GRA Proposal
Spring 2026

Lauren Sabo
McGrath Lab

Development of a Low-Coverage QTL Mapping Pipeline to Identify a Male Sex Determiner in *Aulonocara sp.* "Yellow Head"

1 Background

The African cichlids of Lake Malawi are one of the most unique examples of adaptive radiation among vertebrates. Lake Malawi cichlids possess vast diversity in their phenotypes and adaptations that suit them to the various environments across the lake Cichlid species are further known to utilize complex sex determination systems, including both XY and ZW mechanisms, making them powerful models for understanding how sex-linked variation evolves (Böhne *et al.*, 2025). Recent work by Kumar *et al.* (2025) revealed there are 6 large inversions segregating within Lake Malawi cichlid species. They discovered some of these inversions are heterozygous and could be serving as potential sex determining systems.

The *Aulonocara sp.* "Yellow Head" (YH) species contains all 6 of the inversions discovered by Kumar *et al*. In this species, the inversion on linkage group (LG) 10 shows a pattern: males are consistently heterozygous for the inversion, while all females tested (n = 24) were homozygous for the inverted haplotype. Their work provides evidence that a male-determining factor may be located on the non-inverted copy of LG10.

The McGrath lab has generated hybrid crosses between a YH male (heterozygous for the inversion) and a *Mchenga conophoros* (MC) female (homozygous non-inverted). The MC species was chosen because work by the McGrath lab showed that the MC lineage does not carry a male determiner on LG10. The resulting F1 offspring were genotyped for the inversion and phenotyped for sex. There was a statistically significant association between sex and inheritance of the non-inverted YH haplotype in the F1 generation (X-squared = 84.712, df = 1, p-value < 2.2e-16, unpublished work). An F_1 male was subsequently backcrossed to an MC female to produce a large sample size of backcross descendants (N=276). These backcrosses can be used to map the genomic region linked to sex determination in YH using quantitative approaches.

The goal of this project is to develop a computational pipeline for quantitative trait locus (QTL) mapping (Broman *et al.*, 2003) that can efficiently process ultra-low-coverage (~0.1X) short-read sequencing data from these backcross individuals to map the male sex determiner in YH. Specifically, this study will apply probabilistic genotype inference using Hidden Markov Models (HMMs) and to perform QTL analysis.

Research Question: Can we identify the genomic region linked to sex determination in *Aulonocara sp.* "Yellow Head" by constructing a low-coverage-compatible QTL mapping pipeline?

2 Methodology

This project will create a reproducible QTL mapping pipeline designed for low-coverage whole-genome data from experimental crosses. The workflow includes four major components: (1) aligning sequencing reads and variant calling, (2) phasing high coverage (~15X) parents (3) predicting offspring haplotypes using HMM and QTL mapping

2.1 Alignment and Variant Calling

Illumina short reads from backcross individuals will undergo quality control using FastQC and trimming with Trimmomatic. Cleaned reads will be aligned to the *Metriaclima zebra* reference genome (Mzebra_GT3a) using BWA-MEM (Li *et al.*, 2009). Resulting BAM files will be used for variant calling with the GATK using the *HaplotypeCaller* algorithm in `-ERC GVCF` mode to generate GVCF for each sample. These files will subsequently be used for joint genotyping of the whole cohort using the *GenomicsDBImport* and *GenotypeGVCFs* algorithms.

2.2 Phasing Parents

The YH father, MC mother, and F_1 father will be sequenced to about 15X coverage. Using Beagle 5.5 (Browning *et al.*, 2021) we will phase the F_1 father using his parents. Homozygous variants specific to each parent can be used to phase variants in the F_1 father. A resulting phased VCF file will be utilized in the next stage for phasing BC_1 offspring using HMM imputation.

2.3 BC_1 Offspring genotype Imputation and QTL Mapping

The genotype likelihoods (PLs) from BC_1 offspring will be crucial for determining if they came from the YH genome or the MC genome. Any phased F_1 heterozygous variants where the MC mother is homozygous for the MC allele can be used for determining if the BC_1 offspring inherited YH DNA at that locus. The R/qtl2 *calc_genoprob* algorithm will be to impute the haplotype state of the BC_1 offspring at every variant. The algorithm utilizes the PLs from the BC_1 individuals as the emission probabilities. The phased haplotypes of the F_1 father are used to define the HMM's two hidden states (e.g., YH/MC or MC/MC). The transition probabilities are the recombination frequencies between markers. This map can be generated using R/qtl2 using the *mstmap* algorithm. The output from *calc_genoprob* can be passed to the *scan1* function for correlating the genotypes with BC_1 sex. The output will be a LOD (logarithm of odds) scores plot showing which loci are correlated with sex in our offspring, allowing us to pinpoint the y sex determiner in *Aulonocara* “Yellow Head”.

3 Proposed Work

3.1 My Plan

Over the course of the semester, we will develop, test, and validate the QTL mapping pipeline. During the first month, efforts will focus on data preparation, including retrieving raw sequencing data, performing quality control, and aligning reads to the *M. zebra* reference genome. We will perform variant calling and variant filtering and the parental and F_1 genomes will be phased.

By mid-semester, we will begin implementing and refining the Hidden Markov Model component to impute BC_1 offspring genotypes using the phased parental data. This stage will involve developing an R script using R/qtl2 for generating the recombination frequency map and then imputing BC1 genotypes. This R executable will be integrated within the Python-based framework for seamless processing of variant data generated from the first month's work. At the conclusion of this phase, we will have identified genomic regions associated with sex in *Alonocara sp.* "Yellow Head", with particular attention to LG10 as indicated by prior research.

In the final phase of the semester, LOD score plots and summary statistics will be generated, and the results will be interpreted in the context of cichlid sex chromosome evolution. A list of candidate genes within the identified region will also be compiled using the new Mzebra_GT3a genome annotation generated by RefSeq (GCF_041146795.1). The project will conclude with a fully documented, reproducible pipeline and a comprehensive report detailing the methodology, results, and broader biological implications.

3.2 Deliverables

- A reusable Python-based pipeline integrating R/qtl2 for QTL mapping
- QTL plots showing LOD score distributions across linkage groups.
- A list of potential genes within the sex-linked region in focus.
- Full documentation and open-source code.

3.3 Broader Implications

This project will generate both a computational tool and biological understanding. The resulting pipeline can utilize low-cost, low-coverage sequencing data to perform QTL analysis for other cichlid crosses. Biologically, this work will help clarify how inversions influence sex determination in African cichlids and open the door to studying other phenotypes of interest like well-documented courtship behaviors in cichlid species. By resolving whether the YH male determiner is linked to the LG10 inversion, this study will contribute to a broader understanding of how genomic structural variation influences sex chromosome evolution.

References

Böhne, A., Hsiung, K., & Smith, S. H. (2025). Zw and xy sex chromosomes drive rapid and distinctive evolution of sex‑biased gene expression. *Molecular Ecology*. https://doi.org/10.1111/mec.70152

Broman, K. W., Wu, H., Sen, Ś., & Churchill, G. A. (2003). R/qtl: QTL mapping in experimental crosses. Bioinformatics, 19(7), 889–890. https://doi.org/10.1093/bioinformatics/btg112

Browning, B. L., Tian, X., Zhou, Y., & Browning, S. R. (2021). Fast two-stage phasing of large-scale sequence data. The American Journal of Human Genetics, 108(10), 1880–1890. https://doi.org/10.1016/j.ajhg.2021.08.005

Kumar, N. M., Cooper, T. L., Kocher, T. D., Streelman, J. T., & McGrath, P. T. (2025). Large inversions in Lake Malawi cichlids are associated with habitat preference, lineage, and sex determination. eLife Sciences Publications, Ltd. https://doi.org/10.7554/elife.104923

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics, 25(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324