Expanding the Lake Malawi cichlid genome using high quality long-read sequencing

A Dissertation
Presented to
The Academic Faculty

by

Lauren Sabo
Georgia Institute of Technology
lsabo8@gatech.edu

In Partial Fulfillment of the Requirements for the Degree
of Computer Science in the College of Computing

CHAPTER 1. INTRODUCTION

The family Cichlidae represents one of the most remarkable examples of adaptive radiation in vertebrates, comprising over 2,000 species globally, with 500-860 species endemic to Africa's Lake Malawi alone (Fan et al., 2019; Santos et al., 2023). This extraordinary diversity encompasses dramatic variation in morphology, trophic adaptations, coloration, and behavior. This diversity is shaped by an evolutionary process known as adaptive radiation. Lake Malawi cichlids demonstrate a unique ability to thrive in diverse habitats, from rock-dwelling zooplankivores to pelagic piscivores and benthic sand-dwellers (Malinsky et al., 2018).

Despite differences in mating rituals, coloration, and behavior, this family can hybridize, producing offspring with novel combinations of parental traits (Conte et al., 2019). This hybridization capability fuels further adaptive radiation, leading to the emergence of numerous new species. Consequently, Lake Malawi cichlids provide an extraordinary model for studying evolutionary processes in real-time.

In 2015, computational biologists at the University of Maryland successfully sequenced the genome of *Maylandia zebra* (*M. zebra*), an endemic species to Lake Malawi (Conte et al., 2019). Since its publication, this genome assembly has become the "gold standard" for genomic research on Lake Malawi cichlids, remaining the highest-quality reference for comparative studies. However, relying solely on the genome of a rock dweller is not ideal for analyzing the genetic variation of species with drastically different ecologies, such as those that are pelagic or piscivorous. However, its dominance as the primary resource for cichlid genome analysis has constrained researchers' ability to explore variation within other cichlid species. This study aims to address this limitation by updating the current *M. zebra* genome and introducing four additional cichlid species to the repertoire of reference genomes using a genomic assembler within a custom pipeline.

The advancement of long-read sequencing technology has opened new avenues for generating high-quality genomic assemblies. Unlike short-read sequencing, which often results in fragmented assemblies, long-read sequencing produces continuous, extensive reads that span complex regions of the genome. This capability allows for the accurate assembly of repetitive and structurally variant regions, which are often underrepresented in short-read assemblies. By utilizing long-read sequences, we can create more complete and accurate reference genomes for cichlid species, capturing the full extent of their genetic diversity. This will significantly enhance our understanding of the evolutionary mechanisms driving the diversification of cichlids in Lake Malawi and provide a more comprehensive foundation for future genomic studies.

By expanding the collection of high-quality cichlid reference genomes, this project seeks to enable more comprehensive comparative genomics analyses across multiple species. For instance, the availability of new reference genomes will allow us to identify genetic variations specific to different ecological groups, such as those that influence the brilliant coloration in mbuna cichlids. These key mutations, which may be present in rock-dwelling mbuna but absent in sand-dwelling species, can be resolved and analyzed in detail. This specificity will enhance our understanding of the genetic mechanisms underlying the diverse phenotypic traits observed in Lake Malawi cichlids. Moreover, the availability of multiple reference genomes will improve the precision of population genomic studies, providing insights into the demographic history and adaptive processes shaping cichlid populations within the lake.

CHAPTER 2. LITERATURE REVIEW

In 2019, Conte et al. made significant advancements in genomic research by re-anchoring the Nile tilapia genome and developing a novel genome assembly for the Lake Malawi cichlid, *Metriaclima zebra*, which enabled chromosome-scale comparisons across African cichlid genomes (Conte et al., 2019). Their study revealed frequent large intra-chromosomal structural variations among species, while inter-chromosomal differences were comparatively rare. Additionally, they identified variations in centromere placement, which provided insights into karyotype differences. The task of assembling a fully sequenced genome is immensely complex, and the work of Conte et al. represents a substantial progression in this field.

Genome assembly can be conceptualized through the analogy of reconstructing a jigsaw puzzle under challenging conditions. In post-sequencing bioinformatics, researchers must assemble billions of DNA fragments, the individual puzzle pieces, without access to a reference image and often without clearly defined boundaries. This challenge necessitates the strategic integration of two complementary methodological approaches: genomic sequencing and linkage mapping. While genomic sequencing provides the puzzle pieces, linkage mapping offers crucial information about the length of each chromosome, essentially the size of each puzzle section. For instance, when a large segment of assembled pieces (a chromosome) is identified, linkage mapping can determine that this section is the only one of sufficient size to accommodate the number of puzzle pieces it contains.

The integration of these two data types becomes particularly complex in de novo assembly contexts, those lacking a closely related reference genome. Without a template to guide the positioning of sequence fragments, assembly becomes a computationally intensive problem, requiring sophisticated algorithms to determine genomic structure from “red herring” arrangements. Conte et al. addressed this challenge by developing a comparative framework that leveraged existing tilapia genomic resources to facilitate chromosomal differentiation in *M. zebra*, establishing a methodological template that has proven valuable for subsequent Lake Malawi cichlid genomic studies.

In a study by Cosma et al., the importance of carefully selecting assembly algorithms when utilizing third-generation sequencing technologies, such as Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio), for eukaryotic genome assemblies is emphasized, particularly due to the higher error rates associated with long reads (Cosma et al., 2022). The advent of HiFi reads, which offer significantly reduced error rates, presents a promising approach for achieving more accurate assembly outcomes. By benchmarking state-of-the-art long-read de novo assemblers on both real and simulated datasets, the study identifies Flye as the leading assembler for PacBio CLR and ONT reads, while Hifi.asm and LJA emerge as the optimal choices for PacBio HiFi reads. This finding underscores the positive impact of read length on assembly quality, although the extent of this impact varies depending on the genome's size and complexity.

For this thesis, involving the use of PacBio HiFi reads, the study by Cosma et al. suggests that Hifi.asm and LJA are the most suitable assemblers, providing a valuable starting point for narrowing down the selection of sequencing tools. This study complements the work of Mahmoud et al., who review the significance of structural variants (SVs) in evolutionary, population, and clinical genomics (Mahmoud et al., 2019). Mahmoud et al. evaluate various algorithms for SV detection, highlighting their respective strengths and limitations. While short-read mapping methods are cost-effective for genotyping known SV alleles, they struggle with detecting novel SVs, particularly insertions. In contrast, de novo assembly and long-read sequencing approaches offer a comprehensive means of SV detection but are currently impractical for large-scale studies due to the high costs associated with sequencing. Given that this thesis focuses on using genome sequencing algorithms to assemble the structural variants of African cichlids, Mahmoud et al.'s study is invaluable, as it critically evaluates the advantages and disadvantages of different sequencing methodologies.

CHAPTER 3. METHODS

DNA Extraction and Quality Assessment

To sequence DNA on the Pacbio HiFi platform, DNA was initially extracted using Qiagen MagAttract HMW DNA Kit. Fin clips from the respective species, weighing approximately 10-15 mg, were used as starting material for this extraction. The DNA quality was assessed by a NanoDrop 2000 Spectrophotometer (Thermo Fisher Scientific) according to the manufacturer's instructions. Purity was quantified by determining the A260/A280 and A260/A230 absorbance ratios using a Qubit™ Fluorometer and Qubit™ 1X dsDNA High Sensitivity (HS) Assay Kit.

Transition to Blood-Based DNA Extraction

Due to inconsistencies in data quality from fin clip extractions, the protocol transitioned to using blood samples containing 1.5 million nucleated blood cells as the starting material, following the Bionano SP-G2 Blood DNA isolation protocol. This approach yielded more consistent and higher-quality DNA suitable for long-read sequencing. Blood samples were processed in the same manner as fin clips, with DNA quality and purity evaluated using the NanoDrop 2000 Spectrophotometer and Qubit™ Fluorometer as described above.

PacBio HiFi Sequencing and Data Generation

To generate long-read sequences, PacBio high-fidelity (HiFi) sequencing technology was utilized. This approach was selected due to its ability to produce highly accurate reads with significant depth, which is critical for the subsequent stages of genomic analysis. The samples were processed through the Pacbio Revio sequencer which produced long reads with superior accuracy.

Comparative De novo Assembly

In the process of assembling the cichlid genomes, three different de novo assembly algorithms were compared: Flye (version 2.9.5), Hifi.asm (version 0.7), and LJA (version 0.2). These algorithms were chosen for their specialization in handling long-read data, making them particularly suited for the high-quality reads produced by Pacbio HiFi sequencing.

Flye is known for its ability to efficiently assemble genomes from noisy long reads, making it a robust choice for complex genomes. Hifi.asm is optimized for accuracy and speed, specifically tailored for HiFi data, and can handle large genome assemblies with high precision. Longest Jump Assembly, also known as "LJA", is designed to produce highly contiguous assemblies by leveraging long-range information from the reads.

FASTQ files from PacBio HiFi sequencing for five cichlid species were used as the input for each assembler with default parameters to ensure a standardized evaluation.

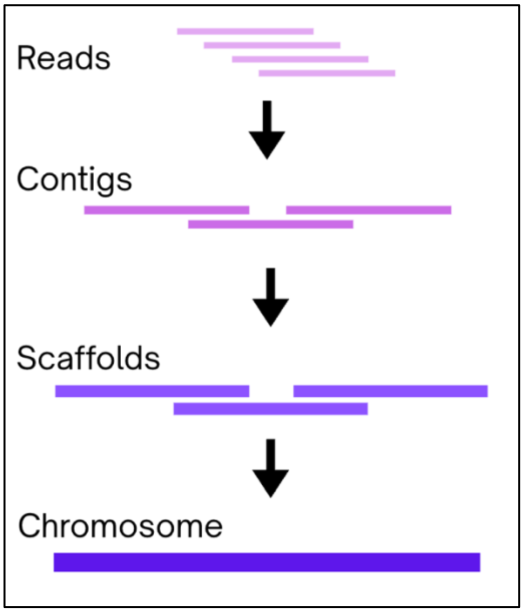

Figure 1. *De novo assembly stages, with overlay*

Computational Pipeline Design and Parallelized Workflow Architecture

A fair portion of this project was the development of a custom computational pipeline designed to automate, parallelize, and manage the full genome-assembly workflow for five cichlid species. Because long-read genome assembly is fundamentally a large-scale computing problem, requiring the creativity to handle multi-gigabyte datasets, the pipeline served as the computational backbone of the project so that a reproducible analysis could be performed across species.

The pipeline was implemented in Python and integrates multiprocessing, file-system organization, and structured data handling to run each assembler (Flye, Hifi.asm, and LJA) under controlled, comparable conditions. To accomplish this, the workflow launches assemblies in parallel across 11 compute cores on the laboratory's Linux server, while logging memory usage, execution time, and intermediate results for downstream evaluation. Python's multiprocessing module was used to distribute independent assembly jobs across species and assemblers, while shared-memory data structures were used to collect assembly metrics for comparison. This computational design assured the assembler testing would be conducted efficiently and reproducibly, avoiding runtime conflicts and allowed for a rapid iteration process during pipeline development.

The pipeline also used data-processing practices to automate the extraction of genome statistics from the resulting FASTA and FASTQ files. These components parsed contig lengths, computed summary statistics (including N50, depth, and size distributions), and returned standardized outputs for external tools such as Inspector, Bionano Solve, and D-GENIES. All steps were organized in Python with documentation, allowing for transparency and collaboration with project mentors and ensuring that the full workflow could be continued or revised by the McGrath Lab after the conclusion of this undergraduate project.

By developing this computational pipeline, the project contributed not only biological results but also a scalable computational framework for long-read assembly benchmarking.

Assembly Quality Evaluation

The completeness of each genome assembly was evaluated using Inspector, a reference-free long-read de novo assembly evaluation algorithm. By inputting the assembly files generated by each assembler along with the corresponding FASTQ files, the performance of each assembler was compared using several metrics: total genome length (base pairs), the total number of contigs, the total length of contigs greater than 1 million base pairs, the length of the longest contig, the length of the second-longest contig, the N50 value, and the sequencing depth (calculated by dividing the total length of aligned reads by the total contig length).

This comprehensive analysis allowed us to identify the assembler that produced the most complete and accurate genome assemblies for the cichlid species under study.

Hybrid Scaffold Construction and Visualization

Following de novo assembly, hybrid scaffolds were constructed by integrating the PacBio HiFi assembly data with Bionano Optical Genome Mapping (BOGM). This integration combined the structural information from optical maps with the high-accuracy sequence data to produce highly contiguous and accurate scaffolds.

Bionano software tools were used to visualize the hybrid scaffold construction, providing insights into the structural organization of the genomes. This step was critical for refining the assemblies and ensuring high-quality representations of the cichlid genomes.

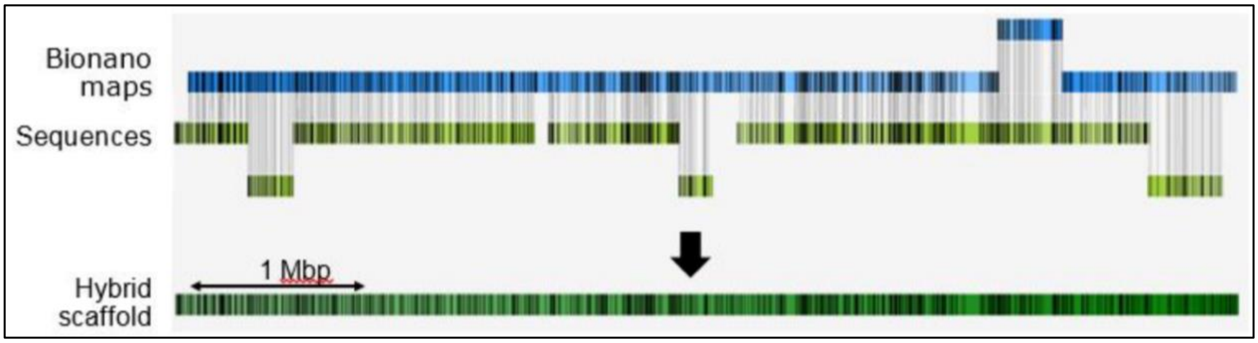

Figure 2. *Bionano hybrid visualization software (Bionano Genomics, 2020, Fig. 1)*

Comparative Analysis with D-GENIES

After hybrid scaffold construction, D-GENIES was used to perform comparative genomic analysis. Each of the hybrid scaffolds generated from Bionano Optical Genome Mapping (BOGM) was used as the query sequence, while a singular *M. zebra* reference genome (in FASTA format) served as the target sequence.

This process generated four new FASTA sequences by aligning and comparing the hybrid scaffolds against the reference genome. These outputs provided additional insights into the structural differences and similarities between the assemblies and the M. zebra reference, facilitating further refinement and evaluation of the hybrid scaffolds.

CHAPTER 4. RESULTS

Inspector Analysis on Assembler Runs using Default Parameters

In this study, we evaluated the performance of two genome assemblers, Hifi.asm and LJA, by comparing key metrics for the five cichlid species sequenced. Each sample was run using 11 computational cores and without any defining parameters (default parameters used only). The comparison was based on total genome length (bp), the total number of contigs, the total length of contigs greater than 1 million bp, the length of the longest contig, the second-longest contig, N50, and depth.

Sample Name	Total Length (Bp)	Total Number of Contigs	Total Length >1m Bp	Longest Contig (Bp)	2nd Longest Contig (Bp)	N50	Depth
MZ4f	958496430	220	945117420	63676297	56938743	31682983	80.7285
MC3m	968257136	204	955103536	56808348	46363082	31612503	95.0988
CV4f	942077428	242	927111976	61172947	54395767	32351000	99.2655
YH7f	963007403	346	934306827	61743854	44274689	24131926	62.42
YH7m	963940697	339	932273567	45123318	9277260	755911	20.5293
LF2f	946102408	363	918077103	34713868	31330946	15303642	40.8995

Figure 3a. *Hifi.asm Run with Default Parameters on 6 Samples*

Sample Name	Total Length (Bp)	Total Number of Contigs	Total Length >1m Bp	Longest Contig (Bp)	2nd Longest Contig (Bp)	N50	Depth
MZ4f	958496430	220	945117420	63676297	56938743	31682983	80.7285
MC3m	968257136	204	955103536	56808348	46363082	31612503	95.0988
CV4f	942077428	242	927111976	61172947	54395767	32351000	99.2655
YH7f	963007403	346	934306827	61743854	44274689	24131926	62.42
YH7m	963940697	339	932273567	45123318	9277260	755911	20.5293
LF2f	946102408	363	918077103	34713868	31330946	15303642	40.8995

Figure 3b. *LJA Run with Default Parameters on 6 Samples*

Figure 3a. & 3b. Key

Total Genome Length (bp)

The total genome length is a critical measure of how well the assembler reconstructs the genome.

Total Number of Contigs

A lower number of contigs indicates a more contiguous assembly, which is generally desirable in de novo assemblies.

Total Length of Contigs Greater Than 1 million bp

Contigs greater than 1 million bp are an indicator of the assembler's ability to resolve large and complex genomic regions.

Length of the Longest Contig (bp)

The longest contig generated by an assembler provides insight into its capacity to produce long, continuous sequences.

Length of the Second Longest Contig (bp)

Similar to the longest contig, the second longest contig also highlights the assembler's performance in producing contiguous genomic regions.

N50

The N50 statistic represents the contig length at which 50% of the total genome length is contained. A higher N50 indicates better assembly contiguity.

Depth

The sequencing depth was measured to assess the completeness of the genome assembly.

Based on these metrics, Hifi.asm generally outperformed LJA for total genome length, longest contig, N50, and total length of contigs over 1 million bp, making it the preferable assembler for generating high-quality, contiguous genome assemblies in cichlid species. However, LJA showed strengths in minimizing the total number of contigs, suggesting it may still be useful for certain genomic regions requiring higher contiguity. For this study, Hifi.asm was selected as the primary assembler due to its superior overall performance in reconstructing the cichlid genomes.

Bionano Hybrid Scaffolding with Hifi.asm Default Results

In the second phase of this study, Hifi.asm assemblies generated using default parameters were combined with optical mapping data through Bionano Hybrid Scaffolding. The scaffolding software highlights the contiguity of assemblies by resolving gaps and scaffolding large genomic regions. The addition of hybrid scaffolding data improved the metrics: total genome length, N50, and the number of large contigs (>1 million bp), while also reducing the total number of contigs.

CHAPTER 5. DISCUSSION

The high-quality genomic assemblies produced for *M. zebra*, *M. conophoros*, *C. virginalis*, *A. sp. 'chitande'*, and *L. fuelleborni* using Pacbio HiFi sequencing represent a significant advancement in cichlid genomics. These assemblies demonstrate the power of long-read sequencing technologies in overcoming the challenges posed by repetitive elements, structural variants, and high levels of heterozygosity, which are hallmark features of cichlid genomes. By achieving assemblies with high N50 values, large contig sizes, and a reduced number of fragmented regions, this study not only reaffirms the effectiveness of Pacbio HiFi sequencing but also establishes a critical resource for future studies aimed at understanding the genomic underpinnings of adaptive radiation and speciation.

The comparison between the HiFi.asm and LJA assemblies demonstrates differences in assembly metrics, including total contig length, the number of contigs over 1 million base pairs, and overall scaffold continuity. The superior contiguity of the HiFi.asm assembly, particularly in terms of N50 and largest contig size, suggests that HiFi.asm is the better choice for cichlid genome assembly, where repetitive regions and high heterozygosity can complicate assembly processes. However, LJA produced competitive results, particularly when resolving smaller contigs, suggesting that hybrid assembly approaches may produce the most complete representations of these genomes in future studies. These findings suggest that future efforts could benefit from hybrid assembly strategies that combine the strengths of both assemblers to achieve the most complete and biologically accurate representations of cichlid genomes.

The limitations of this study include the potential underrepresentation of structural variants and repetitive regions in the assemblies, despite the advantages offered by Pacbio HiFi sequencing. Future studies could address this by integrating complementary technologies, such as optical mapping or chromosome conformation capture (Hi-C), to produce even more refined assemblies. Additionally, while the RNA-Seq data provided valuable insights into gene expression patterns, further functional validation, including proteomic analysis, would be beneficial to fully understand the biological implications of the findings.

Overall, the study contributes to the field of cichlid genomics by providing comprehensive genome assemblies for five species, adding to the growing resources available for studying the genomic basis of adaptive radiation. These genomes also provide a critical framework for downstream analyses, including genome annotation, comparative genomics, and the identification of genes associated with adaptive traits. Annotation efforts using RNA-Seq data, for example, have already highlighted valuable insights into gene expression patterns, but this is just the beginning. Integrating these assemblies with additional datasets, such as proteomics, metabolomics, and epigenomics, will enhance our ability to link genomic variation to phenotypic traits, particularly those underlying ecological adaptation and speciation.

CHAPTER 6. CONCLUSION

This project created a foundation for generating high-quality reference genomes for five Lake Malawi cichlid species using PacBio HiFi sequencing and state-of-the-art assembly tools. Through comprehensive benchmarking of the Hifi.asm and LJA assemblers, this work identified the most effective strategy for maximizing contiguity, long-range accuracy, and structural resolution in cichlid genomes, ultimately demonstrating that Hifi.asm consistently produced the strongest assemblies under default parameters. These findings provided a critical evidence-based rationale for selecting Hifi.asm as the primary assembler for the subsequent hybrid scaffolding and iterative refinement pipeline.

The project’s success is reflected in the completion and submission of the finalized reference genome to NCBI on August 12, 2024, a milestone built on the assembler selection, performance analyses, and pipeline planning completed at the outset. While the later iterative improvement steps were carried out after my involvement, the groundwork laid here ensured that the PhD team could proceed with a validated, high-performance assembly strategy and a clear roadmap for genome polishing and scaffold evaluation.

Overall, this work demonstrates the importance of strong methodological design in genome assembly projects. By identifying optimal assembly tools, generating high-quality initial contigs, and structuring the pipeline for subsequent refinement, this project contributed significantly to the production of a robust, publicly available cichlid reference genome. The framework and insights developed here will support future comparative genomics, evolutionary analyses, and functional annotation efforts across this rapidly diversifying group of species.

REFERENCES

Bionano Genomics. (2020). Bionano Solve Theory of Operation: Hybrid Scaffold (Doc. No. 30073 Rev F). https://bionano.com/wp-content/uploads/2023/01/30073-Bionano-Solve-Theory-of-Operation-Hybrid-Scaffold.pdf

Conte, M. A., Joshi, R., Moore, E. C., Nandamuri, S. P., Gammerdinger, W. J., Roberts, R. B., Carleton, K. L., Lien, S., & Kocher, T. D. (2019). Chromosome-scale assemblies reveal the structural evolution of African cichlid genomes. *GigaScience*, *8*(4). https://doi.org/10.1093/gigascience/giz030

Cosma, B.-M., Shirali Hossein Zade, R., Jordan, E. N., van Lent, P., Peng, C., Pillay, S., & Abeel, T. (2022). Evaluating long-read de novo assembly tools for eukaryotic genomes: Insights and considerations. *GigaScience*, *12*, giad100. https://doi.org/10.1093/gigascience/giad100

Fan, S., Elmer, K. R., & Meyer, A. (2012). Genomics of adaptation and speciation in cichlid fishes: Recent advances and analyses in African and Neotropical lineages. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *367*(1587), 385–394. https://doi.org/10.1098/rstb.2011.0247

Mahmoud, M., Gobet, N., Cruz-Dávalos, D. I., Mounier, N., Dessimoz, C., & Sedlazeck, F. J. (2019). Structural variant calling: The long and the short of it. *Genome Biology*, *20*(1), 246. https://doi.org/10.1186/s13059-019-1828-7

Malinsky, M., Svardal, H., Tyers, A. M., Miska, E. A., Genner, M. J., Turner, G. F., & Durbin, R. (2018). Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. *Nature Ecology & Evolution*, *2*(12), 1940–1955. https://doi.org/10.1038/s41559-018-0717-x

Santos, M. E., Lopes, J. F., & Kratochwil, C. F. (2023). East African cichlid fishes. *EvoDevo*, *14*(1), 1–21. https://doi.org/10.1186/s13227-022-00205-5