Background And Goals for Fellowship Training

A) Doctoral Dissertation and Research Experience

As an undergraduate I had been exposed largely to molecular, developmental, and cellular approaches to studying biology, and gained some research experience in the lab of David Baum at the University of Wisconsin, where I studyied the evolution and development of petal spot polymorphisms in the California wildflower *Clarkia gracilis*. However, I began to realize that the problems that interested me most were those most ammenable to investigation by most quantitative approaches, with which I had very little experience. this motivated my choice of graduate school.

My dissertation research focused in large part on the development of statistical methods and coalescent theory approaches to understand the population genetic signatures of adaptation. In light of the perceived failure of the "classic selective sweep" program to identify more than a handful of genes clearly involved in adaptation events, the inclination in the field has been to take this as an indication that the majority of adatpation does not proceed through the fixation of *de novo* mutations of large effect, but rather via some other population genetic mode. There are many candidates (as well as the possibility that there have simply been fewer adaptive events in the recent history of humans and other well studied species than had been expected), and my dissertation research focused on developing the statistical tools necessary to identify the signals left behind in population genetic data under two particular scenarios in which standing variation plays a major role in the response to adaptation.

Detecting Polygenic Adaptation with GWAS Data

When directional selection acts on highly polygenic phenotypes, the response is expected to be distributed among the many loci underlying the trait, such that no individual site witnesses large changes in alleles frequency. Genetic drift may even be expected to dominate individual allele frequency trajectories, even when a substantial change in the population mean trait value occurs, provided that the genetic architecture is sufficiently diffuse. A major component of my dissertation work focused on the development of statistical methods to combine population genetic datasets with annotations of trait associated alleles from genome wide association studies (GWAS) in order to detect signals of polygenic adaptation in these sort of scenarios. The method relies on the construction of a population's mean polygenic score for a particular phenotype as the sum of the contributions from many different loci. The method then relies on the multivariate normal approximation of genetic drift across multiple popualtions to build a null model for the distribution of polygenic scores expected under neutrality. This model can be leveraged to construct a statistical test for evidence that recent directional selection has caused divergence among populations for the trait of interest.

The method can be applied to an arbitrary number of populations with an arbitrary relatedness structure, and is therefore well suited to human populations and also applicable to any organism in which is it possible to carry out GWAS. Our first paper on this topic presented an extensive characterization of the method and its performance, as well as an application to six traits for which GWAS data was publicly available at the time. I replicated previously observed signals of selection on height and skin pigmentation,

and also showed how the interaction between natural selection, genetic architecture, and demography have the potential to create misleading signals of polygenic adaptation under a naive interpretation.

In a subsequent project which is currently in prepratation for publication, I revisited the question of polygeneic adpatation in human traits by applying the inference tools developed above to a much broader array of traits (over 40), and have further extended the methods to help identify 1) when and where selection acted on the quantitative trait in question, and 2) to determine whether multiple concurrently observed signals of selection can be disentangled to determine which traits are evolving under direct selection, and which are responding to selection on a correlated trait.

This work has shown that in addition to the north-south gradient in height across Europe, selection has also impacted the distribution of height across much of Eurasia and the Americas, but that each of these additional signals can be explained in terms of a selection event which is also ancestral to modern Europeans. This result demonstrates that natural selection was likely responsible for divergence among ancestral human populations long before the recent divergence among the ancestors of modern Europeans which has been directly observed in ancient DNA.

Modeling the Hitchhiking Signature of a Sweep from Standing Variation

One possible explanation for the failure of the "classic hard sweep" paradigm is that while sweep-like behavior may still be common, many sweeps may involve mutations which are already present as standing variants at the time of an environmental change, rather than *de novo* mutations. Previous simulation based work had shown that for many of the summary statistics commonly used to identify sweep candidate regions, selective sweeps of this kind resulted in signatures which were weaker and more variable than those of classic hard sweeps, but little had been done in the way of theoretical characterization.

We recognized an analogy between the infinite sites models and the pattern of early recombination events in the linked region surrounding the beneficial mutation. This allowed me to apply the Ewen's Sampling Formula to describe the hitchhiking signature left behind by any event of this kind. My results showed that sweeps from standing variation are characterized by a more rapid breakdown of haplotype structure as one moves away from the selected site (yielding the weaker signal), and that this breakdown generally proceeds by the creation of multiple haplotypes of intermediate frequency, rather than the single core haplotype accompanied by multiple low frequency haplotypes as expected following a classic hard sweep.

The primary outcomes of this work were 1) an extensive theoretical characterization of the hitchhiking pattern observed following a sweep from standing variation, and 2) a set of guidelines for methods developers regarding the features of DNA sequence data which are likely to be most informative for indentifying sweeps from standing variation, and distinguishing them from other classes of sweep-like phenomena. The aspect of this work which will be most applicable to the presently proposed work is that it provided me with extensive experience modeling and simulating the frequency trajectories of alleles under selection.

Relationship of previous experience to proposed fellowship My dissertation research experience relates to the proposed fellowship in a couple of different ways. The first is that while I did not begin graduate school with this in mind, but experience working on polygenic adaptation captured my interest in the evolution of complex and quantitative traits. Moving forward, I hope to build this topic into a major pillar of my research program, and my choice to study disease phenotypes (rather than adaptive ones), and to do so via a distinct mathematical approach, represents a broadening of my efforts on this topic.

Second, I went to graduate school with the explicit goal of complementing my undergraduate background in biology with a set of statistical, computational and mathematical skills, all three topics in which

I had very little prior experience. During the course of my PhD, I found that I most enjoyed explicit mathematical modeling projects which allowed me to distill some biological process down to its essential parts and then analyze the resulting models to learn something new about the biology. This interest played a large role in motivating my choice of post-doc lab and project.

B) Training Goals and Objectives

I have been a post-doc in the Sella lab since October 2016, and hope to conitinue in this position until approximately Fall of 2019. I aim to begin applying for positions as a principle investigator (PI) at a major research institution (e.g. R01 university) within the next year. My career goals are driven by my interest in applying the tools of mathematical population genetics and evolutionary biology in general in order to understand the biology of human phenotypes and diseases.

The evolution of mutations that affect fitness: My research has and will continued to focus on understanding the evolution of alleles that influence fitness. In my PhD, I pursued this research interest primarily by working with models of allele frequency evolution under neutrality, or in cases of strong selection. Such models are generally fairly easy to analyze and work with, which makes them attractive because they simplify data analysis. It also made sense to focus on these easier to analyze models because of my limited mathematical background prior to beginning graduate school, as a way to ease into the field. However, a comprehensive understanding of the evolution of human traits and diseases will require substantial progress in the analysis of models which have traditionally been challenging for population geneticists, particularly those which include alleles under weak selection, or in which the evolution of many independent alleles are coupled together through their influence on a phenotype.

Progress analyzing and understanding models of this type will require the application of new and creative mathematical approaches, and I will require training in these approaches in order to be successful. In particular, the toolkit of statistical physics, which has been introduced into population genetics over the course of the last 10 years, will be especially valuable. Statistical physics is generally concerned with the study of situations in which a very large number of particles contribute to the dynamics of some larger physical system, and recent work has identified a series of analogies between the problems to which these approaches are typically applied in physics, to the case where a very large number of distinct genetic loci contribute to variation in a given phenotype. Dr Sella was a pioneer in the early application of these approaches in population genetics, and an education in the application of these (and other) approaches to analyze difficult models in population genetics will constitute a major component of my training.

Population genetics and human biology: One of the major goals of my research program as a post-doc and (eventually) as a PI will be to use the tools of population genetics in order to gain fundamental insights into human biology. Given the enormous amount of genetic and phenotypic data that will be generated over the coming decades, we will have unprecedented opportunity to understand how mutation, selection, and other population genetic forces shape human biology (e.g. disease, life history, adaptation, etc), and in turn how our biology gives rise to these population genetic forces. To be effective in this goal, I will need a much deeper understanding of the major questions being addressed in human genetics, and the data being generated to answer them. The environment of the Sella-Przeworski-Pickrell group is uniquely suited to provide this training, which I will receive through regular interaction and meetings with all three PIs, as well as through journal clubs and lab meetings, which are heavily focused on questions at the intersection of human and population genetics.

Collaboration and Mentoring: For my PhD dissertation, I worked mainly on independent projects without collaborators. This was valuable in that I learned how to design and carry out a project from start to finish. However, in order to be a successful PI, I will need to improve my ability to work as an effective

collaborator and as a mentor for future students and post-docs of my own. My postdoctoral training experience in Dr Sella's lab will provide training in both of these skills, as there are multiple lab members working on projects related to some aspects of my proposed work, which will provide opportunities to work in a collaborative or mentoring role where beneficial to do so. I am also mentoring an undergraduate student in the lab, which is a first step toward preparing to mentor my own (under)graduate students and post-docs as a PI.

C) Activities Planned Under This Award

Responsible conduct of research activities (including formal instruction and informal discussion with my mentors) will account for approximately 1% of my total time each year. I will attend an average of 2 scientific conferences per year, which will account for approximately 4% of my time. I will spend another 10% of my time mentoring undergraduate students who will work with me either on some aspect of my proposed research, or potentially on an independent research project of their own design, and an addtional 10% of my time on journal clubs, lab meetings, seminars, and other formally organized training sessions with colleagues. The remaining 75% of my time will be spent directly on research related activities.

My proposed research focuses on two aims: developing pop gen models for the genetic architecture of complex disease, and infering evolutionary parameters across many complex disease datasets from GWAS data. Below is a timeline for these activities over the course of the three years of the training period. Each year totals to 75% in the timeline below, such that when added to the above activities the total for each year is 100%:

Timeline:

1. First Year
    - Expand current model to incorporate pleiotropy/demography/environemtnal change: 45%
    - Begin developing inference machinery for basic single trait no pleiotropy case: 25%
    - Begin compiling and become familiar with GWAS datasets that will be used for inference: 5%
2. Second Year
    - Complete work expanding basic model: 25%
    - Develop inference machinery to include features of expanded model: 30%
    - Apply inference machinery to GWAS datasets: 20%