Project Summary

Overview:

Natural environments are often heterogeneous, with strong spatial variation in both biotic and abiotic characteristics that can impose divergent selection between populations. These populations can then evolve differences in morphology, physiology, behavior, or life history traits that provide a fitness advantage under their respective local conditions. Such local adaptation is often facilitated by structural variants within genomes, such as inversions, which genomic data have revealed to be increasingly ubiquitous across the tree of life. I will integrate data from common garden experiments, cutting-edge genomic analyses, and computational evolutionary modeling to investigate how inversions facilitate local adaptation in wild populations. My proposed research will take place under the guidance of Dr. Nina Overgaard Therkildsen at Cornell University, and will use the large inversion regions and strong locally adapted phenotypes in the Atlantic silverside (*Menidia menidia*) as a model to test how recombination suppression within inversions can facilitate environmentally mediated local adaptation. Key outcomes of this project will be: (1) identification of regions within inversions associated with common garden phenotypes, to investigate whether recombination suppression or another evolutionary mechanism is the primary driver of inversion-related local adaptation; (2) a population-wide selection scan to determine whether differentiation is greatest at inversion breakpoints or at candidate loci within inversions; (3) simulations to quantify the parameter space within which recombination suppression is a viable evolutionary mechanism maintaining local adaptation; (4) meaningful independent research experiences for undergraduate students; and (5) workshops on reproducible research for early career graduate students.

Intellectual Merit:

Adaptive divergence and speciation are fundamental processes underlying the generation and maintenance of biodiversity, and structural genomic variants (such as inversions) are increasingly recognized as important sources of genetic variation and drivers of differentiation across the tree of life. Despite their ubiquity, the rules and mechanisms governing how inversions facilitate local adaptation and population divergence have rarely been tested in wild populations, as it is challenging to investigate these questions at the genomic resolution required and prior studes have led to inconclusive results. I will combine experiments, cutting-edge genomic technologies, and computational modeling to contribute to understanding the underlying principles connecting heterogeneity at the molecular level (i.e., genomic rearrangements) to locally adapted phenotypes.

Broader Impacts:

While conducting my research, I will also work to meet two additional goals: (1) advocating for, practicing, and teaching other researchers about best practices in open science and reproducible research; and (2) facilitating research experiences for undergraduate students. Involving students in authentic research experiences enhances interest in scientific careers and develops critical thinking, communication, and computational skills that will advantage them in future careers. I will mentor three Cornell University undergraduate students in independent research related to **QI** of my project, including working with these students to develop proposals and apply for funding for their research. To educate early-career researchers about best practices in reproducible research, I will host two online workshops for upper-division undergraduate and early-career graduate students. Workshop participants will be recruited from minority-serving institutions in the region and will be taught ethics, workflows, and tools for conducting open science.

Investigating evolutionary mechanisms that facilitate local adaptation via inversions in the Atlantic silverside

1 Introduction

Structural genomic variation in natural populations suppresses recombination and can contribute greatly to local adaptation and speciation. In particular, large genomic inversions have attracted significant attention because inverted sequences suppress recombination, which allows differentiation to accumulate and persist between alternate inversion arrangements. Limited recombination also allows linkage to persist between long tracts of alleles, resulting in alleles on one arrangement evolving independently from alleles on the other [1]. While inversions were first recognized in the 1920s in studies of *Drosophila*, genome-scale data have increasingly demonstrated that these structural variants are taxonomically widespread and important sources of genetic variation within species [2, 3]. Both recent and older examples [e.g., 4, 5] demonstrate the important role that inversions can play in facilitating local adaptation and ecological speciation [3]. **Despite the ubiquity of inversions across the tree of life and their important role in facilitating rapid differentiation between populations, the evolutionary mechanisms contributing to inversion-associated local adaptation remain largely untested.** Here, I will combine experimental, observational, theroretical, and computational approaches to test competing hypotheses for the evolutionary mechanisms contributing to inversion-associated local adaptation, using the Atlantic silverside (*Menidia menidia*) study system.

2 Background

Chromosomal inversions can promote local adaptation via three different mechanisms, as outlined by Coughlan and Willis [6] and Faria et al. [7]: (H1) inversions do not contribute to local adaptation *per se*, but instead either hitchhike along with nearby beneficial mutations or are associated with locally adapted alleles by chance; (H2) inversions play a direct role in adaptation via gene disruption at inversion breakpoints and experience positive selection; or (H3) inversions are favored due to suppressing recombination among sets of locally adapted alleles (Fig. 1). Inversion-related recombination suppression may be particularly important in facilitating both local adaptation and rapid speciation with gene flow via the maintenance of co-adapted gene complexes. This "recombination suppression hypothesis" (*H3*) is often invoked to explain associations between inversions and locally adapted traits [3]. However, few studies move beyond speculation to investigate evidence for this mechanism (but see [6, 8, 9]), and the relative importance of recombination suppression in local adaptation for wild populations remains unclear.

The Atlantic silverside study system. Atlantic silversides (*Menidia menidia*) exhibit an extraordinary degree of local adaptation, high levels of gene flow and very low genome-wide differentiation among populations (median $F_{ST} = 0.006$; [10]), extremely large effective population sizes (N_e >100 million; [11]), and a large amount of standing genetic variation [12]. They are distributed across a steep thermal gradient in estuarine waters of the North American east coast [13] and exhibit countergradient variation: individuals grow faster in the shorter growing seasons of the north, while trade-offs with predator avoidance have selected for slower growth in the south [14]. In addition to growth rate, silversides show latitudinal clines in number of vertebrae, temperature-dependent sex determination, lipid storage, spawning temperature, egg production, and hatch size of offspring [13]. Correlations among locally adapted traits suggest that suites of co-adapted alleles associated with ecological differences may be genetically linked [15].

Recent whole genome sequence data have revealed large blocks of strongly differentiated SNPs (F_{ST} >0.95) fixed for alternate alleles in northern and southern silverside populations [10]. These regions of elevated differentiation correspond to large inversion regions distributed across the silverside genome [12]. A quantitative trait locus (QTL) mapping experiment has demonstrated that several of these inversions do affect key locally adapted traits, but the exact genomic variants contributing to locally adapted phenotypes have not been identified.

Investigating inversion regions. Strong linkage between alleles that individually confer fitness advan-

tages in the local environment can maintain suites of locally adaptive traits despite ongoing gene flow between populations in different selective environments. However, this linkage also makes it difficult to locate genes *within* inversions contributing to local adaptation. This fact renders classical genetic crosses inadequate for identifying the targets of selection within an inverted region, which in turn makes it difficult to identify the evolutionary mechanisms involved in linking inversion genotypes to phenotypes—and therefore to distinguish between the hypotheses proposed above for how inversions can promote local adaptation. However, genetic exchange between alternative inversion types is not completely prevented, but only suppressed. Linkage within the inversion can be disrupted by both gene conversion and double crossovers (termed "gene flux"; [8]), which can occur at low levels, particularly in regions away from inversion breakpoints [16]. While the rate of gene flux in each generation is low, these events can accumulate within populations over many generations when heterokaryotypes persist. This rare gene flux can allow population-level sampling to serve as a natural proxy for crossing experiments, allowing identification of individual loci under selection within inversions in situations where three criteria are met: (1) inversion heterokaryotypes occur within populations, (2) effective population sizes are large, and (3) sufficient time has occurred since origination of inversions to allow for many generations of gene flux [8]—all characteristics of the Atlantic silverside system.

3 Research Aims and Methods

I will use a model study system with clear locally adapted traits linked to large inversion regions to investigate support for the recombination suppression hypothesis as a *rule of life* governing the role of inversions in local adaptation in wild populations (Fig. 1, *H3*). To do this, I will use a combination of common garden experiments, whole genome sequencing, and evolutionary simulations to address the following questions: **(Q1)** What is the genetic architecture of locally adapted traits affected by inversions? **(Q2)** Do patterns of differentiation indicate positive selection within inversions? **(Q3)** Under what scenarios is recombination suppression a viable mechanism for enabling local adaptation?.

Question I. ***What is the genetic architecture of locally adapted traits affected by inversions?***

Hypothesis I. If recombination suppression is facilitating local adaption, there will be multiple physically distant loci within the inversion associated with local adaptation.

Alternative Hypothesis I. If recombination suppression is not facilitating local adaptation, there will not be multiple physically distant loci associated with locally adapted phenotypes.

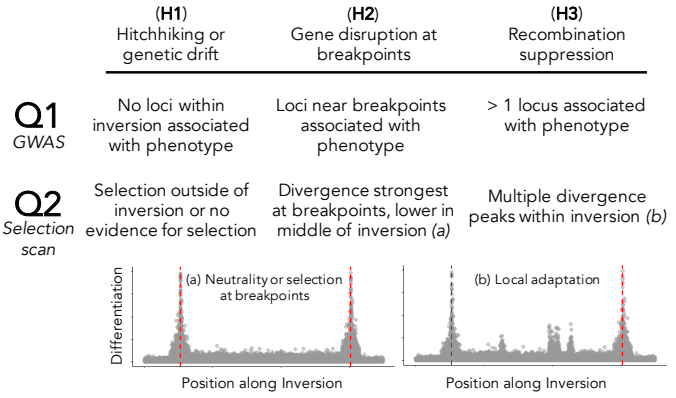

Figure 1: Hypotheses for the evolutionary mechanisms involved in inversion-associated local adaptation, as outlined by [6], with associated predictions for **Question I** and **Question II** under each mechanism. Plots demonstrate expectations for divergence patterns within the inverted region under (a) neutrality or selection only at breakpoints, and (b) local adaptation where multiple loci contribute to increased fitness [17] .

For recombination suppression between locally adapted alleles to contribute to the maintenance of an inversion, there must be multiple locally adapted alleles located within the inversion. To locate genomic regions associated with locally adapted phenotypes in Atlantic silversides, I will take advantage of the gene flux occurring in wild populations of Atlantic silversides by sampling one mid-latitude population harboring all three inversion karyotypes for the two largest inversion regions: chromosome 18 (Chr18) and chromsome 24 (Chr24). Inversions on Chr18 and Chr24 are associated with differences in growth rates and vertebral number between northern and southern karyotypes (Therkildsen Lab, unpublished data). In addition, variants on chromosome 24 are associated

with rapid growth rate shifts following experimental size-selective harvest of silversides [18]. I will use a genome-wide association study (GWAS) framework to identify the location of alleles within Chr18 and Chr24 inversions that are associated with local adaptation in Atlantic silversides.

Methods. I will collect adult Atlantic silversides during peak spawning times (February-March) from Oregon Inlet, NC (OINC). When previously sampled, both Chr18 and Chr24 inversions were at intermediate frequencies at OINC and the homo- and heterokaryotypes were found in roughly Hardy-Weinberg proportions (Fig. 2; [10]). For each collected fish (n≈400), I will take a fin clip for genomic analysis. I will use a TaqMan-based genotyping assay of alleles in the inverted region—developed using the Atlantic silverside reference genome [12]—to determine the inversion karyotype of each fish. These fish will be pooled and strip spawned following established protocols for the species (i.e., as in [19]). The offspring will then be **raised in a common garden experiment** in facilities previously used for raising Atlantic silversides (e.g., [19]) in collaborator Dr. Hannes Baumann's facilities at the Rankin Seawater Facility (University of Connecticut Avery Point). Rearing fish in the lab will allow me to phenotype these fish for environmentally-mediated traits in a controlled environment. Embryos will be randomly divided among two different rearing temperature treatments, with six replicates for each temperature (Fig 2). A subsample of fish will be measured every 2 days to track growth rate. When offspring reach 40mm in length (∼100-135 days post-hatch; [20]), a subset of ∼600 total fish will be genotyped using **haplotagging** [21] coupled with highly cost-effective **low-coverage whole genome sequencing** [22]. They will also be measured for a suite of morphological and physiological traits: body shape, number of vertebrae, and swimming efficiency [13, 15]. Haplotagging is a linked-read sequencing preparation method [21] which enables haplotype-resolved genotyping of these fish, and therefore tracking of which inversion type a given variant occurs on in heterokaryotypic fish. Haplotagging has been successfully implemented in pilot studies in the Therkildsen lab.

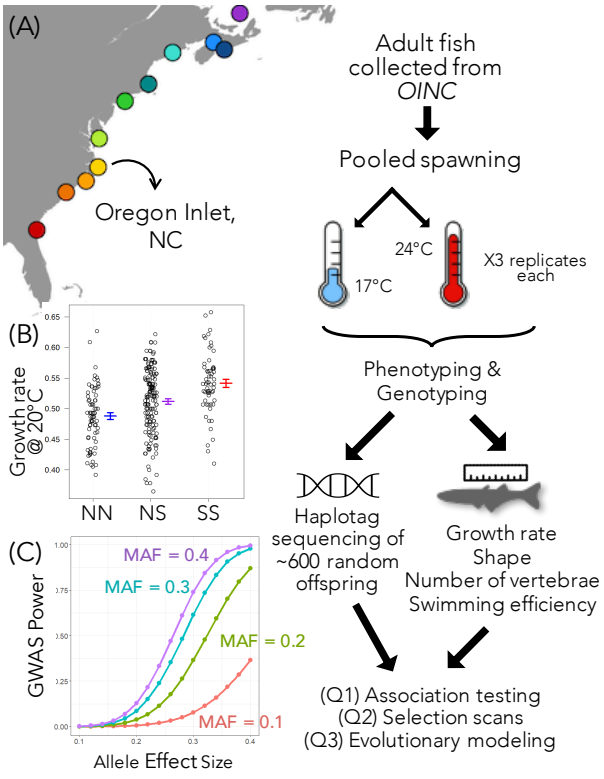

Figure 2: Experimental design for **Question I**. (A) Map dots indicate previously sequenced populations along the East Coast of the USA and Canada, which will be used in **Question II**. Fish will be collected from OINC and pooled for spawning. (B) A recent QTL study found differences in mean growth rate among northern homokaryotypes (NN), heterokaryotpes (NS), and southern karyotypes (SS) for the Chr24 inversion, although substantial variation occurs within each inversion type. Offspring will be randomly divided among two rearing temperatures and three biological replicates, raised to ∼40mm size, genotyped and phenotyped, and used for GWAS to identify loci within inversions associated with the adaptive phenotypes. (C) Power analyses for GWAS indicate that sample sizes are sufficient to detect loci of large effect.

I will align haplotagged WGS reads to the Atlantic silverside reference genome [12], phase and impute haplotypes (following [21]), and calculate genotype likelihoods (e.g, using `STITCH` and `ANGSD`; [23, 24]). From genotype likelihood estimates and phenotypic measurements, I will then conduct **association tests for each phenotypic character using methods appropriate for low-coverage genomic data**, which incorporate uncertainty into a generalized linear modeling framework to identify SNPs that are associated with a given phenotype (e.g., using `ANGSD-asso` and `GEMMA`; [25–27]). Preliminary power

analyses indicate that my sample sizes will give me a high probability of detecting alleles of large effect occurring at frequencies >0.2 (Fig. 2); therefore, we are comfortable with the ability of this study design and association testing method to detect candidate loci of major effect.

Within the GWAS framework, a common garden approach at high and low rearing temperatures will allow me to identify candidate loci and **link genotype to phenotype** for the genotype-by-environment (GxE) interaction that produces local adaptation in Atlantic silversides. The location of local adaptation-associated alleles of major effect *within* the inversions may provide evidence that supports the recombination suppression hypothesis for this inversion, and the combination of sequencing individuals raised at low/high temperatures and using haplotagging will allow these inferences to occur at finer resolutions than possible in most inversion-focused studies.

Question II. ***Do patterns of differentiation indicate positive selection within inversions?*** .

Hypothesis II. If recombination suppression facilitates local adaptation, then populations with divergent phenotypes will have multiple peaks of high genetic divergence across inversion regions.

Alternative Hypothesis II. If genetic drift or positive selection for gene disruption at inversion breakpoints, rather than recombination suppression, is the main mechanism facilitating local adaptation, then the strongest divergence should occur at inversion breakpoints.

To further investigate support for the recombination suppression hypothesis for inversions facilitating local adaptation, I will combine genomic data from **QI** with previously collected sequencing data to characterize population-wide patterns of differentiation within inversions between inversion types. From this, I will weigh evidence for direct selection on sites within the Chr24 inversion, or if selection patterns instead indicate selection on inversion breakpoints (Fig 1).

Methods. For this objective, I will use previously collected and sequenced Atlantic silversides from 12 locations along the Atlantic coast of North America (n = 50 genomes per location; map in Fig. 2). These fish have been sequenced using low-coverage ($\sim$1.3x) whole genome sequencing [10], and these data will be combined with haplotagged WGS data from **QI**.

I will analyze divergence among these WGS data using sliding window approaches along the length of the inverted regions (e.g. using `ANGSD`; [24]). I will assess divergence patterns (F_{ST} and d_{xy}) of loci within inversion regions, comparing groups with different karyotypes. Under neutrality, or when selection acts directly on inversion breakpoints, we expect a pattern with maximal divergence at the breakpoints where recombination is maximally suppressed; under local adaptation, we might expect additional peaks of divergence away from the breakpoints that are shaped by the interplay between selection, mutation, and gene flux (Fig. 1) [17, 28]. Here, I will use candidate regions identified in **QI** to examine whether candidate regions identified in GWAS correspond to regions of peak differentiation. Additionally, I will use coalescent models to determine whether the divergence peaks are outside of expectations under the null model of evolution (as recommended by [28, 29]).

Question III. ***In what scenarios is recombination suppression a viable mechanism for local adaptation?***

Despite the ubiquity with which the recombination suppression hypothesis is invoked to explain the evolutionary roles of inversions, the scenarios in which recombination suppression is sufficient to lead to local adaptation are not well understood. I will use evolutionary models parameterized by the results from **QI** and **QII** to simulate evolutionary scenarios and determine which situations are sufficient for maintaining local adaptation.

Methods. I will use the results from **QI** and **QII** to parameterize models for forward evolutionary simulations that allow heterogeneous recombination rates (e.g., in `SLiM3` or `nemo`; [30, 31]) and determine the parameter space in which recombination suppression maintains an inversion (similar to simulations in [32]). In these simulations, I will vary recombination rates among loci, population sizes, selection on loci within linked regions, and number of loci contributing to locally adapted phenotypes. If **QI** and **QII**

are consistent with the recombination suppression hypothesis, then these models will simulate how the large inversions may have become associated with local adaptation in Atlantic silversides. Alternatively, if the results are not consistent with recombination suppression, then these simulations will quantify what parameter space would allow for recombination suppression to occur and lead to local adaptation associated with the inversions.

4 Significance

Adaptive divergence and speciation are fundamental processes underlying the emergence and maintenance of biodiversity, and the two processes are linked when adaptive divergence promotes speciation (i.e., ecological speciation; [3, 33]). Structural variants are increasingly recognized as important sources of genetic variation and drivers of differentiation across the tree of life, particularly in populations with ongoing gene flow [34]. However, it remains unclear how exactly these structural variants facilitate divergence. Therefore, this research will contribute to understanding the underlying principle connecting heterogeneity at the molecular level (i.e. genomic rearrangements) to environmentally-mediated phenotypes, and therefore to the evolutionary trajectory of a population.

5 Broader Impacts of the Proposed Work

The broader impacts of this project have two main tenets: (1) advocating for and practicing best practices in open science and reproducible research; and (2) facilitating critical thinking, communication, and computational skill learning in high school, undergraduate, and early-career graduate students. Science writ large—and the fields of ecology and evolutionary biology in particular—increasingly recognizes the importance of open science for moving research forward [35]. However, a large barrier remains for most scientists in their beliefs that making their research open and reproducible requires time, effort, and skills that they lack [36]. This common misconception can be mitigated through education and an increased familiarity with the concepts and tools associated with open science.

Data Access and Dissemination: All data and code associated with this research will be made available in online repositories (i.e., *Zenodo, Dryad, GitHub*), making them open access and free to everyone. All manuscripts associated with this work will be posted on preprint servers (i.e., *bioRxiv*) and available open access. I will also present the work at regional and national scientific conferences (e.g., the Evolution Meeting and American Fisheries Society Meeting), and these presentations will be made available virtually.

Mentoring to Facilitate Critical Thinking, Communication, and Computational Skills: Involving students in scientific research increases interest in careers in science fields [37] and strengthens critical thinking and problem solving skills that are essential no matter their future careers. I will engage undergraduate learners of diverse backgrounds with the scientific process by hiring three undergraduate researchers to develop independent projects related to the common garden experiments in **QI**. At least one of these students will be involved with either the McNair Scholars or Cornell Office of Academic Diversity Initiatives Research Scholars Program, and I will work with and mentor all in applying to fund their research. I will preferentially recruit students belonging to marginalized groups for these positions.

I will also host two online three-day workshops over spring break in 2022 and 2023, developed in collaboration with Dr. Therkildsen. In these workshops, I will teach upper division undergraduate and early-career graduate students about reproducible research workflows and tools (e.g., those identified in [36]), building upon previous workshops I have taught on using LaTeX and R, and on Dr. Therkildsen's current data science tools course. Workshop participants will be recruited from minority-serving institutions in the region (i.e., State University of New York colleges) and give special consideration to minority-identifying students. Each of the 8-10 students per workshop will receive a small stipend for attending, funded either from my NSF PRFB or small grants. Hosting these workshops online reduces costs and increases accessibility to students who may not be able to travel to an in-person workshop.

6 Training Objectives

Link to Career Development and Future Research. My ultimate goal is to join the faculty at a liberal arts college where I will combine my interests in (1) empowering students to be creative, critical thinkers, and (2) conducting rigorous scientific research at the interface between phylogenetics and population genetics. The NSF PRFB will prepare me for this goal by providing an unparalleled opportunity to facilitate my development and training as both a researcher and educator.

In future research, I am interested in continuing to investigate the influences of reduced or enhanced gene flow on population differentiation, diversification, and speciation, and quantifying how changing environments influence these processes. For my M.Sc. research, I studied how patterns of genetic diversity changed over the course of three years of harvest in the wolf population in northern Minnesota [38], acquiring a strong foundation in population genetics methods. In my PhD, I transitioned to using next-generation genomic data to investigate both population genomic and phylogenomic questions in the pelagic fishes of Lake Tanganyika, East Africa. I used reduced-representation and WGS to characterize contemporary patterns of genetic diversity [Rick et al., in prep; 39] and infer historical processes leading to these patterns, including environment-mediated admixture and diversification events [Rick & Wagner, in prep]. In both of these projects, I have worked at the interface between population- and species-level investigations to describe patterns of genomic diversity and evolutionary patterns, but found myself limited in the inferences of process leading to these patterns.

My proposed research builds upon my prior work by digging deeper into inferring evolutionary processes contributing to population differentiation by combining of laboratory experiments, population-wide sampling, and simulations. In addition to this conceptual advance, it will help me develop bioinformatic skills for working with whole genome data and probabilistic models based on genotype likelihoods.

Sponsoring Scientist and Host Institution. For this project, I will be working with Dr. Nina Overgaard Therkildsen at Cornell University. Dr. Therkildsen is a population and conservation genomicist researching the genomic basis and evolutionary consequences of local adaptation, with a primary focus on marine fishes. The genomic basis of local adaptation in the Atlantic silverside has been a key focus for her research program in recent years, and the proposed work is an extension of ongoing research in her lab. Cornell additionally offers cutting-edge expertise in population genomics and genotype-phenotype association mapping through its Center for Comparative and Population Genomics and Center for Vertebrate Genomics, as well as professional development opportunities through its Office of Postdoctoral Studies and Center for Teaching Innovation, which I will be well-poised to take advantage of.

7 Timeline

Pilot data related to this work will be collected in Spring and Summer 2021 by current Therkildsen lab members. My NSF PRFB research will begin in September 2021 and follow the schedule in Fig. 3.

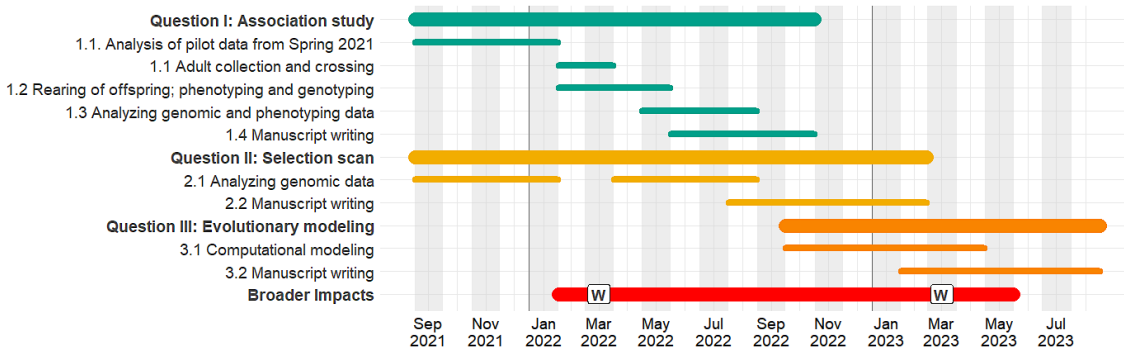

Figure 3: Gantt chart detailing my proposal timeline. For broader impacts, "W" indicates a workshop, while the bar indicates periods spent mentoring undergraduate researchers.

References Cited

[1] Kirkpatrick, M., 2010: How and why chromosome inversions evolve. *PLoS Biology*, **8 (9)**, doi:10.1371/journal.pbio.1000501.

[2] Kirkpatrick, M. and N. Barton, 2006: Chromosome inversions, local adaptation and speciation. *Genetics*, **173 (1)**, 419–434, doi:10.1534/genetics.105.047985.

[3] Wellenreuther, M. and L. Bernatchez, 2018: Eco-Evolutionary Genomics of Chromosomal Inversions. *Trends in Ecology and Evolution*, **33 (6)**, 427–440, doi:10.1016/j.tree.2018.04.002, `http://dx.doi.org/10.1016/j.tree.2018.04.002`.

[4] Noor, M. A. F., K. L. Grams, L. A. Bertucci, and J. Reiland, 2001: Chromosomal inversions and the reproductive isolation of species. *Proceedings of the National Academy of Sciences*, **98 (21)**, 12 084–12 088, `www.pnas.orgcgidoi10.1073pnas.221274498`.

[5] Todesco, M., G. L. Owens, N. Bercovich, J. S. Légaré, S. Soudi, D. O. Burge, K. Huang, K. L. Ostevik, E. B. Drummond, I. Imerovski, K. Lande, M. A. Pascual-Robles, M. Nanavati, M. Jahani, W. Cheung, S. E. Staton, S. Muños, R. Nielsen, L. A. Donovan, J. M. Burke, S. Yeaman, and L. H. Rieseberg, 2020: Massive haplotypes underlie ecotypic differentiation in sunflowers. *Nature*, **584 (7822)**, 602–607, doi:10.1038/s41586-020-2467-6.

[6] Coughlan, J. M. and J. H. Willis, 2019: Dissecting the role of a large chromosomal inversion in life history divergence throughout the Mimulus guttatus species complex. *Molecular Ecology*, **28 (6)**, 1343–1357, doi:10.1111/mec.14804.

[7] Faria, R., K. Johannesson, R. K. Butlin, and A. M. Westram, 2019: Evolving Inversions. *Trends in Ecology and Evolution*, **34 (3)**, 239–248, doi:10.1016/j.tree.2018.12.005.

[8] Ayala, D., S. Zhang, M. Chateau, C. Fouet, I. Morlais, C. Costantini, M. W. Hahn, and N. J. Besansky, 2019: Association mapping desiccation resistance within chromosomal inversions in the African malaria vector Anopheles gambiae. *Molecular Ecology*, **28 (6)**, 1333–1342, doi:10.1111/mec.14880.

[9] Lee, C. R., B. Wang, J. P. Mojica, T. Mandáková, K. V. Prasad, J. L. Goicoechea, N. Perera, U. Hellsten, H. N. Hundley, J. Johnson, J. Grimwood, K. Barry, T. Fairclough, J. W. Jenkins, Y. Yu, D. Kudrna, J. Zhang, J. Talag, W. Golser, K. Ghattas, M. E. Schranz, R. Wing, M. A. Lysak, J. Schmutz, D. S. Rokhsar, and T. Mitchell-Olds, 2017: Young inversion with multiple linked QTLs under selection in a hybrid zone. *Nature Ecology and Evolution*, **1 (5)**, doi:10.1038/s41559-017-0119.

[10] Wilder, A. P., S. R. Palumbi, D. O. Conover, and N. O. Therkildsen, 2020: Footprints of local adaptation span hundreds of linked genes in the Atlantic silverside genome. *Evolution Letters*, **4 (5)**, 430–443, doi:10.1002/evl3.189.

[11] Lou, R. N., N. K. Fletcher, A. P. Wilder, D. O. Conover, N. O. Therkildsen, and J. B. Searle, 2018: Full mitochondrial genome sequences reveal new insights about post-glacial expansion and regional phylogeographic structure in the Atlantic silverside (Menidia menidia). *Marine Biology*, **165**, 124, doi:10.1007/s00227-018-3380-5, `https://doi.org/10.1007/s00227-018-3380-5`.

[12] Tigano, A., A. Jacobs, A. P. Wilder, A. Nand, Y. Zhan, J. Dekker, and N. O. Therkildsen, 2020: Chromosome-level assembly of the Atlantic silverside genome reveals extreme levels of sequence diversity and structural genetic variation. *bioRxiv*, 2020.10.27.357293, doi:10.1101/2020.10.27.357293.

[13] Conover, D. O., S. A. Arnott, M. R. Walsh, and S. B. Munch, 2005: Darwinian fishery science: Lessons from the Atlantic silverside (Menidia menidia). *Canadian Journal of Fisheries and Aquatic Sciences*, **62 (4)**, 730–737, doi:10.1139/f05-069.

[14] Conover, D. O. and T. M. Present, 1990: Countergradient variation in growth rate: compensation for length of the growing season among Atlantic silversides from different latitudes. *Oecologia*, **83 (3)**, 316–324, doi:10.1007/BF00317554.

[15] Hice, L. A., T. A. Duffy, S. B. Munch, and D. O. Conover, 2012: Spatial scale and divergent patterns of variation in adapted traits in the ocean. *Ecology Letters*, **15 (6)**, 568–575, doi:10.1111/j.1461-0248.2012.01769.x.

[16] Navarro, A., E. Betrán, A. Barbadilla, and A. Ruiz, 1997: Recombination and gene flux caused by gene conversion and crossing over in inversion heterokaryotypes. *Genetics*, **146 (2)**, 695–709.

[17] Kapun, M. and T. Flatt, 2019: The adaptive significance of chromosomal inversion polymorphisms in Drosophila melanogaster. *Molecular Ecology*, **28 (6)**, 1263–1282, doi:10.1111/mec.14871.

[18] Therkildsen, N. O., A. P. Wilder, D. O. Conover, S. B. Munch, H. Baumann, and S. R. Palumbi, 2019: Contrasting genomic shifts underlie parallel phenotypic evolution in response to fishing. *Science*, **365 (6452)**, 487–490, doi:10.1126/science.aaw7271.

[19] Murray, C. S. and H. Baumann, 2018: You better repeat it: Complex CO2 Ã temperature effects in Atlantic silverside offspring revealed by serial experimentation. *Diversity*, **10 (3)**, doi:10.3390/d10030069.

[20] Murray, C. S. and H. Baumann, 2020: Are long-term growth responses to elevated pCO2 sex-specific in fish? *PLoS ONE*, **15 (7 July)**, 1–21, doi:10.1371/journal.pone.0235817, `http://dx.doi.org/10.1371/journal.pone.0235817`.

[21] Meier, J., P. Salazar, M. Kučka, R. W. Davies, A. Dréau, I. Aldás, O. B. Power, N. Nadeau, J. Bridle, C. Rolian, N. Barton, W. O. McMillan, C. Jiggins, and Y. F. Chan, 2020: Haplotype tagging reveals parallel formation of hybrid races in two butterfly species. *bioRxiv*, doi:10.1101/2020.05.25.113688.

[22] Therkildsen, N. O. and S. R. Palumbi, 2017: Practical low-coverage genomewide sequencing of hundreds of individually barcoded samples for population and evolutionary genomics in nonmodel species. *Molecular Ecology Resources*, **17 (2)**, 194–208, doi:10.1111/1755-0998.12593.

[23] Davies, R. W., J. Flint, S. Myers, and R. Mott, 2016: Rapid genotype imputation from sequence without reference panels. *Nature Genetics*, **48 (8)**, 965–969, doi:10.1038/ng.3594.

[24] Korneliussen, T. S., A. Albrechtsen, and R. Nielsen, 2014: ANGSD: Analysis of next generation sequencing data. *BMC Bioinformatics*, **15**, 356, doi:10.1007/978-1-4614-1174-1_8.

[25] Jørsboe, E. and A. Albrechtsen, 2020: Efficient approaches for large scale GWAS studies with genotype uncertainty. *bioRxiv*, 1–20, doi:10.1101/786384.

[26] Skotte, L., T. S. Korneliussen, and A. Albrechtsen, 2012: Association Testing for Next-Generation Sequencing Data Using Score Statistics. *Genetic Epidemiology*, **36 (5)**, 430–437, doi:10.1002/gepi.21636.

[27] Zhou, X. and M. Stephens, 2014: Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature Methods*, **11 (4)**, 407–409, doi:10.1038/nmeth.2848.

[28] Guerrero, R. F., F. Rousset, and M. Kirkpatrick, 2012: Coalescent patterns for chromosomal inversions in divergent populations. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **367 (1587)**, 430–438, doi:10.1098/rstb.2011.0246.

[29] Kirkpatrick, M., 2017: The evolution of genome structure by natural and sexual selection. *Journal of Heredity*, Oxford University Press, Vol. 108, 3–11, doi:10.1093/jhered/esw041.

[30] Haller, B. C. and P. W. Messer, 2019: SLiM 3: Forward Genetic Simulations Beyond the Wright-Fisher Model. *Molecular Biology and Evolution*, **36 (3)**, 632–637, doi:10.1093/molbev/msy228.

[31] Guillaume, F. and J. Rougemont, 2006: Nemo: An evolutionary and population genetics programming framework. *Bioinformatics*, **22 (20)**, 2556–2557, doi:10.1093/bioinformatics/btl415.

[32] Yeaman, S., 2013: Genomic rearrangements and the evolution of clusters of locally adaptive loci. *Proceedings of the National Academy of Sciences of the United States of America*, **110 (19)**, doi:10.1073/pnas.1219381110.

[33] Nosil, P., D. J. Funk, and D. Ortiz-Barrientos, 2009: Divergent selection and heterogeneous genomic divergence. *Molecular Ecology*, **18 (3)**, 375–402, doi:10.1111/j.1365-294X.2008.03946.x.

[34] Oomen, R. A., A. Kuparinen, and J. A. Hutchings, 2020: Consequences of Single-Locus and Tightly Linked Genomic Architectures for Evolutionary Responses to Environmental Change. *The Journal of heredity*, **111 (4)**, 319–332, doi:10.1093/jhered/esaa020.

[35] Nosek, B. A., G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, C. D. Chambers, G. Chin, G. Christensen, M. Contestabile, A. Dafoe, E. Eich, J. Freese, R. Glennerster, D. Goroff, D. P. Green, B. Hesse, M. Humphreys, J. Ishiyama, D. Karlan, A. Kraut, A. Lupia, P. Mabry, T. A. Madon, N. Malhotra, E. Mayo-Wilson, M. McNutt, E. Miguel, E. L. Paluck, U. Simonsohn, C. Soderberg, B. A. Spellman, J. Turitto, G. VandenBos, S. Vazire, E. J. Wagenmakers, R. Wilson, and T. Yarkoni, 2015: Promoting an open research culture. American Association for the Advancement of Science, 1422–1425 pp., doi:10.1126/science.aab2374.

[36] Alston, J. M. and J. A. Rick, 2021: A beginner's guide to conducting reproducible research. *Bulletin of the Ecological Society of America*, doi:10.1002/bes2.1801.

[37] Harrison, M., D. Dunbar, L. Ratmansky, K. Boyd, and D. Lopatto, 2011: Classroom-based science research at the introductory level: Changes in career choices and attitude. *CBE Life Sciences Education*, **10 (3)**, 279–286, doi:10.1187/cbe.10-12-0151.

[38] Rick, J. A., R. A. Moen, J. D. Erb, and J. L. Strasburg, 2017: Population structure and gene flow in a newly harvested gray wolf (Canis lupus) population. *Conservation Genetics*, **18 (5)**, 1–14, doi:10.1007/s10592-017-0961-7.

[39] Junker, J., J. A. Rick, P. B. McIntyre, I. Kimirei, E. A. Sweke, J. B. Mosille, B. Wehrli, C. Dinkel, S. Mwaiko, O. Seehausen, and C. E. Wagner, 2020: Structural genomic variation leads to genetic differentiation in Lake Tanganyika's sardines. *Molecular Ecology*, **29 (17)**, 3277–3298, doi:10.1111/mec.15559.

Supplementary Documentation

Data Management Plan

Both myself (PI Rick) and sponsoring scientist Dr. Therkildsen are committed to broad dissemination of all data and metadata during the timeframe of the project, and to make products from this project open and accessible to others in the scientific community. All project personnel will be trained in data management procedures in accordance with NSF requirements. Each type of data will be managed to maximize access, sharing, and preservation.

Data and Materials Produced

The research described in this proposal will lead to the collection of biological specimens (Atlantic silverside fishes), as well as phenotypic and genomic data associated with those specimens. Metadata associated with the wild-caught biological specimens includes collection locality, date of collection, body size, sex, and gamete mass. Lab-reared biological specimens will have associated metadata including spawning and hatch date, rearing temperature, body size at phenotyping, full morphometrics at 40mm size, number of vertebrae, and swimming efficiency. Genomic data for wild-caught fish will include rapid genotyping results for the major inversions. In addition, a subsample of the lab-reared fish will have haplotagged Illumina short-read sequence data (FASTQ files), processed genomic data (BAM files), and mapped genotype data (VCF files). Physical tissue samples of Atlantic silversides and genomic DNA preparations will be archived in long-term -80°C freezers in the Therkildsen lab at Cornell University. Custom analysis scripts will also be produced as necessary for data analysis.

Materials corresponding to reproducible research and data science workshops will be hosted on a dedicated workshop website and FigShare (https://figshare.com/) following workshop completion, where they will be open and freely accessed by anyone. Each workshop will also have an associated GitHub repository, which will be available to all workshop participants during and following the course.

Standards, Formats and Metadata

Where possible, all data will be stored in flat file formats for portability (i.e., phenotypic and metadata will be stored as .csv files). Short read genomic data will be retained in their raw format (i.e., FASTQ files), as well as processed (BAM files) and mapped (VCF files) formats. While experiments are in progress, electronic lab notebooks (i.e., using Open Science Framework, https://osf.io/) will be used, providing both automatic version control and access to all individuals involved in the research. When the project is complete, static versions of these lab notebooks will be saved and archived with the rest of the data. An electronic lab notebook will also be used to record bioinformatic methods used. Where possible, scripting languages (e.g., `R`, `Python`, `bash`) will be used for analyses to ensure reproducibility, and these scripts will kept in a GitHub repository, which will be uploaded to Zenodo with the corresponding data when manuscripts are published. All data will additionally be synced to local versions with automatic daily backup on Dr. Therkildsen's networked computing server. Monthly, the data and scripts will be backed up to external hard drives maintained by PI Rick and stored in a separate physical location from Dr. Therkildsen's server.

Roles and Responsibilities

PI Rick will have primary responsibility for collection and maintenance of the data generated from this project, and these data will be shared equally between PI Rick and Dr. Therkildsen.

Dissemination Methods, Data Sharing, and Public Access

Illumina sequence data from Atlantic silverside specimens (FASTQ files) will be deposited into National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA, http://www.ncbi.nlm.nih.gov/sra) upon publication of the first manuscript associated with these data. Prior to being deposited, genomic data will be stored on Dr. Therkildsen's networked computing server, with regular automated backup of all data as described above. Upon completion of analyses, VCF files from genomic data and associated metadata will be made publicly available, preserved, and citable in publications Cornell University Library's institutional repository, eCommons (https://ecommons.cornell.edu). eCommons provides each item with a persistent identifier and is committed to preserving the binary form of the digital object. These datasets will be available via the world wide web without restriction.

Dissertation Abstract - Jessica A. Rick

Dissertation Title: *Environmental influences on admixture, diversification, and gene flow in freshwater Lates of the East African Rift Lakes*

Variation in climatic conditions in space and time is a key driver of adaptation, speciation, and patterns of biodiversity. However, the links between local or regional environmental conditions and resulting patterns in evolution and species richness remain poorly understood. Theory predicts that more variable climates lead to diversification in some cases and species decline, extinction, or collapse in others, while directional shifts in climate can also lead to either diversification via ecological opportunity or species extinction due to constraints on adaptation. Though much theory exists on how climate and, relatedly, energy availability, affect patterns of biodiversity, mechanistic links between climatic variability and the production and maintenance of biodiversity remain sparse, especially at higher trophic levels.

Our ability to answer questions about links between environmental change and biodiversity has been fundamentally altered by the recent development of technologies that enable the generation of high-resolution genomic data. Recent decreases in sequencing costs and improvements in sequencing and analytical technologies have enabled generation of much more detailed estimates of species' evolutionary histories. Recent advances in DNA sequencing and marker discovery have revolutionized the collection of genome-scale population genetic data for non-model species, enhancing the ability of researchers to measure genetic diversity and characterize evolutionary processes.

Here, I have used emerging genomic tools to elucidate how organisms have responded to environmental and anthropogenic forces in the present and recent past, as well as throughout their evolutionary histories, focusing on the four endemic *Lates* species in Lake Tanganyika, East Africa. **First, I determine the effect of bioinformatic parameter selection on phylogenomic inference using next-generation sequencing data**, demonstrating that the choice of a reference genome (i.e., whether the reference genome is from an in- or outgroup taxon) interacts with choices of minor allele frequency filters to bias resulting phylogenetic trees, particularly when the true phylogeny has high levels of incomplete lineage sorting. **In my second chapter, I use reduced-representation genomic data to investigate the ways in which climatic and environmental factors shape gene flow and diversification in aquatic predators,** demonstrating that the four *Lates* species in Lake Tanganyika have no evidence for contemporary hybridization and generally do not show patterns of geographic population structure. In contrast, in my third chapter, I use a combination of whole genome and reduced-representation data to investigate evidence for historical admixture between each of the four Lake Tanganyika *Lates* species and the nearby riverine *Lates niloticus*. These admixture events may be connected to periods throughout the history of the East African Rift where populations of freshwater organisms have become disconnected and reconnected with one another as basins and waterways have formed, grown, and contracted. These findings help us to understand how historical and contemporary processes affecting gene flow have shaped the evolution of the radiation of *Lates* fishes in Lake Tanganyika, and how the lake's environmental history may have played a role in shaping how these species evolved.