Project Summary/Abstract

The growing abundance of population genomic data creates a *critical need* for inference approaches that can reveal evolutionary history. The PI's *long-term goal* is to understand how natural selection shapes the evolution and function of the molecular networks that comprise life. Toward that goal, the PI's group develops and applies methods for inferring the evolutionary past from population genomic data. The *objectives* of this application are to understand how context affects mutation fitness effects, to develop improved inference methods, and to support the population genomics research community. The *rationale* is that this research program will both reveal new insights into evolution and enhance the ability of colleagues to reveal complementary insights.

The PI's research group has expanded the concept of a distribution of fitness effects to multiple dimensions, focusing on differences in mutation fitness effects among populations. The PI proposes to apply this approach to numerous systems, to elucidate the relative roles of genetic and environmental context in creating differences in fitness effects. The group will also extend this approach to consider differences in fitness effects over time.

The PI developed and maintains the software `dadi`, among the most popular approaches for fitting population genomic models to data. The PI will continue to support and enhance `dadi`, while developing complementary inference approaches. These will include new diffusion methods based on pairs of loci and the linkage among them and a novel deep learning approach for inferring the distribution of fitness effects.

The PI helped found the PopSim consortium, which aims to expand the rigor and transparency of population genomic models for the scientific community. The PI's group will continue to be active in the consortium, particularly leading a new initiative to facilitate rigorous testing of population genomic methods via open competition.

The proposed research program is *innovative* both conceptually and methodologically. The novel concept of a multidimensional distribution of fitness effects has many applications, and the group will develop novel methodology for several population genomics inferences. The *expected outcomes* of the proposed research are new insights into the ecology and biology of mutation fitness effects, new population genomic inference tools, and a framework for blinded evaluation of such tools. These outcomes are expected to have important *positive impact* on the filed of population genomics. The methods will be widely applicable and well-supported, and the inferences will feed into approaches for inferring the evolutionary past and predicting the evolutionary future.

Project Narrative

The proposed research is relevant to public health because understanding natural selection plays a key role in combating genetic and infectious diseases. The proposed research will generate new methods for identifying both negative and positive natural selection in natural populations, including humans. These methods will form a foundation for predicting, for example, rates of adaptation of pathogens and the architecture of human genetic disease.

Facilities and Other Resources

Environment-Contribution to Success

The research will take place in the Department of Molecular & Cellular Biology at the University of Arizona. I also have affiliations with the Department of Ecology & Evolutionary Biology, the Department of Epidemiology and Biostatistics, and the Graduate Interdisciplinary Programs in Applied Biosciences, Applied Mathematics, Cancer Biology, Genetics, and Statistics & Data Science.

The University of Arizona possesses great expertise in areas of relevance to the Gutenkunst group and its research. In particular, the Applied Mathematics program has an NIH T32 Training Grant in Computational and Mathematical Modeling of Biomedical Systems, and I am on the Steering Committee for that grant. In addition, the Gutenkunst group has a strong connection to the group of Prof. Joanna Masel, an expert in theoretical molecular evolution. The Gutenkunst and Masel groups share space and hold a joint weekly lab meeting, providing a stimulating intellectual environment. We are often joined by Dr. David Enard, an expert in applied population genomics, who has a particular interest in adaptation to viral pathogens.

Facilities

Computer

My group's computing facilities consist of Macintosh desktops and laptops with at least 2.66 GHz quad-core processors and 16 GB of memory. All computers have access to 1 Gb switched wired Internet connections and the University wireless network. All Department computing facilities are maintained by full-time staff. To facilitate long distance collaborations, my computer is equipped with modern teleconferencing equipment.

Office

My office is roughly 10 ft x 11 ft and located on the same floor as my group's research space. My group shares research space with the group of Prof. Joanna Masel. This space consists of roughly 600 sq ft of office space that was renovated in summer 2014, including an interaction area with large whiteboards and couches. Also included is another roughly 200 sq ft of office space.

Equipment

Computer

My group has access to the University of Arizona's excellent central high-performance research computing systems.

The Ocelote cluster is a 2.3 GHz Intel Xeon Haswell Dual 14-core cluster with 400 nodes (11,528 total cores) with 192 GB of memory per node. Among those nodes are 46 NVIDIA P100 GPUs. The nodes are connected by FDR Infiniband. On this cluster, my group has 100,000 high-priority and 65,000 standard priority CPU-hours each month and the ability to claim 1,162 cores at a time. The Puma cluster is a 2.3 GHz AMD EPYC Rome Dual 48-core cluster with 244 nodes (23,616 total cores) with 512 GB or memory per node. Among those nodes are 29 NVIDIA V100S GPUs. The nodes are connected by 25Gb/s Ethernet. On this cluster, my group has 280,320 high-priority and 100,000 standard priority CPU-hours each month and the ability to claim up to 3,674 cores at a time. In addition to our half-million allocated monthly CPU hours, we may submit jobs to the windfall queue that will run on any available cores.

Connected to the HPC is a high-speed flash-based Qumulo storage array with 2.29 PB of raw capacity. On this array, each user has 50 GB of space, and my group has a shared 20 TB allocation. My group also has 72 TB of network-attached RAID storage for storing backups and data archives.

OMB No. 0925-0001 and 0925-0002 (Rev. 10/2021 Approved Through 09/30/2024)

BIOGRAPHICAL SKETCH

Provide the following information for the Senior/key personnel and other significant contributors.
Follow this format for each person. **DO NOT EXCEED FIVE PAGES.**

NAME: Ryan N. Gutenkunst

eRA COMMONS USER NAME: RGutenkunst

POSITION TITLE: Associate Professor

EDUCATION/TRAINING *(Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable. Add/delete rows as necessary.)*

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	Completion Date MM/YYYY	FIELD OF STUDY
California Institute of Technology, Pasadena, CA	B.S.	06/2002	Physics
Cornell University, Ithaca, NY	Ph.D.	01/2008	Physics
Cornell University, Ithaca, NY	Postdoctoral	12/2008	Population Genetics
Los Alamos National Lab, Los Alamos, NM	Postdoctoral	08/2010	Computational Immunology

A. Personal Statement

My long-term goal is to understand how natural selection shapes the evolution and function of the molecular networks that comprise life. Toward that goal, my research program currently focuses on population genomics, with an emphasis on developing and applying computational methods for inferring models of population history and natural selection from data. My PhD training was focused on computational modeling in systems biology, during which I discovered a universal pattern of parameter sensitivities in molecular network models (Gutenkunst 2007). Intrigued by the potential evolutionary implications of this pattern, I undertook postdoctoral training in population genetics. I expanded my computational biology training during a second postdoctoral appointment, during which half my time was independent. As principal investigator, my research group has focused on population genomics, while also contributing to evolutionary systems biology (e.g., Mannakee 2016) and cancer genomics (recent highlights below).

I serve the population genomics community through my maintenance of the software dadi (Diffusion Approximations for Demographic Inference; Gutenkunst 2009). I developed dadi as a postdoc to fit models of demographic history and natural selection to population genomic data. Despite the emergence of competing approaches, dadi remains widely used. Since May 2020, dadi has been downloaded over 47,000 times from conda-forge. I believe my active support of the community is key to dadi's longevity. As of April 2022, the dadi-user Google Group has 382 members, and in 2021 I posted 71 messages to the group. All those posts were in response to user questions, most of which were conceptual questions about how to apply dadi to data, rather than technical questions about installation or software errors. As detailed below and in Research Strategy, my group continues to enhance dadi, by integrating new analyses into it and improving its performance and usability.

I also serve the population genomics community through the PopSim consortium, which aims to create standardized and accessible population genomic simulations for the community. In 2018, I co-organized the workshop that kicked off the consortium, and my group and I remain active within it. In 2020, the consortium published its first paper, describing an initial catalog of neutral population genomics simulations for several model organisms (Adrion 2020). Since then, the consortium has worked to expand the catalog to include non-model organisms and to incorporate natural selection into the simulations. The consortium has broad scope in the community. The authors of Adrion (2020) were affiliated with institutions from nine countries, and 44 researchers have made commits to the consortium's primary GitHub repository, as of April 2022. The consortium remains active and will continue to transform population genomics research.

a. **RN Gutenkunst**, JJ Waterfall, FP Casey, KS Brown, CR Myers, JP Sethna (2007) Universally sloppy parameter sensitivities in systems biology. PLoS Computational Biology 3:e189. [I was supported by NIH T32GM08267.]
b. **RN Gutenkunst**, RD Hernandez, SH Williamson, CD Bustamante (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genetics 5:e1000695. [I was supported by startup funds to SH Williamson.]
c. **BK Mannakee**, **RN Gutenkunst** (2016) Selection on network dynamics drives differential rates of protein domain evolution. PLoS Genetics 12:e1006132. [My student was supported by NSF GRF DGE-1143953.]
d. JR Adrion, CB Cole, N Dukler, JG Galloway, AL Gladstein, G Gower, CC Kyriazis, AP Ragsdale, G Tsambos, F Baumdicker, J Carlson, RA Cartwright, A Durvasula, BY Kim, P McKenzie, PW Messer, E Noskova, DO Vecchyo, F Racimo, **TJ Struck**, S Gravel, **RN Gutenkunst**, KE Lohmeuller, PL Ralph, DR Schrider, A Siepel, J Kelleher, AD Kern (2020) A community-maintained standard library of population genetic models. eLife 9:e54967. [My team was supported by NIH R01GM127348.]

Ongoing and recently completed research support that I would like to highlight include:

NIH R01 GM127348	Gutenkunst (PI)	02/12/19-01/31/2024

"Joint inferences of natural selection between sites and populations"
I am sole PI on this grant, the goals of which are to develop and apply novel computational methods for inferring quantitative models of natural selection from population genomic data. I was also awarded an administrative supplement to add distributed cloud computing capability to dadi.

B. Positions, Scientific Appointments, and Honors

Positions and Scientific Appointments

2021	Panelist, National Science Foundation, Understanding the Rules of Life: Emergent Networks
2021,2020,2017, 2013,2010	Panelist, NSF Graduate Research Fellowship Program
2018	Co-organizer, workshop on Population Genomics Simulation at Cold Spring Harbor Laboratory
2017-present	Associate Professor, Department of Molecular & Cellular Biology (MCB), University of Arizona, Tucson, AZ Secondary appointments with the Department of Ecology & Evolutionary Biology, the Department of Epidemiology & Biostatistics, the BIO5 Institute, and the Graduate Interdisciplinary Programs in Applied Biosciences, Applied Mathematics, Cancer Biology, Genetics, and Statistics & Data Science
2017-present	Associate Department Head, Department of MCB, U of Arizona
2017-present	Director of Graduate Studies, MCB Accelerated Master's Degree program and MCB track of Applied Biosciences Professional Science Master's Degree program
2015-present	Associate Editor, BMC Evolutionary Biology
2010-2017	Assistant Professor, Department of MCB, U of Arizona, Tucson, AZ
2009-2010	Postdoctoral Scholar with Byron Goldstein in the Center for Nonlinear Studies at Los Alamos National Laboratory
2007-2008	Postdoctoral Scholar with Scott Williamson and Carlos Bustamante at Cornell University

Honors

2014	Kavli Fellow, US National Academy of Sciences and The Kavli Foundation
2013	Distinguished Early-Career Teaching Award, University of Arizona, College of Science
2004-2006	NIH Molecular Biophysics Training Grant, Cornell University
2002-2004	NSF Integrative Graduate Education and Research Traineeship in Nonlinear Systems, Cornell University

C. Contributions to Science

Members of my group are bolded in author lists.

1. **Multidimensional Distributions of Fitness Effects**
The distribution of mutation fitness effects (DFE) quantifies how likely a mutation is to have a given effect on fitness and is a key concept in evolutionary biology. The properties of this distribution affect many aspects of organismal biology and evolution, such as the genetic architecture of complex traits and the speed of adaptation. My group has recently extended the concept of a DFE beyond a single dimension, to consider the relationship between two or more selection coefficients. We first did this in the context of triallelic sites, where not one but two mutations arise at the same genomic location (Ragsdale 2016). We developed a novel triallelic diffusion equation and inference approach, which we applied to *Drosophila melanogaster* data to quantify the relationship between same-site nonsynonymous mutations. To validate our inferences, we compared with biochemical data. Remarkably, we found quantitative agreement between our population genomic inferences and deep mutational scanning experiments. This highlights the utility and accuracy of our population genomic inferences.
More recently, we quantified the relationship between mutation fitness effects in pairs of populations of humans, *D. melanogaster*, and wild tomatoes (Huang 2021). In simulations, we found the correlation of the DFE could be inferred from allele frequency spectra precisely, even with little genomic data. When applied to real data, we found interesting patterns in which Gene Ontology terms corresponded to genes with high or low correlation in mutation fitness effects. We also found that among these three species, the correlation in selection coefficients decreased with population differentiation, suggesting a strong effect of genetic context on mutation fitness. A key goal of my future research is to quantify this correlation in multiple scenarios, to establish the relative roles of environmental and genomic context in determining mutation fitness effects.
Selection coefficients may change not just between populations but also over time. In population genomics, this has been most well studied in the case of soft selective sweeps, in which mutations become beneficial. But other changes in selection coefficients also likely occur. We have extended our framework to include temporal changes. Using ancient DNA data, we applied this approach to consider the relationship between selection before and after the advent of agriculture in Europe (Marchi 2022). We found no evidence for strong changes, suggesting that the agricultural transition may have changed the selective environment less than previously expected.
    a. **AP Ragsdale**, **AJ Coffman**, **P Hsieh**, **TJ Struck**, **RN Gutenkunst** (2016) Triallelic population genomics for inferring correlated fitness effects of same site nonsynonymous mutations. Genetics 203:513. [My team was supported by NSF DEB-1146074.]
    b. **X Huang**, **AL Fortier**, **AJ Coffman**, **TJ Struck**, **JE James**, **MN Irby**, **JE Leon-Burguete**, **AP Ragsdale**, **RN Gutenkunst** (2021) Inferring genome-wide correlations of mutation fitness effects between populations. Molecular Biology and Evolution 38:4588. [My team was supported by NIH R01GM127348.]
    c. N Marchi, L Winkelbach, I Schulz, M Brami, Z Hofmanová, J Blöcher, CS Reyna-Blanco, Y Diekmann, A Thiéry, A Kapopoulou, V Link, V Piuz, S Kreutzer, SM Figarska, E Ganiatsou, A Pukaj, **TJ Struck**, **RN Gutenkunst**, N Karul, F Gerritsen, J Pechtl, J Peters, A Zeeb-Lanz, E Lenneis, M Teschler-Nicola, S Triantaphyllou, S Stefanović, C Papageorgopoulou, D Wegmann, J Burger, L Excoffier (2022) The genomic origins of the world's first farmers. Cell. [My team was supported by NIH R01GM127348.]

2. **Computational Methods for Population Genomics**
My group continues to innovate in developing computational methods, particularly extensions to dadi. For many years, a major challenge to dadi use was estimating uncertainties. The only known method involved full inference on many bootstrap data sets, so quantifying uncertainties could increase the computational cost of an analysis by a factor of one hundred. To alleviate this burden, we developed an approach based on the Godambe Information Matrix that estimates uncertainties with only a handful of model simulations (Coffman 2016). Notably, this approach can be applied to other population genomic inference tools based on composite likelihood, particularly those in which model simulation dominates the cost.
Another limitation of dadi is that it is based on a single-locus diffusion equation, so it cannot model linkage between loci. To alleviate this limitation, we developed a diffusion model for pairs of linked loci (Ragsdale 2017). In the neutral case we showed that, for single populations, two-locus statistics are more informative than single-locus statistics, particularly for inferring bottleneck parameters. In unpublished work, we have incorporated selection into the method, preparing it for the applications described in the current proposal.

Most recently, I ported the core dadi integration code to operate on Graphics Processing Unit processors (Gutenkunst 2021). This enables a massive speed up of dadi integrations for problems with many sampled individuals or multiple populations. The gain in speed is sufficient to enable consideration of four- and five-population models, which will be important for our upcoming polyploidy work.

a. **AJ Coffman**, **P Hsieh**, S Gravel, **RN Gutenkunst** (2016) Computationally efficient composite likelihood statistics for demographic inference. Molecular Biology and Evolution 33:591. [My team was supported by NSF DEB-1146074.]
b. **AP Ragsdale**, **RN Gutenkunst** (2017) Inferring demographic history using two-locus statistics. Genetics 206:1037. [My team was supported by NSF DEB-1146074.]
c. **RN Gutenkunst** (2021) dadi.CUDA: Accelerating population genetic inference with Graphics Processing Units. Molecular Biology and Evolution 38:2177. [I was supported by NIH R01GM127348.]

3. **Other Applied Population Genomics Inference**

My research group and I often collaborate with empirical colleagues to apply population genomics inference to diverse data. For example, we collaborated with conservation colleagues to use population RNA seq data to determine whether multiple populations of desert tortoise were exchanging genetic material (Edwards 2016). We found no evidence of recent exchange, providing a strong argument that all populations should be conserved.

We also collaborated extensively with Dr. Michael Hammer on the population genetics of indigenous peoples. We developed a comprehensive model of the demographic history of African Pygmy hunter-gatherer populations and neighboring farmers. We then used genome-scale simulations of this model as a null against which to detect natural selection. We thus identified both individual genes and functionally-related sets of genes that have potentially driven adaptation to the jungle environment (Hsieh 2016). In Hsieh (2017) we used similar approaches to consider Siberian populations, identifying sets of genes important for their adaptation to the Arctic environment.

Most recently, I collaborated with colleagues in estimating the demographic history of corals (Prata 2022), a collaboration that began with a question on the dadi-user mailing list. Our work identified ongoing gene flow between corals on different islands, suggesting replenishment may be possible if individual populations go extinct.

a. T Edwards, M Tollis, **P Hsieh**, **RN Gutenkunst**, Z Liu, K Kusumi, M Culver, RW Murphy (2016) Assessing models of speciation under different biogeographic scenarios; an empirical study using multi-locus and RNA-seq analyses. Ecology and Evolution 6:379. [My team was supported by NSF DEB-1146074.]
b. **P Hsieh**, KR Veeramah, J Lachance, SA Tishkoff, JD Wall, MF Hammer, **RN Gutenkunst** (2016) Whole-genome sequence analyses of Western Central African Pygmy hunter-gatherers reveal a complex demographic history and identify candidate genes under positive natural selection. Genome Research 26:279. [My team was supported by NSF DEB-1146074.]
c. **P Hsieh**, B Hallmark, J Watkins, TM Karafet, LP Osipova, **RN Gutenkunst**, MF Hammer (2017) Exome sequencing provides evidence of polygenic adaptation to a fat-rich animal diet in indigenous Siberian populations. Molecular Biology and Evolution 34:2913. [My team was supported by NSF DEB-1146074.]
d. KE Prata, C Riginos, **RN Gutenkunst**, K Latijnhouwers K, JA Sánchez, N Englebert, K Hay, P Bongaerts (2022) Deep connections: divergence histories with gene flow in mesophotic *Agaricia* corals. Molecular Ecology. [I was supported by NIH R01GM127348.]

4. **Population Genomics of Plants**

We have a fruitful ongoing collaboration with Prof. Mike Barker (see Letter of Support) to extend population genomic analyses in plants. This collaboration began with inference of demographic history using dadi in *Brassica* crops (Qi 2017). This was a remarkable validation of population genomics models, because the inferred demographic model agreed quantitatively with the written record of domestication.

Building on this successful collaboration, we then mentored an NSF Postdoctoral Research Fellow in developing new approaches for plant population genomics. Many plants exhibit high levels of inbreeding or selfing, so we extended dadi to directly model recent inbreeding (Blischak 2020). This approach is not limited to plants, because inbreeding occurs in many species, such as the cougars we also studied in that paper.

Plants also frequently hybridize, which motivated our development of an approach based on convolutional neural networks to infer models of hybridization (Blischak 2021). Again, such hybridization is not restricted to plants; our application in the paper was to Heliconius butterflies. This study honed our expertise in the machine learning approaches we will pursue in the next few years.

a. XC Qi, H An, **AP Ragsdale**, TE Hall, **RN Gutenkunst**, JC Pires, MS Barker (2017) Genomic inferences of domestication events are corroborated by written records in *Brassica rapa*. Molecular Ecology 26:3373. [My team was supported by NSF DEB-1146074.]
b. **PD Blischak**, MS Barker, **RN Gutenkunst** (2020) Inferring the demographic history of inbred species from genome-wide SNP frequency data. Molecular Biology and Evolution 37:2124. [My postdoc was supported by NSF PRFB IOS-1811784, and I was supported by NIH R01GM127348.]
c. **PD Blischak**, MS Barker, **RN Gutenkunst** (2021) Chromosome-scale inference of hybrid speciation and admixture with convolutional neural networks. Molecular Ecology Resources 21:2676. [My postdoc was supported by NSF PRFB IOS-1811784, and I was supported by NIH R01GM127348.]

5. **Cancer genomics:**

Cancer is fundamentally an evolutionary process, so concepts and tools from population genomics can find important application. With a former externally funded Ph.D. student, I recently worked on several cancer genomics projects. In our first project, we developed a simple but effective approach for analyzing genomic data from human tumors transplanted into mice (Mannakee 2018). Contamination of tumor sequence with mouse reads is a major issue for analyzing such data, and our approach efficiently filters problematic reads. It also proved useful in non-xenograft settings, because the same filtering also removes potential false positive variants that might be inferred because of paralogs in the human genome.

A unique challenge in cancer genomics is that tumors are heterogeneous, so variants may be found in almost any fraction of reads from a sample. Low frequency variants may be functionally important, but evidence for them may be concealed by sequencing errors. To enable more powerful variant calling, we developed a Bayesian approach that incorporated the fact that every tumor has a unique spectrum of mutational types that it tends to generate (Mannakee 2020). High-frequency variants can be used to identify this spectrum, so it can be applied to improve identification of low-frequency variants. This work helped motivate our recent explorations of how the DFE for germline mutations differs among mutational types that occur at different rates.

Lastly, we have also collaborated with clinical colleagues on data analysis, with a focus on identifying subclonal variants (Shaheen2022).

a. **BK Mannakee**, U Balaji, AK Witkiewicz, **RN Gutenkunst**, ES Knudsen (2018) Sensitive and specific post-call filtering of genetic variants in xenograft and primary tumors. Bioinformatics 34:1713. [My student was supported by NSF GRF DGE-1143953.]
b. **BK Mannakee**, **RN Gutenkunst** (2020) BATCAVE: Calling somatic mutations with a tumor- and site-specific prior. NAR: Genomics and Bioinformatics 2:lqaa004. [My student was supported by NSF GRF DGE-1143953. I was supported by NIH R01GM127348.]
c. MF Shaheen, JY Tse, ES Sokol, M Masterson, P Bansal, I Rabinowitz, CA Tarleton, AS Dobroff, TL Smith, TJ Bocklage, **BK Mannakee**, **RN Gutenkunst**, JE Bischoff, SA Ness, GM Riedlinger, R Groisberg, R Pasqualini, S Ganesan, W Arap (2022) Genomic landscape of lymphatic malformations and durable complete response to the PI3Kα inhibitor Alpelisib. medRxiv 2022.01.03.21267856. [My student was supported by NSF GRF DGE-1143953.]

Complete List of Published Work in MyBibliography

https://www.ncbi.nlm.nih.gov/myncbi/browse/collection/52622043/?sort=date&direction=descending

Budget Justification

Increase in support relative to NIGMS research support over the previous three years

The amount of PI Gutenkunst's salary requested has increased, to reflect the 51% MIRA research commitment. In addition, COVID inhibited recruitment to the Gutenkunst group, but the group has recently grown to a size corresponding to the requested budget increase.

Indirect Costs

The University of Arizona indirect rate per the agreement approved by DHHS on April 5, 2022 is 53.5% Modified Total Direct Cost (MTDC). Equipment, capital expenditures, tuition remission, rental costs, participant support, scholarships and fellowships, and the portion of sub grants and subcontracts in excess of $25,000 are excluded from MTDC. A copy of the University's DHHS approved rate agreement dated April 22, 2022 is available online at: https://www.fso.arizona.edu/sites/default/files/2022-04/rate_agreement_2022_4.22_0.pdf

Genomic inferences of history and selection across populations and time

BACKGROUND

Fitting computational models to genomic data from populations of humans and other species can reveal the evolutionary past and guide efforts to combat genetic (Di Rienzo 2006) and infectious disease (Messer & Petrov 2013). The distribution of genetic variation within and among populations is shaped by multiple factors, including the mutational input of new variation, negative and positive selection, and demographic history. My *long-term goal* is to understand how natural selection shapes the evolution and function of the molecular networks that comprise life. Toward that goal, my group develops and applies methods for inferring the evolutionary past from population genomic data. My *objectives* over the next five years are to quantify the effects of genetic and environmental contexts on selection, to develop more efficient and powerful inference methods, and to lead the population genomics community in testing methods. My *rationale* is that my research program will both reveal new insights into evolution and enhance the ability of colleagues to reveal complementary insights.

The mutational input of genetic variation is characterized by a distribution of fitness effects (DFE), which quantifies how likely any mutation is to have a given effect on fitness (reviewed in Eyre-Walker & Keightley 2007; Fig 1A). Outside of microbes, experimental DFE measurements are limited by the number of mutations that can be assayed, so inferences from patterns of natural genomic variation are essential (e.g., Keightley & Eyre-Walker 2007; Boyko et al. 2008). A key weakness of existing inference approaches is that they assume each mutation has a fixed effect on fitness. But for many mutations the fitness effect depends on environmental or genomic context (Kondrashov & Houle 1994; Wang et al. 2014). My group has generalized the notion of a DFE beyond a single dimension, to consider joint distributions of fitness effects (Ragsdale et al. 2016; Fig. 1C&E). Recently, we have developed novel methodology to infer how fitness effects may change among populations (Huang et al. 2021) or over time (Marchi et al. 2022). Our next step is to apply this methodology widely, to broadly explore how context affects selection.

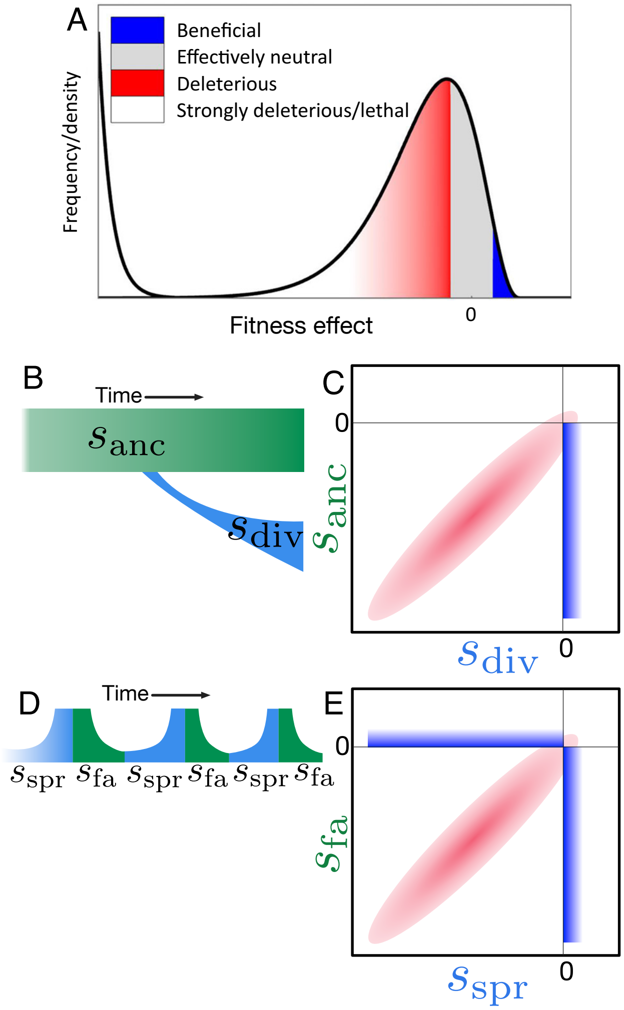

Figure 1: A: The distribution of mutation fitness effects (DFE) quantifies the relative occurrence of mutations with different fitness effects. In general, moderately deleterious mutations are most common and beneficial mutations are rare. (Figure adapted from Bank et al. 2014.) B: The fitness effect of a mutation may differ among populations. For example, a mutation may have fitness effect s_{anc} in the ancestral population and s_{div} in a diverged population. C: In a potential model for the corresponding joint DFE, some fraction of mutations that were neutral or deleterious ($s \leq 0$) in the ancestral population are adaptive ($s > 0$) in the diverged population. D: Mutation fitness effects may also vary over time. For example, a mutation may alternate between fitness effect s_{spr} in Spring and s_{fa} in Fall. Here, the height of the colored region illustrates relative population size. E: In the corresponding joint DFE, most variants have similar effects in both seasons, but some are adaptive in only one season.

My group is best known for developing the software `dadi` for population genomic inference from an allele frequency spectrum (AFS; Gutenkunst et al. 2009). The AFS is a powerful summary of genomic data that records the frequency of each mutation in sequenced samples from one or more populations. Although alternative approaches for inference from such spectra have been developed (Excoffier et al. 2013; Jouganous et al. 2017; Kamm et al. 2020), `dadi` remains widely used. Recently, my group extended the `dadi` framework to calculate the full distribution of configurations for pair of alleles (Ragsdale & Gutenkunst 2017), an innovation I will build upon. Complementary inference approaches based on machine learning are proving successful in population genomics (Schrider & Kern 2018). Thus my group is also developing machine learning approaches to infer demographic history and DFEs.

Thriving scientific communities are built through the exchange of data, methods, and models. Genomics was a pioneer in sharing data, through the Fort Lauderdale agreement (Foster & Sharp 2007). But genomics has lagged behind other fields in sharing computational models. For example, the BioModels database (Le Novère et al. 2006) houses thousands of systems biology models in standardized SBML format (Hucka et al. 2003). In 2018, I helped found the PopSim consortium, which aims to develop a community resource of curated population genomics models and tools to simulate them. We recently published the first release of models, which included a comparison of demographic history inference methods (Adrion et al. 2020). Over the next five years, I will work through the PopSim consortium to further progress the development of population genomics inference.

RECENT PROGRESS

My group is actively exploring the biology of DFEs. In many organisms, the mutation rate of any given genomic base is strongly influenced by the flanking bases (Hodgkinson & Eyre-Walker 2011). And, due to the genetic code, the outcomes of a mutation in coding sequence are strongly constrained by this local context. Thus selection could act to reduce the rates of mutation types that tend to generate more deleterious variants. This is analogous to the drift-barrier hypothesis (Sung et al. 2012), but acting among mutation types within a species, rather than among species. Postdoc Dr. David Castellano has tested this hypothesis using human population genomic data. Strikingly, he found that transition mutations that create more deleterious variants have lower mutations rates, supporting this hypothesis (Fig. 2). He is finalizing his human results, before we apply his workflow to several other species.

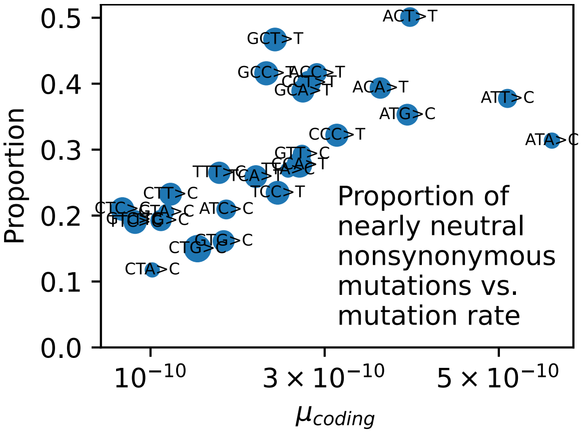

Figure 2: In humans, nonsynonymous transition mutation types that occur more frequently are more likely to have nearly neutral fitness effects. The size of each dot corresponds to the number of those mutations in the 1000 Genomes data. (The 1000 Genomes Project Consortium 2015)

We are also actively expanding models of the joint DFE. With little theory to guide us, our prior work has considered relatively simple models, like bivariate lognormal distributions and mixtures of gamma distributions (Ragsdale et al. 2016; Huang et al. 2021; Marchi et al. 2022). But these models are difficult to generalize to include positive selection. We are collaborating with Prof. Sebastián Ramos-Onsins (see Letter of Support) on a rigorous simulation study to evaluate the ability of joint DFE inference to quantify shifts in selective effects during domestication. To this end, we have developed models that explicitly account for the proportion of mutations that shift from deleterious to advantageous and vice versa. We will apply these new models in our upcoming inferences of joint DFEs.

Collaborating with Prof. Mike Barker (see Letter of Support), we have recently extended `dadi` inference to polyploids, focusing on hybrids where recombination among chromosomes from different parent lineages is restricted. In our model, chromosomes from different lineages are treated as separate populations and recombination among them is treated as migration. In application to data from *Capsella bursa-pastoris* (Douglas et al. 2015), we find strong evidence for such recombination. This work sets the stage for our upcoming analyses of DFEs within polyploids, which will be led by postdoc Dr. Justin Conover, who has been awarded an NSF Postdoctoral Research Fellowship in Biology to fund this research.

In addition to expanding `dadi`'s capabilities, we are also enhancing its speed and usability. My Ph.D. student Ms. Linh Tran has found that multilayer neural networks can infer demographic model parameters (Fig. 3A) from allele frequency spectra with accuracy comparable to likelihood inference with `dadi` (Fig. 3B&C). Importantly, we can also estimate rigorous confidence intervals using a jacknife+ approach (Fig. 3D; Barber et al. 2021). Once trained, application of such networks is almost instantaneous. We anticipate training and distributing networks for the models available within `dadi` and the popular Portik et al. (2017) pipeline. Likelihood optimization will still be necessary when using custom models or estimating parameter covariances (Coffman et al. 2016). Previously, such `dadi` usage required Python scripting, and parallelization required custom workflows. With funding from an Administrative Supplement, we have developed a command-line interface to `dadi` that enables distributed computation using Work Queue (Bui et al. 2011) across heterogenous systems, including local or cloud clusters.

The PopSim consortium is making great strides. With community help, the catalog of models has been expanded dramatically, particularly for non-model organisms, such as mosquito, cow, and orangutans. We are drafting a manuscript laying out rigorous and transparent procedures for adding new organisms to the catalog. We have also developed a YAML format for specifying complex demographic models, called `demes`, which `dadi` and other simulators now support. A manuscript detailing this format is imminent. Lastly and most fundamentally, we are extending the stdpopsim simulation framework to include models of natural selection (using the SLiM simulation engine; Haller & Messer 2019), with the ability to apply different DFEs to different genomic annotations (coding regions, enhancers, etc.). These last two developments enable the inference competitions I will lead in the next period.

FUTURE RESEARCH PROGRAM

In the next funding period, my research program will continue to explore differences in mutation fitness effects, to develop novel population genomic inference methods, and to support the population genomics community. Here I describe some of our plans.

Differences in Mutation Fitness Effects

Our joint DFE approach offers a novel genome-wide perspective on how mutation fitness effects can differ among populations and eras. The central scientific question is now: **What biological factors drive differences in mutation fitness effects?** To address this question, postdoc Dr. Emanuel Fonseca pursue multiple analyses, some detailed below.

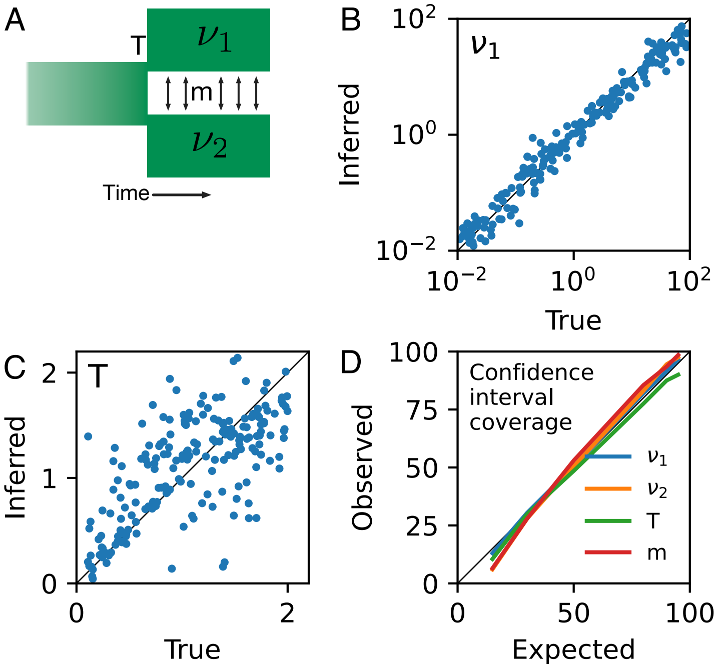

Figure 3: A: In this demographic history model, the ancestral population splits a time T in the past, into contemporary populations with size ν_1 and ν_2 relative to the ancestral, with symmetric migration at rate m between them. B: From an allele frequency spectrum with 10 individuals per population, relative population sizes can be inferred precisely, over a range of simulated true values. C: The divergence time T is inferred less precisely, although this precision matches that of full likelihood optimization in `dadi`. D: Our approach estimates confidence intervals accurately. For example, the estimated 95% confidence interval is observed to cover close to 95% of the true simulated values for all four model parameters.

Understanding how mutation fitness effects depend on context is of fundamental importance. During environmental speciation, divergent selection on mutations with different fitness effects in different environments drives the initial genetic differentiation that eventually leads to reproductive isolation (Schluter 2009; Seehausen et al. 2014). By quantifying divergent selection, our analyses will identify molecular functions most likely to drive speciation. Many human disease variants are mildly deleterious (Di Rienzo 2006), which affects their distribution among populations and thus the power of Genome-Wide Associate Studies (GWAS; Uricchio et al. 2016). Understanding how selection differs among human populations and how those differences depend on gene function may thus help in understanding the transferability of GWAS results among populations (Martin et al. 2017). Mutation fitness effects may also differ over time, because to environmental changes such as climate change or, in humans, the advent of agriculture. Lastly, because sets of genes with different functions may differ in their sensitivity to environmental and genetic context, we often segment our data using various ontologies.

Among Populations

Several studies have found evidence for differences in the DFE among populations of humans and other primates (Boyko et al. 2008; Ma et al. 2013; Kim et al. 2017; Castellano et al. 2019; Tataru & Bataillon 2019). But these studies are weakened by the implicit assumption that the fitness effects of a given mutation in different populations are independent draws from distinct DFEs. Our joint DFE approach removes this restrictive assumption, and so far we have focused our attention on the correlation of fitness effects among populations. Our recent work in humans, *Drosophila*, and wild tomatoes (Huang et al. 2021) shows that the strength of this correlation can vary dramatically. A fundamental question is: To what degree are differences in mutation fitness effects among populations driven by differences in external environmental context or internal genetic context (i.e., epistasis)? Among our three examples, greater genetic differentiation (as assessed by F_{ST}) led to lower correlation in mutation fitness effects, but we cannot rigorously derive general principles from so few examples.

Our next application of joint DFE analysis will be to Mexican sulfidic fishes, in collaboration with Prof. Joanna Kelley (see Letter of Support). In several locations in Mexico, sulfidic springs flow into nearby freshwater streams. The water from these springs is toxic to most fish, but several species have adapted to colonize the inflow. Prof. Kelley's lab has observed parallel phenotypic changes among these species, driven by distinct genomic changes (Greenway et al. 2021). With joint DFE analysis, based on models such as Fig. 1B&C, my group will rigorously quantify the extent to which these adaptations arise from mutations that are nearly neutral or deleterious in the freshwater fishes. We will also quantify the degree to which purifying selection has changed in this new environment, for example, due to reduced predation. The replication of these adaptations across several species and springs will aid in deriving general conclusions about the factors that drive differences in fitness effects.

Beyond the sulfidic fishes, we foresee applications of joint DFE analysis to multiple systems. For example, *Anolis* lizards have undergone an explosive radiation into multiple distinct populations. Bourgeois et al. (2019) resequenced five populations covering the species range, and studying their joint DFEs will reveal broadly how selection has changed across the radiation. *Arabidopsis lyrata* is an emerging model for local adaptation, and Hämälä & Savolainen (2019) have sequenced two pairs of low- and high-altitude populations. Joint DFE analysis will reveal the extent to which altitude-dependent differences in selection are general or location-specific. Lastly, the house mouse is an attractive model because of the extensive functional data available. Harr et al. (2016) sequenced several species of house mouse, offering several pairwise comparisons of selection pressures.

Among Eras

Mutation fitness effects may also change over time. In population genomics, this is most well studied in the case of adaptation from existing neutral genetic variation (which generates soft sweeps; Hermisson & Pennings 2005). Repeated changes in the sign of fitness effects can lead to balancing selection, which preserves diversity on long timescales (Charlesworth 2006). Our joint DFE analysis can rigorously encompass both scenarios, along with other scenarios that, for example, only affect the magnitude of deleterious fitness effects.

A particularly well-studied case of temporal changes in fitness effects is seasonal adaptation in *Drosophila melanogaster*. Bergland et al. (2014) found hundreds of genetics variants that rose and fell in frequency with the seasons in several *D. mel.* orchard populations. This could have been caused by repeated local adaptation or migration, but Rudman et al. (2022) used outdoor cages to provide evidence that these changes are local adaptations. Recently, Machado et al. (2021) published genome-wide pooled sequencing data for *D. mel.* with repeated sampling between Fall and Spring seasons in several locales. Machado et al. (2021) focused their analyses on pairwise comparisons between Spring and Fall seasons at the same location, considering each year separately. But we will fit demographic and joint DFE models (Fig. 1D&E) to the data from five locations that were sampled over multiple years. Our modeling will quantitatively estimate the proportion of mutations that are seasonally responsive, particularly when rigorously calibrated against simulations that include linkage. By comparing among locations, we will also assess how the magnitude of seasonal variation impacts mutation fitness effects.

Ancient DNA data offer another perspective on temporal changes in fitness effects. We have already extended our joint DFE analysis to ancient samples, with application to human populations within Europe (Marchi et al. 2022). We anticipate that during the next five years sufficient Neanderthal genomes will become available to carry out joint DFE analysis between humans and Neanderthals. We expect this analysis will offer insight into how selection differed between Neanderthals and humans and how that affected which genomic segments could introgress into humans.

Overall, we will extend our joint DFE analyses to many biological situations, including those involving adaptation. Our *expected outcome* is that we will begin to infer general principles from these examples for how mutation fitness effects depend on genomic and ecological context. These principles will enhance understanding of short-term population differentiation and long-term speciation.

Population Genomic Inference Methods

My group has long been at the forefront of methods development for population genomics. Our work has typically been based on underlying diffusion theory (Kimura 1964), which incorporates selection more easily than coalescent theory (Wakeley 2009). But recently we have developed expertise in machine learning and neural networks (e.g., Blischak et al. 2021), which is opening up new avenues.

Two-locus Inference

A weakness of dadi is that the underlying theory does not model linkage among variants that are nearby on a chromosome. In the absence of selection, this causes inference precision to be overestimated, which can be rigorously corrected using our Godambe Information Matrix approach (Coffman et al. 2016). But with selection, linkage results in background selection or hitchiking, which can bias inferences. Several years ago, we developed a numerical solution to the two-locus diffusion equation (Ragsdale & Gutenkunst 2017) that can model selection on one or both loci. Since then, others have solved the same equation using moments-based methods, which may be more efficient in some cases (Ragsdale 2021; Friedlander & Steinrücken 2022). Regardless of the solution method, modeling selection on pairs of loci opens new avenues for inference.

Postdoc Dr. David Castellano will first apply two-locus inference to recombination rate estimation. The rate of recombination varies dramatically across the genome of humans and other species (Peñalba & Wolf 2020), and that variation can be inferred from population genomic data (Stumpf & McVean 2003). The most modern method, pyrho (Spence & Song 2019), depends on tables of two-locus probabilities calculated by LDpop (Kamm et al. 2016), which models single-population demographic history but not selection. These methods have recently been tested with neutral simulations (Samuk & Noor 2021; Raynaud et al. 2022), but potential biases caused by selection have not been characterized. Because recombination governs background selection and hitchiking, biased estimates in regions of the genome undergoing substantial selection are problematic. To test for such biases, we will use simulations in SLiM (Haller & Messer 2019) with realistic gene structure and DFEs. We will then correct any biases found by using our two-locus methods to calculate probability tables that directly model natural selection. Our *expected outcome* is a more rigorous set of methods for inferring recombination maps, which can be widely applied.

Deep Learning for DFEs

To infer the DFE from population genetic data, typical approaches begin by fitting a model of demographic history to putatively neutral variants (often synonymous coding variants; Eyre-Walker et al. 2006; Boyko et al. 2008). Model allele frequency spectra are then computed by including selection in addition to that demographic history. The expected AFS for functional sites is then a sum of model spectra over fitness effects, weighted by the DFE (Ragsdale et al. 2016; Kim et al. 2017). This approach has proven powerful, and it is the basis of our current joint DFE inferences. But it neglects substantial information in the data, because it ignores all patterns of linkage among variants. Recently, Johri et al. (2020) developed an Approximate Bayesian Computation (ABC) approach for simultaneously inferring demographic history and the DFE, incorporating some linkage summary statistics. But ABC approaches demand careful curating of summary statistics and discard many of the simulations used to train them (Beaumont 2010). Deep learning approaches can be more powerful, because they implicitly learn the most informative summary statistics and incorporate information from all training simulations (Schrider & Kern 2018). We will thus develop a deep learning approach for inferring DFEs.

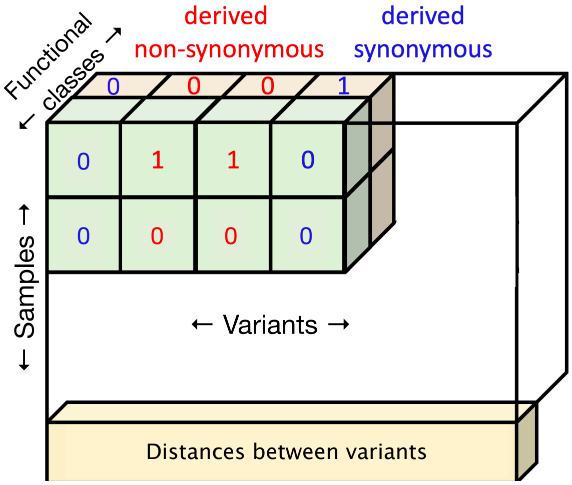

Figure 4: Our novel representation of population genomic data incorporates different types of mutations into different layers of the tensor. If a sampled haplotype contains the derived allele for a mutation of a given type, that entry in the tensor is encoded 1, otherwise 0.

Our DFE inference approach will be based on a novel representation of population genomic data. Previous work has summarized population genomic data into a 2D matrix in which each column represents a variant site, coded as 1 for derived allele versus 0 for ancestral allele, supplemented by a vector containing the distances between variant sites (e.g., Sheehan & Song 2016; Flagel et al. 2019; Sanchez et al. 2021). In DFE inference, we are often interested in how selective forces differ among variants of different functional classes, such as synonymous, nonsynonymous, intronic, etc. To represent these different functional classes, we will expand the 2D matrix into a 3D tensor, in which each layer corresponds to a different functional class (Fig. 4). We will then feed this data representation into a convolutional neural network that is invariant to permutation of the rows (corresponding to the arbitrary ordering of samples in the data; Chan et al. 2018). The filters thus encompass columns, corresponding to nearby variants in the genome, and layers, corresponding to different functional classes. The outputs of these filters will then be passed into a neural network architecture modeled on that of Sanchez et al. (2021), from which the final outputs are estimates of the DFE parameters.

We will train our neural network on data simulated with SLiM (Haller & Messer 2019), using human-like parameters for demographic history, gene architecture, mutation, and recombination rates. Once trained, we will apply our network to human data from the 1000 Genomes Project (The 1000 Genomes Project Consortium 2015), comparing our estimates to more traditional approaches (Kim et al. 2017). We expect to find higher precision of our estimates, given the additional linkage information our method incorporates. We will then explore combined inferences of DFEs for nonsynonymous mutations and mutations in regulatory sequences, including 3' and 5' untranslated regions and promoters. Our *expected outcome* is a fundamentally new approach to infer DFEs that can be applied to many species, along with more precise inferences in humans.

Community Competition

As sequencing improvements have expanded population genomic data, methods for analyzing those data have proliferated (Salojärvi 2019). Understanding which method is best for a given scientific question and data set is a major challenge. While most new methods papers include assessments, they are rarely comparable across papers. Moreover, self-assessments can be compromised, even unconsciously, because the author typically knows the correct result. Blinded inference competitions offer more rigorous and transparent testing. For example, the long-running Critical Assessment of protein Structure Prediction (Moult et al. 1995; Kryshtafovych et al. 2021) has dramatically advanced protein structure prediction, culminating in the revolutionary development of AlphaFold2 (Jumper et al. 2021). For fifteen years, the Dialogue on Reverse Engineering Assessments and Methods (DREAM) has run a diverse range of competitions (Stolovitzky et al. 2007). The initial focus of DREAM was systems biology, but they have branched out into other biological inference problems (Mason et al. 2020). I will work with DREAM to run several competitions for the population genomics community (see Letter of Support from Dr. Pablo Meyer).

Given my expertise, the first competition will be in demographic history inference. A small committee will develop a series of increasingly complex challenges, involving demographic histories with one or more populations and varying patterns of genomic selection. We will use the stdpopsim simulation framework to generate data for these challenges (see Letter of Support from Prof. Kern and Prof. Ralph). For focused tests on inference methods, in some challenges Variant Call Format (VCF; Danecek et al. 2011) files directly from the simulations will be distributed. To test full pipelines, in other challenges simulated read data generated using NEAT (Stephens et al. 2016) will be distributed. Challenges will be advertised on social media, mailing lists such as Evoldir and the Society for Molecular Biology and Evolution, and machine learning forums. Competitors will submit inferred models in the `demes` specification. Two scoring systems will be employed. The first will simply compare inferred and simulated parameter values (divergence times, etc.). The second will compare distributions of coalescent times between inferred and true models. We will organize this competition using the Synapse platform developed by Sage Bionetworks, which provides storage and computing and has hosted many DREAM competitions.

Subsequent competitions will encompass other important realms of population genomics inference, with the inclusion of suitable experts on the design committees. Inference of individual ancestry (Royal et al. 2010) is central to many direct-to-consumer genetic testing companies, and the results can have profound implications for consumers (Roth & Ivemark 2018). An ancestry inference challenge will require simulation of many genomes to provide a reference panel, along with complex test individuals, but it promises to have high impact. Statistical methods for Genome-Wide Association Studies (GWAS) also demand testing, particularly those that attempt to correct for population structure (Sul et al. 2018). This competition will leverage SLiM's ability to simulate spatial models and selection on quantitative traits (Haller & Messer 2019).

Our *expected outcomes* are popular competitions that will not only identify the best-performing existing inference methods, but will also spur the development of new methods from researchers outside the existing population genomics community. The competitions will also provide a powerful training opportunity for new researchers, who can hone their skills with different methods by applying them to the challenge data.

Through these competitions, and my group's own research on multidimensional DFEs and novel inference methods, I expect that my research program will continue to have a large and positive impact on the population genomics community.

References Cited

Adrion JR, Cole CB, Dukler N, Galloway JG, Gladstein AL, Gower G, Kyriazis CC, Ragsdale AP, Tsambos G, Baumdicker F, Carlson J, Cartwright RA, Durvasula A, Gronau I, Kim BY, McKenzie P, Messer PW, Noskova E, Vecchyo DOD, Racimo F, Struck TJ, Gravel S, Gutenkunst RN, Lohmueller KE, Ralph PL, Schrider DR, Siepel A, Kelleher J, Kern AD (2020) A community-maintained standard library of population genetic models. Elife 9:e54967.

Bank C, Ewing GB, Ferrer-Admettla A, Foll M, Jensen JD (2014) Thinking too positive? Revisiting current methods of population genetic selection inference. Trends Genet 30:540.

Barber RF, Candès EJ, Ramdas A, Tibshirani RJ (2021) Predictive inference with the jackknife+. Ann Stat 49:486.

Beaumont Ma (2010) Approximate Bayesian Computation in Evolution and Ecology. Annu Rev Ecol Evol Syst 41:379.

Bergland AO, Behrman EL, O'Brien KR, Schmidt PS, Petrov DA (2014) Genomic Evidence of Rapid and Stable Adaptive Oscillations over Seasonal Time Scales in Drosophila. PLoS Genet 10:e1004775.

Blischak PD, Barker MS, Gutenkunst RN (2021) Chromosome-scale inference of hybrid speciation and admixture with convolutional neural networks. Mol Ecol Resour 21:2676.

Bourgeois Y, Ruggiero RP, Manthey JD, Boissinot S, Gojobori T (2019) Recent Secondary Contacts, Linked Selection, and Variable Recombination Rates Shape Genomic Diversity in the Model Species Anolis carolinensis. Genome Biol Evol 11:2009.

Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, Adams MD, Schmidt S, Sninsky JJ, Sunyaev SR, White TJ, Nielsen R, Clark AG, Bustamante CD (2008) Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet 4:e1000083.

Bui P, Rajan D, Abdul-Wahid B, Izaguirre J, Thain D (2011) Work Queue + Python: A Framework For Scalable Scientific Ensemble Applications. In *Work. Python High Perform. Sci. Comput.*.

Castellano D, MacIà MC, Tataru P, Bataillon T, Munch K (2019) Comparison of the full distribution of fitness effects of new amino acid mutations across great apes. Genetics 213:953.

Chan J, Spence JP, Mathieson S, Perrone V, Jenkins PA, Song YS (2018) A likelihood-free inference framework for population genetic data using exchangeable neural networks. In *Adv. Neural Inf. Process. Syst.*, pages 8594–8605.

Charlesworth D (2006) Balancing selection and its effects on sequences in nearby genome regions. PLoS Genet 2:379.

Coffman AJ, Hsieh P, Gravel S, Gutenkunst RN (2016) Computationally efficient composite likelihood statistics for demographic inference. Mol Biol Evol 33:591.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R (2011) The variant call format and VCFtools. Bioinformatics 27:2156.

Di Rienzo A (2006) Population genetics models of common diseases. Curr Opin Genet Dev 16:630.

Douglas GM, Gos G, Steige KA, Salcedo A, Holm K, Josephs EB, Arunkumar R, Ågren JA, Hazzouri KM, Wang W, Platts AE, Williamson RJ, Neuffer B, Lascoux M, Slotte T, Wright SI (2015) Hybrid origins and the earliest stages of diploidization in the highly successful recent polyploid Capsella bursa-pastoris. Proc Natl Acad Sci U S A 112:2806.

Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M (2013) Robust demographic inference from genomic and SNP data. PLoS Genet 9:e1003905.

Eyre-Walker A, Keightley PD (2007) The distribution of fitness effects of new mutations. Nat Rev Genet 8:61061.

Eyre-Walker A, Woolfit M, Phelps T (2006) The distribution of fitness effects of new deleterious amino acid mutations in humans. Genetics 173:891.

Flagel L, Brandvain Y, Schrider DR (2019) The unreasonable effectiveness of convolutional neural networks in population genetic inference. Mol Biol Evol 36:220.

Foster MW, Sharp RR (2007) Share and share alike: deciding how to distribute the scientific and social benefits of genomic data. Nat Rev Genet 8:633.

Friedlander E, Steinrücken M (2022) A numerical framework for genetic hitchhiking in populations of variable size. Genetics 220:iyac012.

Greenway R, Brown AP, Camarillo H, Delich C, Mcgowan KL, Nelson J, Arias-rodriguez L, Kelley JL, Tobler M (2021) Convergent adaptation and ecological speciation result from unique genomic mechanisms in sympatric extremophile fishes. bioRxiv .

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet 5:e1000695.

Haller BC, Messer PW (2019) SLiM 3: Forward Genetic Simulations Beyond the Wright-Fisher Model. Mol Biol Evol 36:632.

Hämälä T, Savolainen O (2019) Genomic Patterns of Local Adaptation under Gene Flow in Arabidopsis lyrata. Mol Biol Evol 36:2557.

Harr B, Karakoc E, Neme R, Teschke M, Pfeifle C, Pezer Ž, Babiker H, Linnenbrink M, Montero I, Scavetta R, Abai MR, Molins MP, Schlegel M, Ulrich RG, Altmüller J, Franitza M, Büntge A, Künzel S, Tautz D (2016) Genomic resources for wild populations of the house mouse, Mus musculus and its close relative Mus spretus. Sci Data 3:160075.

Hermisson J, Pennings PS (2005) Soft sweeps: Molecular population genetics of adaptation from standing genetic variation. Genetics 169:2335.

Hodgkinson A, Eyre-Walker A (2011) Variation in the mutation rate across mammalian genomes. Nat Rev Genet 12:756.

Huang X, Fortier AL, Coffman AJ, Struck TJ, Irby MN, James JE, León-Burguete JE, Ragsdale AP, Gutenkunst RN (2021) Inferring Genome-Wide Correlations of Mutation Fitness Effects between Populations. Mol Biol Evol 38:4588.

Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, Arkin aP, Bornstein BJ, Bray D, Cornish-Bowden A, Cuellar AA, Dronov S, Gilles ED, Ginkel M, Gor V, Goryanin II, Hedley WJ, Hodgman TC, Hofmeyr JH, Hunter PJ, Juty NS, Kasberger JL, Kremling A, Kummer U, Le Novere N, Loew LM, Lucio D, Mendes P, Minch E, Mjolsness ED, Nakayama Y, Nelson MR, Nielsen PF, Sakurada T, Schaff JC, Shapiro BE, Shimizu TS, Spence HD, Stelling J, Takahashi K, Tomita M, Wagner J, Wang J (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics 19:524.

Johri P, Charlesworth B, Jensen JD (2020) Jointly Inferring Demography and Purifying Selection. Genetics 215:173.

Jouganous J, Long W, Ragsdale AP, Gravel S (2017) Inferring the Joint Demographic History of Multiple Populations: Beyond the Diffusion Approximation. Genetics 206:1549.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D (2021) Highly accurate protein structure prediction with AlphaFold. Nature 596:583.

Kamm J, Terhorst J, Durbin R, Song YS (2020) Efficiently Inferring the Demographic History of Many Populations With Allele Count Data. J Am Stat Assoc 115:1472.

Kamm JA, Spence JP, Chan J, Song YS (2016) Two-locus likelihoods under variable population size and fine-scale recombination rate estimation. Genetics 203:1381.

Keightley PD, Eyre-Walker A (2007) Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. Genetics 177:2251.

Kim BY, Huber CD, Lohmueller KE (2017) Inference of the distribution of selection coefficients for new nonsynonymous mutations using large samples. Genetics 206:345.

Kimura M (1964) Diffusion Models in Population Genetics. J Appl Probab 1:177.

Kondrashov AS, Houle D (1994) Genotype-environment interactions and the estimation of the genomic mutation rate in Drosohpila melanogaster. Proc R Soc B 258:221.

Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J (2021) Critical assessment of methods of protein structure prediction (CASP)—Round XIV. Proteins Struct Funct Bioinforma 89:1607.

Le Novère N, Bornstein B, Broicher A, Courtot M, Donizelli M, Dharuri H, Li L, Sauro H, Schilstra M, Shapiro B, Snoep JL, Hucka M (2006) BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. Nucleic Acids Res 34:D689.

Ma X, Kelley JL, Eilertson K, Musharoff S, Degenhardt JD, Martins AL, Vinar T, Kosiol C, Siepel A, Gutenkunst RN, Bustamante CD (2013) Population genomic analysis reveals a rich speciation and demographic history of orang-utans (Pongo pygmaeus and Pongo abelii). PLoS One 8:e77175.

Machado HE, Bergland AO, Taylor R, Tilk S, Behrman E, Dyer K, Fabian DK, Flatt T, González J, Karasov TL, Kim B, Kozeretska I, Lazzaro BP, Merritt TJ, Pool JE, O'brien K, Rajpurohit S, Roy PR, Schaeffer SW, Serga S, Schmidt P, Petrov DA (2021) Broad geographic sampling reveals the shared basis and environmental correlates of seasonal adaptation in Drosophila. Elife 10:e67577.

Marchi N, Winkelbach L, Schulz I, Brami M, Hofmanová Z, Blöcher J, Reyna-Blanco CS, Diekmann Y, Thiéry A, Kapopoulou A, Link V, Piuz V, Kreutzer S, Figarska SM, Ganiatsou E, Pukaj A, Struck TJ, Gutenkunst RN, Karul N, Gerritsen F, Pechtl J, Peters J, Zeeb-Lanz A, Lenneis E, Teschler-Nicola M, Triantaphyllou S, Stefanović S, Papageorgopoulou C, Wegmann D, Burger J, Excoffier L (2022) The genomic origins of the world's first farmers. Cell .

Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, Daly MJ, Bustamante CD, Kenny EE (2017) Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. Am J Hum Genet 100:635.

Mason MJ, Schinke C, Eng CL, Towfic F, Gruber F, Dervan A, White BS, Pratapa A, Guan Y, Chen H, Cui Y, Li B, Yu T, Chaibub Neto E, Mavrommatis K, Ortiz M, Lyzogubov V, Bisht K, Dai HY, Schmitz F, Flynt E, Dan Rozelle, Danziger SA, Ratushny A, Dalton WS, Goldschmidt H, Avet-Loiseau H, Samur M, Hayete B, Sonneveld P, Shain KH, Munshi N, Auclair D, Hose D, Morgan G, Trotter M, Bassett D, Goke J, Walker BA, Thakurta A, Guinney J (2020) Multiple Myeloma DREAM Challenge reveals epigenetic regulator PHF19 as marker of aggressive disease. Leukemia 34:1866.

Messer PW, Petrov DA (2013) Population genomics of rapid adaptation by soft selective sweeps. Trends Ecol Evol 28:659.

Moult J, Pedersen JT, Judson R, Fidelis K (1995) A large-scale experiment to assess protein structure prediction methods. Proteins Struct Funct Bioinforma 23:ii.

Peñalba JV, Wolf JB (2020) From molecules to populations: appreciating and estimating recombination rate variation. Nat Rev Genet 21:476.

Portik DM, Leaché AD, Rivera D, Barej MF, Burger M, Hirschfeld M, Rödel MO, Blackburn DC, Fujita MK (2017) Evaluating mechanisms of diversification in a Guineo-Congolian tropical forest frog using demographic model selection. Mol Ecol 26:5245.

Ragsdale AP (2021) Can we distinguish modes of selective interactions using linkage disequilibrium? bioRxiv .

Ragsdale AP, Coffman AJ, Hsieh P, Struck TJ, Gutenkunst RN (2016) Triallelic population genomics for inferring correlated fitness effects of same site nonsynonymous mutations. Genetics 203:513.

Ragsdale AP, Gutenkunst RN (2017) Inferring demographic history using two-locus statistics. Genetics 206:1037.

Raynaud M, Gagnaire Pa, Galtier N (2022) Performance and limitations of linkage-disequilibrium-based methods for inferring the genomic landscape of recombination and detecting hotspots: a simulation study. bioRxiv .

Roth WD, Ivemark B (2018) Genetic options: The impact of genetic ancestry testing on consumers' racial and ethnic identities. Am J Sociol 124:150.

Royal CD, Novembre J, Fullerton SM, Goldstein DB, Long JC, Bamshad MJ, Clark AG (2010) Inferring Genetic Ancestry: Opportunities, Challenges, and Implications. Am J Hum Genet 86:661.

Rudman SM, Greenblum SI, Rajpurohit S, Betancourt NJ, Hanna J, Tilk S, Yokoyama T, Petrov DA, Schmidt P (2022) Direct observation of adaptive tracking on ecological time scales in Drosophila. Science 375:eabj7484.

Salojärvi J (2019) *Computational Tools for Population Genomics*, pages 127–160. Springer International Publishing, Cham.

Samuk K, Noor M (2021) Gene flow biases population genetic inference of recombination rate. bioRxiv .

Sanchez T, Cury J, Charpiat G, Jay F (2021) Deep learning for population size history inference: Design, comparison and combination with approximate Bayesian computation. Mol Ecol Resour 21:2645.

Schluter D (2009) Evidence for ecological speciation and its alternative. Science 323:737.

Schrider DR, Kern AD (2018) Supervised Machine Learning for Population Genetics: A New Paradigm. Trends Genet 34:301.

Seehausen O, Butlin RK, Keller I, Wagner CE, Boughman JW, Hohenlohe PA, Peichel CL, Saetre GP, Bank C, Brännström Å, Brelsford A, Clarkson CS, Eroukhmanoff F, Feder JL, Fischer MC, Foote AD, Franchini P, Jiggins CD, Jones FC, Lindholm AK, Lucek K, Maan ME, Marques DA, Martin SH, Matthews B, Meier JI, Möst M, Nachman MW, Nonaka E, Rennison DJ, Schwarzer J, Watson ET, Westram AM, Widmer A (2014) Genomics and the origin of species. Nat Rev Genet 15:176.

Sheehan S, Song YS (2016) Deep Learning for Population Genetic Inference. PLoS Comput Biol 12:e1004845.

Spence JP, Song YS (2019) Inference and analysis of population-specific fine-scale recombination maps across 26 diverse human populations. Sci Adv 5:eaaw9206.

Stephens ZD, Hudson ME, Mainzer LS, Taschuk M, Weber MR, Iyer RK (2016) Simulating next-generation sequencing datasets from empirical mutation and sequencing models. PLoS One 11:e0167047.

Stolovitzky G, Monroe D, Califano A (2007) Dialogue on reverse-engineering assessment and methods: the DREAM of high-throughput pathway inference. Ann NY Acad Sci 1115:1.

Stumpf MPH, McVean GAT (2003) Estimating recombination rates from population-genetic data. Nat Rev Genet 4:959.

Sul JH, Martin LS, Eskin E (2018) Population structure in genetic studies: Confounding factors and mixed models. PLoS Genet 14:e1007309.

Sung W, Ackerman MS, Miller SF, Doak TG, Lynch M (2012) Drift-barrier hypothesis and mutation-rate evolution. Proc Natl Acad Sci U S A 109:18488.

Tataru P, Bataillon T (2019) PolyDFEv2.0: Testing for invariance of the distribution of fitness effects within and across species. Bioinformatics 35:2868.

The 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. Nature 526:68.

Uricchio LH, Zaitlen NA, Ye CJ, Witte JS, Hernandez RD (2016) Selection and explosive growth alter genetic architecture and hamper the detection of causal rare variants. Genome Res 26:863.

Wakeley J (2009) Coalescent theory. Roberts & Company .

Wang AD, Sharp NP, Agrawal AF (2014) Sensitivity of the distribution of mutational fitness effects to environment, genetic background, and adaptedness: A case study with Drosophila. Evolution 68:840.

Resource Sharing Plan

The methods and inferences developed during this research will be disseminated through presentation at meetings such as the Annual Meeting of the Society for Molecular Biology and Evolution. Articles will be submitted to high-impact interdisciplinary journals, as well as deposited on my research group website. To assure rigor and transparency, code implementing the analyses in each article will be archived in repositories such as Dryad.

Data Sharing Plan

No new data are expected to be generated during the proposed research.

Software Sharing Plan

I have a history of developing and disseminating software tools. I developed the systems biology software `SloppyCell` and the population genetics software `dadi`. `dadi` is widely used in the population genetics community, with thousands of downloads and a mailing list that receives roughly 30 posts a month.

All software developed during the proposed research will be freely available to biomedical researchers in both the non-profit and commercial sectors. To ensure this, all software will be licensed under the permissive open-source MIT license. This license enables free dissemination, modification, and commercialization of the software. All software development will occur within version-controlled git repositories, hosted on Bitbucket. Upon initial release, these repositories will be made public, preserving the entire history of the code. Thus other researchers will be able continue to use and enhance the software even if our team cannot provide continued support. In the event Bitbucket ceases operation, our software repositories will be moved to another public hosting service, such as GitHub.

All documentation and tutorials will be distributed as part of the code repository. Active support will be provided through Google Groups mailing lists. The archives of these lists will be publicly searchable, providing an additional permanent resource for users of the software.