Statistical Methods in Population Genetics for Next-Generation Sequencing Data

by

Rebecca Rothwell

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2017

Doctoral Committee:

Associate Professor Sebastian K. Zoellner, Chair
Professor Gonçalo Abecasis
Professor Thomas M. Braun
Associate Professor Jun Li
Assistant Professor Xiaoquan William Wen

To my parents:

"I am an example of what is possible when girls from the very beginning of their lives are loved and nurtured by people around them."- Michelle Obama

ACKNOWLEDGEMENTS

My time at the University of Michigan pursuing my doctoral degree has been a period of intense learning and growth, both within my scientific research and on a personal level. This dissertation could not be completed without the support of my professors, colleagues, friends, and family.

I would like to first express my sincere gratitude to my committee chair and advisor, Dr. Sebastian Zöllner, for his patience, motivation, enthusiasm, and knowledge. Through each challenge in my graduate studies, Dr. Zöllner has been unwavering in his confidence in me, encouraging me to think deeper and work harder in every task. He has provided me with invaluable research and career opportunities, including the phenomenal experience of working at the Max Planck Institute in Leipzig, Germany. I attribute my growth as a researcher to his mentorship and guidance.

In addition, I am extremely grateful for my dissertation committee members. I would like to thank Dr. Thomas Braun, who as both a teacher and a mentor, has taught me how to think, ask questions, and present findings as a statistician. I am also grateful for Dr. Gonçalo Abecasis who is a constant source of insight and novel ideas in statistical genetics. I want to thank Dr. William Wen for always being available for my questions and taking the time for detailed explanations. Finally, I thank Dr. Jun Li who inspires me by his genuine enthusiasm for research in population genetics.

I would like to express my deep appreciation for the past and current members of the Zöllner lab, particularly Matt Zawistowski, Mark Reppell, Keng-Han Lin, and Jed Carlson, for

their stimulating discussion and friendship. I am also indebted to my colleagues at the Center for Statistical Genetics, especially my dear friend Matthew Flickinger. Additionally, I am grateful for my network of support near and far: my classmates in the Department of Biostatistics for sharing my toughest and happiest moments, my Ann Arbor running family for showing me how to push beyond my limits, the Bantles for always welcoming me into their homes with open arms, and my best friends of over twenty years for their unshakeable positivity and loyalty. This experience would be substantially more difficult and certainly less fun without each of you.

Finally, I would like to thank my brother, Matthew Rothwell, for having a sense of humor in times I lost mine, my brother and sister-in-law Nicholas and Gina Rothwell, for their love, encouragement, and adorable children; my boyfriend, Casey Bantle, for being my source of continuous support and laughter; and my parents, Jean and Stephen Rothwell, who have been my biggest fans from my very first steps.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: Introduction

The field of population genetics examines the genetic variation within and between populations and the processes that affect this variation. Such processes include migration of individuals, genetic drift, natural selection, changes in population size, and non-random mating. Though it is not a new field of study, population genetics is constantly evolving due to the rapid improvements in technology for obtaining genetic information. The biological field originated in the early 20th century as a theory-driven discipline that combined Mendelian genetics with Charles Darwin's theory of natural selection [1]. However, it was not until the widespread use of protein electrophoretic variation that population genetics emerged as a data-driven field.[2-5] Statistical-based analyses continued to develop with the invention of polymerase chain reaction technology [6-8], and later, genome-wide single nucleotide polymorphism (SNP) data [9]. Now, with the falling costs of next generation sequencing (NGS), we have access to increasingly large sequencing studies, offering a unique opportunity for innovative methods in statistical analyses.

New developments aim to exploit the wealth of information in rare variation, made available by sequencing more subjects more thoroughly with NGS. Recent studies have shown that rare genetic variants are extremely abundant in the human population. In a resequencing study of 202 genes in 14,002 subjects, Nelson *et al*. found >95% of variants were defined as rare (minor allele frequency less than 0.05%)[10]. Similarly, Tennessen *et al*. sequenced over 15,000 genes in 2,440 subjects, finding 86% of variants found in the subjects were rare[11]. Rare variants have already informed several important population genetics findings. For example, Keinan and

Clark used rare variation from sequencing studies to confirm explosive population growth in human history and estimate the human population has expanded by at least three orders of magnitude in the past 400 generations[12]. Additionally, studying rare variants within and between populations has shaped approaches to testing for disease risk. Gravel *et al* showed that rare variants have extremely limited sharing between diverged populations[13], leading to new recommended adjustments for statistical power in disease association tests. In this dissertation, we present a new collection of population genetics methods, specifically tailored for accessing the signals residing in rare variants. We focus on methods for the processes of migration, changes in population size with subsequent genetic drift, and natural selection.

In Chapter 2 of this dissertation, we present a new method for estimating changing migration rates using the distribution of rare variants among populations. The estimation of migration rates is essential to our understanding of the genetic variation between and within populations. This is valuable for a wide range of studies including preventing false positives in epidemiological studies[14-16], improving matching in case-control association testing [17; 18], ecological and conservation studies [19-21], and historical analyses [22-26]. Furthermore, the migratory history of a population can help to establish the evolutionary origin of a disease. Each of these applications require a reliable, accurate, and realistic method for estimating migration rates.

As rare variant distributions depend only on the migration rate after the mutation-generating event, we can estimate recent migration from very rare variants and ancient migration from more common variants. Using the distribution specific for each minor allele count, we develop a likelihood function to obtain one estimate of the migration rate for variants with the given minor allele count. Therefore, by comparing different estimates of migration based on variants with different minor allele counts, we obtain evidence for changes in migration rate.

Evaluating the performance of our method on simulated data, we can identify migration changes as recent as 20 generation in the past (approximately 400-500 years ago). Additionally, we apply our method to large-scale exonic sequence data from 202 drug target genes sequence data from European and African American samples. We observe an increase in migration rates in recent years from European populations into African American populations, corresponding to the historical record of increased gene flow. In the European samples, we observe generally high migration rates and temporal trends indicating previously higher gene flow between the Northern Europe populations and the rest of Europe, with decreasing rates in more recent years. We hypothesize this could be the result of the recently inferred expansion of Yamnaya steppe herders[27; 28].

In Chapter 3, we present an exact model for modeling population bottlenecks and subsequent genetic drift and provide two different applications for this model. Population bottlenecks are defined as extreme reductions in population size. Populations that experience bottlenecks can exhibit dramatic shifts in population allele frequencies after returning to their original size[29]. This shift can be driven simply by the random sampling process for reproduction known as genetic drift. In some cases, however, selection acting on the variants during regrowth will alter expected allele frequencies. Discerning between drift and selection is essential to understanding functional consequences of the variants and the effects of environmental pressures. Accurately modeling the bottleneck, genetic drift, and selection has many important applications for humans and other organisms, such as studying the results of natural disasters [30-32], captive breeding[33-35] and re-introduction[36-38] of animals, especially endangered species[36; 39; 40], understanding host-pathogen relationships[41-43], and identifying disease patterns[19; 44-48]. Furthermore, studying bottlenecks and genetic drift can help identify populations with reduced

genetic diversity or elevated rare variant frequencies. Modeling the bottleneck and genetic drift can improve the chance of finding rare variant associations that are difficult to identify using general populations[49; 50]. In this chapter, we show two further examples requiring accurate modeling of population bottlenecks: estimating the size of the bottleneck during mtDNA transmission from mother to child and identifying evidence of selection in a cell growth experiment with a known bottleneck size.

In this approach, we construct a flexible, probability-based approach to directly modeling the biological process of population bottleneck and growth and identifying variants with a selection advantage. We model bottleneck effects using binomial sampling and a discrete stochastic process for finite populations. We build upon traditional models of genetic drift[51; 52], allowing for population growth and overlapping generations. With this approach, we can construct a closed-form equation to calculate the probability of observing a shift in allele frequency under genetic drift alone. The primary process of bottleneck and genetic drift can be concisely computed using a discrete Markov chain with two transition matrices. Incorporating sequencing error and genotyping error, we can use this approach to estimate parameters of the model, such as the size of the bottleneck. We also show this model can be easily adjusted to incorporate and estimate a selection coefficient.

We present two unique applications for this mathematical model. First, we apply this approach to 58 variants from next generation sequencing of cell populations isolated from control subjects and individuals with pre-mature aging disorders at the National Institutes of Health. Using the allele frequencies in the cell populations before and after the bottleneck and genetic drift, we find evidence of selection and estimate a selection coefficient in three of these variants. In a second application of this model, we analyze short read sequences of the

mitochondrial DNA of 189 mother-daughter pairs from the Genome of the Netherlands and Biobanking and Biomolecular Research Infrastructure of the Netherlands. We estimate the size and nature of the bottleneck in mtDNA transmission from mother to child based on a maximum likelihood equation and model comparisons.

In the final chapter, we focus on detecting selection signals in autoimmune disease associated loci. Autoimmune diseases are a particularly interesting case for selection because, though detrimental to reproductive fitness, they maintain prevalence in human populations[53; 54]. One hypothesis for this evolutionary phenomenon is that loci known to be associated with multiple autoimmune diseases were previously selected for protection from infectious diseases or pathogens and offered an evolutionary advantage. Therefore, identifying selection signals in these loci could provide important insight into immunity pathways as well as potential medical interventions for autoimmune diseases.

While focusing on autoimmune diseases, we outline and evaluate a new approach to identify selection signals in large-scale whole genome sequencing (WGS). Specifically, we adapt two existing site frequency spectrum statistics, originally developed for small-scale region-based analyses, to applications using WGS data from over 3500 individuals. We identify advantages of using WGS compared to genome wide association studies (GWAS), including avoiding issues of ascertainment bias and missing short range linkage disequilibrium. Furthermore, WGS allows for significance testing using the empirical distributions from SFS statistics genome-wide. Using this approach, we can account for confounding signals of population growth that affect the SFS statistic across the genome. However, the increased information on rare variation from WGS and the large data format requires adjustments in the usage and expected results of these statistics. We evaluate the power of this approach using previously identified positively selected loci under

various parameter settings such as window size and null distributions and discuss issues of dependency across the genome. Therefore, in this final project, we present and evaluate a novel approach to using existing site frequency spectrum statistics.

Next generation sequencing studies are constantly advancing, increasing in scale and number and requiring the development of novel, flexible statistical methods. Population genetics in particular will benefit from the increased access to rare variation with this technology. This dissertation focuses on three population genetics approaches for NGS data and provides important applications of these analyses. These studies provide insight into current and future developments for NGS data, improving the overall understanding of this steadily growing technology.

CHAPTER 2: Estimating Migration Rates Using Distributions of Rare Variants

2.1 Introduction

The estimation of migration rates is essential to understanding the genetic variation between and within populations and therefore, valuable for a wide range of studies. In population genetics, these estimates can elucidate the history of migratory patterns, particularly recent barriers to gene flow[22-26]. Furthermore, migration rate estimates between subpopulations of species are useful in ecological studies and conservation biology[19-21]. Careful modeling of population relationships is also necessary in epidemiological studies to avoid false positives where allele frequencies differ between subpopulations[14-16]. Finally, understanding population structure can improve matching of cases and controls in association testing for public health studies[17; 18]. All of these applications require a reliable, accurate, and realistic method for estimating migration rates.

Traditional estimators of migration rates primarily fall into three categories, which recent methods build upon. The first category, estimators based on Wright's F-statistic, include the most commonly employed and simplest methods[55]. These estimators use comparisons of heterozygosity within and between subpopulations to develop a single estimate of migration based on information across all variants[56-65]. These estimators often rely on several simplifying assumptions including constant and equal population sizes. Second, there are estimators that rely on coalescent theory[66]. These methods are based on modeling a genealogy of sampled

individuals under various population parameters including migration and determining the likelihood of observed data under different parameter choices[67-71]. The coalescent-based methods are typically computationally intense but allow for relaxing some assumptions such as constant population size. More recent updates to the coalescent approach use a Bayesian framework to estimate a distribution of migration rates[72-78]. These approaches are generally able to analyze large amounts of data by applying algorithms such as the Markov Chain Monte Carlo. Third, there are estimators that use a maximum likelihood estimate based on allele frequencies across populations[79; 80]. This strategy requires an accurate probability distribution of allele frequencies. This can be difficult to obtain, particularly in the case of asymmetric migration rates [57]. Generally, the distribution is established through simulations of simple population models such as Wright's island model[81]. In this distribution of allele frequencies, alleles private to populations are informative of previous genetic dispersal[70; 74; 82]. Our method builds on this traditional maximum likelihood approach by focusing on rare and population-specific variants.

In each of these traditional estimators, the methods consider a constant migration parameter and do not provide a temporal picture of migration changes over time. Identifying changes in migration rate is critical not only to understand our demographic history but also to control rare variant association analysis for population stratification. More recent methods, such as the Pairwise Sequential Markovian Coalescent (PSMC) and, later, the Multiple Sequential Markovian Coalescent (MSMC) allow for estimating changes in migration rate, though restricted to pre-historic times, by computing TMRCA distributions across the genome, indicating historical migration events [83; 84]. As this approach relies on a small number of chromosomes it is not sensitive to recent timescale changes less than 10,000 years ago. Another recent method, dadi[85] is a powerful approach to identifying a broad range of demographic parameters, including

changing migration, using diffusion approximations to obtain the site frequency spectrum. However, this method relies on solving partial derivative equations numerically, making it computational intensive for complicated histories and large data sets and its solutions can be unstable[86]. Excoffier et al[87] overcomes this issue by obtaining the expected site frequency spectrum (SFS) under specific demographic histories from coalescent simulations. Our method applies a similar approach, but adds additional information by slicing across the SFS and leveraging the relative rareness of variants in the population to present a temporal picture of changing migration.

A major challenge for estimating migration rates is the confounding effect of population size. Estimates of migration rate depend on the amount of genetic variation shared between populations. For this reason, migration methods traditionally estimate a compound parameter of population size and migration rate. However, many populations in nature have undergone recent dramatic size changes [12; 88; 89]. As genetic variation increases with population size, the amount of shared variation also increases, even if the per individual migration rate remains constant. Estimates of changes in migration rate must account for population growth so that migration rates are appropriately scaled and do not falsely infer higher migration in recent years. Establishing a method that reflects such demographic changes is necessary to understand the effects of historical human events on gene flow in non-constant size populations.

In this study, we leverage the increasing prevalence of large sequencing datasets and develop a method to identify changing migration patterns using rare variants. Previous studies show that rare variants are particularly informative for understanding fine-scale population structure and geographic origin, with a large excess of rare variation existing in current human populations [10; 90-92]. We base our method on the distribution of these rare variants among multiple

populations. Intuitively, rare variants will largely be population specific under low migration, while increasing migration rates will generate a more balanced distribution of alleles. However, only migration events that occur after a variant arises can affect the population distribution of that variant. Hence, rare variants that arose recently are only affected by recent migration events, reflecting the migration rate in the recent past. We develop a likelihood function to obtain one maximum likelihood estimate of the migration rate using variants with a given minor allele count. By comparing different estimates of migration based on variants with different minor allele counts, we obtain evidence for changes in migration rate. Evaluating the performance of our method on simulated data, we can identify migration changes as recent as 20 generation in the past (approximately 400-500 years ago). Furthermore, this method is adjustable for population growth and changes in population size. We also show that the method is robust to model population characteristics including misspecifications in effective population size, ancestral divergence, asymmetric migration, and uneven sample size between the island populations.

As a proof of principle, we estimated migration rates in recent years from European populations to African American populations using counts generated from sequence data. This data set included 7809 individuals with broad consent selected from a previously published sequencing study of 14,002 subjects, resulting in 7470 Europeans and 339 African American subjects.[10] We observe an increase in migration rates in recent years from European populations into African American populations, corresponding to the historical record of increased gene flow. These results provide a real data example in which our method detects realistic migration changes over recent generations in the presence of population growth. We further analyze four geographically-defined subpopulations of Europe: Northwestern (British Isles), Northern

(Iceland, the Faroe Islands, Denmark, Norway, Sweden, Estonia, Latvia, Lithuania), Western (Belgium, France, Luxembourg, and the Netherlands), and Central (Austria, Germany, and Switzerland)[10]. We observe generally high gene flow with slightly lower migration rates between the more geographically distant populations. Interestingly, we observe temporal trends indicating previously high gene flow between the Northern Europe populations and the rest of Europe, with decreasing rates in more recent years. We hypothesize this signal is the result of the recently inferred expansion of Yamnaya steppe herders[27; 28].

2.2 Methods

2.2.1 Dataset Formulation and Notation

Consider a sample of n chromosomes from s subpopulations, each with effective diploid population size N [93]. We define a segregating site as having minor allele count *k* if the minor allele occurs *k* times across all s subpopulations. For a variant with k minor alleles, let k_t be the number of minor alleles in population t ($t = 1, \dots, s$) with $\sum_{t=1}^{s} k_t = k$. There are $a = \binom{k+s-1}{k}$ ways to arrange the *k* observations in s populations (using unordered sampling with replacement, we choose a population t for each of the *k* observation)[94]. Each observed variant will occur in one of these configurations. Across the s populations, let h_k be the number of sites with minor allele count k. The number of times configuration i is observed is h_{ki} ($h_k = \sum_i h_{ki}$). Notice that the probability of a variant with allele count k being in configuration i depends on the migration rate between the populations since the time of the mutation event that generated the variant. For example, a doubleton ($k = 2$) taken from a sample of $s = 2$ populations, will typically be population specific (configurations $[2,0]$ or $[0,2]$) if migration rate is low and often shared between populations (configuration$[1,1]$) if migration rate is high. Our objective is to

estimate a migration rate M, for each minor allele count k based on $(h_{k1}, \dots, h_{ka})$ where $M = 4Nm$ and m is the fraction of each population made up of new migrants each generation. Let $\boldsymbol{M} = (M_1, \dots, M_{\binom{s}{2}})$ be the vector of migration rates between pairs of s populations. For a given variant with minor allele count k, let $p_{\boldsymbol{M},k,i}$ be the probability of observing configuration i for a given migration rate vector $\boldsymbol{M}$. The $p_{\boldsymbol{M},k,i}$ are therefore dependent on the sample configuration such as the sample size and number of populations. Assuming that all sites segregate independently, we calculate the likelihood of migration rate vector $\boldsymbol{M}$using a multinomial distribution:

$$L_k(\boldsymbol{M}|h_{k1}, h_{k2}, \dots h_{ki}) = \frac{h_k!}{h_{k1}!\, h_{k2}! \dots h_{ki}!} \left(p_{\boldsymbol{M},k,1}{}^{h_{k1}} p_{\boldsymbol{M},k,2}{}^{h_{k2}} \dots p_{\boldsymbol{M},k,i}{}^{h_{ki}}\right) \tag{2.1}$$

Maximizing this likelihood provides an estimate for the most likely migration parameter estimate based on the information contained only in sites with allele count k. We note that in many cases we cannot assume all sites segregate independently. In this situation, (2.1) conveniently becomes a consistent and asymptotically normal pseudo-likelihood [95].

We identify the maximum likelihood estimate of the migration rate based on $(h_{k1}, h_{k2}, \dots h_{ki})$ by implementing a grid-search algorithm across a grid of migration rates vectors, $\boldsymbol{G}$. Using (2.1), we calculate the likelihood for each grid point and identify the grid point with the highest likelihood as the maximum likelihood estimate for allele count k.

We notice at higher allele counts, migration estimates are less accurate due to the decreasing number of observed segregating sites and increasing possible number of configurations. We address this problem by collapsing neighboring maximum likelihood estimates (i.e. those for allele counts k, $k+1$, and $k+2$), into "bins" to ensure a minimum number of observations contribute to each estimate. This means the higher allele count bins

include a larger range of allele counts due to the limited number of observations for each configuration. While under this scheme, estimates are no longer allele-specific, we can observe a qualitative trend of migration estimates using rank-based correlation statistics and model comparisons.

2.2.2 Estimating $p_{\boldsymbol{M},k,i}$

To fully construct (2.1), we first obtain the $p_{\boldsymbol{M},k,i}$ the probability of observing each configuration i for a particular migration vector. We do so by simulating sequence data using the coalescent simulator *ms*[96]. For known parameters, such as sample size and relative population sizes, we choose the simulation parameters to match those of the populations in our dataset of interest. We also set simulation parameters which are unknown such as effective population size, exponential population growth, and instantaneous changes in population size, based on our best knowledge of the populations represented in our dataset of interest. For our basic analyses, we simulate 10^7 independent regions of 3.75 kB each without recombination, drawn from two populations of equal and constant effective population size (N_e=10,000 diploid individuals each). This creates a sample of $2.5\text{x}10^4$ diploid individuals with a total of 750 kB each. In this two-population setting with symmetric migration, the migration rate vector is a single element, M. For our grid of possible migration rates for M, we include a denser coverage lower levels of migration where small differences in gene flow are more easily detectable ($\boldsymbol{G}$ =[5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1250, 1500, 1750, 2000, 2250, 2500, 2750, 3000, 3250, 3500, 3750, 4000, 4250, 4500, 4750, 5000, 5250, 5500, 5750, 6000, 6250, 6500, 6750, 7000, 7250, 7500, 7750, 8000, 8250, 8500, 8750, 9000, 9250, 9500, 9750, 10000, 10250, 10500, 10750, 15000, 20000]),

corresponding to $m = 0$ to 0.5. Throughout this analysis, we use the present-day scaled M parameter, as in the coalescent simulator *ms*[96].

For each M in $\boldsymbol{G}$, we record the number of times r_{ki} each configuration i is observed for each allele count k. Using these frequencies, we estimate the probability $p_{\boldsymbol{M},k,i} = r_{ki} / \sum_i r_{ki}$ of observing this distribution under migration parameter $\boldsymbol{M}$.

2.2.3 Identifying Changing Migration Using Spearman's Rho

For each scenario of migration that we simulate, detailed below, we calculate Spearman's rank correlation coefficient (Spearman's rho) between the allele-count bin number and the allele-count-specific estimated migration rates. A positive Spearman's rho indicates lower migration rates for variants with smaller allele counts than for variants with larger allele counts, indicating decreased migration in recent years. In contrast, a negative Spearman's rho indicates higher migration in smaller allele counts, indicating increased migration in recent years. We estimate the power to detect temporal change in migration rate using this approach by counting the number of significant rho values in these models.

2.2.4 Grid Search Algorithm

In addition to using Spearman's rho as a preliminary and immediate test for identifying changing migration, we introduce a grid search and model comparison procedure to obtain precise estimates of the migration rates and time of change. We build on the likelihood defined in (2.1) but now the grid of possible migration rates, $\boldsymbol{G}$, is a grid of demographic histories,$\boldsymbol{G_H}$ including both constant and changing migration and population expansion. As above, let h_k be the number of sites with minor allele count k. The number of times configuration i is observed is h_{ki} ($h_k = \sum_i h_{ki}$). For a given variant with minor allele count k, let $p_{H,k,i}$ be the probability of observing configuration i for a given history H. Let z be the highest minor allele count to be

analyzed. For datasets considered here we use $z = 75$ for optimal power and precision of detecting recent migration changes. The total number of sites with minor allele count less than limit z is $h = \sum_{k=2}^{z} h_k$. Let $p_{H,k}$ be the probability of observing minor allele count k under history H. For each possible history H in $\boldsymbol{G_H}$, we calculate the probability of the observed configurations for each allele count under of history H (2.2) using a multinomial distribution similar to (2.1). In (2.3), we then combine information across allele counts using an additional multinomial distribution based on the total minor allele counts, to obtain a likelihood estimate for observing the complete allele count data under specific demographic history.

$$P(h_{k1}, h_{k2}, \ldots h_{ki}|H) = \frac{h_k!}{h_{k1}!\, h_{k2}! \ldots h_{ki}!}\left(p_{H,k,1}{}^{h_{k1}} p_{H,k,2}{}^{h_{k2}} \ldots p_{H,k,i}{}^{h_{ki}}\right) \tag{2.2}$$

$$L(H|observed\ SFS) = \left[\frac{h!}{h_2!\, h_3! \ldots h_z!}\left(p_{H,2}{}^{h_2} p_{H,3}{}^{h_3} \ldots p_{H,z}{}^{h_z}\right)\right] \prod_{k=2}^{z} P(h_{k1}, h_{k2}, \ldots h_{ki}|H) \tag{2.3}$$

To construct each history in the grid used in this analysis, we simulate 10^7 independent regions of 3.75 kB each without recombination, drawn from two populations of equal and constant effective population size (N_e=10,000 diploid individuals each). This creates a sample of 2.5×10^4 diploid individuals with a total of 750 kB each. In this two-population setting with symmetric migration, we vary three parameters in the grid: a recent migration rate, M_1, a past migration rate, M_2, and a time of migration rate change τ. For each H in $\boldsymbol{G_H}$, we record r_{ki} , the number of times each configuration i is observed for each allele count k. Using these frequencies, we estimate the probability $p_{H,k,i} = r_{ki} / \sum_i r_{ki}$ of observing this distribution and the probability $p_{H,k} = \sum_i r_{ki} / (\sum_k [\sum_i r_{ki}])$ of observing this minor allele count under migration parameter H.

2.2.5 Testing for Changing Migration and Estimating Parameters with Grid Search

To test for changing migration using this approach, we compare the likelihoods of observing the data of interest, D, under each history. We define the history that achieves the maximum likelihood as the alternative model (H_a) and the constant migration history with the largest likelihood as the null model (H_0). We then compute the likelihood ratio test statistic: $\mathcal{L} = -2[\log(L_{H_0}) - \log(L_{H_a})]$. Where these likelihoods are sufficiently different, we reject the null hypothesis of constant migration.

To statistically test this difference, we must identify a critical value. Using the coalescent simulator *ms*, we simulate 1,000 datasets under the null hypothesis constant migration model, using the sample size and data size of D. For each dataset, we calculate the likelihoods for each history and compute the likelihood ratio test statistic as above. The empirical distribution of these likelihood ratio test statistics is the distribution for our test statistic under the null. We define the 95% quantile as C, the critical value for our testing procedure. Where our likelihood ratio statistic, $\mathcal{L}$, is greater than C we reject the null hypothesis of constant migration. In this case, we estimate the parameters of changing migration as those defined in the maximum likelihood history.

2.2.6 Simulation of Populations with Constant and Changing Migration

To assess our method, we generate test datasets with the coalescent simulation program *ms*. We first simulate two populations of constant population size under a two-population island model with effective population sizes of 10,000 individuals each. Initially, we consider a constant migration rate of $M = 100$ ($m = 0.0025$). We generate 10,000 datasets of 1,000 diploid individuals from each population and applied our algorithm to each dataset to estimate the temporal pattern of migration rate as well as the average bias and mean squared error. To

determine the behavior of the method for a range of migration rates, we perform 10,000 simulations for 10 migration parameters from $M = 10$ to $M = 1{,}000$, calculating the mean, median, 95% empirical confidence intervals, and relative bias.

We next focus on the method's ability to detect a temporal change of migration rate. In model (a), we first simulate two populations under the ancestral divergence model. In this model, two populations were previously a single ancestry population. In recent times, a split occurred, creating two distinct populations with low migration, $M = 100$ ($m = 0.0025$). For 10,000 iterations, we simulate this scenario where the time of split occurs at four different times: 0.01, 0.005, 0.001, and 0.0005 coalescent units, corresponding to approximately 400, 200, 40, and 20 generations before present day. With 20 year generations, these time scales are 8,000, 4,000, 800, and 400 years in the past, respectively. Second, in model (b), we simulate a model of two isolated populations with historically zero gene flow with new migration ($M = 100$) beginning at the four different times (0.01, 0.005, 0.001, and 0.0005 coalescent units). Third, we generate two additional models with smaller changes in migration: (c) migration decreases from $M_2 = 100$ in the past to $M_1 = 50$ in recent years and (d) migration increases from $M_2 = 50$ in the past to $M_1 = 100$ in recent years. In each case, we simulate the four different times of change (τ= 0.01, 0.005, 0.001, and 0.0005).

Using these models (c) and (d), we also assess the power and precision of our grid search and model comparison approach to identify the parameters of changing migration. The grid used included 657 histories for each combination of: recent migration (M_1): 0, 25, 50, 75, 100, 125, 150, 175, 200, past migration (M_2): 0, 25, 50, 75, 100, 125, 150, 175, 200, and time of change in coalescent units in the past from present day (τ): constant, 0.00025, 0.0005, 0.00075, 0.001, 0.0025, 0.005, 0.0075, 0.01, 0.025. We simulate 1000 datasets for each model and record the

likelihood of each history in the grid. We obtain the critical value for testing constant migration using the likelihood ratio testing procedure outlined in section 2.2.5. For the 1,000 test data sets, we record how often we reject the null hypothesis of constant migration. We also identify the number of times the highest likelihood belongs to a non-constant model and the number of times all three parameters are estimated exactly correctly.

To consider model adjustments in cases of exponential growth, we simulate a two-population island model with a growth in effective population sizes of 10,000 to 1,000,000 in 500 generations and a constant, symmetric migration of $M = 100$ ($m = 0.000025$). To adjust for this change, we re-estimate the $p_{\boldsymbol{M},k,i}$ in the likelihood equation based coalescent-simulated data under a model of exponential growth. Under these expansion-adjusted parameters, we expect the migration estimates to be close to $M = 100$ across allele count bins, newly scaled based on the present-day population size. We calculate median estimates before and after adjusting our model for this growth.

Finally, we assess our method's robustness to model misspecifications such as deep ancestral divergence, imbalanced effective population size, and asymmetric migration (2.5). Previously, we simulated the $p_{\boldsymbol{M},k,i}$values with parameters corresponding to the populations of interest including sample size, effective population size, ancestral populations, direction of migration, and relative subpopulation sizes. We now consider what happens when these parameters are incorrectly defined. We generate 10,000 simulations of constant migration ($M = 100$), under a range of parameters and counted the number of simulations falsely identified as changing migration by Spearman's rho.

2.2.7 Application

We estimated migration rates in European and African American populations using counts generated from an exome sequencing study of 14,002 individuals focused on 202 drug target genes (351 kb of coding and 323 kb of untranslated (UTR) exon regions).[10] We analyzed 7809 individuals with broad consent, including 7470 Europeans and 339 African Americans. From the European subjects, we down-sampled 2800 individuals to obtain equal sample sizes from four geographically-defined subpopulations (n=700 per population): Northwestern (British Isles), Northern (Iceland, the Faroe Islands, Denmark, Norway, Sweden, Estonia, Latvia, Lithuania), Western (Belgium, France, Luxembourg, and the Netherlands), and Central (Austria, Germany, and Switzerland).We consider pairs of these populations at a time, resulting in six different temporal pictures of migration. In this analysis, we estimate the $p_{\boldsymbol{M},k,i}$ values for the maximum likelihood using simulations based on a symmetric migration model with exponential growth. In each case, we set a growth in effective population sizes of 10,000 to 1,000,000 in 500 generations based on approximate recent estimates of human population expansion [97; 98]. For each pair of populations, after estimating the allele-count-specific migration rates, we calculate Spearman's rho and assessed significance of an increasing or decreasing trend in migration rates.

We also apply the model comparison grid search approach, with histories including an initial migration rate, final migration rate, and time of change. The values included in the grid are based on the range of allele count specific migration estimates, using the same growth in effective population sizes of 10,000 to 1,000,000 in 500 generations. The grid included 512 histories for each combination of: recent migration (M_1): 2500, 5000, 7500, 10000, 12500, 15000, 17500, 20000, past migration (M_2): 2500, 5000, 7500, 10000, 12500, 15000, 17500, 20000, and time of change in coalescent units in the past from present day (τ): constant,

0.0000025, 0.000005, 0.0000075, 0.00001, 0.000025, 0.00005, 0.000075, 0.0001, 0.00025. For each pair of populations, we test for changing migration and estimate these three parameters using this approach.

As an example of known historical gene flow, we use our method to estimate the migration between the 339 African Americans (population one) and the 7470 Europeans (population two). We estimate the $p_{\boldsymbol{M},k,i}$ values used in the likelihood via simulations based on a directional, asymmetric migration pattern with exponential growth (growth in effective population sizes of 10,000 to 1,000,000 in 500 generations). We estimate the allele-count-specific migration rates and calculated Spearman's rho with standard significance testing of an increasing or decreasing trend in migration rates. We repeated this procedure focusing on a down-sampled set of 339 Europeans to understand the effects of equal versus unequal sample sizes from the two populations.

We also apply the model comparison grid search approach, with histories including a recent migration rate, past migration rate, and time of change. The values included in the grid are based on the range of allele count specific migration estimates, with directional migration, exponential growth, and sample size as outlined above. The grid used included 52 histories for each combination of: recent migration (M_1): 500, 5000, 10000, 12500, past migration (M_2): 500, 5000, 10000, 12500, and time of change in coalescent units in the past from present day (τ): constant, 0.000005, 0.00001, 0.00005, 0.0001. For each pair of populations, we test for changing migration and estimate these three parameters using this approach.

2.3 Results

To evaluate this method, we simulated a series of datasets of two populations with varying parameters and applied the estimation method on each dataset to evaluate its ability to

identify changing migration rates. We further assess our method's ability to adjust for and estimate migration under both constant and changing population size, including exponential growth. Finally, we used our method to estimate migration rates in African Americans and Europeans. We estimate migration rates in pairs of four distinct European populations and show the method's ability to detect previously unknown changes in gene flow.

2.3.1 Constant Migration

We first compare the accuracy and precision of the estimator across allele count bins assuming constant migration rates. We simulate 10,000 coalescent-based sequence datasets of 750 kb from 1,000 individuals from two populations of constant population size under a two-population island model with constant, symmetric migration rate of $M = 100$ ($m = 0.0025$). For each dataset, we estimate the migration rate M between these two populations for each allele count bin. We then calculate the mean and median migration estimates across datasets for each allele count bin with corresponding 95% empirical confidence intervals (Figure 2.1). In the first allele count bin, which consists of doubletons, the median migration estimate is 100 with a mean of 101.8 (95% CI: [50,160]). In the highest allele count bin, with minor allele counts between 72 and 100, the median migration estimate is 100 and mean of 103.5 (95% CI: [70,140]). Across all allele count bins, the median estimate of the migration rate is consistently 100. The mean estimates range from 101.24 to 103.67 (Figure 2.1). The mean squared error (MSE) for these estimates is lowest (291.74 to 329.46) in the mid-range allele count bins (bins 8 through 17), with the highest MSE observed in the first allele count bin (809.48) (Figure 2.8). While these median estimates are accurate across allele count bins, the means and the empirical confidence intervals are increasingly skewed slightly upwards with increasing allele count bin number (Figure 2.1). We calculate the average bias as the mean of the bias for each of the 10,000

simulations (Figure 2.8). The magnitude of the average bias increases at higher allele count bins. We observed this upward bias due to the flattened likelihood curve at higher allele counts (Figure 2.9). Particularly at the higher allele count bins, the shape of these likelihood curves fall quickly for migration rates decreasing below the MLE, but decrease slowly for migration rates increasing above the MLE. However, these slight biases do not affect our proposed initial test for changing migration rates using Spearman's rho (See 2.2.3), as 370 of 10,000 simulations resulted in significant ($p<0.05$) test, providing a false positive estimate of 3.7%.

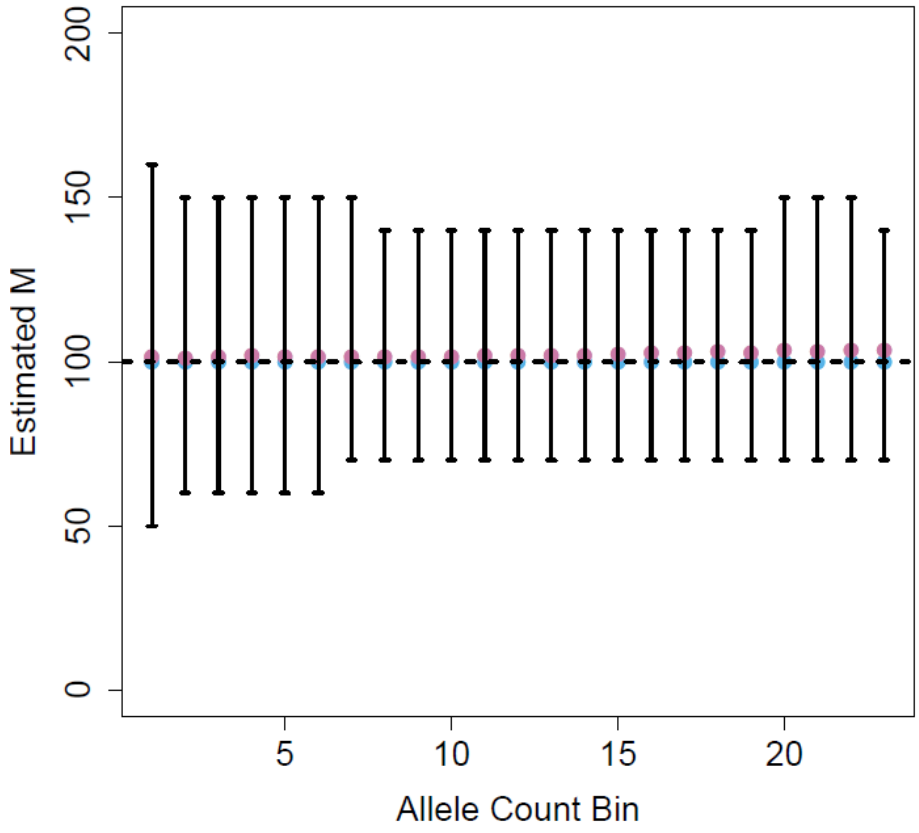

Figure 2.1 Estimates Under Constant Migration Across Allele Count Bins
Median estimates, indicated by the blue dots, and mean estimates, indicted by pink dots, for each allele count bin and the corresponding empirical 95% confidence intervals, shown as vertical error bars, with the true migration parameter of $M = 100$ indicated by the hashed line. The vertical axis shows the estimated $\widehat{\mathbf{M}}$ value and the horizontal axis shows the allele count bin (with decreasing level of rareness with increasing bin number).

To evaluate robustness when the model used to estimate parameters is miss-specified, we simulate a range of scenarios: deep ancestral divergence, asymmetric migration rates, and asymmetric population sizes. We estimate migration rates and calculate Spearman's rho to obtain false positive rates under each of these settings (2.5). For most scenarios, Spearman's rho is well-calibrated. Only for highly asymmetric population sizes with $N_1 = 5{,}000$ and $N_2 = 15{,}000$, miss-specified as symmetric with $N = 10{,}000$ each, is the false positive rate high at 20.58%.

To determine the performance of our method across a range of migration rates, we generate 10,000 datasets for each of nine constant migration parameters from M =100 to M =1,000 and analyzed the resulting datasets. We calculate the median estimates and corresponding empirical 95% confidence intervals for M =100, 200, 300, 400, 500, 600, 700, 800, 900, and 1000 (Figure 2.2). At $M = 100$, the median estimate is the expected $\widehat{M} = 100$ across allele count bins and the average estimates ranges between 101.2 and 103.7 across allele count bins. At the highest simulated migration rate, $M = 1000$, the median estimate is $\widehat{M}$ =1000, with average estimates ranging from 1032.2 to 1319.2 across bins. The relative bias of mutation rate estimates increases with increasing migration rates (Spearman's rho=0.9878, p-value <0.0001), as does the mean squared error, (Spearman's rho=1, p-value<0.0001) (Figure 2.10). However, the median estimates remain accurate at higher levels of migration and false positive rate based on the Spearman's rho test remain low. Based on 10,000 simulations, at the higher constant migration rate of M =1000, the false positive rate is 2.82%, comparable to the false positive rate for the low migration rate of M =100 at 3.70%.

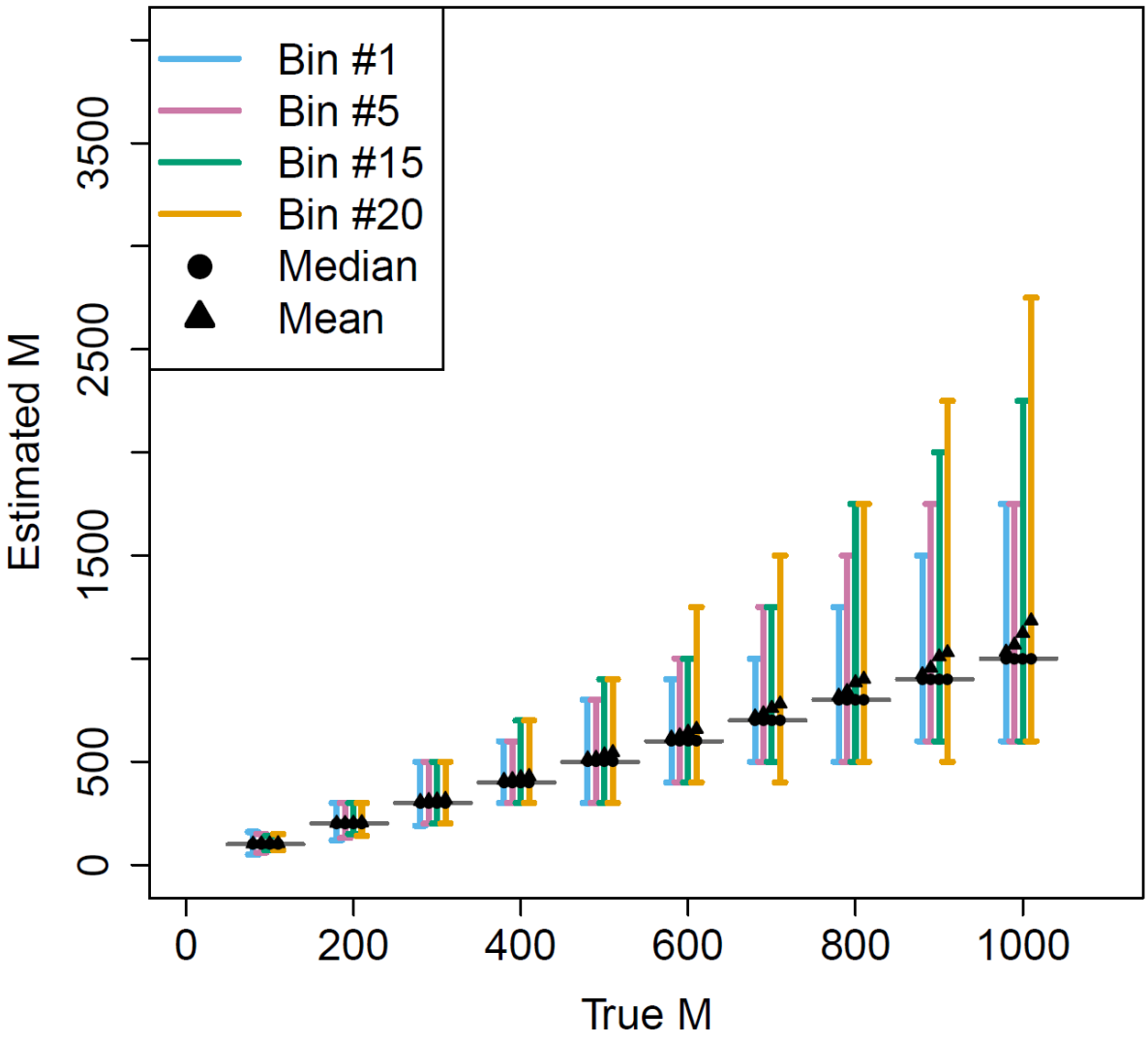

Figure 2.2 Estimated Migration vs. True Migration for Simulated Constant Migration
We show four median estimates (black dots) and four means (black triangles) across the range of allele count bins (from allele count bins 1, 5, 15, and 20) with their corresponding empirical 95% confidence intervals (indicated by error bars). The true parameter estimates M =100, 200, 300, 400, 500, 600, 700, 800, 900, and 1000 are shown by the gray line. The horizontal axis shows the estimated $\widehat{M}$ value and the vertical axis shows the corresponding true parameter estimated in each group.

2.3.2 Temporal change of migration rates

We next considered pairs of populations with increasing or decreasing migration rate to determine this method's ability to detect temporal change. To quantify this change in migration, we calculate Spearman's rank correlation coefficient (Spearman's rho). A positive significant rho value indicates decreased migration, whereas a negative significant rho value indicates increased migration. In addition, we apply the model comparison grid search procedure (see Methods) to quantify the change. We consider 4 scenarios.

In model (a), the ancestral divergence model, we simulate a historically unified population. At time τ, this population splits, becoming two distinct populations. The two new populations maintain a migration rate of M=100 from time τ to present day. We consider the scenarios where the time of split occurs at four different times: 0.01, 0.005, 0.001, and 0.0005 coalescent units, corresponding to approximately 400, 200, 40, and 20 generations before present day. Assuming 20 year generations, these times correspond to 8,000, 4,000, 800, and 400 years in the past, respectively. For each model, we calculate the migration rate estimate for each allele count bin. Repeating this procedure for 10,000 individual data sets, we calculate the mean estimate for each allele count bin (Figure 2.3A). The mean estimates range from 99.92 to 101.22 for τ =0.01, 98.5 to 121.36 for τ =0.005, 143.02 to 351.82 for τ =0.001, 286.3 to 609.99 for τ =0.0005. For every choice of τ, we observe increasing mean estimates of M with increasing allele counts, indicating a downward trend in migration in recent years. To quantify this trend, we calculate the rho estimates of each data set. The mean rho values were: 0.117, 0.327, 0.656, and 0.473 for τ = 0.01, 0.005, 0.001, and 0.0005 respectively (Table 2.1). These positive rho values correctly identify decreased gene flow, with the strongest signal in the more recent time changes (τ =0.001, τ =0.0005).

To assess our power to identify decreasing gene flow, we repeat this procedure for 10,000 simulations and recorded the number of simulations resulting in significant rho values for each model. In this ancestral divergence model, there are 9.8%, 34.6%, 96.8%, and 64.6% significant rho values, for τ = 0.01, 0.005, 0.001, and 0.0005 respectively, at the $\alpha = 0.05$ significance level with a two-sided hypothesis test (Table 2.1). This indicates the change in gene flow is increasingly detectable with more recent changes to a point. When the time of change is too recent (in this case, shown by decreased power at τ=0.0005), our ability to detect a change using Spearman's rho is diminished.

As our second model, model (b), we perform simulations with new migration between two isolated populations with historically zero gene flow. At time τ (τ= 0.01, 0.005, 0.001, and 0.0005), we begin symmetric, constant migration of M=100. In this setting, we expect to see lower migration estimates at higher allele counts, resulting in negative correlation coefficients. For each model, we simulate 100 data sets and estimated migration rates for each allele count bin. The mean migration estimates ranged from 70.0 to 101.62 for τ =0.01, 50.0 to 100.81 for τ =0.005, 10.65 to 81.29 for τ =0.001, 10.0 to 60.68 for τ =0.0005 (Figure 2.3B). The mean rho estimates per dataset are: -0.037, -0.570, -0.903, and -0.845 for time changes τ = 0.01, 0.005, 0.001, and 0.0005, respectively (Table 2.1). In contrast to the first model, we observe negative correlation coefficients, indicating increased migration in recent years. The mean rho value at the most distant time of change (τ= 0.01) is close to zero. Across 10,000 simulations, we observe 7.0%, 86.6%, 100%, and 100% significant rho values, respectively, at the α=0.05 significance level. Like our first model, we observe changing migration is more difficult to detect by significant rho values at more distant time points, due to the longer stretches of constant

migration levels capture by our rare variants. We observe more recent changes (τ =0.001, 0.0005) give consistently significant rho values of negative correlation.

To provide examples of smaller changes in migration, we generate two additional models: (c) migration decreases from M_2=100 in the past to M_1=50 in recent years and (d) migration increases from M_2=50 in the past to M_1=100 in recent years. In each case, we simulate the four different times of change ($\tau = 0.01,\ 0.005,\ 0.001,\ 0.0005$). Under each model we simulate 10,000 individual data sets, estimate migration rates for each allele count bin and calculate the mean estimate for each bin (Figure 2.3C-D). For model (c), the mean migration estimates range from 58.3 to 60.0 for τ =0.01, 51.5 to 61.4 τ =0.005, 58.9 to 87.4 for τ =0.001, 67.2 to 95.1 for τ =0.0005. As expected, we see positive correlation coefficients, indicating higher gene flow in the past than present times. The mean rho values based on these datasets are: 0.137, 0.293, 0.507, 0.427 for time changes t= 0.01, 0.005, 0.001, and 0.0005, respectively (Table 2.1). There are 10.7%, 30.4%, 73.9%, 57.7% significant rho values, respectively, at the α=0.05 significance level. For model (d), the mean migration estimates range from 101.3 to 103.2 for τ =0.01, 93.7 to 100.8 for τ =0.005, 63.9 to 94.2 for τ =0.001, 57.8 to 85.3 for τ =0.0005. This model has primarily negative correlation coefficients, indicating the increased migration in current times. The mean rho values in model (d) are: 0.053, -0.075, -0.473, and -0.405 respectively (Table 2.1). In model (d), there are 6.9%, 7.0%, 66.8%, 53.1% significant rho values, respectively (Table 2.1).

We note that in cases where the migration rate change only affects either the rarest frequency categories or the more common ones, it can be difficult to have power using Spearman's rho as the statistic is most sensitive to a trend that is continuous across all frequency categories.

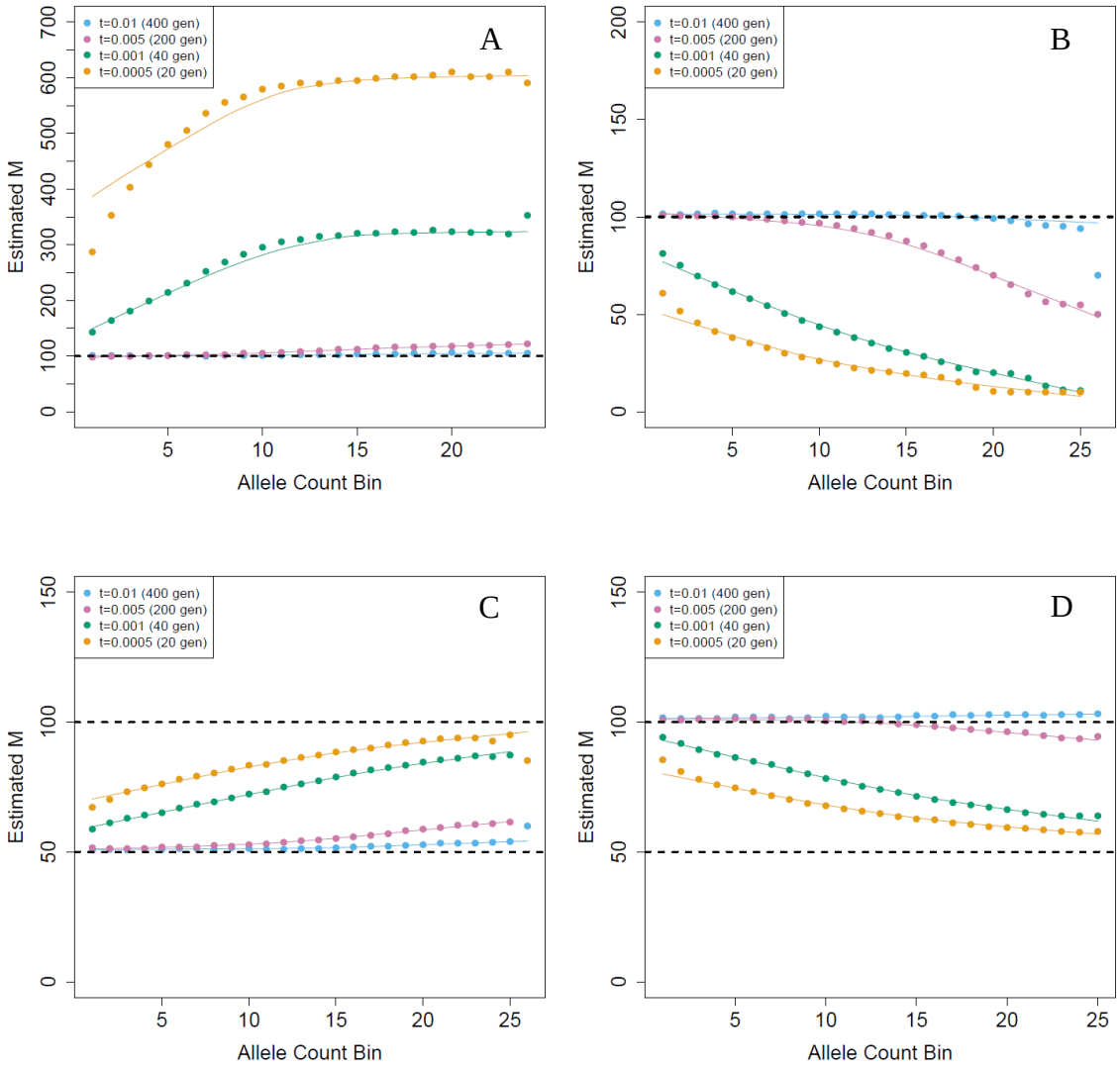

Figure 2.3 Estimated Migration for Change in Migration Models
We show four examples of changing migration. The vertical axis shows the estimated $\widehat{\mathrm{M}}$ value and the horizontal axis shows the allele count bin (with decreasing level of rareness with increasing bin number). Each point indicates the mean value in the allele count bin across the 10,000 simulations. The lines are generated using a loess curve on each set of correspondingly colored points. We use four different times of migration rate change: 0.01, 0.005, 0.001, and 0.0005 coalescent units, corresponding to 400, 200, 40, and 20 generations in the past respectively (constant effective population sizes=10,000 individuals, sample sizes=1,000 individuals each). (A) Ancestral divergence: Previously one population, two distinct populations form with current symmetric migration rates of M =100. (B) New migration: Previously isolated populations, at the time of change, the migration rate shifted to the current rate of M =100. (C) Two populations with past symmetric migration rates of M_2 =100 decrease to a migration of M_1 =50 in recent years. (D) Two populations with past symmetric migration rates of M_2=50 increase to a migration of M_1=100 in recent years.

Table 2.1 Spearman's Rho Test for Temporal Change Models
For 10,000 dataset simulations of the four examples of changing migration, ancestral divergence and increased and decreased migration, and the four different times of migration change, we calculated Spearman's rank correlation coefficient and recorded the number of significant signals (α=0.05). For comparison, in the previous simulations of constant migration, 370 of 10,000 simulations resulted in significant rho values, providing a false positive estimate of 3.70%.

Time of Change (Coalescent Units)	*(A) Ancestral Divergence*		*(B) New Migration*		*(C) M_2=100 to M_1=50*		*(D) M_2=50 to M_1=100*	
	Mean Rho	Power	Mean Rho	Power	Mean Rho	Power	Mean Rho	Power
0.01	0.117	0.098	-0.037	0.070	0.137	0.107	0.053	0.069
0.005	0.327	0.346	-0.570	0.866	0.293	0.304	-0.075	0.070
0.001	0.656	0.968	-0.903	1.000	0.507	0.739	-0.473	0.668
0.0005	0.473	0.646	-0.845	1.000	0.427	0.577	-0.405	0.531

2.3.3 Identifying Parameters of Changing Migration Rates

In addition to Spearman's rho, for models (c) and (d), we also apply the model comparison grid search approach. We simulate 1,000 datasets of 750 kb under each model. Using a grid of 657 histories (see Methods), we calculate the power to reject the null hypothesis of constant migration (Table 2.2). For model (c), migration decreases from M_2=100 in the past to M_1=50 in recent years, the proportion of datasets where the grid search algorithm selects a changing migration as the largest likelihood is 1.0 for τ =0.0005, 1.0 for τ =0.001, 0.996 for τ =0.005, and 0.976 for τ =0.01. The power to reject the null in these datasets was 0.926 for τ =0.0005, 0.831 for τ =0.001, 0.503 for τ =0.005, 0.06 for τ =0.01. For model (d), migration increases from M_2=50 in the past to M_1=100 in recent years, the proportion of datasets where the grid search algorithm selects a changing migration as the largest likelihood is 1.0 for τ =0.0005, 1.0 for τ =0.001, 0.996 for τ =0.005, and 0.985 for τ =0.01. The power to reject the null in these

datasets was 0.998 for τ =0.0005, 0.908 for τ =0.001, 0.12 for τ =0.005, 0.043 for τ =0.01. We observe the power is highest for the most recent time changes and decreases farther in the past. The power is particularly strong in the range of τ=0.0005 and 0.01. Farther in the past, the algorithm often identifies a changing migration as the most likely history, but there is not enough evidence to reject the null hypothesis. We also perform 1,000 simulations of 750 kb datasets with constant migration of M =100. We obtain a false positive rate, where we incorrectly reject the null hypothesis of constant migration of 4.2%. We observe the power to detect changing migration for these models is substantially improved for changes more recent than $\tau = 0.01$ with the grid search algorithm compared to Spearman's rho.

Table 2.2 Grid Search Results for Identifying Changing Migration
For 1,000 dataset simulations of two examples of changing migration with four different times of migration change, we apply the grid search algorithm and record the proportion of the datasets that selected a changing migration model as the most likely. We also record how often there is evidence to reject a constant migration (power) and how often all three parameters are estimated correctly.

Model	Test Data Set Parameters			Proportion with Maximum Likelihood of Changing Migration	Power to Reject Constant Migration	Proportion Correct All Parameters
	M_1	M_2	τ			
(c)	50	100	0.0005	1.00	0.926	0.169
			0.001	1.00	0.831	0.353
			0.005	0.996	0.503	0.128
			0.01	0.976	0.060	0.031
(d)	100	50	0.0005	1.00	0.998	0.352
			0.001	1.00	0.908	0.466
			0.005	0.990	0.120	0.045
			0.01	0.985	0.043	0.010

Of the 1,000 data sets, the proportion that exactly identified the correct parameters in model (c) was 0.169 for τ =0.0005, 0.353 for τ =0.001, 0.128 for τ =0.005, 0.031 for τ =0.01. While all the parameters are not necessarily exactly estimated, the individual average estimated parameters are

consistently close to the true underlying value (Figure 2.4, Figure 2.5). For the time estimates in model (c), the mean across datasets was 0.0011 (median: 0.00075, 95% CI: [0.00025, 0.005]) for τ =0.0005, 0.0014 (median: 0.001, 95% CI: [0.00025, 0.005]) for τ =0.001, 0.0063 (median: 0.005, 95% CI: [0.005, 0.025]) for τ =0.005, 0.011 (median: 0.0075, 95% CI: [0.00025, 0.025]) for τ =0.01. Similarly, for the time estimates in model (d), the mean across datasets was 0.00071 (median: 0.0005, 95% CI:[0.00025, 0.0025]) for τ =0.0005, 0.0016 (median: 0.001, 95% CI:[0.00025, 0.0075]) for τ =0.001, 0.0061 (median: 0.005, 95% CI:[0.00025, 0.025]) for τ =0.005, 0.0064 (median: 0.005, 95% CI:[0.00025, 0.025]) for τ =0.01 (Figure 2.4). We observe the precision for estimating time is best for recent time points.

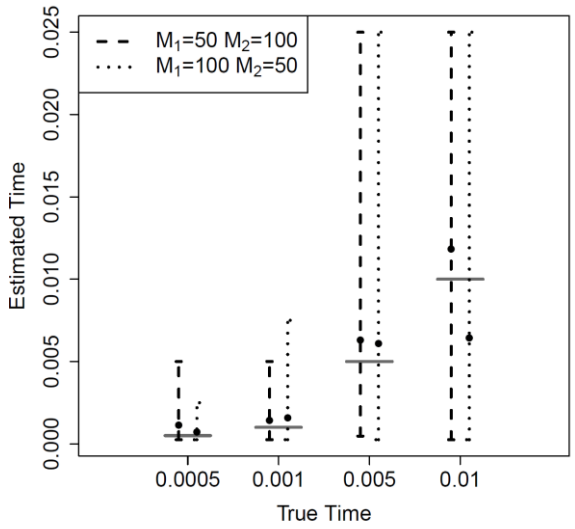

Figure 2.4 Estimation of the Time Parameter for Changing Migration
For 1,000 data simulations of two examples of changing migration with four different times of migration rate change, we attempt to estimate the time of change in each dataset, showing here the mean estimate across datasets and the 95% confidence interval. The times are 0.01, 0.005, 0.001, and 0.0005 coalescent units, corresponding to 400, 200, 40, and 20 generations in the past respectively (constant effective population sizes=10,000 individuals, sample sizes=1,000 individuals each). In the first model, two populations with past symmetric migration rates of M_2=100 decrease to a migration of M_1=50 in recent years (dashed line). In the second model, two populations with past symmetric migration rates of M_2=50 increase to a migration of M_1=100 in recent years (dotted line). We show the true value being estimated in the solid grey horizontal line.

In estimating the two migration rates, M_1 the recent migration, and M_2 the past migration, we observe the precision for both models is high (Figure 2.5). For model (c), the M_1 median estimates are consistently equal to 50 with average M_1 estimates of 49.3 (95% CI: [0, 75]) for τ =0.0005, 46.5 (95% CI: [25, 75]) for τ =0.001, 50.2 (95% CI: [50, 75]) for τ =0.005, 50.2 (95% CI: [25, 75]) for τ =0.01. For model (d), the M_1 median estimates are also consistently equal to the true value of 100 with average M_1 estimates of 107.8 (95% CI: [75, 150]) for τ =0.0005, 103.9 (95% CI: [75, 150]) for τ =0.001, 101.9 (95% CI: [75, 125]) for τ =0.005, 98.7 (95% CI: [75, 125]) for τ =0.01. For the estimates of M_2, the estimates are less precise with larger confidence intervals, but the median and average estimates remain accurate (Figure 2.5). For model (c), the M_2 average estimates are upwardly biased: 114.8 (median: 100, 95% CI: [100, 200]) for τ =0.0005, 113.2 (median: 100, 95% CI: [75, 200]) for τ =0.001, 128.64 (median:125, 95% CI: [50, 200]) for τ =0.005, 100.1 (median: 100, 95% CI: [0, 200]) for τ =0.01. For model (d), the M_2 average estimates are 48.2 (median: 50, 95% CI: [25, 50]) for τ =0.0005, 46.1 (median: 50, 95% CI: [0, 75]) for τ =0.001, 69.5 (median: 75, 95% CI: [0, 200]) for τ =0.005, 101.47 (median: 100, 95% CI: [0, 200]) for τ =0.01. We observe again the precision for estimating both migration rates is best for recent time points.

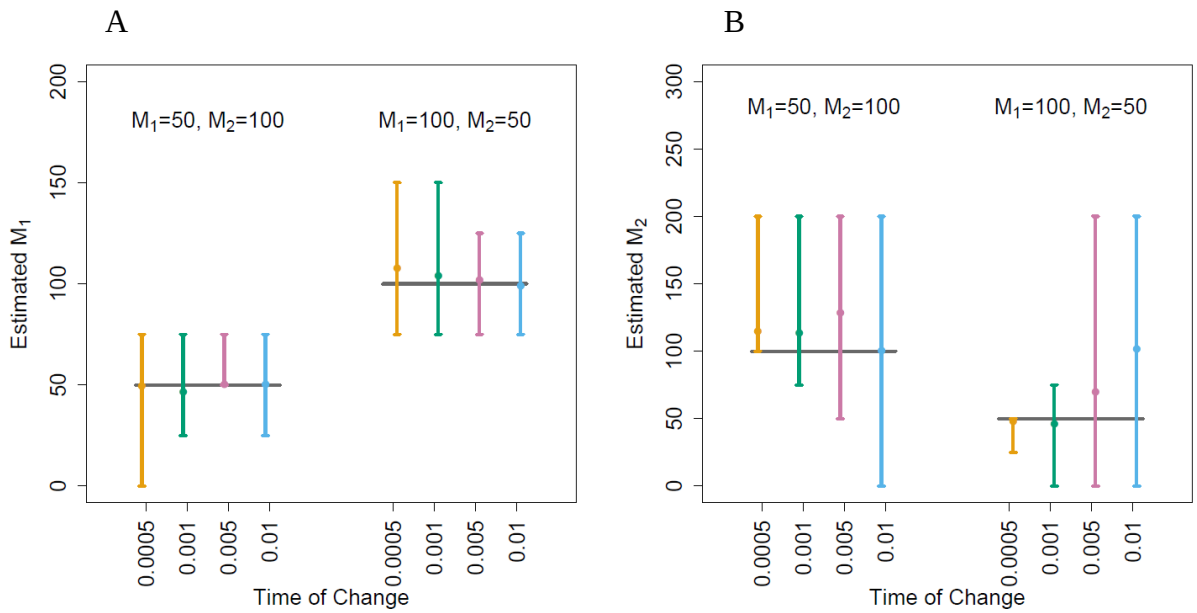

Figure 2.5 Estimation of the Migration Rates in Change of Migration Models
For 1,000 data simulations of two examples of changing migration with four different times of migration rate change, we attempt to estimate the recent migration (M_1) and past migration (M_2) in each dataset, showing here the mean estimate (dot) across datasets and the 95% confidence interval (bar). The times of change are 0.01, 0.005, 0.001, and 0.0005 coalescent units, corresponding to 400, 200, 40, and 20 generations in the past respectively (constant effective population sizes=10,000 individuals, sample sizes=1,000 individuals each). In the first model, two populations with past symmetric migration rates of M_2 =100 decrease to a migration of M_1=50 in recent years. In the second model, two populations with past symmetric migration rates of M_2 =50 increase to a migration of M_1=100 in recent years. We show the true value being estimated in the solid grey horizontal line.

We note that we apply this method to a small dataset of 750 kb. With increased availability of whole exome and whole genome data, this data size can now feasibly be much larger. We repeat this procedure for a sample size equal to that of whole exome sequencing (30 MB) and see that both precision and power increases (Table 2.6). The range of time changes in which there is power to reject constant migration increases, including as distant as τ=0.005. At the time ranges of τ=0.0005 and τ=0.001, this data size allows for precise estimation of parameters. We also perform 1,000 simulations of 30000 kb datasets with constant migration of M =100. We obtain a false positive rate, where we incorrectly reject the null hypothesis of constant migration, of 4.7%.

2.3.4 Exponential Growth

Previously, we assumed population size in both populations to be constant. To consider model adjustments in cases of exponential growth, we simulate a baseline growth model of two islands with growth in effective population sizes from 10,000 to 1,000,000 in 500 generations and a constant, symmetric migration of $M = 100$ ($m = 0.000025$). Without applying adjustments, the method incorrectly indicates a substantial increase in migration in recent years (Figure 2.11). To adjust for population growth, we re-estimate the $p_{\boldsymbol{M},k,i}$ in the likelihood equation based coalescent-simulated data under a model of exponential growth. Under these expansion-adjusted parameters, we expect the migration estimates to be close to $M = 100$ across allele count bins, newly scaled based on the present-day population size. In the first allele count bin, which consists entirely of doubletons, the median migration estimate was 100 with a mean of 101.5 (95% CI: [50,150]). In the highest allele count bin, with minor allele counts between 88 and 100 across datasets, the median migration estimate was 100 and mean of 106.6 (95% CI: [50,150]). Across all allele count bins, the median estimates of the migration rate are consistently at 100. The mean estimates range from 94.56 to 109.89. We observe this change properly adjusts the method to once again provide accurate estimates based on the present-day size. To assess our power to distinguish changing gene flow from exponential growth, we repeat this procedure for 10,000 simulations and recorded the number of simulations resulting in significant rho values for each model. We observe a false positive rate of 2.4%.

We also find these results are robust when the growth rate is underestimated in the adjustment for exponential growth. We simulate a two-population island model with a growth in effective population sizes of 10,000 to 5,000,000 in 500 generations and a constant, symmetric migration of $M = 100$ ($m = 0.000025$). Using $p_{\boldsymbol{M},k,i}$ estimated under this growth model, the

false positive rate is 1.32%. In contrast, using the $p_{\boldsymbol{M},k,i}$ estimated under our baseline growth model, the false positive rate is 1.92%. We also simulate a two-population island model with a growth in effective population sizes of 10,000 to 500,000 in 500 generations and a constant, symmetric migration of $M = 100$ ($m = 0.000025$). Using the $p_{\boldsymbol{M},k,i}$ estimated over our baseline growth model to provide an example of overestimating the adjustment, the false positive rate is 11.36%.

Like the case of constant population size, we generate two additional change of migration models under exponential growth with correctly applied adjustments: (a) migration decreases from M_2=10000 in the past to M_1=5000 in recent years and (b) migration increases from M_2=5,000 in the past to M_1=10,000 in recent years. In each case, we simulate four different times of change ($\tau = 0.0001,\ 0.00005,\ 0.00001,\ 0.000005$ coalescent units), corresponding to approximately 400, 200, 40, and 20 generations before present day. With 20 year generations, these time scales are 8,000, 4,000, 800, and 400 years in the past, respectively. We perform these simulations for 10,000 individual data set simulations and calculated the average estimate under each model for each allele count bin (Figure 2.12). We observe clear signals of changing migration comparable to the case of constant population size within this time scale. For model (a), this signal of decreasing migration is well defined across values of τ, though the point estimates of migration rate for specific minor allele count bins are slightly biased upwards for the most recent times of change ($\tau = 0.00001, 0.000005$). For model (b), we see clear signals of increasing migration rate, except for the most distant time of change ($\tau = 0.0001$). Therefore, while the temporal picture of changing migration is distinct, the estimates are most accurate where much of recent history is spent at lower migration levels (M =5000).

We also apply the grid search algorithm to these two models under exponential growth. We simulate 1,000 datasets of 750 kb under each model with a growth in effective population sizes of 10,000 to 5,000,000 in 500 generations. The grid used included 657 histories with the same growth parameters and each combination of: recent migration (M_1): 0, 2500, 5000, 7500, 10000, 12500, 15000, 17500, 20000, past migration (M_2): 0, 2500, 5000, 7500, 10000, 12500, 15000, 17500, 20000, and time of change in coalescent units in the past from present day (τ): constant, 0.0000025, 0.000005, 0.0000075, 0.00001, 0.000025, 0.00005, 0.000075, 0.0001, 0.00025. We observe power to reject constant migration comparable to the case of constant population size (Table 2.7). Under this model of exponential growth, the power is highest at $\tau = 0.00005$ and $\tau = 0.0001$. As with the constant population size case, we repeat these simulations for the larger data size of 30000 kb for exome sequencing data (Table 2.8). With this larger data size, the power to reject constant migration is extremely high and the precision of estimates is greatly improved.

2.3.5 Application to Sequence Data

African American and European Populations

We next evaluate our method's performance by estimating migration using counts from sequence data for African American and European individuals which enabled us to contrast our method to a known historical record of migration changes. For a large portion of the past twenty thousand years, modern day Africans had very low gene flow with Europeans e.g. [83; 99]. Approximately 200 to 300 years ago, corresponding to ~10 generations in the past, the increased gene flow began between a smaller genetically isolated group of individuals from the larger African population and a subset of modern day European ancestors. This genetic exchange formed the African American population we observe today [100]. This historical model indicates

we should observe higher migration rates between the European sample and the African American sample in recent years and lower migration rates farther back in time.

To estimate this recent gene flow into the African American population from Europeans, we adjusted our method by calculating the probabilities used in the maximum likelihood equation, $p_{\boldsymbol{M},k,i}$, for directional, asymmetric migration with exponential growth (growth in effective population sizes from 10,000 to 1,000,000 in 500 generations) and corresponding sample sizes of 7470 (Europeans) and 339 (African Americans). The resulting migration estimates ranged from M=900 (allele count bin number 34 with counts 73-79) to 9750 (allele count bin number 1 with only doubletons). This corresponds to m=0.0009 to 0.009750 assuming a current effective population of 1,000,000. The graph of the migration estimates across allele count bins indicates higher estimates in the recent past decreasing farther back in time Figure 2.6). We compute a Spearman's rho value of -0.7692, $p < 0.00001$ under a two-sided alternative. We observe a similar trend when this procedure is repeated for equal sample sizes (down-sampling the European sample to 339 randomly selected individuals) (Figure 2.13). This temporal trend is consistent with the increased migration in recent years with a downward trend back in time.

Using the grid search algorithm, we also estimate the parameters of an initial migration rate, time of change, and a final (recent) migration rate. These results indicate the most likely model is of constant migration with M =5000. This may be a function of the sparse grid used which included only 4 migration rates and 4 times of change, with the most recent at 20 generations prior to present day. Future analyses will include a denser grid for more refined estimate of change.

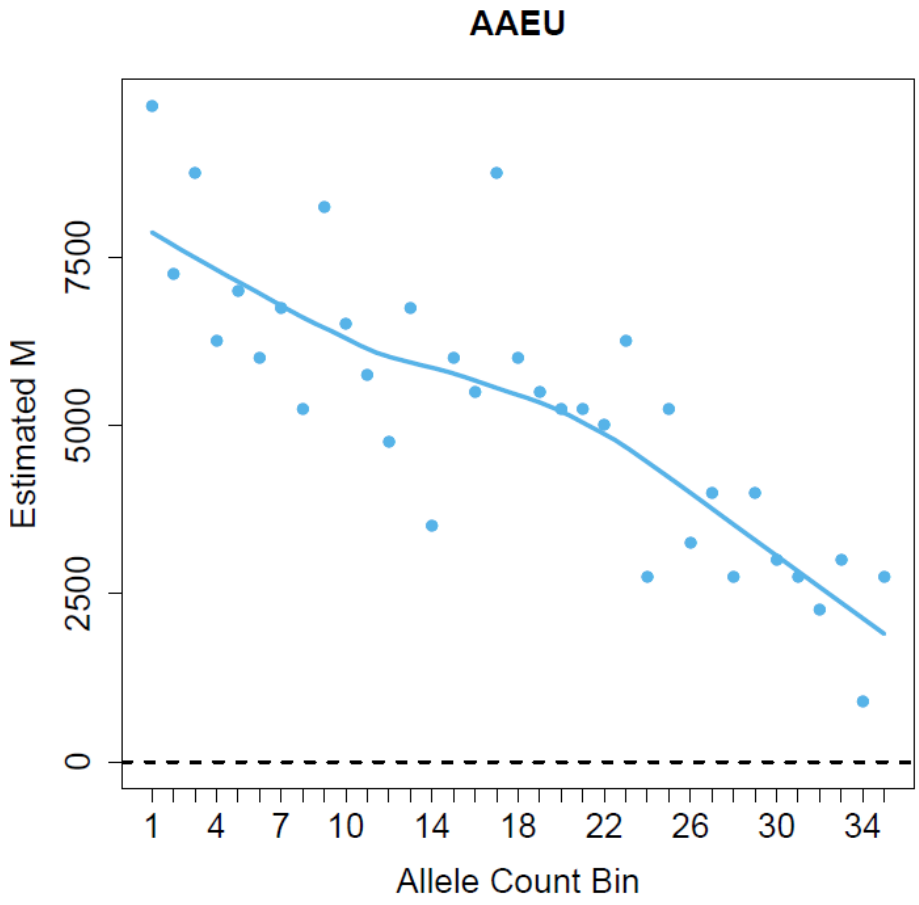

Figure 2.6 Estimated Migration for African-Americans and Europeans
Estimates of the migration rates using probabilities based on populations of sample sizes 7470 and 339, undergoing exponential growth, with migration in a single direction.

European Populations

We use our method to estimate migration rates between four subpopulations within Europe. These subpopulations are defined primarily by the United Nations geo-scheme: Northwestern (British Isles), Northern (Iceland, the Faroe Islands, Denmark, Norway, Sweden, Estonia, Latvia, and Lithuania), Western (Belgium, France, Luxembourg, and the Netherlands), and Central (Austria, Germany, and Switzerland)[10]. The estimates were derived by randomly selecting 700 individuals from each population. For each variant in the sequencing dataset, we calculate the minor allele count both between and within populations. We estimate migration

rates between each pair of populations using probabilities, $p_{\boldsymbol{M},k,i}$, based on symmetric migration with exponential growth (growth in effective population sizes from 10,000 to 1,000,000 in 500 generations) and sample sizes matching those of each population sample (Figure 2.14). Across all the population pairs, the migration rates range from M =3,000 to M =15,000. To quantify these patterns, we calculate Spearman's rho in the six population pairs. For example, the Northern-Western estimates range from M =3,000 to 7,000, with a negative rho value of 0.751 (p = 0.00000172) (Figure 2.7A, Table 2.3) and Western-Northwestern estimates range from M =7,000 to 14,000 with a negative rho value of -0.249 (p = 0.193) (Figure 2.7B, Table 2.3). Three pairs show evidence of decreasing migration rates (Northern-Central, Northern-Western, and Northern-Northwestern), indicated by significantly positive Spearman's rho values. The remaining three pairs indicate increasing migration, (Central-Western, Central-Northwestern, and Western-Northwestern), based on negative rho values (Table 2.3, Figure 2.14). The p-value for Central-Western is significant while these other negative rho values are non-significant.

Using the grid search algorithm, we perform an additional test of changing migration rate and estimate the parameters: initial migration rate, time of change, and a final (recent) migration rate (Table 2.4). We find the parameter estimates indicating increasing and decreasing migration rates match the direction indicated by the sign Spearman's rho. In contrast with Spearman's rho, the algorithm rejects the null hypothesis of constant migration rate in each population pair. We observe, however, the population pairings that did not reject constant migration with Spearman's rho also have the smallest estimated changes in migration rate (Table 2.3, Table 2.4). For example, the Central-Northwestern and Western-Northwestern, the estimated change in parameters is small (M_2 =7,500 to M_1=10,000). The times of change for the increasing migration rates ranged from 0.00001 to 0.000025 coalescent units (40 to 100 generations in the past). The

estimates for the Central-Western pairing indicate a large increase in migration, with M_2=5000 to M_1=17500 and a time of change at 0.000025 coalescent units in the past. As also reflected by the negative Spearman's rho values, the estimates for Northern-Central, Northern-Western, and Northern-Northwestern each indicate decreasing migration rates. The Northern-Northwestern pairing has an estimated strong decrease in migration 0.0001 coalescent units in the past (400 generations) from M_2=20,000 to M_1=5000. The Northern-Central and Northern-Western have smaller estimated changes in migration from M_2=7500 to M_1=2500 at 0.000025 coalescent units in the past (100 generations) and M_2=10000 to M_1=2500 at 0.00005 coalescent units in the past (200 generations).

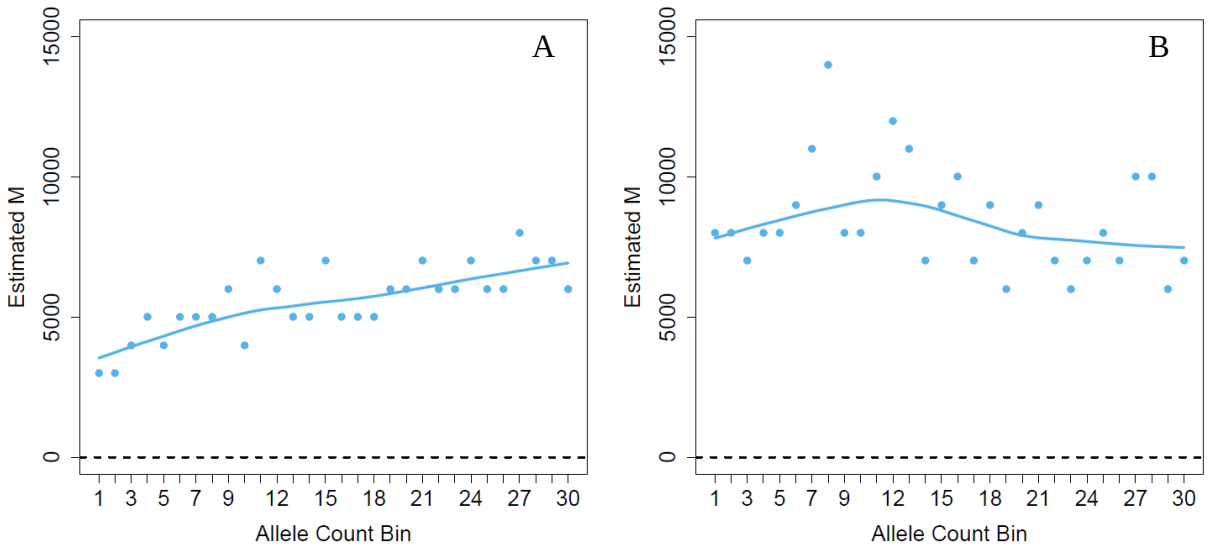

Figure 2.7 Estimated Migration for European Populations
(A) Estimated migration between Western and Northern Europe (Nordic-Baltic). (B) Estimated migration between Western Europe and Northwestern (Britain and Ireland).

Table 2.3 Spearman's Rho Test for Association for European Populations

Populations	*Spearman's Rho*	*P-value*
Northern-Central	0.312	0.0995
Northern-Western	0.751	0.00000172
Northern-Northwestern	0.664	0.0000867
Central-Western	-0.519	0.00392
Central-Northwestern	-0.249	0.193
Western-Northwestern	-0.288	0.1233

Table 2.4 Grid Search Results with Parameter Estimations for European Populations

Populations	*Estimated Parameters*			*Reject*
	M_1	M_2	*Time*	
Northern-Central	2500	7500	0.000025	Yes
Northern-Western	2500	10000	0.00005	Yes
Northern-Northwestern	5000	20000	0.0001	Yes
Central-Western	17500	5000	0.000025	Yes
Central-Northwestern	10000	7500	0.000025	Yes
Western-Northwestern	10000	7500	0.00001	Yes

2.4 Conclusion

We propose a novel method for using rare variants to identify changing migration patterns. Based on the configuration of alleles across populations, we obtain one maximum-likelihood estimate of the migration rate for each minor allele count. By comparing estimated migration rates for a range of minor allele counts, we generate a temporal picture of gene flow between populations. An increase (or decrease) in migration is indicated by the higher (or lower) migration estimates as the variants become rarer. To quantify these changes, we present two options: a quick initial test using Spearman's rank correlation coefficient (Spearman's rho) and a refined model comparison grid search with parameter estimates. We show our method can detect changes of migration rates not only for fundamental changes in migration, such as when two populations were previously a single ancestry population or when new migration begins in two isolated populations with historically zero gene flow, but also for smaller changes in mutation

rate. With samples of 1000 individuals each, our method captures shifts in migration rate that occurred recently (20-40 generations in the past). Our method is less likely to identify more distant shifts over 200 generations in the past in this setting. Similarly, for changes that occurred less than 20 generations ago, our power decreases slightly. In this case, the signal is diluted by longer runs of constant migration estimates up until the time change. We also observed that migration estimates become less reliable at extremely high levels of gene flow. Thus, in populations with consistently high migration, our method has less power to establish accurate estimates.

We apply this methodology to the sequence data for African American and European individuals to evaluate our method in a case where there is a historical record of migration changes. In this case, we incorporate a directional, asymmetric migration pattern with exponential growth. As expected, the temporal picture created from our analysis shows an extreme increase in gene flow from Europeans into the African-American population in recent years. This evidence is consistent with admixture following the transatlantic slave trade, beginning in the mid-fifteenth century [101]. While the grid search algorithm did not identify changing migration, we believe this is a result of the sparse grid applied. Denser grids will give more precise estimates of changing migration and we plan to include this in a future analysis.

We further use this methodology to estimate migration rates between four closely related European populations: Northwestern (British and Irish), Northern (Nordic and Baltic), Western, and Central. Based on the close geographical and historical relationships between these regions, we expect to observe consistently high gene flow between pairs of populations. As Ralph and Coop [102] observe in their identity-by-descent analysis and as indicated in studies of isolation-by-distance effect [103; 104], we expect lower migration from the more geographically distant

populations, such as the Nordic-Baltic region. The results from our method reflect this expectation, with the geographically proximate Central-Western pair indicating overall high gene flow and continued increasing migration into present-day. We also observe significant changes in migration rates in several pairs of populations. For the Northern-Central, Northern-Western, and Northern-Northwestern pairings, our method indicates that migration rates were higher in the past than they are in present-day. This observation is unlikely to be the result of model specifications as we show our method is robust to the underestimation of exponential growth rates. Using the grid search algorithm, our results suggest that these changes occurred for the Northern-Central, Northern-Western, and Northern-Northwestern pairings approximately 100, 200, and 400 generations in the past, respectively. At this time, we observe the migration between these populations reaches a lower rate. We hypothesize this signal is the result of the Yamnaya steppe herders entering Europe from the East, creating new populations and spreading current populations [27]. As this Yamnaya population entered Europe, Allentoft *et al*, based on f_3 statistics, indicate this influx resulted in admixture, forming the Corded Ware population [28]. This Corded Ware population is inferred to have spread north and west through Europe, creating a cline of genetic affinity with Yamnaya[28]. We speculate this is the past high gene flow we observe in the population pairs with Northern Europe. Haak *et al* found this steppe migration continues until approximately 3,000 years ago, aligning with the timescale of the decline in migration rates we observe through recent years[27]. Our results, therefore, provide support for understanding this complex historical admixture of present-day Europeans.

While we show the range of applicability of our method, we recognize there are several limitations to our current approach. The first caveat is the detectable time range is dependent on scale of data (sample size) and the population history. For the time scales we consider, increasing

the data size to exome sequencing substantially improves power to detect change in migration rate and precision in estimating parameters. Also, the initial process focuses on providing a qualitative description of migratory history. Precise quantitative descriptions require a further step, applying the model comparison grid search algorithm. Choosing the histories to include in this grid can be somewhat subjective, though the initial analysis should inform this set. We note the precision of these time estimates is restricted by the histories included in the grid. A follow-up with a denser grid could yield more precise estimates where required. Finally, to establish the appropriate probability values, some knowledge of the population history is necessary. While the method is clearly robust to small misspecifications, we require a general understanding of the previous population size and migration direction to accurately apply this method.

In summary, we establish a flexible method for estimating migration and detecting temporal patterns while allowing for demographic changes. Using simulated and real sequence data we show the applicability of our method for a wide range of population scenarios. Identifying such temporal trends in migration will help to identify history patterns in human demography.

2.5 Appendix

2.5.1 Robustness to Misspecifications

Previously, we simulated the $p_{\boldsymbol{M},k,i}$ values with parameters corresponding to the populations of interest including sample size, effective population size, ancestral populations, direction of migration, and relative subpopulation sizes. We now assess the performance of our method when these parameters are incorrectly defined. We generated 10,000 simulations of constant migration (M =100), under a range of parameters and counted the number of simulations falsely identified as changing migration using Spearman's rho (Table 2.5).

Deep Ancestral Divergence

Most current population genetics models support an out-of-Africa hypothesis, indicating modern human populations derive from a single ancestral population e.g. [105; 106-109]. To maintain the simplicity of the model, it is important to determine if including this population history in calculating $p_{\boldsymbol{M},k,i}$ values is necessary to correctly identify a temporal pattern of migration. We simulated a population model where the two distinct populations were previously one population 246,160 years ago and applied the original maximum likelihood equation using probabilities that do not include this change to see how this affected estimates. These populations have a constant migration rate of M =100 with constant effective population sizes 10,000 individuals and sample sizes of 1000 individuals each. With this ancestral divergence condition, median estimates were consistently M =100 across allele count bins. The false-positive rate, 3.70%, for changing migration was similar to that where ancestral divergence was not included, 4.34% (Table 2.5). Therefore, the inclusion or exclusion of deep ancestral divergence history does not affect these more recent migration estimates.

Imbalanced Effective Population Size

Previously we assumed an effective population size of 10,000 individuals. We next considered the case where the effective population sizes are incorrectly specified, specifically when one population is larger than the other. We simulated two different cases of this misspecification, using simulations of constant migration of $M = 100$. In the case of small inaccuracies in effective population size, with 8,000 and 12,000 instead of 10,000 and 10,000, the false positive rate of changing migration remains well controlled at 5.68%. For the larger discrepancy, 5,000 and 15,000, the false positive rate of changing migration increases to 20.58% (Table 2.5). For small discrepancies, while the magnitude of estimates may no longer be accurate under these misspecifications, the temporal pattern is still identifiable.

Asymmetric Migration

We considered the case of asymmetric migration in which migration occurs from population 1 to population 2 more frequently than migration from population 2 to population 1. We simulated this scenario with constant effective population sizes of 10,000 individuals, sample sizes of 1,000 individuals each, with migration rates in three different ratios: 1:0.25 (M_1=100 and M_2=25), 1:0.5 (M_1=100 and M_2=50), and 1:0.75 (M_1=100 and M_2=75). In the case of mild asymmetry 1:0.75, the false positive rate for changing migration was 4.93%. For 1:0.5, the false positive rate for changing migration was 5.14%. In the extreme directional migration case of 1:0.25, the false positive rate for changing migration was 4.93% (Table 2.5). In each of these scenarios, we note that while estimates may change, the temporal pattern of constant migration remains clear with a controlled false positive rate. Therefore, even under these misspecifications, a temporal trend could be detected with this method.

Table 2.5 False Positive Rates of Changing Migration Under Parameter Specifications
Using Spearman's rho for 10,000 simulations of constant migration (M =100), under a range of parameters, we recorded the number of simulations falsely identified as changing migration.

Parameter Specifications	*False Positive Rate*
Correct Specifications	3.70%
Out-of-Africa Hypothesis	4.34%
Misspecifications of Effective Population Size	
8,000 and 12,000	5.68%
5,000 and 15,000	20.58%
Asymmetric Migration	
1:0.25	4.83%
1:0.50	5.14%
1:0.75	4.93%

2.5.2 Supplementary Figures and Tables

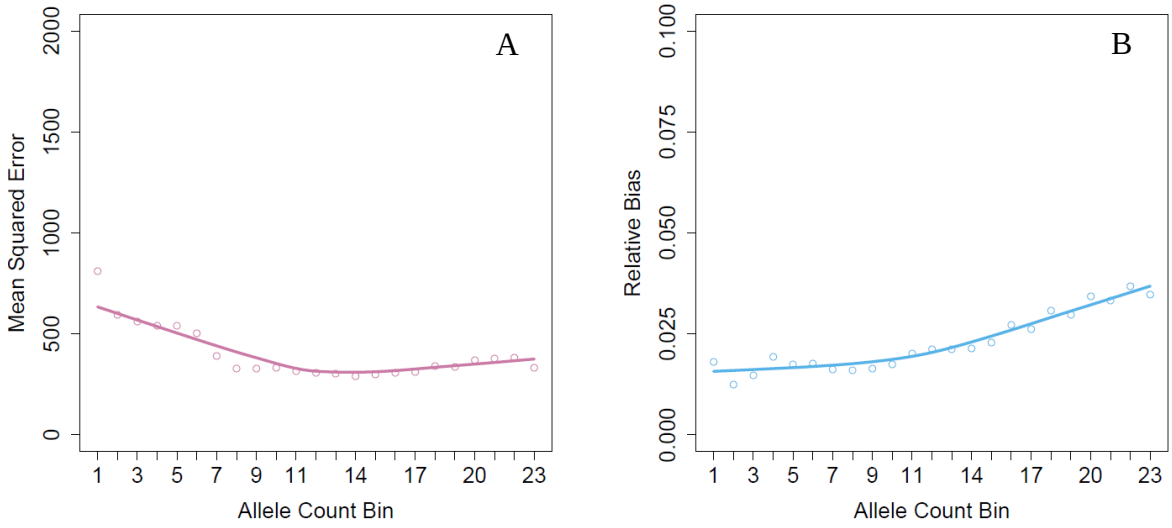

Figure 2.8 Mean Squared Error and Bias by Allele Count Bin Under Constant Migration
(A) Relative mean squared error for each allele count bin for 10,000 simulations of $M = 100$. The MSE is lowest at the intermediate allele count bins but remains stable across allele count bins. (B) Relative bias based on the 100 simulations for each allele count bin. There is an increase in overestimation, resulting in an upward bias, at the higher allele count bins

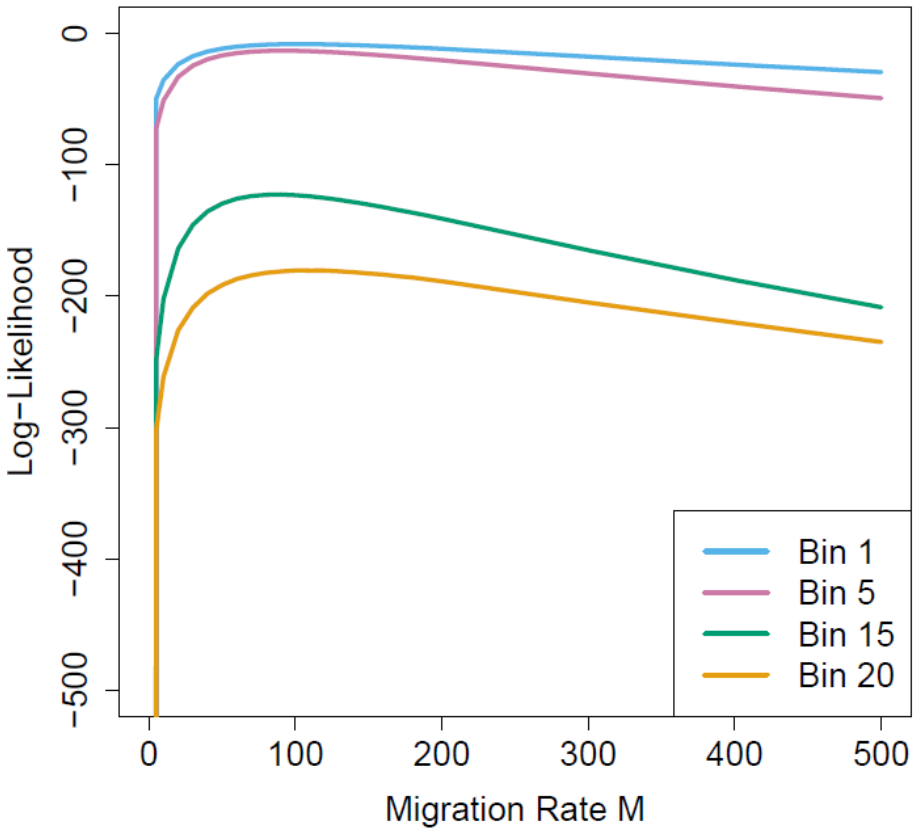

Figure 2.9 Likelihood Curves for One Simulation of Constant Migration $M = 100$
Four likelihood curves from a range of allele count bins. The curve becomes very flat with increasing migration rate, indicating any bias in estimates is likely to be upward.

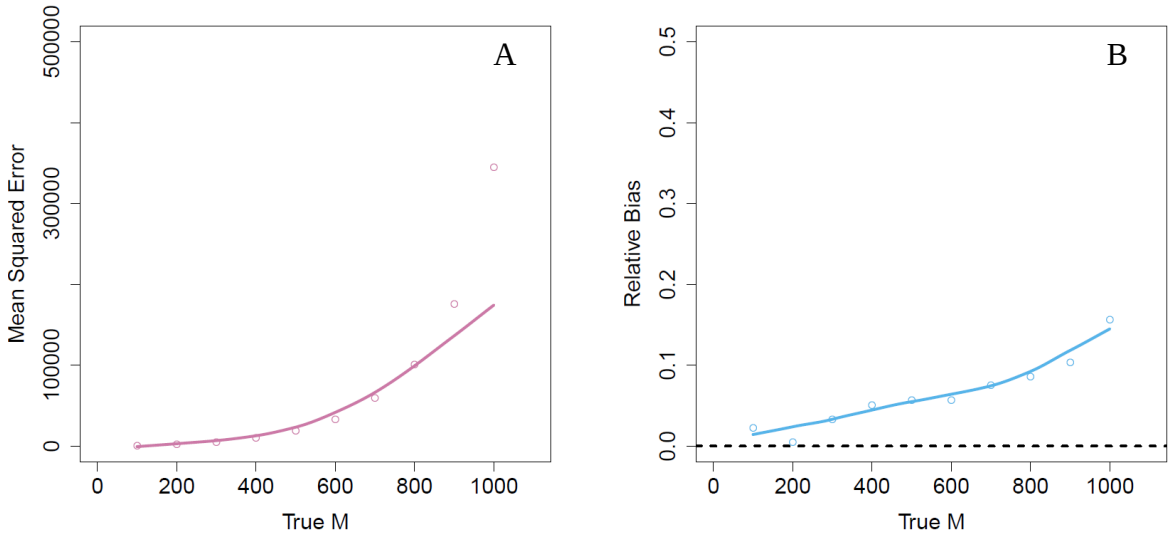

Figure 2.10 Mean Squared Error and Bias by Migration Rate under Constant Migration
(A) Mean squared error across allele count bins based on the 10000 simulations for each migration rate. There is an increase in bias and variance in estimates, resulting in overall precision loss. (B) Relative bias across allele count bins based on 10000 simulations for each migration rate.

Table 2.6 Grid Search Results for Identifying Changing Migration with Whole Exome Data
For 1,000 dataset simulations of 30,000 kb (~whole exome data size) of two examples of changing migration with four different times of migration change, we apply the grid search algorithm and recorded the proportion of the datasets that selected a changing migration model as the most likely. We also record how often there is evidence to reject a constant migration (power) and how often all three parameters are estimated correctly.

Model	Test Data Set Parameters			Proportion with Maximum Likelihood of Changing Migration	Power to Reject Constant M	Proportion Correct All Parameters
	M_1	M_2	τ			
(c)	50	100	0.0005	1.00	1.00	0.985
			0.001	1.00	1.00	1.00
			0.005	1.00	1.00	0.889
			0.01	0.975	0.533	0.17
(d)	100	50	0.0005	1.00	1.00	1.00
			0.001	1.00	1.00	1.00
			0.005	1.00	0.999	0.790
			0.01	0.950	0.064	0.055

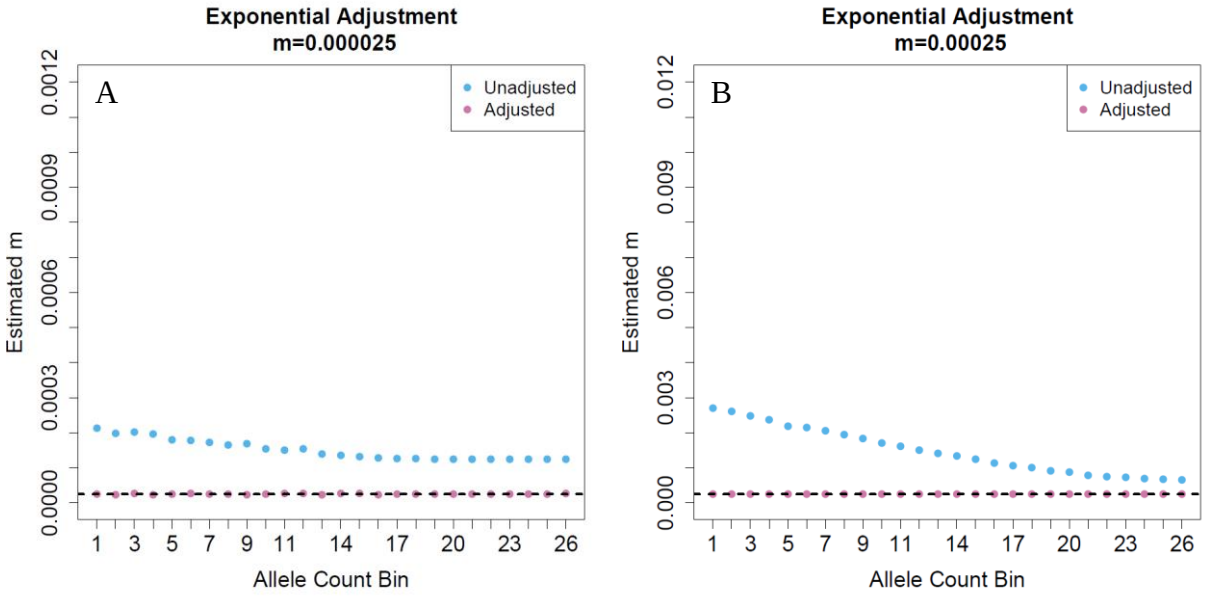

Figure 2.11 Adjustments for Exponential Growth
Two populations with an exponential growth in effective population sizes of 10,000 to 1,000,000 in 500 generations and a constant migration of (A) (left) $M = 100$ $(m = 0.000025)$, (B) (right) $M = 1000$ $(m = 0.00025)$. Without applying any adjustments (blue line), the method incorrectly indicates a high level of migration in the past and a substantial increase in migration in recent years. The adjustment corrects for this bias (pink line), indicating a constant migration at the expected rate (black dashed line).

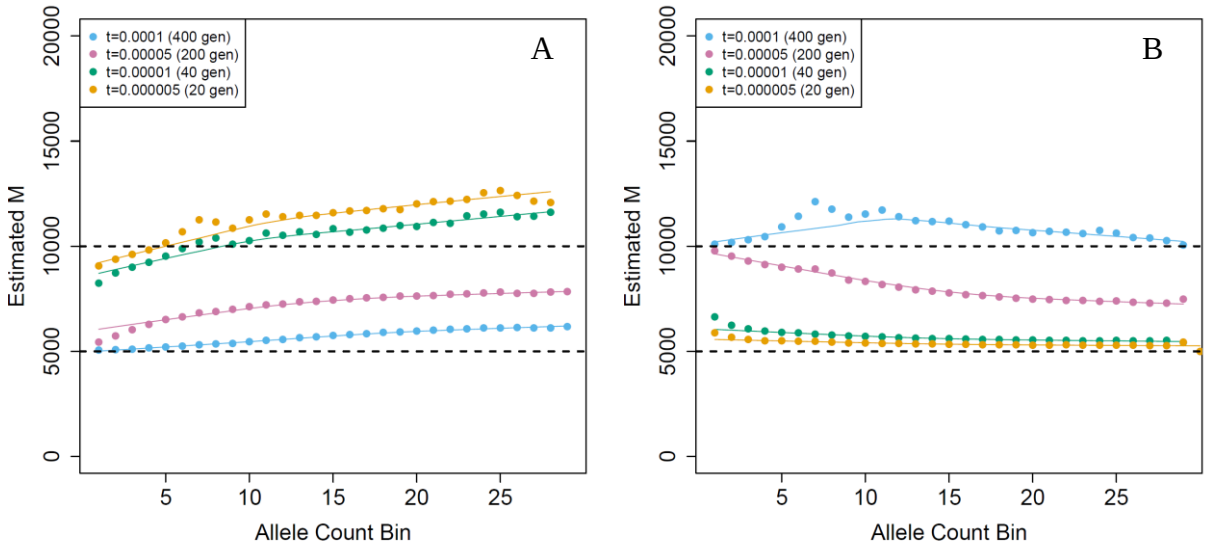

Figure 2.12 Estimated Migration for an Exponential Growth Change in Migration Model
We show two examples of changing migration under exponential growth with the correctly applied adjustments. (A) Two populations with past symmetric migration rates of M_2 =10000 decrease to a migration of M_1 =5000 in recent years. (B) Two populations with past symmetric migration rates of M_2 =5000 increase to a migration of M_1 =10000 in recent years. We use four different times of migration rate change: 0.0001, 0.00005, 0.00001, and 0.000005 coalescent units, corresponding to 400, 200, 40, and 20 generations in the past respectively (population growth from 10,000 to 1,000,000 in 500 generations, sample sizes=1,000 individuals each).

Table 2.7 Grid Search Results for Identifying Changing Migration with Exponential Growth
For 1,000 dataset simulations of 750 kb of two examples of changing migration with exponential growth for four different times of migration change, we apply the grid search algorithm and recorded the proportion of the datasets that selected a changing migration model as the most likely. We also record how often there is evidence to reject a constant migration (power) and how often all three parameters are estimated correctly.

Model	Test Data Set Parameters			Proportion with Maximum Likelihood of Changing Migration	Power to Reject Constant M	Proportion Correct All Parameters
	M_1	M_2	τ			
(a)	5000	10000	0.000005	0.998	0.279	0.082
			0.00001	1.00	0.822	0.114
			0.00005	1.00	1.00	0.513
			0.0001	1.00	0.999	0.423
(b)	10000	5000	0.000005	0.998	0.891	0.124
			0.00001	1.00	1.00	0.186
			0.00005	1.00	0.954	0.507
			0.0001	0.994	0.23	0.143

Table 2.8 Grid Search Results for Identifying Changing Migration with Exponential Growth and Whole Exome Sequencing

For 1,000 dataset simulations of 30000 kb (whole exome) of two examples of changing migration with exponential growth for four different times of migration change, we apply the grid search algorithm and recorded the proportion of the datasets that selected a changing migration model as the most likely. We also record how often there is evidence to reject a constant migration (power) and how often all three parameters are estimated correctly.

Model	Test Data Set Parameters			Proportion with Maximum Likelihood of Changing Migration	Power to Reject Constant M	Proportion Correct All Parameters
	M_1	M_2	τ			
(c)	5000	10000	0.000005	1.00	0.985	0.452
			0.00001	1.00	1.00	0.727
			0.00005	1.00	1.00	0.942
			0.0001	1.00	1.00	0.948
(d)	10000	5000	0.000005	1.00	1.00	0.319
			0.00001	1.00	1.00	0.6
			0.00005	0.999	0.999	0.945
			0.0001	0.999	0.711	0.667

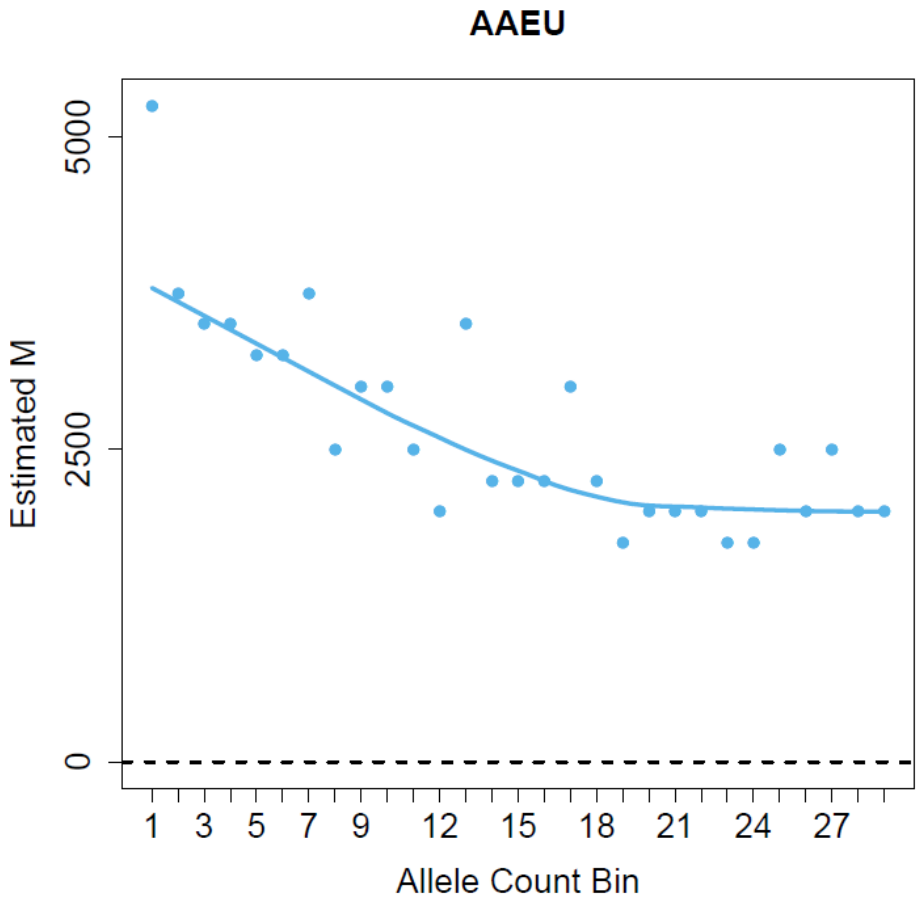

Figure 2.13 Estimated Migration for African-Americans and Europeans.
Estimates of the migration rates using probabilities based on populations of equal sample sizes 339 and 339, undergoing exponential growth, with migration in a single direction.

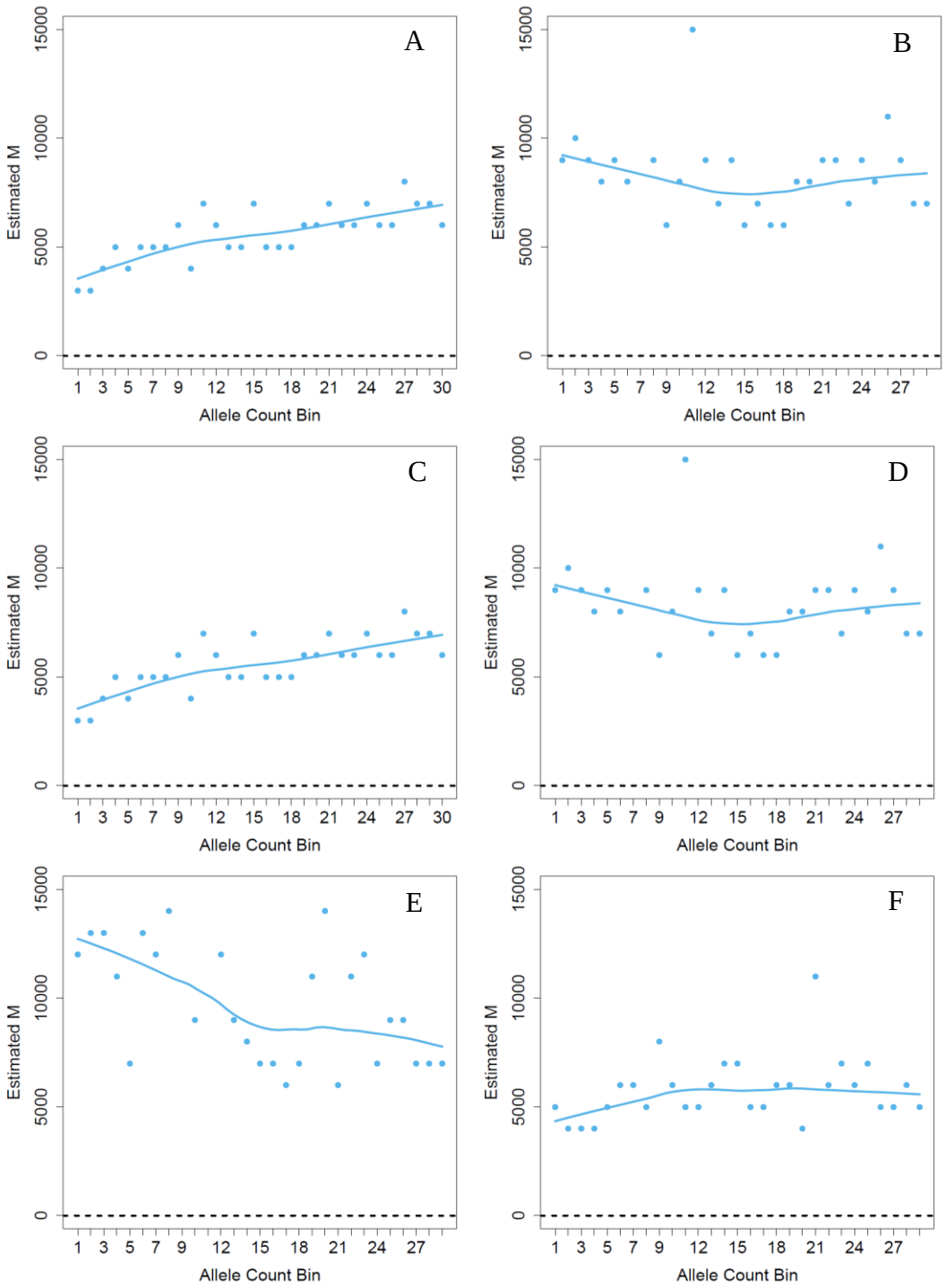

Figure 2.14 Estimated Migration for European Populations
(A) Estimated migration between Northwestern (Britain and Ireland) and Northern Europe (Nordic-Baltic). (B) Estimated migration between Northwestern (Britain and Ireland) and Western Europe. (C) Estimated migration between Northern (Nordic-Baltic) and Western Europe). (D) Estimated migration between Central and Northwestern (Britain and Ireland) Europe. (E) Estimated migration between Central and Western Europe. (F) Estimated migration between Central and Northern (Nordic-Baltic) Europe.

CHAPTER 3: Mathematical Modeling of Population Bottlenecks and Genetic Drift in Next Generation Sequencing Data

3.1 Introduction

Population bottlenecks are defined as drastic reductions in population size. There are many possible causes for bottlenecks, including natural disasters [30-32], captive breeding and re-introduction[36-38], host-pathogen relationships[41-43], and founder populations[19; 44; 45]. After the bottleneck-inducing event, the population begins to regenerate, growing towards its original size. During this period of regrowth, there is a random sampling process for reproduction known as genetic drift. The random nature of genetic drift, compounded with the small size and potential reduction in genetic diversity of the post-bottleneck population, causes dramatic shifts in the regrown population allele frequencies compared to those of the original population[29]. Therefore, accurately modeling the bottleneck and genetic drift has many important applications, such as studying endangered species[36; 39; 40], animal breeding[33-35], and disease patterns[46-48]. In this chapter, we present a mathematical model for population bottlenecks followed by genetic drift and two applications: mtDNA transmission in humans and fibroblast cell growth.

There are several existing basic models describing genealogies that relate to present-day observable sequences. Two models in particular form the foundation of stochastic approaches to modeling reproduction in population genetics. The first of these is the Wright-Fisher model, named after the independent ideas of Sewell Wright[52] and R.A. Fisher[110]. In this model, the ancestors of the genes in the present day generation are drawn randomly with replacement from

the genes in the previous parental generation[29; 111]. Each previous generation is entirely replaced by its offspring[111]. This model is valued for its simplicity, with the assumption of discrete and non-overlapping generations allowing for simple binomial sampling[112]. An alternative to this model which relaxes this assumption was proposed by Patrick Moran[51]. In contrast to the Wright-Fisher model, the Moran model defines steps, or birth-death events, at which two individuals are chosen with replacement: one individual is to give birth (replicate itself) to a new individual and one individual to die[29; 111]. Thus, at each step, the minor allele increases its frequency by one, the major allele increases its frequency by one, or the relative frequencies remain the same[111]. This approach is popular because many calculations that can only be approximated under the Wright-Fisher model are mathematically tractable[29; 111]. However, in the basic formulation of each of these models, the populations represent an idealized scenario with constant population size, no sexual reproduction, random mating, no mutation, and no selection[29; 111].

For this model of population bottlenecks and subsequent genetic drift, we build on these existing models. Maintaining the concept of overlapping generations, we develop a "modified Moran model", now allowing for a growing population size. We show the entire process can be written as a discrete Markov chain with transition matrices corresponding to the bottleneck and subsequent growth. Constructing a closed-form equation, we fully model the probability of observing the shift in allele frequency in populations before and after the bottleneck. Additionally, we develop a framework for incorporating and testing selection in this model. We modify the stochastic process to include changes in the probability of reproduction due to a selection coefficient. Using a grid search, we estimate the most likely selection coefficient given the shift in allele frequencies before and after the population bottleneck.

In my first application of this model, we focus on mtDNA transmission from mother to child. This work is previously published in Genome Research, April 2016[113]. The allele frequencies in mtDNA allele frequencies from mother to child can shift dramatically, indicating the presence of a severe bottleneck during this process. We aim to understand the number of genetic units in this bottleneck and its characteristics, such as variability in size. To this end, we analyze short read sequences of the mitochondrial DNA of 189 mother-child pairs from the Genome of the Netherlands and Biobanking and Biomolecular Research Infrastructure of the Netherlands. Using a maximum likelihood equation and model comparisons, we determine the best fitting model is a variable size bottleneck with a mean of nine individual genetic units transmitted.

In the second application, we apply the mathematical model to cell growth in a laboratory setting. We analyze 58 variants from a set of 1489 variants from next generation sequencing of cell populations isolated from subjects at the National Institutes of Health. During the experimental process, these fibroblast cell populations underwent an extreme bottleneck of known size. The allele frequencies of the population before the bottleneck and after regrowth from the bottleneck differ drastically. While this shift could be driven by genetic drift alone, in some cases, selection is acting on the variants. Discerning between drift and selection is essential to understanding functional consequences of the variants and providing insight into the etiology of the pre-mature aging disorders studied here. Applying the probabilistic approach with a known bottleneck size, we find evidence of positive selection in three of these variants and estimate corresponding selection coefficients. We therefore present a second application of this flexible, probability-based approach to directly modeling the biological process of population bottlenecks and growth and identifying variants with a selection advantage.

3.2 Mathematically Modeling Population Bottlenecks and Genetic Drift

We first describe the basic model: a single bottleneck of constant size n_b with subsequent genetic drift (population growth) and no selection. The observed data set includes the following statistics for each variant: in the initial population (also called "early stage"), we observe a sample of size n_I with observed minor allele count, k_I^{obs} ($0 \leq k_I^{obs} \leq n_I$); in the final population (also called "late stage"), we observe a sample of size, n_F, with observed minor allele count, k_F^{obs}($0 \leq k_F^{obs} \leq n_F$). We aim to directly formulate the probability, $P\left(k_F^{obs} \middle| k_I^{obs}, n_I, n_F, n_b\right)$, the probability of observing the final minor allele count, given the bottleneck size and initial population statistics.

There are four primary components to this model: genotyping, sampling, population bottleneck, and genetic drift. In Figure 3.1, we show a schematic diagram of how these components fit together. The observed initial minor allele count in the sample is obtained from the true minor allele count in the sample by genotyping, with some potential for error. This true minor allele count arises through sampling from the initial population, creating sampling error. The initial population undergoes the bottleneck, followed by subsequent genetic drift or population growth to obtain the final population. Like the initial population, this final population is also sampled and then genotyped, with uncertainty incurred at each step. Finally, this results in the final observed minor allele count. We will discuss each of these components separately, beginning with the primary population genetic processes, the bottleneck and genetic drift, and returning to the full probabilistic model at the end of this section. We include Table 3.1 as a reference for the symbols used to construct this model.

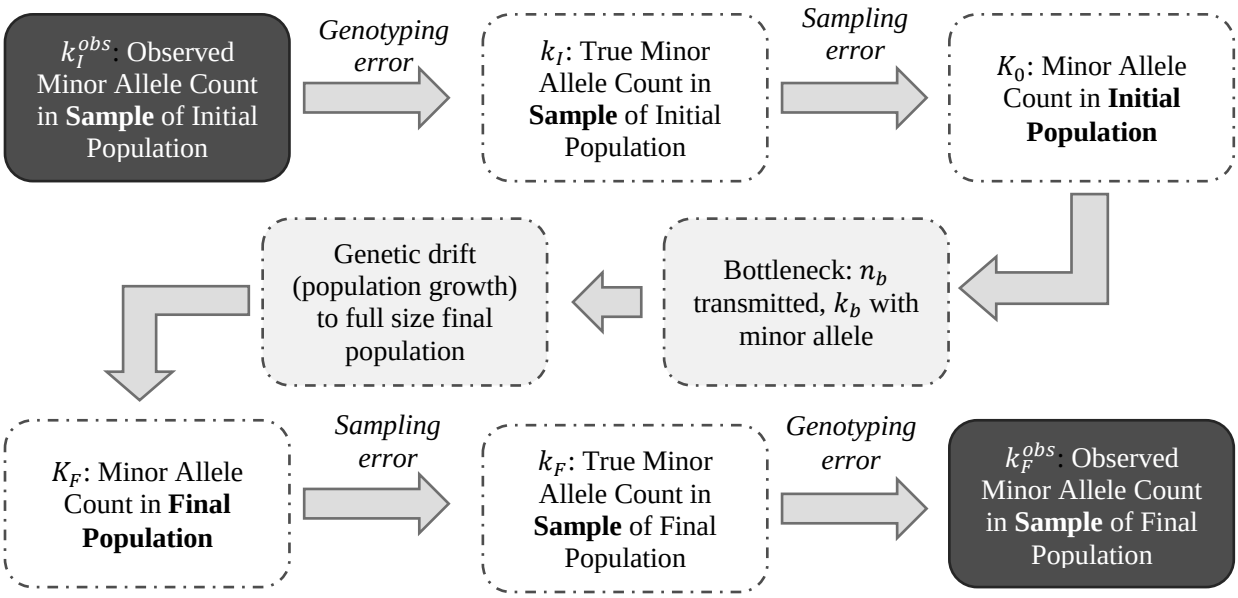

Figure 3.1 Schematic Diagram of Each Component of the Mathematical Model
We model genotyping, sampling, population bottleneck, and genetic drift. The black boxes indicate the observed data: initial population observed minor allele count and final population observed minor allele count. The grey boxes indicate the primary population genetics components of the model: the bottleneck and subsequent genetic drift.

Table 3.1 Symbols for Probabilistic Model

Symbol	Known or Unknown?	Meaning
k_I^{obs}	Known	Observed minor allele count in the sample of the initial population
n_I	Known	Initial sample size
k_I	Unknown	True minor allele count in the sample of the initial population
k_F^{obs}	Known	Observed minor allele count in the sample of the final population
n_F	Known	Final sample size
k_F	Unknown	True minor allele count in the sample of the final population
k_b	Unknown	Minor allele count in the post-bottleneck population
n_b	Known	Post-bottleneck population size
$K_F\ (= k_b + z)$	Unknown	True minor allele count in the full final population
$N_F\ (= n_b + j)$	Known	Final Population Size
j	Known	Number of steps during genetic drift
z	Unknown	Additional minor alleles gained during genetic drift
ε	Known	Position-specific error rate
$p_F\ (= K_F/N_F)$	Unknown	Minor allele frequency in final population
p_0	Unknown	Minor allele frequency in initial population

Population Bottleneck

The bottleneck step is a random sample from the full size initial population. Let N_0 be the number of individuals in the full-size initial population and K_0 ($0 \leq K_0 \leq N_0$) of these individuals carry a minor allele. From this pool of individuals, we assume an unordered draw without replacement for n_b individuals, k_b ($0 \leq k_b \leq n_b$) of which carry the minor allele. In this case, n_b=200 randomly chosen cells from the early stage cell population. Therefore, the number of individuals carrying a minor allele in the post-bottleneck population follows a hypergeometric distribution, $k_b \sim Hypergeometric(n_b, K_0, N_0)$. Where N_0 is much larger than n_b, this hypergeometric distribution converges in distribution to the simpler binomial distribution ($k_b \sim Binomial(n_b, p_0 = \frac{K_0}{N_0})$)[94] as in (3.1). As in this application $n_b << N_0$, we use this binomial distribution to model this bottleneck process, creating a transition matrix to all possible values of k_b given n_b, K_0, N_0, with transition probabilities $P(k_b|n_b, p_0 = K_0/N_0)$.

$$P(k_b|p_0, n_b) = \binom{n_b}{k_b} p_0^{k_b}(1-p_0)^{n_b-k_b} \tag{3.1}$$

Genetic Drift

To model the subsequent genetic drift during the replication or population growth process, we build on the basic Moran model while incorporating a growing population size. Under the original Moran model, at each step, one individual dies and is replaced by the copy of another individual in the population[51], maintaining a constant population size. In our modified Moran model, at each step, one individual is chosen randomly from the current population to be replicated and added to the current generation. This means there are no deaths, causing the population to grow by one individual in each step. Therefore, given a previous population size of n_b with k_b individuals carrying the minor allele, the current population of $n_b + 1$ individuals can

have k_1 individuals carrying the minor allele, where $k_1 = k_b$ or $k_1 = k_b + 1$. This gives a Bernoulli distribution for the number of minor alleles in the population after one step (k_1):

$$P(k_1|n_b, k_b) = \begin{Bmatrix} \frac{k_b}{n_b} & k_1 = k_b + 1 \\ \frac{n_b - k_b}{n_b} & k_1 = k_b \\ 0 & otherwise \end{Bmatrix} \tag{3.2}$$

Similarly, in the next step, the proportions of minor and major alleles in the population are updated and the sampling repeats. Therefore, the number of minor alleles, k_2, in the population after two steps (population size now $n_b + 2$) is:

$$P(k_2|n_b, k_b) = \begin{Bmatrix} \frac{n_b - k_b}{n_b}\left(\frac{n_b - k_b + 1}{n_b + 1}\right) & if\ k_2 = k_b \\ 2\frac{k_b(n_b - k_b)}{n_b(n_b + 1)} & if\ k_2 = k_b + 1 \\ \frac{k_b}{n_b}\left(\frac{k_b + 1}{n_b + 1}\right) & if\ k_2 = k_b + 2 \\ 0 & otherwise \end{Bmatrix} \tag{3.3}$$

Repeating this binomial sampling for each step, at generation *j* the probability of observing *z* additional individuals carrying minor alleles, for a total of $k_b + z$ minor alleles is:

$$P\big(k_j = k_b + z|k_b\big) \tag{3.4}$$
$$= \binom{j}{z}\left(\frac{(k_b + z - 1)!}{(k_b - 1)!}\right)\left(\frac{\big((n_b - k_b) + (j - z - 1)\big)!}{(n_b - k_b - 1)!}\right)\left(\frac{(n_b - 1)!}{\big(n_b + (j - 1)\big)!}\right)$$

This closed-form equation (3.4) provides the probabilities for the transition matrix for moving from the initial k_b minor alleles in the post-bottleneck population of size n_b, to the final number of alleles after genetic drift, k_j, in the final population of size $n_b + j$. The binomial sampling for the bottleneck in (3.1) and the genetic drift in equation (3.4) correspond to two transition matrices of a discrete Markov Chain.

Modeling Sampling Error

The next component of the full probability reflects the sampling process used to obtain the observed data set. Consider the true final minor allele count, K_F in the full size final population of size N_F and a sample of size n_F, the probability of observing k_F ($0 \leq k_F \leq n_F$) minor alleles in this sample follows a hypergeometric distribution, $k_F \sim Hypergeometric(n_F, K_0, N_0)$. As in the bottleneck process, $n_F << N_F$, we use a binomial distribution to model this sampling, ($k_F \sim Binomial(n_F, p_F = \frac{K_F}{N_F})$). Therefore, we calculate:

$$P(k_F|p_F, n_F) = \binom{n_F}{k_F} p_F^{k_F} p_F^{n_F - k_F} \tag{3.5}$$

Additionally, we need to incorporate the sampling process for the initial population sample. In this case, the probability calculation is in the opposite direction and we aim to estimate $P(p_0|k_I, n_I)$, the probability of the minor allele frequency in the initial full size population, p_0, given the minor allele count k_I and sample size n_I. We apply Bayes' Rule and the Total Probability Theorem, followed by the General Product Rule. The final reduction occurs because n_I (the initial population sample size) is a known parameter and independent of p_0 ($P(n_I|p_0) = P(n_I) = 1$). Now $P(k_I|p_0, n_I)$ is simply the sampling error described above in (3.5), the binomial probability of k_I minor alleles in a sample of size n_I, drawn from the underlying frequency of p_0. We assume a uniform prior on p_0.

$$P(p_0|k_I, n_I) = \frac{P(k_I, n_I|p_0)P(p_0)}{\int_0^1 [P(k_I, n_I|p_0)P(p_0)]dp_0} = \frac{[P(k_I|p_0, n_I)P(n_I|p_0)]P(p_0)}{\int_0^1 [P(k_I|p_0, n_I)P(n_I|p_0)P(p_0)]dp_0}$$
$$= \frac{P(k_I|p_0, n_I)P(p_0)}{\int_0^1 [P(k_I|p_0, n_I)P(p_0)]dp_0} \tag{3.6}$$

Modeling Genotyping Error

To model the genotyping errors, we incorporate a position-specific error rate ε of 0.001 based on the base quality scores during sequencing. Consider the final sample of size n_F with the true number of minor alleles k_F ($0 \leq k_F \leq n_F$), the probability that the number of observed minor alleles is k_F^{obs} ($0 \leq k_F^{obs} \leq n_F$) is determined by rate of errors in genotyping, ε. In (3.7), this probability is made up of two binomials that model: the probability that i minor alleles are correctly called minor (with probability $(1-\varepsilon)$); and the probability that the remaining $k_F^{obs} - i$ alleles are incorrectly called minor (with probability ε).

$$P\left(k_F^{obs} | n_F, k_F\right) = \sum_{i=0}^{k_F^{obs}} \binom{k_F^{obs}}{i} (1-\varepsilon)^i (\varepsilon)^{k_F - i} \binom{n_F - k_F}{k_F^{obs} - i} \varepsilon^{k_F^{obs} - i} (1-\varepsilon)^{(n_F - k_F) - (k_F^{obs} - i)} \tag{3.7}$$

Because there is potential for genotyping error on both sides of the model (for the observed initial sample and the observed final sample), we also need to calculate the probability of the true initial minor allele count in the sample, k_I, given the observed value k_I^{obs}. To this end, we apply Bayes' Rule and conditional probability to calculate $P(k_I | k_I^{obs}, n_I)$ as shown in (3.8). We assume a uniform prior on k_I so $P(k_I|n_I)/P\left(k_I^{obs}\middle|n_I\right)$ is constant in k_I. Therefore, this reduces to the genotyping error component as shown in (3.7), now for the initial population sample: the probability of observing k_I^{obs} minor alleles in the initial sample of size n_I given the true minor allele count of k_I.

$$P(k_I | k_I^{obs}, n_I) = \frac{P\left(k_I^{obs}, k_I \middle| n_I\right)}{P\left(k_I^{obs}\middle|n_I\right)} = \frac{P\left(k_I^{obs}\middle|n_I, k_I\right)P(k_I|n_I)}{P\left(k_I^{obs}\middle|n_I\right)} \propto P\left(k_I^{obs}\middle|n_I, k_I\right) \tag{3.8}$$

Combining Components for the Full Model

With each of these components defined, we now return to the full probabilistic model, $P\left(k_F^{obs} \middle| k_I^{obs}, n_I, n_F, n_b\right)$. For reference, Table 3.1 provides a list of the symbols and their meanings used throughout this derivation. First, we write $P\left(k_F^{obs} \middle| k_I^{obs}, n_I, n_F, n_b\right)$ by conditioning on k_b, the number of minor alleles in the bottleneck of size n_b:

$$\begin{aligned} P\left(k_F^{obs} \middle| k_I^{obs}, n_I, n_F, n_b\right) &= \sum_{k_0=0}^{n_0} P\left(k_F^{obs} \middle| k_b, k_I^{obs}, n_I, n_F, n_b\right) P\left(k_b \middle| k_I^{obs}, n_I, n_F, n_b\right) \\ &= \sum_{k_b=0}^{n_b} \underbrace{P\left(k_F^{obs} \middle| k_b, n_F, n_b\right)}_{Part\ B} \underbrace{P\left(k_b \middle| k_I^{obs}, n_I, n_b\right)}_{Part\ A} \end{aligned} \tag{3.9}$$

In (3.9), this probability further simplifies because k_F^{obs} given k_b, n_F, n_b is independent of k_I^{obs}, n_I and k_b given $k_I^{obs}, n_I, n_b$is independent of n_F. Therefore, this probability consists of two expressions: (A) the probability of transmitting k_b minor alleles in a bottleneck of size n_b given the initial observed minor allele frequency k_I^{obs} and initial sample size n_I; and (B) the probability of observing k_F^{obs} minor alleles after genetic drift, conditional on k_b, n_F, n_b. We use conditional probabilities to break these parts into the interpretable components defined above.

To calculate Part A, we condition on p_0, the minor allele frequency in the initial population. Because p_0 is independent of n_b given $k_I^{obs}, n_I$and because k_b given p_0, n_b is independent of k_I^{obs}, n_I and p_0, this simplifies to two additional parts in (3.10).

$$\begin{aligned} P\left(k_b \middle| k_I^{obs}, n_I, n_0\right) &= \int_0^1 P\left(k_b | p_0, k_I^{obs}, n_I, n_0\right) P\left(p_0 | k_I^{obs}, n_I, n_0\right) dp_0 \\ &= \int_0^1 \underbrace{P(k_b | p_0, n_0)}_{A_1} \underbrace{P\left(p_0 | k_I^{obs}, n_I\right)}_{A_2} dp_0 \end{aligned} \tag{3.10}$$

The first part of this expression, A_1, is simply the bottleneck step as in (3.1). The second expression of this equation, A_2, requires further work. We first condition on k_I, the true minor allele count and A_2 simplifies to (3.11) since p_0 given n_I, k_I is independent of k_I^{obs}. This expression now includes the genotyping and sampling processes for the initial population sample as outlined in previous sections.

$$P\left(p_0|k_I^{obs}, n_I\right) = \sum_{k_I=0}^{n_I} P(p_0|\, k_I^{obs}, n_I, k_I)P(k_I|k_I^{obs}, n_I)$$
$$= \sum_{k_I=0}^{n_I} P(p_0|\, n_I, k_I)P(k_I|k_I^{obs}, n_I) \tag{3.11}$$

In (3.12), we incorporate the sampling error as calculated in (3.6), and the genotyping error as calculated in (3.8).

$$\sum_{k_I=0}^{n_I} P(p_0|\, n_I, k_I)P(k_I|k_I^{obs}, n_I)$$
$$\propto \sum_{k_I=0}^{n_I} \left\{ \left(\frac{P(k_I|p_0, n_I)P(p_0)}{\int_0^1 P(k_I|p_0, n_I)P(p_0)dp_0} \right) P\left(k_I^{obs}\middle|k_I, n_I\right) \right\} \tag{3.12}$$

We now focus on Part B of (3.9), the probability of observing k_F^{obs} minor alleles after genetic drift, conditional on k_b, n_b, and n_F. This part models the three processes that occur after the bottleneck: (1) genetic drift (growth) to reach the final minor allele count K_F and final total allele count N_F from the bottleneck size of n_I, (2) sampling from this final population, (3) genotyping error in our sample. To reach these interpretable parts, we start by conditioning on k_F, the true minor allele count in the sample, and then K_F, the final population minor allele count.

$$
\begin{aligned}
P\left(k_F^{obs}\middle|k_b,n_b,n_F\right) &= \sum_{k_F=0}^{n_F} P\left(k_F^{obs}|k_F,n_b,k_b,n_F\right)P(k_F|k_b,n_b,n_F) \\
&= \sum_{k_F=0}^{n_F} P\left(k_F^{obs}|k_F,n_F\right)P(k_F|k_b,n_b,n_F) \\
&= \sum_{k_F=0}^{n_F}\left[P\left(k_F^{obs}|k_F,n_F\right)\sum_{K_F=0}^{N_F=0} P(k_F|k_b,n_b,n_F,K_F,N_F)\,P(K_F|N_F,k_b,n_b,n_F)\right] \\
&= \sum_{k_F=0}^{n_F}\left[\underbrace{P\left(k_F^{obs}|k_F,n_F\right)}_{B_1}\sum_{K_F=0}^{N_F=0}\underbrace{P(k_F|n_F,K_F,N_F)}_{B_2}\underbrace{P(K_F|N_F,k_b,n_b)}_{B_3}\right]
\end{aligned}
\tag{3.13}
$$

The first expression in (3.13), B_1, arises because k_F^{obs} is independent of n_b, k_b given k_F, n_F. Then B_1 is the genotyping error probability, as calculated in (3.7). The second term, B_2, arises by conditioning on K_F, N_F, such that $P(k_F|n_F, K_F, N_F)$ is independent of k_b, n_b (we assume N_F, the final population size, is known). Then B_2 is the sampling error component as in (3.5): a binomial that corresponds to the observing k_F minor alleles after sampling n_F from the full final population where minor alleles are sampled with probability $p_F = K_F/N_F$. The last portion, B_3, models the growth of the population (genetic drift) to the full size final population from the bottleneck size. In B_3, K_F is independent of n_F given k_b, n_b, N_F. As in (3.4), B_3 is calculated using a modified Moran model without replacement, with $K_F = k_b + j$ and $N_F = n_b + j$.

Combining these components gives the overall summation in (3.9). This is the basic model for the bottleneck and subsequent growth.

3.3 Application to mtDNA Transmission

Mitochondria, regarded as the "energy powerhouses" of the cell, are vital to the health of an individual. Previous studies implicate mutations in mitochondrial DNA (mtDNA) as the cause of major health problems, including colorectal cancer susceptibility, tissue aging, and post

lingual deafness [114-117]. Most mtDNA mutations that cause diseases due to defects in mitochondrial function exist as heteroplasmies (intra-individual variation) and only cause disease symptoms when the frequency of the mutant allele exceeds a particular threshold [118]. Below this threshold, individuals are asymptomatic, presumably because there are enough functional mitochondria for normal metabolism. Changes in the frequency of pathogenic mutations during the transmission of heteroplasmies from mothers to offspring can thus play an important role in the disease risk of the offspring [119-122]. However, most of our knowledge concerning the dynamics of heteroplasmy transmission comes from studies of pathogenic mutations [118; 123-125] or from mouse models [126-129]. There are also a few studies in oocytes [130; 131] and placenta [132]. To date there have been limited studies of normal patterns of heteroplasmy transmission in humans [133-137], and several questions remain.

The process of mitochondrial DNA transmission, for example, remains unclear. Previous estimates of the effective number of transmitted mtDNA genomes range widely, from 1 mtDNA genome to 200 mtDNA genomes [134; 135; 138]. These previous studies assumed a constant size for the bottleneck across individuals, ignoring the potential effects of allowing the bottleneck size to vary among individuals. Furthermore, previous biological studies, both microscopic and biochemical[139], suggest that mtDNA genomes may not behave as individual, independent entities, but rather, behave as discrete homoplasmic units called “nucleoids”, each of which contains 5-10 identical mtDNA genomes [140-142]. Under this nucleoid model, mtDNA heteroplasmy at the cellular level would reflect multiple nucleoids that are homoplasmic for different sequence variants. Some studies found nucleoid-based models provide a better fit than simple bottleneck models in the segregation of heteroplasmic mtDNA genomes in cell lines [140;

[142]. Other studies, however, find the opposite[126]. To date nucleoid-based models have not been investigated in the transmission of mtDNA heteroplasmy from mothers to offspring.

In this study, we analyzed short read sequences of the mitochondrial DNA of mothers and children from 189 trios from the Genome of the Netherlands and Biobanking and Biomolecular Research Infrastructure of the Netherlands. The allele frequencies in the mitochondrial DNA between generations differed considerably. Using the probabilistic method described, we estimate the size and nature of the bottleneck based on a maximum likelihood equation and model comparisons.

3.3.1 mtDNA Transmission Data

The characteristics of the study population from the Genome of the Netherlands and the production of the sequence data have been described in detail in previous manuscripts[143; 144]. Briefly, genomic DNA was purified from blood samples from 769 individuals from across The Netherlands and sequenced to an average genomic coverage of ~14X on the Illumina HiSeq2000 platform. This consisted of the 231 trios used for this analysis as well as, 11 monozygotic (MZ) twin quartets, and 8 dizygotic (DZ) twin quartets. Of the 231 trios, 189 mothers exhibited heteroplasmy and 112 of these heteroplasmies were transmitted to their children. For this analysis, we use one offspring and one heteroplasmic position per family to avoid any complications due to potential non-independence of heteroplasmies within families, resulting in a total of 125 independent sites.

3.3.2 mtDNA Transmission Methods

We aim to estimate the size and nature of the bottleneck during the inheritance of mitochondria based on the change in minor allele frequency of these sites transmitted from mother to offspring. we considered four models: a constant size bottleneck model, in which each mtDNA genome is a segregating unit and the bottleneck size does not vary between individuals; a variable size bottleneck

model, in which each mtDNA genome is a segregating unit and the bottleneck size is allowed to vary between individuals; a constant size nucleoid model, in which a nucleoid containing a variable number of identical mtDNA genomes (with mean = 7.5 genomes per nucleoid) is the segregating unit and the bottleneck size does not vary between individuals; and a variable size nucleoid model, in which a nucleoid containing a variable number of identical mtDNA genomes (with mean = 7.5 genomes per nucleoid) is the segregating unit and the bottleneck size is allowed to vary between individuals.

We first describe the most basic model: a constant size bottleneck with the transmission of individual mitochondria. Using the notation of the mathematical model described above, let n_0 be the size of the bottleneck, k_I^{obs} the number of copies of the minor allele in the mother, n_I the total number of reads in the mother, k_F^{obs} the number of copies of the minor allele in the offspring, and n_F the total number of reads in the offspring. We aim to maximize $L(n_0|k_F^{obs}, n_F, k_I^{obs}, n_I)$. To this end, we model the bottleneck as sampling n_0 mtDNA genomes with k_0 copies of the minor allele where each transmitted mtDNA genome is sampled independently from many maternal mtDNA genomes. Therefore, we calculate the probability of observing k_F^{obs}given k_I^{obs} when n_0 mtDNA genomes are transmitted, as described in Section 3.2:

$$\begin{aligned} &L\left(n_0\middle|k_F^{obs}, n_F, k_I^{obs}, n_I\right) \\ &= P\left(k_F^{obs}\middle|k_I^{obs}, n_I, n_F, n_0\right) \sum_{k_F=0}^{n_F} P\left(k_F^{obs}|k_F, n_0, k_0, n_F\right) P(k_F|k_0, n_0, n_F) \end{aligned} \tag{3.14}$$

Building on the most basic model of a constant size bottleneck, we construct three more complex models. The variable size bottleneck model differs from the constant size bottleneck model by modeling n_0, the number of mtDNA genomes transmitted to the child, as a Poisson distributed random variable with mean λ. The estimate of λ can be obtained by maximizing the likelihood of λ while summing over the unknown values of k_0 and n_0 as in (3.15). With a goal of simply obtaining

an approximate value, we use integer values of λ. This allows for a grid search to obtain the maximum value of λ.

$$L\left(\lambda \middle| k_F^{obs}, n_F, \widehat{k}_I, n_I\right) = P\left(k_F^{obs} \middle| \lambda, k_I^{obs}, n_I, n_F, n_0\right) \tag{3.15}$$
$$= \sum P\left(k_F^{obs} \middle| n_0, k_I^{obs}, n_I, n_F, \lambda\right) P(n_0 |\, \lambda, k_I^{obs}, n_I, n_F)$$
$$= \sum P\left(k_F^{obs} \middle| n_0, k_I^{obs}, n_I, n_F\right) P(n_0 |\, \lambda)$$

The final equality in (3.15) arises because k_F^{obs} given n_0, k_I^{obs}, n_I, n_F is independent of λ and n_0 given λ is independent of k_I^{obs}, n_I, n_F. Because the upper limit of n_0 is infinite for a Poisson distribution, we calculate this sum until $P(n_0|\lambda)$ reaches a lower limit (set at 10^{-10}).

The third model, the constant size bottleneck with nucleoids, differs from the first two models in that the estimate of n_0 now represents the number of nucleoids transmitted to the child, with each nucleoid containing only identical copies of either the major allele or the minor allele. We assume each nucleoid i has a random size $g_i, i = 0 \ldots n_0$ modeled as a Poisson-distributed random variable with mean $\lambda = 7.5$ (based on empirical studies that find that each nucleoid has 5-10 mtDNA genomes[140-142]). Without loss of generality, the first k_0 groups contain the minor allele. This gives $\sum_{i=1}^{n_0} g_i$ as the total number of transmitted mitochondria and $\sum_{i=1}^{k_0} g_i$ as the total number of copies of the minor allele. Under this nucleoid model, we adjust B_3 in equation (3.13), which models the replication process to the full-size offspring population from the bottleneck size at transmission. Using the same model of replication, we now assume that in the initial population, there are $\sum_{i=1}^{n_0} g_i$ mtDNA genomes with $\sum_{i=1}^{k_0} g_i$ carrying minor alleles. Because we lack a closed form equation for all possibilities of the Poisson-distributed random sizes of g_i, we use a Monte-Carlo approximation to calculate this term. The other terms of equation (3.13), B_1 and B_2, are again made

up of the genotyping error probability and the sampling error probability. The remainder of the maximum-likelihood estimation was calculated as for the constant size bottleneck model.

Finally, we consider the variable size bottleneck with nucleoids. Like the variable size bottleneck model in which each mtDNA genome is a segregating unit, this model differs from the constant size bottleneck model with nucleoids in that we now estimate λ, the mean of a Poisson distributed random variable that represents the mean number of nucleoids transmitted to the child. The estimate of λ can be obtained by maximizing the likelihood of λ while summing over the unknown values of $k_0$and n_0, as in (3.15) with λ now representing the mean number of nucleoids transmitted to the child. Again, with a goal of simply obtaining an approximate value, we use integer values of λ. This allows for a grid search to obtain the maximum value of λ.

3.3.3 mtDNA Transmission Results

We used the distribution of the changes in minor allele frequency (MAF) between mother and offspring pairs at heteroplasmic sites to investigate the size and nature of the transmission bottleneck. We fit four models to the data: a constant size bottleneck, a variable size bottleneck, a constant size bottleneck with nucleoids, and a variable size bottleneck with nucleoids.

Under the constant size bottleneck model, the maximum-likelihood estimate (MLE) of the number of transmitting mtDNA genomes was 8 (Figure 3.2A) while under the variable size bottleneck model the MLE was a mean of 9 transmitted mtDNA genomes (Figure 3.2B). For the constant size bottleneck with nucleoids the MLE was 7 nucleoids, assuming an average size of 7.5 mtDNA genomes (based on empirical data [139; 140; 142]) (Figure 3.2C). The variable size bottleneck with nucleoids, also assuming an average size of 7.5 mtDNA genomes, the MLE was a mean of 9 transmitted nucleoids (Figure 3.2D). The Akaike Information Criterion (AIC) was 1123.92 for the constant size bottleneck model, 1119.16 for the variable size bottleneck

model,1190.38 for the constant size bottleneck with nucleoids, and 1167.00 for the variable size bottleneck with nucleoids (Table 3.2). This indicates that the variable size bottleneck with nucleoids model provides the best fit to the data (the smaller the AIC value, the smaller the loss in information when fitting the model).

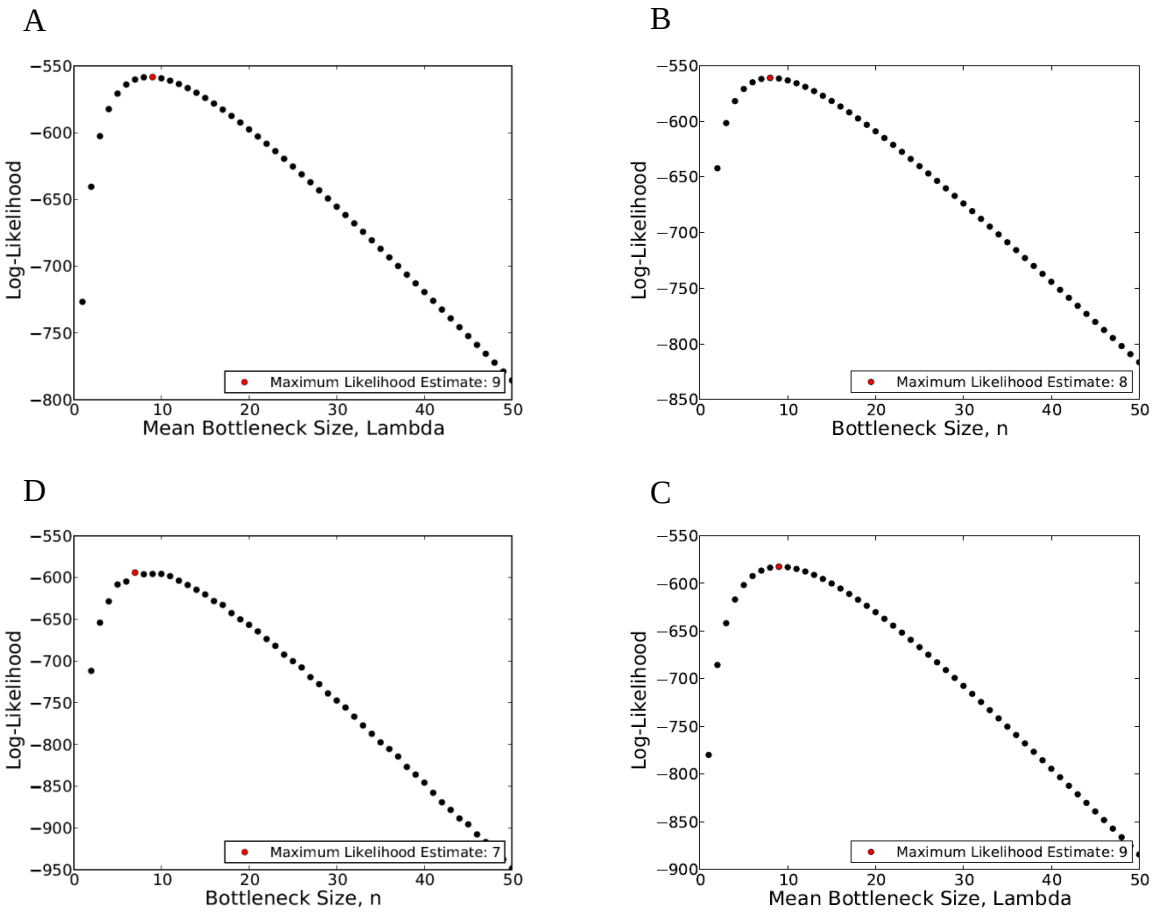

Figure 3.2 Likelihoods for Bottleneck Size Under Four Different Bottleneck Models
The likelihoods combined across the 125 independent sites, with the maximum likelihood estimate indicated with a red dot. The four panels correspond to the four models: (A) a constant size bottleneck model, in which each mtDNA genome is a segregating unit and the bottleneck size does not vary between individuals, (B) a variable size bottleneck model, in which each mtDNA genome is a segregating unit and the bottleneck size is allowed to vary between individuals, (C) a nucleoid model, in which a nucleoid containing a variable number of homoplasmic mtDNA genomes is the segregating unit (mean size 7.5) and the bottleneck size does not vary between individuals, (D)a variable size nucleoid model, in which a nucleoid containing a variable number of homoplasmic mtDNA genomes (mean size 7.5) is the segregating unit and the bottleneck size is allowed to vary between individuals.

Table 3.2 Maximum Likelihood Estimates (MLE) and Akaike's Criterion Information (AIC) for Each Model

Model	MLE	AIC
Constant Size Bottleneck	8	1123.92
Variable Size Bottleneck	9	1119.16
Constant Size Bottleneck with Nucleoids	7	1190.38
Variable Size Bottleneck with Nucleoids	9	1167.00

3.4 Application to Fibroblast Cell Growth

Natural selection is an evolutionary process in which the genetics of some individuals provide an advantage (or disadvantage) in survival and reproduction. Selection on a variant alters the assumption of randomness in reproduction, therefore changing the allele frequencies expected in a population over time[29]. As observed previously, drastic reductions in population size, followed by genetic drift, can also change population allele frequencies[29; 112]. Discerning between drift and selection is essential to understanding functional consequences of the variants. In addition, identifying genes under selective pressure can elucidate the mechanisms of disease, the role of bacterial resistance, and improve the design of gene-driven pharmaceutical interventions.

This application focuses on understanding the somatic variation in aging related genes in fibroblast cell samples. Specifically, this study uses a deep sequencing approach to analyze intra-individual genetic variation at the single nucleotide level in 44 aging related candidate genes to determine if this variation changes during *in vitro* aging. The initial and final populations of primary cells come from unaffected individuals and individuals with premature aging diseases. We aim to identify variants in this dataset that show evidence against the null hypothesis of genetic drift alone and, in those cases, estimate selection coefficients. Applying the mathematical

modeling approach, we identified three variants with evidence of selection including a likely driver mutation in *CDKN2A* in an XPA patient.

3.4.1 Cell Growth Data

This study uses next generation sequencing on primary dermal fibroblast cell cultures from 3 individuals diagnosed with Hutchinson-Gilford progeria syndrome, 3 individuals diagnosed with Xeroderma pigmentosum, and 5 unaffected individuals from the Coriell Cell Repository and the Progeria Research Foundation cell and tissue bank. Target enrichment and sequencing were performed for 44 genes including genes involved in cell cycle regulation, DNA repair, telomere maintenance, and the nuclear lamina on a sample from the initial population (approximately two million cells) from each donor. Using a cell sorter, two hundred cells from this initial population were randomly sampled, creating a post-bottleneck population. This sample was re-plated and allowed to grow through approximately sixteen doublings. Target enrichment and sequencing was repeated for the 44 genes on a sample from the final population (approximately 13.6 million cells) for each donor. In both sequencing procedures, sequence capture was done to enrich for loci totaling 290 kb of sequence with a mean read depth between 600-2800 bp. After sequencing, there were 1489 mutations across individuals identified as somatic based on minor allele frequency. Of these mutations, 58 differed in allele frequency in the initial and final population within a participant's sample (based on collaborator's *lofreq* analysis). The changes in minor allele frequency from initial to final population in these 58 variants are shown in Figure 3.3.

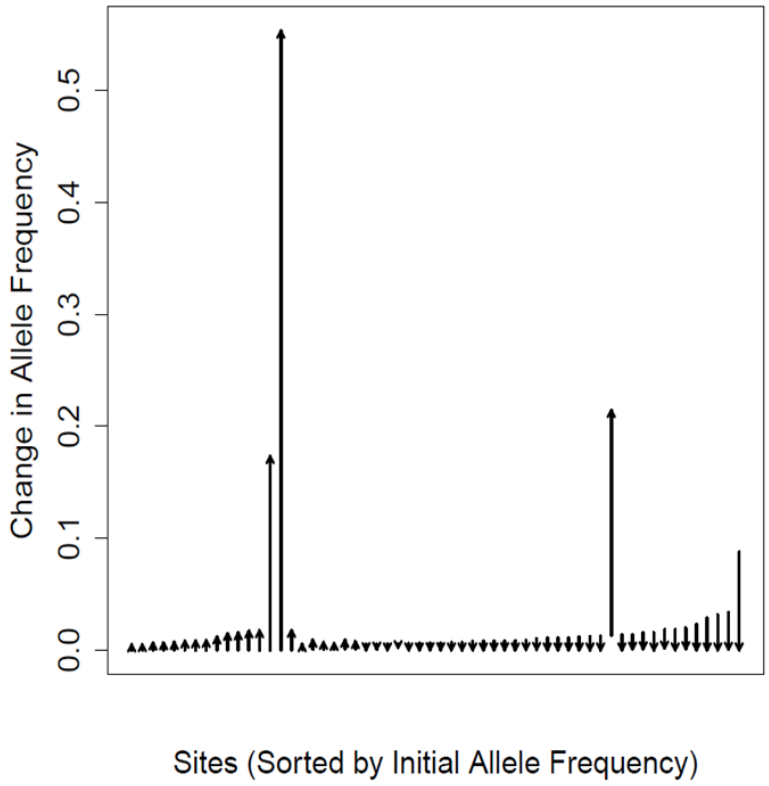

Figure 3.3 Changes in Minor Allele Frequency from Initial to Final Population
Sites are sorted by initial allele frequency, with the direction of change indicated by the arrow.

3.4.2 Cell Growth Methods

We aim to test the null hypothesis of the change in minor allele frequency of a site between the initial to the final population under a model of genetic drift acting alone. To this end, we calculate a p-value: the probability, given the initial minor allele count, of observing a final minor allele count at least as extreme as that observed in the data, under the null model of basic genetic drift. Using the notation of the mathematical model described above, we calculate the probability of the observed final allele count k_F^{obs} given n_b, n_F, k_I^{obs}, and n_I as in (3.9): $P(k_F^{obs}|n_b, n_F, k_I^{obs}, n_I)$. To test the null hypothesis of genetic drift alone, we calculate the individual probabilities of observing each minor allele count at least as extreme as $\hat{k}_F$. For example, if the final minor allele count is greater than the initial minor allele count, indicating

the minor allele count has increased, we calculate each $k_F^* \geq k_F^{obs}$. Summing over these probabilities, we obtain $P\left(k_F^* \geq k_F^{obs} \middle| n_0, n_F, k_I^{obs}, n_I\right) = \sum_{k_F^*=k_F^{obs}}^{k_F^*=n_F} P\left(\ k_F^* \middle| n_0, n_F, k_I^{obs}, n_I\right)$. Similarly, if the observed final minor allele count is less than the initial minor allele count, indicating the minor allele count has decreased, we calculate $P\left(\ k_F^* \leq k_F^{obs} \middle| n_0, n_F, k_I^{obs}, n_I\right) = \sum_{k_F^*=0}^{k_F^*=\hat{k}_F} P\left(\ k_F^* \middle| n_0, n_F, k_I^{obs}, n_I\right)$. This formulation provides a closed-form equation to calculate a p-value: the probability of change in allele frequency at least as extreme as that observed, under the null hypothesis of genetic drift alone. Where this p-value is sufficiently small, we have evidence against this null hypothesis. Because there are 1489 originally detected somatic mutations that were then assessed with lofreq, we adjust for multiple testing by comparing each p-value to a Bonferroni corrected alpha (α=0.05/1489= 3.35 x 10^{-5}).

Where there is significant evidence against the null hypothesis of drift alone, we estimate s, the selection coefficient of the variant. The selection coefficient is a measure of the relative fitness of individuals carrying the minor allele[145]. Individuals carrying the minor allele have an increased probability of reproducing by a factor of $(1 + s)$. We construct a grid of possible values for s: (-1.0, -0.9, -0.8, -0.7, -0.6, -0.5, -0.4, -0.3, -0.25, -0.2, -0.15, -0.1, -0.05, -0.01, 0.0, 0.01, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0) based on the changes in minor allele count in the data set. We use an upper limit of $s = 1.0$ (indicating very strong selection and increasing the probability of reproduction by a factor of 2) because a larger s is unlikely when the observed final minor allele frequencies do not approach 1.0. We have additional points between -0.25 and 0.25 to increase precision where we anticipated the most likely estimates of s. For each possible value for s in the grid, we calculate $L\left(s|\ n_0, k_F^{obs}, n_F, k_I^{obs}, n_I\right)$, the likelihood of this s given the observed data. Maintaining the basic model above, we adjust the step-wise

probabilities in (3.2) such that individuals carrying the minor allele have an increased probability of being selected for replication of $(1 + s)$. After normalization, the probability of k_1 minor alleles in the next generation, given k_b in the current population of size n_b is:

$$P(k_1|n_b, k_b) = \begin{Bmatrix} \frac{n_b - k_b}{n_b + k_b s} & if\ k_1 = k_b \\ \frac{k_b + k_b s}{n_b + k_b s} & if\ k_1 = k_b + 1 \\ 0 & otherwise \end{Bmatrix} \quad (3.16)$$

Under this formulation, the closed-form expression of (3.4) is no longer available. Therefore, to obtain Part B_3 in equation (3.13), we apply Monte Carlo integration based on 10,000 random walks using these step-wise probabilities from the post-bottleneck population size, $n_0 = 200$, to the final population size, N_F. For each walk, we record the simulated final minor allele frequency, K_F, producing a probability distribution over K_F to be used for Part B_3. The rest of the model remains unchanged. The s with the maximum likelihood across this grid is the maximum likelihood estimate (MLE) of the selection coefficient. In addition, we calculate the acceptance region of the MLE using the log-likelihood ratio test with a chi-square value of one degree of freedom. The grid values whose likelihoods are contained in this region make up the reported 95% confidence interval.

3.4.3 Cell Growth Results

For each of the 58 variants identified by *lofreq*, we assess the evidence against a null model of genetic drift. Twenty of the variants showed nominally significant departure from this null hypothesis, but only three of these variants showed significant evidence after correcting for multiple testing (p-value < 3.35×10^{-5}) (Table 3.3,Table 3.4). All three of these significant variants increased in frequency from early to late passage. Two of the three variants were not

detected in the initial sequenced sample, thus likely having a very low initial allele frequency. The most significant p-value was for variant ID chr9:21974774:A, with a minor allele frequency change from 0 to 55.3% (Table 3.3). While they did not pass Bonferroni-corrected significant levels, ten of the twenty nominally significant variants were observed in the early passage sample but not in the final passage sample. Twelve of these twenty variants increase in minor allele frequency, while eight decrease. In these twenty variants, the total count size increased from early to late passage in seven variants and decreased in thirteen.

For the three significant variants, we found the maximum likelihood estimate of the selection coefficient. All three of the variants have estimates of strong positive selection (s greater than 0.25) (Table 3.4). The variant with highest significant value has the strongest selection coefficient estimate of 0.5. For these variants, we also calculate the 95% confidence intervals of the selection coefficient. We observe these intervals are large, but they include contain only high values of positive selection.

Table 3.3 Assessing the Null Hypothesis of Genetic Drift

Variant ID Number	Initial Population Sample		Final Population Sample		P-Value
	Minor Allele Count	Total Count	Minor Allele Count	Total Count	
chr9:21974774:A	0	725	156	282	$3.87*10^{-13}$
chr5:60169670:T	0	2168	122	705	$3.02*10^{-14}$
chr13:32914714:T	42	3122	642	2996	$9.82*10^{-39}$
chr5:1294984:T	29	331	0	195	$2.84*10^{-04}$
chr5:1253376:A	40	1267	0	1653	$2.29*10^{-03}$
chr5:1253377:A	43	1274	1	1641	$3.41*10^{-03}$
chr6:158613635:A	0	2694	48	3109	$9.08*10^{-03}$
chr2:47708066:A	50	1762	0	659	$1.03*10^{-02}$
chr11:108238380:T	0	2103	30	2013	$1.23*10^{-02}$
chr5:1272305:T	1	2180	20	1128	$1.36*10^{-02}$
chr10:69644914:C	0	1389	13	758	$1.39*10^{-02}$
chr5:1255344:C	0	895	45	2587	$1.68*10^{-02}$
chr9:100444492:G	0	2267	9	761	$2.22*10^{-02}$
chr11:108238298:A	0	2452	14	1671	$3.40*10^{-02}$
chr17:7577097:A	22	1423	0	1689	$3.48*10^{-02}$
chr10:50682121:C	0	1735	28	3218	$3.85*10^{-02}$
chr2:128047816:A	0	2894	20	3023	$4.07*10^{-02}$
chr12:21654214:A	37	1871	2	1799	$4.13*10^{-02}$
chr15:91260568:A	27	1188	0	393	$4.24*10^{-02}$
chr9:121929565:A	24	1782	0	1657	$4.94*10^{-02}$
chr13:49033845:T_2	0	1903	16	2335	$5.01*10^{-02}$
chr13:33591216:A	0	1303	12	1469	$5.34*10^{-02}$
chr13:49033845:T_1	0	1914	13	2022	$5.39*10^{-02}$
chr12:102795854:T	0	1840	12	2525	$7.23*10\text{-}^{02}$
chr9:122011379:A	38	2727	2	2892	$7.76*10^{-02}$
chr3:56213963:A	1	1355	17	1917	$7.85*10^{-02}$
chr19:2428438:A	19	1033	0	367	$7.89*10^{-02}$
chr13:49054511:T	2	2005	17	1928	$7.96*10^{-02}$
chr5:1266593:T	26	1418	1	616	$9.20*10^{-02}$
chr13:49033845:T_3	8	522	1	1006	$9.52*10^{-02}$
chr8:31004723:A	1	1343	18	2583	$1.04*10^{-01}$
chr6:158619968:T	0	969	9	1827	$1.08*10^{-01}$
chr4:120242868:A	10	1461	0	2472	$1.11*10^{-01}$
chr10:50747184:C	14	1733	0	1879	$1.14*10^{-01}$
chr15:99507532:T	2	2128	14	2335	$1.17*10^{-01}$
chr13:48921931:T	1	1777	10	2064	$1.18*10^{-01}$

chr13:32921003:T	17	2940	0	2796	$1.22*10^{-01}$
chr10:50678713:A	26	2448	0	885	$1.23*10^{-01}$
chr14:56472467:T	21	1767	0	603	$1.28*10^{-01}$
chr9:122075467:A	22	1781	0	538	$1.29*10^{-01}$
chr17:73645363:A	16	1993	0	1533	$1.33*10^{-01}$
chr6:83526997:T	24	2334	1	1959	$1.37*10^{-01}$
chr13:49055813:A	22	2573	1	2907	$1.46*10^{-01}$
chr11:66611768:A	1	785	18	2354	$1.50*10^{-01}$
chr13:49055814:A	21	2560	1	2915	$1.55*10^{-01}$
chr19:45855602:A	15	1349	0	402	$1.84*10^{-01}$
chr5:1253380:T	12	2033	0	1458	$1.97*10^{-01}$
chr3:189507602:A	11	1748	1	3248	$2.01*10^{-01}$
chr3:189507601:A	11	1751	1	3243	$2.02*10^{-01}$
chr5:126113018:T	11	1302	0	623	$2.06*10^{-01}$
chr19:45855601:A	14	1148	0	269	$2.10*10^{-01}$
chr5:1271115:A	8	721	0	314	$2.11*10^{-01}$
chr5:1253379:T	11	2016	0	1428	$2.16*10^{-01}$
chr5:126113019:T	10	1309	0	623	$2.32*10^{-01}$
chr13:32910631:A	14	2040	0	670	$2.55*10^{-01}$
chr5:60169882:A	13	2613	1	3089	$2.65*10^{-01}$
chr12:102791138:A	11	2040	2	2652	$3.39*10^{-01}$
chr15:99501088:G	9	1586	1	475	$4.92*10^{-01}$

Table 3.4 Variants with Significant Evidence against the Null Hypothesis of Drift Alone and Corresponding Selection Coefficient Estimates

Variant ID Number	Initial Population Sample		Final Population Sample		P-Value	MLE of Selection Coefficient	95% Confidence Interval
	Minor Allele Count	Total Count	Minor Allele Count	Total Count			
chr9:21974774:A	42	3122	642	2996	$3.87*10^{-13}$	0.25	[0.2, 0.4]
chr5:60169670:T	0	2168	122	705	$3.02*10^{-14}$	0.3	[0.2, 0.6]
chr13:32914714:T	0	725	156	282	$9.82*10^{-39}$	0.5	[0.4, 0.8]

3.5 Conclusion

In this chapter, we present a mathematical model for population bottlenecks followed by genetic drift and two applications: mtDNA transmission in humans and fibroblast cell growth. The first application, focused on the transmission of human mtDNA heteroplasmy across the entire mtDNA genome, is one of the largest mtDNA genome studies to date, and provides several important insights. We used the shifts in heteroplasmy minor allele frequency (MAF) from mothers to offspring to estimate the size of the bottleneck that occurs during the transmission of mtDNA genomes. The size of the bottleneck was estimated under four models: a constant size bottleneck model, in which each mtDNA genome is a segregating unit and the bottleneck size does not vary between individuals; a variable size bottleneck model, in which each mtDNA genome is a segregating unit and the bottleneck size is allowed to vary between individuals; a constant size nucleoid model, in which a nucleoid containing a variable number of homoplasmic mtDNA genomes is the segregating unit and the bottleneck size does not vary between individuals; and a variable size nucleoid model, in which a nucleoid containing a variable number of homoplasmic mtDNA genomes is the segregating unit and the bottleneck size is allowed to vary between individuals. The best fitting model (as determined by AIC values) was a variable size bottleneck, with an estimated mean of 9 individual mtDNA genomes transmitted 4 from mothers to offspring.

This number is smaller than a recent estimate of 30-35 mtDNA genomes transmitted, based on 39 mother-offspring pairs[135]. Although this previous study assumed a constant-size bottleneck model, our estimate for a similar constant-size bottleneck model is also smaller, about 8 mtDNA genomes transmitted. The reason for this discrepancy is most likely because we do not assume that the observed minor allele frequency in the child is identical to the minor allele

frequency at transmission (immediately after the bottleneck). Instead, we model the replication process from the bottleneck to the actual mtDNA population in the child, thereby allowing for genetic drift during the replication process. Doing so allows for substantial changes in MAF during the replication process, but such changes will only be substantial if the bottleneck size is small. Incorporating drift in this way has two consequences: first, the same bottleneck model can be consistent with the few variants in the dataset that have drastic changes in allele frequency and with the large set of variants in the dataset that show a smaller change. Second, small MAF in the offspring do not require very large bottleneck sizes. Consider that without a drift model, the smallest nonzero allele frequency possible is 1/n, where n is the bottleneck size. Hence without modeling drift, all descendants with a very low MAF provide strong evidence for a large bottleneck size. However, by including drift, the final MAF in the offspring can be substantially smaller than the frequency at the bottleneck. Because drift can only have substantial effects if the bottleneck size is small, this explains the estimate of a relatively small number of transmitted mtDNA genomes.

A variable-size bottleneck with each mtDNA genome as a segregating 1 unit fit the data better than models involving nucleoids. However, this is not necessarily evidence against nucleoids, as we assumed an average of 7.5 mtDNA genomes per nucleoid, in accordance with some observations[139; 140; 142]. If instead the number of mtDNA genomes per nucleoid is smaller, then the results based on nucleoids will approach the results based on mtDNA genomes as segregating units; in the limit, if each nucleoid contains exactly one mtDNA genome, as suggested by some studies[146] , then both models will give identical results. Our results therefore argue against the existence of nucleoids with several mtDNA genomes, but not necessarily against nucleoids with smaller numbers of mtDNA genomes. The most important conclusion is

that the size of the bottleneck varies among individuals, whereas all previous attempts to model the size of the bottleneck have assumed that it is constant among individuals. Identifying the factors that influence this between-individual variation in bottleneck size would be of great interest and might have consequences for understanding the transmission of mtDNA-related diseases.

One limitation of this approach is that we are using the MAF observed in the mother's blood several years after conception as the estimate for the MAF in the egg at the time of conception. In the absence of data on heteroplasmy in human eggs this limitation is unavoidable, although one way to improve the estimate would be to utilize heteroplasmy data from multiple tissues, as was done recently elsewhere[135]. To further investigate this potential limitation, a previous study of heteroplasmy variation across different tissues[147] calculated the correlation in MAF at heteroplasmic positions in blood and ovarian tissue from the same individual. There were 52 heteroplasmies with MAF>0.02 detected in either blood or ovarian tissue (or both) in individuals with data from both tissues, and the MAF in blood exhibits a modest but nonetheless significant correlation with that in ovarian tissue (Pearson's correlation = 0.62, $p<0.0001$). This would suggest that the MAF in blood is a reasonable proxy for the MAF in ovarian tissue, although data on heteroplasmy in human eggs would still be desirable. A significant correlation between the mother's age at conception and the number of heteroplasmies detected in the offspring was reported previously[135]. However, there is no such correlation in the GoNL data (Pearson's rho = -0.03, $p =0.65$), even though the range of mother's ages at conception is similar between the two studies (range = 18 – 44). The reason for this difference is unclear and further studies are warranted.

In the second application, we explore somatic mutations that exist at low frequencies in tissues of healthy individuals or individuals with premature aging conditions. Applying a mathematical model, we calculate the probability of observing the change in allele frequency from the initial to final population under the null hypothesis of genetic drift acting alone. We find three cases where there is significant evidence to reject this null hypothesis. The most striking significant evidence against the null was observed for the variant chr9:21974774, a mutation in the *CDKN2A* gene, detected in the fibroblast cells from an XPA patient. This mutation is located in a CpG dinucleotide within a CpG island. These islands are mutational hotspots, with fifteen times the mutation rate observed as other sites[148; 149]. This mutation was previously identified in liver carcinoma and suggested to be one of the *CDKN2A* inactivation mechanisms[149; 150]. The mutation is also within the binding site of several transcription factors, including *EZH2*. *EZH2* is a known inhibitor of the *INK4A-ARF* pathway, involved in beta-cell regulation and cellular senescence[151; 152].

The cell populations of this XPA individual also carried another mutation, chr5:60169670, located in the 3'UTR of *ERCC8*. This somatic mutation also showed significant evidence against the null hypothesis of genetic drift. The *ERCC8* mutation may have a possible functional effect based on its location in regions of two microRNA, though this may not explain the extreme shift in allele frequency. The occurrence of these two mutations in the same individual suggest that while the *CDKN2A* mutation could drive the clonal expansion of cells that have this mutation, with the *ERCC8* mutation as a passenger mutation. The *ERCC8* mutation does not appear to have a strong negative effect on the cell and, in agreement with the lower allele frequency in the final population of cells compared to the allele frequency of the *CDKN2A* mutation, could be secondary to the *CDKN2A* mutation.

The third mutation, that indicated by its change in allele frequencies to provide a selective advantage, was chr13:32914714:T, located in *BRCA2*. This mutation was found in a sample from a healthy subject of old age (85 years old). Analysis by collaborators using the Human Splicing Finder and Alamut splicing tools for the functional significance of the *BRCA2* mutation showed that it could potentially cause aberrant splicing of the C-terminal region. The region is essential for BRCA2 function and its interaction with RAD51[153]. Even in the case of loss of BRCA2 function, it is not expected that this mutation alone to provide proliferative advantage to its cells. This mutation, however, could result in increased genomic instability in its cells due to a dysfunctional mechanism of homologous recombination of double-strand breaks. It is also possible that this mutation is a passenger for a variant that was not observed in the dataset.

One criticism of this model is the assumption of no deaths during the period of regrowth. While it is possible some cells are lost during the regrowth period, we believe this to be a minimal amount. Further, if deaths are occurring at random, our results should be largely unaffected. In the case were cells are dying do to a selection pressure, this would be captured in the evidence against the null hypothesis of genetic drift alone. Therefore, the evidence for the three positively selected variants would be maintained. For future analyses, where deaths play a larger role in the evolution of a population, the probabilities of regrowth could be easily adjusted. Additionally, the initial and final cell populations for each variant are derived from a sample form a single individual. While this is an interesting first step, to make broader claims of the selective pressures of these variants, it may be useful to see if similar trends are seen in these variants when the same procedure is applied to samples from many individuals. Another possible extension could be to manipulate the environmental pressures of the cells, applying this model to assess departures from genetic drift alone under these changes. Finally, while this procedure

identifies signals of selection, it is not clear if the variants are the targets themselves or hitch-hiking artifacts. Functional analyses on these variants and variants in the region could clarify these conclusions.

In this chapter, we present a flexible, probability-based approach to model population bottlenecks and genetic drifts, including sequencing and sampling error. With two applications of the model, we show the utility and accessibility of the approach. We provide insight into the mechanisms of mtDNA transmission as well as the functional importance of certain variants during cell growth.

CHAPTER 4: Detecting Positive Selection Signals in Autoimmune Disease Associated Loci with Whole Genome Sequencing Data

4.1 Introduction

Positive selection is the process by which advantageous genetic variants increase in frequency in a population due to improved fitness and reproduction. Also called "Darwinian selection" or "natural selection", selection was the driving force in Darwin's Theory of Evolution, creating the vast amounts of genetic variation within species and the divergence between species[154]. In contrast, Kimura's Neutral Theory of Molecular Evolution states that the majority of this variation is driven by the random process of genetic drift and that most alleles are selectively neutral[155]. One motivating factor for studying positive selection in modern genetics is to distinguish between these theories of evolutionary origins and identify what relative importance drift and selection have in genetic and phenotypic diversity among humans. Additionally, identifying genetic regions that are under selection can provide important functional information of the genetic variants. We are particularly motivated by understanding the underlying genetic model of disease, therefore providing biological insights and potentially leading to future medical interventions.

In this study, we focus on identifying positive selection in autoimmune disease associated loci. Autoimmune diseases are defined by abnormally low activity or over-activity of the immune system. The immune systems in individuals with these diseases are reacting against normally-occurring antigens in the body as if these antigens were foreign. Collectively, the

diseases are a leading cause of death in women in the United States and are known to have a negative effect on reproductive fitness[54; 156]. However, the diseases remain prevalent, with epidemiological studies estimating autoimmune diseases collectively affect at least 5% of individuals worldwide[53]. The diseases included in this category, such as Celiac disease, rheumatoid arthritis, systemic lupus erythematosus, ankylosing spondylitis, type 1 diabetes, and inflammatory bowel disease, present a wide range of symptoms and affect a variety of organ systems. Though phenotypically diverse, the diseases share multiple associated loci. We focus on these loci as they may contribute to a broader, shared immune response and implicate common pathways under selection.

The detrimental effect on fitness paired with the prevalence of autoimmune diseases presents an evolutionary conundrum. The thrifty gene hypothesis is one attempt at explaining this occurrence with positive selection. This hypothesis states that variants in the present day that appear to confer detrimental attributes persist in the population because they previously offered some other evolutionary advantage. James V. Neel first introduced this hypothesis in 1962 to explain the persistence of Type 2 diabetes, suggesting its driving variants were selected during times of food shortages and became detrimental with the ease of modern access to nutrition[157]. A similar concept emerged later in the field of epidemiology. Incidence of autoimmune diseases has increased in the past three decades, particularly for inflammatory bowel disease, Type 1 diabetes, and multiple sclerosis[158]. At the same time, epidemiological studies show a decrease in infectious burden with industrialization[158; 159]. In 1989, Strachan suggested a link may exist between these discordant trends, coining the term ‘the hygiene hypothesis’[160]. Autoimmune disease associated loci could be maintained in the population because they were previously necessary to offer protection from infectious diseases or foreign pathogens. These past selective

events create distinct signals in present day human genetic variation. Therefore, to investigate the 'hygiene hypothesis', we will conduct a comprehensive search for these signals in autoimmune disease associated loci.

A few studies have previously investigated the 'hygiene hypothesis' for autoimmune diseases. For example, human leukocyte antigen (HLA) genes, which are highly associated with autoimmune diseases, are well-studied and show strong evidence of balancing selection[161-164]. Existing studies on non-HLA autoimmune-associated genes driven by pathogen selection are more limited. These studies primarily use a combination of methods from three categories with genome wide association data. The first category identifies regions under balancing selection with correlations between pathogen richness and genetic variability[165; 166]. Fumagalli *et al* found several risk alleles in interleukin genes to be significantly correlated with micropathogen richness[167]. The second approach relies on comparing haplotype length using the integrated haplotype score (iHS)[54; 166; 168-170]. This method aims to identify alleles that have swept to intermediate frequencies by comparing the width of linkage disequilibrium surrounding a derived allele to that of the ancestral allele in the same position[170]. Using this approach followed by functional analysis, Zhernakova et al identified evidence of positive selection in the *SH2B3* locus, a primary risk variant for Celiac disease[169]. The final approach compares allele frequencies between populations. Using population differentiation measures, such as F_{ST}, these studies aim to identify variants affected by different selection pressures[54; 167; 168]. This previous research is primarily restricted to genome-wide association studies (GWAS). There are several drawbacks to utilizing this common variant genotyping data for scans of selection including the presence of ascertainment bias, in which the nonrandom sampling of single nucleotide

polymorphisms on an array distorts measures of human diversity[171]. In the case of haplotype-based tests, the power also relies on the phasing accuracy in the GWAS[170].

To our knowledge, this is the first study searching for evidence of selection in a collective gene set using variants discovered by large-scale whole genome sequencing. Sequencing, compared to GWAS, can better identify small regions of linkage disequilibrium to understand complicated signals of selection, as shown in the hemoglobin beta gene[172; 173]. Furthermore, whole genome sequencing studies, rather than sequencing portions of the genome, have the advantage of obtaining a full picture of variation in an individual's genome. Where certain population histories can confound signals of selection in the genome, whole genome sequencing circumvents this issue. Because population history affects the entire genome, whole genome sequencing provides an internal control. We will show we can adjust for these effects by comparing the loci of interest to an empirical distribution of the genome-wide statistics. Moreover, large scale whole genome sequencing studies increase access to rare variation. The clear majority of genetic variation within genes is rare, arose recently, and is highly population specific[10; 11]. Rare variants provide information on the very recent past and may be able to capture signals of recent positive selection that were previously unattainable. We aim to exploit the power of whole genome sequencing in this scan for selection in autoimmune loci.

To investigate positive selection in autoimmune loci collectively, we will apply two site frequency spectrum (SFS) tests. These two tests are based on theoretical properties of the frequency distribution of variants in a sample. Under the Neutral Theory of Kimura, the fate of mutations under neutral evolution is determined entirely by random genetic drift[112; 174]. In the presence of selection, however, the site frequency spectrum does not follow these theoretical expectations[112]. For instance, when a new, strongly selected advantageous mutation quickly

increases in frequency, the variation in the neighboring ("linked") regions is reduced, limiting the number of intermediate frequency alleles[111; 175]. At the time close to fixation, the site frequency spectrum is characterized by an abundance of rare variants, which arose recently, and high-frequency alleles, which "hitchhiked" during the process of selection[111; 175; 176]. Therefore, examining the site frequency spectrum can identify potential signals of selection.

However, the site frequency spectrum is also affected by population forces other than selection. For example, exponential population growth creates an excess of rare variants in spectrum[29; 177; 178]. Also, previous population bottlenecks and the presence of population structure can affect the spectrum, causing an excess of intermediate frequency alleles[29; 177; 179]. In present day humans, it is therefore difficult to tease apart true selection signals from these effects. In applying site frequency spectrum tests, we must be able to account for these confounding signals.

In this study, we focus on two site frequency spectrum tests: Tajima's D and Fay and Wu's H. Each test is calculated by contrasting different estimators of the scaled mutation rate. The estimators have varying sensitivities to the excess or depletion of low, intermediate, and high frequency alleles[175]. Therefore, obtaining the difference between estimators provides information on which parts of the spectrum are divergent from neutrality and which population forces are most likely responsible. Significant deviations from zero in the differences indicate the standard neutral model should be rejected[29; 111]. Tajima's D[180] measures departures from neutral evolution that are specifically reflected in the difference between low-frequency and intermediate frequency alleles[175]. Fay and Wu's H[181] measures departures from neutral evolution that are specifically reflected in the difference between high-frequency and intermediate frequency alleles[175]. Tajima's D is particularly sensitive to exponential growth, which causes an abundance of rare alleles. Fay and Wu's H does not have this issue; however, it does require

ancestral information to identify derived alleles. Using both tests can help capture information missed or confounded by other factors in one test alone[175].

We apply these tests to the new context of large-scale whole genome sequencing. While we emphasized the benefits of using this data type, there are several issues remaining to be investigated in applying SFS methods to whole-genome data. Applying this approach to previously identified genomic regions of positive selection, we assess its utility in identifying selection signals for a collective gene set. We examine the optimal window size in which to calculate the statistics when constructing a distribution, the relative similarity of genic vs. non-genic windows, and the appropriate way to account for dependency in adjacent windows. After establishing these parameters, we explore signals of selection in 39 autoimmune loci.

4.2 Data

The data used for this project comes from the Bipolar Research in Deep Genome and Epigenome Sequencing (BRIDGES) Consortium. This is a multi-center case-control study aimed at identifying the genetic contributions to bipolar disorder. The dataset includes whole genome sequencing for 3675 individuals of European descent with an average coverage of 9.2x. There are 68,020,887 biallelic SNPs in the dataset. Using the folded site frequency spectrum, most variation is rare: 57.36% of SNPs are singletons (the minor allele is observed only once) and 9.81% of SNPs are doubletons (Figure 4.1). Of these polymorphisms, we can infer ancestral information for 93.84%, allowing for their use in Fay and Wu's H calculations.

We consider three sets of genes for this analysis. We begin by focusing on identifying selection signals in genes previously identified as positively selected. Second, we apply our method to a collection of genes previously implicated in two or more autoimmune diseases. Finally, we study six autoimmune disease-specific gene sets. We analyze gene sets individually

for six autoimmune diseases independently: Celiac disease (CD), rheumatoid arthritis (RA), inflammatory bowel disease (IBD), systemic lupus erythematosus (SLE), Ankylosing spondylitis (AS), and type 1 diabetes (T1D).

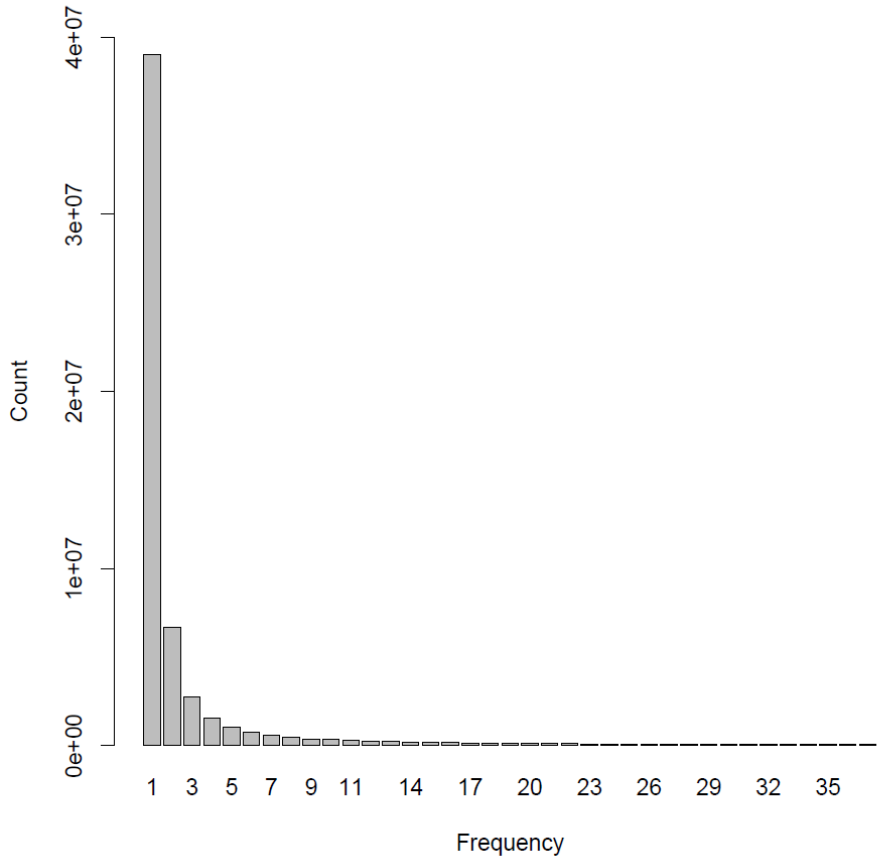

Figure 4.1 Folded Site Frequency Spectrum in BRIDGES Data
We show the folded site frequency spectrum for biallelic SNPs with minor allele frequency less than 0.05%.

4.2.1 Previously Identified Positively Selected Genes

To determine the optimal parameters for applying the SFS tests to whole genome sequencing data, we identified a subset of genes previously identified as positively selected from past studies using an in-depth literature search. Each gene we included in this positively selected gene (PSG) list has at least two studies identifying a signal of positive selection and no existing

conflicting studies (Table 4.1). The genes included are LCT, LYZ, MCPH1, CYP3A5, HFE, BRCA1, and ACKR1 (Duffy blood group). Using the empirical distribution from the whole genome sequencing data, we assessed our ability to identify selection signals with site frequency spectrum tests in this collection of genes under windows of size 1 kb, 10 kb, and 100 kb.

Table 4.1 Previously Identified Positively Selected Genes

Gene Name	Gene Location	Brief Description	Supporting Literature
LCT	2q21.3	Involved in lactase metabolism	Bersaglieri et al, 2004[182]; Liu et al, 2013[183]; Schlebusch et al[184], 2013; Hollox et al, 2001[185]
LYZ	12q15	Encodes an antimicrobial agent found in human milk and organs	Messier and Stewart, 1997[186]; Yang et al, 1998[187]
MCPH1	8p23.1	Involved in regulation of chromosome condensation and DNA damage response	Pulvers et al, 2015[188]; Shi et al, 2013[189]; McGown et al, 2011[190]
CYP3A5	7q33.1	Involved in liver enzymes and metabolism of drugs	Bans et al, 2013[191]; Thompson et al, 2004[192]; Chen et al, 2009[193]
HFE	6p22.2	Involved in iron absorption	Toomajian et al, 2003[194]; Toomajian and Kreitman, 2002[195]; Ajioka et al, 1997[196], Thomas et al 1998[197]
BRCA1	17q21.31	Breast cancer gene 1, critical for DNA repair, cell cycle, and genomic stability	Huttley et al, 2000[198]; Lou et al, 2014[199]
ACKR1	1q23.2	Duffy blood group	Hamblin et al, 2000[200], Hamblin et al, 2002[201]

4.2.2 Autoimmune Genes

For this part of the analysis, we consider a collection of 39 non-HLA loci previously implicated in two or more autoimmune diseases. While HLA genes contribute the strongest signals of association with autoimmune diseases, the strength of these associations and the extensive LD in the MHC region make it difficult to identify independent associations from common haplotypes. Therefore, we focus on the non-MHC pathways that also show strong associations in immune mediated diseases.

Though phenotypically diverse, autoimmune diseases share multiple associated loci and are believed to have common etiopathogenic factors[202]. We focus on these loci as they may contribute to a broader, shared immune response and implicate common pathways under selection. Parkes *et al* recently performed a systematic analysis of autoimmune loci to identify shared susceptibility loci based on individual studies in the ImmunoBase website[202]. In this analysis, we use this list of genes associated with two or more autoimmune diseases. These gene groups and pathways include: IL-23 and T_H1, NF-κB, aminopeptidase, IL-2 and IL-21, IRF family, T-cell co-stimulation, PTPN2 and PTPN22, ubiquitylation, and viral response (Table 4.2)[202].

Table 4.2 Non-HLA Genes Used in Analysis
An abbreviated table from Parkes et al showing genes and pathways implicated in two or more autoimmune diseases[202]. We abbreviate inflammatory bowel disease (IBD), type 1 diabetes (T1D), and systemic lupus erythematosus (SLE). Parkes et al identified genes using individual studies in the ImmunoBase website.

Pathway or Gene Group	Positional candidate genes shared by 2 or more diseases	Diseases associated with this pathway or one or more gene
IL-23 and T_H1	*IL23R* (1p31), *IL12B* (5q33), *IL12A* (3q25), *TYK2* (19p13), *JAK2* (9p24), *STAT3* (17q21), *STAT4* (2q32), *IL27* (16p11) and *CCR6* (6q27)	Ankylosing spondylitis, IBD, psoriasis, coeliac disease, rheumatoid arthritis, T1D, SLE, and multiple sclerosis
NF-κB	*REL* (2p16), *TNFAIP3* (6q23), *NFKB1* (4q24) and *TNIP1* (5q32)	IBD, psoriasis, coeliac disease, rheumatoid arthritis, T1D, SLE, and multiple sclerosis
Aminopeptidase	*ERAP1* (5q15) and *ERAP2* (5q15)	Ankylosing spondylitis, IBD, and psoriasis
IL-2 and IL-21	*IL2*, *IL21* (4q26), *IL2RA* (10p15) and *IL2RB* (22q13)	IBD, coeliac disease, rheumatoid arthritis, T1D, and multiple sclerosis
IRF family	*IRF4* (6p25), *IRF5* (7q32), *IRF7* (11p15) and *IRF8* (16q24)	IBD, psoriasis, coeliac disease, rheumatoid arthritis, SLE, and multiple sclerosis
T-cell co-stimulation	*CD40* (20q12), *CD28*, *CTLA4*, *ICOS* (2q33) and *ICOSLG* (21q22)	Ankylosing spondylitis, IBD, coeliac disease, rheumatoid arthritis, and multiple sclerosis
PTPN2 and PTPN22	*PTPN2* (18p11) and *PTPN22* (1p13)	IBD, coeliac disease, rheumatoid arthritis, T1D, and SLE
Ubiquitylation	*UBE2L3* (22q11)	Ankylosing spondylitis, IBD, psoriasis, coeliac disease, rheumatoid arthritis, SLE and multiple sclerosis
Viral Response	*IFIH1* (2q24)	IBD, psoriasis, T1D and SLE
Other	*IL10* (1q32)	IBD, T1D and SLE
	IL18RAP (2q12)	IBD, coeliac disease and T1D
	FCGR2A (1q23)	Ankylosing spondylitis, IBD (ulcerative colitis), rheumatoid arthritis, T1D, SLE and multiple sclerosis
	PTGER4 (5p13)	Ankylosing spondylitis, IBD and multiple sclerosis
	BACH2 (6q15)	Ankylosing spondylitis, IBD, coeliac disease, T1D and multiple sclerosis
	CARD9 (9q34)	Ankylosing spondylitis and IBD
	ZMIZ1 (10q22)	IBD, psoriasis, coeliac disease and multiple sclerosis
	YDJC (22q11)	IBD, psoriasis, coeliac disease, rheumatoid arthritis and SLE
	TAGAP (6q25)	IBD, psoriasis, coeliac disease, rheumatoid arthritis, T1D and multiple sclerosis
	PRDM1 (6q21)	IBD, rheumatoid arthritis and SLE

4.2.3 Disease-Specific Autoimmune Genes

In addition to studying these shared associated genes, we search for signals of selection that may be specific to one disease. We analyze gene sets individually for six autoimmune diseases independently: Celiac disease (CD), rheumatoid arthritis (RA), inflammatory bowel disease (IBD), systemic lupus erythematosus (SLE), Ankylosing spondylitis (AS), and type 1 diabetes (T1D). We use the Phenotype-Genotype Integrator (PheGenI) website (https://www.ncbi.nlm.nih.gov/gap/phegeni/) from the National Center for Biotechnology Information (NCBI), which combines NHGRI genome-wide association study (GWAS) catalog data with several databases housed at NCBI, including Gene, dbGaP, OMIM, GTEx and dbSNP to list genes associated with a trait. For each of the six diseases studied here, we include genes identified as associated with a p-value of less than 10^{-8}.

4.3 Methods

4.3.1 Site Frequency Spectrum Statistics

We calculate two different site frequency spectrum statistics to determine if a sequence evolved through neutral process or if there is evidence of some non-neutral process, such as selection. Each of these statistics relies on contrasting estimators of the scaled mutation rate, θ. For human diploid populations, $\theta = 4N_e\mu$, where N_e is the effective population size and μ is the average number of new neutral mutations in each generation[112]. Under the Neutral Theory of Kimura, the fate of mutations that are strictly selectively neutral is determined entirely by random genetic drift[112; 174]. On the other hand, mutations that are under selection pressure are more likely to rapidly increase or decrease in population frequency, affecting the estimation of the neutral mutation rate[112]. The different estimators of θ have varying sensitivities to the excess or depletion of low, intermediate, and high frequency alleles. Each of these estimators is

unbiased under neutrality (no selection, no population subdivision, and no changes in effective population size over time), meaning the difference between estimators is expected to be zero. Significant deviations from zero indicate the standard neutral model should be rejected[29; 111]. Since population genetics forces affect parts of the site frequency spectrum differently, the specific estimators that differ can be informative of which forces are acting[175].

We use three different estimators of mutation rate, θ, in these site frequency spectrum statistics, each unbiased under neutrality. Fu *et al* showed that $E(\varepsilon_i) = \frac{\theta}{i}$ for $i = 1, \dots, n-1$, where ε_i is the number of times the derived allele is observed i times in the sample of size n [175; 203]. We refer to "derived" alleles as the mutant allele based on the ancestral state determined by an outgroup, in this case the chimpanzee. Using this framework, there are many possible unbiased estimators of θ [175; 203]. Each unbiased estimator we consider can be written as a linear function of ε_i, while weighting different frequency classes[175]. The first estimator, Tajima's estimator, is calculated as $\hat{\pi} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} i(n-i)\varepsilon_i$ [175]. This weighting of ε_i makes Tajima's estimator particularly sensitive to an excess or depletion of intermediate frequency alleles[175]. More frequently, Tajima's estimator, $\hat{\pi}$, is written as $\frac{2}{n(n-1)} \sum_{i<j} \pi_{ij}$, with π_{ij} the pairwise difference in sequence i and j [52]. This has the convenient interpretation of the average number of pairwise differences in a sample and does not require knowledge of the ancestral state[29; 111]. The second estimator of mutation rate, the Watterson estimator, is calculated as $\hat{\theta}_W = \frac{1}{a_n} \sum_{i=1}^{n-1} \varepsilon_i$ where $a_n = \sum_{j=1}^{n-1} \frac{1}{j}$ [175; 176]. This estimator is particularly sensitive to changes in the number of low frequency alleles[175]. The Watterson estimator is often written as $\hat{\theta}_W = \frac{S_n}{a_n}$ with n the sample size, and S_n the observed number of mutations in the sample[29; 111; 175]. In this form, ancestral

information for the allele is not required. The third estimator, the H estimator, is calculated as $\hat{\theta}_H = \sum_{i=1}^{n-1} \frac{2\varepsilon_i i^2}{n(n-1)}$ [175; 176]. The weighting of ε_i in this estimator makes it sensitive to an excess of high frequency derived alleles[175]. This estimator requires ancestral information for the alleles. These three estimators are contrasted in the two site frequency spectrum statistics focused on in this study.

The first statistic, Tajima's D, shown in equation (4.1), measures the normalized difference between $\hat{\pi}$ and $\hat{\theta}_W$ [29; 111; 180]. Significant deviations from zero for this statistic indicate the basic neutral model does not capture the variation in the data, specifically reflecting a change in the number of low-frequency and intermediate-frequency alleles[111; 177]. After strong selection in the population, most of the tightly linked neutral sites on the haplotypes in the region of the selected variant should be identical, causing a decrease in intermediate frequency variants[177; 179; 204]. Of the mutations that exist in these haplotypes, they arose recently, meaning they are rare in the population. Thus, there will be elevated levels of low frequency alleles[204]. Applying this intuition to equation (4.1), we expect D to be negative in the case of positive selection[177; 179]. Negative values of D can also arise from population growth or background selection, which create an excess of rare variants[29; 177; 178]. Therefore, it is difficult to discern which force is the cause of the negative Tajima's D. Positive D values can arise from population bottlenecks, population structure, or balancing selection, each of these maintaining an excess of intermediate frequency alleles in the population[29; 177; 179].

$$D = \frac{\hat{\pi} - \hat{\theta}_W}{\sqrt{\widehat{Var}(\hat{\pi} - \hat{\theta}_W)}} \tag{4.1}$$

The second statistic, Fay and Wu's H shown in equation (4.2), measures the normalized difference between $\hat{\pi}$ and $\hat{\theta}_H$ [175; 176]. As in Tajima's D, this difference is expected to be zero under neutral evolution. Deviations from zero indicate departures from neutrality that are specifically reflected in the difference between high-frequency and intermediate-frequency alleles[175; 177]. The excess of high-frequency derived alleles is a hallmark of strong positive selection, where under rapid positive selection, the variant will increase in frequency[175]. Tightly linked regions will also be "swept" to higher frequencies in a process called hitchhiking. High-frequency variants are particularly useful because very few are expected under neutrality[175; 176]. Applying this intuition to (4.2), positive H values indicate a deficiency of high frequency variants and a negative H indicates an excess of high-frequency alleles. We expect H to be negative in the case of positive selection. Unlike Tajima's D, Fay and Wu's H is not affected by an excess of rare variants, meaning it can be distinguished from the effects of population growth or background selection[176; 178]. However, Fay and Wu's H has the disadvantage that it requires an outgroup to identify derived alleles[176]. This information is not always readily available, reducing the number of regions that can be studied with this method. Furthermore, ancestral alleles are misidentified will result in falsely extreme values of Fay and Wu's H. For these reasons, we will use both statistics in our analyses here.

$$H = \frac{\hat{\pi} - \hat{\theta}_H}{\sqrt{\widehat{Var}(\hat{\pi} - \hat{\theta}_H)}} \tag{4.2}$$

4.3.2 Empirical Distributions

In several tests for positive selection, population effects such as exponential growth and background selection can confound SFS statistics. Therefore, it is often unclear how to assess significance for a test statistic. To alleviate this problem, we calculate these statistics in

uniformly-sized non-overlapping windows across the genome. For each analysis, we use three different window sizes: 1,000, 10,000, and 100,000 base pairs. For each window size, we construct an empirical distribution of these statistics from the set of windows across the genome. We then compare the windows that overlap with our genes of interest to this distribution. We expect that windows that have selected genic content will exist at the extremes of the genome-wide empirical distribution. We are particularly interested in window that have some genic contents as these could be immediately informative of the function in specific pathways. We obtain a p-value for these windows directly from the corresponding quantile in the empirical distribution. We apply a Bonferroni correction to account for multiple testing when assessing significance of these p-values.

4.3.3 Genic vs. Non-Genic Window Distributions

The windows across the genome fall into two categories: "genic", if the window's base-pair coordinates overlap with a known gene, and "non-genic" otherwise. We are interested in if the empirical distribution of the site frequency spectrum statistics of genic windows differs from that of non-genic windows or all windows. This could provide insight into the differing effects of selection on genes and non-genes. To explore this question, we categorize genic windows using Ensembl gene annotation, defined by the outermost transcript start and end coordinates of the gene[205]. For this analysis, we focus specifically on the protein-coding genes based on their immediate functional impact as targets of selection. We construct empirical distributions for two different datasets: genic windows and all windows across the genome. We calculate and compare summary statistics for each distribution (mean, median, minimum, maximum).

Due to the non-normality of the distributions (each Shapiro-Wilkes and Anderson-Darling tests p-value <0.001), we use the non-parametric Kolmogorov-Smirnov test (K-S test) to

formally compare the distributions. The two-sample K-S test measures the probability that a sample dataset is drawn from the same underlying population as a second sample dataset. The test statistic, D, is defined as the greatest distance between the empirical distribution functions of the two samples[206]. The K-S test requires an assumption of independence in the datasets. This assumption is violated due to linkage disequilibrium throughout the genome, creating a dependency between the neighboring windows for each dataset. To identify the extent of this dependency, we calculate the autocorrelation across windows with lag 1 to 50. To alleviate these issues of dependency, we use a random sample from each dataset of 1000 windows. We calculate autocorrelation within each random subsample to verify the dependency has been eliminated. Finally, we apply the K-S test to these random subsamples. We repeat this procedure with larger sample sizes of 5,000 and 10,000 to confirm the results are not a function of sample size.

4.3.4 Rank-Based Testing

In addition to testing individual gene windows, we are interested in identifying selection signals that affect the collections of genes. We aim to determine if the SFS test statistics for windows overlapping autoimmune genes are significantly different than a random subset of windows across the genome. We apply a variation of the Wilcoxon Rank Sum Test, a non-parametric alternative to a t-test. The Wilcoxon Rank Sum statistic is obtained by ordering the data values across groups and summing these ranks within groups. The exact p-value is easily obtained through permutations for small samples. In this case, we are interested in an excess of both extremely high and low SFS test statistics. Using the sum of ranks as the test statistic can mask extreme signals that are present in the sample at both ends of the distribution. For this reason, we instead use the sum of the absolute difference from the mean rank. Consider the set of

windows overlapping the genes of interest: $X_1 \dots X_{n_1}$ and the sample from the remaining windows genome wide $Y_1 \dots Y_{n_2}$ with N total number of windows genome-wide. The testing procedure is as follows:

1. Obtain rank of each value genome-wide: $R(X_1) \dots R(X_{n_1}), R(Y_1) \dots R(Y_{n_2})$
2. Calculate mean rank $M = \frac{N}{2}$ genome-wide.
3. Calculate the absolute difference from the mean rank $W(X_1), \dots W(X_{n_1}), W(Y_1), \dots W(Y_{n_2})$ for each data value: $|M - R(X_1)| \dots |M - R(X_{n_1})|, |M - R(Y_1)| \dots |M - R(Y_{n_1})|$
4. Sum rank differences in group X, to obtain the observed test statistic: $S^* = \sum_1^{n_1} W(X_i)$.
5. Sum rank differences in group Y, to obtain $S = \sum_1^{n_1} W(Y_i)$
6. Randomly choose n_2 new windows among N ranks and assign as group Y. Re-compute S.
7. Repeat step (6) for each of the $\binom{N}{n_2}$ possible permutations or a large subset of these permutations for larger samples.
8. Record the number of times the observed value is greater than the permuted value to obtain the p-value, $p = P(S^* \geq S)$

We apply this testing procedure to each subset of genes outlined in section 4.2. Because many genes overlap multiple windows, dependency between neighboring windows can inflate the test statistic. To account for this dependency, the sample used for each group Y is drawn to mimic the dependency structure in our gene subset, by randomly sampling windows with the

same frequency of consecutive windows as in $X_1 \ldots X_{n_1}$. We perform 10,000 permutations to obtain a p-value, the probability a more extreme set of SFS test statistics than observed in the subset of interest.

4.3.5 Determining Window Size

To determine the optimal window size to power this approach, we apply the rank-based testing procedure to the known positively selected gene subset. This requires a balance between windows that are too large (averaging out any potential signal across the region) and windows that are too small (breaking up the potential signal between windows). We calculate individual p-values and apply the rank-based testing for windows of 1000 bp, 10,000 bp, and 100,000 bp. In each case, we expect to find significant p-values, identifying a signal for the positive selection known to exist in this collection of genes. We compare the results for each rank-based test (Tajima's D and Fay and Wu's H) for each window size and select the window with the most significant result as the optimal window size. This window size is used throughout the remainder of the results.

4.4 Results

4.4.1 Optimal Window Size

To determine the optimal window-size, we focus on the set of previously identified positively selected genes. We expect the SFS statistics for these genes to fall into the tails of the empirical distributions. Using window sizes of 1 kb, 10 kb, and 100 kb, we first calculate the empirical p-value for the Tajima's D statistic and Fay and Wu's H statistic in each window overlapping these genes. The QQ-plots for each of these sets of empirical p-values indicate that, for each window size, there are many points that reach nominal significance (hashed line) but

most points are within the limits of the Bonferroni-corrected significance line (black dotted lines) (Figure 4.2). Comparing the 1 kb window (Figure 4.2A, B) and 100 kb window plots (Figure 4.2E, F) shows the striking reduction in sample size with larger window size. In the 1 kb statistics (Figure 4.2A, B) the points pick up from the diagonal, potentially indicating the high level of dependency between these windows. The Tajima's D statistics (Figure 4.2A, C, E), fall primarily below the diagonal, then bowing slightly back up towards or over the diagonal. This indicates these samples are skewed slightly to the left compared to the empirical distributions or the distribution has a lighter tail on the right-side of the distribution. In contrast, the Fay and Wu's H statistics (Figure 4.2B, D, F), are primarily above the line, then bow slightly down or over the diagonal, indicating a possibly right skew.

To account for dependency and formally test the best possible window size, we apply the rank-based testing procedure (See Section 4.3.4) to each of the six window-size and statistic combinations. In each case, we expect to see a significant p-value ($\alpha = 0.05$), indicating departure from the null genome-wide empirical distribution. Tajima's D is highly underpowered at each window level, indicating the nominally significant value observed in the QQ-plots were inflated by dependent windows (Table 4.3, Figure 4.2). The tests for Fay and Wu's H, and in particular the 10 kb window size, appear to have the best power (Table 4.3). For this reason, we will show this window size for all future results.

Table 4.3 Rank-Based Test Results for Set of Positively Selected Genes

Window Size (kb)	Rank-Based Test P-Value	
	Tajima's D	Fay and Wu's H
100	0.9983	0.3394
10	0.9081	0.0529
1	0.8712	0.3197

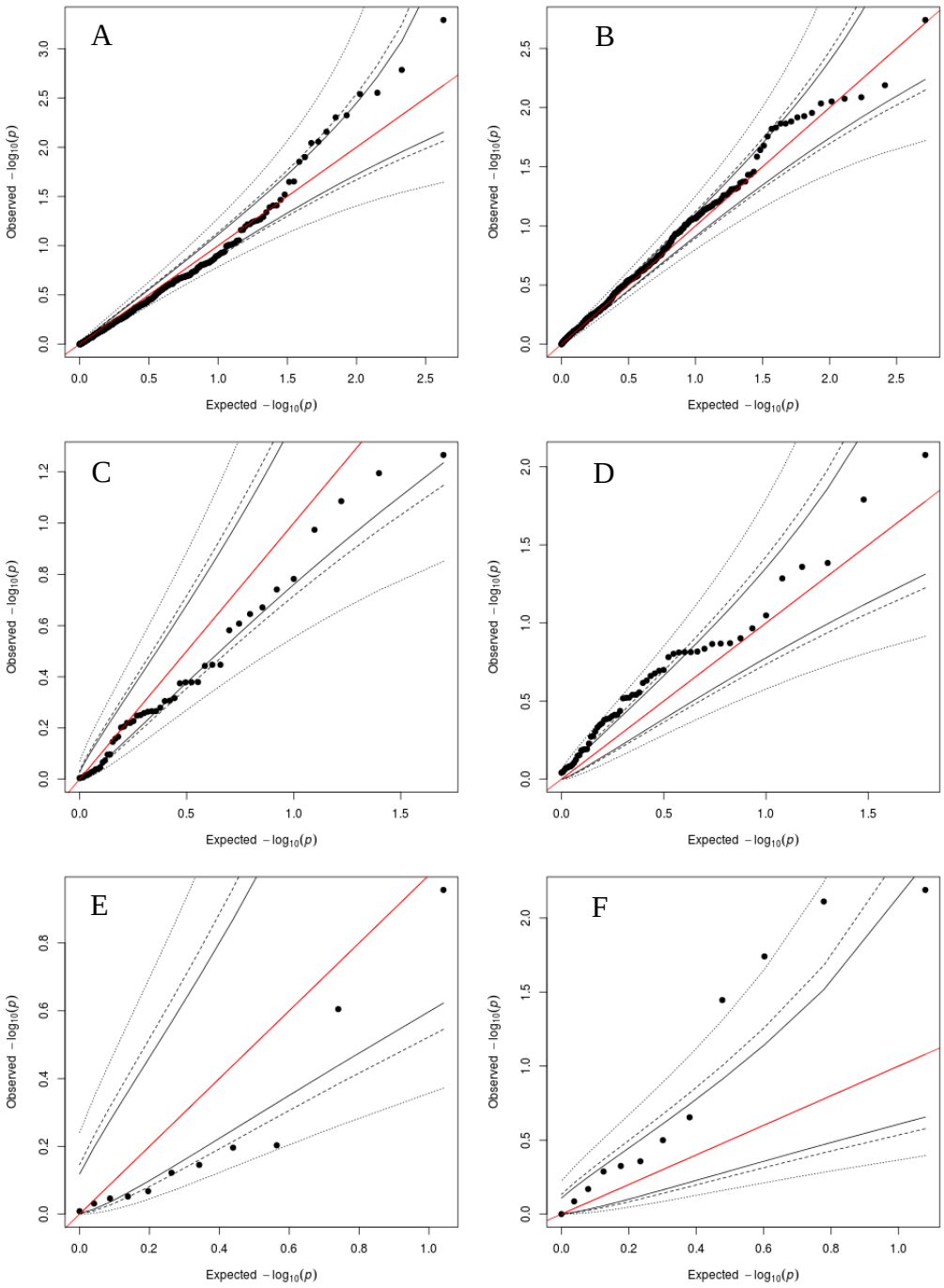

Figure 4.2 QQ-Plots of Empirical P-values Previously Identified Positively Selected Genes Tajima's D (left) and Fay and Wu's H (right) empirical p-values for windows of size 1 kb (A, B), 10 kb (C, D), and 100 kb (E, F). The red line shows equality between observed and expected$-\log_{10}(p)$. The black lines show quantiles under the expected distribution: 5% and 95% (solid), 2.5% and 97.5% (hashed), and Bonferroni-corrected 2.5% and 97.5%.

4.4.2 Comparing Genic and Non-genic Windows

To identify if selection signals in genic windows differ from those in non-genic windows, we calculate the empirical distributions for two different datasets: genic windows and all windows across the genome. We observe the summary statistics (mean, median, minimum, and maximum) are extremely similar between the distributions for both SFS statistics (Table 4.4). The maximum Tajima's D for the genome-wide collection of windows, 8.12, is slightly higher than the genic windows, 7.36. However, the centrality measures (mean and median) are separated by less than 0.01 for both statistics.

Table 4.4 Summary Statistics for Tajima's D Distribution in Genic Windows and All Windows (10 kb)

Statistic	Tajima's D		Fay and Wu's H	
	Genic Windows Only	All Windows	Genic Windows Only	All Windows
Maximum	7.36	8.12	1.52	1.52
Minimum	-2.67	-2.67	-15.9	-15.9
Mean	-1.83	-1.82	-0.63	-0.64
Median	-1.91	-1.90	-0.51	-0.52

To formally compare these distributions, we apply the Kolmogorov-Smirnov (K-S) test. Without any adjustments for dependency in the windows, the KS test for comparing Tajima's D distribution of genic windows to that of all windows appears highly significant (D=0.19, p-value= 3.1×10^{-29}). Similarly, this test for Fay and Wu's H statistic also returns a highly significant result (D=0.0088, p-value= 2.8×10^{-6}). This indicates there is a significant difference between the SFS statistics of genic windows vs. non-genic windows. However, in each of these distributions, there is a high level of dependency between windows, invalidating these tests. We assess the assumption of independence between windows by calculating the autocorrelation for neighboring windows with lag 1 to 50 (shown for 10 kb windows in chromosome 20 in Figure

4.3). For the Tajima's D statistics, the autocorrelation remains high (>0.10) until lag 24 (Figure 4.3). For the Fay and Wu's H statistics, we observe that, though dependency exists in the neighboring windows, they are less strongly affected by autocorrelation, due to the potential gaps in ancestral information needed for the statistics calculation (Figure 4.3B).

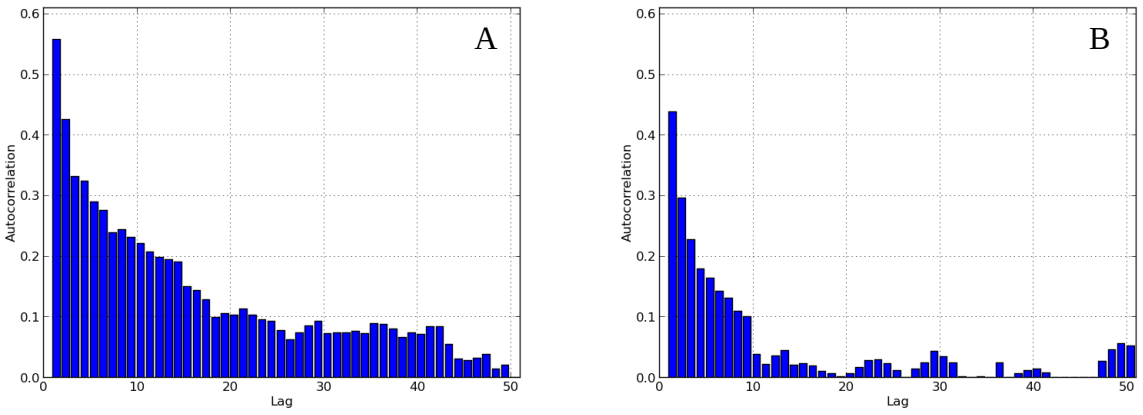

Figure 4.3 Autocorrelation of SFS Statistics in Genic Windows
We show the lag 1 to 50 autocorrelations of the (A) Tajima's D statistics and (B) Fay and Wu's H statistics in 10 kb genic windows across chromosome 22.

Because of this dependency, we apply the K-S test to random sample of 1,000 windows from each distribution. In the sample of Tajima's D statistic windows, the autocorrelation in the sample with lag-1 is eliminated (autocorrelation in genic windows=-0.028, all windows=0.052) and the Spearman's rho across the sample is highly non-significant (genic windows: ρ=-0.0058, p-value=0.86, all windows: ρ=-0.0024, p-value=0.094). In the sample of Fay and Wu's H statistic windows, we observe a similar reduction in correlation (autocorrelation in genic windows=-0.031, all windows=0.072) and Spearman's rho (genic windows: ρ=-0.055, p-value=0.082, all windows: ρ=-0.019, p-value=0.54). Applying the K-S test to the Tajima's D samples gives a significant result (D=0.069, p-value=0.016), indicating these two samples are not drawn from the same underlying distribution. In contrast, the Fay and Wu's H samples give a

strongly non-significant result (D=0.020, p-value=0.987). These tests indicate that for the Tajima's D statistics, there is a significant difference between the empirical distributions of the genic windows versus all genome-wide windows. For Fay and Wu's H, there is no evidence of this significant difference.

To determine if this lack of signal was due to smaller sample size, we repeated this procedure for larger sample sizes of 5,000 and 10,000. At a sample size of 5,000, we again obtain significant results for Tajima's D (D=0.0265, p-value=0.029) and non-significant results for Fay and Wu's H (D=0.024, p-value=0.11). Further increasing the sample size to 10,000, these results are again confirmed but the autocorrelation is no longer well-controlled (lag-1 AC > 0.12 across distributions).

4.4.3 SFS Statistic Distributions

To understand the overall behavior of the SFS statistics across the genome and specifically within the collection of shared autoimmune genes, we compare the probability density functions and cumulative probability distributions. The distribution of Tajima's D values across the genome is highly skewed to the right, reflecting a high number of negative values (Figure 4.4). This reflects our expectation for a population that has undergone exponential growth. The distribution of Tajima's D in autoimmune genes is also right-skewed, but has shorter upper tails, most likely a result of smaller sample size (Figure 4.4). The centrality measures for these distributions are very similar, with a strongly negative mean and median (Table 4.5). Similarly, the minimum values of the distributions are close (-2.60 for autoimmune genes, -2.67 for all windows) (Table 4.5). The difference in maximums is substantial (0.17 for autoimmune genes, 8.12 for all windows) (Table 4.5), reflecting the longer upper tail of the genome-wide distribution.

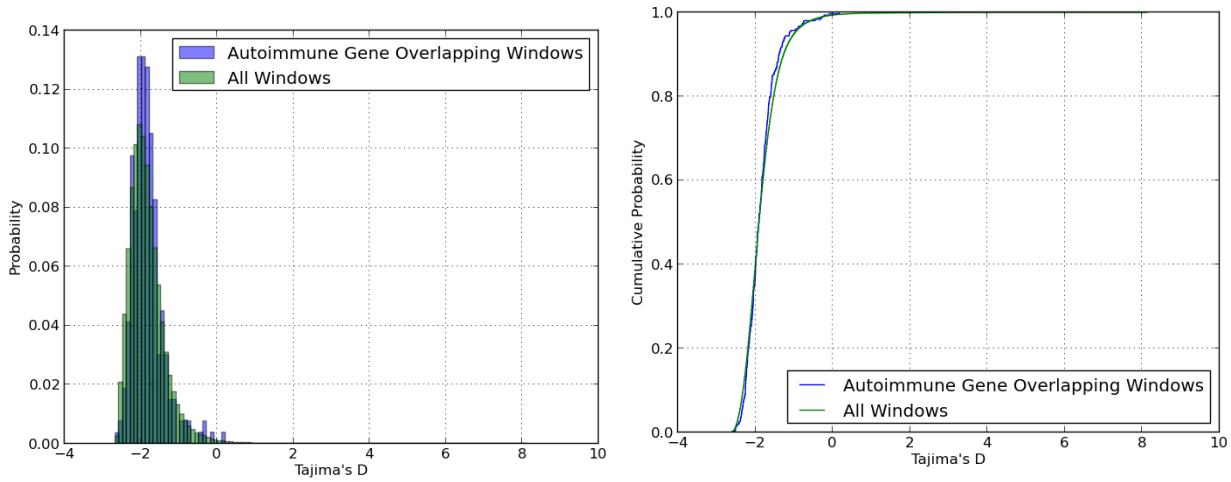

Figure 4.4 Probability distributions of Tajima's D in 10 kb Windows
We show the probability distribution (left) and cumulative probability distribution (right) of Tajima's D in all windows (green) and in windows overlapping autoimmune genes only (blue).

Table 4.5 Summary Statistics for Tajima's D Distributions (10 kb)

	Autoimmune Genes Windows	All Windows
Total Number of Windows	267	268431
Maximum	0.17	8.12
Minimum	-2.60	-2.67
Mean	-1.82	-1.82
Median	-1.88	-1.90

In contrast to Tajima's D, the distribution of Fay and Wu's H statistics across the genome does not have a right skewed but is characterized by a long left tail (Figure 4.5). The distribution in the autoimmune gene windows does not contain this long tail and the mass is shifted to the right (Figure 4.5). The measures of centrality are similar between the two distributions, though both the mean and median are slightly larger for the autoimmune gene windows (Table 4.6). The minimum Fay and Wu's H in the genome-wide distribution (-15.9) is much smaller than that of the autoimmune gene distribution (-2.68). The maximums also differ, with 1.52 in the genome-wide distribution and 0.56 in the autoimmune gene distributions (Table 4.6). The extreme values in the genome-wide distribution are likely reflective of misidentified ancestral alleles.

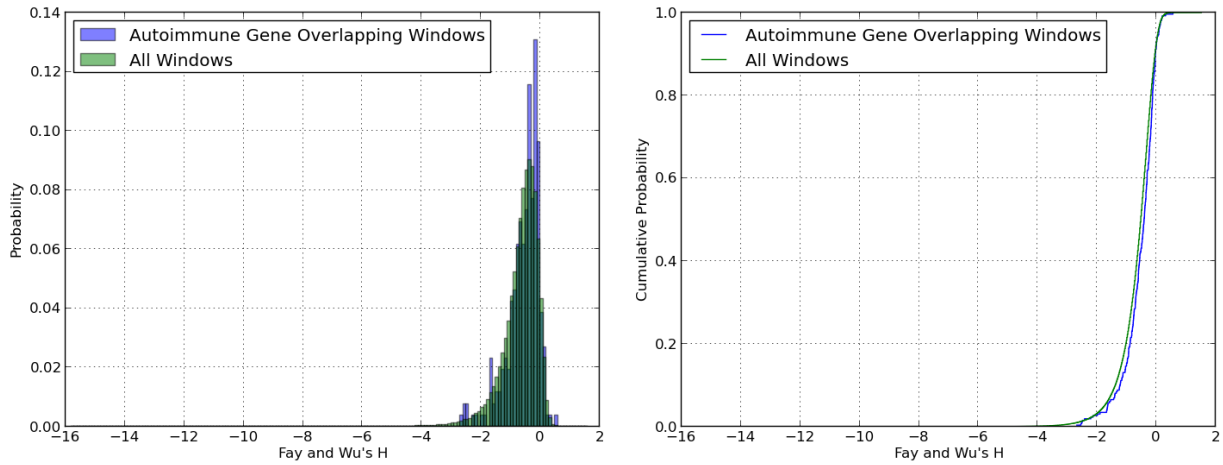

Figure 4.5 Probability distributions of Fay and Wu's H in 10 kb Windows
We show the probability distribution (top) and cumulative probability distribution (bottom) of Fay and Wu's H in all windows (green) and in windows overlapping autoimmune genes only (blue).

Table 4.6 Summary Statistics for Fay and Wu's H Distributions (10 kb)

	Autoimmune Genes Windows	All Windows
Total Number of Windows	292	263408
Maximum	0.56	1.52
Minimum	-2.68	-15.9
Mean	-0.52	-0.64
Median	-0.39	-0.52

4.4.4 Identifying Selection Signals in Autoimmune Genes

Using the genome-wide distribution of SFS statistics, we calculate the empirical p-values for each window in the autoimmune gene set. The QQ-plot for the Tajima's D statistic shows there are several points that reach nominal significance (hashed line) but all points are within the limits of the Bonferroni-corrected significance line (black dotted lines) (Figure 4.6). The top ten autoimmune gene windows, shown in Table 4.7, come from both tails of the distribution. Four of the top ten windows fall below the genome-wide mean of -1.82 with highly negative Tajima's D values. In the QQ-plot of autoimmune gene windows, the points fall primarily below the diagonal, then bowing slightly back over the diagonal. As in Figure 4.4, this indicates the distribution of autoimmune genes is skewed slightly to the left compared to the empirical

distributions and the distribution has a lighter tail on the right-side of the distribution. To formally test if selection signals exist in the collection of shared autoimmune genes, we apply a rank-based test to compare the distributions of the autoimmune gene windows and the genome-wide windows (Section 4.3.4). This test result was strongly non-significant (p-value= 0.9923), indicating there is not significant evidence that these distributions differ.

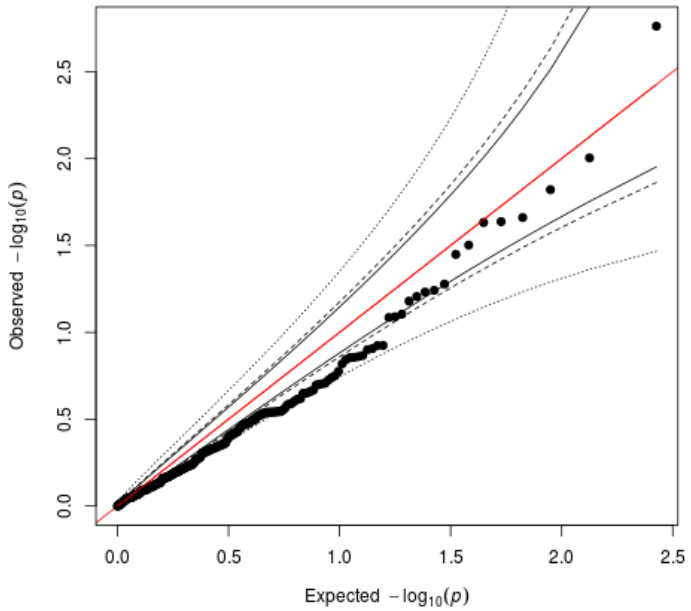

Figure 4.6 QQ-Plots of Empirical P-values from Tajima's D for Autoimmune Genes
We calculate Tajima's D empirical p-values for windows of size 10 kb. The red line shows equality between observed and expected $-\log_{10}(p)$. The black lines show quantiles under the expected distribution: 5% and 95% (solid), 2.5% and 97.5% (hashed), and Bonferroni-corrected 2.5% and 97.5%.

Table 4.7 Top Ten Windows by P-Value for Tajima's D Statistic in Autoimmune Gene Set

Gene	Tajima's D	P-Value
IL27RA	-2.60	0.0017
UBE2L3	0.17	0.0099
YDJC	-0.09	0.015
UBE2L3	-0.28	0.022
UBE2L3	-0.30	0.023
STAT4	-2.51	0.023
BACH2	-0.46	0.031
IL12A	-2.49	0.036
STAT4	-2.46	0.053
UBE2L3	-0.75	0.057

We repeat these procedures for the Fay and Wu's H statistics in each window. Like the Tajima's D windows, the QQ-plot for the Fay and Wu's H statistic shows there are a few points that reach nominal significance (hashed line) but all points are within the limits of the Bonferroni-corrected significance line (black dotted lines) (Figure 4.7). The points fall relatively tightly around the line expectation (red), indicating the autoimmune gene window distribution fits well with the empirical distribution (Figure 4.7). The top ten autoimmune gene windows, shown in Table 4.8, come from both tails of the distribution. Five of the top ten windows fall below the genome-wide mean of -0.64 with highly negative Fay and Wu's H values. Applying the rank-based testing procedure, we find a strongly non-significant (p-value= 0.3363), indicating there is not significant evidence that these distributions differ. Based on the results of both statistics, we do not observe significant evidence to reject the null hypothesis of neutral evolution at these genes.

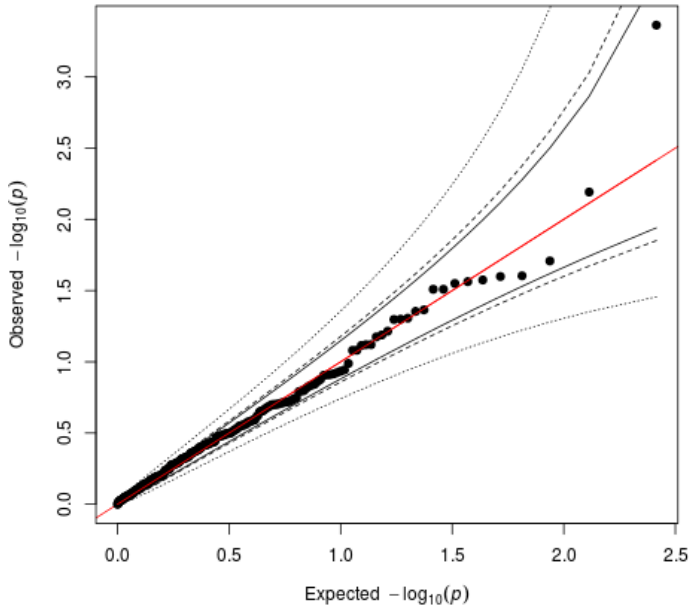

Figure 4.7 QQ-Plots of Empirical P-values from Fay and Wu's H for Autoimmune Genes
We calculate Fay and Wu's H empirical p-values for windows of size 10 kb. The red line shows equality between observed and expected $-\log_{10}(p)$. The black lines show quantiles under the expected distribution: 5% and 95% (solid), 2.5% and 97.5% (hashed), and Bonferroni-corrected 2.5% and 97.5%.

Table 4.8 Top Ten Windows by P-Value for Fay and Wu's H in Autoimmune Gene Set

Gene	Fay and Wu's H	P-Value
ZMIZ1	0.56	0.00043
UBE2L3	0.31	0.0064
CARD9	-2.68	0.020
TAGAP	0.20	0.025
ERAP2	-2.54	0.025
NFKB1	-2.51	0.027
NFKB1	-2.49	0.027
PTPN2	0.19	0.028
FCGR2A	-2.42	0.031
UBE2L3	0.18	0.031

4.4.5 Individual Autoimmune Diseases

We analyze gene sets for six autoimmune diseases independently: Celiac disease (CD), rheumatoid arthritis (RA), inflammatory bowel disease (IBD), systemic lupus erythematosus (SLE), Ankylosing spondylitis (AS), and type 1 diabetes (T1D). We show the QQ-Plots for each disease with 10 kb windows for both statistics, Tajima's D (Figure 4.8) and Fay and Wu's H (Figure 4.9). For Tajima's D, most the points remain within the Bonferroni-corrected significance level (dotted line). However, Celiac disease and SLE have an excess of points beyond this limit (Figure 4.8B, E). For Fay and Wu's H, the QQ-plots also show points staying within nominal significance levels (Figure 4.9). Again, the exception is SLE, with several points beyond the Bonferroni-corrected significance level (dotted line, Figure 4.9E). To formally test if selection signals exist in each collection of disease associated genes, we apply a rank-based test to compare the distributions to the genome-wide windows (Section 4.3.4). Apart from SLE, each of these tests results are strongly non-significant, indicating there is not significant evidence that these distributions differ from the genome-wide windows (Table 4.9). The test for Fay and Wu's H for SLE was nominally significant (p-value=0.047), providing weak evidence that these windows may differ from those genome-wide (Table 4.9).

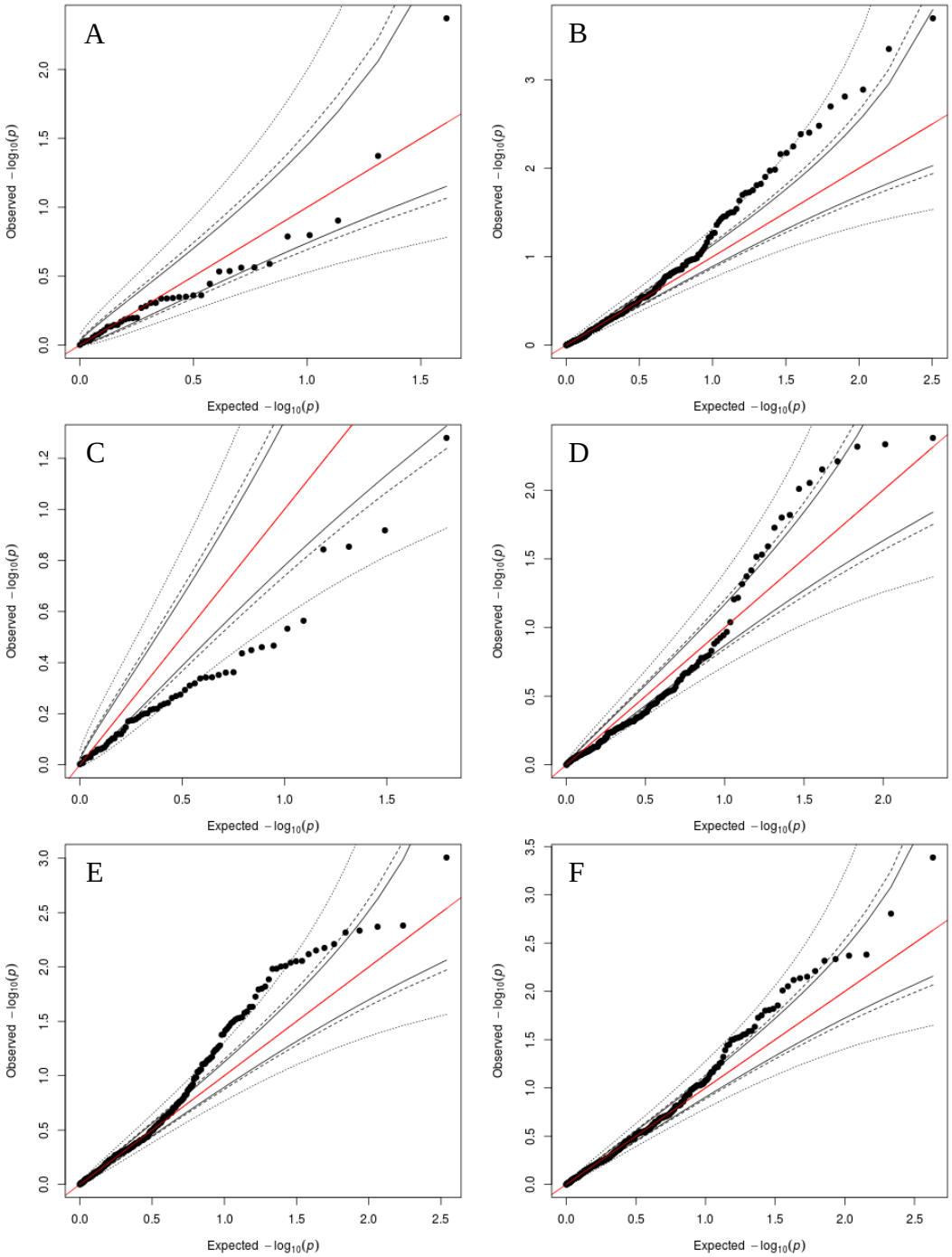

Figure 4.8 QQ-Plots of Empirical P-values from Tajima's D for Each Disease
We calculate Tajima's D empirical p-values for windows of size 10 kb for each disease set: AS (A), Celiac (B), IBD (C), RA (D), SLE (E), T1D (F). The red line shows equality between observed and expected $-\log_{10}(p)$. Black lines show quantiles under the expected distribution: 5% and 95% (solid), 2.5% and 97.5% (hashed), and Bonferroni-corrected 2.5% and 97.5%.

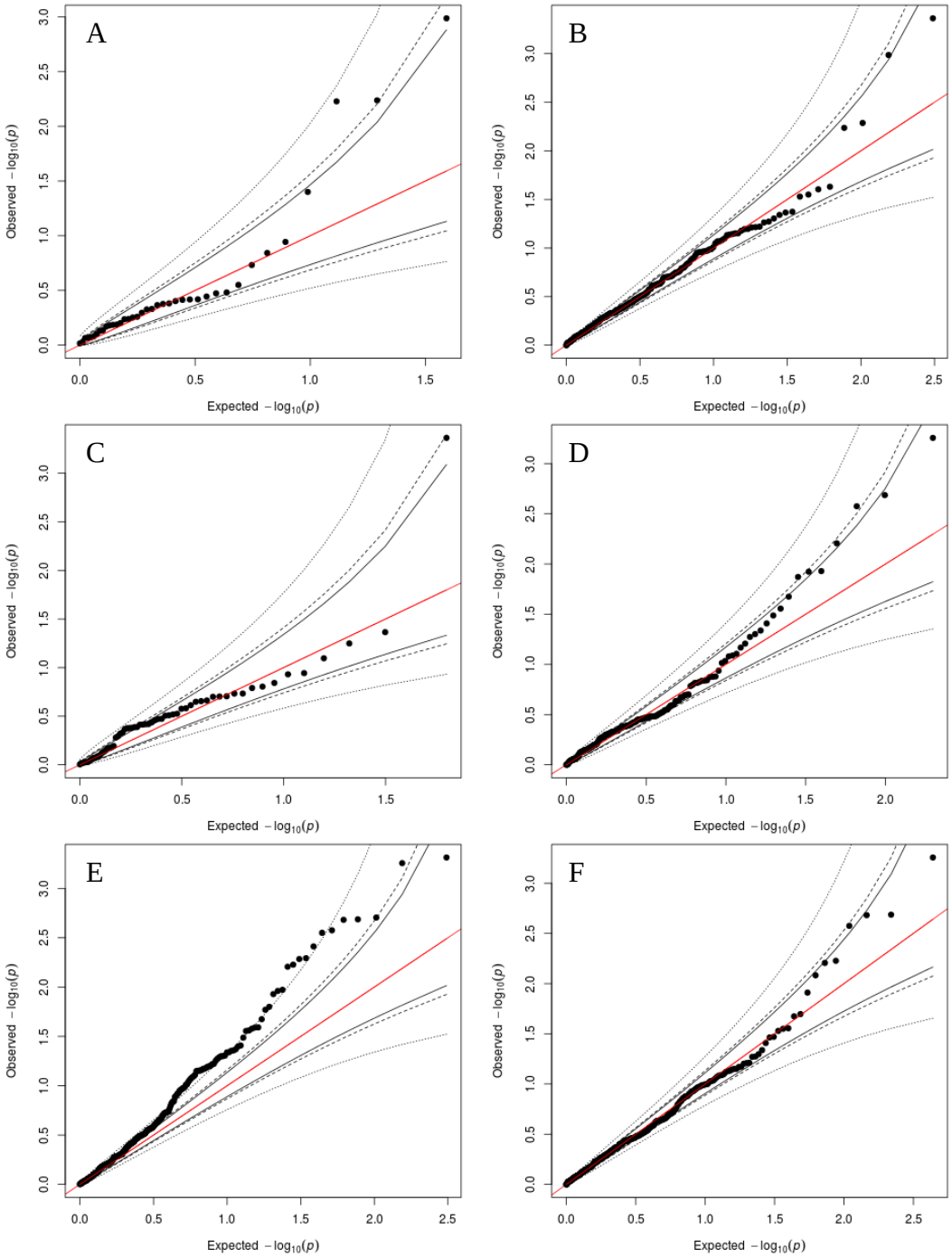

Figure 4.9 QQ-Plots of Empirical P-values of Fay and Wu's H for Each Disease
We calculate Fay and Wu's H empirical p-values for windows of size 10 kb for each disease set: AS (A), Celiac (B), IBD (C), RA (D), SLE (E), T1D (F). The red line shows equality between observed and expected $-\log_{10}(p)$. Black lines show quantiles under the expected distribution: 5% and 95% (solid), 2.5% and 97.5% (hashed), and Bonferroni-corrected 2.5% and 97.5%.

Table 4.9 Rank-Based Test Results for Each Disease Gene Set (10 kb windows)

Disease	Rank-Based Test P-Value	
	Tajima's D	Fay and Wu's H
AS	0.85	0.67
CD	0.56	0.28
IBD	0.99	0.34
RA	0.94	0.38
SLE	0.38	0.047
T1D	0.53	0.48

4.5 Conclusion

In this study, we design and conduct a comprehensive search using site frequency spectrum tests for signals of positive selection in shared non-HLA autoimmune disease-associated loci. With this specific class of tests, we do not find evidence that positive selection has been the driving force for the prevalence of autoimmunity associated loci. While certain genic windows show nominal evidence of selection, these windows do not pass multiple testing corrections. In particular, the Tajima's D and Fay and Wu's H statistics calculated in genomic windows overlapping these loci compared to those in random windows across the genome do not significantly differ. We also study windows overlapping individual autoimmune disease gene sets. These tests indicate there is not significant evidence that these distributions differ from the genome-wide windows.

In addition to the analysis of autoimmune genes, we compare the distributions of SFS statistics in protein coding genic windows to windows genome-wide. These tests indicate that there a significant difference between the Tajima's D statistics of the genic windows versus all genome-wide windows. This could indicate selection acts differently on protein-coding genes compared to non-protein coding genes. We note that we apply these tests to genic regions defined as protein-coding because it is immediately interpretable for functional response to external pathogens. The implications of different SFS statistics in genic, non-protein-coding

regions is less easily understood. When we expand this definition to all genic regions, including regions coding for RNA, we no longer obtain a significant result in any tests.

With the increasing availability of whole genome sequencing data, this study provides important insight into the use of SFS tests for this context. We develop a rank-based approach to compare distributions that accounts for dependency in neighboring windows. Applying the tests over a range of window sizes, we obtain the strongest power for identifying positive selection in 10 kb windows using Fay and Wu's H. We determine this approach for whole genome sequencing data to be highly underpowered for identifying gene sets that have undergone positive selection.

There are a few limitations in this study, identifying areas where further research is needed for applications to whole genome sequencing data. One limitation for this study is the reliance on previously identified positively selected genes to determine power and optimal window size. While each gene selected for power assessment was chosen carefully based on multiple existing studies, the number of undisputed positively selected genes is limited and our knowledge of the evolutionary history of many of these genes, such as the precise timing of selection pressures, is unknown. Furthermore, our dataset is made up entirely of European ancestry. Some positively selected genes are known to strongly show evidence of selection in non-European populations, such as FADS2 in Indian populations[207]. Comparing site frequency spectrum statistic distributions between different ancestry groups could yield further information on differing selection effects. Finally, in the case of Fay and Wu's H statistic, we are limited by the availability of ancestral data. While this includes approximately ninety-four percent of the full dataset, this restricts our ability to study certain areas of the genome.

For this study, we focus on identifying past positive selection based on the "hygiene hypothesis". An extension of this study could assess the power of this approach in identifying other non-neutral evolution histories, such as ongoing positive and negative selection. This approach could also be extended to other disease models or gene sets. For example, tumor suppression and apoptosis genes have shown signals of positive selection in comparisons of human and chimpanzee polymorphism data[208]. In addition, this test may useful in combination with computationally intensive non-SFS tests for selection such as the extended haplotype homozygosity test, long-range haplotype test, and singleton density score test. Our approach is computationally efficient to identify candidate regions based on nominal significance to apply such tests in further steps. Finally, these tests are confounded by different periods of selection across the variants in the gene set. Nakagome *et al* recently presented an approach using approximate Bayesian computation to estimate the time of selection on variants[209]. This study applied the approach to three different autoimmune risk alleles and identified three different probable epochs of selection (before out-of-Africa, after out-of-Africa, and after onset of agriculture). Future studies could refine our approach by focusing on variants expected to be selected within the same time period and identifying the time range where power is optimized. In this study, we present an important contribution to the field of population genetics by evaluating the adaptation of existing methods to developing technologies and identifying areas of future improvements.

CHAPTER 5: Discussion

The continually increasing availability of next generation sequencing (NGS) data has drastically changed classic approaches to population genetic analyses. The new data format provides both unique opportunities for methods development, as well as challenges in adapting previous approaches. New developments aim to exploit the wealth of information in rare variation made available by NGS, while differentiating from signals of noise. In this dissertation, we present a collection of population genetics methods, specifically tailored for next generation sequencing data and the signals residing in rare variants. This section will focus on the overarching impact of these methods, lessons learned in their developments, and adaptations in the field that will further improve their usage. We will also identify possible future extensions and applications for each method.

In Chapter 2, we present a novel method for estimating changing migration rates between populations. This project relies on the intuition of shared variation between populations reflecting historical interactions between populations. Rare variation carry two crucial pieces of information for which common variants do not have proper resolution: the relative time that the variant arose based on its rareness in the sample and the population or populations in which it resides. Using these two pieces of information together, we can create a temporal picture of migration history. This method emphasizes flexibility, as we show our approach is robust to misspecifications of several kinds (effective population size, ancestral migration, and imbalanced migration) and allows for adjustments for exponential population growth.

We find this method's power to detect change and precision to identify specific parameters relies on the extent of information available on rare variation. In the context of sequenced target gene data presented here, we can identify changing migration within the approximate range of 20 to 400 generations. We see that this power and precision improves with whole exome sequencing data, providing more information on rare variation. Therefore, we can expect this range can be extended further with a focus on further increasing sample sizes in diverse populations for whole genome sequencing studies. As these studies grow, more unique questions on population interactions can be approached.

This method also highlights the utility of coalescent simulation. We rely on a grid search to identify parameters of changing migration, with grid parameters based on repeated coalescent simulations. Therefore, detecting minute migration signatures in variation is heavily dependent on well-controlled Monte Carlo error, which decreased with increasing numbers of simulations. Improving speed, access, available options, and relative ease of coalescent simulators will continue to improve the applicability of this approach. Since the conception of this project, several new coalescent simulators have been developed, such as FTEC[210] and fastcoalsim[87]. This approach may be refined based on the new parameters allowed under these coalescent simulations and the growing availability of parallel computing.

Our method for detecting changing migration rates focuses on counts of individual rare variants for evidence of gene flow between and within populations. A possible extension for this method could incorporate information from local patterns of genetic diversity, indicating stretches of the genome rather than individual variants that are shared between individuals. For instance, several existing methods including Browning and Browning and Gusev *et al* use long-range shared haplotypes between and within populations to identify past demographic events[211;

[212]. Furthermore, the Pairwise Sequential Markovian Coalescent (PSMC) and, later, the Multiple Sequential Markovian Coalescent (MSMC) use the local density of heterozygous sites across the genome to identify regions of constant TMRCA separated by historical recombination events [83; 84]. Incorporating local genetic information into our model could further refine our parameter estimation and improve power. The drawback of this addition is the increased computational burden required for phasing individuals, making the approach less accommodating to the increasing sample sizes of genetic data. Therefore, future work will need to investigate compromises between computational efficiency and increased information from shared local genetic variation.

There are several other possible extensions for the migration method presented in Chapter 2. First, the method relies on the number of variants with a particular minor allele count and how these variants are distributed between populations. However, the method does not incorporate potential uncertainty in this minor allele count. Allowing for uncertainty could improve migration estimates, particularly at higher minor allele counts where observations are fewer and spread across many possible configurations between populations. In this work, we focus on detecting and identifying the simplest models of increasing or decreasing migration rates. More complicated patterns are possible, such as U-shapes of alternating high and low migration rates and models involving greater than two populations. Future studies could investigate identifying these different patterns of migration rates and the data sizes or adjustments to the method they require.

The methods in this dissertation stress the relevance of population genetics in medical and biological studies. Because this field focuses on the large-scale view of populations, the immediate medical impact is often overlooked and underemphasized. In Chapter 3, we show two

examples where properly modeling population bottlenecks and genetic drift are important for predicting and understanding disease. In the first application, we model mtDNA transmission from mothers to offspring and estimate the size and nature of the bottleneck in the process. Mitochondria are considered the powerhouse of the cell and mtDNA mutations can be highly detrimental to the health of an individual. Most mutations that cause diseases due to defects in mitochondrial function exist as heteroplasmies (intra-individual variation) and only cause disease symptoms when the frequency of the mutant allele exceeds a particular threshold [118]. Therefore, accurately modeling the transmission of mtDNA and the expected frequency of pathogenic mutations is particularly relevant to the medical community and the disease risk of the offspring. Second, we apply this approach to somatic mutations in individuals with pre-mature aging disorders. We show our mathematical model can be adjusted to discern between effects of genetic drift and selection. Using this model to identify selection signals among mutations provides insight into their functional impact, potentially improving understanding for future medical treatments. Specifically, we identify three variants with evidence of positive selection for further functional investigation.

An additional strength of the approaches discussed in this dissertation is their emphasis on flexibility to experimental design. In Chapter 3 specifically we present a mathematical model that is adapted for two different experimental situations. We can adjust parameters within the model to account for known parameters, such as bottleneck size, as well as incorporate sources of error from sequencing and sampling procedures. As this project developed, we could build on the most basic model to include these details.

There are several possible extensions and future applications for the bottleneck and genetic drift model proposed in Chapter 3. First, in the current model, we model each variant

independently, ignoring multiple variants with dramatic allele frequency shifts in the same individual. Future extensions could investigate incorporating linkage disequilibrium across variants. This additional information will improve our understanding of passenger variants where we observe evidence of selection and refine our estimates of population bottleneck sizes. We also currently focus on two specific populations but this model can be more broadly applied. This approach could be particularly useful in conservation biology where subsets of populations are used to grow new populations[213; 214]. In this context, our model can create probability distributions of final allele frequencies, given the bottleneck population size and composition, to minimize inbreeding and maintain genetic diversity. In Chapter 3, we present a flexible model for population bottlenecks and genetic drift, allowing for the development of several future adjustments.

In Chapter 4, we develop an approach to detecting selection signals in whole genome sequencing data, focusing on genes associated with multiple autoimmune disorders. We aim to assess the 'hygiene hypothesis', which suggests that autoimmune disease associated loci could be maintained in the population because they were previously necessary to offer protection from infectious diseases or foreign pathogens[160]. We focus on loci shared across autoimmune diseases specifically, as they may contribute to a broader, shared immune response and implicate common pathways under selection.

In addition to another instance of population genetics with medical and biological relevance, this chapter shows the need for assessing classical approaches in new data types. We adapt existing site frequency spectrum (SFS) tests to detect positive selection in whole genome sequencing (WGS) data. There are many advantages of using WGS for SFS tests over genome-wide association studies, such as eliminating ascertainment bias and allowing access to small

regions of linkage disequilibrium. Furthermore, using WGS, we can use an empirical distribution for significance testing of SFS statistics and to account for confounding effects of population history. Applying the approach to previously identified signals of selection, we find the approach is not well-powered to detect the selection signals of interest. While disappointing in its lack of statistically significant results, this observation is in itself an important contribution to the field. We identify tests that are not immediately applicable to newly emerging data and emphasize the impact of noise and dependency among observations in analyzing whole genome sequencing data. We also learn about SFS statistics across the genome, showing the severe non-normality and skewness in the distributions genome-wide and identifying differences in the SFS statistic distributions of genic and non-genic regions. We suggest further research into the power of adapting existing approaches for detecting selection with WGS data, such as the integrated haplotype scores, extended haplotype homozygosity, and singleton density scores.

The methods discussed in Chapter 4 could also be extended by studying correlations between the SFS test statistics. Zeng *et al* previously investigated the sensitivity of site frequency spectrum tests under different models of selection, showing Tajima's *D* and Fay and Wu's *H* in combination can be most effective in detecting positive selection while being insensitive to other demographic effects[175]. Future work is needed to study these tests in combination for different models of selection within the context of whole genome sequencing data. In chapter 4 we also focus specifically on genes previously identified as associated with autoimmune disease loci based on the hygiene hypothesis. This approach could be extended to identify selection in any collection of disease genes to assess particular hypotheses. Gene sets of interest may include cancer-susceptibility genes or Neanderthal genes maintained in the present-day human

population. The adaptation and evaluation of site frequency spectrum tests for whole genome sequencing presented in Chapter 4 encourage many future analyses.

Increasingly large sequencing studies provide a new frontier for methods development and specifically, population genetics, with access to large sources of rare variation. With innovative statistical methods, population genetics studies can help to answer a wide spectrum of genetic, medical, and biological questions. In this dissertation work, we present several new methods and applications for population genetics on next generation sequencing data. We therefore provide several important contributions to approaching genetic questions and understanding the underlying genetic bases of disease.

BIBLIOGRAPHY

1. Fisher, R.A. (1930). The genetical theory of natural selection: a complete variorum edition.(Oxford University Press).
2. Zouros, E. (1979). Mutation Rates, Population Sizes and Amounts of Electrophoretic Variation of Enzyme Loci in Natural Populations. Genetics 92, 623-646.
3. Lewontin, R.C., and Hubby, J.L. (1966). A MOLECULAR APPROACH TO THE STUDY OF GENIC HETEROZYGOSITY IN NATURAL POPULATIONS. II. AMOUNT OF VARIATION AND DEGREE OF HETEROZYGOSITY IN NATURAL POPULATIONS OF DROSOPHILA PSEUDOOBSCURA. Genetics 54, 595-609.
4. Harris, H., Hopkinson, D.A., and Luffman, J. (1968). ENZYME DIVERSITY IN HUMAN POPULATIONS. Annals of the New York Academy of Sciences 151, 232-242.
5. Harris, H. (1966). Enzyme polymorphisms in man. Proceedings of the Royal Society of London Series B, Biological sciences 164, 298-310.
6. Kreitman, M. (1983). Nucleotide polymorphism at the alcohol dehydrogenase locus of Drosophila melanogaster. Nature 304, 412-417.
7. Sawyer, S.A., Dykhuizen, D.E., and Hartl, D.L. (1987). Confidence interval for the number of selectively neutral amino acid polymorphisms. Proceedings of the National Academy of Sciences of the United States of America 84, 6225-6228.
8. Whittam, T.S., Clark, A.G., Stoneking, M., Cann, R.L., and Wilson, A.C. (1986). Allelic variation in human mitochondrial genes based on patterns of restriction site polymorphism. Proceedings of the National Academy of Sciences of the United States of America 83, 9611-9615.
9. Pool, J.E., Hellmann, I., Jensen, J.D., and Nielsen, R. (2010). Population genetic inference from genomic sequence variation. Genome research 20, 291-300.
10. Nelson, M.R., Wegmann, D., Ehm, M.G., Kessner, D., St Jean, P., Verzilli, C., Shen, J., Tang, Z., Bacanu, S.A., Fraser, D., et al. (2012). An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science 337, 100-104.
11. Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., et al. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 337, 64-69.
12. Keinan, A., and Clark, A.G. (2012). Recent explosive human population growth has resulted in an excess of rare genetic variants. Science 336, 740-743.
13. Gravel, S., Henn, B.M., Gutenkunst, R.N., Indap, A.R., Marth, G.T., Clark, A.G., Yu, F., Gibbs, R.A., Bustamante, C.D., and Project, G. (2011). Demographic history and rare allele sharing among human populations. Proc Natl Acad Sci U S A 108, 11983-11988.

14. Kittles, R.A., Chen, W., Panguluri, R.K., Ahaghotu, C., Jackson, A., Adebamowo, C.A., Griffin, R., Williams, T., Ukoli, F., Adams-Campbell, L., et al. (2002). CYP3A4-V and prostate cancer in African Americans: causal or confounding association because of population stratification? Human genetics 110, 553-560.
15. Rosenberg, N.A., and Nordborg, M. (2006). A General Population-Genetic Model for the Production by Population Structure of Spurious Genotype–Phenotype Associations in Discrete, Admixed or Spatially Distributed Populations. Genetics 173, 1665-1678.
16. Clayton, D.G., Walker, N.M., Smyth, D.J., Pask, R., Cooper, J.D., Maier, L.M., Smink, L.J., Lam, A.C., Ovington, N.R., Stevens, H.E., et al. (2005). Population structure, differential bias and genomic control in a large-scale, case-control association study. Nature genetics 37, 1243-1246.
17. Campbell, C.D., Ogburn, E.L., Lunetta, K.L., Lyon, H.N., Freedman, M.L., Groop, L.C., Altshuler, D., Ardlie, K.G., and Hirschhorn, J.N. (2005). Demonstrating stratification in a European American population. Nature genetics 37, 868-872.
18. Wang, C., Zhan, X., Bragg-Gresham, J., Kang, H.M., Stambolian, D., Chew, E.Y., Branham, K.E., Heckenlively, J., The, F.S., Fulton, R., et al. (2014). Ancestry Estimation and Control of Population Stratification for Sequence-based Association Studies. Nature genetics 46, 409-415.
19. Shama, L.N., Kubow, K.B., Jokela, J., and Robinson, C.T. (2011). Bottlenecks drive temporal and spatial genetic changes in alpine caddisfly metapopulations. BMC evolutionary biology 11, 278.
20. Morjan, C.L., and Rieseberg, L.H. (2004). How species evolve collectively: implications of gene flow and selection for the spread of advantageous alleles. Molecular ecology 13, 1341-1356.
21. Hitchings, S.P., and Beebee, T.J.C. (1997). Genetic substructuring as a result of barriers to gene flow in urban Rana temporaria (common frog) populations: implications for biodiversity conservation. Heredity 79, 117-127.
22. Su, H., Qu, L.J., He, K., Zhang, Z., Wang, J., Chen, Z., and Gu, H. (2003). The Great Wall of China: a physical barrier to gene flow? Heredity 90, 212-219.
23. Le Corre, V., Dumolin-Lapègue, S., and Kremer, A. (1997). Genetic variation at allozyme and RAPD loci in sessile oak Quercus petraea (Matt.) Liebl.: the role of history and geography. Molecular ecology 6, 519-529.
24. Sá-Pinto, A., Branco, M.S., Alexandrino, P.B., Fontaine, M.C., and Baird, S.J.E. (2012). Barriers to Gene Flow in the Marine Environment: Insights from Two Common Intertidal Limpet Species of the Atlantic and Mediterranean. PloS one 7, e50330.
25. Gayden, T., Cadenas, A.M., Regueiro, M., Singh, N.B., Zhivotovsky, L.A., Underhill, P.A., Cavalli-Sforza, L.L., and Herrera, R.J. (2007). The Himalayas as a Directional Barrier to Gene Flow. American journal of human genetics 80, 884-894.
26. Bosch, E., Calafell, F., Comas, D., Oefner, P.J., Underhill, P.A., and Bertranpetit, J. (2001). High-Resolution Analysis of Human Y-Chromosome Variation Shows a Sharp Discontinuity and Limited Gene Flow between Northwestern Africa and the Iberian Peninsula. The American Journal of Human Genetics 68, 1019-1029.
27. Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., Brandt, G., Nordenfelt, S., Harney, E., Stewardson, K., et al. (2015). Massive migration from the steppe is a source for Indo-European languages in Europe.

28. Allentoft, M.E., Sikora, M., Sjogren, K.-G., Rasmussen, S., Rasmussen, M., Stenderup, J., Damgaard, P.B., Schroeder, H., Ahlstrom, T., Vinner, L., et al. (2015). Population genomics of Bronze Age Eurasia. Nature 522, 167-172.
29. Hein, J., Schierup, M.H., and Wiuf, C. (2005). Gene genealogies, variation and evolution : a primer in coalescent theory.(Oxford ; New York: Oxford University Press).
30. Pujolar, J.M., Vincenzi, S., Zane, L., Jesensek, D., De Leo, G.A., and Crivelli, A.J. (2011). The effect of recurrent floods on genetic composition of marble trout populations. PLoS One 6, e23822.
31. Mkize, L.S., Mukaratirwa, S., and Zishiri, O.T. (2016). Population genetic structure of the freshwater snail, Bulinus globosus, (Gastropoda: Planorbidae) from selected habitats of KwaZulu-Natal, South Africa. Acta Trop.
32. Premoli, A.C., and Kitzberger, T. (2005). Regeneration mode affects spatial genetic structure of Nothofagus dombeyi forests. Mol Ecol 14, 2319-2329.
33. Rahimi-Mianji, G., Nejati-Javaremi, A., and Farhadi, A. (2015). GENETIC DIVERSITY, PARENTAGE VERIFICATION AND GENETIC BOTTLENECKS EVALUATION IN IRANIAN TURKMEN HORSE BREED. Genetika 51, 1066-1074.
34. Marsden, C.D., Ortega-Del Vecchyo, D., O'Brien, D.P., Taylor, J.F., Ramirez, O., Vilà, C., Marques-Bonet, T., Schnabel, R.D., Wayne, R.K., and Lohmueller, K.E. (2016). Bottlenecks and selective sweeps during domestication have increased deleterious genetic variation in dogs. Proc Natl Acad Sci U S A 113, 152-157.
35. Schubert, M., Jónsson, H., Chang, D., Der Sarkissian, C., Ermini, L., Ginolhac, A., Albrechtsen, A., Dupanloup, I., Foucal, A., Petersen, B., et al. (2014). Prehistoric genomes reveal the genetic foundation and cost of horse domestication. Proc Natl Acad Sci U S A 111, E5661-5669.
36. Oliveira-Jr, P.R., Costa, M.C., Silveira, L.F., and Francisco, M.R. (2016). Genetic guidelines for captive breeding and reintroductions of the endangered Black-fronted Piping Guan, Aburria jacutinga (galliformes, cracidae), an Atlantic Forest endemic. Zoo Biol.
37. Servanty, S., Converse, S.J., and Bailey, L.L. (2014). Demography of a reintroduced population: moving toward management models for an endangered species, the Whooping Crane. Ecol Appl 24, 927-937.
38. Forstmeier, W., Segelbacher, G., Mueller, J.C., and Kempenaers, B. (2007). Genetic variation and differentiation in captive and wild zebra finches (Taeniopygia guttata). Mol Ecol 16, 4039-4050.
39. Bryant, J.V., Gottelli, D., Zeng, X., Hong, X., Chan, B.P., Fellowes, J.R., Zhang, Y., Luo, J., Durrant, C., Geissmann, T., et al. (2016). Assessing current genetic status of the Hainan gibbon using historical and demographic baselines: implications for conservation management of species of extreme rarity. Mol Ecol.
40. Osborne, A.J., Negro, S.S., Chilvers, B.L., Robertson, B.C., Kennedy, M.A., and Gemmell, N.J. (2016). Genetic Evidence of a Population Bottleneck and Inbreeding in the Endangered New Zealand Sea Lion, Phocarctos hookeri. J Hered.
41. González-Tortuero, E., Rusek, J., Maayan, I., Petrusek, A., Piálek, L., Laurent, S., and Wolinska, J. (2016). Genetic diversity of two Daphnia-infecting microsporidian parasites, based on sequence variation in the internal transcribed spacer region. Parasit Vectors 9, 293.
42. Couce, A., Rodríguez-Rojas, A., and Blázquez, J. (2016). Determinants of Genetic Diversity of Spontaneous Drug-Resistance in Bacteria. Genetics.

43. Bull, R.A., Luciani, F., McElroy, K., Gaudieri, S., Pham, S.T., Chopra, A., Cameron, B., Maher, L., Dore, G.J., White, P.A., et al. (2011). Sequential bottlenecks drive viral evolution in early acute hepatitis C virus infection. PLoS Pathog 7, e1002243.
44. Boycott, K.M., Parboosingh, J.S., Chodirker, B.N., Lowry, R.B., McLeod, D.R., Morris, J., Greenberg, C.R., Chudley, A.E., Bernier, F.P., Midgley, J., et al. (2008). Clinical genetics and the Hutterite population: a review of Mendelian disorders. Am J Med Genet A 146A, 1088-1098.
45. Smith, D.C., Atadzhanov, M., Mwaba, M., and Greenberg, L.J. (2015). Evidence for a common founder effect amongst South African and Zambian individuals with Spinocerebellar ataxia type 7. J Neurol Sci 354, 75-78.
46. Rees, E.E., Pond, B.A., Cullingham, C.I., Tinline, R.R., Ball, D., Kyle, C.J., and White, B.N. (2009). Landscape modelling spatial bottlenecks: implications for raccoon rabies disease spread. Biol Lett 5, 387-390.
47. Biek, R. (2007). Evolutionary dynamics and spatial genetic structure of epizootic hemorrhagic disease virus in the eastern United States. Infect Genet Evol 7, 651-655.
48. Shoda-Kagaya, E. (2007). Genetic differentiation of the pine wilt disease vector Monochamus alternatus (Coleoptera: Cerambycidae) over a mountain range - revealed from microsatellite DNA markers. Bull Entomol Res 97, 167-174.
49. Hatzikotoulas, K., Gilly, A., and Zeggini, E. (2014). Using population isolates in genetic association studies. Brief Funct Genomics 13, 371-377.
50. Levy-Lahad, E., Catane, R., Eisenberg, S., Kaufman, B., Hornreich, G., Lishinsky, E., Shohat, M., Weber, B.L., Beller, U., Lahad, A., et al. (1997). Founder BRCA1 and BRCA2 mutations in Ashkenazi Jews in Israel: frequency and differential penetrance in ovarian cancer and in breast-ovarian cancer families. Am J Hum Genet 60, 1059-1067.
51. Moran, P.A.P. (1962). The statistical processes of evolutionary theory.(Oxford,: Clarendon Press).
52. Wright, S. (1990). Evolution in Mendelian populations. 1931. Bull Math Biol 52, 241-295; discussion 201-247.
53. Cooper, G.S., Bynum, M.L.K., and Somers, E.C. (2009). Recent Insights in the Epidemiology of Autoimmune Diseases: Improved Prevalence Estimates and Understanding of Clustering of Diseases. J Autoimmun 33, 197-207.
54. Brinkworth, J.F., and Barreiro, L.B. (2014). The contribution of natural selection to present-day susceptibility to chronic inflammatory and autoimmune disease. Curr Opin Immunol 31, 66-78.
55. Wright, S. (1951). The genetical structure of populations. Annals of eugenics 15, 323-354.
56. Michalakis, Y., and Excoffier, L. (1996). A generic estimation of population subdivision using distances between alleles with special reference for microsatellite loci. Genetics 142, 1061-1064.
57. Beerli, P. (1998). Estimation of migration rates and population sizes in geographically structured populations. NATO ASI SERIES A LIFE SCIENCES 306, 39-54.
58. Weir, B.S. (1996). Genetic data analysis II: methods for discrete population genetic data.(Sunderland, Mass.: Sinauer Associates).
59. Hudson, R.R., Slatkin, M., and Maddison, W.P. (1992). Estimation of Levels of Gene Flow from DNA Sequence Data. Genetics 132, 583-589.
60. Lynch, M., and Crease, T.J. (1990). The analysis of population survey data on DNA sequence variation. Molecular Biology and Evolution 7, 377-394.

61. Excoffier, L., Smouse, P.E., and Quattro, J.M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131, 479-491.
62. Rousset, F. (1996). Equilibrium Values of Measures of Population Subdivision for Stepwise Mutation Processes. Genetics 142, 1357-1362.
63. Nei, M. (1973). Analysis of gene diversity in subdivided populations. Proc Natl Acad Sci U S A 70, 3321-3323.
64. Oconnell, N., and Slatkin, M. (1993). High Mutation Rate Loci in a Subdivided Population. Theoretical Population Biology 44, 110-127.
65. Weir, B.S., and Cockerham, C.C. (1984). Estimating F-Statistics for the Analysis of Population Structure. Evolution; international journal of organic evolution 38, 1358-1370.
66. Kingman, J.F.C. (1982). The coalescent. Stochastic Processes and their Applications 13, 235-248.
67. Pearse, D.E., and Crandall, K.A. (2004). Beyond F(ST): Analysis of population genetic data for conservation. Conserv Genet 5, 585-602.
68. Wakeley, J. (1998). Segregating sites in Wright's island model. Theoretical population biology 53, 166-174.
69. Beerli, P., and Felsenstein, J. (1999). Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. Genetics 152, 763-773.
70. Slatkin, M., and Maddison, W.P. (1989). A cladistic measure of gene flow inferred from the phylogenies of alleles. Genetics 123, 603-613.
71. Beerli, P. (1997). Analysis of geographically structured populations: Estimators based on coalescence. In. (University of Washington.
72. Li, S., and Jakobsson, M. (2012). Estimating demographic parameters from large-scale population genomic data using Approximate Bayesian Computation. BMC genetics 13, 22.
73. Gao, H., Williamson, S., and Bustamante, C.D. (2007). A Markov chain Monte Carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype data. Genetics 176, 1635-1651.
74. Novembre, J., and Slatkin, M. (2009). Likelihood-based inference in isolation-by-distance models using the spatial distribution of low-frequency alleles. Evolution; international journal of organic evolution 63, 2914-2925.
75. Hey, J., and Nielsen, R. (2007). Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. Proceedings of the National Academy of Sciences of the United States of America 104, 2785-2790.
76. De Iorio, M., Griffiths, R.C., Leblois, R., and Rousset, F. (2005). Stepwise mutation likelihood computation by sequential importance sampling in subdivided population models. Theoretical population biology 68, 41-53.
77. Nielsen, R., and Wakeley, J. (2001). Distinguishing migration from isolation: a Markov chain Monte Carlo approach. Genetics 158, 885-896.
78. Beerli, P., and Felsenstein, J. (2001). Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. Proceedings of the National Academy of Sciences of the United States of America 98, 4563-4568.

79. Tufto, J., Engen, S., and Hindar, K. (1996). Inferring patterns of migration from gene frequencies under equilibrium conditions. Genetics 144, 1911-1921.
80. Rannala, B., and Hartigan, J.A. (1996). Estimating gene flow in island populations. Genet Res 67, 147-158.
81. Wright, S. (1931). Evolution in Mendelian Populations. Genetics 16, 97-159.
82. Slatkin, M. (1987). Gene flow and the geographic structure of natural populations. Science 236, 787-792.
83. Li, H., and Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. Nature 475, 493-496.
84. Schiffels, S., and Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences.
85. Gutenkunst, R.N., Hernandez, R.D., Williamson, S.H., and Bustamante, C.D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet 5, e1000695.
86. Schraiber, J.G., and Akey, J.M. (2015). Methods and models for unravelling human evolutionary history. Nat Rev Genet 16, 727-740.
87. Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V.C., and Foll, M. (2013). Robust demographic inference from genomic and SNP data. PLoS Genet 9, e1003905.
88. Mila, B., Girman, D.J., Kimura, M., and Smith, T.B. (2000). Genetic evidence for the effect of a postglacial population expansion on the phylogeography of a North American songbird. Proceedings Biological sciences / The Royal Society 267, 1033-1040.
89. Guiher, T.J., and Burbrink, F.T. (2008). Demographic and phylogeographic histories of two venomous North American snakes of the genus Agkistrodon. Molecular phylogenetics and evolution 48, 543-553.
90. O'Connor, T.D., Fu, W., Project, N.G.E.S., Genetics, E.S.P.P., Statistical Analysis Working Group, E.T., Mychaleckyj, J.C., Logsdon, B., Auer, P., Carlson, C.S., Leal, S.M., et al. (2015). Rare variation facilitates inferences of fine-scale population structure in humans. Molecular biology and evolution 32, 653-660.
91. Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., et al. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 337, 64-69.
92. Gravel, S., Henn, B.M., Gutenkunst, R.N., Indap, A.R., Marth, G.T., Clark, A.G., Yu, F., Gibbs, R.A., Genomes, P., and Bustamante, C.D. (2011). Demographic history and rare allele sharing among human populations. Proceedings of the National Academy of Sciences of the United States of America 108, 11983-11988.
93. Hein, J., Schierup, M.H., and Wiuf, C. (2005). Gene genealogies, variation and evolution: a primer in coalescent theory.(Oxford ; New York: Oxford University Press).
94. Casella, G., and Berger, R.L. (2002). Statistical Inference.(Thomson Learning).
95. Arnold, B.C., and Strauss, D. (1991). Pseudolikelihood Estimation: Some Examples. Sankhyā: The Indian Journal of Statistics, Series B (1960-2002) 53, 233-243.
96. Hudson, R. (1990). Gene Genealogies and the Coalescent. In Oxford Surveys in Evolutionary Biology, D.J. Futuyma and J. Antonovics, eds. (Oxford Univ. Press, Oxford.
97. Coventry, A., Bull-Otterson, L.M., Liu, X., Clark, A.G., Maxwell, T.J., Crosby, J., Hixson, J.E., Rea, T.J., Muzny, D.M., Lewis, L.R., et al. Deep resequencing reveals excess rare recent variants consistent with explosive population growth. Nature Communications 1, 131-.

98. Gravel, S., Henn, B.M., Gutenkunst, R.N., Indap, A.R., Marth, G.T., Clark, A.G., Yu, F., Gibbs, R.A., Bustamante, C.D., Altshuler, D.L., et al. (2011). Demographic history and rare allele sharing among human populations. Proceedings of the National Academy of Sciences of the United States of America 108, 11983-11988.
99. Tishkoff, S.A., and Verrelli, B.C. (2003). PATTERNS OF HUMAN GENETIC DIVERSITY: Implications for Human Evolutionary History and Disease. Annual Review of Genomics and Human Genetics 4, 293-340.
100. Botigué, L.R., Henn, B.M., Gravel, S., Maples, B.K., Gignoux, C.R., Corona, E., Atzmon, G., Burns, E., Ostrer, H., Flores, C., et al. (2013). Gene flow from North Africa contributes to differential human genetic diversity in southern Europe. Proceedings of the National Academy of Sciences 110, 11791-11796.
101. Berlin, I. (2005). African Immigration to Colonial America. History Now.
102. Ralph, P., and Coop, G. (2013). The Geography of Recent Genetic Ancestry across Europe. PLoS Biol 11, e1001555.
103. Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A.R., Auton, A., Indap, A., King, K.S., Bergmann, S., Nelson, M.R., et al. (2008). Genes mirror geography within Europe. Nature 456, 98-101.
104. Lao, O., Lu, T.T., Nothnagel, M., Junge, O., Freitag-Wolf, S., Caliebe, A., Balascakova, M., Bertranpetit, J., Bindoff, L.A., Comas, D., et al. (2008). Correlation between genetic and geographic structure in Europe. Current biology : CB 18, 1241-1248.
105. Plagnol, V., and Wall, J.D. (2006). Possible Ancestral Structure in Human Populations. PLoS genetics 2, e105.
106. Vigilant, L., Stoneking, M., Harpending, H., Hawkes, K., and Wilson, A.C. (1991). African populations and the evolution of human mitochondrial DNA. Science 253, 1503-1507.
107. Penny, D., Steel, M., Waddell, P.J., and Hendy, M.D. (1995). Improved analyses of human mtDNA sequences support a recent African origin for Homo sapiens. Molecular biology and evolution 12, 863-882.
108. Keinan, A., Mullikin, J.C., Patterson, N., and Reich, D. (2007). Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. Nature genetics 39, 1251-1255.
109. Akey, J.M., Eberle, M.A., Rieder, M.J., Carlson, C.S., Shriver, M.D., Nickerson, D.A., and Kruglyak, L. (2004). Population history and natural selection shape patterns of genetic variation in 132 genes. PLoS Biol 2, e286.
110. Fisher, R.A. (1930). The genetical theory of natural selection.(Oxford,: The Clarendon press).
111. Wakeley, J. (2009). Coalescent theory : an introduction.(Greenwood Village, Colo.: Roberts & Co. Publishers).
112. Hartl, D.L., and Clark, A.G. (1997). Principles of population genetics.(Sunderland, Mass.: Sinauer Associates).
113. Li, M., Rothwell, R., Vermaat, M., Wachsmuth, M., Schröder, R., Laros, J.F., van Oven, M., de Bakker, P.I., Bovenberg, J.A., van Duijn, C.M., et al. (2016). Transmission of human mtDNA heteroplasmy in the Genome of the Netherlands families: support for a variable-size bottleneck. Genome Res 26, 417-426.
114. Chinnery, P.F., and Hudson, G. (2013). Mitochondrial genetics. British medical bulletin 106, 135-159.

115. Greaves, L.C., Reeve, A.K., Taylor, R.W., and Turnbull, D.M. (2012). Mitochondrial DNA and disease. The Journal of pathology 226, 274-286.
116. Lombes, A., Aure, K., Bellane-Chantelot, C., Gilleron, M., and Jardel, C. (2014). Unsolved issues related to human mitochondrial diseases. Biochimie 100, 171-176.
117. Wallace, D.C. (2012). Mitochondria and cancer. Nature reviews Cancer 12, 685-698.
118. Wallace, D.C., and Chalkia, D. (2013). Mitochondrial DNA genetics and the heteroplasmy conundrum in evolution and disease. Cold Spring Harbor perspectives in medicine 3, a021220.
119. Cree, L.M., Samuels, D.C., and Chinnery, P.F. (2009). The inheritance of pathogenic mitochondrial DNA mutations. Biochimica et biophysica acta 1792, 1097-1102.
120. Poulton, J., and Morten, K. (1993). Noninvasive diagnosis of the MELAS syndrome from blood DNA. Annals of neurology 34, 116.
121. t Hart, L.M., Jansen, J.J., Lemkes, H.H., de Knijff, P., and Maassen, J.A. (1996). Heteroplasmy levels of a mitochondrial gene mutation associated with diabetes mellitus decrease in leucocyte DNA upon aging. Human mutation 7, 193-197.
122. Rajasimha, H.K., Chinnery, P.F., and Samuels, D.C. (2008). Selection against pathogenic mtDNA mutations in a stem cell population leads to the loss of the 3243A-->G mutation in blood. American journal of human genetics 82, 333-343.
123. de Laat, P., Koene, S., Heuvel, L.P., Rodenburg, R.J., Janssen, M.C., and Smeitink, J.A. (2013). Inheritance of the m.3243A>G mutation. JIMD reports 8, 47-50.
124. Monnot, S., Gigarel, N., Samuels, D.C., Burlet, P., Hesters, L., Frydman, N., Frydman, R., Kerbrat, V., Funalot, B., Martinovic, J., et al. (2011). Segregation of mtDNA throughout human embryofetal development: m.3243A>G as a model system. Human mutation 32, 116-125.
125. Shen, S.S., Liu, C., Xu, Z.Y., Hu, Y.H., Gao, G.F., and Wang, S.Y. (2012). Heteroplasmy levels of mtDNA1555A>G mutation is positively associated with diverse phenotypes and mutation transmission in a Chinese family. Biochemical and biophysical research communications 420, 907-912.
126. Cree, L.M., Samuels, D.C., de Sousa Lopes, S.C., Rajasimha, H.K., Wonnapinij, P., Mann, J.R., Dahl, H.H., and Chinnery, P.F. (2008). A reduction of mitochondrial DNA molecules during embryogenesis explains the rapid segregation of genotypes. Nature genetics 40, 249-254.
127. Fan, W., Waymire, K.G., Narula, N., Li, P., Rocher, C., Coskun, P.E., Vannan, M.A., Narula, J., Macgregor, G.R., and Wallace, D.C. (2008). A mouse model of mitochondrial disease reveals germline selection against severe mtDNA mutations. Science 319, 958-962.
128. Freyer, C., Cree, L.M., Mourier, A., Stewart, J.B., Koolmeister, C., Milenkovic, D., Wai, T., Floros, V.I., Hagstrom, E., Chatzidaki, E.E., et al. (2012). Variation in germline mtDNA heteroplasmy is determined prenatally but modified during subsequent transmission. Nature genetics 44, 1282-1285.
129. Ross, J.M., Stewart, J.B., Hagstrom, E., Brene, S., Mourier, A., Coppotelli, G., Freyer, C., Lagouge, M., Hoffer, B.J., Olson, L., et al. (2013). Germline mitochondrial DNA mutations aggravate ageing and can impair brain development. Nature 501, 412-415.
130. Jacobs, L., Gerards, M., Chinnery, P., Dumoulin, J., de Coo, I., Geraedts, J., and Smeets, H. (2007). mtDNA point mutations are present at various levels of heteroplasmy in human oocytes. Molecular human reproduction 13, 149-154.

131. Marchington, D.R., Hartshorne, G.M., Barlow, D., and Poulton, J. (1997). Homopolymeric tract heteroplasmy in mtDNA from tissues and single oocytes: support for a genetic bottleneck. American journal of human genetics 60, 408-416.
132. Marchington, D.R., Scott Brown, M.S., Lamb, V.K., van Golde, R.J., Kremer, J.A., Tuerlings, J.H., Mariman, E.C., Balen, A.H., and Poulton, J. (2002). No evidence for paternal mtDNA transmission to offspring or extra-embryonic tissues after ICSI. Molecular human reproduction 8, 1046-1049.
133. Goto, H., Dickins, B., Afgan, E., Paul, I.M., Taylor, J., Makova, K.D., and Nekrutenko, A. (2011). Dynamics of mitochondrial heteroplasmy in three families investigated via a repeatable re-sequencing study. Genome biology 12, R59.
134. Guo, Y., Li, C.I., Sheng, Q., Winther, J.F., Cai, Q., Boice, J.D., and Shyr, Y. (2013). Very low-level heteroplasmy mtDNA variations are inherited in humans. Journal of genetics and genomics = Yi chuan xue bao 40, 607-615.
135. Rebolledo-Jaramillo, B., Su, M.S., Stoler, N., McElhoe, J.A., Dickins, B., Blankenberg, D., Korneliussen, T.S., Chiaromonte, F., Nielsen, R., Holland, M.M., et al. (2014). Maternal age effect and severe germ-line bottleneck in the inheritance of human mitochondrial DNA. Proceedings of the National Academy of Sciences of the United States of America.
136. Sekiguchi, K., Kasai, K., and Levin, B.C. (2003). Inter- and intragenerational transmission of a human mitochondrial DNA heteroplasmy among 13 maternally-related individuals and differences between and within tissues in two family members. Mitochondrion 2, 401-414.
137. Sondheimer, N., Glatz, C.E., Tirone, J.E., Deardorff, M.A., Krieger, A.M., and Hakonarson, H. (2011). Neutral mitochondrial heteroplasmy and the influence of aging. Human molecular genetics 20, 1653-1659.
138. Brown, D.T., Samuels, D.C., Michael, E.M., Turnbull, D.M., and Chinnery, P.F. (2001). Random genetic drift determines the level of mutant mtDNA in human primary oocytes. American journal of human genetics 68, 533-536.
139. Bogenhagen, D.F. (2012). Mitochondrial DNA nucleoid structure. Biochimica et biophysica acta 1819, 914-920.
140. Cao, L., Shitara, H., Horii, T., Nagao, Y., Imai, H., Abe, K., Hara, T., Hayashi, J., and Yonekawa, H. (2007). The mitochondrial bottleneck occurs without reduction of mtDNA content in female mouse germ cells. Nature genetics 39, 386-390.
141. Jacobs, H.T., Lehtinen, S.K., and Spelbrink, J.N. (2000). No sex please, we're mitochondria: a hypothesis on the somatic unit of inheritance of mammalian mtDNA. BioEssays : news and reviews in molecular, cellular and developmental biology 22, 564-572.
142. Khrapko, K. (2008). Two ways to make an mtDNA bottleneck. Nature genetics 40, 134-135.
143. Boomsma, D.I., Wijmenga, C., Slagboom, E.P., Swertz, M.A., Karssen, L.C., Abdellaoui, A., Ye, K., Guryev, V., Vermaat, M., van Dijk, F., et al. (2014). The Genome of the Netherlands: design, and project goals. Eur J Hum Genet 22, 221-227.
144. Consortium, G.o.t.N. (2014). Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat Genet 46, 818-825.
145. Gillespie, J.H. (1998). Population genetics : a concise guide.(Baltimore, Md: The Johns Hopkins University Press).
146. Kukat, C., Wurm, C.A., Spåhr, H., Falkenberg, M., Larsson, N.G., and Jakobs, S. (2011). Super-resolution microscopy reveals that mammalian mitochondrial nucleoids have a

uniform size and frequently contain a single copy of mtDNA. Proc Natl Acad Sci U S A 108, 13534-13539.
147. Li, M., Schröder, R., Ni, S., Madea, B., and Stoneking, M. (2015). Extensive tissue-related and allele-related mtDNA heteroplasmy suggests positive selection for somatic mutations. Proc Natl Acad Sci U S A 112, 2491-2496.
148. Kondrashov, A.S. (2003). Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. Hum Mutat 21, 12-27.
149. Kim, S.H., Elango, N., Warden, C., Vigoda, E., and Yi, S.V. (2006). Heterogeneous genomic molecular clocks in primates. PLoS Genet 2, e163.
150. Anzola, M., Cuevas, N., Lopez-Martinez, M., Martinez de Pancorbo, M., and Burgos, J.J. (2004). p16INK4A gene alterations are not a prognostic indicator for survival in patients with hepatocellular carcinoma undergoing curative hepatectomy. J Gastroenterol Hepatol 19, 397-405.
151. Chen, H., Gu, X., Su, I.H., Bottino, R., Contreras, J.L., Tarakhovsky, A., and Kim, S.K. (2009). Polycomb protein Ezh2 regulates pancreatic beta-cell Ink4a/Arf expression and regeneration in diabetes mellitus. Genes Dev 23, 975-985.
152. Bracken, A.P., Kleine-Kohlbrecher, D., Dietrich, N., Pasini, D., Gargiulo, G., Beekman, C., Theilgaard-Mönch, K., Minucci, S., Porse, B.T., Marine, J.C., et al. (2007). The Polycomb group proteins bind throughout the INK4A-ARF locus and are disassociated in senescent cells. Genes Dev 21, 525-530.
153. Davies, O.R., and Pellegrini, L. (2007). Interaction with the BRCA2 C terminus protects RAD51-DNA filaments from disassembly by BRC repeats. Nat Struct Mol Biol 14, 475-483.
154. Darwin, C. (1859). On the origin of species by means of natural selection, or, The preservation of favoured races in the struggle for life.(London,: J. Murray).
155. Kimura, M. (1979). The neutral theory of molecular evolution. Sci Am 241, 98-100, 102, 108 passim.
156. Cooper, G.S., and Stroehla, B.C. (2003). The epidemiology of autoimmune diseases. Autoimmun Rev 2, 119-125.
157. NEEL, J.V. (1962). Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"? Am J Hum Genet 14, 353-362.
158. Bach, J.F. (2002). The effect of infections on susceptibility to autoimmune and allergic diseases. N Engl J Med 347, 911-920.
159. Okada, H., Kuhn, C., Feillet, H., and Bach, J.F. (2010). The 'hygiene hypothesis' for autoimmune and allergic diseases: an update. Clin Exp Immunol 160, 1-9.
160. Strachan, D.P. (1989). Hay fever, hygiene, and household size. BMJ 299, 1259-1260.
161. Prugnolle, F., Manica, A., Charpentier, M., Guégan, J.F., Guernier, V., and Balloux, F. (2005). Pathogen-driven selection and worldwide HLA class I diversity. Curr Biol 15, 1022-1027.
162. Hertz, T., Nolan, D., James, I., John, M., Gaudieri, S., Phillips, E., Huang, J.C., Riadi, G., Mallal, S., and Jojic, N. (2011). Mapping the landscape of host-pathogen coevolution: HLA class I binding and its relationship with evolutionary conservation in human and viral proteins. J Virol 85, 1310-1321.
163. McClelland, E.E., Penn, D.J., and Potts, W.K. (2003). Major histocompatibility complex heterozygote superiority during coinfection. Infect Immun 71, 2079-2086.

164. Aguilar, A., Roemer, G., Debenham, S., Binns, M., Garcelon, D., and Wayne, R.K. (2004). High MHC diversity maintained by balancing selection in an otherwise genetically monomorphic mammal. Proc Natl Acad Sci U S A 101, 3490-3494.
165. Fumagalli, M., Cagliani, R., Pozzoli, U., Riva, S., Comi, G.P., Menozzi, G., Bresolin, N., and Sironi, M. (2009). Widespread balancing selection and pathogen-driven selection at blood group antigen genes. Genome Res 19, 199-212.
166. Barreiro, L.B., and Quintana-Murci, L. (2010). From evolutionary genetics to human immunology: how selection shapes host defence genes. Nat Rev Genet 11, 17-30.
167. Fumagalli, M., Pozzoli, U., Cagliani, R., Comi, G.P., Riva, S., Clerici, M., Bresolin, N., and Sironi, M. (2009). Parasites represent a major selective force for interleukin genes and shape the genetic predisposition to autoimmune conditions. J Exp Med 206, 1395-1408.
168. Raj, T., Kuchroo, M., Replogle, J.M., Raychaudhuri, S., Stranger, B.E., and De Jager, P.L. (2013). Common risk alleles for inflammatory diseases are targets of recent positive selection. Am J Hum Genet 92, 517-529.
169. Zhernakova, A., Elbers, C.C., Ferwerda, B., Romanos, J., Trynka, G., Dubois, P.C., de Kovel, C.G., Franke, L., Oosting, M., Barisani, D., et al. (2010). Evolutionary and functional analysis of celiac risk loci reveals SH2B3 as a protective factor against bacterial infection. Am J Hum Genet 86, 970-977.
170. Voight, B.F., Kudaravalli, S., Wen, X., and Pritchard, J.K. (2006). A map of recent positive selection in the human genome. PLoS Biol 4, e72.
171. Albrechtsen, A., Nielsen, F.C., and Nielsen, R. (2010). Ascertainment biases in SNP chips affect measures of population divergence. Mol Biol Evol 27, 2534-2547.
172. Karlsson, E.K., Kwiatkowski, D.P., and Sabeti, P.C. (2014). Natural selection and infectious disease in human populations. Nat Rev Genet 15, 379-393.
173. Jallow, M., Teo, Y.Y., Small, K.S., Rockett, K.A., Deloukas, P., Clark, T.G., Kivinen, K., Bojang, K.A., Conway, D.J., Pinder, M., et al. (2009). Genome-wide and fine-resolution association analysis of malaria in West Africa. Nat Genet 41, 657-665.
174. Kimura, M. (1968). Evolutionary rate at the molecular level. Nature 217, 624-626.
175. Zeng, K., Fu, Y.X., Shi, S., and Wu, C.I. (2006). Statistical tests for detecting positive selection by utilizing high-frequency variants. Genetics 174, 1431-1439.
176. Fay, J.C., and Wu, C.I. (2000). Hitchhiking under positive Darwinian selection. Genetics 155, 1405-1413.
177. Biswas, S., and Akey, J.M. (2006). Genomic insights into positive selection. Trends Genet 22, 437-446.
178. Charlesworth, B., Morgan, M.T., and Charlesworth, D. (1993). The effect of deleterious mutations on neutral molecular variation. Genetics 134, 1289-1303.
179. Nielsen, R. (2005). Molecular signatures of natural selection. Annu Rev Genet 39, 197-218.
180. Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123, 585-595.
181. Fay, J.C., Wyckoff, G.J., and Wu, C.I. (2001). Positive and negative selection on the human genome. Genetics 158, 1227-1234.
182. Bersaglieri, T., Sabeti, P.C., Patterson, N., Vanderploeg, T., Schaffner, S.F., Drake, J.A., Rhodes, M., Reich, D.E., and Hirschhorn, J.N. (2004). Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet 74, 1111-1120.

183. Liu, X., Ong, R.T., Pillai, E.N., Elzein, A.M., Small, K.S., Clark, T.G., Kwiatkowski, D.P., and Teo, Y.Y. (2013). Detecting and characterizing genomic signatures of positive selection in global populations. Am J Hum Genet 92, 866-881.
184. Schlebusch, C.M., Sjödin, P., Skoglund, P., and Jakobsson, M. (2013). Stronger signal of recent selection for lactase persistence in Maasai than in Europeans. Eur J Hum Genet 21, 550-553.
185. Hollox, E.J., Poulter, M., Zvarik, M., Ferak, V., Krause, A., Jenkins, T., Saha, N., Kozlov, A.I., and Swallow, D.M. (2001). Lactase haplotype diversity in the Old World. Am J Hum Genet 68, 160-172.
186. Messier, W., and Stewart, C.B. (1997). Episodic adaptive evolution of primate lysozymes. Nature 385, 151-154.
187. Yang, Z. (1998). Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15, 568-573.
188. Pulvers, J.N., Journiac, N., Arai, Y., and Nardelli, J. (2015). MCPH1: a window into brain development and evolution. Front Cell Neurosci 9, 92.
189. Shi, L., Li, M., Lin, Q., Qi, X., and Su, B. (2013). Functional divergence of the brain-size regulating gene MCPH1 during primate evolution and the origin of humans. BMC Biol 11, 62.
190. McGowen, M.R., Montgomery, S.H., Clark, C., and Gatesy, J. (2011). Phylogeny and adaptive evolution of the brain-development gene microcephalin (MCPH1) in cetaceans. BMC Evol Biol 11, 98.
191. Bains, R.K., Kovacevic, M., Plaster, C.A., Tarekegn, A., Bekele, E., Bradman, N.N., and Thomas, M.G. (2013). Molecular diversity and population structure at the Cytochrome P450 3A5 gene in Africa. BMC Genet 14, 34.
192. Thompson, E.E., Kuttab-Boulos, H., Witonsky, D., Yang, L., Roe, B.A., and Di Rienzo, A. (2004). CYP3A variation and the evolution of salt-sensitivity variants. Am J Hum Genet 75, 1059-1069.
193. Chen, X., Wang, H., Zhou, G., Zhang, X., Dong, X., Zhi, L., Jin, L., and He, F. (2009). Molecular population genetics of human CYP3A locus: signatures of positive selection and implications for evolutionary environmental medicine. Environ Health Perspect 117, 1541-1548.
194. Toomajian, C., Ajioka, R.S., Jorde, L.B., Kushner, J.P., and Kreitman, M. (2003). A method for detecting recent selection in the human genome from allele age estimates. Genetics 165, 287-297.
195. Toomajian, C., and Kreitman, M. (2002). Sequence variation and haplotype structure at the human HFE locus. Genetics 161, 1609-1623.
196. Ajioka, R.S., Jorde, L.B., Gruen, J.R., Yu, P., Dimitrova, D., Barrow, J., Radisky, E., Edwards, C.Q., Griffen, L.M., and Kushner, J.P. (1997). Haplotype analysis of hemochromatosis: evaluation of different linkage-disequilibrium approaches and evolution of disease chromosomes. Am J Hum Genet 60, 1439-1447.
197. Thomas, W., Fullan, A., Loeb, D.B., McClelland, E.E., Bacon, B.R., and Wolff, R.K. (1998). A haplotype and linkage disequilibrium analysis of the hereditary hemochromatosis gene region. Hum Genet 102, 517-525.
198. Huttley, G.A., Easteal, S., Southey, M.C., Tesoriero, A., Giles, G.G., McCredie, M.R., Hopper, J.L., and Venter, D.J. (2000). Adaptive evolution of the tumour suppressor

BRCA1 in humans and chimpanzees. Australian Breast Cancer Family Study. Nat Genet 25, 410-413.
199. Lou, D.I., McBee, R.M., Le, U.Q., Stone, A.C., Wilkerson, G.K., Demogines, A.M., and Sawyer, S.L. (2014). Rapid evolution of BRCA1 and BRCA2 in humans and other primates. BMC Evol Biol 14, 155.
200. Hamblin, M.T., and Di Rienzo, A. (2000). Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. Am J Hum Genet 66, 1669-1679.
201. Hamblin, M.T., Thompson, E.E., and Di Rienzo, A. (2002). Complex signatures of natural selection at the Duffy blood group locus. Am J Hum Genet 70, 369-383.
202. Parkes, M., Cortes, A., van Heel, D.A., and Brown, M.A. (2013). Genetic insights into common pathways and complex relationships among immune-mediated diseases. Nat Rev Genet 14, 661-673.
203. Fu, Y.X. (1995). Statistical properties of segregating sites. Theor Popul Biol 48, 172-197.
204. Hancock, A.M., and Rienzo, A.D. (2008). Detecting the Genetic Signature of Natural Selection in Human Populations: Models, Methods, and Data. Annu Rev Anthropol 37, 197-217.
205. Yates, A., Akanni, W., Amode, M.R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L., et al. (2016). Ensembl 2016. Nucleic Acids Res 44, D710-716.
206. Chakravarti, I.M., Laha, R.G., and Roy, J. (1967). Handbook of methods of applied statistics.(New York,: Wiley).
207. Kothapalli, K.S., Ye, K., Gadgil, M.S., Carlson, S.E., O'Brien, K.O., Zhang, J.Y., Park, H.G., Ojukwu, K., Zou, J., Hyon, S.S., et al. (2016). Positive Selection on a Regulatory Insertion-Deletion Polymorphism in FADS2 Influences Apparent Endogenous Synthesis of Arachidonic Acid. Mol Biol Evol 33, 1726-1739.
208. Nielsen, R., Bustamante, C., Clark, A.G., Glanowski, S., Sackton, T.B., Hubisz, M.J., Fledel-Alon, A., Tanenbaum, D.M., Civello, D., White, T.J., et al. (2005). A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol 3, e170.
209. Nakagome, S., Alkorta-Aranburu, G., Amato, R., Howie, B., Peter, B.M., Hudson, R.R., and Di Rienzo, A. (2016). Estimating the Ages of Selection Signals from Different Epochs in Human History. Mol Biol Evol 33, 657-669.
210. Reppell, M., Boehnke, M., and Zöllner, S. (2012). FTEC: a coalescent simulator for modeling faster than exponential growth. Bioinformatics 28, 1282-1283.
211. Gusev, A., Palamara, P.F., Aponte, G., Zhuang, Z., Darvasi, A., Gregersen, P., and Pe'er, I. (2012). The architecture of long-range haplotypes shared within and across populations. Mol Biol Evol 29, 473-486.
212. Browning, S.R., and Browning, B.L. (2007). 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, 1084-1097.
213. Bláha, M., Žurovcová, M., Kouba, A., Policar, T., and Kozák, P. (2016). Founder event and its effect on genetic variation in translocated populations of noble crayfish (Astacus astacus). J Appl Genet 57, 99-106.
214. Jamieson, I.G. (2011). Founder effects, inbreeding, and loss of genetic diversity in four avian reintroduction programs. Conserv Biol 25, 115-123.