UNIVERSITY OF CALIFORNIA SAN DIEGO

Analyzing Genetic Adaptation in Action: Identifying the Evolutionary Mechanisms Rescuing Stressed Populations

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Electrical Engineering (Intelligent Systems, Robotics, and Control)

by

Arya Iranmehr

Committee in charge:

Professor Vineet Bafna, Co-Chair
Professor Siavash Mirarab, Co-Chair
Professor Gabriel G. Haddad
Professor Sergey Kryazhimskiy
Professor Nuno Vasconcelos

2019

The dissertation of Arya Iranmehr is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____
Co-Chair

_____
Co-Chair

University of California San Diego

2019

DEDICATION

To my loving wife, Shohreh,

my sweethearts, Liam and Nick

and my caring parents.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to express my gratitude to my advisor, Vineet Bafna, for his support and guidance, and encouragement over the past years. I consider myself fortunate to be mentored by such a distinguished scientist with an amazing personality.

I would like to thank my thesis committee for their advice and encouragement along the way; Vineet Bafna, Siavash Mirarab, Gabriel Haddad, Sergey Kryazhimskiy, and Nuno Vasconcelos. In particular, I would like to thank Gabriel Haddad for his support and valuable advice during the past years and for being a wonderful collaborator.

I would like to thank my beloved wife, Shohreh, for her unconditional emotional support and encouragement and being there during hard times.

Finally, everything I have accomplished so far would not have been possible without relentless devotion of my hardworking parents, who loved and provided with all the supports I needed.

Chapter 2, in full, contains material from Arya Iranmehr, Ali Akbari, Christian Schltterer, Vineet Bafna. "Clear: composition of likelihoods for evolve and resequence experiments". Genetics. 2017 Jan 1:genetics-116 [1]. I was the primary investigator and author of this paper.

Chapter 3, in full, contains material from Arya Iranmehr, Tsering Stobdan, Dan Zhou, Vineet Bafna, Gabriel G Haddad and Helen Zhao. "Revealing Evolutionary Forces via Experimental Evolution". In preparation. I was the primary investigator and author of this paper.

Chapter 4, in full, contains material from Arya Iranmehr, Tsering Stobdan, Dan Zhou, Orit Poulsen, Kingman P. Strohl, Almaz Aldashev, Amalio Telenti, Emily HM Wong, Ewen F Kirkness, J Craig Venter, Vineet Bafna, Gabriel G Haddad. "Novel insight into the genetic basis of high-altitude pulmonary hypertension in Kyrgyz highlanders. European Journal of Human Genetics". 2019 Jan;27(1):150 [2]. I was the primary investigator and author of this paper.

VITA

2006	B.Sc. in Computer Engineering, Islamic Azad University, Shiraz, Iran
2011	M.Sc. in Software Engineering, Shiraz University, Iran
2019	Ph.D. in Electrical Engineering (Intelligent Systems, Robotics, and Control), University of California San Diego, USA

PUBLICATIONS

Arya Iranmehr, Tsering Stobdan, Dan Zhou, Sergey Kryazhimskiy, Vineet Bafna, Gabriel G Haddad. "Long-term experimental reveals the involment of multiple evolutionary mechanisms in rapid adaptation." *In preperation*.

Michael Hicks, Claire YC Hou, **Arya Iranmehr**, Krisztina Marosi, Ewen Kirkness. Target discovery using biobanks and next-generation sequencing data. *Drug Discovery Today* (Under review)

Hamed Masnadi-Shirazi, **Arya Iranmehr**, Nuno Vasconcelos. "Cost-Sensitive Support Vector Machines." *Neurocomputing* (2019).

Ali Akbari, Joseph J Vitti, **Arya Iranmehr**, Mehrdad Bakhtiari, Pardis C Sabeti, Siavash Mirarab, and Vineet Bafna. "Identifying the favored mutation in a positive selective sweep." *Nature methods* (2018).

Arya Iranmehr, Ali Akbari, Christian Schlötterer, and Vineet Bafna. "Clear: composition of likelihoods for evolve and resequence experiments." *Genetics* (2017).

Arya Iranmehr, Tsering Stobdan, Dan Zhou, Orit Poulsen, Kingman P Strohl, Almaz Aldashev, Amalio Telenti, Emily HM Wong, Ewen F Kirkness, J Craig Venter, Vineet Bafna, Gabriel G Haddad. "Novel insight into the genetic basis of high-altitude pulmonary hypertension in Kyrgyz highlanders." *European Journal of Human Genetics* (2018).

ABSTRACT OF THE DISSERTATION

Analyzing Genetic Adaptation in Action: Identifying the Evolutionary Mechanisms Rescuing Stressed Populations

by

Arya Iranmehr

Doctor of Philosophy in Electrical Engineering (Intelligent Systems, Robotics, and Control)

University of California San Diego, 2019

Professor Vineet Bafna, Co-Chair
Professor Siavash Mirarab, Co-Chair

Genetic adaptation is central to shaping phenotype diversity among populations. If there was not any genetic adaptation, Homo sapiens was not able to migrate out of their original habitat, east of Africa, to colonize the planet. Interestingly, adaptation has enabled us to occupy a wider range of adverse environments such as, arctic, high-altitude, and highly pathogenic (e.g. areas with high rates of transmission of Malaria) regions. Other phenotypes such as skin pigmentation, size of stature, lactose intolerance and several disease susceptibility are directly linked to genetic adaptation.

Adaptation also play an important role in global burden of disease and mortality. One every three deaths worldwide is attributed to the evolution of large asexual cell populations. Adaptation provide pathogens to ability to persist to the immune system or exogenous drug to avoid recovery of the host. It also enables them to revive and relapse the disease after obtaining the drug resistance allele, that is the case for Cancer, HIV, Malaria and many other lethal disease. Moreover, some of the ethnic populations also have a significantly lower susceptibility to an specific disease, such as pulmonary hypertension, Malaria, cardiovascular disease, etc.

In all the cases, better understanding of mechanisms of adaptation and the genomic targets of the selection can provide actionable information. For instance, the tedious and expensive process of drug discovery can be facilitated by taking into account of disease susceptibility targets. Better therapy drugs can be made by targeting the adapted loci on the pathogen. Finally, mapping the targets of adaptation provides insights into cryptic biological processes.

While human biology is the center of attention, model organisms provide convenient, inexpensive, and salable (to populations) framework to test evolutionary hypotheses. This owes to the fact that the molecular mechanisms of evolution are predominantly similar between any living organism.

Here in this dissertation, I utilize experimental evolution of *D. melanogaster* to test multiple evolutionary hypotheses regarding the mechanisms, targets, modes and tempo of adaptation. To answer these questions, I develop genomic time-series models to describe data and find targets of selection. Using the evolutionary models I analyzed an ethnic population to find disease susceptibility genes.

Chapter 1

Introduction

All the living organisms die, and before they die, they need to pass the biological instructions to the next generation. The so called biological instructions provide the cookbook for making endogenous products such as proteins and enzymes for the biological processes such as oxygen metabolism. Life has managed to encode these information in very long molecules, DeoxyriboNucleic Acid (DNA). With a gross simplification, the DNA can be regarded as a long string of letters $\{$A,T,C,G$\}$. From an information theoretic prospective, each character can encode 2 bits of information and the DNA of length N, creates 4^N contingencies. This fundamental rules provides enough flexibility for creation and diversification of complicated forms of life.

To avoid extinction, living organisms reproduce more offsprings than what will survive to reproduce. Additionally, the process of reproduction is error-prune in the sense that the genetic material passed to the progenies is not identical to the parent(s). Specifically, errors in DNA replication creates phenotypic diversity at a population by introducing genetic variation. Finally, as the resources of environments are limited, the excess of reproduction and phenotype variability leads to the competition of the descendants to survive. These three principles are the basis pf Darwin's natural selection theorem in which the fittest (with respect to the environment) in the population has a higher chance of surviving and reproducing. Statistically, positive(negative)

selection can be regarded as over(under)-sampling of the selected genomes from one generation to the next generation. These biases in sampling creates distinct signatures when assessing genome of a population.

Genetic adaptation is one of the main evolutionary forces and its outcomes are nontrivial. In living organisms, it give rise to phenotypic diversity among populations and with reproductive isolation, it makes one of the speciation mechanisms [3]. If there was not any selection, forms of life would not survive and speciate during history of life, of which some eras were extremely hostile [4]. According to fossil, archaeological, and genetic data, Our own species, Homo sapiens, has originally originated in from sub-Saharan Africa[5–7] around 300Kya. Around 100Kya, they started to migrate out of Africa to colonize the planet [8]. Adaptation has enabled us to occupy a wide range of adverse environments such as, arctic, high-altitude [9, 10], and highly pathogenic [11] (e.g. areas with high rates of transmission of Malaria) regions. Other phenotypes such as skin pigmentation [12], size of stature [13], lactase persistence [14] and several disease susceptibility [2, 15, 16] are directly linked to genetic adaptation.

Adaptation also play an important role in global burden of disease, mortality, morbidity and disability-adjusted life years (DALYs) [17]. One every three deaths worldwide is attributed to the evolution of large asexual cell populations [18]. Adaptation provide pathogens to ability to persist to the immune system or exogenous drug to avoid recovery of the host. It also enables them to revive and relapse the disease after obtaining the drug resistance allele, that is the case for Cancer [19], HIV [20], Malaria [21] and many other lethal disease [22]. Moreover, some of the ethnic populations also have a significantly higher susceptibility to an specific disease, such as pulmonary hypertension, Malaria, cardiovascular disease, etc.

In all the cases, better understanding of mechanisms of adaptation and the genomic targets of the selection can provide actionable information. For instance, the tedious and expensive process of drug discovery can be facilitated by taking into account of disease susceptibility targets. Better therapy drugs can be made by targeting the adapted loci on the pathogen. Finally, mapping

the targets of adaptation provides insights into cryptic biological processes.

More importantly, the phenotypic variability can be attributed to two main contributing factor: genetic and environmental [23]. In general,contribution of each factor is not identifiable. This is because of the fact that only phenotypic variance can be measured easily. The genetic variance breaks down to additive, dominance, and (interaction between genetic loci) components. Also, In theory, the environment can interact with each of the genetic sub-components. As a result, distinguishing contribution of different components into phenotypic variance is a very difficult problem. A trick to solve this problem, is to enforce some of these component to be zero, and solve for the rest. For instance, identical twin studies [24], minimize the genetic variance to estimate the environmental variance, and subsequently estimate the genetic Heritability for the rest (non-twin) of the population. Similar, strategy can be used for model organisms to use inbred lines [23, 25].

Even with the adequate adjustment for external factors, human studies are subject to unobserved confounders or environmental factors such as socio-economic class that can not be cancelled out [26]. As a result, model organisms provide a less confounded, convenient, inexpensive, and salable (to populations) framework to test evolutionary hypotheses. This owes to the fact that the molecular mechanisms of evolution are predominantly similar between any living organism. To this end, *D. melanogaster* is known to be one of the most widely used model organisms for scientific and exploratory studies [23]. Moreover, short generation-time allows scientist to perform longitudinal studies on fly populations and perform *experimental evolution*.

Experimental evolution refers to the study of the evolutionary processes of a model organism in a controlled [27–33] or natural [34–40] environment. Recent advances in whole genome sequencing have enabled us to sequence populations at a reasonable cost, even for large genomes. Perhaps more important for experimental evolution studies, we can now evolve and resequence (E&R) multiple replicates of a population to obtain *longitudinal time-series data*, in order to investigate the dynamics of evolution at molecular level. Although constraints such

as small sizes, limited timescales, and oversimplified laboratory environments may limit the interpretation of E&R results, these studies are increasingly being used to test a wide range of hypotheses [41] and have been shown to be more predictive than static data analysis [42–44]. In particular, longitudinal E&R data is being used to estimate model parameters including population size [45–50], strength of selection [49, 51–56], allele age [55] recombination rate [49], mutation rate [49, 57], quantitative trait loci [58] and for tests of neutrality hypotheses [40, 49, 59, 60].

Here in this dissertation, I utilize experimental evolution of *D. melanogaster* to test multiple evolutionary hypotheses regarding the mechanisms, targets, modes and tempo of adaptation. To answer these questions, I develop genomic time-series models to describe data and find targets of selection. Using the evolutionary models I analyzed an ethnic human population to find disease susceptibility genes.

1.1 Background

Evolutionary mechanisms are predominantly similar across organisms. Here, I provide an overview on some of the evolutionary forces and phenomena.

Wright-Fisher model. Throughout this manuscript, we consider the Wright-Fisher model [61] that is characterized by random mating, non-overlapping generations, equal proportion and fitness of genders, and infinite site assumption (no recurrent mutation). Wright-Fisher model, that provide a tractable evolutionary model for a evolving population of size N. In particular, it models the process of reproduction as if genomes are being binomially sampled with replacement from the existing pool genomes.

Genetic Drift. Under this mode, genetic drift is fully characterized by the population size parameter (N). Specifically, allele frequencies change from a generation to the next generation according to

$$2N\nu_{t+1} \sim \text{Binomial}(2N, \nu_t). \tag{1.1}$$

Mutation. Without mutation, there is no genetic variation and all the genomes in the population are identical, aka clonal. Mutation is a major source of variation. For simplicity, I only take into account of germline mutations, in which passed to offsprings. Mutations accumulate for an on the genomes and the abundance of each mutation is being referred as frequency (of the non-reference allele). When observing genetic drift and mutation process jointly, it is become clear how genetic drift effects allele frequency. Due to sampling, carriers of a mutation can be over/under-sampled, leading to increase/decrease of the frequency. Each mutation has a fitness that is determined by the locus, substituted allele, sequence context, environment, and physiology.

Linkage Disequilibrium. All the co-occuring mutations on a chromosome are passed from parents to offspring. This law of inheritance creates non-random association between alleles, linkage disequilibrium (LD). LD creates correlations between co-occurring mutations in genetic data. For instance, two fully linked mutations have identical allele-frequency trajectories.

Recombination. Genetic recombination is another evolutionary force to create genetic diversity. Sexual organisms, pass *a* combination for their genomes to their offspring. This can decouple two co-occuring mutation on a genome and as a result reduces the LD between any two mutations. Roughly, the number of recombinations that is function of distance and number of generations, is inversely related to LD.

Negative Selection. Life has originated more than 4.5 *Bya*, and given genetic drift and mutation mechanisms and large population sizes, we should be able to observe variation at each or many loci. However, when we observe a population we see a depletion of mutation, especially is some of the functional regions. One of the explanations for this observations is that some of the mutations have negative fitness and their carriers are not viable to live and reproduce, i.e., being negatively selected. In other words, we only observe the survivors of the mutation process. In a broader sense, *mutation selection balance* [62, 63] provides an an equilibrium such that the number (or frequency) of deleterious alleles in a population are kept limited.

Positive Selection. When the fitness of mutation is positive, the carriers are selected with a higher chance. As a result, the frequency of the beneficial mutations can be high or they can be fixed in the population. While only sign of the fitness is different than negative selection, the consequences are tremendous. A beneficial allele is rapidly propagate in the population along with its linked loci. This process creates distinct footprints on the population genomes data.

Demographic Change. Along with selection pressure, demographic change makes the major external force that can influence genetic variation. While there are many forms of demographic events such as admixture, migration etc., in this dissertation I only focus on changes in population size, i.e., bottleneck and recovery. The population size is one of the important parameters of introducing variation by mutation, genetic drift. Additionally, the genetic make up of the population can also be determined by the survivors (founders) of a bottleneck. Hence, when analyzing genomic data fro selection, it is important to adjust for the most powerful confounder in selection studies, demographic change.

1.2 Dissertation overview

As experimental evolution and inexpensive sequencing technologies enable us to observe genotype and phenotype in time, a new type of data is being created: time-series genomic-phenomic data. Inferring selection and demographic based on variation data, single sample in time, has been thoroughly studied in the literature [64, 65]. However, the time-series methods for sexual model organisms is understudied. In this theses, I focus on the experiments in which *D. melanogaster* is evolved and sequenced for up to 200 generations under a selection pressure.

In chapter 2, I start by outlining the probabilistic generative model, Wright-Fisher Markov chain, that basis of my predictive models. Then, I test my model for genetic drift only and drift with selection to show that the probabilistic model can distinguish selection and neutral evolution. Next, I boost the model by taking into account of linked-loci to the target allele, by computing

the composite likelihood for a region. Finally, I derive the likelihood model for estimating the population size.

Chapter 3 builds makes models to make higher level inference based upon models of chapter 2. First, I show how the estimates of population size in time can be indicator of demographic history, e.g. bottleneck or recovery. In real data, I show that the so called "fixed-population size" assumption is overly simplistic and collapses when the selection pressure is strong. Then, I develop models to infer the mode of adaptation, i.e., hard or soft sweep (new mutation or standing variation). Next, I aim to find epistatic interactions in time-series data. I characterized epistasis as change of fitness of a mutation in a fixed environment. I develop models to identify such signatures and test it on real data.

Chapter 4 is devoted to analysis of ethnic human populations for disease susceptibility genes. I develop a robust pipeline for quality control and statistical analysis of such populations to discover targets of selection.

Chapter 2

Identifying Selection in Time-Series Data

The advent of next generation sequencing technologies has made whole-genome and whole-population sampling possible, even for eukaryotes with large genomes. With this development, experimental evolution studies can be designed to observe molecular evolution "in-action" via Evolve-and-Resequence (E&R) experiments. Among other applications, E&R studies can be used to locate the genes and variants responsible for genetic adaptation. Most of existing literature on time-series data analysis often assumes large population size, accurate allele frequency estimates, or wide time spans. These assumptions do not hold in many E&R studies.

In this article, we propose a method–Composition of Likelihoods for Evolve-And-Resequence experiments (CLEAR)–to identify signatures of selection in small population E&R experiments. CLEAR takes whole-genome sequence of pool of individuals (pool-seq) as input, and properly addresses heterogeneous ascertainment bias resulting from uneven coverage. CLEAR also provides unbiased estimates of model parameters, including population size, selection strength and dominance, while being computationally efficient. Extensive simulations show that CLEAR achieves higher power in detecting and localizing selection over a wide range of parameters, and is robust to variation of coverage. We applied CLEAR statistic to multiple E&R experiments, including, data from a study of *D. melanogaster* adaptation to alternating

temperatures and a study of outcrossing yeast populations, and identified multiple regions under selection with genome-wide significance.

2.1 Introduction

Natural selection is a key force in evolution, and a mechanism by which populations can adapt to external 'selection' pressure. Examples of adaptation abound in the natural world [66], including for example, classic examples like lactose tolerance in Northern Europeans [67], human adaptation to high altitudes [9, 68], but also drug resistance in pests [69], HIV [70], cancer [71, 72], malarial parasite [73, 74], and others [75]. In these examples, understanding the genetic basis of adaptation can provide valuable information, underscoring the importance of the problem.

Experimental evolution refers to the study of the evolutionary processes of a model organism in a controlled [27–33] or natural [34–40] environment. Recent advances in whole genome sequencing have enabled us to sequence populations at a reasonable cost, even for large genomes. Perhaps more important for experimental evolution studies, we can now evolve and resequence (E&R) multiple replicates of a population to obtain *longitudinal time-series data*, in order to investigate the dynamics of evolution at molecular level. Although constraints such as small sizes, limited timescales, and oversimplified laboratory environments may limit the interpretation of E&R results, these studies are increasingly being used to test a wide range of hypotheses [41] and have been shown to be more predictive than static data analysis [42–44]. In particular, longitudinal E&R data is being used to estimate model parameters including population size [45–50], strength of selection [49, 51–56], allele age [55] recombination rate [49], mutation rate [49, 57], quantitative trait loci [58] and for tests of neutrality hypotheses [40, 49, 59, 60].

While many E&R study designs are being used [57, 76], we restrict our attention to the adaptive evolution due to standing variation in fixed size populations. This regime has been considered earlier, typically with *D. melanogaster* as the model organism of choice, to identify adaptive

genes in longevity and aging [60, 77] (600 generations), courtship song [78] (100 generations), hypoxia tolerance [79] (200 generations), adaptation to new laboratory environments [29, 80] (59 generations), egg size [81] (40 generations), C virus resistance [82] (20 generations), and dark-fly [83] (49 generations).

The task of identifying selection signatures can be addressed at different levels of specificity. At the coarsest level, identification could simply refer to deciding whether some genomic region (or a gene) is under selection or not. In the following, we refer to this task as *detection*. In contrast, the task of *site-identification* corresponds to the process of finding the favored mutation/allele at nucleotide level. Finally, *estimation of model parameters*, such as strength of selection and dominance at the site, can provide a comprehensive description of the selection process.

In the effort to analyze E&R selection experiments, many authors chose to adapt existing tests that were originally used for static data, pairwise comparisons (two time-points) and single replicates to perform a null scan. For instance, Zhu *et al.* [79] used the ratio of the estimated population size of case and control populations to compute test statistic for each genomic region. Burke *et al.* [60] applied Fisher exact test to the last observation of data on case and control populations. Orozco-terWengel *et al.* [29] used the Cochran-Mantel-Haenszel (CMH) test [84] to detect SNPs whose read counts change consistently across all replicates of two time-point data. Turner *et al.* [78] proposed the diffStat statistic to test whether the change in allele frequencies of two populations deviate from the distribution of change in allele frequencies of two drifting populations. Bergland *et al.* [40] calculated F_{st} to populations throughout time to signify their differentiation from ancestral (two time-point data) as well as geographically different populations. Jha *et al.* [81] computed test statistic of generalized linear-mixed model directly from read counts.

Alternatively, *direct* methods have been developed to analyze time-series data by taking a likelihood approach, and estimating population genetics parameters. Bollback *et al.* [53] proposed a Hidden Markov Model (HMM) to estimate the selection coefficient s and population size by

using a diffusion approximation to the Wright Fisher process. Steinrücken *et al.* [56] proposed a general diploid selection model which takes into account of dominance of the favored allele and approximates likelihood analytically. Recently, Schraiber *et al.* [85] proposed a Bayesian framework to estimate parameters using Monte Carlo Markov chain sampling. Mathieson and McVean [51] adopted HMMs to structured populations and estimated parameters using an Expectation Maximization (EM) procedure on discretized allele frequency. Feder *et al.* [59] modeled increments in allele frequency with a Brownian motion process, proposed the Frequency Increment Test (FIT). More recently, Topa *et al.* [86] proposed a Gaussian Process (GP) for modeling single-locus time-series pool-seq data. Terhorst *et al.* [49] extended GP to compute joint likelihood of multiple loci under null and alternative hypotheses. Finally, Levy *et al.* [18] proposed a Bayesian model to handle sequencing, amplification and growth noise in a large population of barcoded lineages.

Among the methods specifically designed for time-series data, many make assumptions which may not hold in E&R studies. One common assumption is that the underlying population size is large, so it is reasonable to model dynamics of allele frequencies using continuous state models [49, 53, 59]. Second, many existing methods were originally designed to process wider time spans seen in ancient DNA studies, an assumption that does not hold for E&R experiments [56, 85]. Finally, many E&R analysis tools assume that allele frequencies in the input data are unbiased (e.g. [53]), which may not be valid for shotgun sequencing experiments.

Here, we consider a Hidden Markov Model (HMM), similar to Williamson *et al.* [45] and Bollback *et al.*'s [53] but under a "small-population-size" regime. Specifically, we use a discrete state (frequency) model. We show that for small population sizes, discrete models can compute likelihood exactly, which improves statistical performance, especially for short time-span experiments. Additionally, we add another level of sampling-noise to the traditional HMM model, allowing for heterogeneous ascertainment bias due to uneven coverage among variants. We show that for a wide range of parameters, CLEAR provides higher power for detecting selection,

estimates model parameters consistently, and localizes favored allele more accurately compared to the state-of-the-art methods, while being computationally efficient.

2.2 Materials and Methods

Consider a panmictic diploid population with fixed size of N individuals. Let $\nu = \{\nu_t\}_{t \in T}$ be frequencies of the derived allele at generations $t \in T$ for a given variant, where at generations $T = \{\tau_i : 0 \leq \tau_0 < \tau_1 \ldots < \tau_T\}$ samples of n individuals are chosen for pooled sequencing. The experiment is replicated R times. We denote allele frequencies of the R replicates by the set $\{\nu\}_R$. To identify the genes and variants that are responding to selection pressure, we use the following procedure:

1. **Estimating population size.** The procedure starts by estimating the effective population size, $\widehat{N}$, under the assumption that much of the genome is evolving neutrally.

2. **Estimating selection parameters.** For each polymorphic site, selection and dominance parameters s, h are estimated so as to maximize the likelihood of the time series data, given $\widehat{N}$.

3. **Computing likelihood statistics.** For each variant, a log-odds ratio of the likelihood of selection model ($s > 0$) to the likelihood of neutral evolution/drift model is computed. Likelihood ratios in a genomic region are combined to compute the CLEAR statistic for the region.

4. **Hypothesis testing.** An empirical null distribution of the CLEAR statistic is calculated using genome-wide drift simulations, and used to compute p-values and thresholds for a specified FDR. We perform single locus hypothesis testing within selected regions to identify significant variants and report genes that intersect with the selected variants.

These steps are described in detail below.

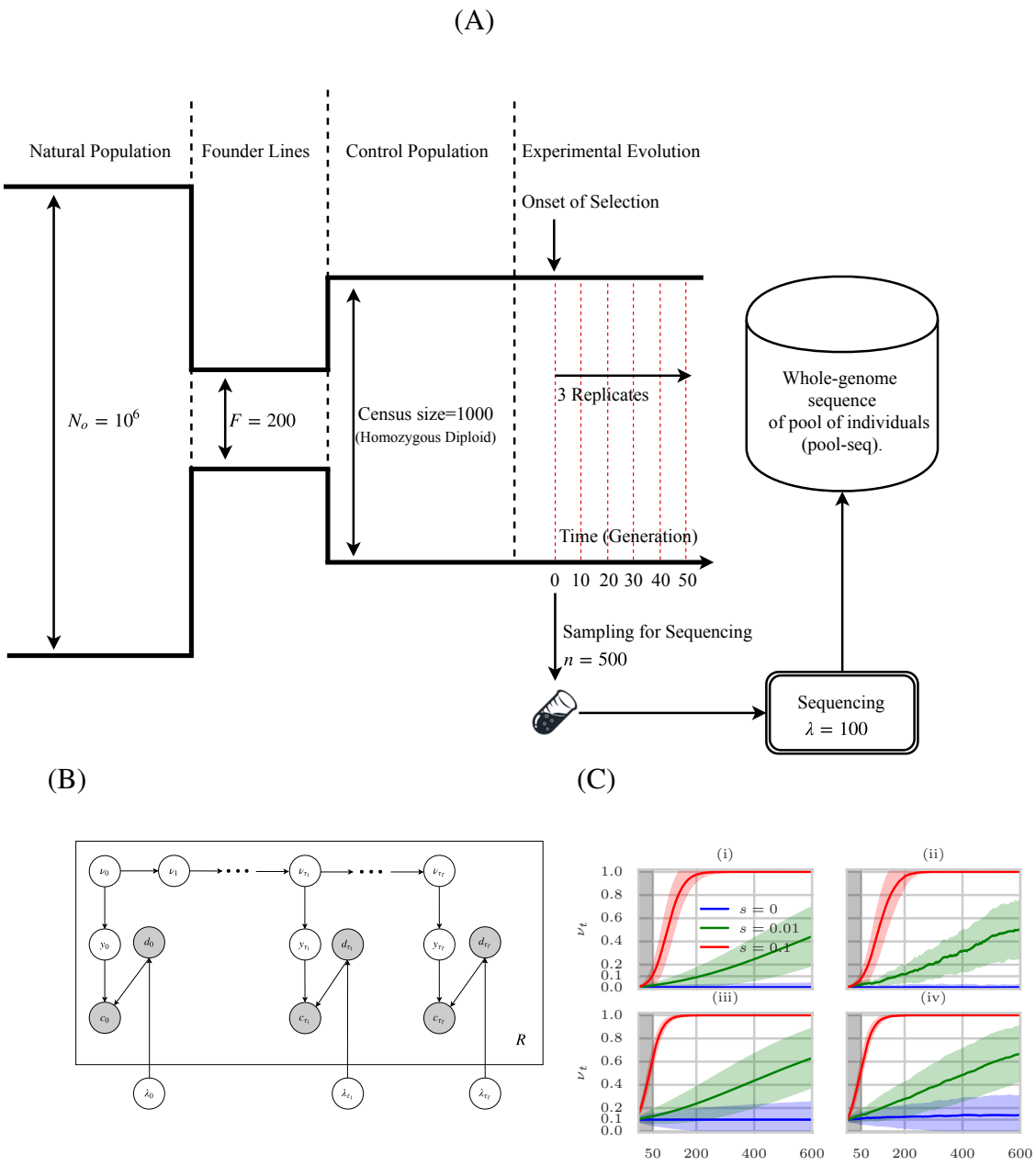

Figure 2.1: **Evolve and Resequence Selection Experiments on *D. melanogaster* .** (A) Typical configuration in which time-series data is collected for *D. melanogaster* . A small set of founder lines ($F = 200$) is selected from a large population ($N_o = 10^6$), and used to create a sub-population of isofemale lines. Multiple replicates of the population are evolved and resequenced to collect time-series genomic data. For sequencing, n individuals are randomly sampled and sequenced with coverage λ. (B) Graphical model showing dependence of the random variables in the single-locus model used to compute CLEAR statistics. Observed variables, c (derived allele read count) and d (total read count) are shaded. The variables ν, y, λ denote allele frequency, sampled allele frequency, and mean sequencing coverage, respectively. (C) Mean and 95% confidence interval of the theoretical (i,iii) and empirical (ii,iv) trajectories of the favored allele for hard (i,ii) and soft (iii,iv) sweep scenarios and $N = 1000$. The first 50 generations are shaded in gray to represent the sampling span of sampling in short-term experiments, illustrating the difficulty in predicting selection at early stages of selective sweep.

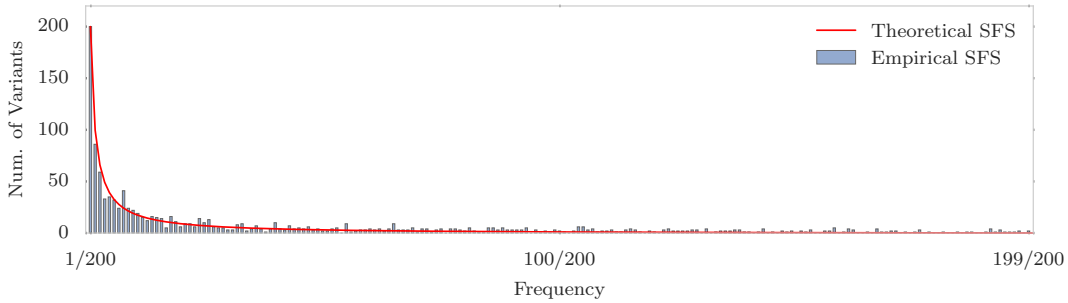

Figure 2.2: **Site Frequency Spectrum.**
Theoretical and Empirical SFS in a 50Kbp region for a neutral population of 200 individuals when $N_e = 10^6$ and $\mu = 10^{-9}$. The x-axis corresponds to site frequency, and the y-axis to the number of variants with a specific frequency. In a neural population, majority of the variations stand in low frequency.

2.2.1 Estimating Population Size

Methods for estimating population sizes from temporal neutral evolution data have been developed [45, 49, 50, 53, 87]. Here, we aim to extend these models to explicitly model the sampling noise that arise in pool-seq data. Specifically, we model the variation in sequence coverage over different locations, and the noise due to sequencing only a subset of the individuals in the population. In addition, many existing methods [49, 53, 59, 86] are designed for large populations, and model frequency as a continuous quantity. We observed that using Brownian motion to model frequency drift may be inadequate for small populations, low starting frequencies and sparse sampling (in time), factors that are common in experimental evolution (see Results, 2.4A-C, and 2.3). To this end, we model the Wright-Fisher Markov process for generating pool-seq data (2.11) via a discrete HMM (2.1-B). We start by computing a likelihood function for the population size given neutral pool-seq data.

Likelihood for Neutral Model. We model the allele frequency counts $2N\nu_t$ as being sampled from a Binomial distribution. Specifically,

$$\begin{aligned} \nu_0 &\sim \pi, \\ 2N\nu_t|\nu_{t-1} &\sim \text{Binomial}(2N, \nu_{t-1}) \end{aligned}$$

where π is the global distribution of allele frequencies in the base population. Note that π depends on the demographic history of the founder lines and can be estimated from site frequency spectrum(see 2.2) of the initial population. For notational convenience, henceforth we omit the dependence of likelihoods to the parameter π.

To estimate frequency after τ transitions, it is enough to specify the $2N \times 2N$ transition matrix $P^{(\tau)}$, where $P^{(\tau)}[i,j]$ denotes probability of change in allele frequency from $i/2N$ to $j/2N$ in τ generations:

$$\begin{aligned} P^{(1)}[i,j] &= \Pr\left(\nu_{t+1} = \frac{j}{2N} \mid \nu_t = \frac{i}{2N}\right) \\ &= \binom{2N}{j} s\nu_t^j (1-\nu_t)^{2N-j}, \end{aligned} \tag{2.1}$$

$$P^{(\tau)} = P^{(\tau-1)} P^{(1)} \tag{2.2}$$

Furthermore, in an E&R experiment, $n \leq N$ individuals are randomly selected for sequencing. The sampled allele frequencies, $\{y_t\}_{t\in T}$, are also Binomially distributed

$$2ny_t \sim \text{Binomial}(2n, \nu_t) \tag{2.3}$$

We introduce the $2N \times 2n$ sampling matrix Y, where $Y[i,j]$ stores the probability that the sample allele frequency is $j/2n$ given that the true allele frequency is $i/2N$.

We denote the pool-seq data for that variant as $\{x_t = \langle c_t, d_t \rangle\}_{t\in T}$ where d_t, c_t represent the coverage, and the read count of the derived allele, respectively. Let $\{\lambda_t\}_{t\in T}$ be the sequencing coverage at different generations. Then, the observed data are sampled according to

$$d_t \sim \text{Poisson}(\lambda_t), \qquad\qquad c_t \sim \text{Binomial}(d_t, y_t) \tag{2.4}$$

The emission probability for a observed tuple $x_t = \langle d_t, c_t \rangle$ is

$$\mathbf{e}_i(x_t) = \binom{d_t}{c_t} \left(\frac{i}{2n}\right)^{c_t} \left(1 - \frac{i}{2n}\right)^{d_t - c_t}. \tag{2.5}$$

For $1 \leq t \leq T, 1 \leq j \leq 2N$, let $\alpha_{t,j}$ denote the probability of emitting $x_1, x_2, \ldots, x_t$ and reaching state j at τ_t. Then, α_t can be computed using the forward-procedure [88]:

$$\alpha_t^T = \alpha_{t-1}^T P^{(\delta_t)} \text{diag}(Y\mathbf{e}(x_t)) \tag{2.6}$$

where $\delta_t = \tau_t - \tau_{t-1}$. The joint likelihood of the observed data from R independent observations is given by

$$\begin{aligned} L(N|\{x\}_R, n) &= \prod_{r=1}^{R} L(N|x^{(r)}, n) = \Pr(\{x\}_R|N, n) \\ &= \prod_{r=1}^{R} \sum_i \alpha_{T,i}^{(r)} \end{aligned} \tag{2.7}$$

where $x = \{x_t\}_{t \in T}$. The graphical model and the generative process for which data is being generated is depicted in 2.1-B and 2.11, respectively.

Finally, the last step is to compute an estimate $\widehat{N}$ that maximizes the likelihood of all M variants in whole genome. Let $x_i^{(r)}$ denote the time-series data of the i-th variant in replicate r. Then,

$$\widehat{N} = \arg\max_N \prod_{i=1}^{M} \prod_{r=1}^{R} L(N|x_i^{(r)}) \tag{2.8}$$

2.2.2 Estimating Selection Parameters

Likelihood for Selection Model. Assume that the site is evolving under selection constraints $s \in \mathbb{R}$, $h \in \mathbb{R}_+$, where s and h denote selection strength and dominance parameters , respectively. By definition, the relative fitness values of genotypes $0|0$, $0|1$ and $1|1$ are given by $w_{00} = 1$, $w_{01} = 1 + hs$ and $w_{11} = 1 + s$. Then, , the frequency at time $\tau_t + 1$ (one generation

ahead), can be estimated using:

$$\begin{aligned}\hat{\nu}_{t^+} = \mathbb{E}[|s,h,\nu_t] &= \frac{w_{11}\nu_t^2 + w_{01}\nu_t(1-\nu_t)}{w_{11}\nu_t^2 + 2w_{01}\nu_t(1-\nu_t) + w_{00}(1-\nu_t)^2} \\ &= \nu_t + \frac{s(h+(1-2h)\nu_t)\nu_t(1-\nu_t)}{1+s\nu_t(2h+(1-2h)\nu_t))}.\end{aligned} \tag{2.9}$$

The machinery for computing likelihood of the selection parameters is identical to that of population size, except for transition matrices. Hence, here we only describe the definition transition matrix $Q_{s,h}$ of the selection model. Let $Q_{s,h}^{(\tau)}[i,j]$ denote the probability of transition from $i/2N$ to $j/2N$ in τ generations, then (See [89], Pg. 24, Eqn. 1.58-1.59):

$$\begin{aligned}Q_{s,h}^{(1)}[i,j] &= \Pr\left(= \frac{j}{2N} \,\middle|\, \nu_t = \frac{i}{2N}; s,h,N\right) \\ &= \binom{2N}{j} \hat{\nu}_{t^+}^{j} (1-\hat{\nu}_{t^+})^{2N-j}\end{aligned} \tag{2.10}$$

$$Q_{s,h}^{(\tau)} = Q_{s,h}^{(\tau-1)} Q_{s,h}^{(1)} \tag{2.11}$$

The maximum likelihood estimates are given by

$$\widehat{s},\widehat{h} = \arg\max_{s,h} \prod_{r=1}^{R} L(s,h|x^{(r)},\widehat{N}) \tag{2.12}$$

Using grid search, we first estimate N (Eq. 2.8), and subsequently, we estimate parameters s,h (Eq. 2.12, 2.12). By broadcasting and vectorizing the grid search operations across all variants, the genome scan on millions of polymorphisms can be done in significantly smaller time than iterating a numerical optimization routine for each variant(see Results and 2.7).

2.2.3 Empirical Likelihood Ratio Statistics

The likelihood ratio statistic for testing directional selection, to be computed for each variant, is given by

$$H = -2\log\left(\frac{L(\bar{s},0.5|\{x\}_R,\widehat{N})}{L(0,0.5|\{x\}_R,\widehat{N})}\right), \tag{2.13}$$

where $\bar{s} = \arg\max_s \prod_{r=1}^{R} L(s,0.5|x^{(r)},\widehat{N})$. Similarly we can define a test statistic for testing if selection is dominant by

$$D = -2\log\left(\frac{L(\widehat{s},\widehat{h}|\{x\}_R,\widehat{N})}{L(\bar{s},0.5|\{x\}_R,\widehat{N})}\right). \tag{2.14}$$

While extending the single-locus WF model to a multiple linked-loci can improve the power of the model [49], it is computationally and statistically expensive to compute exact likelihood. In addition, computing linked-loci joint likelihood requires haplotype resolved data, which pool-seq does not provide. Here, similar to Nielsen *et al* [90], we calculate *composite likelihood ratio* score for a genomic region.

$$H = \frac{1}{|L|}\sum_{\ell\in L} H_\ell. \tag{2.15}$$

where L is a collection of segregating sites and H_ℓ is the likelihood ratio score based for each variant ℓ in L. The optimal value of the hyper-parameter L depends upon a number of factors, including initial frequency of the favored allele, recombination rates, linkage of the favored allele to neighboring variants, population size, coverage, and time since the onset of selection (duration of the experiment). In 2.5, we provide a heuristic to compute a reasonable value of L, based on experimental data.

We work with a normalized value of H, given by

$$H_i^* = \frac{H_i - \mu_C}{\sigma_C}, \qquad \forall i \in C, \tag{2.16}$$

where μ_C and σ_C are the mean and standard deviation of H values in a large region C. We found different chromosomes to have different distribution of H_i values, and therefore decided to use single chromosomes as C.

2.2.4 Hypothesis Testing

Single-Locus tests. Under neutrality, Log-likelihood ratios can be approximated by X^2 distribution [91], and p-values can be computed directly. However, Feder *et al.* [59] showed that when the number of independent samples (replicates) is small, X^2 is a crude approximation to the true null distribution and results in more false positive. Following their suggestion, we first compute the empirical null distribution using simulations with the estimated population size (See 2.11). The empirical null distribution of statistic H is used to compute p-values as the fraction of null values that exceed the test score. Finally, we use Storey and Tibshirani's method [92] to control for False Discovery Rate in multiple testing.

Composite likelihood tests.

Similar to single-locus tests, we compute the null distribution of the H^* statistic using whole-genome simulations with the estimated population size, and subsequently compute FDR. The simulations for generating the null distribution of H^* are described next.

2.2.5 Simulations

We use the same simulation procedure for two purposes. First, we use them to test the power of CLEAR against other methods in small genomic windows. Second, we use the simulations to generate the distribution of null values for the statistic to compute empirical p-values. We mainly chose parameters that are relevant to *D. melanogaster* experimental evolution [93]. See also 2.1-A for illustration.

1. **Creating initial founder line haplotypes.** Using `msms` [94], we created neutral populations for F founding haplotypes with command `$./msms <F> 1 -t <2`$\mu W N_o$`> -r <2`rWN_o`> <W>`, where $F = 200$ is number of lines, $N_o = 10^6$ is effective founder population size, $r = 2 \times 10^{-8}$ is recombination rate, $\mu = 2 \times 10^{-9}$ is mutation rate. The window size W is used to compute $\theta = 2\mu N_o W$ and $\rho = 2N_o rW$. We chose $W = 50$Kbp for simulating individual windows for performance evaluations, and $W = 20$Mbp for simulating *D. melanogaster* chromosomes for p-value computations.

2. **Creating initial diploid population.** An initial set of $F = 200$ haplotypes was created from step I, and duplicated to create F homozygous diploid individuals to simulate generation of inbred lines. N diploid individuals were generated by sampling with replacement from the F individuals.

3. **Forward Simulation.** We used forward simulations for evolving populations under selection. We also consider selection regimes which the favored allele is chosen from standing variation (not *de novo* mutations). Given initial diploid population, position of the site under selection, selection strength s, number of replicates $R = 3$, recombination rate $r = 2 \times 10^{-8}$ and sampling times $T = \{0, 10, 20, 30, 40, 50\}$, `simuPop` [95] was used to perform forward simulation and compute allele frequencies for all of the R replicates. For hard sweep (respectively, soft sweep) simulations we randomly chose a site with initial frequency of $\nu_0 = 0.005$ (respectively, $\nu_0 = 0.1$) to be the favored allele. For generating the null distribution with drift for p-value computations, we used this procedure with $s = 0$.

4. **Sequencing Simulation.** Given allele frequency trajectories we sampled depth of each site in each replicate identically and independently from Poisson(λ), where $\lambda \in \{30, 100, 300\}$ is the coverage for the experiment. Once depth d is drawn for the site with frequency ν, the number of reads c carrying the derived allele are sampled according to Binomial(d, ν). For experiments with finite depth the tuple $\langle c, d \rangle$ is the input data for each site.

2.3 Results

Modeling Allele Frequency Trajectories in Small Populations. We first tested the goodness of fit of the discrete versus Brownian motion (a continuous-state model) in modeling allele frequency trajectories, under general E&R parameters. For this purpose, we conducted 100K simulations with two time samples $T = \{0, \tau\}$ where $\tau \in \{1, 10, 100\}$ is the parameter controlling the density of sampling in time. In addition, we repeated simulations for different values of starting frequency $\nu_0 \in \{0.005, 0.1\}$ (i.e., hard and soft sweep) and selection strength $s \in \{0, 0.1\}$ (i.e., neutral and selection). Then, given initial frequency ν_0, we computed the expected distribution of the frequency of the next sample ν_τ under two models to make a comparison. 2.3A-F shows that Brownian motion (continuous model) is inadequate when ν_0 is far from 0.5, or when sampling times are sparse ($\tau > 1$). If the favored allele arises from standing variation in a neutral population, it is unlikely to have frequency close to 0.5, and the starting frequencies are usually much smaller (see 2.2). Moreover, in typical *D. melanogaster* experiments for example, sampling is sparse. Often, the experiment is designed so that $10 \leq \tau \leq 100$ [29, 79, 80, 93].

In contrast to the Brownian motion approximation, discrete Markov chain predictions (Eq. 2.11) are highly consistent with empirical data for a wide range of simulation parameters (2.3A-M). Moreover, the discrete markov chain can be modified to model the case when the the allele is under selection.

Detection Power. We compared the performance of CLEAR against other methods for detecting selection. For each method we calculated detection power as the percentage of true-positives identified with false-positive rate ≤ 0.05. For each configuration (specified with values for selection coefficient s, starting allele frequency ν_0 and coverage λ), power of each method is evaluated over 2000 distinct simulations, half of which modeled neutral evolution and the rest modeled positive selection.

We compared the power of CLEAR with Gaussian process (GP) [49], FIT [59], and

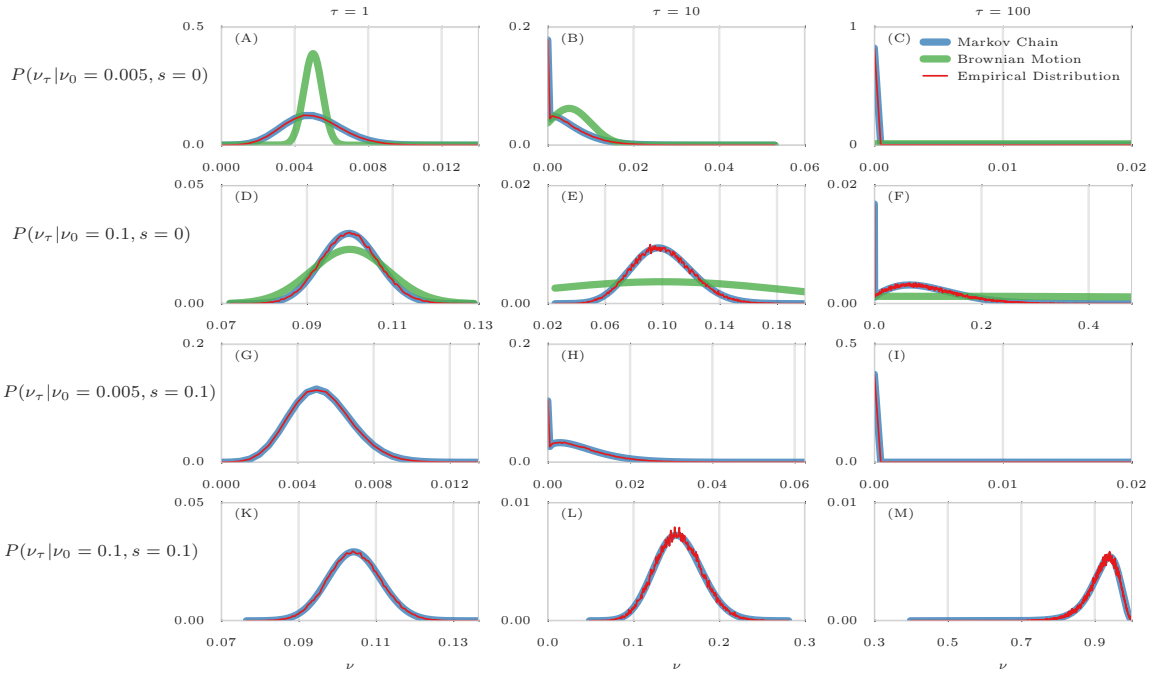

Figure 2.3: **Comparison of empirical distributions of allele frequencies (red) versus predictions from Brownian Motion (green), and Markov chain (blue).**
Comparison of empirical and theoretical distributions under neutral evolution (panels A-F) and selection (panels G-M) with different starting frequencies $\nu_0 \in \{0.005, 0.1\}$ and sampling times of $T = \{0, \tau\}$, where $\tau \in \{1, 10, 100\}$ and $N = 1000$. For each panel, the empirical distribution was computed over 100,000 simulations. Brownian motion (Gaussian approximation) provides poor approximations when initial frequency is far from 0.5 (A) or sampling is sparse (B,C,E,F). In addition, Brownian motion can only provide approximations under neutral evolution. In contrast, Markov chain consistently provides a good approximation in all cases.

CMH [84] statistics. FIT and GP convert read counts to allele frequencies prior to computing the test statistic. CLEAR shows the highest power in all cases and the power stays relatively high even for low coverage (2.4 and 2.1). In particular, the difference in performance of CLEAR with other methods is pronounced when starting frequency is low. The advantage of CLEAR stems from the fact that favored allele with low starting frequency might be missed by low coverage sequencing. In this case, incorporating the signal from linked sites becomes increasingly important. We note that methods using only two time points, such as CMH, do relatively well for high selection values and high coverage. However, the use of time-series data can increase detection power in low coverage experiments or when starting frequency is low. Moreover, time-series data provide means for estimating selection parameters s, h (see below). Finally, as CLEAR is robust to change of coverage, our results (2.4B,C) suggest that taking many samples with lower coverage is preferable to sparse sampling with higher coverage. For comparison purposes, we also tested CLEAR using the single locus statistic ($L = 1$). For the most part, CLEAR showed an improvement over other methods even with $L = 1$, or showed similar performance. The performance improved with higher L.

Site-identification. In general, localizing the favored variant, using pool-seq data is a nontrivial task due to extensive linkage disequilibrium [96]. To measure performance, we sorted variants by their H scores and computed rank of the favored allele for each method. For each setting of ν_0 and s, we conducted 1000 simulations and computed the rank of the favored mutation in each simulation. The cumulative distribution of the rank of the favored allele in 1000 simulation for each setting (2.5) shows that CLEAR outperforms other statistics.

An interesting observation is revisiting the contrast between site-identification and detection [96, 97]. When selection strength is high, detection is easier (2.4A-F), but site-identification is harder, due to the high LD between flanking variants and the favored allele (2.5A-F). Moreover, site-identification becomes more difficult whenever the initial frequency of the favored allele is low, i.e., at the onset of selection, LD between favored allele and its nearby variants is high. For

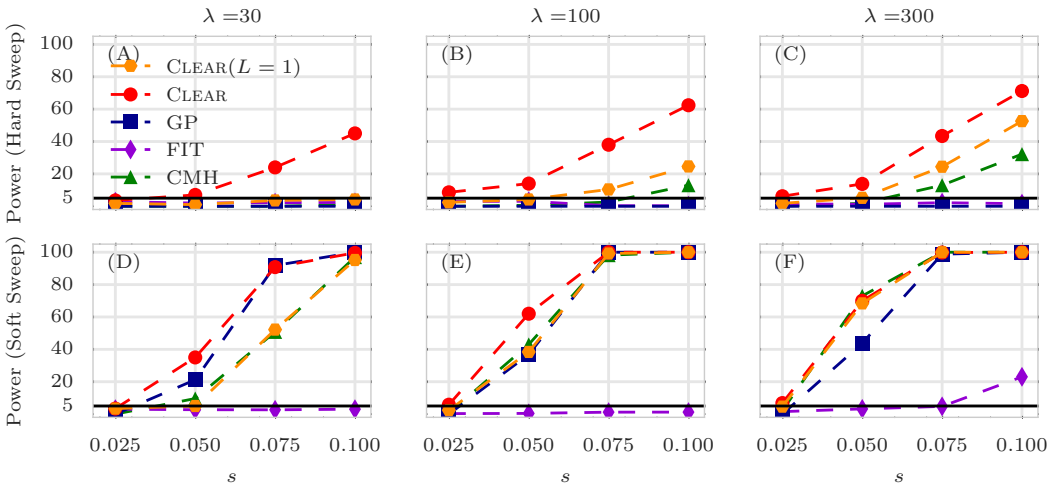

Figure 2.4: **Power calculations for detection of selection.**
Detection power for CLEAR(H), Frequency Increment Test (FIT), Gaussian Process (GP), and CMH under hard (A-C) and soft sweep (D-F) scenarios. λ, s denote the mean coverage and selection coefficient, respectively. Orange hexagons represent the performance of CLEAR when the maximum of the single-locus statistic is used to make a decision for the genomic region, while the red circle corresponds to the performance of CLEAR when single locus statistics are averaged over the region. The y-axis measures power – sensitivity with false positive rate FPR ≤ 0.05 – for $2,000$ simulations with $N = 1,000$, $L = 50$Kbp. The horizontal line reflects the power of a random classifier. In all simulations, 3 replicates are evolved and sampled at generations $T = \{0, 10, 20, 30, 40, 50\}$.

example, when coverage $\lambda = 100$ and selection coefficient $s = 0.1$, the detection power is 75% for hard sweep, but 100% for soft sweep (2.4B-E). In contrast, the favored site was ranked as the top in 14% of hard sweep cases, compared to and 95% of soft sweep simulations.

Estimating Parameters. CLEAR estimates effective population size $\widehat{N}$ and selection parameters, $\hat{s}$ and $\hat{h}$, as a byproduct of the hypothesis testing. We computed bias of selection fitness $(s - \hat{s})$ and dominance $(h - \hat{h})$ for of CLEAR and GP for 1000 simulations in each setting. The distribution of the error (bias) for 100× coverage is presented in 2.6 for different configurations. 2.13 and 2.14 provide the distribution of estimation errors for 30×, and 300× coverage, respectively. For hard sweep, CLEAR provides estimates of s with lower variance of bias (2.6A and 2.15). In soft sweep, GP and CLEAR both provide unbiased estimates of s with low variance (2.6B). 2.6 C-D shows that CLEAR provides unbiased estimates of h as well when $h \in \{0, 0.5, 1, 2\}$ and $s = 0.1$. We also tested if CLEAR provide unbiased estimates of N, by estimating population size on 1000 simulations when $N \in \{200, 600, 1000\}$. As shown in 2.8-A

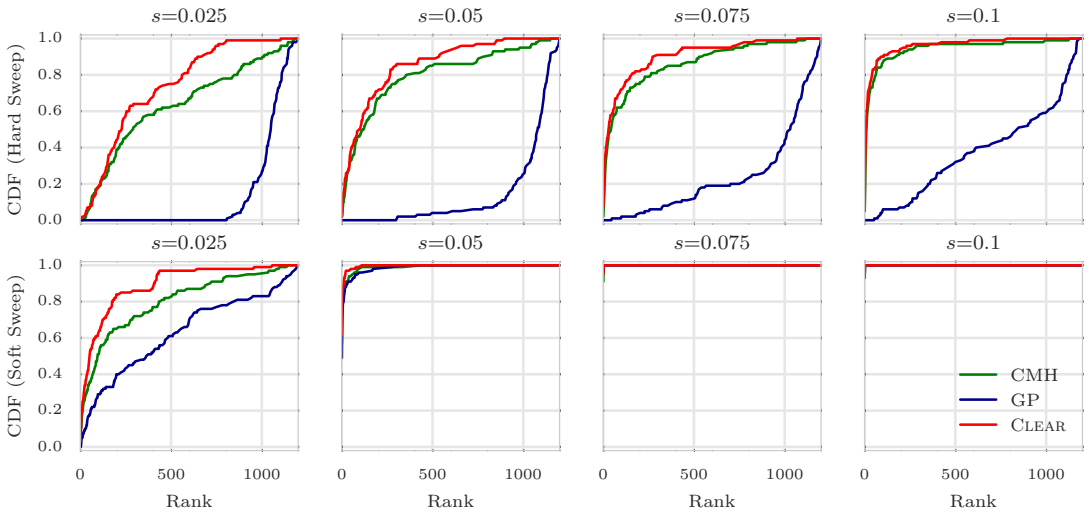

Figure 2.5: **Ranking performance for 100× coverage.**
Cumulative Distribution Function (CDF) of the distribution of the rank of the favored allele in 1000 simulations for CLEAR (H), Gaussian Process (GP), CMH, and Frequency Increment Test (FIT), for different values of selection coefficient s and initial carrier frequency. Note that the individual variant CLEAR score (H) is used to rank variants. The Area Under Curve (AUC) is computed as an overall quantitative measure to compare the performance of methods for each configuration. In all simulations, 3 replicates with $N = 1000$ are evolved and sampled at generations $T = \{0, 10, 20, 30, 40, 50\}$.

and 2.18A-C, maximum likelihood is attained at true value of the parameter.

Running Time. As CLEAR does not compute exact likelihood of a region (i.e., does not explicitly model linkage between sites), the complexity of scanning a genome is linear in number of polymorphisms. Calculating score of each variant requires and $O(TRN^3)$ computation for H. However, most of the operations are can be vectorized for all replicates to make the effective running time for each variant. We conducted 1000 simulations and measured running times for computing site statistics H, FIT, CMH and GP with different number of linked-loci. Our analysis reveals (2.7) that CLEAR is orders of magnitude faster than GP, and comparable to FIT. While slower than CMH on the time per variant, the actual running times are comparable after vectorization and broadcasting over variants (see below).

These times can have a practical consequence. For instance, to run GP in the single locus mode on the entire pool-seq data of the *D. melanogaster* genome from a small sample (≈1.6M variant sites), it would take 1444 CPU-hours (≈ 1 CPU-month). In contrast, after vectorizing and

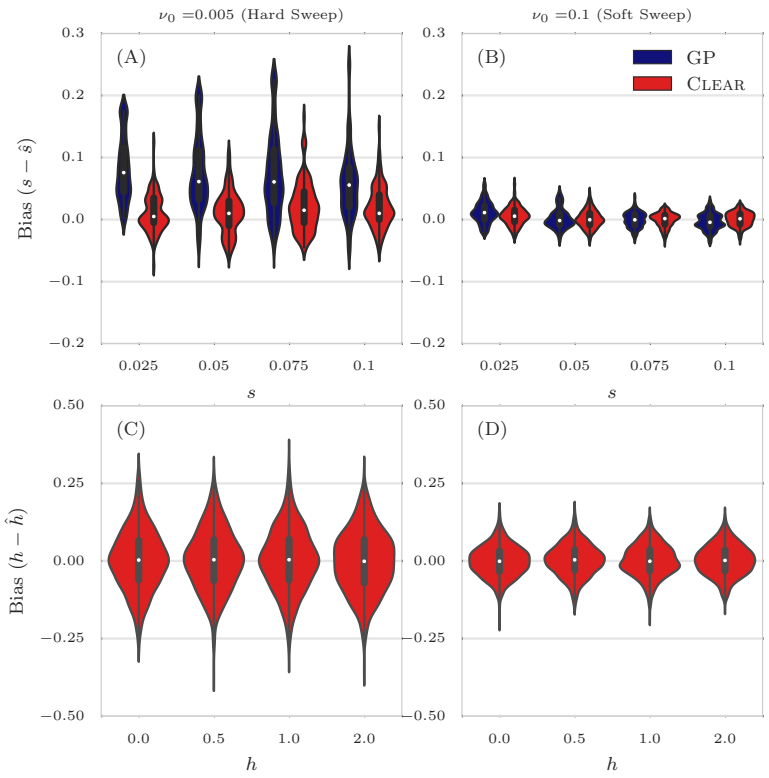

Figure 2.6: **Distribution of bias for 100× coverage.**
The distribution of bias $(s - \hat{s})$ in estimating selection coefficient over 1000 simulations using Gaussian Process (GP) and CLEAR (H) is shown for a range of choices for the selection coefficient s and starting carrier frequency ν_0, when coverage $\lambda = 100$ (Panels A,B). GP and CLEAR have similar variance in estimates of s for soft sweep, while CLEAR provides lower variance in hard sweep. Also see 2.2. Panels C,D show the variance in the estimation of h. In all simulations, 3 replicates are evolved and sampled at generations $T = \{0, 10, 20, 30, 40, 50\}$.

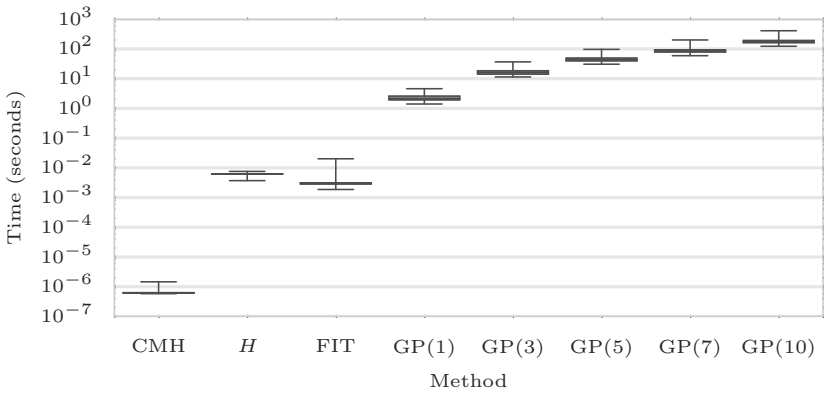

Figure 2.7: **Running time.**
Box plots of running time per variant (CPU-secs.) of CLEAR(H), CMH, FIT, and GP with single, 3, 5, 7, and 10 loci over 1000 simulations conducted on a workstation with Intel Core i7 processor. The average running time for each method is shown on the x-axis. In all simulations, 3 replicates are evolved and sampled at generations $T = \{0, 10, 20, 30, 40, 50\}$.

broadcasting operations for all variants operations using `numba` package, CLEAR took 75 minutes to perform an scan, including precomputation, while the fastest method, CMH, took 17 minutes.

2.3.1 Analysis of a *D. melanogaster* Adaptation to Alternating Temperatures

We applied CLEAR to the data from a study of *D. melanogaster* adaptation to alternating temperatures [29, 80], where 3 replicate samples were chosen from a population of *D. melanogaster* for 59 generations under alternating 12-hour cycles of hot stressful (28°C) and non-stressful (18°C) temperatures and sequenced. In this dataset, sequencing coverage is different across replicates and generations (see S2 Fig of [49]) which makes variant depths highly heterogeneous (2.19).

We first filtered out heterochromatic, centromeric and telomeric regions [98], and those variants that have collective coverage of more that 1500 in all 13 populations: three replicates at the base population, two replicates at generation 15, one replicate at generation 23, one replicate at generation 27, three replicates at generation 37 and three replicates at generation 59. After

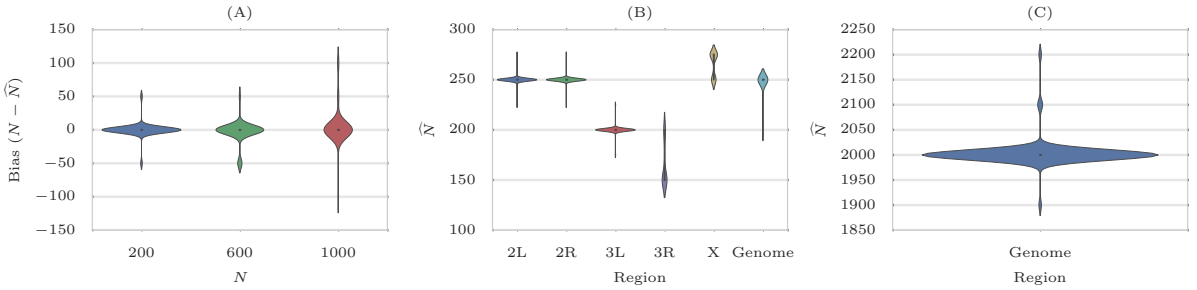

Figure 2.8: **Estimating population size.**
(A) Distribution of bias in estimating N, computed on 1000 neutral simulations for each $N \in \{200, 600, 1000\}$ when $W = 10$Mbp and $r = 2 \times 10^{-8}$. (B) Estimates of population size for data from a study of *D. melanogaster* adaptation to alternating temperatures. For each case, the distribution of estimator is computed by 100 bootstrap computations using 1000 variants each. The multiple modes are an artifact of grid search used to speed up computation. (C) Distribution of the population size estimates on the yeast dataset. Despite large census population size ($10^6 - 10^7$ [99]), this dataset exhibits much smaller effective population size ($\widehat{N} = 2000$).

filtering, we ended up with 1,605,714 variants.

Next, we estimated genome-wide population size $\widehat{N} = 250$ (2.8-B and 2.18-E) which is consistent with previous studies [29, 50]. The likelihood curves of CLEAR are sharper around the optimum compared to that of Bollback et. al [53]'s method (see Supplementary Fig. 1 in [29]). Also, chromosomes 3L and 3R appear to have smaller population size, $\widehat{N} = 200, 150$, respectively. Others have made similar observations on this data. In particular, Jónás *et al.* [50] shown that the chromosome-wise population size varies even more when it is computed for each replicate separately (see Table 1 in [50]). For instance, $\widehat{N}$ is 131 for chromosome 3R replicate 1, while it is 328 for chromosome X replicate 2.

While it would be ideal to compute CLEAR statistic for each replicate and chromosome separately, computing empirical p-values and significant regions become computationally intensive as empirical null distribution of each replicate and each chromosome needs to be computed. Hence, we use a single genome-wide estimate $\widehat{N} = 250$ in all analyses, but we normalize statistic H^* separately for each chromosome.

We use a heuristic calculation (See 2.5) to choose the sliding window size L as the distance where the LD between the favored mutation and a site $L/2$bp away remains strong. For

D. melanogaster parameters, we obtained $L = 30$Kbp. We computed the normalized test statistic H^* on sliding windows of size of 30Kbp and step size of 5Kbp over the genome (See 2.9-A).

Empirical null distribution of H^* was estimated by creating 100 whole genome simulations (400K statistic values) as described in Section 2.2.5. Then, p-value of the test statistic in each region in the experimental data was calculated as the fraction of the null statistic values that are greater than or equal to the test statistic(see 2.20). After correcting for multiple testing, we identified 5 contiguous intervals (2.9) satisfying FDR≤ 0.05, and covering $2,829$ polymorphic sites. We further performed single-locus hypothesis testing on the $2,829$ sites to identify 174 individual variants with FDR ≤ 0.01 (2.9-B).

The final set of 174 variants fall within 32 genes(2.3) including many Serine inhibitory proteases (serpins), and other genes involved in endocytosis. Recycling of synaptic vesicles is seen to be blocked at high temperature in temperature sensitive Drosophila mutants [100]. This is also supported by GO enrichment analysis, where a single GO term 'inhibition of proteolysis' is found to enriched (corrected p-value:0.0041). To test for dominant selection, we computed D statistic on simulated neutral and experimental data, and computed p-values accordingly. After correcting for multiple testing, 96 variants were discovered with FDR≤ 0.01 (2.21).

2.3.2 Analysis of Outcrossing Yeast Populations

We also applied CLEAR to 12 replicate samples of outcrossing yeast populations [99], where samples are taken at generations $T = \{0, 180, 360, 540\}$. We observed a significant variation in the genome-wide site frequency spectrum of certain populations over different time points for some replicates (2.22). The variation does not have an easily identifiable cause. Therefore, we focused analysis on seven replicates $r \in \{3, 7, 8, 9, 10, 11, 12\}$ with genome-wide site-frequency spectrum over the time range (2.23).

We estimated population size to be $\widehat{N} = 2000$ haplotypes (2.8-C and 2.18-F), and computed $\widehat{s}$, $\widehat{h}$ and H statistic accordingly. To compute p-values, we created 1M single-locus neutral

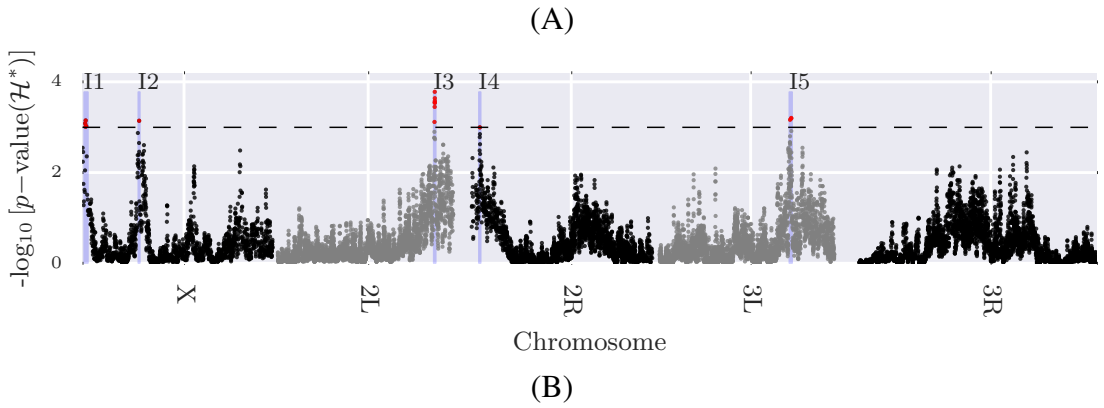

Figure 2.9: **Scan of CLEAR statistic on data from a study of *D. melanogaster* adaptation to alternating temperatures.**
(A) Manhattan plot of scan for H^* statistic using sliding window of size $L = 3000$ over the genome. The dashed line represents cutoff for genome-wide FDR≤ 0.05, and identifies 5 contiguous intervals, I1-I5, which are shaded in blue. (B) Trajectories of the selected variants within intervals I1-I5.

simulations according to experimental data's initial frequency and coverage. By setting FDR cutoff to 0.05, only 18 and 16 variants show significant signal for directional and dominant selection, respectively (2.21). Selected variants for directional selection are clustered in two regions, which match 2 of the 5 regions (regions C and E in Fig. 2-a in [99]) identified by Burke *et al.* in their preliminary analysis.

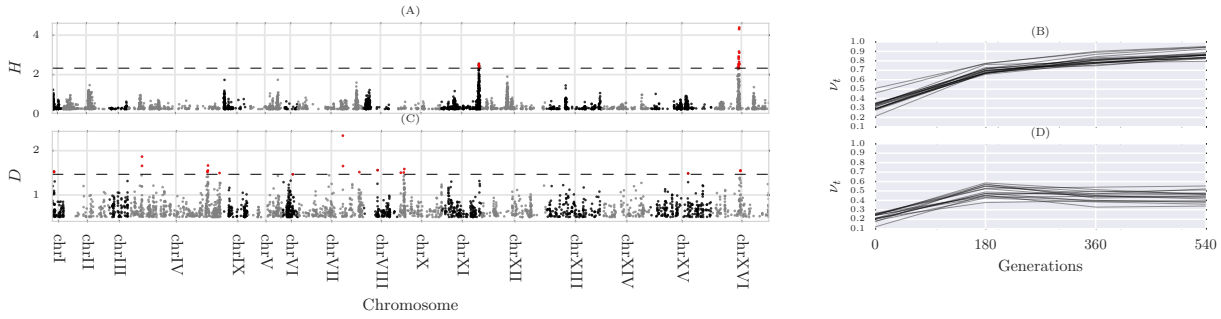

Figure 2.10: **Single locus analysis of the yeast outcrossing populations.**
Manhattan plot of scan single locus CLEAR statistic ($L = 1$) for testing directional selection (A) and dominant selection (C). The dashed line represents cutoff for genome-wide FDR≤ 0.05. Trajectories of the selected variants are depicted in panels (B) and (D).

2.4 Discussion

We developed a computational tool, CLEAR, that can detect regions and variants under selection E&R experiments. Using extensive simulations, we show that CLEAR outperforms existing methods in detecting selection, locating the favored allele, and estimating model parameters. Also, while being computationally efficient, CLEAR provide means for estimating populations size and hypothesis testing.

Many factors such as small population size, finite coverage, linkage disequilibrium, finite sampling for sequencing, duration of the experiment and the small number of replicates can limit the power of tools for analyzing E&R. Here, by an discrete modeling, CLEAR estimates population size, and provides unbiased estimates of s, h. It adjusts for heterogeneous coverage of pool-seq data, and exploits presence of linkage within a region to compute composite likelihood ratio statistic.

It should be noted that, even though we described CLEAR for small fixed-size populations, the statistic can be adjusted for other scenarios, including changing population sizes when the demography is known. For large populations, transitions can be computed on sparse data structures, as for large N the transition matrices become increasingly sparse. Alternatively, frequencies can be binned to reduce dimensionality.

The comparison of hard and soft sweep scenarios showed that initial frequency of the favored allele can have an nontrivial effect on the statistical power for identifying selection. Interestingly, while it is easier to detect a region undergoing strong selection, it is harder to locate the favored allele in that region.

There are many directions to improve the analyses presented here. In particular, we plan to focus our attention on other organisms with more complex life cycles, experiments with variable population size and longer sampling-time-spans. As evolve and resequencing experiments continue to grow, deeper insights into adaptation will go hand in hand with improved

computational analysis.

Software and Data Availability.

The source code and running scripts for CLEAR are publicly available at `https://github.com/airanmehr/clear`.

D. melanogaster data originally published [29, 80]. The dataset of the *D. melanogaster* study, until generation 37, is obtained from Dryad digital repository (`http://datadryad.org`) under accession DOI: 10.5061/dryad.60k68. Generation 59 of the *D. melanogaster* study is accessed from European Sequence Read Archive (http://www.ebi.ac.uk/ena/) under the project accession number: PRJEB6340. The dataset containing experimental evolution of Yeast populations [99] is downloaded from `http://wfitch.bio.uci.edu/~tdlong/PapersRawData/BurkeYeast.gz` (last accessed 01/24/2017). UCSC browser tracks for *D. melanogaster* and Yeast data analysis are found in Suppl. Data 1 and 2, respectively

2.5 Choosing Window Size

In genome-wide scans for detecting selection, we apply the CLEAR statistic on sliding windows of length *L*bp. The single locus statistic values within the window are averaged to get the composite statistic. While the statistic is robust to variation in window-size, choosing a very large window where LD has decayed will weaken the composite signal, and choosing a small window will decrease the power of composite likelihoods. Here, we use a systematic calculation to choose L as the distance where the LD between the favored mutation and a site $L/2$bp away remains strong.

Consider a segregating site l bp away from the favored allele in a selective sweep. Let ρ_τ be the LD between the favored allele and the site, τ generations after the onset of selection. Then,

we have (see Eqs. 30-31 in [101]):

$$\rho_\tau = \alpha_\tau \beta_\tau \rho_0 = e^{-r\tau l} \left(\frac{K^{(\tau)}}{K^{(0)}} \right) \rho_0, \tag{2.17}$$

where $K^{(\tau)} = 2\nu_\tau(1-\nu_\tau)$ is the heterozygosity at the selected site, r is the recombination rate (crossovers/bp/gen). The 'decay factor', $\alpha_\tau = e^{-r\tau l}$, and 'growth factor', β_τ, are due to recombination and selection, respectively. Under regular parameter settings, linkage to the favored allele is expected to increase after onset of selection and then decreases due to crossover events (See 2.24-A). While ρ_0 is unknown in pool-seq E&R experiments, we compute the value of l so that

$$\alpha_\tau \beta_\tau = 1. \tag{2.18}$$

In E&R scenarios, we let τ be the time of the last sampling. For given s, we aim to compute the smallest window size L over all possible starting frequencies. Specifically,

$$L = 2 \min_{\nu_0} \left\{ \frac{1}{r\tau} \log \left(\frac{\hat{\nu}_\tau(1-\hat{\nu}_\tau)}{\nu_0(1-\nu_0)} \right) \right\}, \tag{2.19}$$

where the term $\hat{\nu}_\tau$ depends on initial frequency ν_0 and selection strength s (Eq. 2.9).

We used *D. melanogaster* dataset parameters, $N = 250, r = 2 \times 10^{-8}$ and $\tau = 59$ to compute the optimal window size for different values of Ns, ranging from weak selection to strong selection: $Ns \in \{20, 100, 200, 500\}$, or $s \in \{0.08, 0.4, 0.8, 2\}$. We set $L = 30$Kbp (See 2.24-B) to provide good resolution for detecting weak selection.

2.5.1 Acknowledgments

Chapter 2, in full, contains material from Arya Iranmehr, Ali Akbari, Christian Schltterer, Vineet Bafna. "Clear: composition of likelihoods for evolve and resequence experiments". Genetics. 2017 Jan 1:genetics-116 [1]. I was the primary investigator and author of this paper.

Input: $N, n, R, \{\lambda_t\}_{t \in T}, T = \{\tau_0, \ldots \tau_T\}$
Output: Time-series pool-seq data for R replicates of a single locus $\{\mathbf{c}\}_R$ and $\{\mathbf{d}\}_R$.
for $r \leftarrow 1$ **to** R **do**
  for $t \leftarrow \tau_0$ **to** τ_T **do**
    $2N\nu_t \sim \text{Binomial}(2N, \nu_{t-1})$;
    if $t \in T$ **then**
      $d_t^{(r)} \sim \text{Poisson}(\lambda_{\tau_i})$;
      $2ny_t \sim \text{Binomial}(2n, \nu_t)$;
      $c_t^{(r)} \sim \text{Binomial}(d_t^{(r)}, y_t)$;
    end
  end
end

Figure 2.11: **The Generative Process for Neutral Wright-Fisher Time-series Pool-seq Data.**

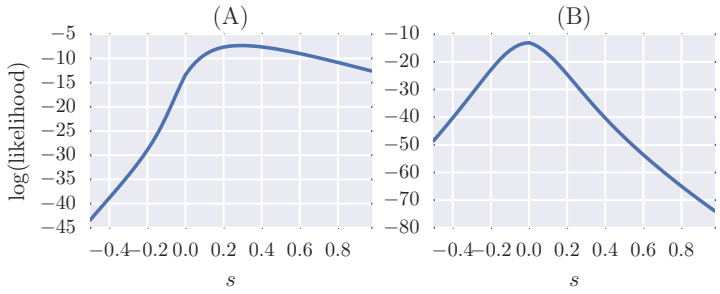

Figure 2.12: **Likelihoods of the parameter s.**
Likelihood of the parameter s in *D. melanogaster* data for a variant with $\widehat{s} = 0.2$ (A) and $\widehat{s} = 0$ (B).

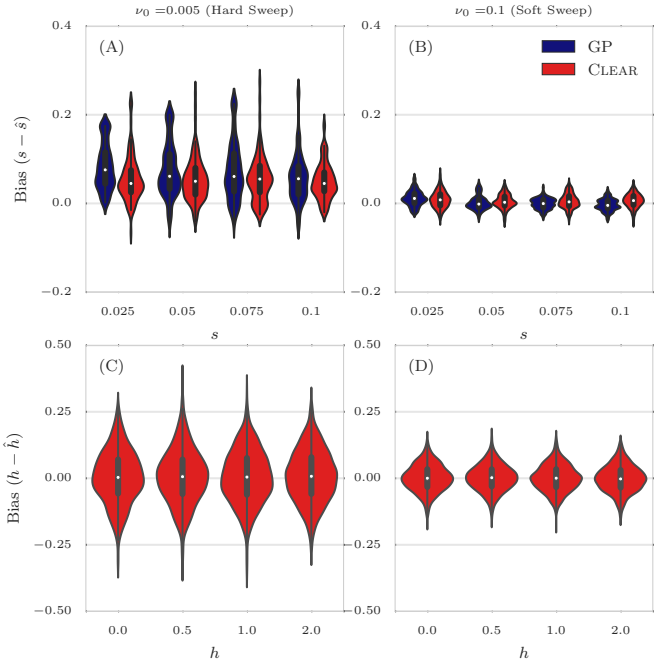

Figure 2.13: **Distribution of bias for 30× coverage.**
The distribution of bias $(s - \hat{s})$ in estimating selection coefficient over 1000 simulations using Gaussian Process (GP) and CLEAR (H) is shown for a range of choices for the selection coefficient s and starting carrier frequency ν_0, when coverage $\lambda = 30$ (Panels A,B). GP and CLEAR have similar variance in estimates of s for soft sweep, while CLEAR provides lower variance in hard sweep. Also see 2.2. Panels C,D show the variance in the estimation of h.

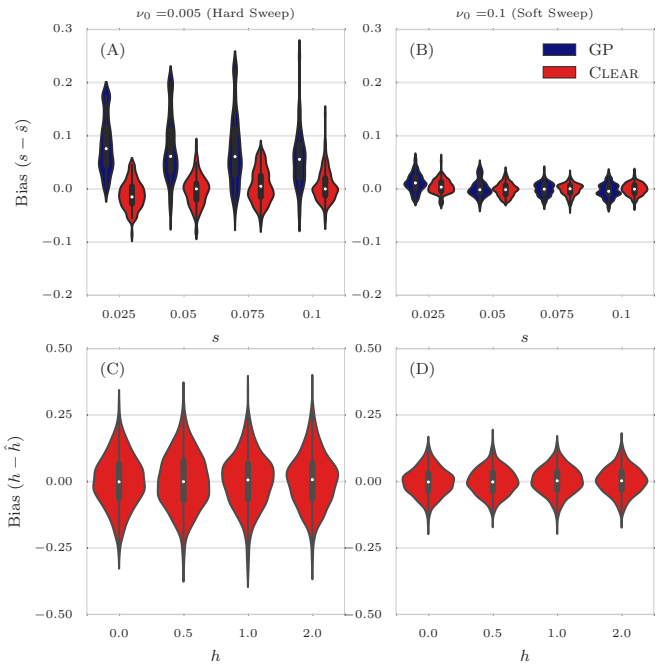

Figure 2.14: **Distribution of bias for 300× coverage.**
The distribution of bias $(s - \hat{s})$ in estimating selection coefficient over 1000 simulations using Gaussian Process (GP) and CLEAR (H) is shown for a range of choices for the selection coefficient s and starting carrier frequency ν_0, when coverage $\lambda = \infty$ (Panels A,B). GP and CLEAR have similar variance in estimates of s for soft sweep, while CLEAR provides lower variance in hard sweep. Also see 2.2. Panels C,D show the variance in the estimation of h.

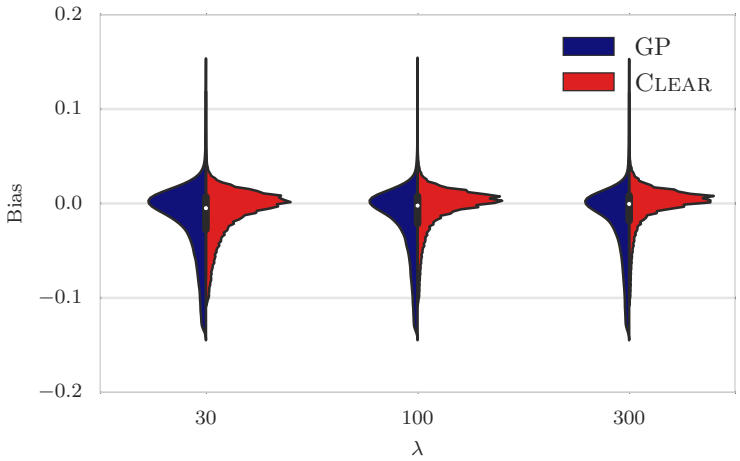

Figure 2.15: **Distribution of bias for null simulations.**
Distribution of bias for null simulations with coverage $\lambda \in \{30, 100, 300\}$.

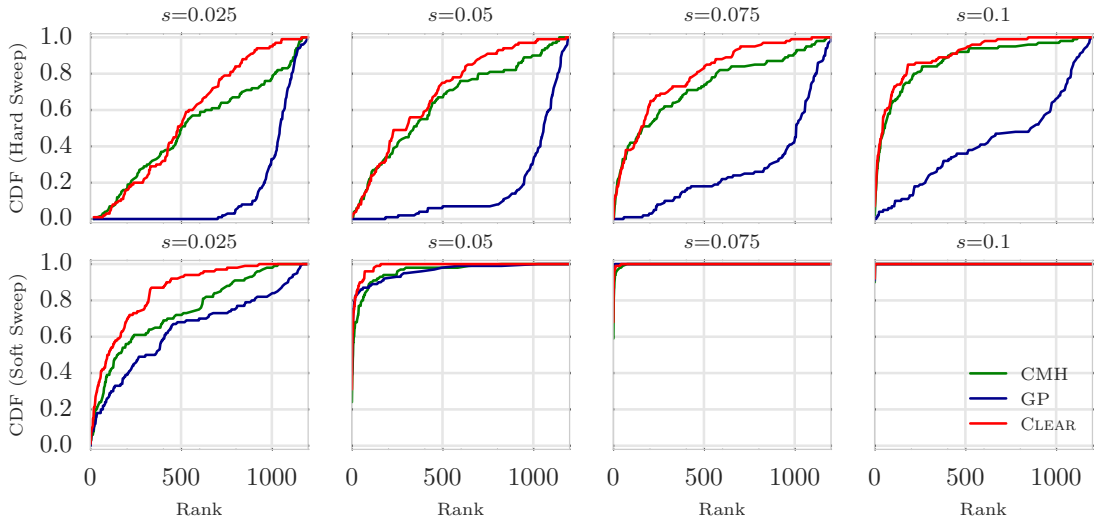

Figure 2.16: **Ranking performance for 30× coverage.**
Cumulative Distribution Function (CDF) of the distribution of the rank of the favored allele in 1000 simulations for CLEAR (*H* score), Gaussian Process (GP), and Cochran Mantel Haenszel (CMH), for different values of selection coefficient *s* and initial carrier frequency.

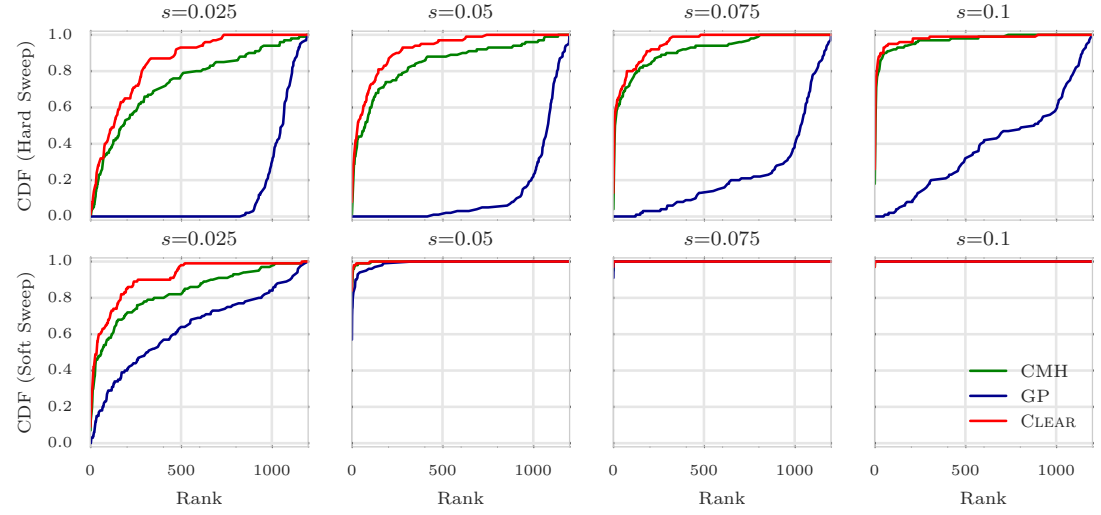

Figure 2.17: **Ranking performance for 300× coverage.**
Cumulative Distribution Function (CDF) of the distribution of the rank of the favored allele in 1000 simulations for CLEAR (*H* score), Gaussian Process (GP), and Cochran Mantel Haenszel (CMH), for different values of selection coefficient *s* and initial carrier frequency.

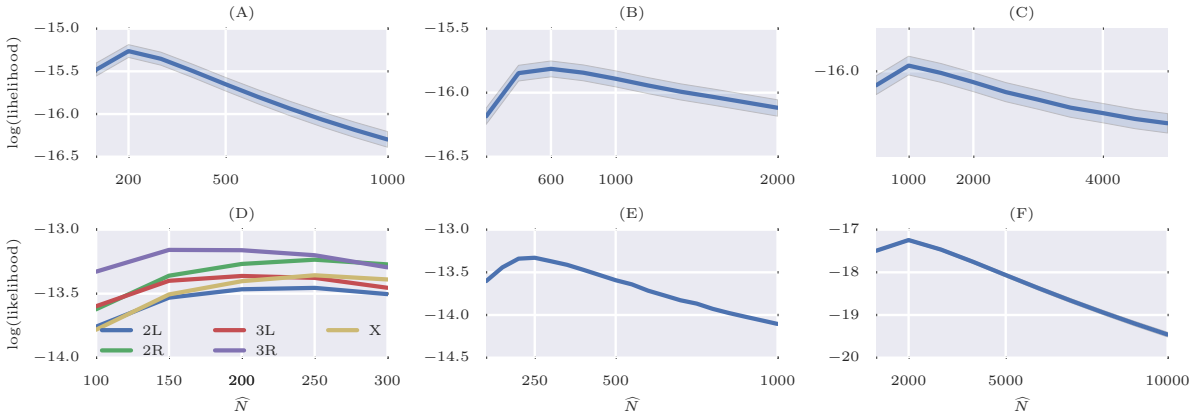

Figure 2.18: **Maximum likelihood Estimates of N.**
Mean and 95% confidence interval of likelihoods of N on simulated data with $N = 200$ (A), $N = 600$(B), and $N = 1000$ individuals, over 1000 simulations. Chromosome-wise (D) and genome-wide (E) likelihood of population size for data from a study of *D. melanogaster* adaptation to alternating temperatures. Likelihood of the Chromosome 3R is attained at 150, while genome-wide maximum likelihood estimate for population size is 250. (F) Likelihood of the population size with respect to all the variants in the yeast dataset. Despite large census population size ($10^6 - 10^7$ [99]), this dataset exhibits much smaller effective population size ($\widehat{N} = 2000$).

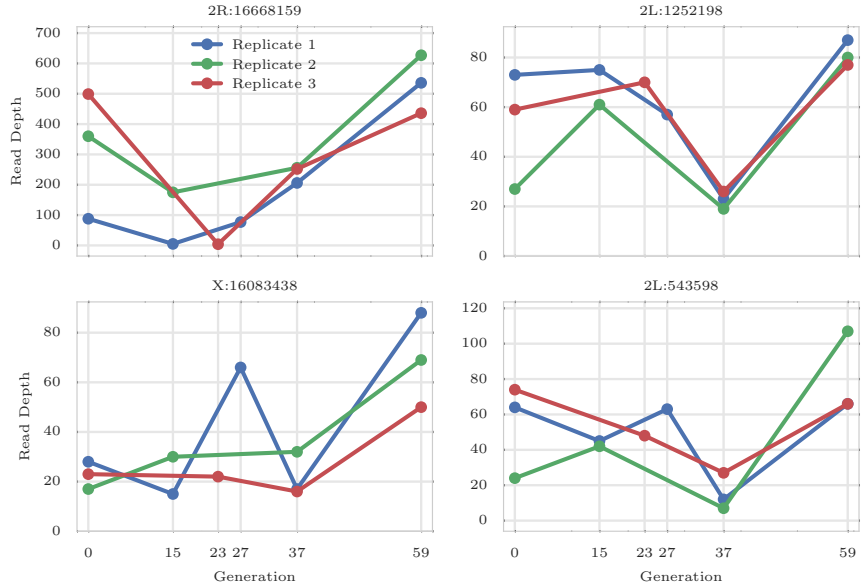

Figure 2.19: **Coverage heterogeneity in time series data.**
Each panel shows the read depth for 3 replicates of the data from a study of *D. melanogaster* adaptation to alternating temperatures data (see section 2.3.1). Heterogeneity in depth of coverage is seen between replicates, and also at different time points, in all 4 variants. None of these sites pass the the hard filtering with minimum depth of 30.

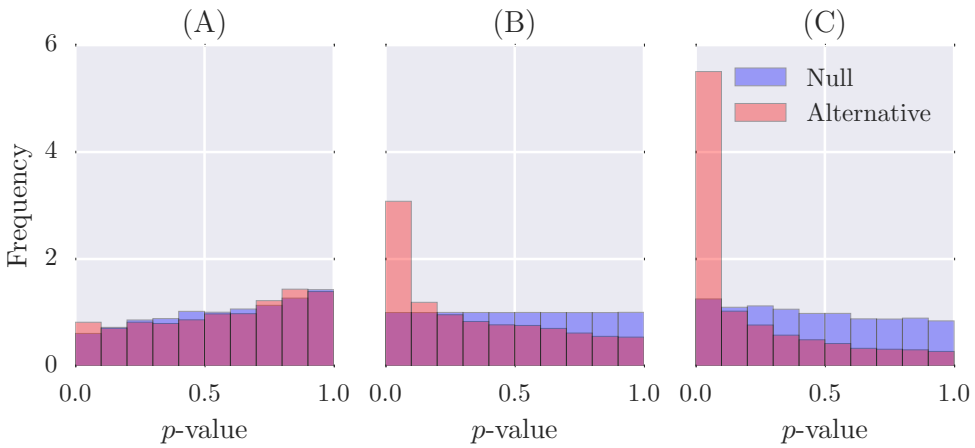

Figure 2.20: **Distribution of p-values.**
Distribution of p-values of CLEAR in null simulations and experimental data when $N = 250$. Panel (A),(C) shows the effect of under estimations ($\widehat{N} = 100$) and over-estimation ($\widehat{N} = 500$) of population size in computing p-values, and panel (B) shows the distribution of $p-$values when unbiased estimate is used to create simulations. .

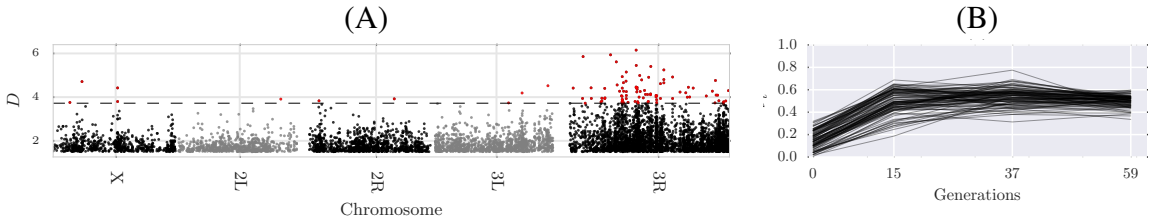

Figure 2.21: **Single locus analysis of the data from a study of *D. melanogaster* adaptation to alternating temperatures.**
Manhattan plot of scan for testing dominant selection (A). Significant variants with FDR ≤ 0.01 are denoted in red, and their trajectories are depicted in panel (B).

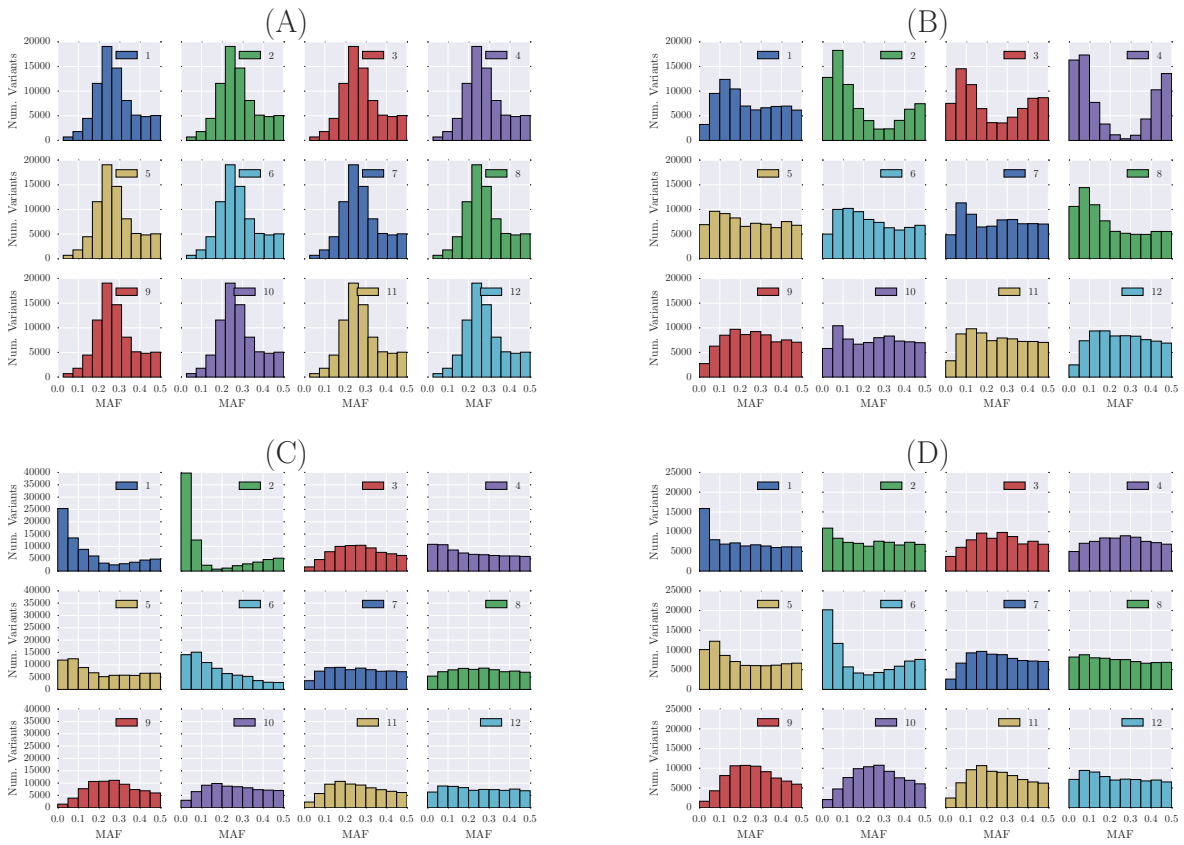

Figure 2.22: **Site frequency spectrum of the Yeast dataset.** Whole-genome site frequency spectrum of the Yeast dataset at generations 0 (A), 180 (B), 360 (C) and 540 (D). Some replicates, e.g. replicate 2, undergoing severe demographic events.

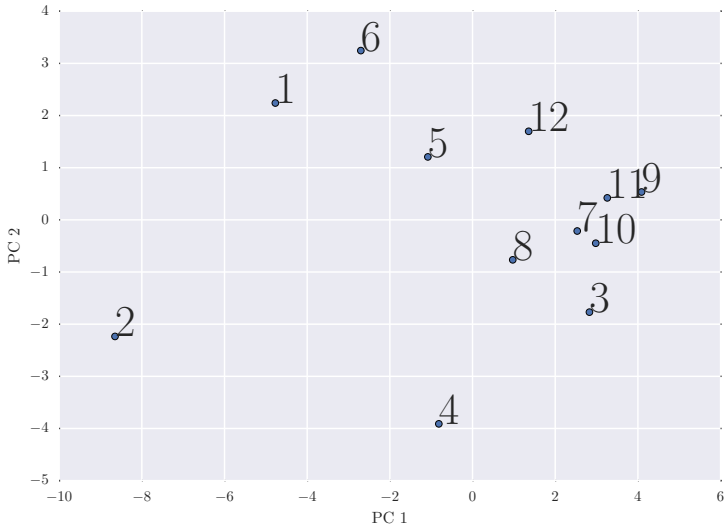

Figure 2.23: **Population similarity.**
Principle component analysis of the 12 replicates throughout the experiment, showing that some populations exhibiting distinct frequency spectra.

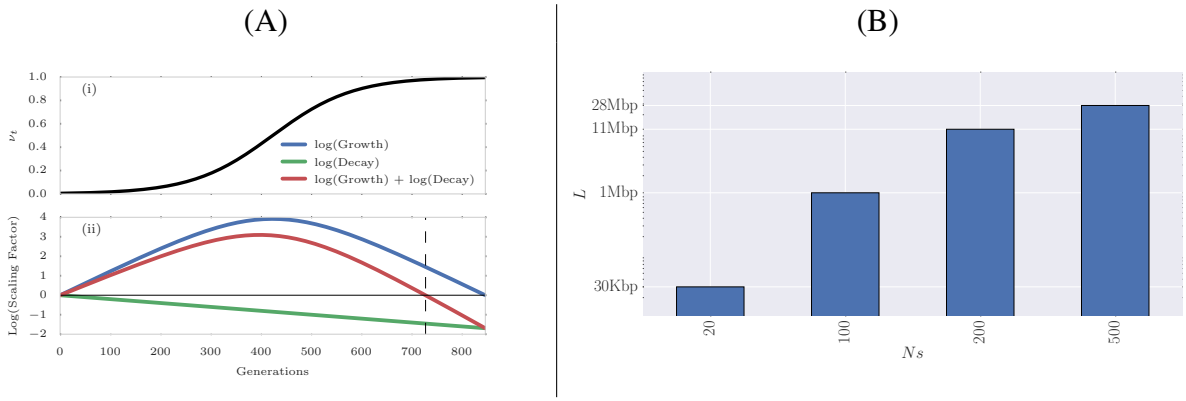

Figure 2.24: **Choosing window size for CLEAR statistic.**
(A) Expected dynamics of LD between favored allele (s = 0.025) and a variant 50Kbp away, with initial frequency ν_0 = 0.01. (A-i) depicts the dynamic of the favored allele during the selective sweep. (A-ii) illustrates interaction of the growth and decay factors introduced in Eq. 2.17, with the red line describing overall effect of selection and recombination on LD. The vertical dashed line points to the time when the LD value starts to decrease below original LD. (B) Alternatively, we can fix time, and find the window-size at which LD decays below the original LD (Eq. 2.19). The plot shows the window size as a function of Ns (20,100,200,500), after fixing other model parameters to match *D. melanogaster* E&R experiments ($N = 250$, $r = 2 \times 10^8, \tau = 59$).

Table 2.1: **Average power.**
Average power is computed for 8000 simulations with $s \in \{0.025, 0.05, 0.075, 0.1\}$. Frequency Increment Test (FIT), Gaussian Process (GP), CLEAR (H statistic) and Cochran Mantel Haenszel (CMH) are compared for different initial carrier frequency ν_0. For all sequencing coverages, CLEAR outperform other methods. When coverage is not high ($\lambda \in \{30, 100\}$) and initial frequency is low (hard sweep), CLEAR significantly perform better than others.

Hard Sweep			Soft Sweep		
λ	Method	Avg Power	λ	Method	Avg Power
300	CLEAR	34	300	CLEAR	69
300	CLEAR$(L=1)$	21	300	CMH	69
300	CMH	12	300	CLEAR$(L=1)$	68
300	FIT	2	300	GP	61
300	GP	0	300	FIT	8
100	CLEAR	31	100	CLEAR	67
100	CLEAR$(L=1)$	10	100	CMH	60
100	CMH	4	100	CLEAR$(L=1)$	60
100	FIT	2	100	GP	59
100	GP	0	100	FIT	1
30	CLEAR	20	30	CLEAR	57
30	CLEAR$(L=1)$	3	30	GP	53
30	FIT	2	30	CMH	39
30	CMH	0	30	CLEAR$(L=1)$	39
30	GP	0	30	FIT	3

Table 2.2: **Mean and standard deviation of the distribution of bias ($s - \hat{s}$) of 8000 simulations with coverage $\lambda = 100\times$ and $s \in \{0.025, 0.05, 0.075, 0.1\}$.**

Method	ν_0	Mean	STD
GP	0.005	0.073	0.061
CLEAR	0.005	0.016	0.035
GP	0.1	0.002	0.016
CLEAR	0.1	0.002	0.013

Table 2.3: **Overlapping genes with the 174 candidate variants.**

Interval	Position	FBgn	Gene Name	GO Function
I1	X:1.567-1.824M	FBgn0023531	CG32809	NA
		FBgn0023130	a6	embryonic development via the syncytial blastoderm
		FBgn0025378	CG3795	serine-type endopeptidase activity
		FBgn0025391	Scgdelta	heart contraction, mesoderm development
		FBgn0261548	CG42666	NA
		FBgn0026086	Adar	RNA editing
		FBgn0026090	CG14812	negative regulation of cysteine-type endopeptidase activity involved in apoptotic process
		FBgn0023522	CG11596	NA
I2	X:7.175-7.241M	FBgn0029941	CG1677	NA
		FBgn0029944	Dok	stress activated protein kinase signaling
		FBgn0029946	CG15034	NA
I3	2L:16.878-16.993M	FBgn0052832	CG32832	mitochondrial pyruvate transport
		FBgn0032618	CG31743	sulfotransferase activity
		FBgn0085342	CG34313	NA
		FBgn0040985	CG6115	NA
		FBgn0261671	tweek	synaptic vesicle endocytosis
		FBgn0026150	ApepP	metalloaminopeptidase activity
		FBgn0262355	CR43053	NA
		FBgn0053179	beat-IIIb	NA
I4	2R:2.725-2.810M	FBgn0040674	CG9445	NA
		FBgn0265935	coro	adult somatic muscle development
		FBgn0033110	CG9447	NA
		FBgn0033113	Spn42Dc	Inhibitory Serpins
		FBgn0028988	Spn42Dd	Inhibitory Serpins
		FBgn0033115	Spn42De	Inhibitory Serpins
		FBgn0050158	CG30158	small GTPase mediated signal transduction
I5	3L:14.362-14.514M	FBgn0036421	CG13481	ubiquitin-protein transferase activity
		FBgn0262580	CG43120	NA
		FBgn0036422	CG3868	NA
		FBgn0087007	bbg	PDZ domain
		FBgn0036426	CG9592	NA
		FBgn0036427	CG4613	serine-type endopeptidase activity

Chapter 3

Analysis of Long-term Experimental Evolution

3.1 Introduction

Oxygen homeostasis is at the basis of human health and is central to the development of the top leading causes of deaths including cancer [102, 103], ischemic [104–106], and respiratory disease [107]. Aging [108], other phenotypes such as body size [109, 110], associated to chronic Hyperoxic stress. Using experimental model systems of evolution and adaptation to these stresses can allow us to understand the genes and pathways involved in maintaining oxygen homeostasis. Importantly, observing evolution in action reveals the underlying nature of adaptive evolution, determined by mode, tempo and mechanism of fixation. We have conducted a long-term study of *D. melanogaster* evolving under hyperoxia for more than 180 generations over the course of 10 years to learn more about oxygen metabolism. its limits and protective alleles by observing trajectories of common variations and the associated phenotype. Additionally, we revisit and test some of the central evolutionary hypotheses: whether selection acts on *de novo*, or standing variation; estimation the magnitude of effect size; and whether the effect of beneficial allele

changes under different genomic backgrounds. Our results suggest that acquiring the beneficial allele via standing variation, rare-variation, recombination and epistatic interactions are of the evolutionary mechanisms by which a stressed population rapidly adapts to an adverse environment. Our findings provide new insights for future selection experiments to map local interactions.

3.2 Population Differentiation

To create experimental founder populations, 27 isogenic *D. melanogaster* founder lines were crossed and larger populations (n=1080) created. 3 replicate fly populations were evolved under increasingly hyperoxic conditions, with oxygen levels gradually increased from 20% to 90%. DNA from each replicate population was pooled and WGS acquired at generations 1, 7, 12, 31, 61, 114, and 180 (Fig. 1A, 1-B-I). To test if toxicity due to extreme hyperoxia led to change in population demography, we used a Wright-Fisher Hidden Markov model to compute effective population sizes from pooled allele frequencies (Methods). The calculations showed a severe population bottleneck in all 3 hyperoxia populations between generations 12 to 61, followed by a rapid recovery consistent with genetic adaptation to hyperoxic stress (Fig. 1B). Additionally, the computationally estimated population sizes were in concordance with hand-counted census population sizes (Fig. 1-B-II). To examine the temporal evolution of the populations, we performed a PCA including all hyperoxia evolving populations, and also replicate control populations evolving under hypoxia and normoxia (orange square, green triangles, and blue circles in Fig. 1-C-I). The first principle component (PC1 Fig. 1-C-II) captured deviations of the Hyperoxia populations while the second principle component encoded the evolution and differentiation of control and hypoxia populations (Fig. 1-C-III,IV). Notably, successive generations of hyperoxia evolving populations differentiated progressively from the starting populations, and to a larger extent than the distance between replicate, non-mixing populations. At generation 180, the genetic distance of Hyperoxia populations from the initial populations was

significantly greater than that of control and Hypoxia populations at generation 180, suggestive of a severe stress. We also observed that the rate at which Hyperoxia populations differentiated, defined by genetic distance per generation (Methods), accelerated with the initiation of bottleneck at generations 7-12, and slowed down as the populations recover, after generations 61 (Fig S9).

3.3 Adaptation and Mechanisms

In addition to the strong selection pressure, genetic drift on small population (Ne¡200) also caused allele frequencies to fluctuate (Fig. S3). Therefore, identifying the genomic targets of adaptation reduced to distinguishing between genetic drift and selection trajectories (Fig. S3). Also, as strength of the genetic drift is inversely correlated with population size, we adjusted for the demographic changes (Fig 1-B) (Methods) to correct for the common confounder in tests of selection.

The causal genetic locus tags its surrounding variation via linkage-disequilibrium. Whether the linked mutations have occurred before or after the causal loci, determines its fate, by which they reach incomplete or complete fixation, respectively (Fig. S1). As a result, the genetic makeup of the target loci at the onset of selection, dictates the patterns by which the linked loci converge to a single or multiple frequencies Fig 2-A-III, 2-B-III. If a unique carrier background exists at the onset of selection, that is no mutation occurs after the causal loci to differentiate carrier backgrounds, all the linked variation diminishes or fixate at fixation (Fig 2-A-II). This corresponds to hard sweep, in which all the carriers coalesce before the onset of selection (Fig 2-A-I). On the other hand, if many mutations occur on the carrier background before the onset of selection (blue and orange circles in Fig 2-B), creates unique signature in time-series data from soft sweep, in which each cluster of highly linked mutation correspond to a different beneficial background.

To maximize statistical power, we scanned the genome for the signals that are replicated in

all three hyperoxia replicates (Methods). Also, as in small populations genetic drift acts strongly and can fixate neutral mutations, we limit the time-span for a selective sweep, an only consider those intervals that are longer that 20 generations and shorter than 100 generations.

Our calculations identified three intervals with consistent signals of selection across replicates (Fig. 3, Fig S4,5). The most significant interval (I1, Fig. 3) corresponds to a rapid and synchronous fixation 603 mutations in 3 populations and elimination of 940 polymorphic sites (Fig S6), implying presence of single beneficial haplotype (hard sweep). Similar signatures observed in the interval I3 with a weaker effect size (Fig. 3, C). We also find a replicated signal of soft sweep in which 146 mutations fixate and 102 mutations remain polymorphic and remain lined, after fixation. Interestingly, all the replicable signals are observed at the time series starting at generations 12 or earlier (Fig S2), that are the generations that the 3 replicates are the most similar (Fig. 1 C).

3.4 Epistasis

When only a few haplotypes exist in the population as a result of selection or demographic changes, the standing variation is tightly linked and forms distinct mutational clusters in time-series data (Fig S10,12). The mutational clusters can be used to identify haplotypes in a low diversity regime. Also, time-series data allows us to measure the fitness of a mutations between an interval. Hence, we can measure the fitness before and after of any middle time point. We use these principles to identify rapid and severe changes in the fitness of haplotypes that could be a signal for an non-additive genetic interactions (epistasis). When a haplotype is at a low frequency, then its fitness is low or negative. The rapid raise of such haplotypes corresponds to gain of fitness that could be due to a *de novo* mutation or recombination (black mutations in Fig 4-A-I,III). If the *de novo* exists on the other replicates and it has a low or negative fitness then it is epistatic. The characteristic signature of this model is the raise of many low frequency mutations

to fixation. Additionally, if more than one haplotype exists in the region in form of highly linked cluster of polymorphic mutations, the mutational clusters split. Those mutations that co-occur on the low-frequency background (green circles Fig 4-A-III), also hitchhike to fixation. Such a signal can be created either by *de novo* mutation(s) or a recombination. When the former is the case, the fixing and diminishing mutations expected to be distributed uniformly in the epistatic region Fig. 4A-IV. The latter creates similar signal in trajectories, the genomic distribution of the fixed and absorbed mutation is completely asymmetric Fig. 4A-V.

To scan for epistatic regions, we analyzed each replicate independently for signals of selection, and identified 7 intervals (Table S1), 5 of which correspond to the epistasis due to *de novo* mutation (Fig S11) and two intervals provide evidence for an epistatic recombination. For instance, one epistatic recombination is observed in the replicate 1 (Fig. 4B), of which a cluster of mutations split at generation 114, and simultaneously, a cluster of low frequency mutations raise to fixation. The splitting, fixation and extinction pattern in trajectories and their genomic distributions are consistent with the model (Fig. 4 I,II) and simulation (Fig. 4 III,IV).

3.5 Conclusion and Discussion

Identifying the genetic basis of rapid adaptation is central focus of evolutionary biology. Existing models predict population rescue itself by finding the beneficial allele from new mutations or standing variation. On the other hand, deleterious mutations contribute negatively to the fitness effect of a genetic background [111] and accumulation of deleterious mutations increases the genetic load. In sexual organisms, recombination decouples deleterious mutations from the neutral and beneficial variation and breaks the Mullers ratchet. Recombination can also act as a mechanism to bring alleles together and make a new beneficial haplotype. In another experimental evolution on *D. melanogaster* [112], it has been shown that epistasis is prevalent that makes effect-size estimates irreplicable. In this experiment, we observed evidence of for a wide range of

adaptive evolutionary mechanisms including standing variation, *de novo* mutation, recombination and epistasis.

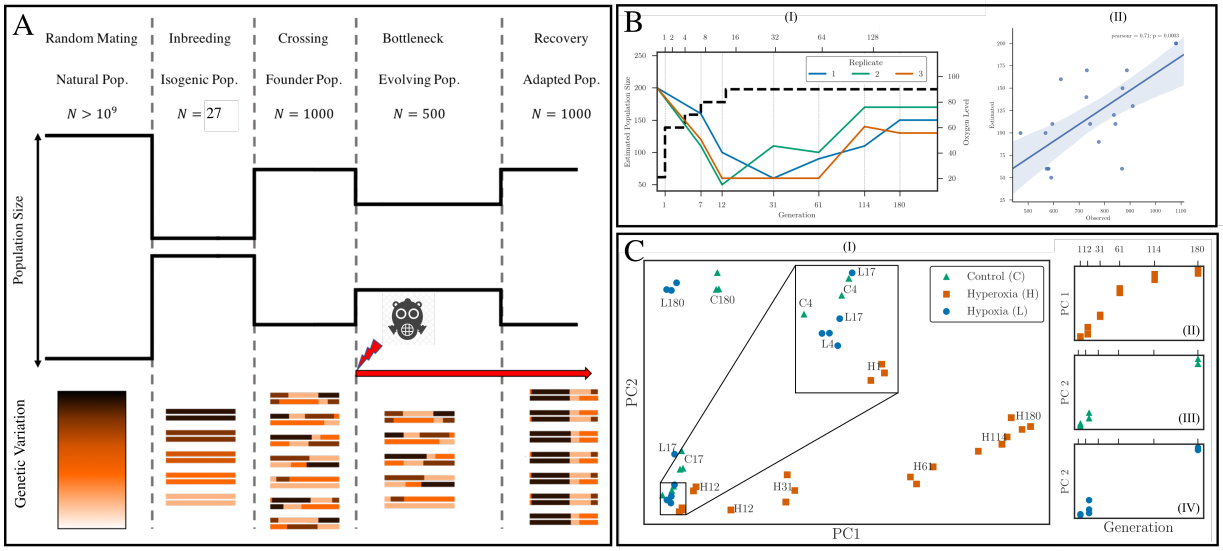

Figure 3.1: **Long-term *D. melanogaster* experimental evolution.**
(A) *D. melanogaster* selection experiment. 27 flies are collected from natural environment to create isogenic lines, and then founder populations. Exposing the founder population to high oxygen levels, causes a bottleneck and eventually fixation of the carriers of the beneficial allele in the population. (B) Estimated effective population size between every consecutive samples provide evidence for a bottleneck (I). The estimates are consistent with the experimental observations (II). (C) Differentiation of experimental populations (I), in which the first principle component captures sequential differentiation of hyperoxia populations.

3.5.1 Acknowledgments

Chapter 3, in full, contains material from Arya Iranmehr, Tsering Stobdan, Dan Zhou, Vineet Bafna, Gabriel G Haddad and Helen Zhao. “Revealing Evolutionary Forces via Experimental Evolution”. In preparation. I was the primary investigator and author of this paper.

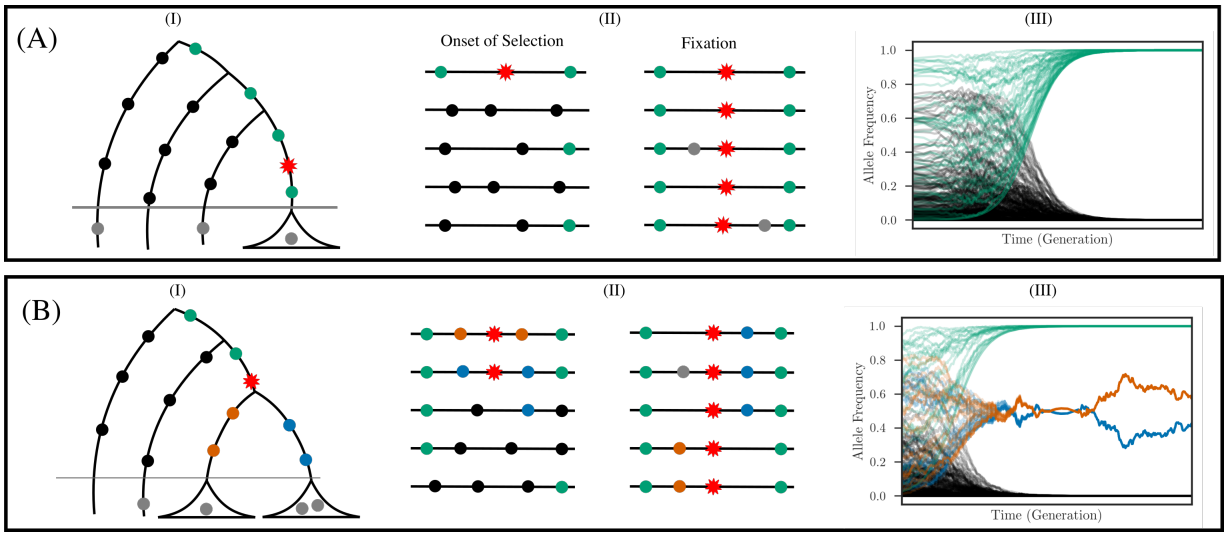

Figure 3.2: **Models of positive selection.**
(A) Hard seep, in which all the carriers of the beneficial allele coalesce before the onset of selection (I), leading to fixation of a single haplotype in the population (II) and elimination of all the genetic variation (III). (B) In a soft sweep due to standing variation, carriers coalesce before the onset of selection (I) and more than one haplotype exist at fixation (II). Vitiation exclusive to each beneficial haplotype forms a cluster that is observable in time-series frequency data (III).

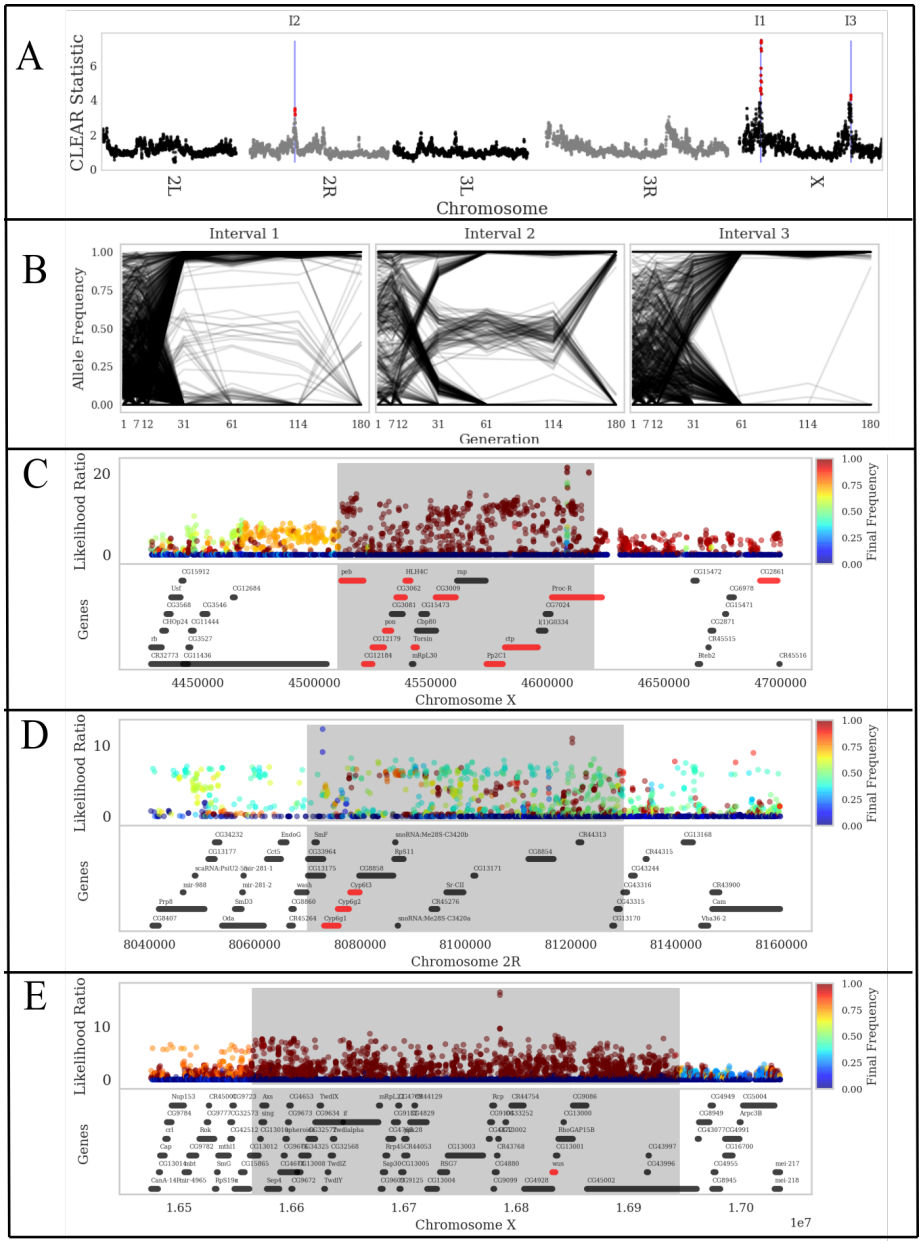

Figure 3.3: **Genomic scan of replicable signals of selection.**
(A) Manhattan plot of the CLEAR statistic. 3 intervals pass the significance level. (B) Trajectories of the mutations provide evidence for hard, soft and hard sweep for intervals I1,2,3, respectively. (C,D,E) Likelihood ratio statistic for each variant draws the distribution of the responding variants to the selection pressure for intervals I1,2,3 respectively. Color of each mutation corresponds to the final frequency at fixation (generation 61).

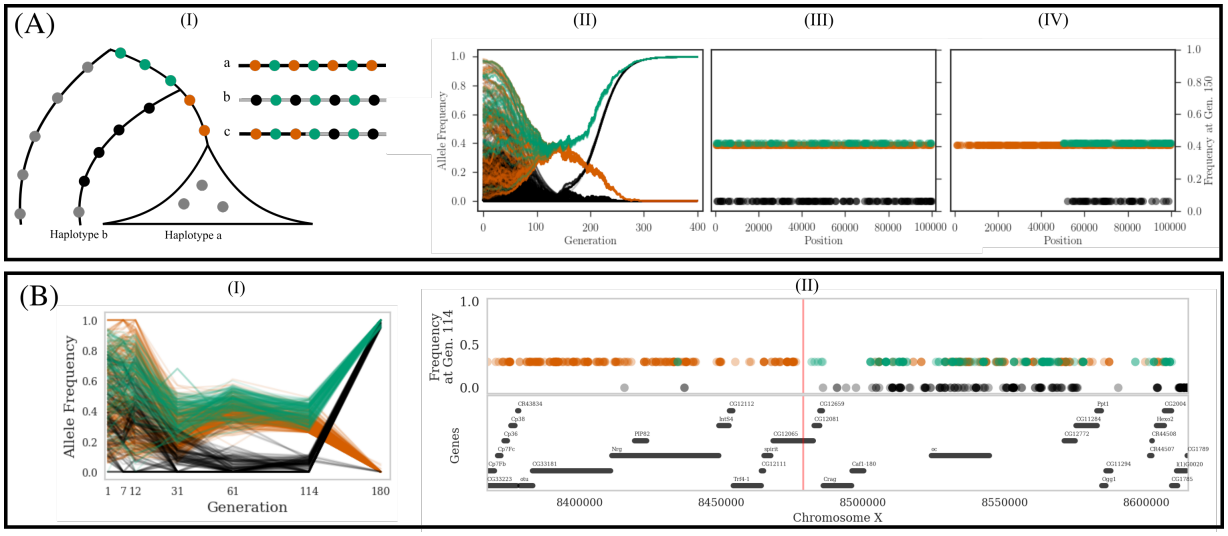

Figure 3.4: **Sudden change of fitness of a haplotype in reduced diversity populations.**
(A) In a reduce diversity population, dominant haplotype(s) represent itself via a cluster of mutations (green and orange in (II)). If in such a population change of fitness of a haplotype are noticeable time-series frequency data, especially when a low frequency haplotype (II) fixates in the population. While change of fitness due to de novo variation and recombination create indistinguishable signals in frequency data, mapping clusters of mutations to genome, specifies the mechanism of action: that could be via mutation (III) or recombination (IV). (B) A 250Kb interval on chromosome X of replicate 1 exhibit evidence of epistatic recombination.

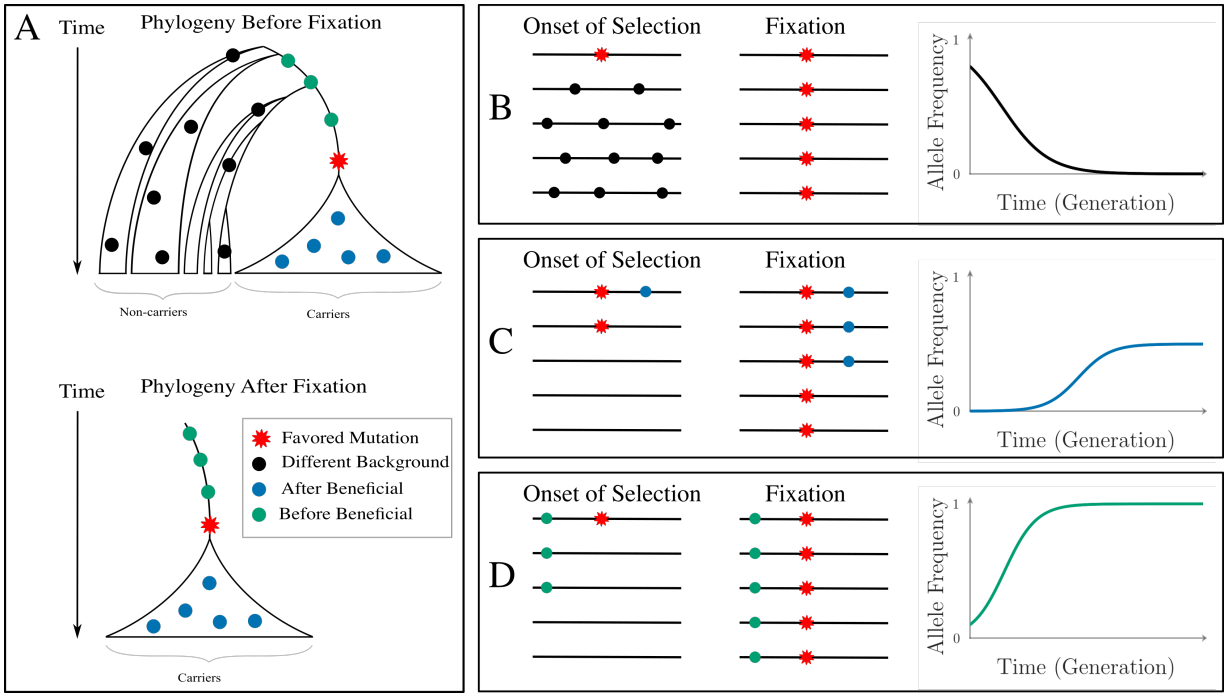

Figure 3.5: **Fate of the linked mutations.**

Chapter 4

Analyzing Human Ethnic Population for Selection on Disease Susceptibility

The Central Asian Kyrgyz highland population provides a unique opportunity to address genetic diversity and understand the genetic mechanisms underlying high altitude pulmonary hypertension (HAPH). While a significant fraction of the population is unaffected, there are susceptible individuals who display HAPH in the absence of any lung, cardiac or hematologic disease. We report herein the analysis of the whole genome sequencing of healthy individuals compared with HAPH patients and other controls (total n=33). Genome scans reveal selection signals in various regions, encompassing multiple genes from the first whole genome sequences focusing on HAPH. We show here evidence of three candidate genes MTMR4, TMOD3 and VCAM1 that are functionally associated with well-known molecular and pathophysiological processes and which likely lead to HAPH in this population. These processes are a) dysfunctional BMP-signaling, b) disrupted tissue repair processes and c) abnormal endothelial cell function. Whole genome sequence of well characterized patients and controls and using multiple statistical tools uncovered novel candidate genes that belong to pathways central to the pathogenesis of HAPH. These studies on high altitude human populations are pertinent to the understanding of

sea level diseases involving hypoxia as a main element of their pathophysiology.

4.1 Introduction

Pulmonary hypertension (PH) is a condition with an abnormally high blood pressure in the pulmonary arteries due to arterial resistance to the pulmonary blood flow. This may be due to a variety of causes and combination of factors such as endothelial dysfunction, vasoconstriction of small pulmonary arteries and endothelial and smooth muscle cells proliferation [113]. Clinically, PH is categorized into five groups ranging from idiopathic, familial/heritable, to presenting as a secondary disease e.g., congenital heart disease with left-to-right shunt, chronic obstructive pulmonary disease (COPD), chronic thromboembolic pulmonary disease and chronic exposure to high altitude (HA) hypoxia or high altitude pulmonary hypertension (HAPH) [113]. Ernst von Romberg, a German physician, described pulmonary vascular sclerosis, as far back as 1891. Despite the recognition of this disorder more than a century ago, some form of PH, e.g., pulmonary arterial hypertension (PAH), has no known cure. The available treatments are only to relieve the symptoms and slow the progress of the disease. Fortunately, through these efforts, the 3-year mortality rate has decreased over the past two decades to $\leq 30\%$ [113]. Additionally, a recent study using whole genome sequencing has established the rate of gene mutation to be about 24% of PAH cases [114] providing additional therapeutic targets for the treatment of this disorder. In countries like the Kyrgyz republic where 90% of the area is at altitude $\geq$3000m (Fig. 1A and 1B), HAPH is a public health issue [115]. Previous studies on Kyrgyz highlanders have reported 14-20% of the population to have signs of HAPH [115]. This was based on signs of cor pulmonale in high altitude dwellers with HAPH or mean pulmonary pressures (mPAP) of $\geq$ 25mmHg [115]. Remarkably, an additional 21% manifest a $\geq$ 2 fold increase in mPAP on exposure to acute hypoxia, i.e. 30 mins of 11% oxygen breathing [115] even though they had normal resting mPAP of $\leq$ 25 mmHg (Hyper-responder, HR). This high prevalence rate of HAPH in Kyrgyz highlanders

provides an opportunity to understand the biology of susceptibility to HAPH or mal-adaptation to HA. In general, HA studies provide one of the best natural experiments in humans to study gene versus environment interactions [116]. Since hypoxia is involved in the pathophysiology of many diseases including PH, it also provides a unique opportunity to identify the genes involved in hypoxia regulation which can be explored for therapeutic purposes. With this insight, numerous studies, including some from our group, were conducted to study human adaptation/mal-adaptation to HA in Ethiopians, Tibetans and Andeans, the three major highland populations [10, 16, 68, 117, 118]. Using whole genome sequencing, we were able to identify and functionally validate novel genes involved in altitude adaptation in these populations [10, 16, 117, 119, 120]. When describing adaptation/mal-adaptation to HA in Kyrgyz highlanders, it is important to distinguish this unique population from the rest of the HA populations, i.e., Ethiopians, Tibetans and Andeans. Studies conducted on the other three HA populations successfully identified different physiological and genetic modes of adaptation, some common across populations [121] and others exclusive to one population [117]. However, very few studies have been conducted on Kyrgyz highlanders where the studies are either focused on single-gene/single-variant i.e., Angiotensin Converting Enzyme Insertion/Deletion polymorphism [115], or on using whole exome sequencing [122]. It is important to note that the Kyrgyz population offers a unique advantage for the study of HAPH as compared to other HA populations. For example, the Kyrgyz subjects do not present with any other feature of chronic hypoxia besides HAPH, unlike the Andean population. Indeed, the Andean population, where HAPH was first described [123], does show other potentially confounding characteristics such as polycythemia which, in turn, can lead to PH. In the current study we wanted to test the hypothesis that, due to genetic selection, healthy Kyrgyz highlanders are biologically protected from developing HAPH. Towards that end, we sequenced and analyzed the whole genome of Kyrgyz highlanders, in a case-control study design, and identified novel genes involved in HAPH. In addition, the candidate genes identified may also be related to PH in general as there is no study ever attempted to identify novel genes for

HAPH using whole genome sequence analysis.

4.2 Methods for detecting selection in ethnic populations

We included volunteers from a Kyrgyz highlander population. All individuals gave informed consent. A detailed phenotypic characterization of the cohort, including right heart catheterization studies are reported elsewhere [115, 124]. In brief, the cohort consisted of four groups which included a) high altitude pulmonary hypertension (HAPH, n=9) b) controls (No-HAPH, n=9) consisting of healthy highlanders that age-matched to the HAPH group, c) hyper-responders (HR, n=7), consisting of relatively younger individuals showing a significant increase in their pulmonary artery pressure (PAP, mmHg) when exposed to 11% oxygen, and d) normal-responders (NR, n=9), were the control group (age matched to HR) not hyper-responding to lower oxygen. All the subjects were carefully analyzed by an expert person. The study was approved by an institutional review board. Library construction and sequencing Blood sample (10mL) was collected from each subject for DNA extraction. The whole genome sequencing was carried out at HLI (Human Longevity Inc. San Diego). Library preparation was carried out using the TruSeq Nano DNA HT kit (Illumina Inc.). Manufacturer's instruction was strictly followed at all steps. Whole genome sequencing were done at a mean coverage of 40.3x on the Illumina HiSeqX sequencer utilizing a 150 base paired-end single index read format. The additional details of the library construction and the quality control are described in the Supplementary Material. The sequencing data from this study have been submitted to the European Genome-phenome Archive (EGA; https://www.ebi.ac.uk/ega/datasets/) under accession number EGAD00001004285. Admixture analysis and population structure ADMIXTURE [125] was used to measure the genetic affinity of the Kyrgyz individuals with other major populations from the 1000 genome project [126]. To calculate ancestry proportions, we first pooled Kyrgyz samples with individuals from 1000 Genome project populations, including YRI, CEU, PJL (SAS), and JPT. Then we filtered

out variants with allele frequency of less than 0.05 and ran ADMIXTURE program with the parameter K=3. Additional details on population structure is described in the Supplementary Material. Selection scan A detailed explanation on selection scan is provided in the Supplementary Material. Briefly, we computed the test statistics using a sliding window of 50Kb, with steps of 10Kb, over the autosomal genome. D [127] and H [128] statistics were computed on a site-frequency-spectrum of a 50Kb window and did not require any post-processing. However, in iHS, nSL, PBS and XP-EHH scans where the statistic was computed for a single variant, we computed the average of the scores for each window. Subsequently, for each method, we took genome-wide top 0.1 percentile to be our significance cutoff, and merged overlapping significant windows to obtain distinct set of significant intervals for each scan. Prioritization of the selected intervals To prioritize genomic intervals of different lengths and levels of variation, we assigned a p-value to each of the 71 selected intervals, using a case-control style association. We then used False Discovery Rate (FDR) [129] of 10% to choose a significance cutoff. Under the null model, we expect that allele frequencies of Case and Control groups to be similar. Additionally, because Case and Control individuals belong to ethnic and local populations, our null model posits that allele frequencies of the Control population is closer to the allele frequencies of the Case population than that of the outgroup population, a genetically distant population. We use three populations: case (Healthy), control (Sick), and outgroup (JPT), as defined previously. We use the PBS statistic, which uses the length of the case-lineage with Fst as a measure of genomic distance [9]. Specifically, PBS = -log[1 - Fst(case, control)] - log[1 - Fst(case, outgroup)] + log[1 - Fst(control, outgroup)]. We used all variants in the selected region to compute Fst , defined by Weir et al [130]. For each selected interval, we computed a windowed PBS statistic and then computed its significance by testing against an empirical null distribution of that interval. We calculated the empirical null distribution by permuting samples of case and control populations (1000 times) and keeping the outgroup population fixed. Haplotyping and other pre-processing In order to evaluate iHS, nSL and XP-EHH statistics for the Kyrgyz dataset, we first inferred

the phased haplotypes. We used Beagle v4.1 program [131] where a reference panel of EAS population from 1000 Genome project is provided as an input parameter. We computed the frequency of each haplotype in a population as the statistical mode of the frequencies of the SNPs carried by the haplotype. Also, to identify the derived allele, we used the Ensembl Compara 59 database [132], which has inferred the ancestral allele on 6 primates. eQTL analysis Expression quantitative trait locus (eQTL) for different tissues were obtained from GTEx database [133]. We limited our eQTL analysis to tissues that were in the GTEx data and were directly related to PH. The list of eQTL SNPs with significant P values ($P \leq 0.05$) were obtained for each candidate genes. The sample size for related tissues reported in the current study are Lung (n=383) and whole blood (n=369).

4.3 Results

To our knowledge, this is the first whole genome sequence study performed on a Central Asian population, Kyrgyz highlanders (Fig. 1A and 1B). Additionally, we used a case-control study design, comparing HAPH to healthy controls. The subjects in these cohorts have been very well phenotyped (Fig. 1C) and cardiac catheterization has been done on all subjects [115]. Since this is a study of only 33 subjects, a question could be asked about a potential bias since the sample size is relatively small. While large-scale association studies of urban populations could potentially provide means for determining genetic architecture of common complex traits, studying of locally adapted ethnic populations can be used to increase statistical power and target less common phenotypes [9, 10, 116, 134–136]. This paradigm reduces the number of loci for association from millions to tens or thousands of genetic loci. We note that while our sample size is limited, we have access to whole genome sequence covering all variants (40.3x coverage) to perform selection scan on 66 haplotypes to reduce the number of loci for statistical test for association. By utilizing a pipeline that identifies regions under positive selection through an

analysis of all variants in large (≥50kb) segments, and an integrated statistical test, we have been successful in the past identifying candidate genes involved in HA adaptations both in Ethiopian and Andean populations and these were subsequently validated [10, 117, 119, 120]. We use a similar approach here, with some modifications (Methods). Ancestry and Population Structure of Kyrgyz Individuals To identify the genetic history of the Kyrgyz population and to find a close outgroup population for selection scan, we performed Admixture and principal component analysis (PCA). Chromosome-wide admixture analysis of Kyrgyz samples, along with samples from African (YRI, n=108), European (CEU, n=99), South Asian (PJL, n=96) and East Asian (JPT, n=104) (Fig. 2A), demonstrated that Kyrgyz highlanders (KGZ, n=33) have a strong East Asian component (=0.76, =0.09 for the distribution of the proportion of the East Asian Ancestry in Kyrgyz samples, where is mean and is standard deviation) along with some European ancestry (=0.24, =0.09 for the distribution of the proportion of the European Ancestry in Kyrgyz samples). As East Asian ancestry is significantly higher in the Kyrgyz samples (two sample Kolmogorov-Smirnov P=0.00015), we used JPT as reference population in our current study because we and others have found that this population was relatively closer to Kyrgyz than the other East Asian reference population, i.e., Han Chinese [137] (Supplementary Fig. S1). The admixture from East and West is consistent with the mitochondrial DNA ancestry analysis of the Kyrgyz highlanders [138]. No sign of shared ancestry was detected with YRI (two sample Kolmogorov-Smirnov P=3.98E-20). Similar to KGZ, PJL showed European and East Asian ancestry, but with inverse proportions that were relatively inverted (=0.82, =0.18 for European and East Asian ancestry, respectively; Fig. 2A). Interestingly, despite the prehistoric and historic eastward migrations, the JPT proportion of admixture dominated both the mitochondrial/maternal lineage [138] as well as the nuclear DNA lineage (Fig. 2A). We then performed PCA on the Kyrgyz sample along with 1000 Genome project super-populations AFR (n=661), EAS (n=504), EUR (n=503) and SAS (n=489), and observed that the Kyrgyz individuals were located between EAS and SAS super-populations (Fig. 2B). A previous study on the population structure and

genetic ancestry of Central Asia which also included the Kyrgyz population has revealed similar findings [137]. Finally, we tested if there is a population structure within Kyrgyz populations that is correlated with phenotypic groups. To do this, we computed the PCA of the Kyrgyz samples and stratified the PCA projection by four phenotypic groups: No-HAPH, HAPH, HR and NR. As shown in Fig. 2C, the Kyrgyz subgroups do not reveal any significant substructure (one-way ANOVA P=0.3, 0.29 for the first and second principal components, respectively). Selected Intervals We computed six different statistics that captured regions under selection over sliding windows of 50Kbp with steps of 10Kbp over each autosomal genome (Supplementary Fig. 2 and Methods). The phased and unphased dataset contain 6,841,212 and 11,703,698 variants, respectively, resulting in 284,906 overlapping 50Kbp windows. For each test statistic, we identified windows that scored in the top 0.1 percentile (top 285) among all windows. We then merged the identified windows from the six different methods. This resulted in 71 distinct intervals. Next, we computed empirical p-values for each interval (Methods) and used 10% FDR cut-off [129] to identify top intervals (Table 1) for further analysis. The most significant interval, interval 1 (Fig. 3A and 3B), located on chromosome 17, contains 147 SNPs with the haplotype extending to 892Kbp. The haplotype frequency difference between the cases and the controls was 60% (Fig. 3B and Methods). The region contained 10 genes (Table 1, Fig. 3A). In order to make biological sense of the candidate interval, we performed an extensive literature survey looking for all possible aspects that would connect this interval to HAPH. Accumulating evidence suggests MTMR4 (myotubularin-related protein 4, highlighted in Fig. 3A) as one of the biologically plausible candidate gene of HAPH. It contains tyrosine/dual-specificity phosphatase activity and is known to dephosphorylate SMAD1/2/3 [139, 140]. Previous studies have shown that transforming growth factor (TGF)/Smad2/3 signaling is disrupted in a monocrotaline based rodent model of pulmonary arterial hypertension (PAH) [141]. Similarly, the disruption of bone morphogenetic proteins (BMPs) signaling, also a member of TGF superfamily, is a known factor that initiate PH [139]. Studies have also shown that MTMR4 is an essential negative

regulator of BMP signaling pathway [140]. Unlike TGF/Smad2/3 signaling, here it binds and dephosphorylates the activated Smad1 to attenuate BMP signaling [139, 140]. This activity was indeed confirmed by overexpressing MTMR4 that led to repressed BMP-induced gene expression, and by MTMR4 specific siRNA enhancing BMP signaling [140]. Therefore genetic selection involving selective inhibition of this gene would lead to an enhancement of both TGF and BMP signaling which may protect No-HAPH from developing PH under chronic environmental hypoxia. Additionally, the haplotype interval also consists of two missense mutations located in HSF5 (Heat Shock Transcription Factor 5, rs3803752 and rs117817367). Not much is reported on HSF5, but being from the HSF family and having transcription factor activity [142], its role in gene regulation events can be explored. We also performed eQTL analysis for these SNPs with gene expression levels in the tissues related to cardio-respiratory system. Interestingly, a large number of these SNPs i.e., from interval 1, were identified as eQTLs where the ancestral allele was significantly associated with lower RAD51C levels in the lungs (Fig. 3C; Supplementary Fig. S3 and Supplementary Table S1). A higher frequency of derived allele among the healthy Kyrgyz highlanders can be correlated with a higher RAD51C expression in the lungs. Previous studies have shown that an increased RAD51C expression is associated with lung cancer [143]. Given that immune and inflammatory processes triggered by cancer cells can also lead to PH [144], it is intriguing to discover a significantly higher frequency of derived alleles in the No-HAPH group as compared to all the other populations (Fig. 3B). The second significant interval (Table 1 and Fig. 4A) is located on chromosome 15 (position 52007217-52268916, hg19) spanning 315 Kbp. It overlays five genes i.e., SCG3, LYSMD2, TMOD2, TMOD3 and LEO1 (Fig. 4A and Table 1). The haplotype consisted of 61 SNPs differing significantly (P0.05) between the No-HAPH and controls (Fig. 4B). All the SNPs were in the non-coding regions and two SNPs located in the promoter region of TMOD3 (rs11637876 and rs12913583) had CADDs PHRED value of 17.2 and 14.2 respectively. Large number of SNPs, which also included rs11637876 and rs12913583, were identified as eQTLs linked to TMOD3 expression (Fig. 4C and 4D; Supplementary Fig.

S3) in specific tissues associated with PH e.g., whole blood (Fig. 4C; P= 5.5e-7), aorta (Fig. 4D; P=1.0e-7). When we explored the biological significance of this gene in term of HAPH, we found that the expression of TMOD3 is increased in patients with idiopathic PAH [145]. This gene is expressed in the motile endothelial cells where it caps the pointed ends of actin filaments [146]. The capping inversely correlates with endothelial cell migration rates and the overexpression of Tmod3 inhibits cell migration [147]. Keeping in mind the importance of cell migration in tissue repair responses, e.g., vascular repair and regeneration in PH [147], a higher expression of TMOD3 in the PH patients [145] may be linked to TMOD3-related delay in cell migration for tissue repair process. A complete knockout of this gene in mice is lethal due to embryonic anemia and defects in fetal erythropoiesis [148] and therefore further studies involving tissue specific differential expression will be needed to specify its role in PH pathogenesis. The involvement of TMOD2, also from the same tropomodulin family, in HAPH is less likely because of its restricted expression in neural tissues [149]. Other candidate intervals: The third top interval consists of 7 genes. Interestingly, SIGLEC11 (Sialic Acid Binding Ig Like Lectin 11) which mediates anti-inflammatory and immunosuppressive signaling was previously found to be associated with HAPH [122]. There was also ample evidence that VCAM1 (vascular cell adhesion molecule 1) from interval 4 (Table 1) has a role in HAPH. The gene is a member of the immunoglobulin superfamily and is known to be a marker of endothelial cell inflammation, and mediate adhesion of white blood cells to the endothelium. VCAM1 has three different splice variants and the transcription levels when measured in different cell lines with ENCODE reveals higher expressions in the Homo sapiens endothelial of umbilical vein primary cell (HUVEC) and Homo sapiens skeletal muscle myoblast primary (HSMM) cells (Supplementary Fig. S4). Interestingly, the H3K4me1 and H3K27Ac tag densities which often found near regulatory elements are higher only in the HUVEC cells and the peaks also align with few of the selected SNPs (Supplementary Fig. S4). This could indicate differential regulation of the selected SNPs specific to endothelial cells. Higher levels of VCAM1 are associated with renal dysfunction,

hepatic impairment and more importantly correlated with the severity of PH in patients with sickle cell disease [150]. The levels were also found high in peripheral blood [150] and lung fibroblasts [151] of patients suffering from idiopathic pulmonary fibrosis. The study also shows that TGF-1 treatment leads to an increase in VCAM1 levels at transcriptional as well as protein levels while silencing VCAM1 expression inhibits fibroblast proliferation [151] and the role of PH in idiopathic pulmonary fibrosis etiology is well recognized. Furthermore, its role in allergic lung diseases [152] or in systemic sclerosis [153], where PAH is the leading cause of death [154], clearly indicates evidence of a positive selection sweep on genetic variants that would protect the No-HAPH subjects from developing PH. Furthermore, studies in mice with LPS-induced acute lung injury have clearly shown an upregulation of VCAM1 [155]. Similarly, transcriptome analysis of patients diagnosed with high altitude pulmonary edema (HAPE), also an acute lung injury, depicts maximum fold change for VCAM1 along with MAPK10 [156]. This upregulation in HAPE patients provides a direct link between VCAM-1 and HAPH because PH is a hallmark in patients with HAPE. In-vitro analysis modeling PAH by CypA-induced oxidative stress also revealed increased VCAM1 [157]. Overall, our analysis of whole genome sequence clearly indicates that there are multiple genomic regions under selection (Table 1). This is also supported by a recent whole exome analysis of the same population where they have identified 33 candidate genes linked to HAPH [122]. Interestingly, only one gene out of these 33 candidates was among the top candidate interval in the current study. These differences could be due to the input data and computational pipelines which are completely different in the two studies. For example, a stronger genetic selection in the regulatory regions such as promoters and enhancers, which the whole exome sequencing fails to capture. In our previous studies on selection scan at HA, we observed similar signals i.e., variants in the non-coding region, which were subsequently validated in different model systems [10, 116, 117, 119, 120]. Additionally, because of weaker signals of selection, other genes may not have passed multiple scans of selection and the 3-way association filtering criteria that we applied in the current analysis. The genes like DST, SDK1 and OSBPL9

from the previous study were part of our initial gene list but were subsequently dropped out as they failed the subsequent stringent PBS association filter. In order to recapitulate previous findings in our study, we took the specific variants (chromosome position and rsID#) from the previous study [122] and systematically scan all the 33 signals in our current subjects/samples. In all cases (including SIGLEC11), we found that the frequency of heterozygotes was at most 3/33 (Supplementary Table 2) and did not meet the criterion established by the previous study [122]. Finally, there were additional known genes located in different intervals where the haplotype frequencies were significantly different between No-HAPH vs HAPH comparisons (Empirical Fst, $P \leq 0.05$) but not significant when compared with the outgroup population, such as with the gene EDNRB (Supplementary Fig. S5). Even though the difference remains significant (Empirical Fst, $P \leq 0.05$) when No-HAPH is compared to other major populations, i.e., AFR, EUR and SAS, it is not consistent with our local adaptation model, i.e., natural selection is specific to the adapted (No-HAPH) populations environment. A simple explanation could be an involvement of other selection pressures on this region in the JPT, totally unrelated to HAPH that may have impacted the genomic structure in this region among JPT to behave in a manner similar to our No-HAPH group.

4.4 Conclusions

Our study provides the first whole genome sequence analysis of a Central Asian population from the Kyrgyz highland and also for HAPH with prospect of identifying novel genes for this disease. Admixture analysis revealed that the Kyrgyz highlanders consisted of both European and East Asian ancestry with an overwhelming contribution from the East Asian gene pool. Evolutionarily, the typical form of HAPH is distinctive to Kyrgyz highlanders and in the current study, we utilized this feature and discover novel genetic markers with a potential to give insight into therapy for PH. From the top candidate genes detected, we raise here three possibilities

that may individually or together lead to HAPH in Kyrgyz highlanders. This includes, first, a dysfunctional BMP signaling involving MTMR4 overexpression as an alternate gene regulating BMP signaling. Second, TMOD3-related delayed cell migration for tissue repair process. And third, an abnormal endothelial cell function with elevated VCAM1 (a schema is presented in Supplementary Fig. S6). However, despite an unbiased identification approach with ample literature evidence, we cannot rule out the involvement of additional genes or gene interactions in HAPH. Future studies targeting these genes may strengthen our findings and will provide a better understanding of HAPH.

4.4.1 Acknowledgments

Chapter 4, in full, contains material from Arya Iranmehr, Tsering Stobdan, Dan Zhou, Orit Poulsen, Kingman P. Strohl, Almaz Aldashev, Amalio Telenti, Emily HM Wong, Ewen F Kirkness, J Craig Venter, Vineet Bafna, Gabriel G Haddad. "Novel insight into the genetic basis of high-altitude pulmonary hypertension in Kyrgyz highlanders. European Journal of Human Genetics". 2019 Jan;27(1):150 [2]. I was the primary investigator and author of this paper.

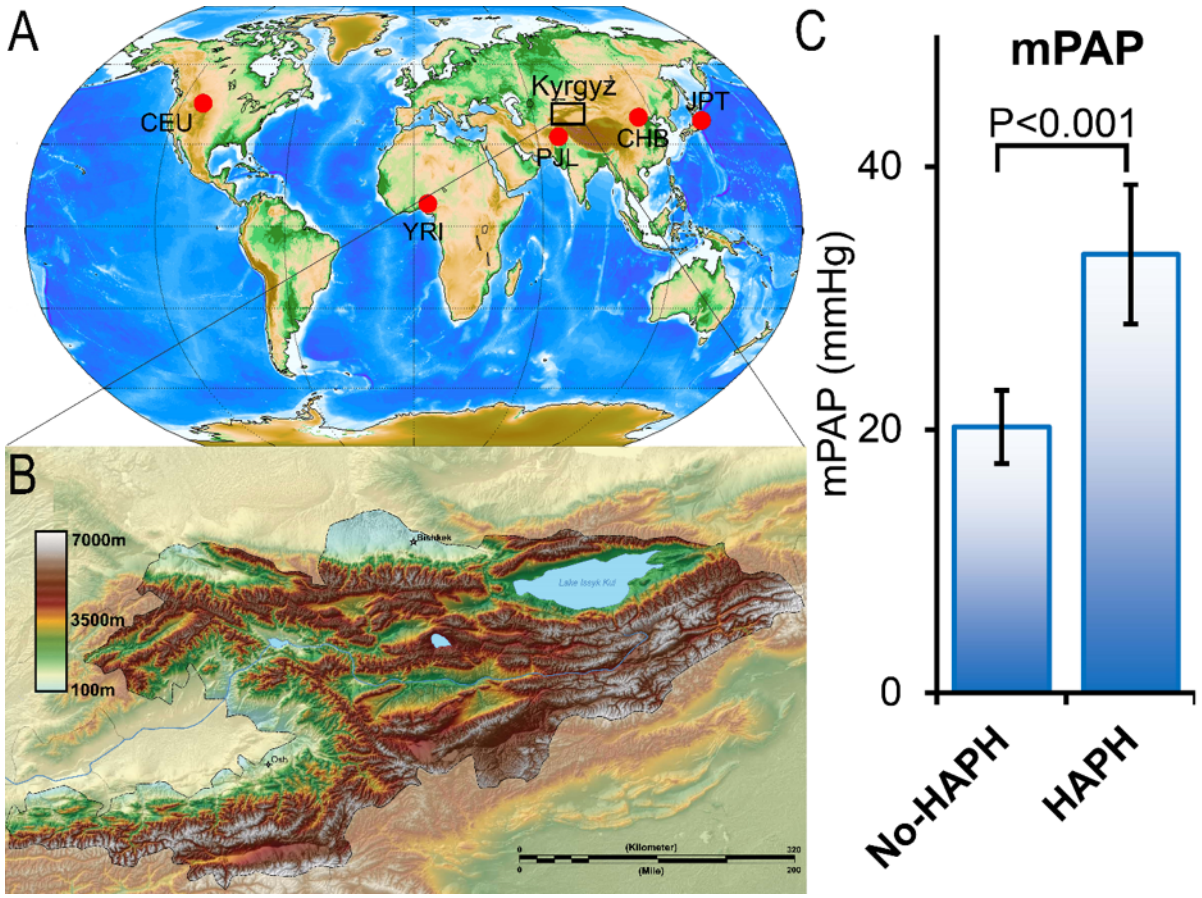

Figure 4.1: **Geographic location of Kyrgyz population.**
Geographic location of Kyrgyz population (box) relative to other major populations from the 1000 genome project used in the Admixture analysis of current study (A). YRI, Yoruba in Ibadan, Nigeria; CEU, Utah Residents with Northern and Western European Ancestry; JPT, Japanese in Tokyo, Japan; CHB, Han Chinese in Beijing, China; SAS (PJL), Punjabi from Lahore, Pakistan. (B) Topography of Kyrgyz republic with $\geq$90% of the area at altitude $\geq$ 3000m above sea level. (C) Mean pulmonary artery pressure (mPAP) in healthy Kyrgyz highlanders (No-HAPH) is significantly lower than the age-matched HAPH patients. Error bar represents standard error.

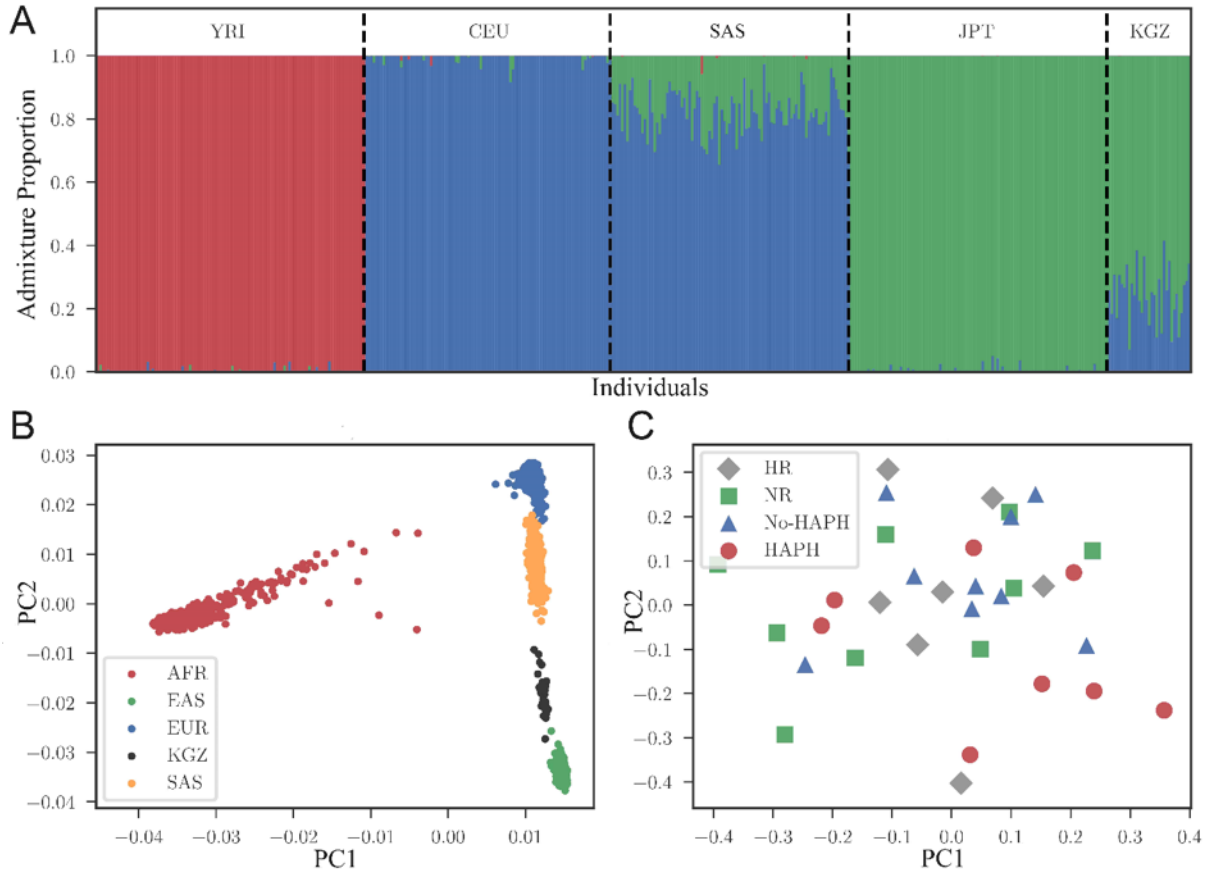

Figure 4.2: **Admixture and PCA of Kyrgyz Population.**
Admixture and PCA depicting the genetic relatedness of Kyrgyz Population (KGZ), which includes HAPH, No-HAPH, HR and NR groups, to other major populations. (A) Admixture analysis shows that Kyrgyz population consists of major genetic proportion from East Asian lineage with minute contributions from the European genetic ancestry. (B) PCA reveals that the Kyrgyz population is located between SAS and EAS but more closely related to EAS. (C) Within Kyrgyz cluster the subjects are randomly distributed. SAS, South Asian; EAS, East Asian.

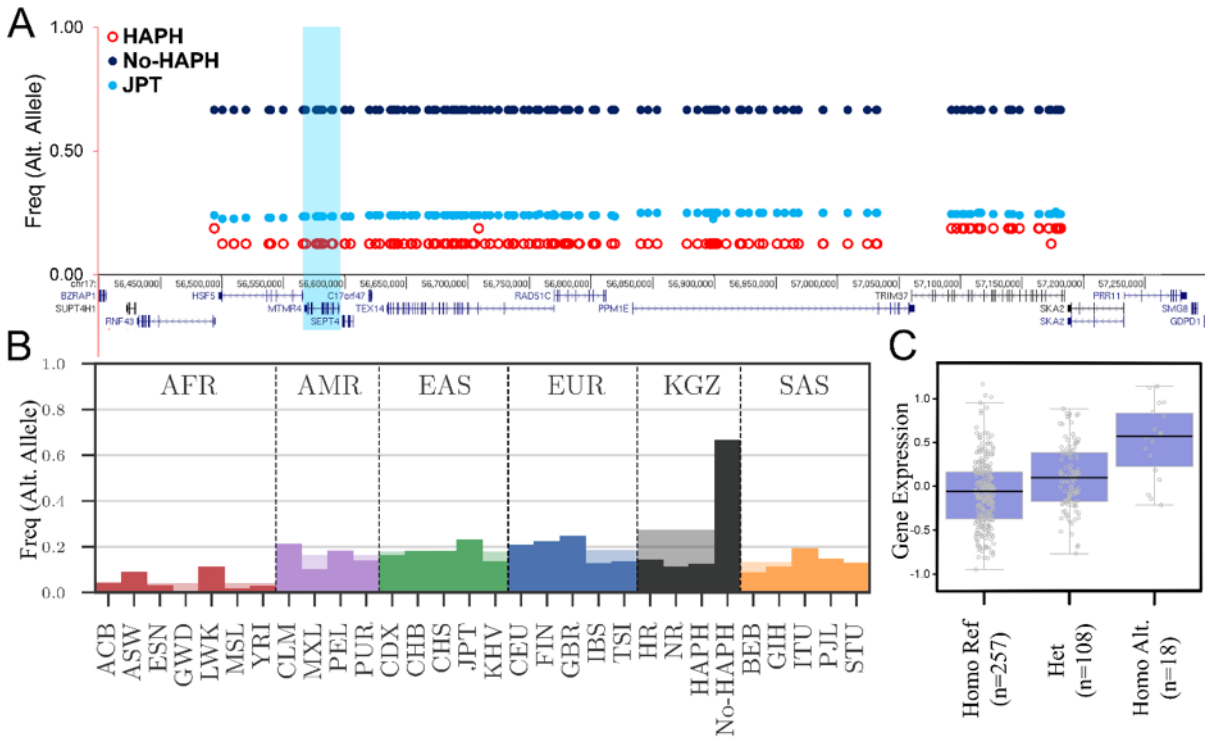

Figure 4.3: **Target region of selection in Kyrgyz population, Interval 1.**
Layout of genetic variation in the top selected interval-1 in the HAPH, No-HAPH and JPT (outgroup) populations (A). The haplotype frequencies among No-HAPH is higher compared to HAPH and JPT. (B) Frequency of the top selected haplotype (interval 1) among Kyrgyz highlanders and populations from the 1000 Genome Project. The y-axis is frequency of one of the SNPs (out of 147 fully linked SNPs) of the selected haplotype. (C) A representative box plot showing the genotype of an eQTL SNP in interval-1 and the respective expression of gene in Lung (P = 1.5e-11).

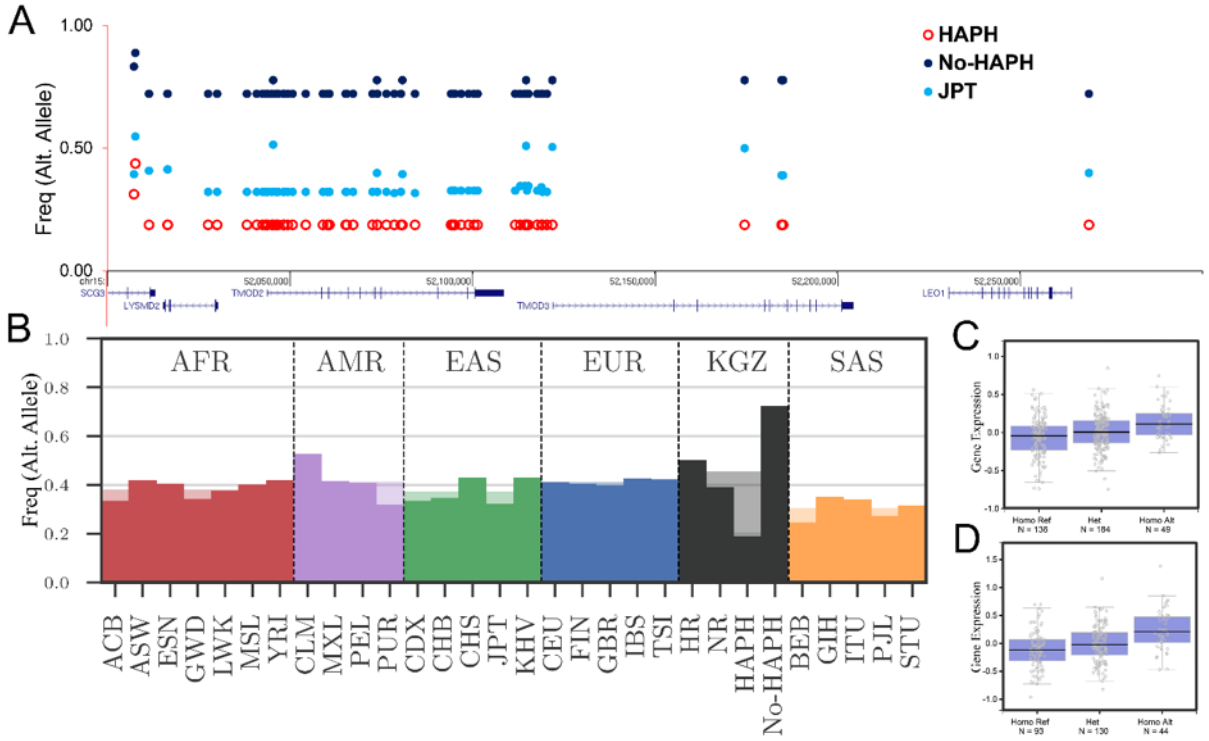

Figure 4.4: **Target region of selection in Kyrgyz population, Interval 2.**
Layout of genetic variation in selected interval-2 in the HAPH, No-HAPH and JPT (outgroup) populations (A). Haplotype frequencies among No-HAPH are higher compared to HAPH and JPT. (B) Frequency of a SNP from a set of perfectly linked SNP of interval-2 among Kyrgyz highlanders and populations from the 1000 Genome Project (B). Box plot of SNPs, rs11637876 (C) and rs12913583 (D) from interval-2 that was identified as eQTL for the expression of TMOD3 (P = 5.5e-7 for both SNPs).

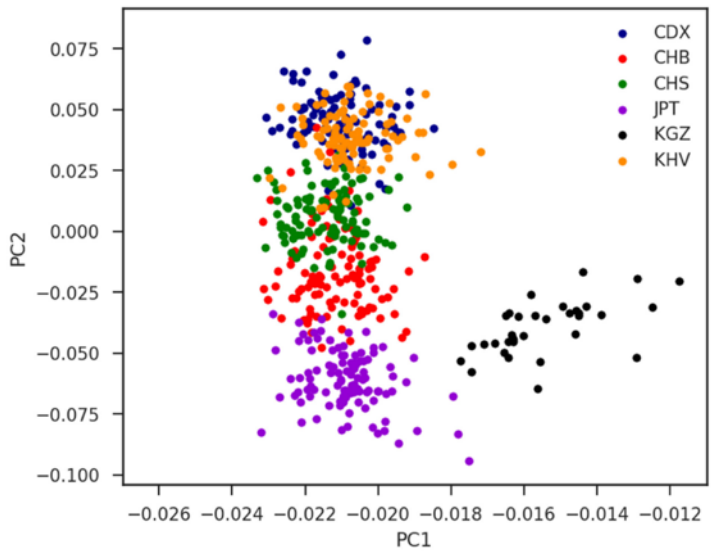

Figure 4.5: **PCA lay out of Kyrgyz population along with 1000 genome sub-populations**
Representation of the Kyrgyz (KGZ) population along with East Asian (EAS) populations i.e., Han Chinese in Beijing from China (CHB), Southern Han Chinese from China (CHS), Chinese Dai in Xishuangbanna from China (CDX), Japanese in Tokyo from Japan (JPT) and Kinh in Ho Chi Minh City from Vietnam (KHV), of the 1000 Genome Project. We chose JPT population in our current study as the outgroup population throughout this paper, as its center is relatively closer to the Kyrgyz population than other EAS populations.

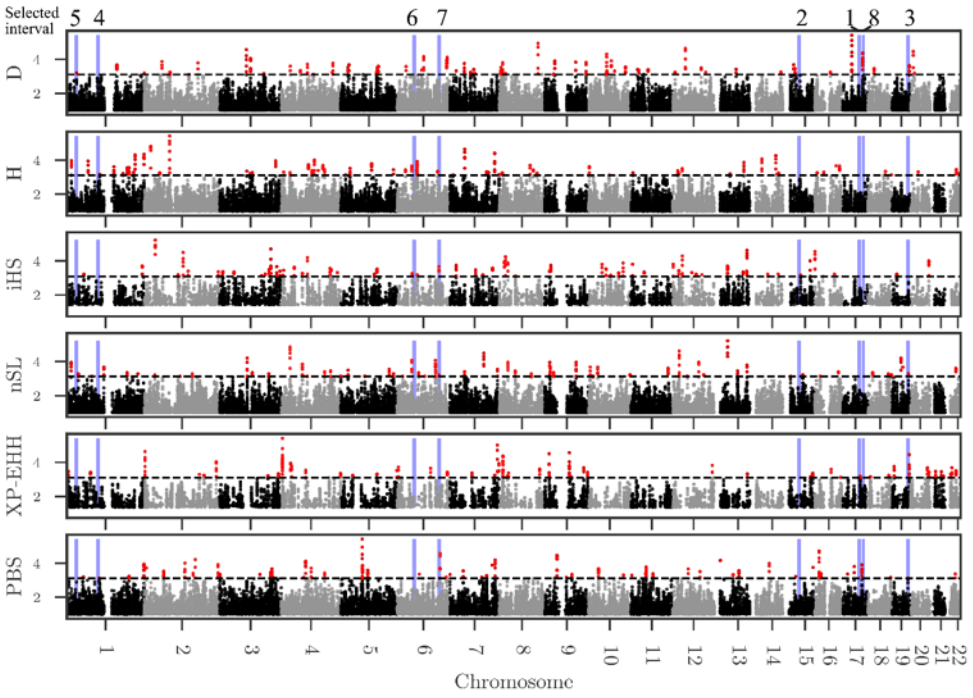

Figure 4.6: **Genome-wide scans of selection**
Genome-wide scan to detect genomic regions under selection. Different statistical tests utilized are indicated on Y-axis. The top selected intervals are highlighted and the ranks are accordingly labelled.

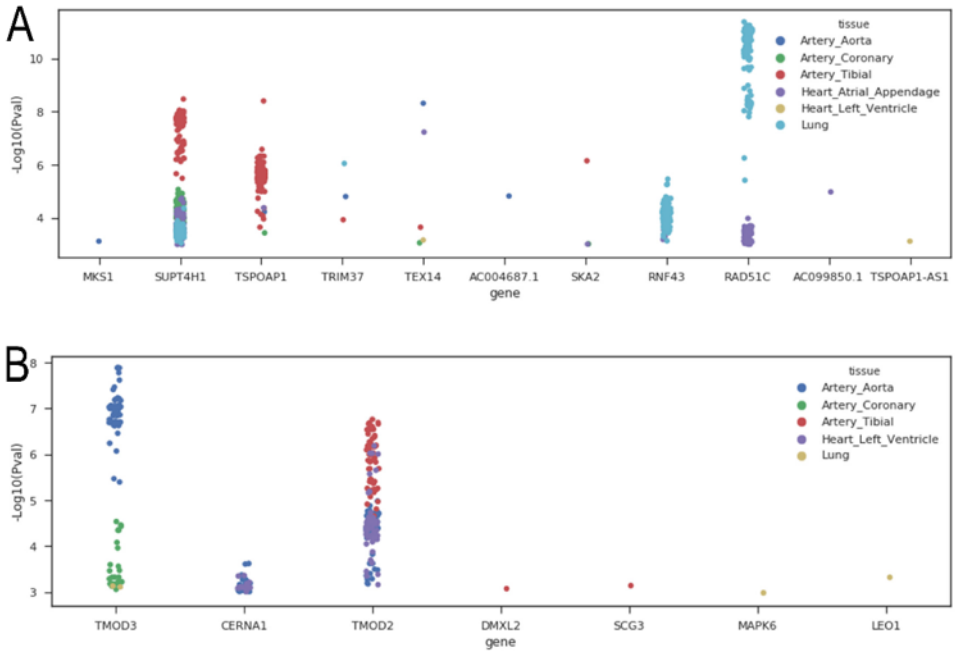

Figure 4.7: **eQTL analysis of the SNPs of target of selection.**
eQTL analysis of the SNPs in interval-1 (Top) and interval-2 (below). X-axis depicts that P-values of the genotype vs gene expression.

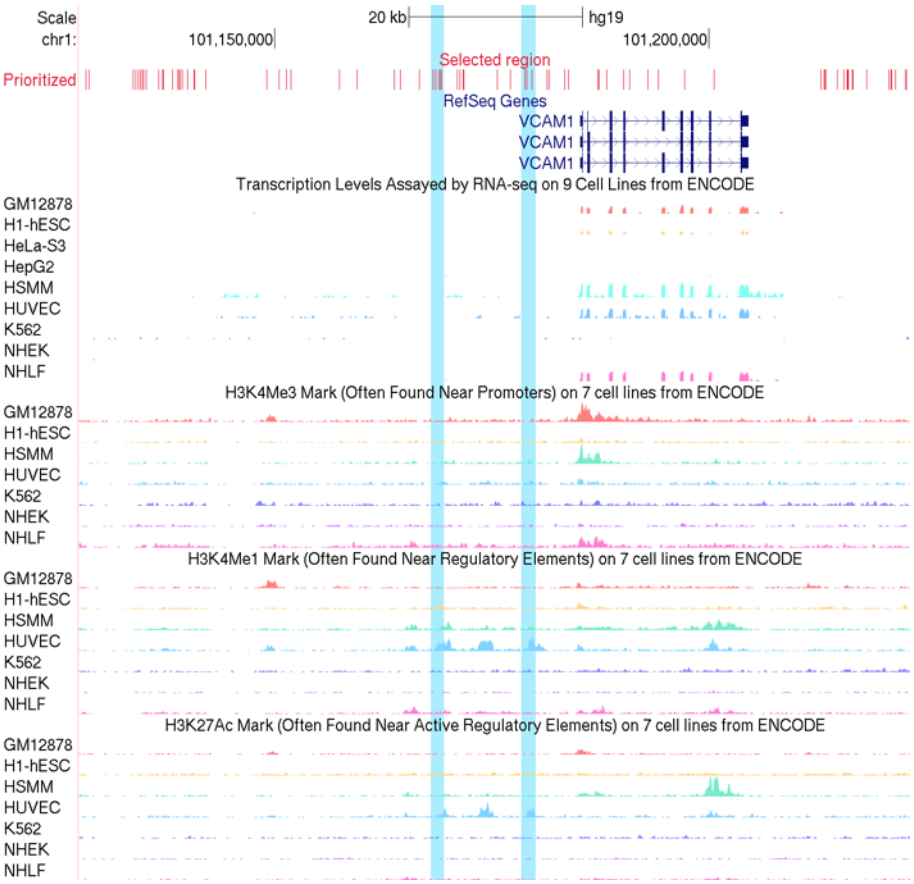

Figure 4.8: **Functional analysis of VCAM1.**
Comparative footprinting of the region around vascular adhesion molecule 1 (VCAM1) in different cell types with ENCODE. The H3K4me1 and H3K27Ac tag densities are higher only in the HUVEC cells and the peaks align with few of the selected SNPs (blue highlighted).

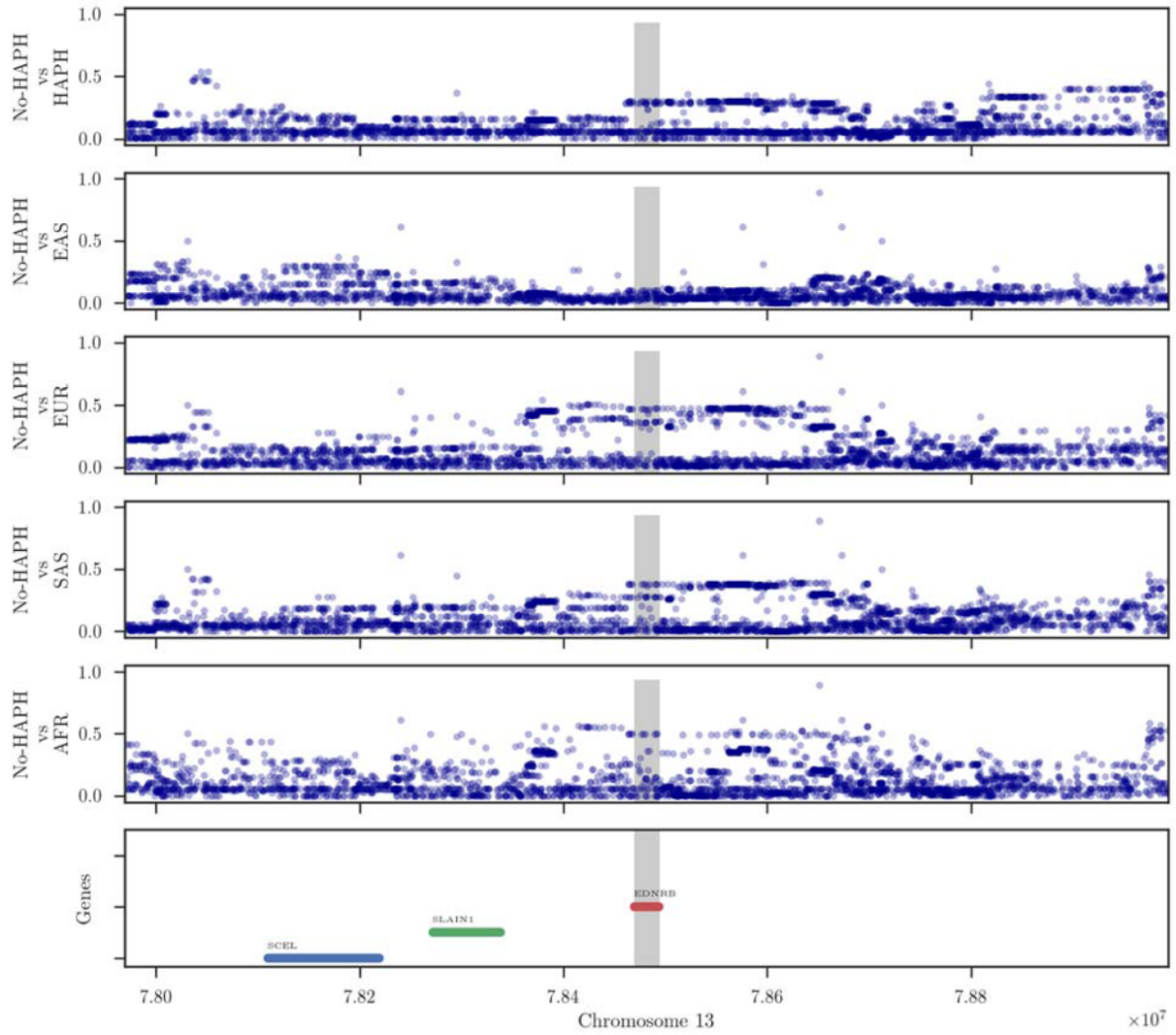

Figure 4.9: **Allele-frequency difference between cases and controls**
Absolute value of allele-frequency-difference between No-HAPH vs HAPH and No-HAPH vs super populations of the 1000 Genome Project for EDNRB gene and its flanking regions.

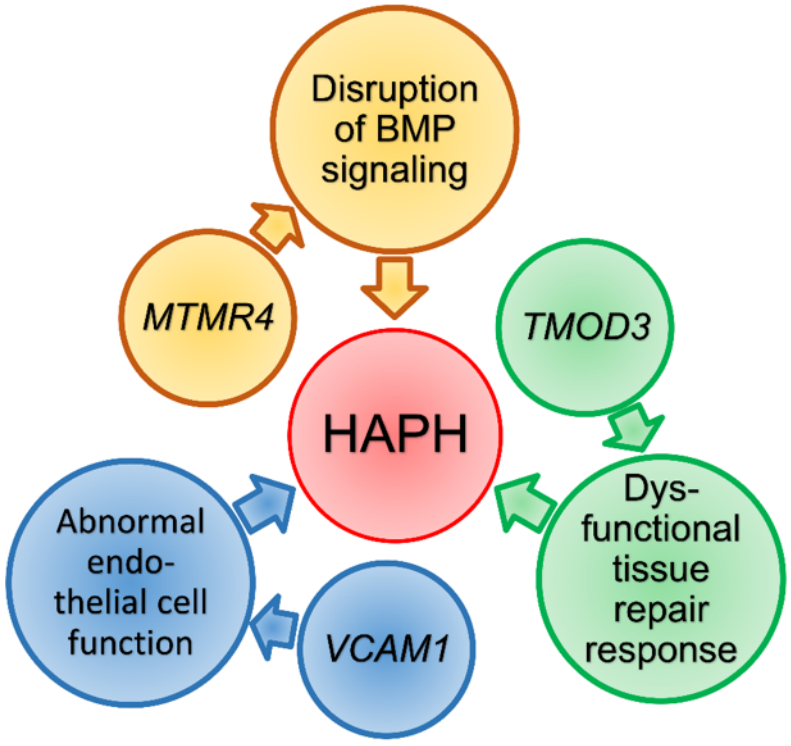

Figure 4.10: **Pathogenesis of HAPH.**
Plausible role of overexpressed MTMR4, TMOD3 and VCAM1 in the pathogenesis of HAPH.

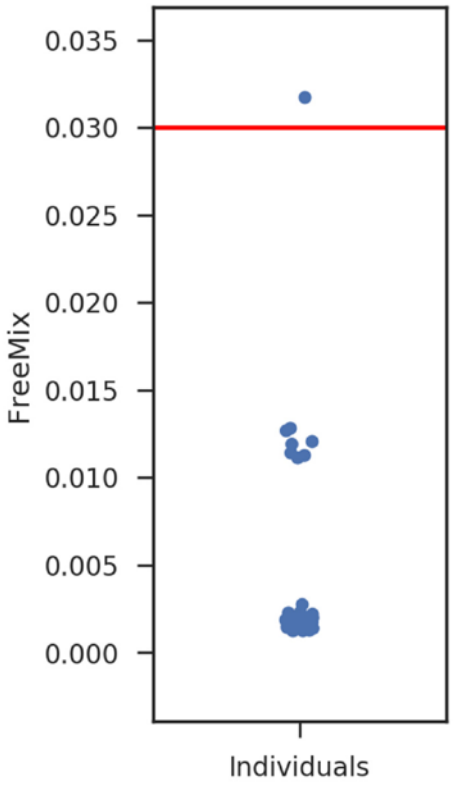

Figure 4.11: **Freemix of whole-genome samples.**
Freemix of Kyrgyz individuals. Empirical cutoff of 0.03 denoted with red line. One of the sample (H11) had score $\geq$0.03.

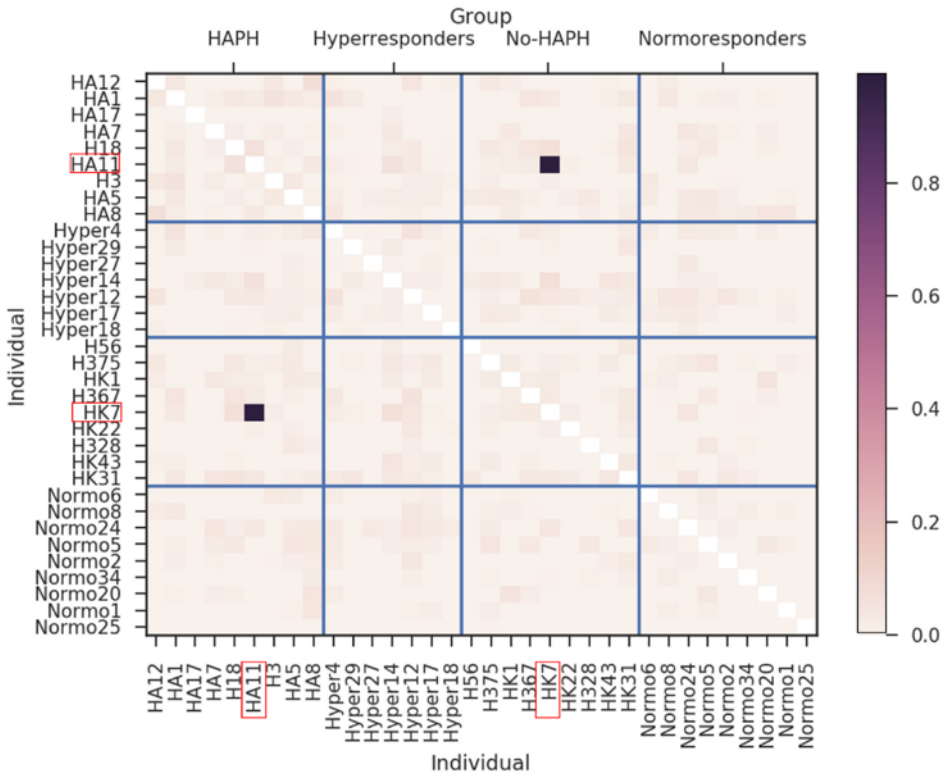

Figure 4.12: **Pairwise identity-by-state matrix.**
Pairwise Identity-by-state (IBS) of the Kyrgyz individuals. Sample H11 had a very high IBS score to sample HK7 (IBP score = 0.99).

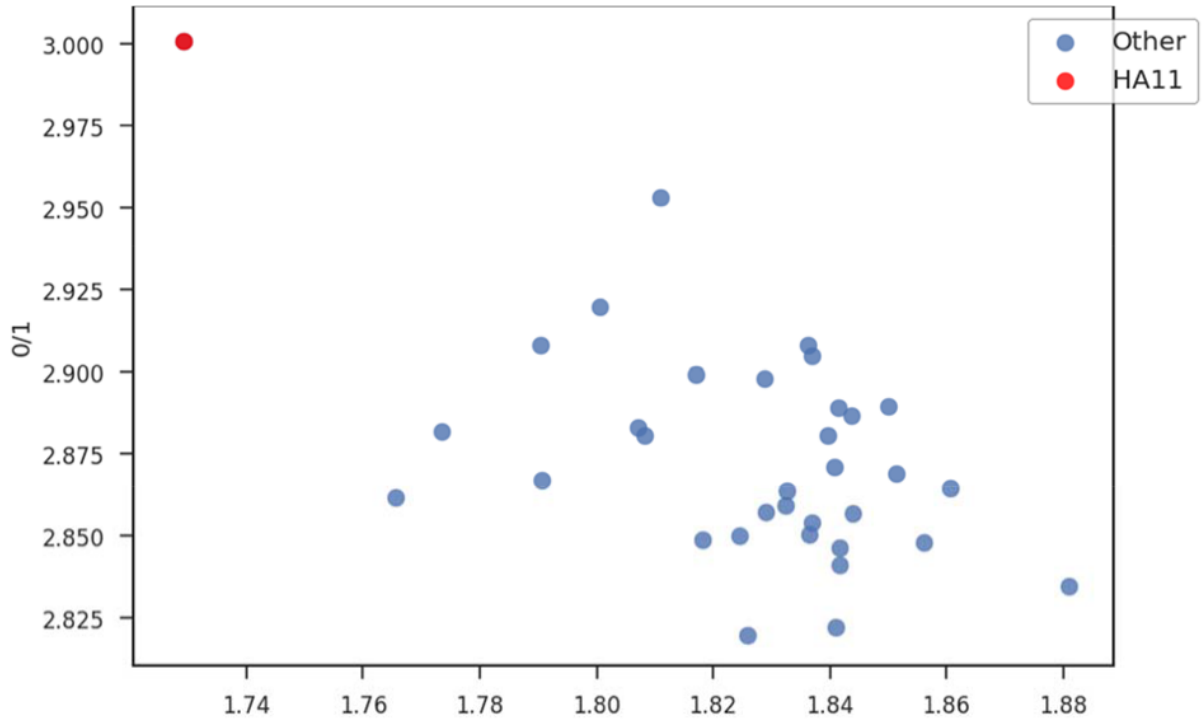

Figure 4.13: **Number of heterozygote vs homozygote calls per individual.**
Sample H11 contains very high heterozygote call.

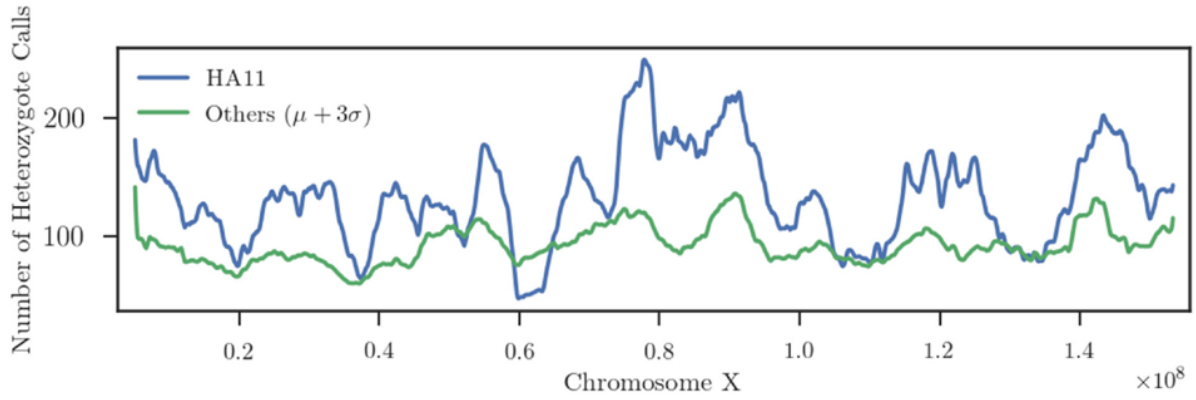

Figure 4.14: **Distribution of Heterozygote Calls in X Chromosome.**
The number heterozygote calls of the contaminated sample HA11 compare to other individuals (mean + 3 standard deviations) in a 100Kbp sliding window on chromosome X. Variant caller is set to make diploid calls for X chromosome. Given that all the samples are male, outside of pseudo-autosomal region heterozygous calls should be rare for a non-contaminated sample.

Bibliography

[1] Iranmehr, A., Akbari, A., Schlötterer, C. & Bafna, V. CLEAR: Composition of Likelihoods for Evolve And Resequence Experiments. *Genetics* **206**, 1011–1023 (2017).

[2] Iranmehr, A., Stobdan, T., Zhou, D., Poulsen, O., Strohl, K. P., Aldashev, A., Telenti, A., Wong, E. H. M., Kirkness, E. F., Venter, J. C. & others. Novel insight into the genetic basis of high-altitude pulmonary hypertension in Kyrgyz highlanders. *European Journal of Human Genetics* **27**, 150 (2019).

[3] Lande, R. Models of speciation by sexual selection on polygenic traits. *Proceedings of the National Academy of Sciences* **78**, 3721–3725 (1981).

[4] Raup, D. M. Biological extinction in earth history. *Science* **231**, 1528–1533 (1986).

[5] White, T. D., Asfaw, B., DeGusta, D., Gilbert, H., Richards, G. D., Suwa, G. & Howell, F. C. Pleistocene homo sapiens from middle awash, ethiopia. *Nature* **423**, 742 (2003).

[6] Nielsen, R., Akey, J. M., Jakobsson, M., Pritchard, J. K., Tishkoff, S. & Willerslev, E. Tracing the peopling of the world through genomics. *Nature* **541**, 302 (2017).

[7] Schlebusch, C. M., Malmström, H., Günther, T., Sjödin, P., Coutinho, A., Edlund, H., Munters, A. R., Vicente, M., Steyn, M., Soodyall, H. & others. Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. *Science* **358**, 652–655 (2017).

[8] Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G. & Siepel, A. Bayesian inference of ancient human demography from individual genome sequences. *Nature genetics* **43**, 1031 (2011).

[9] Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z. X. P., Pool, J. E., Xu, X., Jiang, H., Vinckenbosch, N., Korneliussen, T. S. & others. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329**, 75–78 (2010).

[10] Zhou, D., Udpa, N., Ronen, R., Stobdan, T., Liang, J., Appenzeller, O., Zhao, H. W., Yin, Y., Du, Y., Guo, L. & others. Whole-genome sequencing uncovers the genetic basis of chronic mountain sickness in Andean highlanders. *The American Journal of Human Genetics* **93**, 452–462 (2013).

[11] Tishkoff, S. A., Varkonyi, R., Cahinhinan, N., Abbes, S., Argyropoulos, G., Destro-Bisol, G., Drousiotou, A., Dangerfield, B., Lefranc, G., Loiselet, J. & others. Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. *Science* **293**, 455–462 (2001).

[12] Lao, O., De Gruijter, J. M., van Duijn, K., Navarro, A. & Kayser, M. Signatures of positive selection in genes associated with human skin pigmentation as revealed from analyses of single nucleotide polymorphisms. *Annals of human genetics* **71**, 354–369 (2007).

[13] Jarvis, J. P., Scheinfeldt, L. B., Soi, S., Lambert, C., Omberg, L., Ferwerda, B., Froment, A., Bodo, J.-M., Beggs, W., Hoffman, G. & others. Patterns of ancestry, signatures of natural selection, and genetic association with stature in Western African pygmies. *PLoS genetics* **8**, e1002641 (2012).

[14] Tishkoff, S. A., Reed, F. A., Ranciaro, A., Voight, B. F., Babbitt, C. C., Silverman, J. S., Powell, K., Mortensen, H. M., Hirbo, J. B., Osman, M. & others. Convergent adaptation of human lactase persistence in Africa and Europe. *Nature genetics* **39**, 31 (2007).

[15] Cohen, J. C., Boerwinkle, E., Mosley Jr, T. H. & Hobbs, H. H. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *New England Journal of Medicine* **354**, 1264–1272 (2006).

[16] Stobdan, T., Akbari, A. & Azad, P. New insights into the genetic basis of Monges disease and adaptation to high-altitude. *Mol Biol Evol* **34** (2017). URL `https://doi.org/10.1093/molbev/msx239`.

[17] Murray, C. J. L., Vos, T., Lozano, R., Naghavi, M., Flaxman, A. D., Michaud, C., Ezzati, M., Shibuya, K., Salomon, J. A., Abdalla, S. & others. Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *The lancet* **380**, 2197–2223 (2012).

[18] Levy, S. F., Blundell, J. R., Venkataram, S., Petrov, D. A., Fisher, D. S. & Sherlock, G. Quantitative evolutionary dynamics using high-resolution lineage tracking. *Nature* **519**, 181–186 (2015).

[19] Holohan, C., Van Schaeybroeck, S., Longley, D. B. & Johnston, P. G. Cancer drug resistance: an evolving paradigm. *Nature Reviews Cancer* **13**, 714 (2013).

[20] Clavel, F. & Hance, A. J. HIV drug resistance. *New England Journal of Medicine* **350**, 1023–1035 (2004).

[21] Dondorp, A. M., Nosten, F., Yi, P., Das, D., Phyo, A. P., Tarning, J., Lwin, K. M., Ariey, F., Hanpithakpong, W., Lee, S. J. & others. Artemisinin resistance in Plasmodium falciparum malaria. *New England Journal of Medicine* **361**, 455–467 (2009).

[22] Gold, H. S. & Moellering Jr, R. C. Antimicrobial-drug resistance. *New England journal of medicine* **335**, 1445–1453 (1996).

[23] Mackay, T. F. C. Quantitative trait loci in Drosophila. *Nature reviews genetics* **2**, 11 (2001).

[24] Brass, L. M., Isaacsohn, J. L., Merikangas, K. R. & Robinette, C. D. A study of twins and stroke. *Stroke* **23**, 221–223 (1992).

[25] Mackay, T. F. C. & Lyman, R. F. Drosophila bristles and the nature of quantitative genetic variation. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **360**, 1513–1527 (2005).

[26] Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics eraconcepts and misconceptions. *Nature reviews genetics* **9**, 255 (2008).

[27] Hegreness, M., Shoresh, N., Hartl, D. & Kishony, R. An equivalence principle for the incorporation of favorable mutations in asexual populations. *Science* **311**, 1615–1617 (2006).

[28] Lang, G. I., Rice, D. P., Hickman, M. J., Sodergren, E., Weinstock, G. M., Botstein, D. & Desai, M. M. Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. *Nature* **500**, 571–574 (2013).

[29] Orozco-ter Wengel, P., Kapun, M., Nolte, V., Kofler, R., Flatt, T. & Schlötterer, C. Adaptation of Drosophila to a novel laboratory environment reveals temporally heterogeneous trajectories of selected alleles. *Molecular ecology* **21**, 4931–4941 (2012).

[30] Lang, G. I., Botstein, D. & Desai, M. M. Genetic variation and the fate of beneficial mutations in asexual populations. *Genetics* **188**, 647–661 (2011).

[31] Barrick, J. E., Yu, D. S., Yoon, S. H., Jeong, H., Oh, T. K., Schneider, D., Lenski, R. E. & Kim, J. F. Genome evolution and adaptation in a long-term experiment with Escherichia coli. *Nature* **461**, 1243–1247 (2009).

[32] Bollback, J. P. & Huelsenbeck, J. P. Clonal interference is alleviated by high mutation rates in large populations. *Molecular biology and evolution* **24**, 1397–1406 (2007).

[33] Oz, T., Guvenek, A., Yildiz, S., Karaboga, E., Tamer, Y. T., Mumcuyan, N., Ozan, V. B., Senturk, G. H., Cokol, M., Yeh, P. & others. Strength of selection pressure is an important parameter contributing to the complexity of antibiotic resistance evolution. *Molecular biology and evolution* msu191 (2014).

[34] Maldarelli, F., Kearney, M., Palmer, S., Stephens, R., Mican, J., Polis, M. A., Davey, R. T., Kovacs, J., Shao, W., Rock-Kress, D. & others. HIV populations are large and accumulate high genetic diversity in a nonlinear fashion. *Journal of virology* **87**, 10313–10323 (2013).

[35] Reid, B. J., Kostadinov, R. & Maley, C. C. New strategies in Barrett's esophagus: integrating clonal evolutionary theory with clinical management. *Clinical Cancer Research* **17**, 3512–3519 (2011).

[36] Denef, V. J. & Banfield, J. F. In situ evolutionary rate measurements show ecological success of recently emerged bacterial hybrids. *Science* **336**, 462–466 (2012).

[37] Winters, M. A., Lloyd Jr, R. M., Shafer, R. W., Kozal, M. J., Miller, M. D. & Holodniy, M. Development of elvitegravir resistance and linkage of integrase inhibitor mutations with protease and reverse transcriptase resistance mutations. *PloS one* **7**, e40514 (2012).

[38] Daniels, R., Chang, H.-H., Séne, P. D., Park, D. C., Neafsey, D. E., Schaffner, S. F., Hamilton, E. J., Lukens, A. K., Van Tyne, D., Mboup, S. & others. Genetic surveillance detects both clonal and epidemic transmission of malaria following enhanced intervention in Senegal. *PLoS One* **8**, e60780 (2013).

[39] Barrett, R. D. H., Rogers, S. M. & Schluter, D. Natural selection on a major armor gene in threespine stickleback. *Science* **322**, 255–257 (2008).

[40] Bergland, A. O., Behrman, E. L., O'Brien, K. R., Schmidt, P. S. & Petrov, D. A. Genomic evidence of rapid and stable adaptive oscillations over seasonal time scales in Drosophila. *PLoS Genet* **10**, e1004775 (2014).

[41] Kawecki, T. J., Lenski, R. E., Ebert, D., Hollis, B., Olivieri, I. & Whitlock, M. C. Experimental evolution. *Trends in ecology & evolution* **27**, 547–560 (2012).

[42] Boyko, A. R., Williamson, S. H., Indap, A. R., Degenhardt, J. D., Hernandez, R. D., Lohmueller, K. E., Adams, M. D., Schmidt, S., Sninsky, J. J., Sunyaev, S. R. & others. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet* **4**, e1000083 (2008).

[43] Desai, M. M. & Plotkin, J. B. The polymorphism frequency spectrum of finitely many sites under selection. *Genetics* **180**, 2175–2191 (2008).

[44] Sawyer, S. A. & Hartl, D. L. Population genetics of polymorphism and divergence. *Genetics* **132**, 1161–1176 (1992).

[45] Williamson, E. G. & Slatkin, M. Using maximum likelihood to estimate population size from temporal changes in allele frequencies. *Genetics* **152**, 755–761 (1999).

[46] Wang, J. A pseudo-likelihood method for estimating effective population size from temporally spaced samples. *Genetical research* **78**, 243–257 (2001).

[47] Pollak, E. A new method for estimating the effective population size from allele frequency changes. *Genetics* **104**, 531–548 (1983).

[48] Waples, R. S. A generalized approach for estimating effective population size from temporal changes in allele frequency. *Genetics* **121**, 379–391 (1989).

[49] Terhorst, J., Schlötterer, C. & Song, Y. S. Multi-locus Analysis of Genomic Time Series Data from Experimental Evolution. *PLoS Genet* **11**, e1005069 (2015).

[50] Jónás, ., Taus, T., Kosiol, C., Schlötterer, C. & Futschik, A. Estimating the Effective Population Size from Temporal Allele Frequency Changes in Experimental Evolution. *Genetics* (2016). URL http://www.genetics.org/content/early/2016/08/19/genetics.116.191197.abstract.

[51] Mathieson, I. & McVean, G. Estimating selection coefficients in spatially structured populations from time series data of allele frequencies. *Genetics* **193**, 973–984 (2013).

[52] Illingworth, C. J. R. & Mustonen, V. Distinguishing driver and passenger mutations in an evolutionary history categorized by interference. *Genetics* **189**, 989–1000 (2011).

[53] Bollback, J. P., York, T. L. & Nielsen, R. Estimation of 2Nes from temporal allele frequency data. *Genetics* **179**, 497–502 (2008).

[54] Illingworth, C. J. R., Parts, L., Schiffels, S., Liti, G. & Mustonen, V. Quantifying selection acting on a complex trait using allele frequency time series data. *Molecular biology and evolution* **29**, 1187–1197 (2012).

[55] Malaspinas, A.-S., Malaspinas, O., Evans, S. N. & Slatkin, M. Estimating allele age and selection coefficient from time-serial data. *Genetics* **192**, 599–607 (2012).

[56] Steinrücken, M., Bhaskar, A. & Song, Y. S. A novel spectral method for inferring general diploid selection from time series genetic data. *The annals of applied statistics* **8**, 2203 (2014).

[57] Barrick, J. E. & Lenski, R. E. Genome dynamics during experimental evolution. *Nature Reviews Genetics* **14**, 827–839 (2013).

[58] Baldwin-Brown, J. G., Long, A. D. & Thornton, K. R. The power to detect quantitative trait loci using resequenced, experimentally evolved populations of diploid, sexual organisms. *Molecular biology and evolution* msu048 (2014).

[59] Feder, A. F., Kryazhimskiy, S. & Plotkin, J. B. Identifying signatures of selection in genetic time series. *Genetics* **196**, 509–522 (2014).

[60] Burke, M. K., Dunham, J. P., Shahrestani, P., Thornton, K. R., Rose, M. R. & Long, A. D. Genome-wide analysis of a long-term evolution experiment with Drosophila. *Nature* **467**, 587–590 (2010).

[61] Kingman, J. F. C. On the genealogy of large populations. *Journal of applied probability* **19**, 27–43 (1982).

[62] Crow, J. F., Kimura, M. & others. An introduction to population genetics theory. *An introduction to population genetics theory.* (1970).

[63] Desai, M. M. & Fisher, D. S. Beneficial mutation–selection balance and the effect of linkage on positive selection. *Genetics* **176**, 1759–1798 (2007).

[64] Vitti, J. J., Grossman, S. R. & Sabeti, P. C. Detecting natural selection in genomic data. *Annual review of genetics* **47**, 97–120 (2013).

[65] Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nature genetics* **46**, 919 (2014).

[66] Fan, S., Hansen, M. E. B., Lo, Y. & Tishkoff, S. A. Going global by adapting local: A review of recent human adaptation. *Science* **354**, 54–59 (2016). URL `http://science.sciencemag.org/content/354/6308/54`.

[67] Bersaglieri, T., Sabeti, P. C., Patterson, N., Vanderploeg, T., Schaffner, S. F., Drake, J. A., Rhodes, M., Reich, D. E. & Hirschhorn, J. N. Genetic signatures of strong recent positive selection at the lactase gene. *The American Journal of Human Genetics* **74**, 1111–1120 (2004).

[68] Simonson, T. S., Yang, Y., Huff, C. D., Yun, H., Qin, G., Witherspoon, D. J., Bai, Z., Lorenzo, F. R., Xing, J., Jorde, L. B. & others. Genetic evidence for high-altitude adaptation in Tibet. *Science* **329**, 72–75 (2010).

[69] Daborn, P., Boundy, S., Yen, J., Pittendrigh, B. & others. DDT resistance in Drosophila correlates with Cyp6g1 over-expression and confers cross-resistance to the neonicotinoid imidacloprid. *Molecular Genetics and Genomics* **266**, 556–563 (2001).

[70] Feder, A. F., Rhee, S.-Y., Holmes, S. P., Shafer, R. W., Petrov, D. A. & Pennings, P. S. More effective drugs lead to harder selective sweeps in the evolution of drug resistance in HIV-1. *eLife* **5** (2016). URL `http://www.ncbi.nlm.nih.gov/pubmed/26882502`.

[71] Gottesman, M. M. Mechanisms of cancer drug resistance. *Annual review of medicine* **53**, 615–627 (2002).

[72] Zahreddine, H. & Borden, K. L. Mechanisms and insights into drug resistance in cancer. *Front Pharmacol* **4**, 3389 (2013).

[73] Ariey, F., Witkowski, B., Amaratunga, C., Beghain, J., Langlois, A.-C., Khim, N., Kim, S., Duru, V., Bouchier, C., Ma, L. & others. A molecular marker of artemisinin-resistant Plasmodium falciparum malaria. *Nature* **505**, 50–55 (2014).

[74] Nair, S., Nash, D., Sudimack, D., Jaidee, A., Barends, M., Uhlemann, A.-C., Krishna, S., Nosten, F. & Anderson, T. J. C. Recurrent gene amplification and soft selective sweeps during evolution of multidrug resistance in malaria parasites. *Molecular Biology and Evolution* **24**, 562–573 (2007).

[75] Spellberg, B., Guidos, R., Gilbert, D., Bradley, J., Boucher, H. W., Scheld, W. M., Bartlett, J. G., Edwards, J., of America, I. D. S. & others. The epidemic of antibiotic-resistant infections: a call to action for the medical community from the Infectious Diseases Society of America. *Clinical Infectious Diseases* **46**, 155–164 (2008).

[76] Schlötterer, C., Kofler, R., Versace, E., Tobler, R. & Franssen, S. U. Combining experimental evolution with next-generation sequencing: a powerful tool to study adaptation from standing genetic variation. *Heredity* **114**, 431–440 (2015).

[77] Remolina, S. C., Chang, P. L., Leips, J., Nuzhdin, S. V. & Hughes, K. A. Genomic basis of aging and life-history evolution in Drosophila melanogaster. *Evolution* **66**, 3390–3403 (2012).

[78] Turner, T. L., Stewart, A. D., Fields, A. T., Rice, W. R. & Tarone, A. M. Population-based resequencing of experimentally evolved populations reveals the genetic basis of body size variation in Drosophila melanogaster. *PLoS Genet* **7**, e1001336 (2011).

[79] Zhou, D., Udpa, N., Gersten, M., Visk, D. W., Bashir, A., Xue, J., Frazer, K. A., Posakony, J. W., Subramaniam, S., Bafna, V. & Gabriel G. Haddad. Experimental selection of hypoxia-tolerant Drosophila melanogaster. *Proceedings of the National Academy of Sciences* **108**, 2349–2354 (2011).

[80] Franssen, S. U., Nolte, V., Tobler, R. & Schlötterer, C. Patterns of linkage disequilibrium and long range hitchhiking in evolving experimental Drosophila melanogaster populations. *Molecular biology and evolution* **32**, 495–509 (2015).

[81] Jha, A. R., Miles, C. M., Lippert, N. R., Brown, C. D., White, K. P. & Kreitman, M. Whole-genome resequencing of experimental populations reveals polygenic basis of egg-size variation in Drosophila melanogaster. *Molecular biology and evolution* **32**, 2616–2632 (2015).

[82] Martins, N. E., Faria, V. G., Nolte, V., Schlötterer, C., Teixeira, L., Sucena, . & Magalhães, S. Host adaptation to viruses relies on few genes with different cross-resistance properties. *Proceedings of the National Academy of Sciences* **111**, 5938–5943 (2014).

[83] Izutsu, M., Toyoda, A., Fujiyama, A., Agata, K. & Fuse, N. Dynamics of Dark-Fly Genome Under Environmental Selections. *G3: Genes— Genomes— Genetics* g3–115 (2015).

[84] Agresti, A. & Kateri, M. *Categorical data analysis* (Springer, 2011).

[85] Schraiber, J. G., Evans, S. N. & Slatkin, M. Bayesian inference of natural selection from allele frequency time series. *Genetics* **203**, 493–511 (2016).

[86] Topa, H., Jónás, ., Kofler, R., Kosiol, C. & Honkela, A. Gaussian process test for high-throughput sequencing time series: application to experimental evolution. *Bioinformatics* btv014 (2015).

[87] Anderson, E. C., Williamson, E. G. & Thompson, E. A. Monte Carlo evaluation of the likelihood for Ne from temporally spaced samples. *Genetics* **156**, 2109–2118 (2000).

[88] Durbin, R., Eddy, S. R., Krogh, A. & Mitchison, G. *Biological sequence analysis: probabilistic models of proteins and nucleic acids* (Cambridge university press, 1998).

[89] Ewens, W. J. *Mathematical Population Genetics 1: Theoretical Introduction*, vol. 27 (Springer Science & Business Media, 2012).

[90] Nielsen, R., Williamson, S., Kim, Y., Hubisz, M. J., Clark, A. G. & Bustamante, C. Genomic scans for selective sweeps using SNP data. *Genome research* **15**, 1566–1575 (2005).

[91] Williams, D. & Williams, D. *Weighing the odds: a course in probability and statistics*, vol. 548 (Springer, 2001).

[92] Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences* **100**, 9440–9445 (2003).

[93] Kofler, R. & Schlötterer, C. A guide for the design of evolve and resequencing studies. *Molecular biology and evolution* mst221 (2013).

[94] Ewing, G. & Hermisson, J. MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics* **26**, 2064–2065 (2010).

[95] Peng, B. & Kimmel, M. simuPOP: a forward-time population genetics simulation environment. *Bioinformatics* **21**, 3686–3687 (2005).

[96] Tobler, R., Franssen, S. U., Kofler, R., Orozco-terWengel, P., Nolte, V., Hermisson, J. & Schlötterer, C. Massive habitat-specific genomic response in D. melanogaster populations during experimental evolution in hot and cold environments. *Molecular biology and evolution* **31**, 364–375 (2014).

[97] Long, Q., Rabanal, F. A., Meng, D., Huber, C. D., Farlow, A., Platzer, A., Zhang, Q., Vilhjálmsson, B. J., Korte, A., Nizhynska, V. & others. Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden. *Nature genetics* **45**, 884–890 (2013).

[98] Fiston-Lavier, A.-S., Singh, N. D., Lipatov, M. & Petrov, D. A. Drosophila melanogaster recombination rate calculator. *Gene* **463**, 18–20 (2010).

[99] Burke, M. K., Liti, G. & Long, A. D. Standing genetic variation drives repeatable experimental evolution in outcrossing populations of Saccharomyces cerevisiae. *Molecular biology and evolution* msu256 (2014).

[100] Kosaka, T. & Ikeda, K. Reversible Blockage of Membrane Retrieval and Endocytosis in the Garland Cell of the Temperature-sensitive. *The Journal of cell biology* **97** (1983).

[101] Stephan, W., Song, Y. S. & Langley, C. H. The hitchhiking effect on linkage disequilibrium between linked neutral loci. *Genetics* **172**, 2647–2663 (2006).

[102] Gorrini, C., Harris, I. S. & Mak, T. W. Modulation of oxidative stress as an anticancer strategy. *Nature reviews Drug discovery* **12**, 931 (2013).

[103] Waris, G. & Ahsan, H. Reactive oxygen species: role in the development of cancer and various chronic conditions. *Journal of carcinogenesis* **5**, 14 (2006).

[104] Love, S. Oxidative stress in brain ischemia. *Brain pathology* **9**, 119–131 (1999).

[105] Misra, M. K., Sarwat, M., Bhakuni, P., Tuteja, R. & Tuteja, N. Oxidative stress and ischemic myocardial syndromes. *Medical Science Monitor* **15**, RA209–RA219 (2009).

[106] Dhalla, N. S., Temsah, R. M. & Netticadan, T. Role of oxidative stress in cardiovascular diseases. *Journal of hypertension* **18**, 655–673 (2000).

[107] Semenza, G. L. Oxygen sensing, homeostasis, and disease. *New England Journal of Medicine* **365**, 537–547 (2011).

[108] Gioscia-Ryan, R. A., Battson, M. L., Cuevas, L. M., Eng, J. S., Murphy, M. P. & Seals, D. R. Mitochondria-targeted antioxidant therapy with MitoQ ameliorates aortic stiffening in old mice. *Journal of Applied Physiology* **124**, 1194–1202 (2017).

[109] Harrison, J. F., Kaiser, A. & VandenBrooks, J. M. Atmospheric oxygen level and the evolution of insect body size. *Proceedings of the Royal Society of London B: Biological Sciences* **277**, 1937–1946 (2010).

[110] Berner, R. A., VandenBrooks, J. M. & Ward, P. D. Oxygen and evolution. *Science* **316**, 557–558 (2007).

[111] Eyre-Walker, A. & Keightley, P. D. The distribution of fitness effects of new mutations. *Nature Reviews Genetics* **8**, 610 (2007).

[112] Huang, W., Richards, S., Carbone, M. A., Zhu, D., Anholt, R. R. H., Ayroles, J. F., Duncan, L., Jordan, K. W., Lawrence, F., Magwire, M. M. & others. Epistasis dominates the genetic architecture of Drosophila quantitative traits. *Proceedings of the National Academy of Sciences* **109**, 15553–15559 (2012).

[113] Lai, Y. C., Potoka, K. C., Champion, H. C., Mora, A. L. & Gladwin, M. T. Pulmonary arterial hypertension: the clinical syndrome. *Circ Res* **115** (2014). URL https://doi.org/10.1161/CIRCRESAHA.115.301146.

[114] Gräf, S., Haimel, M. & Bleda, M. Identification of rare sequence variation underlying heritable pulmonary arterial hypertension. *Nat Commun* **9** (2018). URL https://doi.org/10.1038/s41467-018-03672-4.

[115] Aldashev, A. A., Sarybaev, A. S. & Sydykov, A. S. Characterization of high-altitude pulmonary hypertension in the Kyrgyz: association with angiotensin-converting enzyme genotype. *Am J Respir Crit Care Med* **166** (2002). URL https://doi.org/10.1164/rccm.200204-345OC.

[116] Stobdan, T., Karar, J. & Pasha, M. A. Q. High altitude adaptation: genetic perspectives. *High Alt Med Biol* **9** (2008). URL https://doi.org/10.1089/ham.2007.1076.

[117] Udpa, N., Ronen, R. & Zhou, D. Whole genome sequencing of Ethiopian highlanders reveals conserved hypoxia tolerance genes. *Genome Biol* **15** (2014). URL https://doi.org/10.1186/gb-2014-15-2-r36.

[118] Bigham, A., Bauchet, M. & Pinto, D. Identifying signatures of natural selection in Tibetan and Andean populations using dense genome scan data. *PLoS Genet* **6** (2010). URL https://doi.org/10.1371/journal.pgen.1001116.

[119] Stobdan, T., Zhou, D. & Ao-Ieong, E. Endothelin receptor B, a candidate gene from human studies at high altitude, improves cardiac tolerance to hypoxia in genetically engineered heterozygote mice. *Proc Natl Acad Sci USA* **112** (2015). URL https://doi.org/10.1073/pnas.1507486112.

[120] Azad, P., Zhao, H. W. & Cabrales, P. J. Senp1 drives hypoxia-induced polycythemia via GATA1 and Bcl-xL in subjects with Monges disease. *J Exp Med* **213** (2016). URL https://doi.org/10.1084/jem.20151920.

[121] Azad, P., Stobdan, T. & Zhou, D. High-altitude adaptation in humans: from genomics to integrative physiology. *J Mol Med* **95** (2017). URL https://doi.org/10.1007/s00109-017-1584-7.

[122] Wilkins, M. R., Aldashev, A. A. & Wharton, J. α1-A680T variant in GUCY1A3 as a candidate conferring protection from pulmonary hypertension among Kyrgyz highlanders. *Circ Cardiovasc Genet* **7** (2014). URL https://doi.org/10.1161/CIRCGENETICS.114.000763.

[123] Rotta, A., Cánepa, A., Hurtado, A., Velásquez, T. & Chávez, R. Pulmonary circulation at sea level and at high altitudes. *J Appl Physiol* **9** (1956). URL https://doi.org/10.1152/jappl.1956.9.3.328.

[124] Kojonazarov, B. K., Imanov, B. Z. & Amatov, T. A. Noninvasive and invasive evaluation of pulmonary arterial pressure in highlanders. *Eur Respir J* **29** (2007). URL https://doi.org/10.1183/09031936.00137605.

[125] Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19** (2009). URL https://doi.org/10.1101/gr.094052.109.

[126] 1000 Genomes Project Consortium and others. A global reference for human genetic variation. *Nature* **526** (2015). URL https://doi.org/10.1038/nature15393.

[127] Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989).

[128] Fay, J. C. & Wu, C.-I. Hitchhiking under positive Darwinian selection. *Genetics* **155**, 1405–1413 (2000).

[129] Storey, J. D. & Tibshirani, R. Statistical significance for genome wide studies. *Proc Natl Acad Sci USA* **100** (2003). URL https://doi.org/10.1073/pnas.1530509100.

[130] Weir, B. S. & Cockerham, C. C. Estimating F-statistics for the analysis of population structure. *Evol [Soc Study Evol, Wiley]* **38** (1984).

[131] Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* **81** (2007). URL https://doi.org/10.1086/521987.

[132] Flicek, P., Amode, M. R. & Barrell, D. Ensembl 2012. *Nucleic Acids Res* **40** (2012). URL https://doi.org/10.1093/nar/gkr991.

[133] Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N. & others. The genotype-tissue expression (GTEx) project. *Nature genetics* **45**, 580 (2013).

[134] Tishkoff, S. Strength in small numbers. *Science* **349** (2015). URL https://doi.org/10.1126/science.aad0584.

[135] Fumagalli, M., Moltke, I. & Grarup, N. Greenlandic Inuit show genetic signatures of diet and climate adaptation. *Science* **349** (2015). URL https://doi.org/10.1126/science.aab2319.

[136] Ilardo, M. A., Moltke, I. & Korneliussen, T. S. Physiological and genetic adaptations to diving in sea nomads. *Cell* **173** (2018). URL https://doi.org/10.1016/j.cell.2018.03.054.

[137] Hodoğlugil, U. & Mahley, R. W. Turkish population structure and genetic ancestry reveal relatedness among Eurasian populations. *Ann Hum Genet* **76** (2012). URL https://doi.org/10.1111/j.1469-1809.2011.00701.x.

[138] Peng, M.-S., Xu, W., Song, J.-J., Chen, X., Sulaiman, X., Cai, L., Liu, H.-Q., Wu, S.-F., Gao, Y., Abdulloevich, N. T. & others. Mitochondrial genomes uncover the maternal history of the Pamir populations. *European Journal of Human Genetics* **26**, 124 (2018).

[139] Newman, J. H., Wheeler, L. & Lane, K. B. Mutation in the gene for bone morphogenetic protein receptor II as a cause of primary pulmonary hypertension in a large kindred. *N Engl J Med* **345** (2001). URL https://doi.org/10.1056/NEJM200108023450502.

[140] Yu, J., He, X. & Chen, Y. G. Myotubularin-related protein 4 (MTMR4) attenuates BMP/Dpp signaling by dephosphorylation of Smad proteins. *J Biol Chem* **288** (2013). URL https://doi.org/10.1074/jbc.M112.413856.

[141] Zakrzewicz, A., Kouri, F. M. & Nejman, B. The transforming growth factor-beta/Smad2,3 signalling axis is impaired in experimental pulmonary hypertension. *Eur Respir J* **29** (2007). URL https://doi.org/10.1183/09031936.00138206.

[142] Gomez-Pastor, R., Burchfiel, E. T. & Thiele, D. J. Regulation of heat shock transcription factors and their roles in physiology and disease. *Nat Rev Mol Cell Biol* **19** (2018). URL https://doi.org/10.1038/nrm.2017.73.

[143] Chen, X., Qian, D. & Cheng, J. High expression of Rad51c predicts poor prognostic outcome and induces cell resistance to cisplatin and radiation in non-small cell lung cancer. *Tumour Biol* **37** (2016). URL https://doi.org/10.1007/s13277-016-5192-x.

[144] Pullamsetti, S. S., Kojonazarov, B., Storn, S., Gall, H., Salazar, Y., Wolf, J., Weigert, A., El-Nikhely, N., Ghofrani, H. A., Krombach, G. A. & others. Lung cancer–associated pulmonary hypertension: Role of microenvironmental inflammation based on tumor cell–immune cell cross-talk. *Science translational medicine* **9**, eaai9048 (2017).

[145] Rajkumar, R., Konishi, K. & Richards, T. J. Genome wide RNA expression profiling in lung identifies distinct signatures in idiopathic pulmonary arterial hypertension and secondary pulmonary hypertension. *Am J Physiol Heart Circ Physiol* **298** (2010). URL https://doi.org/10.1152/ajpheart.00254.2009.

[146] Fischer, R. S., Fritz-Six, K. L. & Fowler, V. M. Pointed-end capping by tropomodulin3 negatively regulates endothelial cell motility. *J Cell Biol* **161** (2003). URL https://doi.org/10.1083/jcb.200209057.

[147] Farkas, L. & Kolb, M. Vascular repair and regeneration as a therapeutic target for pulmonary arterial hypertension. *Respiration* **85** (2013). URL https://doi.org/10.1159/000350177.

[148] Sui, Z., Nowak, R. B. & Bacconi, A. Tropomodulin3-null mice are embryonic lethal with anemia due to impaired erythroid terminal differentiation in the fetal liver. *Blood* **123** (2014). URL https://doi.org/10.1182/blood-2013-03-492710.

[149] Cox, P. R. & Zoghbi, H. Y. Sequencing, expression analysis, and mapping of three unique human tropomodulin genes and their mouse orthologs. *Genomics* **63** (2000). URL https://doi.org/10.1006/geno.1999.6061.

[150] Richards, T. J., Kaminski, N. & Baribaud, F. Peripheral blood proteins predict mortality in idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* **185** (2012). URL https://doi.org/10.1164/rccm.201101-0058OC.

[151] Agassandian, M., Tedrow, J. R. & Sembrat, J. VCAM-1 is a TGF-β1 inducible gene upregulated in idiopathic pulmonary fibrosis. *Cell Signal* **27** (2015). URL https://doi.org/10.1016/j.cellsig.2015.09.003.

[152] Popper, H. H., Pailer, S., Wurzinger, G., Feldner, H., Hesse, C. & Eber, E. Expression of adhesion molecules in allergic lung diseases. *Virchows Arch* **440** (2002). URL https://doi.org/10.1007/s004280100507.

[153] Shahin, A. A., Anwar, S. & Elawar, A. H. Circulating soluble adhesion molecules in patients with systemic sclerosis: correlation between circulating soluble vascular cell adhesion molecule-1 (sVCAM-1) and impaired left ventricular diastolic function. *Rheumatol Int* **20** (2000). URL https://doi.org/10.1007/s002960000072.

[154] Chaisson, N. F. & Hassoun, P. M. Systemic sclerosis-associated pulmonary arterial hypertension. *Chest* **144** (2013). URL https://doi.org/10.1378/chest.12-2396.

[155] Tao, J., Nie, Y. & Hou, Y. Chemomics-integrated proteomics analysis of Jie-Geng-Tang to ameliorate lipopolysaccharide-induced acute lung injury in mice. *Evid Based Complement Altern Med* **2016** (2016). URL https://doi.org/10.1155/2016/7379146.

[156] Sharma, M., Singh, S. B. & Sarkar, S. Genome wide expression analysis suggests perturbation of vascular homeostasis during high altitude pulmonary edema. *PLoS ONE* **9** (2014). URL https://doi.org/10.1371/journal.pone.0085902.

[157] Xue, C., Sowden, M. & Berk, B. C. Extracellular cyclophilin A, especially acetylated, causes pulmonary hypertension by stimulating endothelial apoptosis, redox stress, and inflammation. *Arterioscler Thromb Vasc Biol* **37** (2017). URL https://doi.org/10.1161/ATVBAHA.117.309212.