CS 5970: Bioinformatics Project

Genomic Basis of Exaggerated Dorsal Fin Morphology in Livebearing Fishes

Tyler J. Reich

Abstract:

Morphological diversity in closely related species provides powerful opportunities to understand how genetic variation shapes phenotypic evolution. In livebearing fishes of the genus *Poecilia*, dorsal fin length varies widely across species, with some exhibiting “sailfin” morphologies and others possessing short, rounded fins. Hybrids between long- and short-finned species display intermediate phenotypes, suggesting a quantitative genetic basis for this trait. Despite its ecological and evolutionary significance, the genetic architecture underlying dorsal fin length remains unknown.

This project aims to develop and implement a bioinformatics workflow to identify genomic regions associated with dorsal fin length variation by integrating low-coverage whole-genome sequencing (lcWGS) with quantitative trait loci (QTL) mapping in hybrid populations. I have already generated F1 hybrids between *Poecilia latipinna* (Sailfin molly) and *P. mexicana* (Atlantic molly) and am producing F2 hybrids for genomic analysis. Building on approaches established by Powell et al., (2021), this project will focus on designing and testing a complete computational pipeline for processing lcWGS data, performing variant calling, generating ancestry-informative markers, and conducting QTL scans to identify chromosomes and genomic intervals influencing dorsal fin length.

The outcome of this work will be a reproducible, well-documented analysis pipeline ready for deployment once sequencing data become available. This project will directly contribute to understanding genotype-phenotype relationships in Poecilia, inform future transcriptomic analyses of fin development, and support my dissertation research on the evolution of morphological variation in livebearing fishes.

Specific Aims:

Variation in dorsal fin morphology across species of *Poecilia* represents an evolutionarily important trait whose genetic basis remains unknown. Long-finned species (e.g., *Poecilia latipinna*) and short-finned species (e.g., *P. mexicana*) differ markedly in dorsal fin length; and hybrids show intermediate phenotypes, indicating a quantitative genetic basis. Identifying the genomic regions responsible for this variation will provide crucial insight into how morphological traits evolve and diversify in livebearing fishes.

The goal of this project is to develop and implement a complete computational workflow for analyzing low-coverage whole-genome sequencing (lcWGS) data from F2 hybrids to identify quantitative trait loci (QTL) associated with dorsal fin length. This work builds directly on established methods used in the closely related *Xiphophorus* and *Poecilia* hybrid systems and lays the foundation for the genomic analyses that will form a central component of my dissertation research.

Aim I: Develop a reproducible pipeline for processing low-coverage whole-genome sequencing data and generating high-quality genotype likelihoods.

I will implement a bioinformatics workflow for trimming, aligning, and filtering lcWGS data from F2 hybrids generated by crossing *P. latipinna* and *P. mexicana*, then performing F1 hybrid intercrosses. This aim includes identifying ancestry-informative markers (AIMs), performing variant calling using genotype-likelihood-based approaches (e.g., GATK), and generating the input files required for QTL analysis. Innovation: Applying low-coverage, likelihood-based genomic methods to this system reduces sequencing cost while retaining power for detection. Outcome: A validated, modular pipeline for lcWGS processing and variant calling.

Aim II: Perform quantitative trait locus (QTL) mapping to identify genomic regions associated with dorsal fin length.

Using phenotypic measurements from F2 hybrids and the genotypic data generated in Aim I, I will conduct QTL mapping to detect chromosomes and genomic intervals underlying dorsal fin length variation. This includes model selection, permutation testing, and visualization of QTL effect sizes. Innovation: This will represent the first QTL analysis of dorsal fin morphology in *Poecilia* using F2 hybrids and establishes a novel framework for mapping morphological traits in this group. Outcome: Identification of candidate chromosomes and genomic regions influencing fin length and a foundation for future fine-mapping and gene expression analyses.

Impact:

Completion of these aims will produce a fully documented, analysis-ready workflow for lcWGS-based QTL mapping and generate genomic insights into the genetic architecture of dorsal fin variation in mollies. This work will directly support downstream RNA-seq studies of candidate genes, strengthen the genomic tools available for *Poecilia*, and accelerate progress toward understanding how complex morphological traits evolve.

Research Strategy:

Significance:

Exaggerated morphological traits provide powerful systems for uncovering how genetic variation gives rise to phenotypic diversity, yet the genetic architecture of many sexually selected ornaments remains poorly understood. In livebearing fishes of the genus Poecilia, dorsal fin morphology varies dramatically among species, with Sailfin mollies (*P. latipinna*) exhibiting a hypertrophied, sexually selected fin while closely related species such as *P. mexicana* retain short, unornamented fins (Figure 1) (Farr et al., 1986; Reznick et al., 2017; Goldberg et al., 2019; Reznick et al., 2021). Hybridization between these species produces intermediate dorsal fin lengths, strongly suggesting a quantitative genetic basis for this trait and highlighting the system as an ideal model for dissecting genotype-phenotype relationships (Berbel-Filho et al., in prep).

Figure 1: Males of *Poecilia latipinna* (top) and *P. mexicana* (bottom) exhibiting drastic dorsal fin morphology, particularly in length.

Despite this promise, only one study has attempted to identify genomic regions influencing dorsal fin exaggeration in *Poecilia* (Keong et al., 2014), and that work relied solely on a small microsatellite (SSR) panel of 29 loci – only 18 of which were informative and grouped into four linkage groups – rather than whole-genome sequencing, limiting both genomic resolution and power to detect polygenic effects. As a result, the field still lacks fundamental knowledge of: (1) how many loci contribute to dorsal fin length variation, (2) whether the trait is governed by a few large-effect regions or a polygenic architecture, and (3) whether the genetic basis of fin exaggeration overlaps with known sexually selected ornaments in other poeciliids, such as swordtails (*Xiphophorus*) (Powell et al., 2021).

This lack of genomic resolution represents a major barrier to progress. Without high-density markers and genome-wide scans, the evolutionary and developmental mechanisms underlying dorsal fin exaggeration remain unresolved, limiting our ability to connect phenotypic evolution to underlying genetic processes. By applying low-coverage whole-genome sequencing (lcWGS) and QTL mapping in F2 hybrids, this project directly addresses this gap and will establish the first high-resolution genomic framework for understanding the evolution of exaggerated fin morphology in *Poecilia*.

Innovation:

This project introduces several methodological and conceptual innovations that advance the study of morphological evolution in *Poecilia* and extend current research paradigms in fish genomics.

1. **First application of whole-genome-based QTL mapping for dorsal fin morphology in *Poecilia*:**

Previous work on dorsal fin exaggeration relied exclusively on low-resolution microsatellite markers (Keong et al., 2014), which limited the ability to detect small-effect loci or define precise genomic intervals. By leveraging low-coverage whole-genome sequencing (lcWGS), this project provides the first high-density, genome-wide characterization of the genetic architecture underlying dorsal fin length in mollies.

2. **Implementation of a cost-efficient, likelihood-based genomic workflow:**

The project applies cutting-edge lcWGS pipelines that use genotype likelihoods rather than hard-called SNPs, allowing accurate QTL detection even at low sequencing depth. This approach, recently adopted in other teleost hybrid systems (Powell et al., 2021), substantially reduces sequencing costs while preserving statistical power, making it an innovative model for morphological mapping in non-model fishes.

3. Use of F2 hybrids to dissect quantitative trait architecture across species boundaries:

By mapping dorsal fin variation in F2 hybrids between *P. latipinna* and *P. mexicana*, this project captures recombination across divergent genomes, enabling fine-scale detection of loci influencing fin morphology. This design provides stronger resolution than within-species mapping and represents a novel comparative genetic framework for *Poecilia* ornamentation.

4. Integration with downstream functional genomics:

The pipeline developed here is explicitly designed to interface with future RNA-seq and candidate gene expression analyses of fin development. This forward-looking integration distinguishes the project from earlier work that ended at marker identification (Keong et al., 2014), opening the door to mechanistic studies of the developmental pathways shaping exaggerated fin structures.

5. Creation of a generalizable, reproducible pipeline through collaboration with Dan Powell:

This project leverages an active collaboration with Dr. Dan Powell, whose team developed a robust lcWGS and QTL-mapping pipeline for identifying the genetic architecture of sexually selected fin ornaments in *Xiphophorus* (Powell et al., 2021). By adapting and extending this proven workflow to analyze dorsal fin length in *Poecilia*, this project transforms an ornament-mapping framework originally designed for swordtail fish into a generalizable, cross-taxa toolset for studying morphological evolution. The resulting pipeline – documented, modular, and compatible with low-coverage genomic data – will provide a reusable resource not only for the *Poecilia* community but also for researchers working on hybrid systems, sexual selection, and trait evolution in other fish lineages. This represents the first systematic translation of the *Xiphophorus* ornament-mapping pipeline to another genus, demonstrating its broader applicability and significantly expanding the genomic toolkit available for poeciliid evo-devo research.

Together, these innovations modernize genetic mapping in *Poecilia*, expand the toolkit for investigating sexually selected ornamentation, and establish a new methodological foundation for linking genotype to phenotype in livebearing fishes. By integrating low-coverage whole-genome sequencing, genotype-likelihood-based variant calling, and hybrid-mapping designs, this project brings Poecilia genetics in line with contemporary standards used in model systems. The resulting pipeline not only increases power and resolution relative to earlier microsatellite-based approaches (Keong et al., 2014) but also creates a scalable framework that can incorporate future datasets, including RNA-seq and long-read assemblies. In doing so, this work positions *Poecilia* as a tractable genomic system for dissecting the molecular basis of complex traits, enabling downstream research into regulatory evolution, convergent genetic pathways with *Xiphophorus* (Powell et al., 2021), and the broader mechanisms by which sexual selection shapes morphological diversity.

Approach:

To generate the mapping population, *Poecilia latipinna* (sailfin) will be crossed with *P. mexicana* (shortfin/Atlantic molly) to produce F1 hybrids, which will then be intercrossed to generate an F2 population segregating for dorsal fin length (Figure 2) (Lander & Botstein, 1989). A sufficiently large F2 cohort (target n = 200-300) will be maintained to ensure adequate statistical power for detecting QTL, including small-effect loci in a potentially polygenic

architecture. All F2 individuals will be phenotyped for dorsal fin length, height, and overall shape using standardized digital imaging and morphological analysis such as ImageJ (Schneider et al., 2012). Additional covariates, including total body length and sex, will be recorded to account for allometric scaling and sexual dimorphism in downstream QTL analyses.

Genomic DNA will be extracted from all F2 individuals and used to prepare sequencing libraries for low-coverage whole-genome sequencing (lcWGS) at approximately 1x coverage per individual (Li et al., 2009). Reads will be aligned to both parental reference genomes (*P. latipinna* and *P. mexicana*) using Burrows-Wheeler Aligner (BWA) (Li & Durbin, 2009) and processed with Sequence Alignment/Map (SAM) tools (Li et al., 2009) to maximize accuracy in ancestry assignment and variant detection. This dual-reference alignment improves assignment of ancestry at each locus, enhances detection of species-specific alleles, and minimizes mapping bias that can occur when using a single reference, thereby increasing the power and precision of downstream QTL analyses. Standard quality-filtering procedures will be applied to remove low-quality or ambiguous reads, ensuring reliable results while maintaining cost efficiency for the large F2 population.

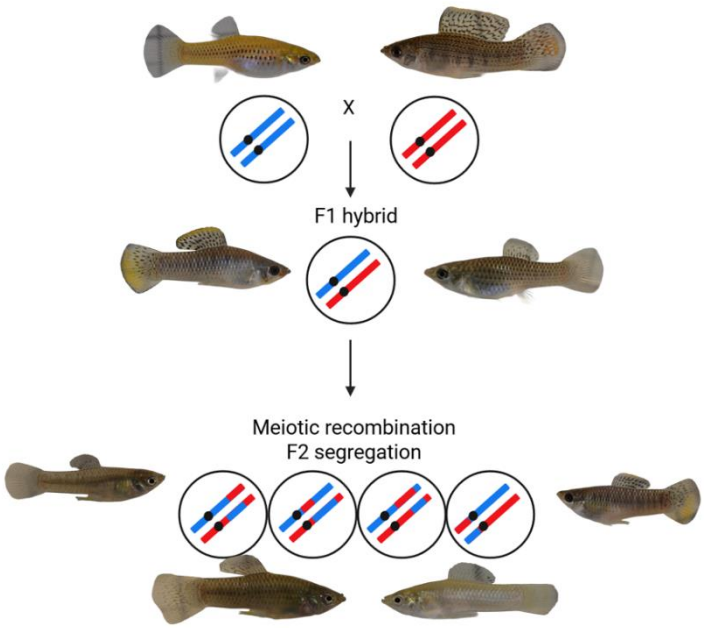

Figure 2: Schematic demonstrating F2 intercross design between a *Poecilia latipinna* male (top right) and *P. mexicana* female (top left) with F2s segregating for dorsal fin size (bottom).

Genotype calling will be performed using Genome Analysis Toolkit (GATK)'s genotype-likelihood-based methods (McKenna et al., 2010), allowing robust Single Nucleotide Polymorphism (SNP) identification from low-coverage sequencing data. Ancestry-informative markers (AIMs) that distinguish *P. latipinna* and *P. mexicana* alleles will be generated and local ancestry of F2s will be inferred using a Hidden Markov Model (HMM) framework (Liu et al., 2014; Powell et al., 2021). The resulting genotype data will be combined with phenotypic measurements of dorsal fin traits to produce the input files necessary for QTL mapping. This approach leverages GATK's established framework, ensuring accurate variant calling while maintaining compatibility with downstream analyses.

Genome-wide QTL scans will be conducted using R/qtl to identify loci associated with dorsal fin traits (Broman et al., 2003). Significance thresholds will be determined through permutation testing. Confidence intervals for detected QTL will be estimated to assess the precision of each signal. Effect sizes and potential interactions among loci will be modeled to distinguish major-effect loci from polygenic contributions, providing insight into the genetic architecture of dorsal fin exaggeration. Identified QTL will be compared with candidate regions previously reported in *Xiphophorus* (Powell et al., 2021) to investigate potential convergent genetic mechanisms underlying sexually selected fin traits across poeciliids. These loci will help reveal how sexually selected traits are encoded genetically and how they may evolve under selection.

Several challenges may arise during this project, along with strategies to mitigate their impact. Low sequencing coverage could reduce the accuracy of rare allele detection; to address this, coverage can be increased for a subset of individuals to validate critical loci, or imputation strategies leveraging parental genotypes can be applied. The potentially complex or polygenic

architecture of dorsal fin traits may reduce power to detect individual QTL; multi-locus or Bayesian mapping approaches can be employed, and the F2 sample size can be increased if initial scans suggest widespread polygenic effects. Finally, environmental variation may influence phenotypic measurements, so rearing conditions will be standardized to control for non-genetic sources of variation.

Results:

To establish a robust pipeline for processing low-coverage whole-genome sequencing (lcWGS) data and generating high-quality genotype likelihoods, I first validated the workflow using human genomic data as a proof-of-concept. Reference genomes were indexed using Burrows-Wheeler Aligner (BWA) (Li & Durbin, 2009) and paired-end reads were aligned using the BWA-MEM algorithm (which identifies Maximal Exact Matches), producing Binary Alignment/Map (BAM) files (Li, 2013). These BAM files were subsequently sorted, duplicates marked, read groups added, and indexed using the Sequence Alignment/Map (SAM) toolkit (Li et al., 2009) to ensure compatibility with downstream variant calling. Variant calling was performed using the Genome Analysis Toolkit (GATK; McKenna et al., 2010; DePristo et al., 2011) with the HaplotypeCaller algorithm run in reference-confidence mode to produce genomic Variant Call Format (gVCF) files, which contain per-site genotype likelihoods. These single-sample gVCFs were subsequently genotyped into standard Variant Call Format (VCF) files. This process successfully traversed multiple chromosomes, confirming the pipeline's ability to handle large eukaryotic genomes. Test runs verified that all components – BWA, SAMtools, and GATK – function correctly in a WSL2 Linux environment on a Windows workstation, and outputs matched expected formats, demonstrating reproducibility and reliability.

To validate the QTL mapping portion of the pipeline, I used the *listeria* dataset, an example F2 intercross dataset provided in R/qtl (Broman et al., 2003). Using this dataset, I confirmed that genotypes and phenotypes could be correctly imported and inspected, missing data visualized, and segregation ratios checked. I then calculated genotype probabilities, performed genome scans using Haley-Knott regression (Haley & Knott, 1992), and conducted permutation tests to determine genome-wide significance thresholds. For significant loci, I estimated 95% confidence intervals and visualized marker effects, including the position of peak markers within confidence intervals (Figures 3-5). This analysis demonstrates that the workflow is fully operational for F2 intercross data and can be directly applied to my *Poecilia* sequencing data once it becomes available.

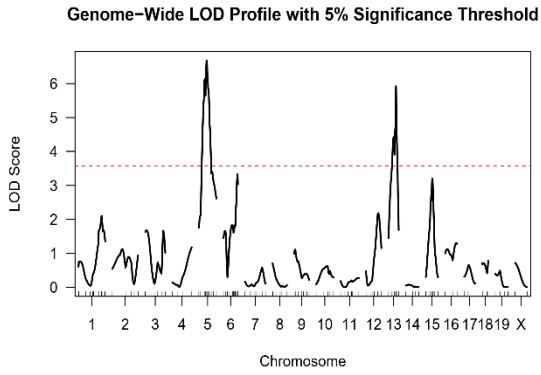

Figure 3: The genome-wide QTL scan with the permutation-derived 5% significance threshold overlaid (red horizontal line). This figure highlights genomic regions indicating statistically significant QTL (5 & 13).

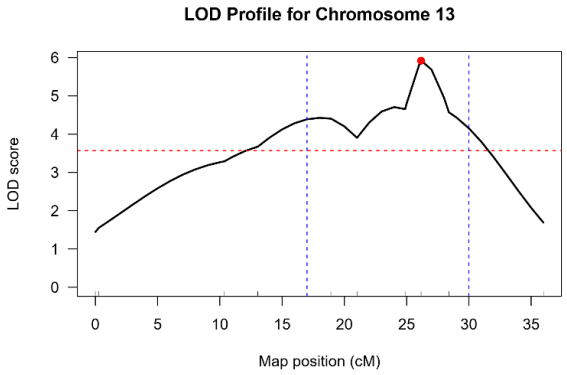

Figure 4: Chromosome-specific LOD profile for chromosome 13 including confidence interval boundaries (blue vertial lines) and the peak marker (red circle) associated with the strongest signal, D13M147.

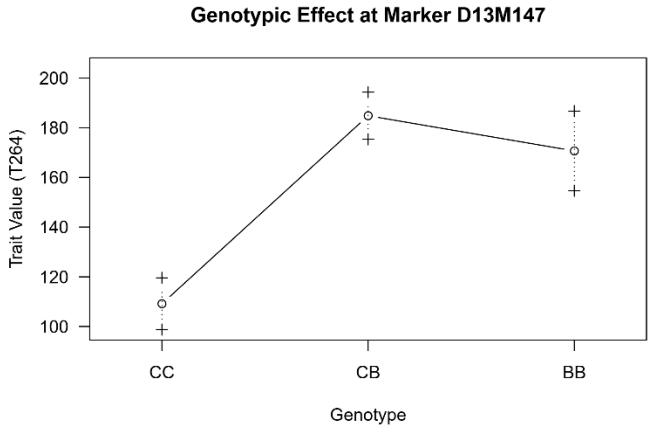

Figure 5: Effect of genotype classes at peak marker (D13M147) on the phenotype (T264). This plot visualizes how phenotypic values differ among genotypes, providing insight into the direction and magnitude of the QTL effect.

With the pipeline established and validated, application to my F2 *Poecilia* sequencing data will require updating only the reference genomes, input FASTQ filenames, and phenotype data. Once these inputs are available, the workflow can efficiently generate high-quality genotype likelihoods, enable accurate inference of ancestry-informative markers, and prepare input files for genome-wide QTL mapping of dorsal fin traits. The full workflow, including scripts and example data, is available on my GitHub page.

Future Directions:

If this project is expanded into a full research program with publication goals, several avenues could be pursued. First, candidate loci identified through QTL mapping could be validated and functionally characterized using RNA-seq in developing dorsal fins of F2 hybrids and parental species, linking genotype to gene expression and developmental mechanisms. Second, high-priority loci could be further investigated with targeted functional assays, leveraging existing poeciliid genetic tools, to directly test their contributions to dorsal fin exaggeration and sexually selected ornamentation. Finally, comparative analyses across other poeciliid species or independent ornamental traits (e.g., *Xiphophorus* swordtails) could explore whether similar genetic and developmental pathways underlie convergent evolution of fin exaggeration, providing broader insight into the evolution of sexually selected traits in livebearing fishes.

References:

Berbel-Filho, W.M., Reich, T., Ubeda, F., Fyon, F., Schlupp, I. Making a species in the lab: attempts to recreate the origin of the Amazon molly. (Preparing for submission to *Evolution*).

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

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., & Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–498. https://doi.org/10.1038/ng.806

Farr, J. A., Travis, J., & Trexler, J. C. (1986). Behavioural allometry and interdemic variation in sexual behaviour of the sailfin molly, *Poecilia latipinna* (Pisces: Poeciliidae). *Animal Behaviour*, *34*(2), 497–509. https://doi.org/10.1016/S0003-3472(86)80118-X

Goldberg, D. L., Landy, J. A., Travis, J., Springer, M. S., & Reznick, D. N. (2019). In love and war: The morphometric and phylogenetic basis of ornamentation, and the evolution of male display behavior, in the livebearer genus *Poecilia*. *Evolution*, *73*(2), 360–377. https://doi.org/10.1111/evo.13671

Haley, C. S., & Knott, S. A. (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity*, *69*(4), 315–324. https://doi.org/10.1038/hdy.1992.131

Keong, B. P., Siraj, S. S., Daud, S. K., Panandam, J. M., & Rahman, A. N. A. (2014). Identification of quantitative trait locus (QTL) linked to dorsal fin length from preliminary linkage map of molly fish, *Poecilia* sp. *Gene*, *536*(1), 114–117. https://doi.org/10.1016/j.gene.2013.11.068

Lander, E. S., & Botstein, D. (1989). Mapping Mendelian Factors Underlying Quantitative Traits Using RFLP Linkage Maps. *Genetics*, *121*(1), 185–199. https://doi.org/10.1093/genetics/121.1.185

Li, H. (2013). *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM* (No. arXiv:1303.3997). arXiv. https://doi.org/10.48550/arXiv.1303.3997

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

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Liu, K. J., Dai, J., Truong, K., Song, Y., Kohn, M. H., & Nakhleh, L. (2014). An HMM-Based Comparative Genomic Framework for Detecting Introgression in Eukaryotes. *PLOS Computational Biology*, *10*(6), e1003649. https://doi.org/10.1371/journal.pcbi.1003649

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Powell, D. L., Payne, C., Banerjee, S. M., Keegan, M., Bashkirova, E., Cui, R., Andolfatto, P., Rosenthal, G. G., & Schumer, M. (2021). The Genetic Architecture of Variation in the Sexually Selected Sword Ornament and Its Evolution in Hybrid Populations. *Current Biology*, *31*(5), 923-935.e11. https://doi.org/10.1016/j.cub.2020.12.049

Reznick, D. N., Furness, A. I., Meredith, R. W., & Springer, M. S. (2017). The origin and biogeographic diversification of fishes in the family Poeciliidae. *PLOS ONE*, *12*(3), e0172546. https://doi.org/10.1371/journal.pone.0172546

Reznick, D. N., Travis, J., Pollux, B. J. A., & Furness, A. I. (2021). Reproductive Mode and Conflict Shape the Evolution of Male Attributes and Rate of Speciation in the Fish Family Poeciliidae. *Frontiers in Ecology and Evolution*, *9*, 639751. https://doi.org/10.3389/fevo.2021.639751

Schneider, C. A., Rasband, W. S., & Eliceiri, K. W. (2012). NIH Image to ImageJ: 25 years of image analysis. *Nature Methods*, *9*(7), 671–675. https://doi.org/10.1038/nmeth.2089