Research Statement

Patterns of genetic variation in natural populations are shaped by demography, selection, and mutation and genome biology. I am broadly interested in using whole-genome sequencing data to understand how these evolutionary processes contribute to molecular and organismal diversity. To do this, I use creative mathematical and computational approaches for predicting patterns of genetic diversity to learn about demographic history, mechanisms of diversity maintenance, molecular consequences of selection, and the biology of recombination. I focus on developing theory and tools that can handle large sample sizes, many populations, and flexible model assumptions.

My lab's research will focus on the development of theory and statistical methods to study genomic data to gain insights into the evolutionary processes that shape diversity. Specifically, I plan to (Aim 1) improve methods for computing linkage disequilibrium and diversity measures to understand the effect of selection across multiple populations, and (Aim 2) study the impact of background selection on patterns of diversity and genomic inference.

1) Computational methods for evolutionary inference and association studies

The genomes of individuals living today carry the signatures of past population size changes, splits, mergers, and migrations, so that we can learn about those events by comparing model predictions to data. In addition to clarifying anthropological and ecological questions, demographic inference serves as a baseline for downstream inference of natural selection, mutation and recombination rates, and the architecture of complex traits.

Past and ongoing research

During my Ph.D., I focused on multi-allelic models to ask novel questions about patterns of selection and to improve power for inferring demography. This research combined the implementation of classical diffusion theory using new numerical methods with creative statistical inference approaches. In one application, using joint allele frequency and linkage disequilibrium measures within a single model framework, I showed that we gain significant statistical power over standard methods to infer population size change history, and I used this approach to refine size-history estimates for well-studied fruit fly populations (Ragsdale & Gutenkunst, 2017).

During my post-doc I developed both theory and computational tools to predict patterns of linkage disequilibrium and common diversity measure for large numbers of populations with complex demography (Ragsdale & Gravel, 2019, *Plos Genetics*). Using this new method, I found that standard models of human expansion history poorly describe patterns of LD in many populations. This discrepancy between model prediction and data is missed when using classical approaches, but by studying a wide array of genetic diversity statistics I was able to infer historical events that leave subtle signatures on patterns of diversity. This framework is powerful for detecting deep historical events, such as ancient population structure and archaic admixture, and I used it to infer a model of human expansion that includes the known Neanderthal admixture in Eurasian populations and a similarly deeply diverged but unidentified lineage that contributes to African populations (Fig. 1).

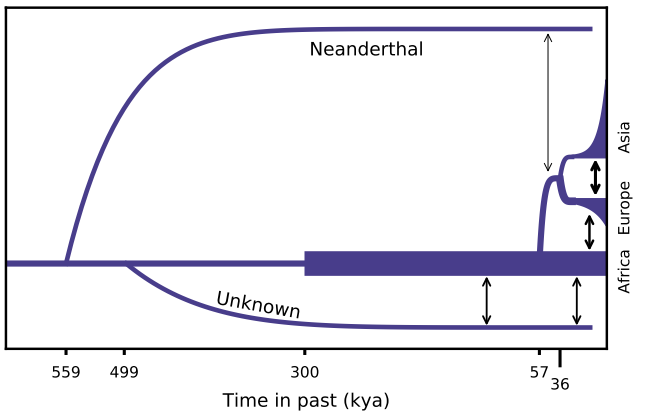

Figure 1: Inferred human expansion model, including archaic contributions to Eurasian and African populations, using multi-population LD models. Such methods are powerful for learning about deep historical events.

To make reliable inferences, we need to be able to accurately estimate relevant statistics from noisy sequencing data. This is a challenging task for LD measures. I derived unbiased estimators for a large set of relevant two-locus statistics which resolve many problems with current estimation of LD and allow for robust inference from LD decay (Ragsdale & Gravel, 2019, *MBE*). These estimators are well suited for inferring small population sizes from LD, a common analysis in conservation and domestication genetics. In ongoing work, we are finding that they also show promise in reducing bias when adjusting test statistics for confounding effects in genome-wide association studies, highlighting the diverse and sometimes unexpected directions opened up by developing novel statistical methods.

Future directions: Improving models of LD and selection for multiple populations

The model and statistical framework I developed for studying LD across multiple populations lends itself to a wide range of applications and relevant methodological extensions. This method's scalability to many populations and inclusion of new diversity statistics make it ideally suited to explore models of deep human history within Africa. I am now collaborating with Brenna Henn (UC Davis) to test competing models of human origins using a large

dataset of geographically and genetically diverse populations within Africa, with the aim of quantifying the deep population structure within human lineages that extends hundreds of thousands of years in the past. These methods will also be fruitful in the study of selection acting in multiple populations, helping in the design of multi-population association studies and understanding the limits to the transferability of risk scores across cohorts.

Aim 1A) Combining selection and recombination in models of multiple populations
Selection is known to alter patterns of LD, allowing us to locate regions impacted by natural selection. However, incorporating selection in multi-locus models is notoriously difficult, and the LD-signals of selection acting in multiple populations are largely unexplored. I implemented selection in two-locus models for single populations with flexible demography, and I have also made progress in deriving theoretical results for the multi-population case. While its implementation will be challenging, my computational background is well suited to the task, and such methods will be valuable for studying the joint effects of selection and recombination on shared patterns of LD, including when selective pressures differ across populations (Fortier et al., 2019).

Aim 1B) Quantifying the transferability of association scores across populations
Most human GWAS are performed within single-ancestry cohorts, and their transferability depends on the shared patterns of LD and allele frequencies between discovery and target populations. The predictive ability of risk scores is known to decline linearly with population divergence, and failing to account for differences in diversity patterns biases risk and trait prediction. Because LD patterns vary between populations (i.e. $\mathrm{Corr}(r_1, r_2) < 1$), our ability to accurately predict risk scores in target populations is reduced. If traits are under selection, differences in LD may be inflated beyond neutral expectations (Aim 1A), lowering that correlation and leading to reduced transferability. My multi-population LD approach provides accurate predictions of shared LD between populations, which can be used to place theoretical limits on risk score predictions in different populations. I am interested in exploring ways to use LD these models to reduce bias in polygenic trait prediction and to provide model-based recommendations for recruitment and sampling strategies for populations that are under-represented in GWAS.

2) The genomic signature of background selection and its effect on inference
Mutations that affect organismal fitness are a major driver of genetic diversity. A key parameter in evolution is the distribution of fitness effects (DFE) for new mutations, which influences everything from rates of adaptation, to the evolution of sex and recombination, and the architecture of complex traits.

Past and ongoing research
Many population genetic inferences of the DFE aggregate the effects of nonsynonymous mutations in protein-coding genes from across the genome. In my Ph.D., I refined this picture by considering the effects of different mutations that occur at the same position in a gene. By studying multi-allelic loci within protein-coding genes, I characterized the correlation of fitness effects of different mutations that occur at the same site and decomposed the contributing factors that determine a mutation's fitness effect (Ragsdale et al., 2016).

DFE analyses typically use synonymous variants as a neutral control for studying selection on nonsynonymous variants. However, background selection (BGS) is known to distort neutral variation linked to selected regions, and there is evidence that some synonymous variants may themselves be under direct selection. In an exploratory study, I used whole-genome sequences of ~ 800 French Canadian individuals to assess this assumption of neutrality of synonymous variants (Fig. 2). Using this data to infer the DFE for synonymous variants, I found that up to one half of synonymous mutations may be at least weakly selected, and that downstream analyses of selection on nonsynonymous mutations are sensitive to the assumption of neutrality for synonymous variants (Ragsdale et al., 2018). Unraveling the confounding effects of background selection, direct selection, and demography will be a key issue in evolutionary inference moving forward, and one that I plan to pursue.

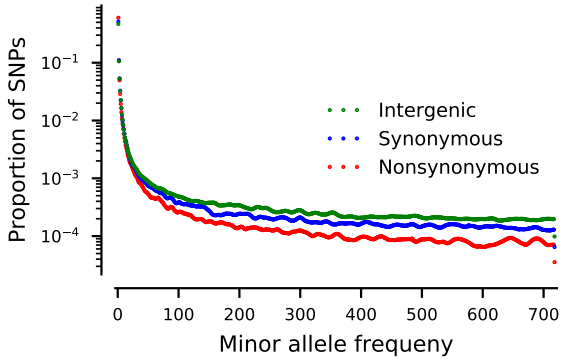

Figure 2: Nonsynonymous mutations are skewed to lower frequencies than synonymous mutations, characteristic of negative selection acting on protein-coding changes. Synonymous mutations are also skewed to lower frequencies compared to neutrally evolving intergenic variants, due to either background or direct selection on synonymous variants.

Future directions: Understanding how background selection shapes diversity in natural populations

Background selection is known to distort patterns of diversity, sometimes dramatically. Previous theoretical and simulation studies have been limited in scope, making unrealistic simplifying assumptions or focusing on summary statistics that don't reflect the full effect of BGS in large sample sizes. Furthermore, its effect in non-standard models, such as non-random mating, inbreeding and selfing, or with spacial structure, are poorly understood.

Aim 2A) Characterizing the joint effects of demography and background selection on genomic inference

Background selection is known to bias demographic and DFE inference (e.g., Ragsdale et al., 2018), but the severity of this bias is unknown and methods to correct for it are lacking. While mathematical models of BGS quickly become intractable, we are entering the era of efficient large-scale whole-genome simulation. Such methods allow for the realistic simulation of whole genomes with arbitrary selection models and quantitative traits, which we can use to build predictions for the effect of BGS on diversity and explore how it may bias genomic inferences. My ongoing collaborations with simulation software developers and work to improve statistical approaches using simulated data place me in a good position to make headway on this problem. Using large-scale simulations and whole-genome sequencing data, I plan to i) characterize the dynamics of mutation trajectories in the presence of BGS, and ii) develop and test methods for mitigating downstream biases in inference caused by BGS.

Aim 2B) Studying background selection in non-random mating, inbreeding, and spatially structured populations

Most population and quantitative genetics models ignore the spatial distribution of individuals and assume that populations can be described as large, randomly mating groups. Natural populations break these rules. The spatial distribution of individuals and their mate choices can strongly affect patterns of local genetic diversity, and it can alter rates of adaptive evolution, the efficacy of selection, and distortions of genetic diversity due to linked selection. Again making use of recent developments for large-scale whole-genome simulations, along with theoretical developments for LD and diversity statistics under non-random mating (Aim 1), I am interested in studying the joint effects of population structure, mating system, and linked selection on patterns of diversity.

Teaching Statement

Throughout my educational career I have been taught and mentored by people who encouraged me to think critically and be actively engaged in my own learning. Their enthusiastic guidance helped to shape my research trajectory and approach to tackling new questions, and it is the approach I try to bring to guiding others through their own classroom and research training.

I taught undergraduate math classes as instructor-of-record at the during my Ph.D. at University of Arizona. I found that a mixture of traditional lecturing followed by group and individual activity in the classroom provided both exposure to relevant material and structured guidance for developing problem solving skills. At the undergraduate level, this encouraged attendance and engagement, and it helped to develop the skills needed to tackle problems independently outside the classroom. This structure also gave me the opportunity to assess students in real time to identify those that needed extra attention, through direct invitation to office hours or free tutoring resources offered by the department.

At the graduate level, my teaching approach would be quite similar, but with the recognition that the material I would be teaching often has direct application to their own research. Quantitative reasoning and statistical methods are central to much of biology and genomics, and my teaching would include hands-on computational modules. I see integrating mathematical and probabilistic thinking in the biology curriculum as a critical need, and my background makes me well suited to teach courses that combine genetics, evolution, and quantitative reasoning. Much of the foundational theory in genetics and evolutionary biology is used in everyday research, from developing experimental design, to assessing quality of data, bioinformatics, and statistical analysis. Mastering the practical applicability of concepts in population and quantitative genetics not only helps students gain a fuller understanding of the underlying theory, but also prepares them with a set of tools that they can bring to their independent research.

I would be interested in teaching graduate courses in population genetics, evolutionary biology, and mathematical and statistical methods in biology. In each case, learning basic programming and computational skills relevant to the coursework would be a central focus. I would also be interested in leading special topics reading courses in evolutionary biology, the modern synthesis, or population genetics theory. At the undergraduate level, my diverse background makes me well suited to teach a range of classes, including introductory genetics, statistics for biology majors, and more advanced statistical and population genetics courses.

My research experiences as a graduate student and a post-doc have been productive and enjoyable largely due to the quality of guidance and thoughtfulness from my mentors. I came into graduate school with very little background in

biology or genetics, so I relied on my graduate advisor and committee to help me propose a research program that would play to my strengths and develop my biological and evolutionary thinking. I have now had the opportunity as a post-doc to mentor a number of undergraduate students, and I try to suggest and refine research projects that reflect their strengths, interests, and enthusiasm. This is the same approach to mentoring I would bring to my own lab. I would encourage a collaborative role with post-docs, in which they develop their own independent research program and have the opportunity to take leadership roles in mentoring graduate and undergraduate students. For graduate and undergraduate students in my lab, I would take a more active role in proposing projects that suit their strengths and interests.

Diversity Statement

Faculty members have many roles within the university, as mentors and teachers, researchers, and members of the broader scientific community. Each of these roles carries unique responsibilities, which in turn provide opportunities to affirm our commitment to inclusivity in learning and research endeavors. I believe that actively encouraging inclusive learning and research environments not only provides space for people from historically marginalized groups to thrive, but it also results in more principled and equitable scientific production. Below I outline how the consideration of diversity and inclusion will influence my approaches to teaching, mentoring, and research.

Teaching

As instructor for undergraduate math classes at a large public university, my students came from a wide range of backgrounds and with a diverse set of needs to ensure their success. My primary goal was to ensure that course material and instructional resources were accessible to every student. I took seriously my responsibility to accommodate disabilities to ensure equal access and experience in my classes. I also gave regular check-in quizzes throughout the term, asking students what could be changed (or continued) to facilitate their success in the class. Students' thoughtful responses were helpful in guiding how I structured lectures and classroom activities, and it gave me a chance to follow up directly and privately with students who expressed concerns if needed. Through this and direct observation of students' work during in-class group or individual activities, I could also identify students who were struggling with the material and encourage them to seek out office hours and free tutoring services provided by the math department.

It is also our responsibility as instructors to provide a respectful and welcoming environment in the classroom and office hours. At the most basic level, this includes reiterating university policy on harassment both in the syllabus and at the beginning of the semester, and providing information about available resources and how to report harassment. I tend to stress these points, because it's important to me that my students know that I take these matters very seriously. In the same "welcome quiz" I also provided space for the student to inform me of their preferred name and pronouns, especially if they differ from the official class roster.

Mentoring

In the same way, I will maintain an inclusive environment and anti-harassment policy within my own lab and through my broader department mentoring responsibilities. Students and postdocs enter the research community from a diverse set of backgrounds, and the difficulties they face and hurdles they have had to clear are not always apparent to faculty. I recognize the importance of carefully listening to and accommodating what they need to succeed as a developing researchers.

As a graduate student and postdoc, I have been in lab groups with differing levels of diversity. One issue that has stood out is the difficulty that some mentors have had in recruiting students and postdocs from marginalized or minority groups when their current lab does not reflect that diversity. For example, a lab group composed entirely of men is not a welcoming environment for women who are considering joining that lab, a situation I have seen play out too often. Diversity within my research group and equity in the hiring and recruiting process require active steps, and I plan to make use of training and workshops offered by the university toward this goal.

Research

Much of my research focuses on human population genetics and evolutionary history. This field continues to focus much of our resources on a small portion of the world's population (those of European ancestries) and misuse data or ignore directives from marginalized communities. I recently attended the "Closing the Genomics Research Gap" workshop to discuss inequalities in genomics research and what to do about them, and I plan to continue attending similar workshops and conferences on this topic. Additionally, it is important to communicate results back to studied communities in a clear and timely manner and to the public in unambiguous and conscientious language. With growing public interest in human population genetics, our engagement with the public is more important than ever to prevent the misinterpretation, unintentional or otherwise, of human genetics research.