I am a computational biologist seeking to understand disease by fitting biologically realistic, evolution-informed models to biobank-scale genotype and phenotype data using machine learning. My research group will combine ideas from statistical and population genetics to answer questions in disease biology. The standard statistical genetics approach conditions on the observed genotypes and probes the relationship between variants and disease. This approach does not care why a particular pattern of genetic variation is observed, only whether genetic variants correlate with disease. Population genetics, however, seeks to understand patterns of genetic variation – exactly the information ignored in statistical genetics. Yet, we *should* care about patterns of genetic variation if we want to fully understand disease. For example, natural selection shapes genetic variation by removing disease-causing variants, encoding information about disease in patterns of genetic variation.

The main obstacle to combining statistical and population genetics is the intractability of models combining genotypes, phenotypes, and natural selection. Existing work uses simplified models that can be understood by pen-and-paper analysis at the cost of being biologically unrealistic. While more complex models remain analytically out of reach, recent advances in population genetics enable the simulation of biologically realistic models that include genotypes, phenotypes, natural selection, and complex demographic forces.

I have shown in ongoing work that using supervised machine learning on data simulated from these intractable models is statistically sound. Intuitively, this allows us to perform valid statistical inference by "black boxing" the mathematical analysis that would traditionally need to be done using pen and paper. This differs from more familiar uses of machine learning in medicine, where supervised learning is performed on labeled real data. In that setting, models learn a difficult to interpret "black box" mapping from inputs to labels, but that mapping from inputs to labels is exactly the disease biology we are interested in understanding. In contrast, my approach retains the interpretability of the underlying models and only "black boxes" biologically uninteresting statistical inference.

The time is ripe to combine these advances in population genetics and machine learning with biobank-scale datasets to deepen our understanding of human disease. **My research group will build interpretable but biologically realistic models of the genetic and environmental causes of disease and fit them to massive human genetics datasets by using cutting-edge population genetic simulation software and developing purpose-built machine learning tools.** My expertise in statistical genetics, machine learning, and population genetics and my ability to make connections across these fields make me uniquely positioned to capitalize on these advances.

Previous and Current Work

Population Genetics

In one line of work, I have used patterns of genetic variation to learn about human history. In one project, I used ancient DNA data collected in collaboration with a large team of geneticists and anthropologists to shed light on the peopling of the Americas [1]. In work that is currently in revision at *Cell* (co-first author), I analyzed a large dataset of individuals from across Africa to better understand the demographic events that have shaped present-day genetic diversity across the continent.

In another line of work, I showed that demographic events can cause standard methods to infer dramatic changes in the rate of recombination along the genome even when no such changes exist [2]. Such effects have confounded previous analyses of recombination rate variation in humans, so I developed a method to properly account for these differences in demography, implemented it in a widely-used software package, `pyrho`, and applied this method to a diverse set of human populations [3].

Statistical Genetics

For one of my main postdoc projects, I have built a flexible modeling framework for using GWAS data to simultaneously estimate how individual variants affect disease risk and the overall distribution of those effect sizes [4]. Most existing methods either infer the effects of each variant or infer the distribution of effect sizes, but not both simultaneously. Yet, each is informative of the other indicating that both can be improved by simultaneous inference. I have applied my method to many disease-relevant traits to construct highly accurate polygenic scores that are significantly more predictive of trait values than competing methods for some traits.

Machine Learning

For many biological questions of interest there are no suitable off-the-shelf machine learning methods, so I have also developed novel machine learning approaches. In statistical genetics, I found that models frequently possess a special structure that makes them especially amenable to variational inference. I used this structure to derive more computationally efficient variational inference schemes resulting in a broadly applicable approach that I presented as a spotlight presentation at NeurIPS, a top machine learning conference [5]. In population genetics, I developed a deep learning architecture that exploits a symmetry ubiquitous in population genetic datasets [6]. This architecture trains faster and is more accurate than standard deep learning architectures.

In the past few years, population geneticists have begun training machine learning models on simulated data and then applying these models to real data. While this approach is intuitive, its statistical interpretation in unclear. In ongoing work, I have given this simulation-based approach to population genetic inference a rigorous statistical interpretation. This statistical interpretation has suggested a number of ways that current approaches can be improved, which I am now implementing.

Future Research Program

My research group will focus on understanding disease by using tools from machine learning to analyze biologically realistic models from population genetics with biobank-scale data. In my previous work, I incorporated more biological complexity into models from population and statistical genetics using analytical techniques like deriving bespoke variational inference schemes. Yet, to tackle the kinds of complex, biologically realistic models that we now have the data to fit, we need a fundamentally different approach.

My research group will answer questions about the genetics of human diseases by using the advances in population genetic simulation software coupled with simulation-based statistical inference using machine learning. Below I list some examples of problems in the genetics of human health that have not been solved by traditional statistical genetics approaches and how my proposed research framework provides natural, novel avenues to approach these problems. Each problem represents an example of the types of projects that I envision pursuing.

Interpretation of rare variants and disease gene prioritization

Rare genetic variants can cause severe disorders and are difficult to analyze using traditional statistical genetics approaches. By definition, few individuals have any particular rare variant, so standard association tests are underpowered to estimate the effect of individual variants. My research group will construct biologically realistic models where genes have different contributions to fitness, and different types of variants (e.g., coding variants in particular protein contexts like solvent-exposed α-helices) have different effects on genes. Since natural selection shapes the variation at these loci, we can use patterns of genetic variation across the genome to learn to what extent different types of variants affect genes. Meanwhile, we can use local patterns of genetic variation within a gene to estimate how much that gene contributes to fitness. While such models are utterly intractable analytically, it is possible to simulate from them to build machine learning models which we can then apply to real data. The resulting estimates of gene-specific effects on fitness will help prioritize which genes are under strong selection and hence likely to play a role in disease, and the estimates of which types of variants have strong effects on genes will help prioritize rare variants for clinical followup.

Leveraging diverse global datasets to improve prediction of disease and discovery of disease-associated variants

The existing datasets used by statistical genetics are not representative of global genetic diversity – they are primarily composed of individuals of European ancestries. These datasets are used because they contain rich phenotyping data in addition to genotypes. Meanwhile several datasets have been collected in the context of human population genetics with the explicit intention of better capturing global genetic diversity. While these datasets do not contain phenotypes, the genotypes alone contain rich information about the effects of natural selection across the genome, which are informative about which loci contribute to disease. Can we leverage this wealth of global genetic data to improve prediction of disease or even improve the discovery of disease-associated variants?

My research group will use simulations to investigate the extent to which diverse genotype-only datasets are informative for traits that are under selection. We will then develop analytical and simulation-based machine learning approaches to combine this information with standard genotype+phenotype biobank data to improve disease prediction (e.g., building better polygenic scores) and increase power for discovering disease associations.

Modes of selection acting on disease traits

Why do common diseases exist? One standard population genetic explanation is mutation-selection balance – disease is universally deleterious to fitness, but mutations arise often enough to persist in the population. In contrast, a number of diseases, such as schizophrenia, appear to be under stabilizing selection perhaps due to pleiotropic effects. That is, these diseases may exist because they share a genetic basis with a beneficial trait. A third possibility is that selective pressures have changed as our environments have changed. Even the time since the agricultural revolution, about 12 thousand years, is almost negligible on an evolutionary time-scale, and some diseases may arise due to the mismatch between present-day environments and the environments in which genetic variation was shaped. While these explanations have all been debated for various diseases, evidence in favor of one explanation or another for any particular disease is generally based on simplified models or non-quantitative evolutionary explanations like the “thrifty gene hypothesis”.

My research group will use simulations to explore how different modes of selection affect patterns of genetic variation. We will also train machine learning models to solve the inverse problem of learning the past and present modes of selection on disease-relevant traits using biobank-scale data.

Summary

My research program aims to understand disease using machine learning to analyze biologically realistic models that incorporate ideas from population and statistical genetics. This research program presents a novel approach to understanding disease, and the above examples show how this framework opens up new avenues to approach problems in statistical genetics. I plan to collaborate with applications-focused research groups to use these approaches on a broad range of problems across biology.

Given the potential impact on our understanding of human disease, I expect my research program will be able to be funded through NHGRI and NIGMS. At the same time, by developing novel machine learning and statistical approaches, I also expect to apply for funding from the NSF as well as smaller grants from the private sector, such as the Amazon Research Award program.

My past and present work combines ideas from across statistical genetics, population genetics, and machine learning. My expertise in these fields makes me well-suited to improve our understanding of human health by developing powerful machine learning methods to harness increasingly flexible population genetic simulation software and capitalize on emerging massive datasets.

[1] J. Víctor Moreno-Mayar*, Lasse Vinner*, Peter de Barros Damgaard*, Constanza de la Fuente*, Jeffrey Chan*, **Jeffrey P. Spence***, et al. Early human dispersals within the Americas. *Science*, 2018.

[2] John A Kamm*, **Jeffrey P. Spence***, Jeffrey Chan, and Yun S Song. Two-locus likelihoods under variable population size and fine-scale recombination rate estimation. *Genetics*, 2016.

[3] **Jeffrey P. Spence** and Yun S. Song. Inference and analysis of population-specific fine-scale recombination maps across 26 diverse human populations. *Science Advances*, 2019.

[4] **Jeffrey P. Spence**, N. Sinnott-Armstrong, T. Assimes, and Pritchard J. K. A flexible modeling and inference framework for estimating variant effect sizes from GWAS summary statistics. *bioRxiv preprint (in revision at eLife)*, 2022.

[5] **Jeffrey P. Spence**. Flexible mean field variational inference using mixtures of non-overlapping exponential families. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[6] Jeffrey Chan, Valerio Perrone, **Jeffrey P. Spence**, Paul Jenkins, Sara Mathieson, and Yun S. Song. A likelihood-free inference framework for population genetic data using exchangeable neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Working in science has been one of the great joys of my life, and any person should be able to pursue a career in the sciences without facing barriers because of their race, gender, sexual orientation, the abledness of their body, or any other aspect of their identity. Unfortunately, at every level of academia structural barriers as well as biases – ranging from the implicit to the overt – act to push individuals from underrepresented groups out. To counteract this and ensure a vibrant, diverse community, we must work to acknowledge and dismantle these barriers so that we can foster a greater sense of inclusion and strive for true equity. Throughout my academic career I have worked to understand these systemic challenges and I have worked with organizations that specifically aim to dismantle these barriers. At a smaller – but no less important – scale, I have mentored several students and postdocs including women and non-binary people. I have helped fellow postdocs with job application materials, I have taught Ph.D. students background material relevant to their research, and I have helped students navigate the emotional challenges of grad school. Going forward, I will continue to learn about these barriers and work to dismantle them, and I will actively work to recruit and mentor students from groups that higher-education has historically excluded.

Systemic barriers act to exclude different groups of people from academia. These barriers range from grad school application fees to the burden of childcare disproportionately affecting women (especially during the pandemic). Unless we face these barriers ourselves it can be all too easy to pretend they do not exist. I have worked to understand the barriers facing individuals from underrepresented groups. During grad school, I attended seminars through an organization, Bias Busters, which seeks to make Berkeley a more inclusive community by identifying and addressing implicit biases. One particularly memorable seminar highlighted that campus climate surveys have only recently begun even asking relevant questions about LGBTQ+ identities, such as allowing respondents to choose to identify as trans or non-binary. Without allowing students to express their gender identities on campus climate surveys, we would never have known that gender-nonconforming students are disproportionately affected by mental health issues ranging from depression to eating disorders. During my postdoc I attended a number of "Town Hall" events related to social justice and the relationship between campus police and students of color. By listening to students share their stories, I was able to better understand the extent of the barriers acting to exclude certain groups from feeling safe on campus.

I have also worked to broaden the diversity and improve the inclusiveness of academia. During grad school, I was a member of Bay Area Scientists Inspiring Students (formerly known as Bay Areas Scientists In Schools). We went to elementary schools in the Oakland metro area, interacted with students, and performed science experiments. The goal was to show students from diverse backgrounds what being a scientist is really like, and that scientists are real people so that the students would be inspired to pursue STEM careers. I served twice as a science fair judge for the Oakland Unified School District, where I was able to interact with a diverse group of students, ask about their projects, and answer their questions about what it is like to be a scientist. During my postdoc, I have worked with the Society for Advancement of Chicanos/Hispanics and Native Americans in Science (SACNAS), serving as a reviewer for the society's annual conference. The SACNAS conference is a large multicultural STEM diversity event, designed to provide students from underrepresented backgrounds an inclusive and welcoming space to present and be immersed in scientific research. I reviewed travel grant applications so that students could attend the conference without worrying about incurring financial burdens. During my postdoc, I also served as a judge for talks through the Stanford Summer Research Program, which seeks to give undergraduate research experience to individuals who would contribute to diversity in biology.

Going forward, I will continue to work to understand and counteract the barriers that individuals from underrepresented groups face in academia. I will deepen my involvement with SACNAS and work with the local chapter at UCSF. Throughout my career I have acted as a mentor to graduate students and postdocs, including women, people of color, and non-binary people. As I transition to a formal advising role, I will work to recruit students and postdocs from diverse backgrounds by giving talks at and collaborating with faculty at historically Black colleges and universities and primarily undergraduate institutions. To ensure that my research group is equitable and inclusive I will institute a lab code of conduct, and I will attend seminars and participate in activities about bias and encourage members of my group to attend with me.

I love being a scientist, and everyone should have the opportunities that I have had. While we unfortunately live in a world where systemic barriers and biases push people out of our field, I will work to counteract these forces to make our field more diverse, equitable, and inclusive.

Teaching is a fundamental part of the academic endeavor. In addition to advancing knowledge through research, we should also prepare the next generation of doctors, researchers, and scientists. I find teaching extremely rewarding, especially seeing students have a "eureka moment" as a difficult concept suddenly becomes clear. In order to achieve this, students need to be active participants in their learning. My teaching philosophy prioritizes engagement and creating an environment where students are comfortable and feel included so that they can focus on actively learning.

During my Ph.D. I was a graduate student instructor for an upper division introduction to mathematical statistics. The professor for the course was my Ph.D. advisor, Professor Yun Song, and it was his first time teaching this course. As a result, the other graduate student instructor and I were deeply involved in planning the course. We developed all of the assignments, quizzes, exams, and discussion section lesson plans with Professor Song. I used this opportunity to hone and implement my teaching philosophy of making students feel comfortable and actively engaged in their learning. For many students, this was the first course in which they encountered proofs, and many students were intimidated by the idea of writing formal arguments. During discussion sections, I made students comfortable with the material by first emphasizing intuition. I engaged them by having them describe to each other the meaning of theorems in plain , non-mathematical language (think-pair-share) or by having them draw pictures to visualize geometric aspects of the material. Then, I would have students discuss with each other why a result should be true at an intuitive level before turning to formalizing that intuition in a rigorous proof. The focus of the course was on proofs and derivations but we included two R programming exercises to engage students by applying theory to real-life datasets. **I received an Outstanding Graduate Student Instructor award for my work in this class.**

The ongoing COVID-19 pandemic has highlighted both the challenges and opportunities of online and hybrid learning. Online learning is not going anywhere, and many classes will have at least some online component going forward. Online learning is challenging and one of the primary challenges is engagement. One powerful strategy ideal for online courses is the flipped classroom: provide students with a pre-recorded lecture or pre-assigned reading and then use lecture time to discuss the material. Then, students must be actively engaged during lecture time to discuss the material instead of just passively listening. Furthermore, having access to course materials online allows students with different learning styles to interact with the material in ways that are most comfortable to them. Students can re-watch recorded lectures, ask and answer questions on discussion platforms like Piazza, or read lecture notes. When I was a graduate student instructor, I used Piazza to make all of my discussion section notes available, and I encouraged students to learn by answering each other's questions. Students really engaged with Piazza because it is a low pressure, comfortable environment where they can ask questions – even anonymously – to other students. Used correctly, online learning can be deeply useful for teaching, especially for students with learning styles that are not well-suited to traditional lecture-based courses.

Going forward, I would be thrilled to teach and develop courses across biology, computing, and statistics. If I were to design a course, it would be "Probabilistic Modeling in Biology" targeted toward not necessarily quantitative graduate-level students. It would be a project-based course where students would create a statistical model for publicly available data or data that they collected in the course of their dissertation research. Students would frame their research questions as inference problems in their statistical models. They would then implement and fit their models using a probabilistic programming language. Finally, students would interpret their results in terms of the original biology. The course material would mirror the steps of this project. The first part would cover fundamentals of probabilistic modeling. Next would be a practical introduction to probabilistic programming languages and inference algorithms. The course would end with model interpretation. This course would engage students to learn new techniques from statistics and machine learning because they would be solving problems arising in their own research. By working with their own data, on their own projects, students would already be experts on some aspects the project, ensuring that they have a comfortable knowledge set from which to approach the unfamiliar statistical material. This course would also require students to solve the real-world open-ended challenges presented to practicing computational biologists.

My teaching philosophy is to actively engage students and create a comfortable classroom environment where they can fully participate in their learning. By making students active participants in their learning, we become partners in the goal of developing the skills and knowledge that they need to succeed.