and genome complexity to name a few. The authors seemed to be favourably placed in most of these departments and have gone with what is considered to be the most advanced assembler with proven superior performance over the other assemblers, notably even with HiFi reads as is the case here. [7] Despite this, it would be good practice to see how it compares to Hicanu in terms of output, to be sure. I will be trying to build my own assembly using the wtdbg2 assembler which is less accurate but very time-efficient using fuzzy-Bruijn graph (FBG) but I don't expect better results than Hifiasm. [8] After further research I came across a paper where the authors state that they have developed new algorithms which show competitive accuracy and computation efficiency compared to currently used software like Hifiasm and Hicanu. [9] They have implemented these algorithms as part of the pre-existing NGSEP software. [10] This integration has led to the development of an easy-to-use tool for analysing long read and short read sequencing data. All things considered, Hifiasm was a good choice.

There were over 1.9 million HiFi reads from the sequencing performed. The mean read length was seen to be 7.2 kb and most reads were between 5-15 kb which seems to be a shorter length of reads than what PacBio advertises for the sequencer used. The authors used GenomeQC to evaluate the contiguity of the assembly. Benchmarking universal single-copy orthologs (BUSCO) v.5.2.0 was used to assess the completeness with putative genes from the lepidopteraodb10.2019-11-20 database. [11] [12] The primary assembly was 500 Mb in size with a reported value of 15.78 Mbs as N50 for 602 contigs. The Phred quality score (Q) of some of the reads, was also seen to be 93 which is the equivalent of 1 error in the base-calling for 2 billion bases which is exceptionally high accuracy. Extensive efforts were made to purge duplicates in the Hifiasm assembler using the 'l 3' option but BUSCO analysis still reported a 6% rate of duplicated orthologs in the primary assembly. It is also worthwhile to mention that BUSCO v5.7 is the latest stable release.

An additional pipeline 'Purge Haplotigs' was used to further collapse allelic variation. This pipeline is tailored solely for third-generation sequencing assemblies. [13] It aids in the manual curation of these assemblies by automating the reassignment of allelic contigs. Haplotig identification and filtering were done as follows. The raw reads were first mapped (aligned) to the above primary assembly using minimap v.2.21. [14] After this a coverage histogram was generated which tells us how frequently different regions were covered by the reads. This was done to select the maximum, median and minimum read depth cutoff values for the purging pipeline. The minimum value is the number of times a region needs to be covered to be deemed reliable. The median value gives us an idea of the typical read-depth coverage and the maximum value is the threshold at which a region is considered to be repetitive or overrepresented. After this the contigs that showed 80% diploid-level coverage were considered as haplotigs, using the 's 80' option, and were removed if the coverage exceeded 80% or was less than the read depth cutoffs, using 'j 80' option. To ensure a well-refined final assembly, this purging step was performed twice. I assume this was done twice due to the size of the genome and the high rate of duplicated orthologs detected by BUSCO in the primary assembly. Ultimately the use of the Purge Haplotigs pipeline was beneficial as it helped reduce the duplicated BUSCOs to 4.7% with only a 1.96% reduction in overall size to 490 Mb in the final assembly. The GC content of 36.3% and N50 value for the final assembly was consistent with other Saturniidae family assemblies

like *Bombyx mori* and *Saturnia povia* (Table 1).

	Automeris io		***Bombyx mori***	***Saturnia povia***
Assembly name	***Curated***	***First draft***	**Bmori 2016v1.0**	**ilSatPavo1.1**
Total length (bps)	490,212,539	500,025,378	460,349,660	489,898,868
N50 (Mbs)	15.78	15.78	16.8	17.68
Contigs	204	602	697	72
GC content	36.3	36.35	36.3	35.8

Table 1: Assembly comparison between *Automeris io* and 2 other related saturniid moths. [1]

Skojec *et al.* say that the uneven taxonomic representation in the repeated element reference database could explain the high 50.36% of repeated elements in the assembly detected by the RepeatModeler 2.0.4 [15] which is above some of the other published lepidoptera genomes. Looking at the supplementary table provided by the authors it is observed that 20.81% of the repeated elements are unclassified. Contamination check results using BlobTools produced positive results. [16] The identified sequences did not match plants or fungi which indicates that the assembly is uncontaminated.

The authors felt that the best chromosome-level assembled genome of choice for a synteny analysis to assess the assembly contiguity was that of the small emperor moth (*Saturnia pavonia*) although it has 3 more chromosomes than the reported 29 of the Io moth. [17] A synteny analysis between genome species is a useful tool to discern the complex evolutionary processes that govern the heterogeneity of chromosome number and structure across different lineages. MUMmer was used to perform the synteny analysis in this paper and high contiguity was seen between them as seen in Supplementary Fig. 3 of the paper. [18] [1]

BUSCO analysis of the final assembly had 98.4% completeness of which 1.7% was missing BUSCOs, 4.7% was duplicated and 93.7% was single-copy. The authors suggest that this high percent of duplication could be due to the heterozygosity which was not successfully collapsed in the final assembly and also refer to the kmer plot coverage which had a high heterozygosity of 1.72%. They say that this could be because they used a wild-caught Io moth. I did not understand this even after viewing the Supplementary Fig. 1. [1] The atypically shorter PacBio Hifi reads also affected the potential to collapse the variation in the assembly across chromosomes.

Skojec *et al.* decided to retain certain areas of high heterozygosity where only the synteny analysis of a chromosome-level assembly with a closer relative of the Io moths would facilitate collapsing those areas confidently. As stated by them, it was a good decision as this prevented the loss of potentially useful information.

Assembly Using wtdbg2 & Additional Analysis

Github Link to View Reports (Click)

In this section I will be attempting to perform my own assembly using the wtdbg2 assembler. [8] The initial QC will be done using the NanoPlot tool from the NanoPack suite for long read processing and analysis. [19] Finally, to assess assembly statistics and quality QUAST will be used. [20]

Note: Every command was prefaced with time during execution to assess how much compute time was required but not shown here. A high thread count of atleast 200 was also used.

The raw reads were downloaded from the Sequence Read Archive using the SRA Toolkit as the total number of bases exceeded 5 Gbases. There were 13.9 Gbases in total. [21]

The command used was:

```
$ fasterq-dump SRR25038556
```

Here SRR2503855 is the run accession number. The final fastq file size was 27 Gigabytes.

Note: Every command has been prefaced with time to assess how much compute time was required. A high thread count of atleast 200 was also used.

The standard procedure with any assembly strategy starts with the quality analysis of the raw read data. I have decided to go with NanoPlot as it is a proven tool and produces the output in quick time. The command used to generate the report is as given below:

The command used to generate the report was:

```
$ NanoPlot -t 220 --verbose --fastq SRR25038556.fastq -o ./
     NanoPlot_output_2/NanoPlot_ICA_genome --loglength --N50
```

It took 16 mins to compute.

Mean read length	7,200.2
Mean read quality	48.9
Median read length	6,631.0
Median read quality	42.8
Number of reads	1,936,546.0
Read length N50	7,228.0
STDEV read length	2,178.0
Total bases	13,943,577,978.0

Figure 4: NanoPlot Report Summary Statistics [19]

The complete report can be downloaded from the GitHub link shared above but the key take-aways from the report were that the reads have a high mean Phred score (Q) of 42.8 and all the reads were having Q¿15. This is typical of the PacBio sequencer used which is a good sign that these reads are suitable for assembly. We can also see that there were 1.9 million reads with a mean read length of 7.2 kb which is consistent with what the authors found as well.

After QC, the assembly is done using the wtdbg2 assembler which assembles raw reads without performing error correction and then builds a consensus from the intermediate assembly output. It first chops the reads into 1024bp segments. It then merges similar segments into a vertex and connects these vertices on the basis of the segment adjacency on reads. The graph which is obtained is known as a fuzzy Bruijn graph (FBG) which bears resemblance to a De Bruijn graph but the critical difference is that FBG allows mismatches/gaps and retains read paths when it collapses k-mers.

This assembler has two main components:

1. The assembler **wtdbg2** and
2. The consenser **wtpoa-cns.**

In practice, the assembler performs the assembly on the raw reads and generates the contig layout which is stored in a file "prefix.ctg.lay.gz". The consenser uses this file to give us the final consensus in FASTA format.

The code that was used for the first step was:

```
$ time wtdbg2 -i SRR25038556.fastq -o ./ICA_wtdbg2_output -g 500m
      -t 200 -x sq
```

This took 101 minutes to complete and generated 3363 contigs. The details of each config is visible in the 'contigs_after_wtdbg2_step_1.txt' file in the GitHub repository. The '-x sq' parameter is the preset that is recommended for reads sequenced by PacBio Sequel machines.

The command which gave us the consensus file was:

```
$ wtpoa-cns -v -t 220 -i ICA_wtdbg2_output.ctg.lay.gz -fo
     ICA_wtdbg2_output.ctg.fa
```

The last step in the assembly pipeline is to assess the assembly quality and I have decided to use the QUAST tool. It gives us several useful, widely used metrics. The command used to generate the QUAST reports was:

```
$ python quast.py --threads 220 --large ../ICA_wtdbg2/
     ICA_wtdbg2_output.ctg.fa
```

The 'large' parameter was chosen because of the size of the genome as recommended by the manual. The entire QUAST report is also available in the GitHub repository.

Looking at the QUAST report we can see that wtdbg2 has not managed to outperform Hifiasm and generated 3363 contigs compared to 600 by Hifiasm's primary assembly. The total length of the assembly is 463 Mbs which indicates that a lot of bases are missing from the assembly and this will result in a huge loss of useful genomic information. It is also to be noted that only 1149 contigs are above 10 kb in length with a N50 value of 1.1 Mbs which is significantly less than the 15.78 Mb of the primary assembly obtained by Skojec *et al* before duplicate purging. The GC content is somewhat similar to what was reported by Hifiasm with a value of 35.9%.

Statistic	Value
# contigs	3362
# contigs (>= 0 bp)	3363
# contigs (>= 1000 bp)	3363
# contigs (>= 5000 bp)	2974
# contigs (>= 10000 bp)	1149
# contigs (>= 25000 bp)	660
# contigs (>= 50000 bp)	614
Largest contig	6059318
Total length	462923864
Total length (>= 0 bp)	462925594
Total length (>= 1000 bp)	462925594
Total length (>= 5000 bp)	461068555
Total length (>= 10000 bp)	448499668
Total length (>= 25000 bp)	441861854
Total length (>= 50000 bp)	440239485
N50	1132852
N90	237633
auN	1449851
L50	121
L90	432
GC (%)	35.9

Table 2: QUAST Statistics without reference for ICA_wtdbg2_output.ctg

I tried to plot a Bandage graph [22] to visualise the assembly but the output was not what I expected it to be. There were no connections between any of the nodes which points to the fact that the assembly achieved is not of good quality.

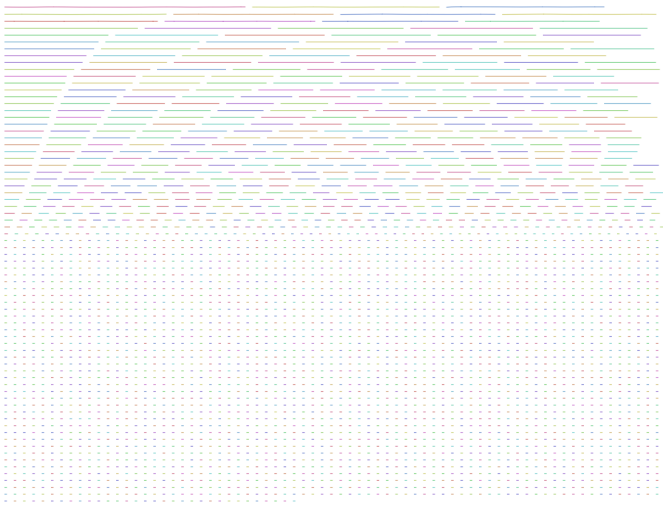

Figure 5: Bandage Graph of wtdbg2 Assembly [22]

In conclusion, Skojec et al. managed to assemble a much better genome from the raw reads compared to the wtdbg2 assembler. Hicanu could have potentially given Hifiasm a challenge but it was not something that I was able to explore in this essay.

References

[1] Chelsea Skojec, R Keating Godfrey, and Akito Y Kawahara. Long read genome assembly of *Automeris io* (*Lepidoptera: Saturniidae*) an emerging model for the evolution of deimatic displays. *G3: Genes, Genomes, Genetics*, 14(3):jkad292, March 2024.

[2] Seong-Ryul Kim, Woori Kwak, Hyaekang Kim, Kelsey Caetano-Anolles, Kee-Young Kim, Su-Bae Kim, Kwang-Ho Choi, Seong-Wan Kim, Jae-Sam Hwang, Minjee Kim, Iksoo Kim, Tae-Won Goo, and Seung-Won Park. Genome sequence of the Japanese oak silk moth, Antheraea yamamai: the first draft genome in the family Saturniidae. *GigaScience*, 7(1):gix113, January 2018.

[3] Diana Victoria Marín, Diana Katherine Castillo, Luis Augusto Becerra López-Lavalle, Jairo Rodríguez Chalarca, and Cristo Rafael Pérez. An optimized high-quality DNA isolation protocol for spodoptera frugiperda J. E. smith (Lepidoptera: Noctuidae). *MethodsX*, 8:101255, February 2021.

[4] Ting Hon, Kristin Mars, Greg Young, Yu-Chih Tsai, Joseph W. Karalius, Jane M. Landolin, Nicholas Maurer, David Kudrna, Michael A. Hardigan, Cynthia C. Steiner, Steven J. Knapp, Doreen Ware, Beth Shapiro, Paul Peluso, and David R. Rank. Highly accurate long-read HiFi sequencing data for five complex genomes. *Scientific Data*, 7(1):399, November 2020. Publisher: Nature Publishing Group.

[5] T. Rhyker Ranallo-Benavidez, Kamil S. Jaron, and Michael C. Schatz. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications*, 11(1):1432, March 2020. Publisher: Nature Publishing Group.

[6] Haoyu Cheng, Gregory T. Concepcion, Xiaowen Feng, Haowen Zhang, and Heng Li. Haplotype resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods*, 18(2):170–175, February 2021. Publisher: Nature Publishing Group.

[7] Yuqiu Wang. A Comparative Study of HiCanu and Hifiasm. In *Proceedings of the 2022 5th International Conference on Mathematics and Statistics*, ICoMS '22, pages 100–104, New York, NY, USA, September 2022. Association for Computing Machinery.

[8] Jue Ruan and Heng Li. Fast and accurate long-read assembly with wtdbg2. *Nature Methods*, 17(2):155–158, February 2020. Publisher: Nature Publishing Group.

[9] Laura Gonzalez-Garcia, David Guevara-Barrientos, Daniela Lozano-Arce, Juanita Gil, Jorge Diaz-Riaño, Erick Duarte, German Andrade, Juan Camilo Bojaca, Maria Camila Hoyos-Sanchez, Christian Chavarro, Natalia Guayazan, Luis Alberto Chica, Maria Camila Buitrago Acosta, Edwin Bautista, Miller Trujillo, and Jorge Duitama. New algorithms for accurate and efficient de novo genome assembly from long DNA sequencing reads. *Life Science Alliance*, 6(5):e202201719, February 2023.

[10] Daniel Tello, Juanita Gil, Cristian D Loaiza, John J Riascos, Nicolas Cardozo, and Jorge Duitama. NGSEP3: accurate variant calling across species and sequencing protocols. *Bioinformatics*, 35(22):4716–4723, November 2019.

[11] Felipe A. Simao, Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19):3210–3212, October 2015.

[12] Mose Manni, Matthew R Berkeley, Mathieu Seppey, Felipe A Simao, and Evgeny M Zdobnov. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution*, 38(10):4647–4654, October 2021.

[13] Michael J. Roach, Simon A. Schmidt, and Anthony R. Borneman. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics*, 19(1):460, November 2018.

[14] Minimap2: pairwise alignment for nucleotide sequences | Bioinformatics | Oxford Academic.

[15] Jullien M. Flynn, Robert Hubley, Clément Goubert, Jeb Rosen, Andrew G. Clark, Cédric Feschotte, and Arian F. Smit. RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences of the United States of America*, 117(17):9451–9457, April 2020.

[16] Dominik R. Laetsch and Mark L. Blaxter. BlobTools: Interrogation of genome assemblies. Technical Report 6:1287, F1000Research, July 2017. Type: article.

[17] Margaret Harris Cook. Spermatogenesis in Lepidoptera. *Proceedings of the Academy of Natural Sciences of Philadelphia*, 62(2):294–327, 1910. Publisher: Academy of Natural Sciences.

[18] Guillaume Marçais, Arthur L. Delcher, Adam M. Phillippy, Rachel Coston, Steven L. Salzberg, and Aleksey Zimin. MUMmer4: A fast and versatile genome alignment system. *PLOS Computational Biology*, 14(1):e1005944, January 2018. Publisher: Public Library of Science.

[19] Wouter De Coster, Svenn D'Hert, Darrin T Schultz, Marc Cruts, and Christine Van Broeckhoven. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics*, 34(15):2666–2669, August 2018.

[20] Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, 29(8):1072–1075, April 2013.

[21] SRR25038556 : Run Browser : SRA Archive : NCBI.

[22] Ryan R. Wick, Mark B. Schultz, Justin Zobel, and Kathryn E. Holt. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics (Oxford, England)*, 31(20):3350–3352, October 2015.