Section 1 Assessing the work of Filipović et al.

This essay focuses on the BMC Genomics paper ‘A high-quality de novo genome assembly based on nanopore sequencing of a wild-caught coconut rhinoceros beetle (*Oryctes rhinoceros*)’ (Filipović *et al.,* 2022), which employed the ONT minION platform to sequence genetic material from a single beetle of and Order which is historically underrepresented in published datasets (Hotaling *et al.,* 2021).

### Sample collection

Assessing the paper’s techniques can start as early as the material collection from the subject species. For the DNA isolation procedures, Filipović et al. sourced tissue from the legs and thorax only rather than the abdomen of the subject beetle. This most likely reduced the potential for downstream contamination from gut microflora, as prokaryotic genetic info would have disrupted the true signals of interest if included (although data on the *O.rhinoceros* microbiome was available at the time and this could have been used to clean data later on (Shelomi, Lin and Liu, 2019). The use of an SPRI bead solution during the sample preparation allowed sample washing and thus sample purification for reduced contamination (*SPRI.* ).

Despite the relatively poor read quality from the experiment (shown in Section 2 of this essay), there appear to have been efforts from the authors to assess and improve it. Electrophoresis was used to assess the quality of DNA prepared in the wet-lab stage of the experiment with the Agilent TapeStation platform which is suited to experiments such as this with 12 or fewer samples (Zascavage *et al.,* 2019).

### ONT library preparations and sequencing

The wet-lab work allowed HMW DNA to be used to create each of the four ONT libraries, accessible at NCBI SRA (NCBI and Tsatsia, 2019) and HMW nucleic acid molecules facilitate ONT sequencing’s long read capabilities. Initially, the motivation for splitting the HWM DNA to generate four libraries on a flow cell and creating four libraries was unclear as they were later merged, however, the benefits of this became apparent: This allowed higher throughput to generate the overall 27 Gigabytes of fastq data more time-efficiently whilst also granting the ability to identify potential technical error from the flow cells, which could have acted as confounding influences on the sequences generated on a particular flow cell. This also provided more reads overall for higher expected coverage and depth.

In this experimental setup, the sequencing yield was ~900,000-1.5 million raw reads across the four flow cells. To my understanding, this is less than the potential yield of a flow cell (around 10 million reads (Srivathsan *et al.,* 2021)) so a repeat analysis might warrant more raw information to be input at the start of the pipeline. This would hopefully improve the depth of assembly later.

### Genome assembly

Table 1-reported statistics of the read data and final assembly

|  |  |
| --- | --- |
| Size of raw reads | 26.4Gb |
| Read length N50 | ~11.3kb |
| Longest read | 143.6kb |
| L50 | 12 (highly contiguous) |

Filipović et al. used Flye to assemble their reads, which is advertised as more accurate than my later used tool wtdbg2 which uses Fuzzy de Bruijn graphs, and therefore expects a lower quality assembly overall but is more time-efficient (Ruan and Li, 2020). Given that their analysis was part of a larger project with a research group, time and efficiency may not have been a required metric of Filipović et al.’s pipeline, whilst, for my work later, wtdbg2 was more suited due to its faster performance. The group used a parameter for Flye specifying the approximate size of the *O. rhinoceros* as 430 Mb based on another species of Orcytes. This is the only other member of the Oryctes genus which has been sequenced until the present, so this was a sensible reference to use, however, retrospectively we see that the length of the O.rhinoceros assembly was far shorter at 377.36 and parameters could be tuned in future repeats of the analysis to make the pipeline more efficient.

Whilst using Flye, Filipović et al. set a parameter for overlap between two reads at a minimum of 5kb. This likely eliminated reads with length<5kb, which I showed to be a significant number of reads in the QC section of my analysis later. Setting a minimum read overlap goes hand in hand with the quality of assembly: this step likely removes erroneous reads which arose in technical faults during base calling by setting a bar for base coverage. This also ensures higher contiguity and higher quality contigs. The 5kb overlap parameter is the default for Flye and could be considered a standard length for minimum overlap, however, the research group also used an iterative assembly procedure, characterised by incrementing the minimum overlap parameter. Increasing the overlap threshold would have exacerbated the benefits and drawbacks mentioned above, so a balanced result was sought. The group determined that the assembly with the highest contiguity possible was generated from the overlap parameter of 7kb.

Filipović et al. worked to identify and remove contaminating genetic information to produce an assembly which represents the *O.rhinoceros* genome accurately. Using assembly visualisation, they identified viral and mitochondrial genome contamination allowing them to disqualify the responsible reads from the assembly, specifically, the mitochondrial genome was identified by searching for a typical circular node in a visualisation of a draft assembly. A node with the traits of around 20kb with over 10,000X coverage was analysed in blastn to check against records of beetle mitochondria and was confirmed as a mitogenome. Using a similar process, a virus which was historically introduced to control the *O.rhinoceros* population was identified due to its coverage exceeding 1000X in the reads.

To highlight other potential contaminating genomes, the group used BLASTx to identify protein-coding sequences in the contigs which have not been recorded in Arthropods, relying on the metagenome tool MEGAN’s annotation. They cross-referenced subject sequences to NCBI’s non-redundant databases to confirm the nature of these contaminating protein-coding regions and they were removed from downstream analysis. If a sequence was not found in any non-redundant record by BLAST, there was no evidence to suggest that it belonged to a contaminating organism and was therefore included in further iterations of the assembly.

Thorough procedures were used to polish the genome. Paired-end short-read data was used to improve genome quality (around 39.4Gb at length 150bp). The approach of adding short-read, high-quality Illumina sequencing reads to cover areas of non-contiguity and highly repetitive regions is a well-evidenced method of improving the quality of long-read genome assemblies. However, as stated by the authors, the Illumina WGS DNA data sequencing data were not derived from the individual beetle which supplied the ONT data and this limited the use of the short-read data to indel polishing instead of covering gaps between contigs or SNP adjustment (Filipović *et al.,* 2022). This is because the authors would be assuming that the data was representative of the same individual if they integrated the reads into the ONT data. A HiSeq X10 was used, which could be improved upon by using a machine with higher sequencing coverage, however, the use of paired-end sequencing and the filtering of reads to those with Phred scores above 20 were high-quality standards. Given that the data’s purpose was to cover indels which can be very short and precise changes between reads, the highest quality of base calling should be desired to improve this in the first instance.

Filipović et al. aligned the ONT data to RNA-seq reads which were de novo assembled in previous literature due to a lack of reference genome using BWA-MEM (which is better than other aligners such as bowtie but could have been superseded by hisat2 in speed and accuracy (Becht *et al.,* 2021)). To justify adding this data (also Illumina, HiSeq 2500) to their own and cover contig gaps in the assembly rather than using it to correct indels, the group went through extensive means to identify the most accurate alignment: The reads had pre-existing structural annotation which helped to align to the ONT data after they were polished by removing Illumina adapters. They were mapped to this paper’s de novo assembly, meaning only reads possessing some homology to the de novo assembly were included.

BUSCO analysis included outgroup comparison with 39 assemblies from within the Order Coleoptera, but this de novo assembly possessed 99.1% of the BUSCO genes from the endopterygota\_odb10 database. This high BUSCO score indicates a high-quality assembly overall. Although the final assembly had a high contiguity (L50 12), employing Hi-C-based proximity-guided scaffolding could further improve the assembly quality by covering gaps between contigs.

## Section 2 [com]

Having read the origin paper, I endeavoured to perform an additional assessment of the raw data used to generate the *de novo* *O.rhinoceros* genome.

#### Methods in brief

Given that the data was generated using minION, my analysis workflow preferred tools which are tailored to long-read raw data. The code snippets below are adapted from NGG Tutorial 3 to take four ONT libraries and produce a QC report, assembly and assembly assessment.

The accessions for four libraries’ raw files were obtained from the NCBI SRA file accession database (NCBI, ), they were converted from SRA format to FASTQ and combined into a single reads file which was then zipped. I combined the four libraries into a single reads file to maximise the number of reads available, though this would have removed the ability to detect the influence of each flow cell on the read quality/ accuracy had I reached this level of assessment.

Nanoplot was used to generate quality metrics of the raw data, and wtdbg2 was used to assemble the reads, with its accompanying programme wtpoa-cns producing the contigs from my assembly. Summary metrics of the assembly were generated with the BBMap script stats.sh, which was recommended over scaffold.sh for the sake of speed.

Below is a code snippet including most of the commands run for this analysis. Note the high thread count used to perform the analysis in good time.

#Run Nanoplot for QC of reads transform data logarithmically for better analysis

NanoPlot --verbose --fastq ./merged.fastq.gz -o Nanoplot\_Phred\_cutoff\_8 --loglength --N50 -t175

#Run Readbean to assemble the reads (resource intensive)

wtdbg2 -x ont -i ./merged.fastq.gz -o ./Orhinoceros\_wtdg2 -g377m -t170

#Generated the contigs

wtpoa-cns -t170 -i ./Orhinoceros\_wtdg2.ctg.lay.gz -fo ./Orhinoceros\_wtdg2.ctg.fa

#Determined the quality of the assembly with BBMap's command stats.sh (gives number of contigs and N50 among others)

stats.sh ./Orhinoceros\_wtdg2.ctg.fa > Assembly\_assessment.txt

#### Results and discussion in brief

**Chart, histogram

Description automatically generated**

Figure 1-A graph of the unfiltered raw data plotted by read length and average read quality. The overall quality was typical of a minION experiment.

The origin paper filtered data in their assembly to reads with an average Phred score of 8 or above. This is also done in previous literature to aid proper taxonomic classification in downstream analyses (Tyler *et al.,* 2018) however I struggled to find a justification to do so in my analysis. Running Nanoplot with the --minqual 8 score filtering parameter showed what happens to the quality statistics of the data when eliminating reads with an average Phred score of less than 8.

Table 2- Summary statistics of the Nanoplot, showing the difference in quality stats when filtering out data with a Phred score of less than 8

|  |  |  |
| --- | --- | --- |
|  | Filtered | Unfiltered |
| Mean read length | 6,275.30 | 6,284.10 |
| Mean read quality | 13.7 | 13.5 |
| Median read length | 3,882.00 | 3,876.00 |
| Median read quality | 13.9 | 13.9 |
| Number of reads | 4,248,731.00 | 4,328,665.00 |
| Read length N50 | 11,147.00 | 11,184.00 |
| STDEV read length | 6,746.50 | 6,786.90 |
| Total bases | 26,661,928,682.00 | 27,201,635,900.00 |

Given that the metrics of the quality of the raw data did not differ substantially (mean read quality changing by 0.2, read length N50 only moving by ~40), I believe that it was most sensible to retain the unfiltered data for the assembly, as filtering these out would be omitting 1 billion bases from further analysis. Given that the bases which were retained had a Phred score of 8 or less, it could be argued that they do not add meaningful data (base call has a probability of being incorrect > 1 in 10), however, this is an expected quality level for minION data so I feel the retention was justified.

Chart, histogram

Description automatically generated

Figure 2- Nanoplot graph showing the distribution of read lengths over the raw reads

*Figure 2* shows the expected peak from 0-1000 bases for the ONT minION platform. This identifies short reads which are caused by technical error. We can assume that some data cleaning would be required for proper assembly. There is a strong peak at the 3500-4000 bases bin of read length, most likely caused by contaminating genomes such as the mitochondria or virus previously mentioned.

Phred score was not great, so other metrics can be used to determine read quality:

|  |  |
| --- | --- |
| What is expected coverage of 337 mb O.rhinoceros genome? | 27201635900 bases /377000000 genome size≈ 75X coverage |
| Base composition (%) | A: 0.3191  C: 0.1811  G: 0.1811  T: 0.3187 |
| basic metrics of the assembly from wtdg2 | Estimated: TOT 429148416, CNT 2540, AVG 168957, MAX 22724608, N50 7323648, L50 16, N90 66048, L90 278, Min 4608 |

Chart

Description automatically generated with medium confidence

Figure 3- The erroneous Bandage graph

Running Bandage (Wick *et al.,* 2015) with default settings to create a visualisation of the DBG graph gave 2545 nodes with no edges. I consider this a failed analysis, however, I was unable to determine the source of the error. If the visualisation had worked, it might have been possible to identify SNP in the reads (characterised by paths which have a read depth of 1, present in bubbles which are caused by bifurcations which re-join later. There was a very high number of contigs indicating poor assembly quality.

I would have liked metagenomic analysis to identify mitochondrial and pathogenic genomes using wtdbg2x before progressing to any potential analysis such as gene annotation. A hybrid assembly combining Illumina reads with this data would increase the quality of the reads from a mean of 13.4 Phred. Finally, an NG50 or NGA50 would be useful to identify misassembly.

## References

*SPRI.* Available at: <https://www.beckman.com/landing/ppc/genomics/cleanup-and-size-selection/size-selection-products> (Accessed: Mar 22, 2023).

Becht, C., Schmidt, J., Blessing, F. and Wenzel, F. (2021) 'Comparative analysis of alignment tools for application on Nanopore sequencing data', *Current directions in biomedical engineering,* 7(2), pp. 831-834. doi: 10.1515/cdbme-2021-2212.

Filipović, I., Rašić, G., Hereward, J., Gharuka, M., Devine, G.J., Furlong, M.J. and Etebari, K. (2022) 'A high-quality de novo genome assembly based on nanopore sequencing of a wild-caught coconut rhinoceros beetle (Oryctes rhinoceros)', *BMC genomics,* 23(1), pp. 426. doi: 10.1186/s12864-022-08628-z.

Hotaling, S., Sproul, J.S., Heckenhauer, J., Powell, A., Larracuente, A.M., Pauls, S.U., Kelley, J.L. and Frandsen, P.B. (2021) 'Long Reads Are Revolutionizing 20 Years of Insect Genome Sequencing', *Genome biology and evolution,* 13(8). doi: 10.1093/gbe/evab138.

NCBI *Run Selector NCBI.* Available at: <https://www.ncbi.nlm.nih.gov/Traces/study/?acc=SRS9765795&o=acc_s%3Aa&s=SRR15436048,SRR15436049,SRR15436050,SRR15436051> (Accessed: Mar 28, 2023).

NCBI and Tsatsia, H. (2019) *SRA data selector.* Available at: <https://www.ncbi.nlm.nih.gov/Traces/study/?acc=SRP332386&o=acc_s%3Aa> (Accessed: Mar 22, 2023).

Ruan, J. and Li, H. (2020) 'Fast and accurate long-read assembly with wtdbg2', *Nature methods,* 17(2), pp. 155-158. doi: 10.1038/s41592-019-0669-3.

Shelomi, M., Lin, S. and Liu, L. (2019) 'Transcriptome and microbiome of coconut rhinoceros beetle (Oryctes rhinoceros) larvae', *BMC genomics,* 20(1), pp. 957. doi: 10.1186/s12864-019-6352-3.

Srivathsan, A., Lee, L., Katoh, K., Hartop, E., Kutty, S.N., Wong, J., Yeo, D. and Meier, R. (2021) 'ONTbarcoder and MinION barcodes aid biodiversity discovery and identification by everyone, for everyone', *BMC biology,* 19(1), pp. 217. doi: 10.1186/s12915-021-01141-x.

Tyler, A.D., Mataseje, L., Urfano, C.J., Schmidt, L., Antonation, K.S., Mulvey, M.R. and Corbett, C.R. (2018) 'Evaluation of Oxford Nanopore’s MinION Sequencing Device for Microbial Whole Genome Sequencing Applications', *Scientific Reports,* 8(1), pp. 1-12. doi: 10.1038/s41598-018-29334-5.

Wick, R.R., Schultz, M.B., Zobel, J. and Holt, K.E. (2015) 'Bandage: interactive visualization of de novo genome assemblies', *Bioinformatics,* 31(20), pp. 3350-3352. doi: 10.1093/bioinformatics/btv383.

Zascavage, R.R., Hall, C.L., Thorson, K., Mahmoud, M., Sedlazeck, F.J. and Planz, J.V. (2019) 'Approaches to Whole Mitochondrial Genome Sequencing on the Oxford Nanopore MinION', *Current Protocols in Human Genetics,* 104(1), pp. e94-n/a. doi: 10.1002/cphg.94.

stylefix