Using Rare Genetic Variation to Understand Human Demography and the Etiology of Complex Traits

by

Mark T. Reppell

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in the University of Michigan
2014

Doctoral Committee:

Associate Professor Sebastian Zöllner, Chair
Professor Gonçalo Abecasis
Professor Michael Boehnke
Professor David Burke
Associate Professor Bhramar Mukherjee

TABLE OF CONTENTS

Chapter

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

ABSTRACT

Using Rare Genetic Variation to Understand Human Demography and the Etiology of Complex Traits

by

Mark Reppell

Chair: Associate Professor Sebastian Zöllner

Modern sequencing technology has revolutionized almost every aspect of human genetics research. Among the novel findings made possible by the sequencing of large samples is how abundant extremely rare genetic variation is in the human genome. Rare genetic variants are likely to have arisen recently. Thus, they provide novel information about recent population history, and because selection has had little time to act on them, sets of rare variants are potentially enriched with important regulatory and biologically functional variants.

Detecting associations between rare variants and genetic traits is challenging; conventional single marker association statistics have little power at low allele counts. Several statistics that aggregate information from multiple variants to increase power and detect group-wise associations have been proposed. In chapter 2 we address the robustness of these group-based tests to population stratification. Using the joint site frequency spectrum of samples from multiple European populations, we show that group-based tests cluster into two classes, and p-value inflation in each class is correlated with a specific form of population structure.

An abundance of rare genetic variation is evidence of recent population growth. Large sequencing studies have found the frequency spectra they observe in their samples are inconsistent with models of simple exponential growth, likely due to a recent acceleration in the growth rate. To address this, in chapter 3 we propose a two-parameter model of accelerating, faster-than-exponential population growth and incorporate it into the coalescent. We show that our model can generate samples containing large quantities of rare genetic variants without inflating the quantity of more common variants, making them well suited to modeling the recent history of humans.

In chapter 4 we develop a series of analytic calculations that allow us to directly sample internal and external branches from a sample's genealogy without resorting to full coalescent simulations. We show that for constant size populations an exact probability function can be defined for branch lengths, and that by using the expected times between coalescent events we can expand our method to a broader range of demographic models.

CHAPTER 1

Introduction

The development of technology to rapidly and cost effectively sequence the DNA of progressively larger study samples has revolutionized almost every aspect of human genetics research. Among the surprising and novel findings driven by this revolution, the characterization of the full human frequency spectrum stands out for both its importance, and the distance between previous belief and current observation. Large scale sequencing projects have revealed that the human genome is characterized by an abundance of extremely rare genetic variation. In Nelson *et al.* (2012) the sequencing of 202 drug target genes in 14,001 individuals revealed $> 60\%$ of all single-nucleotide polymorphisms (SNPs) were present in a single individual, and an additional 14% were seen in only two subjects. Subsequent studies (Tennessen *et al.*, 2012; Gazave *et al.*, 2014) confirmed the finding that the vast majority of genetic variants observed in humans are extremely rare. The ability to assay and recognize extremely rare variation in human samples adds critical new information to both the search for genetic causes of complex disease and efforts to reconstruct human history.

The role of rare variants in disease and trait etiology is a question of pressing concern (Maher *et al.*, 2012; He *et al.*, 2013; Lohmueller *et al.*, 2013). While their rarity makes any individual variant unlikely to play a substantial role in the prevalence of common diseases, in aggregate, their quantity and potential for larger effects relative to common variants makes an important role in disease etiology likely. Unfortunately, single variant association tests, employed with great success to uncover associations between common variants and disease (Wellcome Trust Case Control Consortium, 2007; Teslovich *et al.*, 2010; Peden *et al.*, 2011), have very little power when applied to genetic variants with low minor allele counts. One approach to overcome this limitation has been the development of statistics to aggregate information from multiple variant sites in order to detect a group-based association with a phenotype of interest. Many group-based methods have been proposed, for example: Li and Leal (2008), Madsen and Browning (2009), Morris and Zeggini (2010), Zawistowski *et al.* (2010), Neal *et al.* (2011), and Wu *et al.* (2011). These methods are known to differ in their power, and their robustness to the inclusion of non-causal variants or variants with different effect directions. An open question is the robustness of these methods to sources of confounding likely to result in spurious associations, like population stratification.

Population stratification occurs when cases and controls for a study are selected at different rates from populations with differences in allele frequencies. For single variant association tests, effective methods for controlling population stratification—like genomic control (Devlin and Roeder, 1999) and principal components analysis (Price *et al.*, 2006)—are well known and widely used (Feenstra *et al.*, 2013; Bentley *et al.*, 2014). However, the impact of population stratification on group-based association tests, where each statistic contains multiple alleles, each with its own population specific history, is a more complex problem.

In chapter 2 we examine how different group-based association tests respond to the introduction of rare variant population structure. We find that group-based tests can be broadly divided into two classes, with tests of the same class responding to population stratification similarly. We quantify two forms of rare variant population structure, using allele sharing (Gravel *et al.*, 2011) and a new statistic we develop named weighted symmetry. We show that each form of population structure is correlated with p-value inflation in one class of group-based statistics. These findings lead to a strategy for detecting p-value inflation due to population stratification in real study settings where whole-genome data are unavailable. The results also suggest that correction for population stratification in group-based tests may best be approached on a case-by-case basis, with appropriate strategies differing between tests and datasets.

In addition to expanding our understanding of disease etiology, rare variants provide us with novel information about recent human demography. At neutral genetic sites, variant frequency is a function of variant age, and consequently the rarest variants give us insight into the most recent past. Thus, it is only with the recent characterization of the rarest portions of the frequency spectrum, requiring both large sample sizes and deep sequencing coverage for accurate genotype calling, that inferences about humanity's most recent past are possible. An abundance of rare genetic variation is a signature of recent population growth (Tajima, 1989). Previously, studies using small samples (Marth *et al.*, 2004; Schaffner *et al.*, 2005; Voight *et al.*, 2005; Gutenkunst *et al.*, 2009; Gravel *et al.*, 2011) were unable to detect the quantity of rare variants present in human samples, and estimated recent growth rates of $< 0.75\%$ per generation. This stands in stark contrast to the estimates from recent large sequencing studies, with Nelson *et al.* (2012) estimating a recent growth rate of 1.7% per generation, Tennessen *et al.* (2012) estimating 1.95%, and Coventry *et al.* (2010) more than 9% per generation. To arrive at their estimates all of these studies except Tennessen *et al.* (2012) rely on either instantaneous growth models (Marth *et al.*, 2004; Schaffner *et al.*, 2005; Voight *et al.*, 2005), or continuous models with a single exponential growth parameter (Gutenkunst *et al.*, 2009; Coventry *et al.*, 2010; Gravel *et al.*, 2011; Nelson *et al.*, 2012). Coventry *et al.* (2010) questioned the suitability of such models to generate results consistent with the observed abundance of rare variation. When fitting the exponential growth model which they inferred to have the highest likelihood, Coventry *et al.* (2010) found the resulting frequency spectrum contained significantly fewer singletons than were observed in their samples. The authors suggested a recent acceleration in population growth could account for their findings. This suggestion was bolstered by Tennessen *et al.* (2012), who found that by using a piecewise exponential growth model, one that allowed for a recent phase of faster population growth, they were able to significantly improve the fit of the model to their data. However, piecewise models suffer from several severe limitations: an *a priori* unknown number of growth phases, a parameter space that expands with every additional growth phase, and an unrealistic discontinuous population growth rate.

In chapter 3 we propose a method to model recent accelerating population growth using a continuous two-parameter model that avoids the limitations of a piecewise approach. In addition to the standard exponential growth parameter, our models incorporate an "acceleration" parameter, which causes not only a population's size, but the rate of growth itself, to increase with time. After implementation in a coalescent framework we show that the second parameter in our model can lead to an abundance of singletons in samples without inflating the quantity of more common variants, impossible with simple exponential growth. In chapter 3 we also demonstrate the importance of large samples for distinguishing between growth models, and discusses how models like ours, incorporating accelerating growth, may shrink the large gap between current census estimates of human population size and the effective population size estimates used by population geneticists.

Finding demographic models that accurately reflect the history of real human populations is a challenging problem. Selecting between models with different parameters to find one that best fits observed data is often accomplished using likelihood estimates based on summary statistics. Recent studies with large datasets have generally used either of two approaches. The first method relies on the coalescent, which is used to simulate large numbers of sample ancestries under different demographic models. Then, summary statistics from the simulated sequences are compared to find the model that gives results most closely resembling real observations. Gazave *et al.* (2014), Nelson *et al.* (2012), Coventry *et al.* (2010), and Schaffner *et al.* (2005) all use this coalescent based approach. The second method, introduced in Gutenkunst *et al.* (2009) and used by both Gravel *et al.* (2011)

and Tennessen *et al.* (2012), uses a diffusion approximation to simultaneously calculate multi-population expected allele frequency spectra under a demographic model of growth and migration.

The diffusion approach to demographic inference explicitly models only the expected frequency spectrum of samples, making additional summary statistics unavailable for inclusion in likelihood calculations. In contrast, coalescent simulations model haplotype sequence, making possible the calculation of summary statistics beyond the frequency spectrum for comparison with observed data. As discussed in both chapter 3 and Gutenkunst *et al.* (2009), the pattern of linkage disequilibrium in a sample contains information about recent demography lacking from the frequency spectrum. The inclusion of pairwise r^2 in likelihood estimation would increase power to distinguish between models of recent growth, and is something we are currently researching (Appendix C.0.12). While it is possible to use coalescent simulations in this effort, they are computationally burdensome, making desirable the development of more efficient alternatives.

In chapter 4 we present one such alternative to the coalescent, a series of equations that allow direct sampling of internal branch lengths from a sample's ancestral tree. The calculations are designed to allow accurate inference of recent population growth rates, including the generation of individual branch lengths necessary for inclusion of pairwise r^2 into likelihood estimation, without requiring simulation of entire genealogies. We show that for a population with constant size, the probability distribution function of internal branch lengths can be written explicitly. Building on the work of Rosenberg (2006), we recursively calculate the distribution of the number of branches with a given number of descendants in an ancestral tree. Finally, we show that by using the expected values for times between coalescent events we can expand our calculations to models with varying past population sizes.

It is only very recently that the abundance of rare variation that characterizes the human genome has been uncovered. The new insight into human genetics and history made possible by this discovery are only beginning to take shape, already our estimates of recent growth have been greatly altered (Keinan and Clark, 2012) and important new disease associations have been detected (Guerreiro *et al.*, 2012; Cruchaga *et al.*, 2014; Ortega *et al.*, 2014). In this dissertation I seek to add our efforts to this new understanding, and present three novel methods making use of rare variation to shed light on human demography and disease.

CHAPTER 2

Sources of population stratification in gene-based rare variant tests identified using the joint site frequency spectrum

2.1 Introduction

Recent large-scale sequencing studies have identified an abundance of rare variation in the human genome, likely resulting from recent rapid population expansion and purifying selection against deleterious variants (Keinan and Clark, 2012; Reppell *et al.*, 2012). Coding regions of the genome are thus enriched for rare, putatively functional variants (Nelson *et al.*, 2012; Tennessen *et al.*, 2012), attractive candidates for explaining some of the missing heritability in complex diseases (Pritchard, 2001; Stahl *et al.*, 2012; Huyghe *et al.*, 2013). A variety of group-based tests that simultaneously analyze multiple rare variants have been proposed to assess the role of rare variation in the etiology of complex disease. The majority these tests can be partitioned into two categories based on the underlying assumptions of the genotype-phenotype model (Liu *et al.*, 2013a). The first category, based on the concept of rare variant "burden" tests for a significant correlation between a disease phenotype and an aggregate rare variant summary statistic computed for each individual in a dataset. Example burden test summary statistics include an indicator for presence of at least one rare allele (Li and Leal, 2008), the total count of rare alleles (Morris and Zeggini, 2010; Zawistowski *et al.*, 2010), and a weighted count of rare alleles (Madsen and Browning, 2009). In contrast, "dispersion" tests model the marginal effects of individual rare alleles and combine this information across multiple sites to test for association, specifically modeling variants with opposite directions of risk effect. Two popular examples of dispersion tests include the Sequence Kernel Association Test (Wu *et al.*, 2011) (SKAT) and C-Alpha (Neale *et al.*, 2011).

Comparative analyses have shown that performance can vary dramatically among rare variant tests, particularly with respect to the underlying phenotype model and the inclusion of non-causal variants (Ladouceur *et al.*, 2012; Liu *et al.*, 2013c). For example, dispersion tests have more power to identify regions containing a mix of risk and protective rare variants while burden tests can have more power when all rare variants either increase or decrease risk. Thus, both classes of tests will routinely be applied to the same sequencing dataset, and understanding the behaviors of each test is critical for interpreting results. In this chapter we demonstrate that the two classes of group-based tests respond differently to rare variant population structure, leading to unique patterns of population stratification.

Population stratification arises when cases and controls are sampled at differential rates from genetically divergent populations (Li, 1969; Devlin and Roeder, 1999). Frequencies of individual rare alleles differ between populations due to geographic localization and limited sharing of rare variation (Gravel *et al.*, 2011; Nelson *et al.*, 2012). Also, populations can differ in the total quantity of rare alleles they harbor due to differences in effective population sizes, demographic events, bottlenecks, or selective pressures (1000 Genomes Project Consortium *et al.*, 2010; Gravel *et al.*, 2011; Nelson *et al.*, 2012). For example, African populations contain a larger number of rare variant sites than European populations, and within Europe, there is an increasing gradient of cumulative rare variation moving from north to south (1000 Genomes Project Consortium *et al.*, 2010; Nelson *et al.*, 2012). Stratification in single marker tests depends only on differences in population allele frequencies at individual sites (Price *et al.*, 2006; Gravel *et al.*, 2011). In contrast, group-based tests, which aggregate information across multiple sites, must contend with population differences in both individual allele frequencies and the total quantity of rare alleles.

Several recent papers address stratification in group-based tests. Mathieson and McVean (2012) and Kiezun *et al.* (2012) initially demonstrated that burden-style tests are prone to inflation due to underlying population structure, and that the degree of inflation can differ from single-marker tests. Liu *et al.* (2013) reported differential levels of stratification between C-Alpha and a collapsing test in data simulated using a specific coalescent model. In addition, burden tests had lower levels of inflation relative to C-Alpha in a recent analysis of rare variation in autism spectrum disorders (Liu *et al.*, 2013b). In this chapter, we investigated the specific patterns of rare variant population structure that affect the type I error of group-based tests. In particular, we find that frequency differences of individual rare variants have a much stronger effect on dispersion tests than burden tests. In contrast, population differences in overall abundance of rare alleles inflate only burden tests. This difference leads to differential inflation between group-based rare variant tests. We quantified the rare variant patterns in European populations and conclude that the pattern responsible for inflating dispersion tests is likely more common in real data.

We designed an analysis around the joint site frequency spectra (JSFS) of rare, non-synonymous variants identified as part of a previously published sequencing study initially designed to identify and characterize variation in 202 drug-target genes in 14,002 world-wide individuals (Nelson *et al.*, 2012). The JSFS is a common tool in population genetics to summarize the configuration of observed allele counts between two groups of samples, typically from different populations (Gravel *et al.*, 2011). Here, we used the JSFS as probabilistic models from which we generated examples of case-control data sets containing realistic patterns of population structure, but without any true genotype-phenotype association. We focused on the JSFS from four geographically-defined European populations: Central, Western, Northwestern, and Northern Europeans (see map in Figure 2.1). The genetic diversity in our JSFS reflects population structure that could reasonably be present in an association study of European samples, and provides an ideal method to study realistic group-based test inflation. In addition to the empirical data we developed an analytic model of the JSFS. This model was motivated by the hierarchical beta model for population-specific allele frequencies introduced by Balding and Nichols (1995) and yielded simulated results qualitatively identical to the empirical analysis. More details of the analytic model are provided as as appendix A.0.1.

Our JSFS-based simulation strategy was motivated by the fact that although the Nelson *et al.* (2012) data set contains sequence data from many populations, the number of samples within individual populations does not allow for standard resampling techniques. The joint distribution of rare alleles between pairs of populations, summarized in the JSFS, provided a means for unlimited sampling of population allele counts from their empirical distributions. As group-based tests operate directly on the JSFS of cases and controls, our approach retained the critical population-level information that confounds group-based tests without requiring individual-level sequence data.

2.2 Methods

2.2.1 The joint site frequency spectrum

Consider a sequencing dataset with N haplotypes sampled from population 1 and N haplotypes sampled from population 2. For a given polymorphic site in the data set, let $\phi(i, j)$ denote the probability that i copies of the non-reference allele are observed among the N population 1 haplotypes and j copies are observed among the the N population 2 haplotypes. Then, we define $\Phi = \{\phi(i, j) | i, j \in (0, N)\}$ to the the JSFS of populations 1 and 2

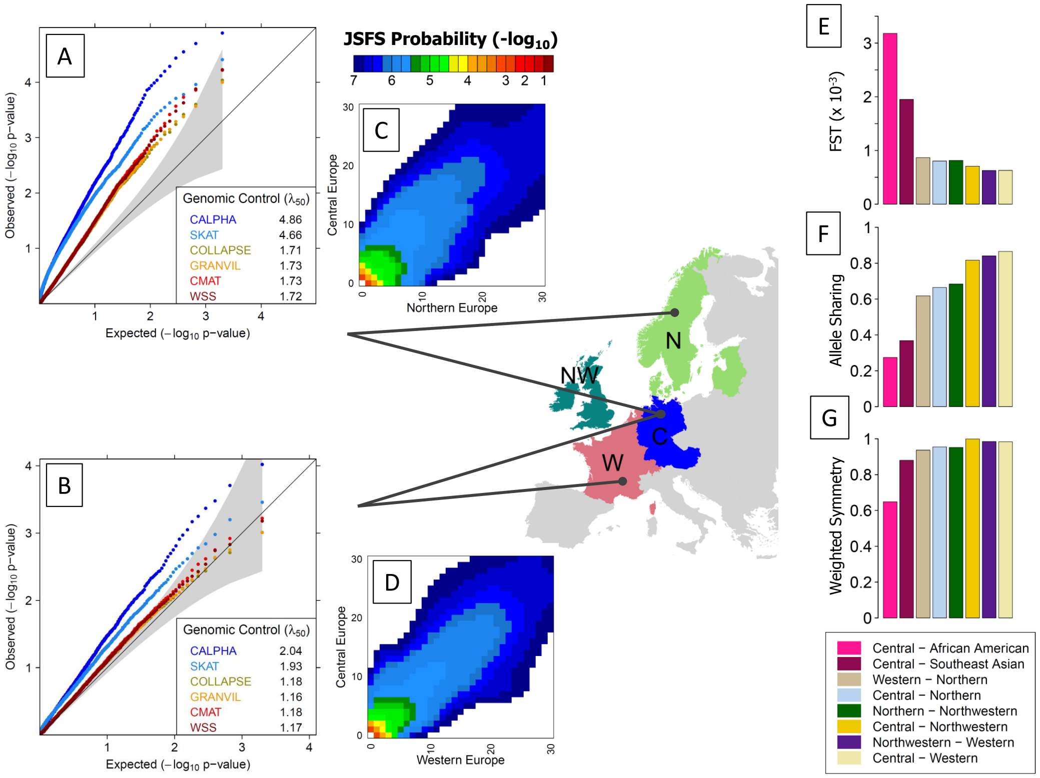

Figure 2.1: Rare variant diversity statistics and p-value distributions for group-based tests in structured European populations. We focused on the empirical joint site frequency spectra (JSFS) of rare, nonsynonymous variants identified by sequencing of Northern, Northwestern, Western and Central European population samples (labeled N, NW, W and C in insert). Heatmaps of the JSFS, pictured for **(C)** Central and Northern European and **(D)** Central and Western European comparisons, provide a graphical representation of the distribution of rare alleles between populations. We quantified aspects of rare variant structure using: **(E)** the F_{ST} statistic of population divergence, **(F)** the allele sharing statistic, measuring variation in population-specific allele frequencies, and **(G)** the weighted symmetry statistic, measuring the evenness of cumulative rare variant load between populations. We analyzed datasets simulated from each JSFS containing population structure but no genotype-phenotype association using several group-based rare variant tests. QQ-plots of the resulting p-value distributions are shown for **(A)** Central and Northern Europeans and **(B)** Central and Western Europeans. Genomic control (λ_{50}) quantifies p-value inflation relative to a uniform null distribution. For illustrative purposes, we display the QQ-plots for an extreme sampling scenario where all cases are sampled from one population and all controls from the other population. Results for less extreme scenarios are shown in Figure 2.3. We find that datasets from more divergent populations produce higher levels of p-value inflation for each group-based test than datasets from more closely related populations. Furthermore, SKAT and C-alpha (blue dots in QQ-plots) consistently show much higher inflation than the Collapsing, GRANVIL, CMAT and WSS tests (red dots in QQ-plots).

for our sample.

The empirical JSFS for multiple worldwide populations were previously estimated and reported as part of Nelson *et al.* (2012). In this study, 202 drug-target genes were deep sequenced in 14,002 samples, including European ($N = 12,514$), African-American ($N = 594$) and Southern Asian ($N = 566$, mostly from India) individuals. The sequenced samples were derived from several case-control data sets. Within each disease study, individuals with pairwise relatedness, $\hat{\pi}$, of > 0.0625 were removed to eliminate closely related individuals. Previous rare variant analyses of these disease studies discovered no significant phenotype associations (Nelson *et al.*, 2012). We focused our analysis on four European subpopulations that were geographically classified according to the UN geoscheme for Europe: Northwestern European (Great Britain and Ireland), Northern Europeans (Norway and Sweden), Western European (Belgium, France, Luxembourg, and The Netherlands) and Central Europe (Austria, Germany, and Switzerland). To account for differences in population sample sizes, the JSFS were computed by averaging over downsampled realizations of 474 individuals per population. We focused on rare, putatively functional variants likely to be included in group-based tests by restricting attention to the JSFS of nonsynonymous variants with sample minor allele frequency $< 1\%$.

2.2.2 JSFS summary statistics

We quantified rare variant population structure within a JSFS using three summary statistics. To focus on rare variants, we computed each summary statistic over allele counts i, j for which the pooled sample allele frequency $(i + j)/2N \leq 0.01$. We calculated an overall measure of genetic diversity using a variation of the standard F_{ST} statistic:

$$F_{ST} = 1 - \frac{\sum_i \sum_j \phi(i,j) \frac{1}{2} [2\frac{i}{N}(1 - \frac{i}{N}) + 2\frac{j}{N}(1 - \frac{j}{N})]}{\sum_i \sum_j \phi(i,j) 2 \frac{i+j}{2N}(1 - \frac{i+j}{2N})}. \tag{2.1}$$

Allele sharing (Gravel *et al.*, 2011) is the probability that two individuals carrying an allele of count n in the sample come from different populations, normalized by the expected frequency in a panmictic population, it is calculated as

$$AS_n = \frac{\sum_{i+j=n} 2ij\phi(i,j)}{\sum_{i+j=n} \binom{n}{2} \phi(i,j)}. \tag{2.2}$$

The allele sharing statistic (AS) for an entire JSFS of rare alleles (with n_{rare} defined by the frequency cutoff above) is defined as the weighted average of AS_n:

$$AS = \frac{\sum_{n \leq n_{rare}} \left[AS_n \sum_{i+j=n} \phi(i,j) \right]}{\sum_{i+j \leq n_{rare}} \phi(i,j)}. \tag{2.3}$$

where $n_{rare} = 2N \times 0.01$ denotes a 1% sample allele frequency threshold.

Weighted symmetry (WS) measures how evenly rare alleles are distributed between the two populations:

$$WS = \frac{min(\sum_i \sum_j \left[i \times \phi(i,j) \right], \ \sum_i \sum_j \left[j \times \phi(i,j) \right])}{\frac{1}{2} \sum_i \sum_j (i+j) \phi(i,j)}. \tag{2.4}$$

In a JSFS for two populations, allele sharing measures the relative weight of probability away from the $x = y$ line. Greater probability weight off the JSFS diagonal means a greater probability that two copies of a randomly sampled allele will be found in members of the same population rather than in different populations, yielding a lower allele sharing statistic. Weighted symmetry measures the balance of probability weight across the $x = y$ line of a JSFS. When probability is greater on one half of the JSFS, there is an imbalance in the quantity of variation between the populations, and weighted symmetry decreases. Figure 2.2 provides a graphical interpretation of the statistics.

2.2.3 JSFS transformations

To isolate the effects of allele sharing and weighted symmetry on test statistic inflation, we designed two transformations that redistribute probability within a JSFS. For each transformation we began with a panmictic JSFS (Gravel *et al.*, 2011) with weighted symmetry $WS = 1$ and allele sharing $AS = 1$.

The first transformation created a sequence of JSFS, each with the same weighted symmetry but with decreasing allele sharing by iteratively applying the following function:

$$\begin{aligned} \phi(i,j)_{(k+1)} =& (1 - \alpha_{\phi(i,j)}) \times \phi(i,j)_{(k)} + \alpha_{\phi(i-1,j+1)} \times \phi(i-1,j+1)_{(k)} \times I_{(i-1) \geq (j+1)} \\ & + \alpha_{\phi(i,j)} \times \phi(i,j)_{(k)} \times I_{i=0} \end{aligned} \tag{2.5}$$

for $i > j$ and

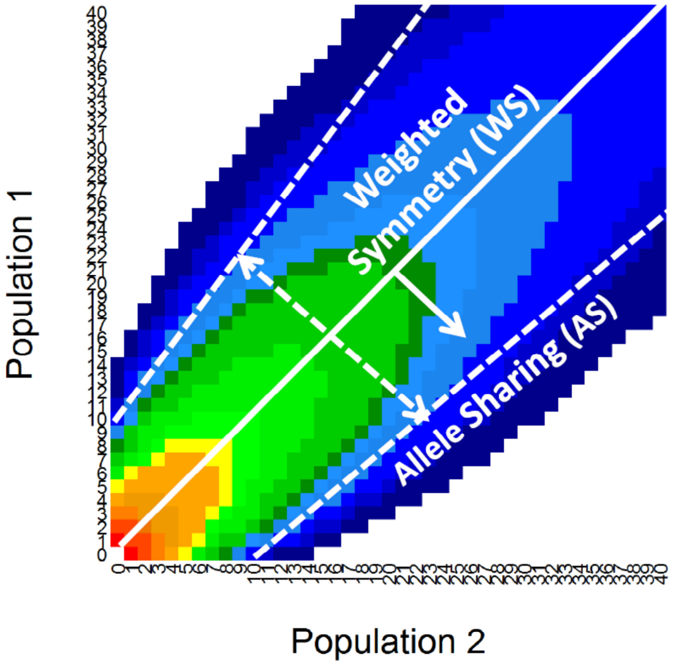

Figure 2.2: Graphical interpretation of allele sharing and weighted symmetry JSFS summary statistics. The heatmap of a JSFS for two closely related populations is characterized by a "cloud" of probability that is heaviest near the origin and dissipates as you move away.The probability cloud of a JSFS based on panmictic population sampling is symmetric about the $x = y$ line ($WS = 1$) and has a thin width corresponding to a hypergeometric distribution ($AS = 1$). Allele sharing measures the width of the probability cloud while weighted symmetry measures the symmetry of the cloud with respect to the $x = y$ line. For divergent populations the probability cloud increases in width as population allele frequencies become more dispersed, corresponding to reduced allele sharing; if either population contains a higher load of rare alleles, corresponding to reduced weighted symmetry, mass within the cloud shifts across the $x = y$ line towards that population.

$$\begin{aligned}\phi(i,j)_{(k+1)} =& (1-\alpha_{\phi(i,j)}) \times \phi(i,j)_{(k)} + \alpha_{\phi(i+1,j-1)} \times \phi(i+1,j-1)_{(k)} \times I_{(i+1)\leq(j-1)} \\ &+ \alpha_{\phi(i,j)} \times \phi(i,j)_{(k)} \times I_{i=0}\end{aligned} \tag{2.6}$$

for $i < j$. Here $\phi(i,j)_{(k)}$ is the (i,j)th element of the kth iteration of the sequence of JSFS and $\alpha_{\phi(i,j)}$ is a weight which decreases as the transformation moves away from the $x = y$ line. This transformation moves probability away from the $x = y$ line, increasing the probability of observing larger differences between population allele counts i and j.

Second, we created a sequence of JSFS with a fixed value of allele sharing but decreasing weighted symmetry by iteratively moving probability across the $x = y$ line from one half of the spectrum to the other using the following transformation:

$$\begin{aligned} \phi(i,j)_{(k+1)} &= \phi(i,j)_{(k)} + \alpha\phi(j,i)_{(k)} \\ \phi(j,i)_{(k+1)} &= \phi(j,i)_{(k)} - \alpha\phi(j,i)_{(k)}. \end{aligned} \tag{2.7}$$

Where, $\phi(i,j)_{(k)}$ is the (i,j)th element of the kth iteration in the sequence of JSFS. As in the previous transformation, the probability of observing a variant with n total minor alleles in the $2N$ haplotypes does not change with this transformation. However, the probability of observing $i > j$ where i and j are the number of minor alleles observed in populations 1 and 2, respectively, increases.

2.2.4 Data simulation and association testing

For each of the six Nelson et al. (2012) inter-European comparisons, we treated the respective JSFS as a joint probability distribution from which we simulated sequence data. The JSFS depends on sample size, so as our empirical JSFS were computed using 474 individuals from each population, we simulated genotypes for 948 total individuals at each gene: 474 individuals (N=948 haplotypes) from each of the two populations. For each genic realization, we sampled pairs of allele counts for S different rare variant sites $(i_s, j_s | 1 \leq s \leq S)$, each with probability according to the JSFS. At the sth site, we randomly distributed the i_s copies of the minor allele among $N = 948$ population 1 haplotypes and j_s copies among $N = 948$ population 2 haplotypes. Allele counts for the S different sites were independently drawn from the JSFS, implicitly assuming a lack of correlation between rare variants in a gene. Although this may not reflect the true relationship between all rare variants, it does not affect test performance as each test is designed to account for correlation between sites.

To induce varying degrees of population structure, we first created diploid samples by randomly pairing haplotypes within each population group. We then assigned a phenotype status to each diploid sample based solely on population affiliation. Treating r as a mixing parameter, we randomly selected $r \times N/2$ samples from the first population to be cases and the remaining $(1 - r) \times N/2$ to be controls. We then assigned $(1 - r) \times N/2$ and $r \times N/2$ haplotypes from the second population to be cases and controls, respectively. Data sets constructed with $r = 0.5$ contained equal numbers of cases and controls from each population. Alternatively, $r = 1.0$ indicated an extreme sampling scenario where all cases were from one population and all controls were from the other population.

We treated each set of S allele counts as an independent "gene" and test for association with the assigned phenotype using 6 tests: Collapsing (Li and Leal, 2008), CMAT (Zawistowski *et al.*, 2010), GRANVIL (Morris and Zeggini, 2010), Weighted Sum Statistic (WSS) (Madsen and Browning, 2009), SKAT (Wu *et al.*, 2011), and C-alpha (Neale *et al.*, 2011). We report p-value distributions for 1,000 genes averaged across 10 replicate runs. With our sample size fixed (474 cases, 474 controls), we varied S between 10 and 60, and used a range of r values between 0.5 and 1.0. We quantified inflation in the distribution of p-values of each test relative to the expected uniform null distribution using a variation on the genomic control statistic of Devlin and Roeder (1995). For $p_{(50)}$ and $p_{(90)}$, the median and 90th percentile p-value for a test statistic's observed p-value distribution, we define

$$\lambda_{50} = \frac{f_{\chi^2}^{-1}(p_{(50)})}{f_{\chi^2}^{-1}(0.5)} \text{ and } \lambda_{90} = \frac{f_{\chi^2}^{-1}(p_{(90)})}{f_{\chi^2}^{-1}(0.9)}$$

where $f_{\chi^2}^{-1}()$ is the quantile function for a one-degree of freedom chi-squared random variable.

2.3 Results

We simulated data sets containing various degrees of rare variant population differentiation using the empirical JSFS of rare, nonsynonymous variation in four geographically-defined European populations: Central, Western, Northwestern and Northern Europeans (Nelson *et al.*, 2012). Pairwise variant F_{ST} computed on the JSFS ranged from 6.26×10^{-4} to 8.66×10^{-4} (Figure 2.1), indicating low overall genetic divergence (Weir *et al.*, 2005).

We analyzed the data sets with four burden tests—Collapsing (Li and Leal, 2008), CMAT (Zawistowski *et al.*, 2010), GRANVIL (Morris and Zeggini, 2010), and the Weighted Sum Statistic (WSS) (Madsen and Browning, 2009)—and two dispersion tests: SKAT (Wu *et al.*, 2011), and C-alpha (Neale *et al.*, 2011). With a fixed sample size of 474 cases and 474 controls, and a fixed number of variants in each gene ($S = 30$), we allowed r, the mixing parameter, to vary between 0.5 and 1.0. We summarized the resulting p-value inflation using genomic control values (Devlin and Roeder, 1999), and reported the inflation in both the medians (λ_{50}) and toward the right tails (λ_{90}) of the p-value distributions (Figure 2.3).

Data sets simulated with balanced population sampling ($r = 0.5$) yielded median genomic control values of $\lambda_{50} \approx 1.00$ for all tests in each population comparison, indicating no inflation. Dispersion tests were deflated for λ_{90} and above, consistent with the conservative nature of these tests for smaller samples sizes and at stringent alpha levels (Wu *et al.*, 2011). Genomic control values increased for each test and population comparison

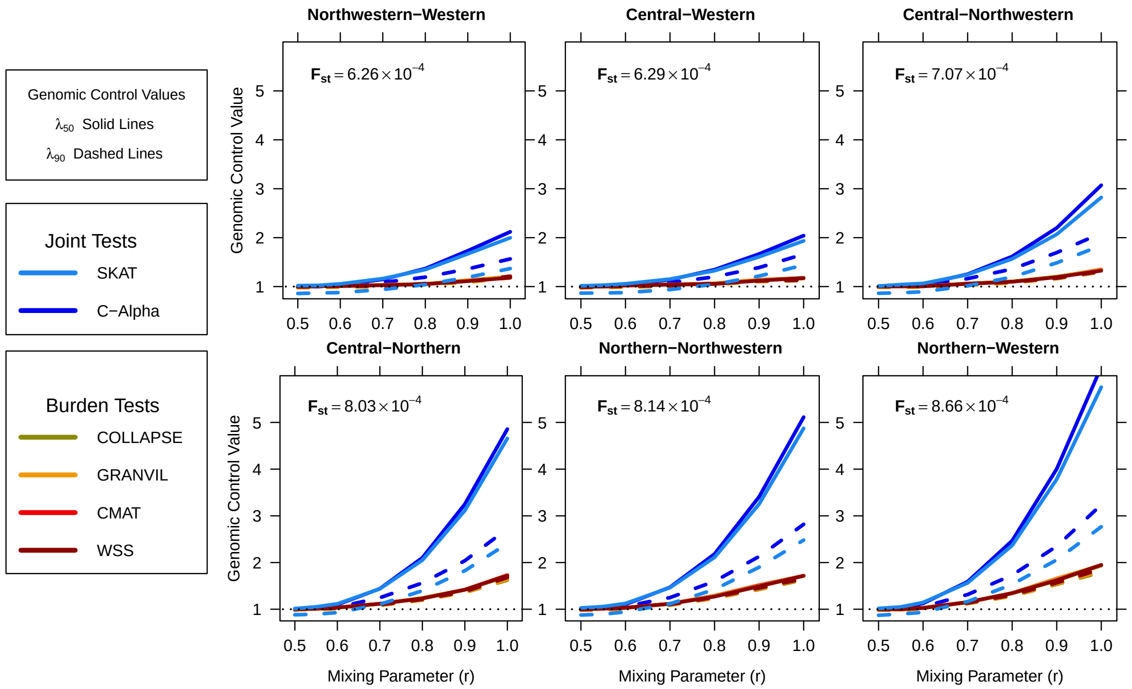

Figure 2.3: Genomic control (GC) values for group-based rare variant tests in structured European datasets. Median GC values (λ_{50}, solid lines) and 90^{th} percentile GC values ($\lambda_{(90)}$, dashed lines) are shown at a range of mixing parameters (r) for each inter-European population comparison. For scenarios containing population structure ($r > 0.5$), the dispersion tests (blue lines) consistently have higher λ_{50} values than the burden tests (red lines) in all population scenarios. In addition, $\lambda_{(90)} << \lambda_{50}$ in many scenarios for the dispersion tests, indicating that inflation in the dispersion tests is not consistent across the p-value distribution.

as the mixing ratio increased from $r = 0.5$ to $r = 1.0$, indicating p-value inflation due to population structure. More divergent populations, as quantified by F_{ST}, showed higher levels of inflation for each test. For example, at mixing ratio $r = 0.8$, the genomic control of the Collapsing test was $\lambda_{50} = 1.05$ in the Central European and Western European comparison ($F_{ST} = 6.29 \times 10^{-4}$) but $\lambda_{50} = 1.23$ in the more divergent Central European and Northwestern European comparison ($F_{ST} = 7.07 \times 10^{-4}$). In many cases, inflation in the medians of the p-value distributions was larger than in the tails (ie $\lambda_{50} > \lambda_{90}$) as evidenced by the difference between dashed (λ_{90}) and solid lines (λ_{50}) in each panel of Figure 2.3. The inconsistent inflation was more pronounced in dispersion tests, and increased in magnitude with both increasing r and increasing population diversity. As a result, standard genomic control severely over corrected inflated p-values more often for dispersion tests than for burden tests (Appendix A.0.2).

Comparing inflation statistics between tests, we observed two consistent patterns across all scenarios. First, the level of p-value inflation for the different tests clustered into two distinct groups, one consisting of the dispersion tests: SKAT and C-Alpha, and the other containing the burden tests: CMAT, Collapsing, WSS, and GRANVIL. Within each group, the level of inflation was similar between tests. For example, in data sets of Central and Western Europeans with $r = 0.7$, each burden test had $\lambda_{50} \approx 1.04$, whereas SKAT and C-Alpha had λ_{50} values near 1.15. The distinct patterns of inflation for the two classes of tests can be seen in Figures 2.1a and 2.1b where burden tests (red dots) clustered together tightly, and are clearly separated from dispersion tests (blue dots). The second consistent pattern in the analysis was higher inflation for dispersion test statistics relative to burden test statistics; the difference increasing with both the divergence of the underlying populations and the mixing parameter r. For example, the difference in inflation between CMAT and SKAT rose from $\lambda_{50} = 1.04$ and 1.13, respectively, in Central and Western data sets to $\lambda_{50} = 1.15$ and 1.56 for the JSFS of the more divergent Northern and Western Europeans at $r = 0.7$.

We hypothesized that the observed patterns of p-value inflation for burden and dispersion tests could be explained by specific underlying rare variant population structures. To test this, we quantified specific patterns of population structure within the JSFS using two statistics: allele sharing and weighted symmetry (see Methods, Figure 2.2). The allele sharing (AS) statistic (Gravel *et al.*, 2011) quantifies inter-population differences in individual allele frequencies for a JSFS. $AS = 1$ indicates allele frequency differences consistent with panmictic population sampling and the statistic decreases towards zero as differences in population allele frequencies increase. We developed the weighted symmetry (WS) statistic to summarize the difference in overall rare allele abundance between populations. Weighted symmetry of 1 indicates an equal quantity of rare alleles in each population, and WS decreases towards zero with increasing inequality in rare allele abundance.

We isolated the effects of the population structures quantified by weighted symmetry and allele sharing on test statistic inflation by analyzing data sets simulated from JSFS where one statistic was fixed and the other decreased in value (see 2.2.3). We first analyzed JSFS with weighted symmetry fixed at $WS = 1$ (Figure 2.4a). When allele sharing also equaled one, the JSFS is equivalent to panmictic sampling and there is no inflation for any test. As allele sharing decreased, genomic control values quickly increased for the dispersion tests, indicating p-value inflation. In comparison, there was only a slight increase in inflation for the burden tests. Next, we considered JSFS with allele sharing fixed at $AS = 1$ and allowed weighted symmetry to decrease. p-value inflation for every burden test increased with decreasing WS, but both SKAT and C-Alpha were unaffected (Figure 2.4b). Taken together, these results imply that the two classes of tests have opposite responses to the rare variant structures quantified by decreasing weighted symmetry and decreasing allele sharing. Inflation in burden tests is primarily due to unequal contributions of rare alleles between the two populations, whereas dispersion test inflation is driven solely by differences in population-specific frequencies of individual rare alleles.

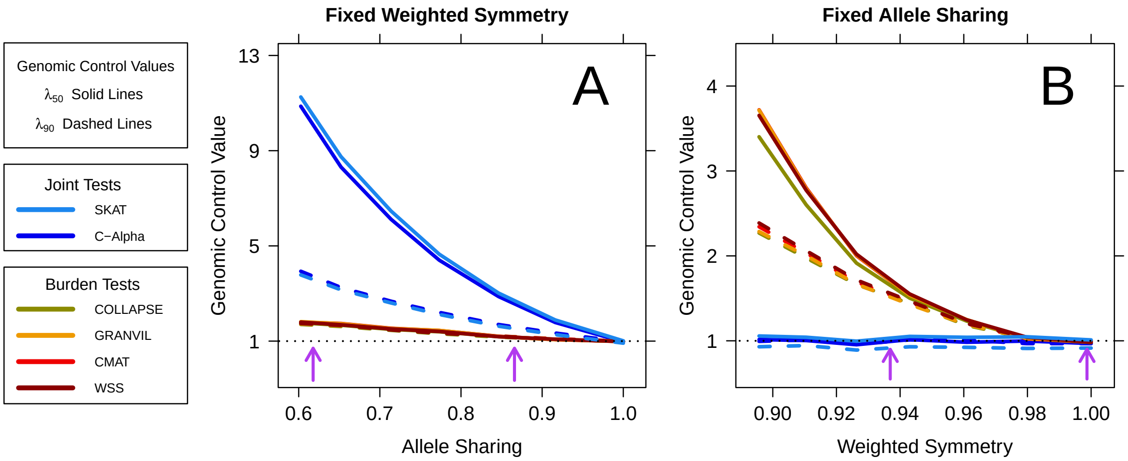

Figure 2.4: The isolated effects of weighted symmetry and allele sharing on p-value inflation in group-based rare variant tests. **(A)** For dataset simulated with weighted symmetry fixed at $WS = 1$ and decreasing allele sharing, inflation grows much larger for the dispersion tests than for the burden tests. **(B)** In contrast, for datasets simulated with allele sharing fixed at $AS = 1$ and decreasing values of weighted symmetry, inflation in each burden test increases while the dispersion tests remain well-controlled. Thus, the two classes of group-based tests have differing responses to these patterns of rare variant population structure. The purple arrows in each plot indicate the minimum and maximum values of that statistic observed in the European joint site frequency spectra. The range of empirical values explains why we observed higher levels of inflation in the dispersion tests.

Having established that burden test inflation correlates strongly with weighted symmetry and dispersion test inflation with allele sharing, we computed these quantities for our European JSFS (Figures 2.1f and 2.1g). Allele sharing ranged between 0.86 and 0.62, with the lowest values observed for JSFS containing the Northern Europeans. We observed weighted symmetry values as high as 0.99 for the JSFS of Central and Northwestern populations and as low as 0.94 for Northern and Western Europeans. The lower weighted symmetry values for JSFS containing Northern Europeans are indicative of fewer rare alleles in that population, consistent with the hypothesis that a historical bottleneck event decreased the population's effective size. Our simulations with fixed weighted symmetry and allele sharing provide context for the differential inflation observed in the inter-European data sets. Allele sharing between European populations was sufficiently low to produce large inflation in the dispersion tests (purple arrows in Figure 2.4a). Alternatively, weighted symmetry between European populations did not decrease to levels that produced substantial inflation in burden tests (Figure 2.4b).

For comparison, we also computed allele sharing and weighted symmetry for the JSFS between our European samples and both African-American and South Asian samples from the same Nelson *et al.* (2012) data set. As expected we saw smaller values of both statistics for these intercontinental population comparisons (Figures 2.1f and 2.1g). Allele sharing between Europeans and African-Americans ranged from 0.22 to 0.28, and from 0.27 to 0.37 between Europeans and the South Asians. Weighted symmetry between the European populations and South Asian took values of 0.90, slightly less than the inter-European comparisons. Weighted symmetry between the African-Americans and Europeans however was much lower, between 0.62 and 0.66, highlighting the larger difference in the total number of rare alleles between these populations. Extrapolating on the theoretical results in Figure 2.4, the values of weighted symmetry between Europeans and African-Americans or Europeans and South Asians are capable of significantly inflating burden tests. However, for these comparisons, allele sharing is even lower and inflation would still be larger for the dispersion tests.

For the previous results we assumed a fixed number of rare variants within each gene ($S = 30$). In reality, the number of rare variants combined into a group-based test varies depending on several factors, including gene length, sample size, population genetic diversity, annotation, and frequency thresholds. To understand the impact that the number of variants per gene has on stratification we repeated our simulations over a range of values for the number of pooled variants S (Figure 2.5). The two classes of tests responded quite differently to a varying number of pooled sites: dispersion tests showed a clear increase in inflation as S increased, whereas inflation in burden tests remained effectively constant. The differential sources of stratification explain this result. In these closely related European populations, the cumulative quantity of rare alleles is quite similar ($WS \approx 1$) but most individual allele frequencies vary slightly between populations. Additional variants do not alter the cumulative allele balance tested for by burden statistics. However, each additional variant provides further evidence of differing allele frequencies between cases and controls, leading to the increasing inflation for dispersion tests. We expect that in a scenario of populations with smaller WS, inflation in burden tests would also increase with the number of variants.

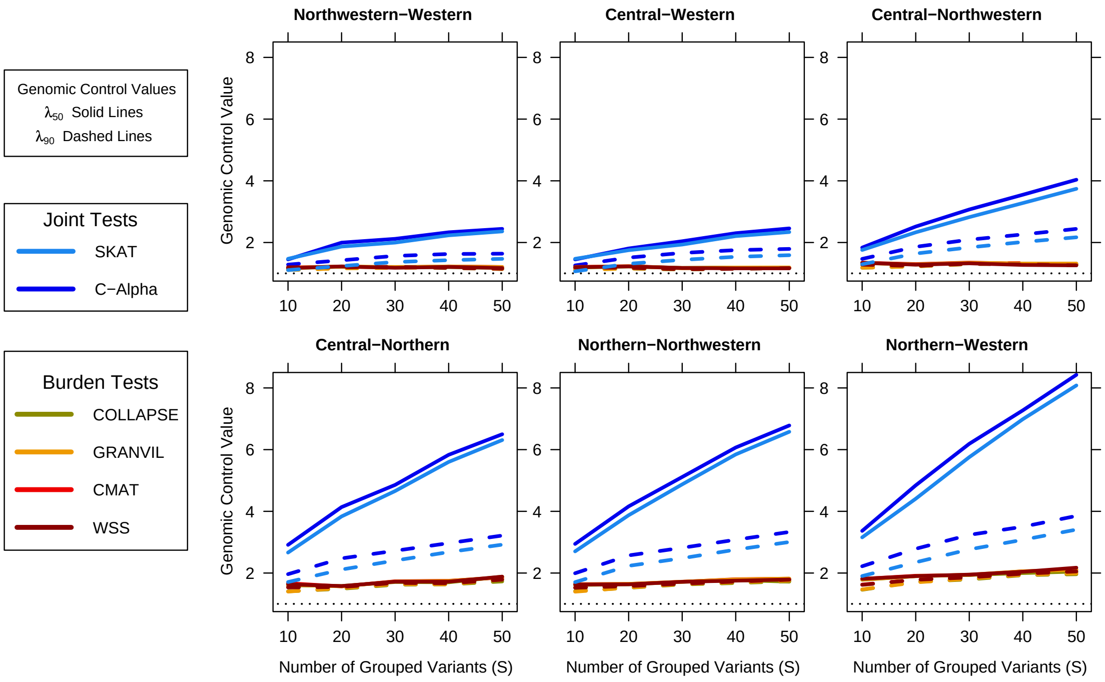

Figure 2.5: The effect of number of rare variant sites (S) pooled together in a group-based tests on p-value inflation (shown for mixing parameter $r = 1.0$). There is a clear increase in inflation for the dispersion tests (blue lines) as the number of rare variant sites pooled into a group-based test increases. Inflation in the burden tests (red lines) remains relatively consistent as the number of sites increases.

2.4 Discussion

We used the JSFS as a model to study the structure of rare variants within European populations and its effect on group-based tests. By quantifying specific patterns in the JSFS, we established that different aspects of population differentiation are responsible for inflating the type I error rates in the two classes of group-based tests. Our results build on those of previous studies examining rare variant population stratification. We independently demonstrated different levels of inflation in C-Alpha and burden tests previously reported in both coalescent simulations (Liu *et al.*, 2013c) and real sequencing data (Liu *et al.*, 2013b). We found that the pattern of differential inflation held more broadly for burden and dispersion tests over a range of population sampling scenarios. Modeling our data sets using the empirical JSFS from several European populations illustrated the magnitude of stratification in realistic samples. Moreover, we identified the precise underlying characteristics of rare variant population structure responsible for the differential stratification, namely, imbalance in rare allele load and overdispersion of individual rare allele frequencies. By looking at the empirical weighted symmetry and allele sharing values observed between multiple

European populations we explained the patterns of population stratification observed in group-based rare variant tests.

A major advantage of dispersion tests over burden tests is a greater power to detect association in genes containing rare variants with opposite directions of effect. Interestingly, it is precisely this ability to accommodate a mix of risk and protective variants that makes dispersion tests more vulnerable to stratification in real population scenarios. Dispersion tests view the population-specific differences in allele frequency at each variant site, regardless of direction, as signal for association. Alternatively, burden tests require the population-specific differences to be predominantly in the same direction, a more stringent criterion. Intuitively, differences in allele frequency (low allele sharing) are more pronounced between populations than differences in the number of rare variants (low weighted symmetry) because all forms of population differentiation resulting in genetic drift lead to allele frequency differences. Creating a significant imbalance in the total quantity of rare variants requires more specific models, for example, a recent bottleneck, unequal migration, or differential growth rates. Thus, the forms of population structure that produce inflation in dispersion tests are more prevalent in real data and we predict that, although dispersion tests provide more power to detect many true rare variant associations, they also require more caution to avoid spurious results.

We anticipate that differential inflation will be particularly problematic for interpreting burden and dispersion test results of sequencing studies that target only a handful of candidate genes. It is straightforward to determine if differential inflation exists when many genes are sequenced (i.e. exome sequencing) by comparing the distributions of p-values for the dispersion and burden tests. This may not be possible in a targeted sequencing data set, and if population structure exists, smaller p-values in dispersion tests could easily be interpreted as stronger signals of true associations rather than increased susceptibility to inflation.

The effect of the number of variants per gene on dispersion statistic inflation provides a practical approach for recognizing stratification. Typically, only rare variants predicted to be deleterious (e.g. nonsynonymous) are included in a gene-level analysis. The remaining excluded rare variants, which likely outnumber the predicted deleterious variants, are predominantly null with respect to phenotype status, yet still contain signal for population structure (Appendix A.0.3). Thus, a dispersion test analysis of the excluded variants is more powerful for detecting population stratification than the analysis of the fewer predicted deleterious variants. We therefore recommend performing the same dispersion test analysis planned for the predicted deleterious variants on the excluded variants as a method to test for population stratification. This method could be particularly helpful for interpreting dispersion test p-values in targeted sequencing studies.

Previous studies have emphasized the challenge of correcting for rare variant population structure in multi-marker group-based tests. Kiezun *et al.* (2012) corrected the stratification using a modified permutation algorithm requiring that population labels be both discrete and either known or accurately estimated, neither of which may be satisfied in real data sets. Mathieson and McVean (2012) and Liu *et al.* (2013b) each showed the standard application of principle components could not correct for all scenarios in either single marker or group-based analyses of rare variants. In light of our finding that inflation differs according to the type of group-based test, the appropriate correction strategy may be context specific depending on the test and populations. We illustrated this point using genomic control. Even in a set of homogeneous genes with a uniform number of variants and identical underlying JSFS, we often observed that the median of the p-value distribution was more highly inflated than the tail of the distribution ($\lambda_{50} > \lambda_{90}$). Under these conditions applying a standard genomic control correction based on λ_{50} over corrects the most significant genes in the analysis (Appendix A.0.2), which reduces power for real associations. The overcorrection was more severe for dispersion tests, implying that genomic control may be more appropriate for burden tests.

Presently, attempts to identify rare risk variants using the pooling approach of group-based tests have had limited success. Despite the potential for stratification seen here, real data sets have often identified no statistically significant genes rather than too many. This lack of significant findings, even false positives, is likely the result of current studies being underpowered due to insufficient sample sizes. Larger sample sizes in future sequencing studies will increase power to find true signals, but will also increase the likelihood of subtle population structure and the number of variants pooled within genes, both of which increase the potential for rare variant population stratification.

CHAPTER 3

The impact of accelerating, faster than exponential population growth on genetic variation

3.1 Introduction

Large-scale, deep sequencing studies have revealed a previously unknown abundance of genetic variation in the human genome, the majority of which is very rare. Nelson *et al.* (2012) sequenced 202 drug target genes in 14,002 individuals and found an alternate allele at 1 in 17 sequenced sites. These variants were overwhelmingly rare, with $> 60\%$ of the observed single nucleotide variants (SNVs) being singletons, present in just one individual. The Exome Sequencing Project (ESP) sequenced the exome of 2,440 individuals and found $> 500,000$ different variant sites, with 57% of these variants singletons (Tennessen *et al.*, 2012). Coventry *et al.* (2010) sequenced the genes *HHEX* and *KCNJ11* in 13,715 individuals and estimated 579 sites with alternate alleles in 12.3 kilobases (kb); singletons made up $> 28\%$ of variants at *HHEX* and $> 17\%$ of variants at *KCNJ11*. This newly discovered abundance of variation in humans, composed primarily of very rare variants, is consistent with recent massive population growth (Tajima, 1989). This result is in stark contrast to older studies where only weak signals of population growth were observed and the precise model of population growth (instantaneous or exponential growth) had little impact on the expected pattern of diversity (Adams and Hudson, 2004; Marth *et al.*, 2004; Williamson *et al.*, 2005; Gutenkunst *et al.*, 2009; Tennessen *et al.*, 2012). Keinan and Clark (2012) explained that many of these studies (Schaffner *et al.*, 2005; Gutenkunst *et al.*, 2009; Gravel *et al.*, 2011) were based on genotype or small scale resequencing data and thus were unable to assay the rarest portions of the frequency spectrum. As a result, models fit with these data failed to account for very recent history, and arrived at low estimates of human growth rates and small current effective population sizes. Coventry *et al.* (2010), with a much larger sample size, using a method of likelihoods based on the observed site frequency spectrum and coalescent simulations, arrived at a faster estimated exponential growth rate in their European origin samples than previous studies. However, when they generated simulated samples using their best estimate growth rate and compared the results to the site frequency spectrum observed at *HHEX* they found the exponential model was still lacking: the number of singleton variants in their sequenced samples significantly exceeded model predictions. Such an abundance of singletons suggests a very large recent population size, but the observed frequencies of less rare variants are consistent with exponential growth rates that fail to achieve a sufficient population size to explain the quantity of singletons. Coventry *et al.* (2010) hypothesized that a recent acceleration in growth could address the discrepancies between their models and their data. To address this issue in their data, Tennessen *et al.* (2012) fit a piecewise growth model with two periods of growth to model a recent acceleration, improving the fit of the simulated frequency spectrum to the observed data. A technical weakness of modeling accelerating growth by piecewise exponential functions is that the resulting growth curve is not continuous. Moreover, piecewise model's parameter space increases in size with every growth rate change and *a priori* the correct number of changes is unknown. Additionally, the exponential models fit by Coventry *et al.* (2010), Tennessen *et al.* (2012), and Nelson *et al.* (2012) all arrived at current effective population size estimates for Europe of < 5 million individuals, a seemingly small number compared to the recent census estimates of a continental population > 738 million (United Nations Department of Economic and Social Affairs Population Division, 2011). The potential for continuous models incorporating faster than exponential (FTE) growth to accurately reflect recent acceleration in population growth, explain the abundance of singleton variation discovered in human samples, maintain accurate prediction of the observed patterns of common variants, and avoid a drastic expansion of the growth parameter space is an important open question. Such FTE models may also result in much larger estimates of present day effective population size than the numbers obtained by Coventry *et al.* (2010), Tennessen *et al.* (2012), and Nelson *et al.* (2012) and thus conveniently provide a way to reduce the gap between census population sizes and effective population size estimates.

To address this question, we propose a two-parameter class of continuous population growth functions which allow accelerating, faster than exponential (FTE) growth. We integrated this model into the coalescent framework (Reppell *et al.*, 2012), a widely used stochastic model which traces the ancestry of a sample backwards through time to its most recent common ancestor (Kingman, 1982a; Hudson, 1983). The original coalescent assumed a constant population size; Donnelly and Tavaré (1995) extended the model to allow for populations with size varying deterministically over time. Using our model, we study rare variant frequency spectra and population genetic summary statistics under a wide range of model parameters. In models with the same initial growth rate, where faster acceleration in FTE models results in larger current population sizes, we find that our model produces an abundance of singleton variation in samples without a subsequent increase in the quantity of more common variants. We also find that when time of growth, ancestral size, and current population size are fixed, FTE growth actually results in fewer total variants, fewer rare variants, and slower decay of LD than exponential growth. With a fixed current size we find that pairwise linkage disequilibrium between very rare variants contains information about recent growth rates. Our work highlights the importance of sample size in distinguishing between growth models and explores the impact of the duration of growth. We find that the addition of the acceleration parameter to growth time and present population size in our models creates a parameter space where sequence summary statistics are consistent with multiple growth times and current sizes, and in particular samples from populations which underwent accelerating growth to larger current sizes share characteristics with samples from exponentially growing populations with smaller current sizes.

3.2 Methods

3.2.1 Faster than exponential growth

We model faster than exponential growth with a set of functions that fit well into the coalescent framework and are motivated by the differential equation suggested by Tolle (2003)

$$\frac{dP}{dt} = \alpha P_t^{\beta}. \tag{3.1}$$

Here, P is population size in number of haplotypes, t is time in generations, and the model parameters α and β are constants. When $\beta = 1$, the solution to this equation is an exponential growth function, for $\beta < 1$ the solution results in slower than exponential growth, while for $\beta > 1$, the solution results in FTE growth, where not just the population size but also the rate of growth increases with time. If we solve equation (3.1) with initial population size P_0 the result is

$$P_t = \begin{cases} [\frac{P_0^{\beta-1}}{1+P_0^{\beta-1}(\beta-1)\alpha t}]^{\frac{1}{\beta-1}} & \text{for } \beta \neq 1 \\ P_0 e^{-\alpha t} & \text{for } \beta = 1 \end{cases}. \tag{3.2}$$

With this parametrization, we interpret α as the exponential growth constant and β as an "acceleration" parameter (Figure 3.1). Under this model, populations can achieve infinite population size in finite time (Appendix B.0.4). As the coalescent conditions on a finite present-day population size, this property does not affect our model or results.

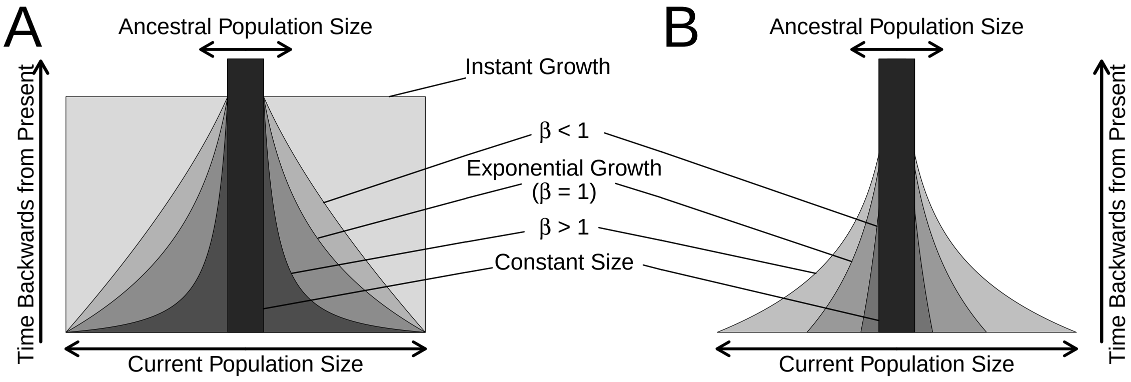

Figure 3.1: **(A)** When conditioning on growth time, ancestral population size and current population size, and comparing to exponential growth ($\beta = 1$), a population where $\beta > 1$ was smaller in size during its past, while growth with $\beta < 1$ results in a population that was larger. **(B)** When current population size is allowed to vary, and the exponential growth parameter α is fixed, values $\beta > 1$ result in current population larger than an exponentially growing population ($\beta = 1$), while values of $\beta < 1$ results in smaller current population sizes.

3.2.2 FTE growth in the coalescent

In the coalescent model we begin with a sample of haplotypes drawn from a current population. Moving backwards in time, coalescent events occur where two sample lineages coalesce into a single lineage through a common ancestor, decreasing the number of distinct lineages by one. Coalescent events are observed until the most recent common ancestor of the entire sample has been reached, and only a single lineage remains. The distribution of coalescent event times depends on the population size and the number of distinct sample lineages remaining. Donnelly and Tavaré (1995) showed that in a population with deterministically varying past size, given an original sample of n haplotypes with j distinct lineages remaining and previous times between coalescent events $T_i(i = n, n-1, ..., j+1)$, the probability of the time to next coalescent event T_j is defined by

$$Prob(T_j > k | T_n + ... + T_{j+1} = S) = exp(-\begin{pmatrix} j \\ 2 \end{pmatrix} [\Lambda(k + S) - \Lambda(S)]). \tag{3.3}$$

Here $\Lambda(t)$ determines how the population size changes over time, and is defined as:

$$\Lambda(t) = \int_0^t \frac{1}{\lambda(s)} ds \tag{3.4}$$

where

$$\lambda(s) = \frac{P_s}{P_0} \tag{3.5}$$

is the ratio of population size at time s and the present time. Once $\Lambda(t)$ is specified, it can be substituted into (3.3) and the time between coalescent events $T_i, i = n, n-1, ..., 2$ can be drawn. For our model, because the coalescent models populations backwards in time, (3.1) must be changed to

$$\frac{dP}{dt} = -\alpha P_t^{\beta} \tag{3.6}$$

before it can be transformed into the corresponding $\Lambda(t)$. For $\beta \neq 1$:

$$\Lambda(t) = \int_0^t (1 + \alpha s(\beta - 1)P_0^{\beta-1})^{\frac{1}{\beta-1}} = \frac{(1 + \alpha t(\beta - 1)P_0^{\beta-1})^{\frac{\beta}{\beta-1}} - 1}{\alpha\beta P_0^{\beta-1}}. \tag{3.7}$$

And when $\beta = 1$:

$$\Lambda(t) = \int_0^t \frac{1}{e^{\alpha s}} ds = \frac{e^{\alpha t} - 1}{\alpha}. \tag{3.8}$$

By substituting $\Lambda(t)$ into (3.3) we can generate coalescent event times for a population that has been growing or contracting according to our two-parameter model.

3.2.3 Simulations

In our simulations, we assume an ancestral population of 20,000 haplotypes which grows over the most recent 100 to 3,000 generations. We simulate 30 kb from samples of between 100 and 20,000 haplotypes and assume an ancestral mutation parameter $\theta = 16.8 = 2N_e m$ where N_e is the effective ancestral population size in haplotypes and m is the product of per base mutation rate and the number of bases analyzed. We also assume a uniform ancestral recombination parameter $\rho = 12 = 2N_e r$. Here, N_e is again the effective ancestral population size in haplotypes and r is the product of per base recombination rate and number of bases analyzed. The parameter value θ corresponds to a per base mutation rate of 1.4×10^{-8} (Campbell *et al.*, 2012; Nelson *et al.*, 2012) and ρ to a recombination rate of 1 cM/Mb (Kong *et al.*, 2002). In our initial analyses we assume a current population size of 8×10^6 haplotypes and 500 generations of growth, approximately consistent with estimates from Coventry *et al.* (2010). For comparison we also simulate samples drawn from (1) populations which maintain a constant size of 20,000 haplotypes throughout history and (2) populations that grow from the ancestral 20,000 haplotypes to a size of 8×10^6 haplotypes instantaneously 500 generations in the past. For each pair of α and β values, we report results for 1,000 independent simulation replicates. We also use simulation to verify that for the range of β values investigated, the coalescent assumption that sample size remains much smaller than effective population size is not violated in a way that changes our findings or conclusions (Appendix B.0.5).

To capture the scale of patterns in linkage disequilibrium (LD) we simulate longer sequences of 100 kb and calculate pairwise r^2 and the absolute value of D' (Appendix B.0.6). We bin variants based on minor allele count and sample a subset for which we calculate all pairwise statistics. Binning allows us to compare LD across models without having the results determined solely by differences in the abundance of very rare variants.

3.3 Results

Using coalescent simulations, we generate samples from populations that grow according to our two-parameter model in which α is an exponential like growth parameter and β an acceleration parameter. We then quantify how changes in growth patterns affect different portions of the sample allele frequency spectrum and linkage disequilibrium. We also demonstrate how varying sample sizes and growth times alter the quantity and nature of genetic diversity within our growth models.

3.3.1 Accelerating growth with fixed current population size

To isolate the trajectory of the population size under a model of accelerating growth we first simulate samples under a simple model where ancestral size, growth time, and current population size are fixed while β varies between 0.1 and 3.5. We use an ancestral size of 20,000 haplotypes which grows over a period of 500 generations to a current effective size of 8×10^6 haplotypes. We also fit models where the population remains at its ancestral size throughout and where the population instantaneously grows to 8,000,000 haplotypes 500 generations in the past.

When conditioning on a fixed current size, a population with accelerating growth initially grows more slowly than a population with constant growth, accelerating to exceed the constant rate only during the very recent past (Figure 3.1). With increasing β, large changes of population size occur more recently, and the population size is closer to its ancestral size for longer. Hence the total variation decreases as β increases (Figure 3.2A). For example, samples of 10,000 haplotypes drawn from a model where $\beta = 0.5$ have on average 54.0 variants per kilobase (kb), 1.4 times greater than the average of 39.6 variants/kb under exponential growth, and 5.9 times greater than the average of 9.1 variants/kb under accelerating growth with $\beta = 3.5$.

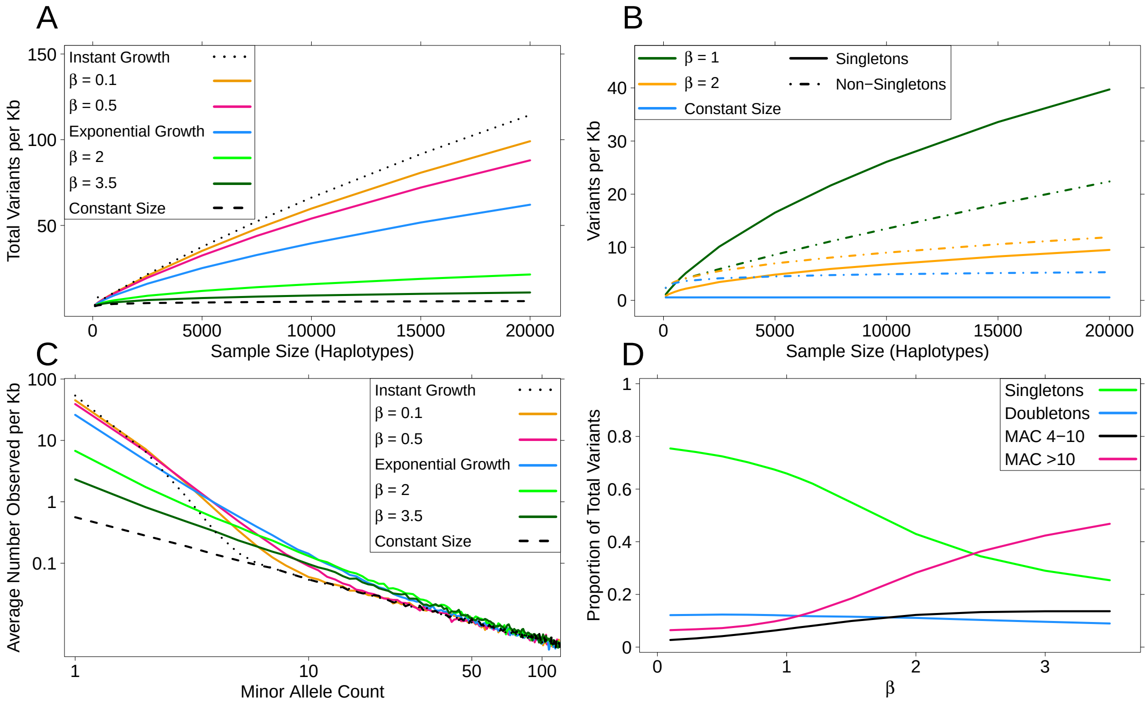

Figure 3.2: Sequence properties for growth models with a current size fixed at 8×10^6 haplotypes reached after 500 generations of growth **(A)** Average number of variant sites per kilobase by sample size. Each colored line represents a different growth model. **(B)** Average number of singleton versus non-singleton variants per kilobase under different growth models as sample size increases. **(C)** Site frequency spectrum showing the average number of variants with a given minor allele count per kilobase of sequence in a sample of 10,000 haplotypes. **(D)** The change in the proportion of all variants that have the given allele counts as growth accelerates in a sample of 10,000 haplotypes.

Rare variants drive the divergence in the total number of variant sites between models (Figure 3.2B). In a sample of 10,000 haplotypes, variants with minor allele count ≤ 10 account for 99% of the reduction in variable sites between models with $\beta = 1$ and $\beta = 2$ and 99.8% of the reduction between models with $\beta = 1$ and $\beta = 3.5$. Correspondingly, the proportion of variants that are very rare also decreases with increasing β (Figure 3.2D). The proportion of variants that are singletons drops from 0.66 to 0.25 between models where $\beta = 1$ and $\beta = 3.5$.

Models with constant population size and models with instantaneous growth bound the family of FTE models with a fixed current population size (Figure 3.2). As β increases, the average frequency spectrum more closely resembles the average frequency spectrum of a model without growth. Conversely, as β approaches 0, the average frequency spectrum more closely resembles the average frequency spectrum from a model with instantaneous growth. For larger values of β, population size is very close to its ancestral value for almost the entire population history. Consequently, genetic sequences simulated under these models share many similarities to those generated under a constant population size model. As β decreases, achieving the current size requires initially fast growth that is slowing over time and sequences simulated under these models bear resemblance to those from instantaneous growth models. Differences between growth models become apparent only with increasing sample size (Figure 3.2, A and B). In a sample of 100 haplotypes, the most extreme models we simulate differ by <1 variant/kb: a model with no growth averages 2.9 variants/kb compared with 3.6 variants/kb with instantaneous growth. And samples with $\beta = 1$ contain <0.1 more variants/kb on average than $\beta = 1.1$, both resulting in an average of 3.4 variants/kb. When we expand sample size to 20,000 haplotypes, instantaneous growth averages 108.5 more variants/kb than a model without growth (114.4 variants/kb vs. 5.9 variants/kb), and the differences between $\beta = 1$ with 62.1 variants/kb and $\beta = 1.1$ with 56.1 variants/ kb become apparent.

3.3.2 Accelerating growth with a variable current population size

In practice, accelerating growth should lead to a larger present day population size than growth at a constant rate. To examine this, we sample 100 to 20,000 haplotypes from populations where the initial growth rate α has been fixed at a value of 100 and β is allowed to vary between exponential growth ($\beta = 1$) and 1.1. Growth accelerates very quickly with increasing β; when we fix the time of growth at 500 generations our parameter settings result in growth from an ancestral size of 20,000 haplotypes to a current size of between 2.4×10^5 and 1.4×10^9 haplotypes, with most of the change in population size occurring in the very recent past.

As growth accelerates, the total quantity of genetic variation in samples quickly increases (Figure 3.3A). For example, in a sample of 10,000 haplotypes there are on average 15.5 variants/kb with exponential growth, 25.5 variants/kb when growth accelerates with $\beta = 1.04$, and 50.8 variants/kb when $\beta = 1.1$. This increase in variation is driven almost entirely by an increase in the quantity of singleton variants (Figure 3.3, B and C), which proportionally come to dominate the frequency spectrum (Figure 3.3D). For example, in 10,000 haplotypes, there are on average 4.9 singletons/kb with exponential growth, 12.3 singletons/kb when growth accelerates with $\beta = 1.04$, and 40.2 singletons/kb when $\beta = 1.1$. Between $\beta = 1$ and $\beta = 1.1$ the proportion of all variants which are singletons rises from 0.32 to 0.79. For non-singleton rare variants, the effect of accelerating growth is relatively small, and in fact the average number observed in a sample can be smaller under faster accelerating models (Figure 3.3C, and Appendix B.0.7). In samples of 10,000 haplotypes the number of variants per kb with minor allele counts between 4 and 10 drops from 2.6 to 2.0 between $\beta = 1$ and $\beta = 1.1$. Additionally, the number of variants with a given minor allele count does not follow a linear trend as growth accelerates. For example, with 10,000 haplotypes, the number of doubletons increases between $\beta = 1$ and $\beta = 1.04$ from 2 to 3.7/kb before decreasing to 3.4/kb when $\beta = 1.1$, and the proportion of all variants that are doubletons follows the same trend ($\beta_1 = 0.14$, $\beta_{1.04} = 0.15$, $\beta_{1.1} = 0.07$) (Figure 3.3D).

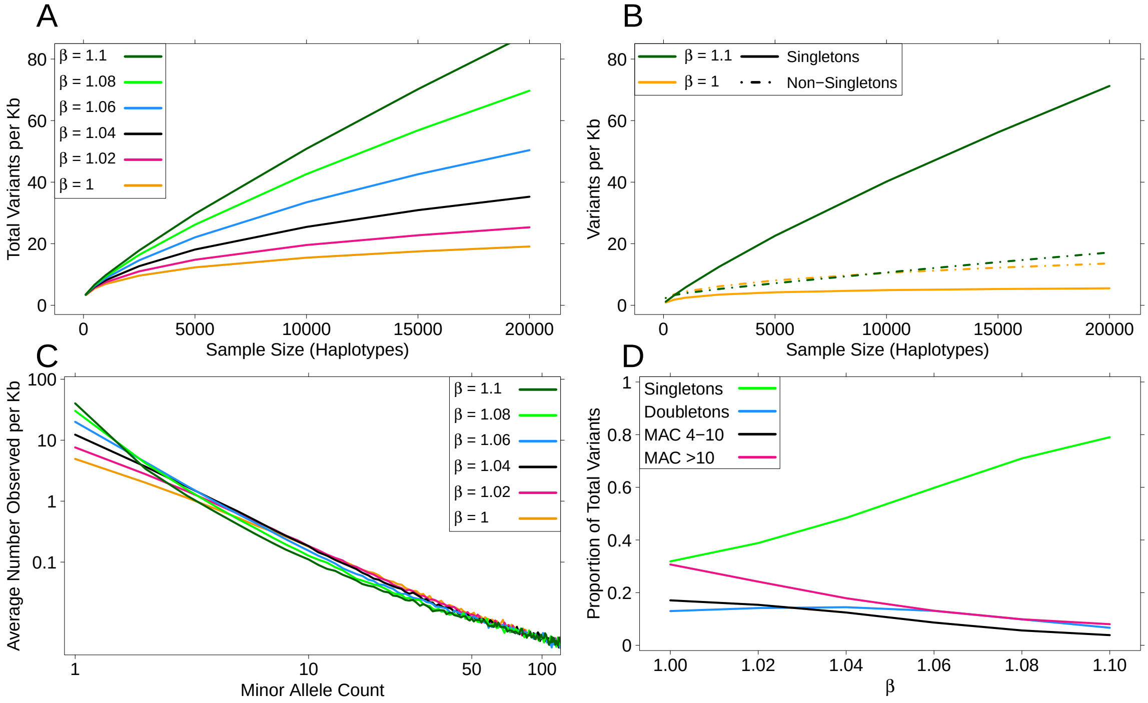

Figure 3.3: Sequence properties for growth models with same initial growth rate $\alpha = 100$ and 500 generations of growth **(A)** Average number of variant sites per kilobase by sample size. Each colored line represents a different growth model. **(B)** Average number of singleton versus non-singleton variants per kilobase as sample size increases for an exponential growth model and a model where $\beta = 1.1$. **(C)** Site frequency spectrum showing the average number of variants with a given minor allele count per kilobase of sequence in a sample of 10,000 haplotypes. **(D)** The change in the proportion of all variants that have the given allele counts as growth accelerates in a sample of 10,000 haplotypes.

As is the case with a fixed current size, differences between growth models become apparent only as sample size increases (Figure 3.3A). With samples of 100 haplotypes, a model where $\beta = 1$ has an average of 3.3 variants/kb, only 0.2 variants/kb smaller than when $\beta = 1.1$. This contrasts sharply with the situation with 10,000 haplotypes, where the difference grows to 35.3 more variants/kb ($\beta_1 = 15.5$ variants/kb, $\beta_{1.1} = 50.8$ variants/kb).

3.3.3 The duration of accelerating growth

To quantify how the duration of population growth impacts our results we perform simulations with samples of 10,000 haplotypes where the initial growth rate α is fixed at 20, β is allowed to vary between 1 and 1.1, and growth duration varies between 100 and 3,000 generations. When we calculate the site frequency spectra for these samples we observe that for all growth durations, the patterns observed in the section above hold: singleton variation increases with accelerating growth, common variation is relatively unaffected, and nonsingleton rare variation follows nonmonotonic patterns where maxima occur at intermediate growth rates. The shorter the duration of growth the less pronounced these patterns are, and the smaller the differences between the growth models. When we compare models with the same rate of acceleration but different growth durations, we find that shorter durations result in frequency spectra with fewer of all variant types, and proportionally fewer singletons and other rare variants (Appendix B.0.8).

Growth time, present population size, and acceleration create a parameter space where sequence characteristics are consistent with multiple sets of growth parameters. With exponential growth, once growth time and growth rate are defined, a population's current size is fixed, and the frequency spectrum can be predicted from these parameters. The addition of an acceleration parameter changes this determinism; the β parameter allows a model where growth time and a have been defined to grow to any current size. Therefore, different combinations of acceleration and current population size can generate the same values of a sequence summary statistic (Figure 3.4). In particular, in a model of accelerated growth the same summary statistic can be generated by a much larger present-day population size compared to a model of constant growth. Assuming an ancestral size of 20,000 haplotypes, 500 generations of growth, and a sample of 10,000 haplotypes, an average of 40 variants/kb of sequence is consistent with both a population that grew to a current size of 8.4×10^6 haplotypes at a constant exponential rate or 18.9×10^6 haplotypes following $\beta = 1.1$. In this example the initial growth rate is slower under accelerating growth (Appendix B.0.9) but the β parameter results in a rapid acceleration and much larger current population size. Note that even though growth follows very different trajectories between these models, the ancestral trees have the same mean total branch length; consequently the same average number of variants is found in the resulting samples.

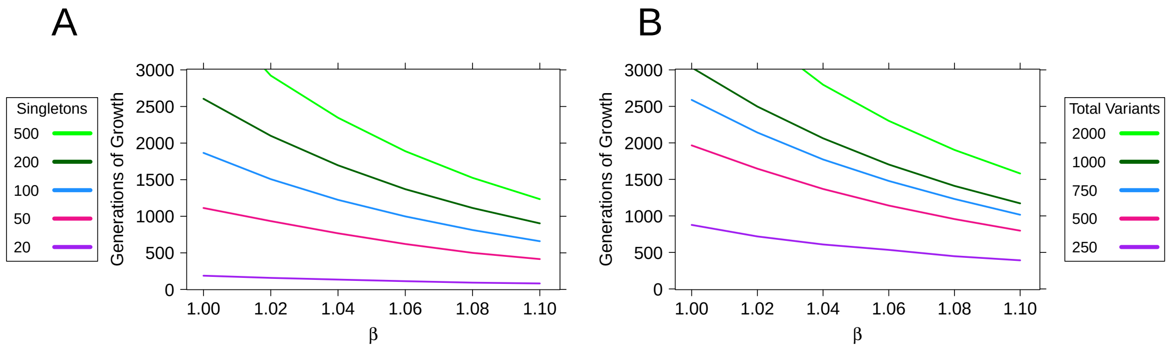

Figure 3.4: Contour plots showing that under different growth parameters summary statistics are consistent with multiple current population sizes. In 30 kilobases of sequence from samples of 10,000 haplotypes, after 500 generations of growth the average quantities of **(A)** singletons and **(B)** all variants from models with larger current population sizes that grew at faster than exponential rates are the same as the average quantities in models with smaller current sizes where the population grew at a slower rate. Each colored line represents a different number of singletons or total variants. Contour lines were generated using linear interpolation over a fine-scale range of α determined using a grid search. For each α in the grid, average parameter values were calculated from 100 independent simulation replicates.

Similarly, different combinations of acceleration and growth time can generate the same values of a sequence summary statistics (Figure 3.5). For example, assuming an ancestral size of 20,000 haplotypes and 30 kb sequenced in a samples of 10,000 haplotypes, we expect about 750 variant sites both from a model of constant growth with rate of $\alpha = 20$ for 2,589 generations and a model where $\alpha = 20$ but growth accelerates with $\beta = 1.1$ for 1,064 generations. Likewise, under the same scenario, we expect 200 singletons both from a model of exponential growth for 2,605 generations and from a model with accelerating growth ($\beta = 1.1$) for 903 generations.

3.3.4 Linkage disequilibrium

To understand the impact of growth on linkage disequilibrium (LD) decay, we simulate 10,000 haplotypes of length 0.1 cM for $\alpha = 100$ and 500 generations of growth and calculate pairwise r^2 and the absolute value of D'. Even though these growth models result in current population sizes that differ by three orders of magnitude, historically their sizes are very similar, and as a result the pattern of LD decay with faster acceleration is sub-

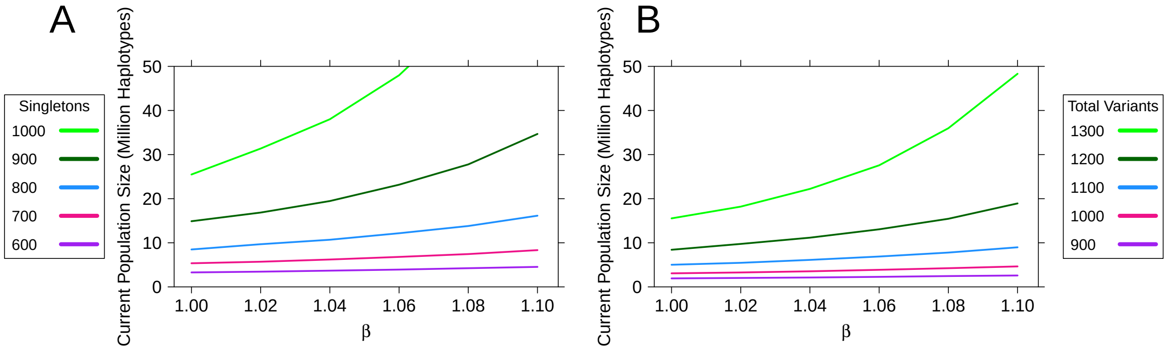

Figure 3.5: Contour plots showing that under different growth parameters summary statistics are consistent with multiple growth durations. In 30 kilobases of sequence from samples of 10,000 haplotypes, after beginning with the same initial growth rate of $\alpha = 20$, the average quantities of **(A)** singletons and **(B)** all variants from models that grow for a shorter amount of time at an accelerated rate are the same as those from models that grow for longer at a slower rate. Each colored line represents a different number of singletons or total variants. Contour lines were generated using linear interpolation over a fine-scale range of α determined using a grid search. Average parameter values were calculated from 100 independent simulation replicates at each grid point.

tle. For variants with minor allele count > 50 the rate of pairwise LD decay increases as β increases and accelerated growth results in larger current population sizes. Between variants with minor allele frequencies between 0.1 and 0.15, at a distance of 40 kb r^2 is on average 0.046 in an exponential growth model, 0.045 for $\beta = 1.04$, and 0.041 when $\beta = 1.1$ (Appendix B.0.10). In the same frequency bin, at 40 kb D' is on average 0.598 in populations with growth at $\beta = 1$ compared to 0.597 for populations with growth at $\beta = 1.04$, and 0.569 when $\beta = 1.1$. In rarer variants, independent of growth model, at a given physical distance the average r^2 is lower and D' is higher (Devlin and Risch, 1995; VanLiere and Rosenberg, 2008). For example, both with $\beta = 1$ and $\beta = 1.1$ the average pairwise r^2 decays below 0.1 in 2 kb for variants with minor allele frequencies between 0.01 and 0.02, compared with 22 kb in variants with minor allele frequencies between 0.1 and 0.15. However, when we look at the relationship between LD decay and acceleration of population growth, it is qualitatively the same across allele counts. Pairwise r^2 in variants with minor allele frequencies between 0.01 and 0.02 from a sample of 10,000 haplotypes and 500 generations of growth with $\alpha = 100$, at 40 kb is on average 0.026 when $\beta = 1$ versus 0.023 when $\beta = 1.1$, and for the same comparison with D' the values are 0.944 for $\beta = 1$ and 0.852 for $\beta = 1.1$. he pattern of faster decay as β increases holds true in smaller sample sizes, but is less pronounced. In samples of 1,000 haplotypes, at 40 kb, variants with minor allele frequencies between 0.01 and 0.02 have an average pairwise r^2 of 0.024

when $\beta = 1$ compared with 0.023 for $\beta = 1.1$, and average D' is 0.018 greater ($D'_{(\beta=1)}$ = 0.955, $D'_{(\beta=1.1)}$ = 0.973).

Varying growth duration between 250 and 750 generations, we find that for any given duration and α, LD decays faster in models where β is greater. When we compare models with the same α and β values but different growth durations we find that shorter growth times, with smaller current sizes, have slower LD decay. For example, in samples of 10,000 haplotypes from models with $\alpha = 100$ and $\beta = 1.06$, the average pairwise r^2 of variants with minor allele frequencies between 0.01 and 0.02 decays below 0.03 at distances of 42 kb, 28 kb, and 24 kb in models with 250, 500, and 750 generations of growth, respectively.

By returning to models with fixed current size we are able to compare pairwise LD decay over a larger range of accelerating growth models. As was described in the section above, with a fixed current size, larger β and hence faster acceleration of growth means a population has been large for a shorter period of time. We find the same pattern of faster LD decay in larger populations with these models; only their relationship with β is reversed. Lower values of β correspond to larger recent population sizes and consequently faster decay in LD (Figure 3.6).

In samples of 10,000 haplotypes from populations which expand from an ancestral size of 20,000 haplotypes to a current size of 8×10^6 haplotypes over 500 generations when we look at variants with minor allele frequencies between 0.1 and 0.15 average pairwise r^2 decays below 0.1 at a distance of 18 kb, 20 kb, and 21 kb for models where $\beta = 0.5, 1, or 3.5$ respectively, and D' falls below 0.75 at distances of 17 kb, 18 kb, and 23 kb. As above, the trends in LD decay are qualitatively the same across variant frequencies, but variants with lower frequencies have lower average pairwise r^2 and the D' between them decays over greater distances. In variants with minor allele frequencies between 0.01 and 0.02 at a distance of 40 kb, pairwise r^2 is 27% greater in a model where $\beta = 3.5$ than when $\beta = 0.5$ ($r^2_{(\beta=0.5)} = 0.022, r^2_{(\beta=3.5)} = 0.028$), and likewise D' is 16% greater ($D'_{(\beta=0.5)} = 0.835, D'_{(\beta=3.5)} = 0.968$).

The range of models we are able to consider with a fixed current size allows us to observe an interesting pattern in variants with minor allele count between 2 and 20. For these variants, D' maintains the same pattern of decay occurring over increasingly long genetic distances, but faster decay in models with smaller β and larger recent population sizes. In contrast, average pairwise r^2 for the rarest variants is lower in models with intermediate rates of growth (Figure 3.6C). For example, at a minor allele count of 10, average r^2 for variants 20kb apart where $\beta = 1$ is 38% lower than when $\beta = 0.5$ and 70% lower than for $\beta = 3.5$ ($r^2_{(0.5)} = 0.013, r^2_{(1)} = 0.008, r^2_{(3.5)} = 0.027$).

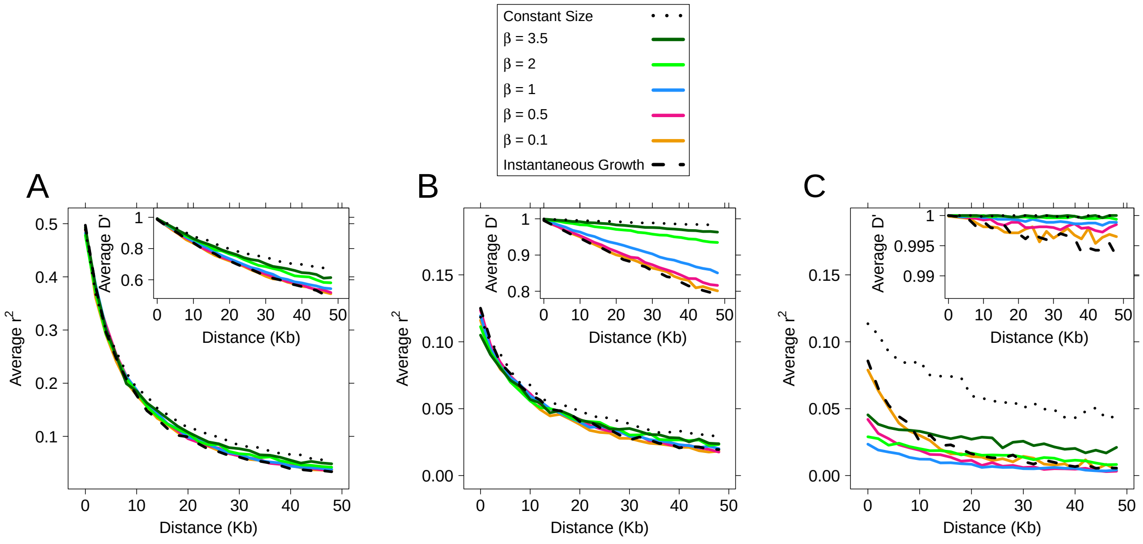

Figure 3.6: Pairwise linkage disequilibrium decay measured by r^2 and D' in samples of 10,000 haplotypes from populations with current population sizes of 8×10^6 haplotypes reached after 500 generations of growth. The panels show LD decay in variants with a sample minor allele frequency of **(A)** 10% to 15%; **(B)** 1% to 2%; **(C)** a fixed allele count of 10. Each colored line represents a different growth model.

3.4 Discussion

Recent sequencing studies suggest a recent acceleration of growth in human populations, as exponential growth models cannot capture the observed excess of singleton variants (Coventry *et al.*, 2010; Tennessen *et al.*, 2012). The frequency spectra in these studies can be better modeled by defining growth as a discontinuous piecewise function with an arbitrarily selected number of segments. However, even the simplest of such models require several additional parameters. Instead, we provide an approach for including an acceleration parameter β into the population growth model, thus obtaining a continuous model that can allow for a wide range of growth trajectories.

When modeling accelerating growth while conditioning on current population size, we observed that the total amount of variation and the amount of rare variation both decrease with increasing β. With a fixed current size, when $\beta > 1$ the initial growth parameter a must be large, and as β increases a correspondingly decreases. Large values of α result in populations that quickly expand at the onset of growth and have a larger size for much of their history. Small α values yield populations with slow initial growth, which achieve only sizes significantly larger than their ancestral size in the recent past. In smaller populations common ancestors are found more quickly, coalescent trees are smaller, and fewer variants are found in samples, explaining our results as β increases. For large β the population history closely resembles a model without growth for all but a small fraction of its history, explaining why models without growth bound our FTE models as β approaches infinity. Likewise, the closer β is to 0, the faster a population's size diverges from its ancestral state, and so more closely resembles a model of instantaneous growth. In this study we assume an ancestral size of 20,000 haplotypes and a period of growth lasting 500 generations; however, the manner in which the demographic history of a population is altered by our growth model, as described in this paragraph, is completely independent of the exact parameter values.

Under an alternate set of conditions, where current size is not fixed and accelerating growth leads to a greater current population size, we observe that the total amount of variation increases with increasing β, and that this increase is driven almost entirely by the amount of singleton variation present in samples. Under these conditions, when we compare a population which grows exponentially to one which grows with $\beta > 1$, initially, growth and population size are determined by the initial growth rate α. Acceleration only manifests itself in the most recent generations. Consequently, for both common and nonsingleton rare variants, the frequency spectra from FTE models closely resemble those from exponentially growing models; however, the very recent acceleration to a much larger current size gives these models an abundance of singletons unmatched by simple exponential models.

Our work shows that the quantity of nonsingleton rare variants changes in a nonmonotonic fashion as β increases. This is best explained by looking at the average population size at the time these variants arise. The number of alleles with a given allele count is a function of the total length of coalescent branches with that number of descendants, a value dependent on population size. In a large population branches are longer; however, coalescent events are also less common, so branches with a given number of descendants occur further in the past than they would in a smaller population. As β increases the population size and mutation age change at different rates. As β increases from low to intermediate values, population size at the times nonsingleton rare variants arise increases even as average mutation age decreases. However, as β increases further, even as average mutation age continues to decrease, the population size at the time nonsingleton rare variants arise decreases more rapidly and is smaller than in the intermediate growth models. Thus, the total length of branches with a given number of descendants is larger at intermediate growth rates and the quantity of variants changes in the observed nonmonotonic manner.

In both the fixed and variable current size contexts we illustrate the importance of sample size for detecting differences between growth models. Small samples contain a limited amount of genetic variation for use in inference. And the inability to distinguish between variants with allele frequencies $<$1/(sample size) means that small samples lack resolution for studying rare and recently arisen variants that contain information about recent demographics. This finding supports the work of Keinan and Clark (2012) in their explanation of why earlier, smaller studies failed to detect evidence of recent massive population growth. It also underpins why more accurate models are needed as study sample sizes increase: in 100 haplotypes failing to include any growth has limited consequences, in 20,000 haplotypes the difference between a model where $\beta = 1$ and $\beta = 1.02$ are substantial.

The addition of the acceleration parameter allows better modeling of very rare variants; the resulting model can "bend" growth trajectories between any ancestral and current size, over any amount of time, flexibility impossible with exponential growth alone. Our work shows how samples from populations which have undergone accelerating growth share the summary statistics of samples from populations with much smaller current sizes achieved via slower growth. Similarly, samples from populations which have grown at an accelerating FTE rate for a shorter length of time share summary characteristics with samples that grew at a slower rate for a longer period of time. Thus, estimates of current population size and the duration of growth estimated from genetic data may differ greatly if accelerating growth is considered.

Our linkage disequilibrium results are also a direct result of how demographic history changes with β. When we look at our models where current population size is allowed to expand with accelerating growth we find that both r^2 and D' decay more quickly in models with larger β, but the effect is small. In larger populations the average time since samples have shared a common ancestor is longer, increasing the likelihood of recombination events that break down D' and r^2. In our models with accelerating growth, it is only over a very brief recent time that the models with larger β have been substantially larger than those with smaller β, and consequently while they show faster decay, it is a very modest difference. However, differences between models are detectable, and this suggests that pairwise LD could be used to help assess the fit of growth models.

The effects of accelerated growth on LD are larger in our models with fixed current size, as these populations substantially differ in size for much longer. We observe that as β increases LD decays more slowly. This pattern holds across the full frequency spectrum for D', and the majority of the frequency spectrum for r^2; however, as minor allele count approaches 1 the pattern for r^2 changes. When we look at the ancestry of a sample, extremely rare variants likely arose only recently and are carried by very few lineages, so the probability of a recombination event occurring between them is small. For these variants every pair of variants has an r^2 of either 1 if they arose on the same lineage or near 0 if they did not. Therefore, average pairwise r^2 becomes equivalent to the probability that the two variants arose on the same lineage. The total quantity of variants with a given minor allele count in a sample is a function of the total length of all lineage branches with exactly that number of descendants, which is itself a function of population size. The probability that two variants with a given minor allele count arise on the same branch is a function of the variability of the branch lengths with exactly that number of descendants. As variability decreases, the probability that two variants arose on the same branch also decreases. The distribution of branch lengths for a given minor allele count reflects how the population size changes while ancestral lineages with that number of descendants exist. For rare variants, in both models with small β and those with large β we observe that the majority of lengths are very short, and the variance is lower than under models with intermediate β. For large β, the population size shrinks so fast that branch lengths corresponding to every variant count are predominantly short. For small β the population size has been large for some time, so lineages rarely find common ancestors during the period of growth, and nonsingleton variants arise mostly in the ancestral population, where all branch lengths are again short. Intermediate β yield a Goldilocks situation, where population sizes result in lineages that have found enough common ancestors that mutations are found in multiple present-day samples, but branch lengths remain relatively long because the time between common ancestors is longer than it would be at the ancestral population size. This observation gives rise to the idea, to be explored in chapter 5, that pairwise r^2 between very rare variants can be used as a tool to estimate recent growth rates.

Presently, the amount of high-quality deep coverage sequencing data available to the research community is increasing by the day. These data offer not only the potential for a better understanding of the genetic underpinnings of a multitude of diseases and traits, but also important insight into the demographic history of our species. The accuracy of these insights will be directly dependent on the models they are based on. The general two-parameter models introduced by this study can substantially revise estimates of current population sizes while simultaneously modeling the excess of very rare variants observed in large human resequencing studies, providing a valuable tool for modeling the complex history of humanity.

CHAPTER 4

Sampling from the distribution of internal branch lengths of a Kingman coalescent

4.1 Introduction

While written history records an inexorable and increasingly rapid expansion of human population sizes, until very recently we lacked genetic evidence of this growth. Large human sequencing studies (Coventry *et al.*, 2010; Nelson *et al.*, 2012; Tennessen *et al.*, 2012; Gazave *et al.*, 2014) have revealed a frequency spectrum characterized by an abundance of extremely rare genetic variation, a signature of recent population growth (Tajima, 1989). This discovery has led to substantial revisions of estimated human population growth rates. Generally, recent large studies have performed inference of growth rates using either of two methods. The first, pioneered by Schaffner *et al.* (2005), relies on coalescent simulations, while the second, from Gutenkunst *et al.* (2009), uses diffusion approximations. The goal of both methods is to estimate the likelihood of observed data under different demographic models. Although widely used (Coventry *et al.*, 2010; Gravel *et al.*, 2011; Tennessen *et al.*, 2012; Nelson *et al.*, 2012; Gazave *et al.*, 2014), the methods are not without substantial drawbacks: the diffusion approach is incapable of using any summary statistics other than the frequency spectrum to estimate likelihoods, while the coalescent approach is computationally burdensome, especially with large samples.

Here we propose a novel method that analytically realizes portions of a sample's genealogy without modeling the entire genealogy. Our method allow us to directly sample external and internal branches from the Kingman coalescent (Kingman, 1982a; Hudson, 1983), providing us with the genealogical features necessary for demographic inference without the computational burden of full coalescent simulations. Additionally, this method retains the individual branch lengths necessary for calculating the probability of observed patterns of pairwise r^2 as outlined in appendix C.0.12, an advantage over the diffusion approach of Gutenkunst *et al.* (2009).

The Kingman coalescent models the ancestry of a sample backwards in time. The model follows chromosomes back through common ancestors; when two ancestries in a sample reach a common ancestor they join into a single shared ancestry. The process continues until the most recent common ancestor of the entire sample is reached. Mutations are modeled as occurring at a constant rate across the resulting tree-shaped genealogy, with the number of mutations on any given branch a function of branch length. Consequently, longer branches on a tree, corresponding to longer waiting times between common ancestors, have more mutation events. Mutations appear as a variants in the final sample on the chromosomes which descend from the branch where the mutation occurred. The waiting times between common ancestors are a function of the size of the population from which the sample was drawn, with greater times in larger populations. Thus, when a population has recently grown, we expect the genealogy of samples to have proportionally longer branches near the terminal nodes, and correspondingly more mutation events along the lower branches. The lower branches of a coalescent tree have relatively few descendants in the final sample, so mutations that occur along them appear at very low minor allele counts, hence an abundance of rare variation is a signal of recent growth.

While the lengths of coalescent branches are a function of population size, the topology of a tree is not. For a given sample size, the probability of observing a given tree topology are identical for every model of varying population size (Kingman, 1982b). In the traditional coalescent, tree topology and length are simulated together for every genealogical realization. Our method separates the topology from the generation of individual branch lengths, allowing us to calculate and store the probabilities of the number of branches with a given number of descendants in a topology, as well as the probabilities of branches starting and ending at specific coalescent events. We can then reuse these calculations for any demographic history from which we are interested in sampling branch lengths.

Once we have the probabilities of different topologies for a sample size, we can generate a realized genealogy under any demographic model by sampling the lengths of individual branches. In constant size populations, the branch lengths are a sum of independent exponential variables, and an exact probability distribution can be written and sampled from. For variable size populations, we show it is possible to first calculate expected times between coalescent events, and use these to generate branch lengths. By generating distributions of individual branches we can calculate the likelihood of observed frequency spectra, patterns of linkage disequilibrium, or any other statistic found to contain information about recent demography. In addition to decoupling tree topology from branch lengths, out method generates branch length distributions for each minor allele count (MAC) separately. As discussed above, information about recent growth is contained in the branches with few descendants in the final sample, the branches occurring low on a genealogical tree. Thus, our method allows us to generate individual branch lengths for MAC below a threshold, facilitating inference of recent growth, while summarizing the rest of the genealogy with a single value, saving the computational effort required to simulate the portions of the genealogy unrelated to the research questions we are interested in.

Previous work on the distribution of internal branches of the Kingman coalescent was focused on their summed length (Fu and Li, 1993; Kersting and Stanciu, 2013). Summed length was of interest because the number of mutations observed in a sample with a given MAC is a function of the total length of branches with a number of descendants equal to the MAC. Fu and Li (1993) presented the expectation and variance of the total summed length of both external and internal branches along a Kingman coalescent without recombination. Recently, Kersting and Stanciu (2013) published the asymptotic distribution of the summed length of all branches with the same number of descendants. With our interest in finite sample sizes and individual branch lengths, our work more closely builds on the findings of Rosenberg (2006), who derived the expectation and variance for the number of internal branches with a given number of descendants in a sample's ancestry.

With our method we calculate the probability of different topological features for a genealogy, we then sample from these probabilities without simulating the entire genealogy. First, we use a recursive algorithm to expand on the results of Rosenberg (2006) and calculate the exact probabilities for the number of branches with a given number of descendants in a sample's genealogy. Our second set of equations allow us to calculate the probability that a branch with a given number of descendants arose at a specific coalescent event, and conditional on its origin, the probability of the branch ending at a later specific event. For a constant size model, by combining our topological probabilities with the exponentially distributed waiting times between coalescent events, we derive an exact probability distribution function for internal and external branch lengths. For demographic models with varying past sizes we substitute expected waiting times in place of the exponential random variables to generate individual branch lengths. We show that using an accept-reject algorithm we are able to directly sample from the branch length distribution in the case of the constant size population. We compare the output of our method with coalescent simulations to show that summary statistics from the branch length distributions are the same between methods. Finally, we discuss how our method provides an important step towards the estimation of demographic likelihoods using an expanded range of summary statistics without resorting to full coalescent simulations.

4.2 Methods

4.2.1 Realizing a genealogy through the sampling of branch lengths from the Kingman coalescent

Our method for generating the realized genealogy of a sample proceeds through several stages. Before beginning, we select a minor allele count (MAC) threshold. Below this threshold we generate individual branch lengths, above the threshold we summarize the entire genealogy using a single length. For every MAC below our threshold, we first sample the number of individual branches seen in the realized genealogy, according to equation 4.1 in section 4.2.2. Then, for each individual branch we sample a beginning and ending coalescent event. The probability distribution for beginning and ending events are given by the equations in sections 4.2.3 and 4.2.4. The beginning and ending coalescent events define a series of random variables: the waiting times for all the coalescent events between the origination and termination of the branch. In the case of a constant size population, these waiting times are independent exponential random variables, and their sum follows a hypoexponential distribution with a rate vector defined by the rates of the individual exponential random variables. In section 4.2.5 we give an explicit formula for the distribution of branch lengths in a constant size population. In a variable size population, following the method of Donnelly and Tavaré (1995), the waiting times are no longer independent, and an exact calculation of branch length requires integration over all possible previous times for each term in the series, problematic for series with dozens or hundreds of terms. To expand our method to variable size populations we therefore estimate the expected waiting times under a model of interest, and the sum these values as determined by the sampled beginning and ending coalescent events. By combining our distributions of individual branches with the single value that summarizes the total length of the genealogy above our MAC threshold, we have the features of a genealogy necessary for inference, without simulating the entire ancestry of a sample.

4.2.2 The number of branches with j descendants in a genealogy

The distribution of the number of branches with j descendants in a sample's ancestry can be calculated recursively. A genealogy of size n can be divided $\lfloor n/2 \rfloor$ ways, where $\lfloor\rfloor$ denotes the floor function, and observing x branches with j descendants in the full sample, n, is equivalent to observing $x-y$ and y branches with j descendants in the divided genealogies. We define $P_{n,j}(x)$ as the probability of observing x branches with j descendants in a sample of size n, then

$$P_{n,j}(x) = \begin{cases} 0 & \text{for } n < j \\ 1 & \text{for } n = j \\ \sum\limits_{i \leq \frac{n}{2}} \frac{2 - I[i = \frac{n}{2}]}{n-1} \sum\limits_{y=1}^{x-1} P_{i,j}(x-y) P_{n-i,j}(y) & \text{for } n > j \end{cases} \tag{4.1}$$

4.2.3 The probability that a branch with j descendants originated at coalescent event $n - k + 1$

Given a branch with j descendants, we are interested in the probability it originated at a specific coalescent event in a sample's ancestry. To calculate this value, we first calculate the related probability: at a given coalescent event what is the probability a branch with j descendants forms? At an arbitrary coalescent event, where two ancestral lines are chosen to coalesce and k ancestral lines become $k - 1$ lines (event $n - k + 1$), a branch with j descendants is created if the number of descendants of the two lines coalescing sum to j. The number of ways such a coalescence can occur is a function of the number of descendants of the k ancestral lines present at the coalescent event. Thus, to calculate the probability of a branch with j descendants forming, we must sum over all possible genealogical histories preceding even $n - k + 1$, corresponding to all the different ways of dividing $n - k$ descendants among k ancestral lines. We calculate the probability of each history using the formula presented in Kingman (1982a) and then multiply it by the number of ways a branch with j descendants can form from the distribution of descendants in the history. Define λ_d as the number of branches with d descendants in the final sample for a given history. Then, the probability that the $n - k + 1$ event gives rise to a branch with j descendants in the final sample is

$$P(\text{Branch formed at event } n - k + 1 \text{ has } j \text{ descendants}) =$$
$$\sum_{\vec{\lambda}} \frac{2(n-k)!(k-1)!(k-2)!}{(n-1)! \prod_{d=1}^{n-k+1} \lambda_d!} \left[\binom{\lambda_{\frac{j}{2}}}{2} I[j \text{ is even}] + \sum_{m=1}^{\frac{j-1}{2}} \binom{\lambda_{j-m}}{1} \binom{\lambda_m}{1} \right] \tag{4.2}$$

Here $\vec{\lambda} = (\lambda_d : 1 \leq d \leq n - k + 1, \sum\limits_{d=1}^{n-k+1} \lambda_d = k, \sum\limits_{d=1}^{n-k+1} d\lambda_d = n)$ are the constraints that define a legitimate history, one with the correct number of lines and ancestors present at the $n - k + 1$ event. In equation 4.2, the term outside the brackets is the probability of the history; inside the brackets is the number of ways a branch with j descendants can form from the ancestral lines present in the history. Independent of our work, Spouge (2014) used a formulation very similar to equation 4.2 as a transition probability in a Markov chain designed to calculate the probability that the most recent common ancestor of a nested subsample has additional descendants in the larger sample from which the subsample was drawn.

Equation 4.2 gives us the conditional probability that at event $n-k+1$ the branch that originates has j descendants. However, for our sampling algorithm we are interested in the opposite, conditional on observing a branch with j descendants, what is the probability the branch arose at event $n-k+1$?

$$P(\text{Branch began at event } n-k+1|\text{Branch has MAC } j) = \frac{P(\text{Branch has MAC } j|\text{Event } n-k+1)}{\sum\limits_{i=1}^{n-1} P(\text{Branch has MAC } j|\text{Event } i)} \tag{4.3}$$

Using equation 4.2 this can be written as

$$P_{start,j}(n-k+1) = \frac{\sum\limits_{\vec{\lambda}} \frac{2(n-k)!(k-1)!(k-2)!}{(n-1)!\prod_{d=1}^{n-k+1}\lambda_d!}[\binom{\lambda_{\frac{j}{2}}}{2} I[j \text{ is even}] + \sum\limits_{m=1}^{\frac{j-1}{2}} \binom{\lambda_{j-m}}{1}\binom{\lambda_m}{1}]}{\sum\limits_{h=3}^{n-j+1} \sum\limits_{\vec{\lambda}} \frac{2(n-h)!(h-1)!(h-2)!}{(n-1)!\prod\limits_{d=1}^{n-h+1}\lambda_d!}[\binom{\lambda_{\frac{j}{2}}}{2} I[j \text{ is even}] + \sum\limits_{m=1}^{\frac{j-1}{2}} \binom{\lambda_{j-m}}{1}\binom{\lambda_m}{1}]}. \tag{4.4}$$

4.2.4 The conditional probability that a branch with j descendants ends at coalescent event $n-b+1$

The conditional probability that a branch originating at event $n-k+1$ finds a common ancestor and ends at event $n-b+1$ is straightforward, and similar to a geometric probability. At each coalescence the probability that the branch ends is a function of the number of remaining ancestral lines:

$$P_{end,j}(n-b+1) = \frac{2}{b}\prod_{a=b+1}^{k-1}(1-\frac{2}{a}) \tag{4.5}$$

4.2.5 A probability distribution function for coalescent branches in a model with constant population size

The probability that a coalescent tree branch has length l is the product of three probabilities: the probability the branch begins at event $n-k+1$, conditional on its starting event the probability that the branch ends at event $n-b+1$, and then conditional on its starting and ending events the probability that the sum of the exponential random variables comprising its length sum to l:

$$P(L_j = l) = \sum_{Start}\sum_{End} P(Length = l|Start, End)P(End|Start)P(Start). \tag{4.6}$$

$P(End|Start)$ and $P(Start)$ are given by 4.5 and 4.4, respectively. The sum of exponential random variables with different rates is a hypoexponential distribution. In the case of coalescent times from a constant size population, the exponential rates are determined by the number of ancestral lines remaining. Referring to 4.4 as $P_{Start,j}(n-k+1)$, we can write the exact distribution of branch lengths with j descendants as

$$P(L_j = l) =$$
$$\sum_{k=3}^{n} P_{Start,j}(n-k+1) \sum_{b=2}^{k-1} \left[\left(\frac{2}{b} \prod_{s=b+1}^{k-1} (1-\frac{2}{s}) \right) \sum_{z=b}^{k-1} \frac{e^{-\binom{z}{2}l} \prod_{v=b}^{z} \binom{v}{2}}{\prod_{v=b, v\neq z}^{k-1} \left(\binom{v}{2} - \binom{z}{2} \right)} \right] \tag{4.7}$$

With expected value

$$E(L_j) = \sum_{k=3}^{n} P_{Start,j}(n-k+1) \sum_{b=2}^{k-1} \left[\left[\frac{2}{b} \prod_{s=b+1}^{k-1} (1-\frac{2}{s}) \right] \sum_{z=b}^{k-1} \frac{2}{z(z-1)} \right] \tag{4.8}$$

4.2.6 Implementation and simulations

By combining the methods of the previous sections it is possible to generate the features of a genealogy necessary for demographic inference without performing a full coalescent simulation. Equations 4.1, 4.4, and 4.5 give the probabilities of a tree topology. These values are independent of the branch lengths, and can be calculated, stored, and reused for samples with the same size. For a constant size population, branch lengths can be sampled directly from equation 4.7 using an accept-reject algorithm (Robert and Casella, 2004) with an exponential proposal distribution. For variable size populations, branch lengths are the sum of expected waiting times. We estimate these times as the average value of 100,000 simulations, using the method of Donnelly and Tavaré (1995), where times from a constant size model are "shifted" to the correct values for a model of changing population size (Section 3.2.2). In addition to our individual branch lengths, we generate a length for the genealogy above our MAC threshold. The proposed likelihoods of Boyko *et al.* (2008), Keinan *et al.* (2007), and appendix C.0.12 all require the total length of the genealogy, so an estimate of time remaining in the ancestry above our threshold is essential. Originally, we calculated the time remaining in a genealogy using the sum of the expected total lengths for all branches above our MAC threshold ($\sum_{r=C_{MAC}+1}^{2} 2/r$, where C_{MAC} is our threshold). However, the distribution of summed total branch lengths is highly skewed, and has a large variance during the period of a sample's genealogy when few ancestral lines remain. Using the expected value significantly changed our inference results relative to coalescent simulations. To overcome this, we again use the method of Donnelly and Tavaré (1995), simulating realizations of the total time in a sample's genealogy above our MAC threshold, dependent on past population sizes. Instead of averaging across realizations to estimate the expected value, we instead store the individual realizations and sample from them, combining this value with the summed total length of our sampled branches below the MAC threshold to get the total tree length for a realized genealogy.

We compare our method with the full coalescent using simulations. For individual branch lengths, we generate values from equation 4.7 using an accept-reject algorithm with an exponential proposal distribution, and compare these values with branch lengths from coalescent simulations with a constant size population. We also use our method with expected waiting times in place of exact branch lengths and compare the results with coalescent simulations. We compare the number of branches in each genealogy, the summed length of branches, and the inter-branch variances between genealogies realized using our method versus the coalescent.

4.3 Results

For constant size populations we can explicitly write the probability distribution function for coalescent branch lengths (equation 4.7). With an accept-reject algorithm, using an exponential proposal distribution, we can then sample individual coalescent branch lengths. Using our method we sample 3 million branch lengths from the genealogies of a current sample of 50 haploid individuals; 1 million each for branches with 3, 6, or 9 descendants.

For comparison, we generate 1 million independent coalescent trees and record the length of all observed branches with 3, 6, or 9 descendants. In table 4.1 we compare the distribution of lengths, in coalescent units, between the methods.

		Proportion of Branches						
Descendants	Method	< 0.001	< 0.005	< 0.01	< 0.05	< 0.1	< 0.5	< 1
3	Analytic	0.0280	0.127	0.230	0.638	0.806	0.978	0.993
	Coalescent	0.0280	0.128	0.231	0.639	0.807	0.978	0.993
6	Analytic	0.0170	0.0800	0.149	0.486	0.674	0.945	0.981
	Coalescent	0.0168	0.0795	0.149	0.486	0.674	0.945	0.981
9	Analytic	0.0119	0.0566	0.108	0.388	0.573	0.907	0.965
	Coalescent	0.0120	0.0567	0.107	0.387	0.573	0.907	0.965

Table 4.1: The distribution of branch lengths generated using our analytic formulas compared with those generated via coalescent simulations in a sample of 50 haploid individuals. Lengths are given in coalescent units of $2N$ generations. For each analytic row 1 million random variables were sampled according to an accept-reject algorithm with an exponential(2) proposal distribution. For each coalescent row, 1 million independent simulations were run, and the length of all branches with the specified number of descendants was recorded.

For a further comparison we calculate summary statistics for the set of branches with the same number of descendants, again comparing the output of our method with full coalescent simulations. Instead of sampling exact lengths, we use the sum of expected waiting times. We realize 1 million genealogies for a sample of 50 haploids, drawn from a population of 30,000 haploids, and for branches with 3, 6, or 9 descendants we recorded the number of branches in the genealogy, the summed length of the branches, and the variance in lengths (Figure 4.1A).

We find that the number of branches we observe in genealogies realized with our method and with coalescent simulations match very closely. For branches with 3 descendants, we observe between 1 and 15 branches in $> 99.99\%$ of genealogies with both methods, and the probability of observing each value between 1 and 15 never differs by more than 0.12%. For branches with 6 or 9 descendants the methods also produce very similar results, with the differences between the probabilities of a given number of branches never exceeding 0.10%. When we investigate the summed length of branches and the variance between branch lengths, we again find that the methods give similar results, however, there is one distinct difference. Our method, relying on the sum of expected values, does not produce genealogies with extremely long waiting times like the coalescent (Figure 4.1B). For branches with 3 descendants, using our method, we never observe a summed branch

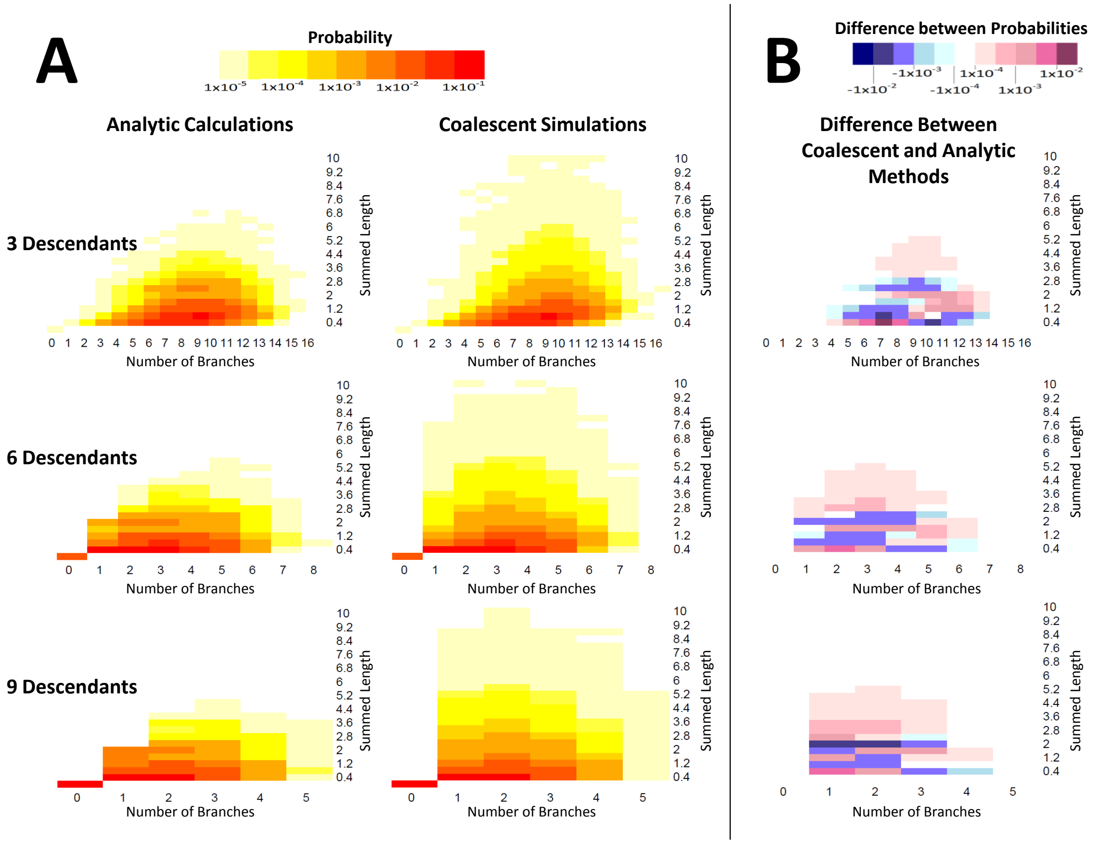

Figure 4.1: Using our analytic method and coalescent simulations we calculate summary statistics across 1 million realized genealogies for a sample with 50 haploids drawn from a population of 30,000 haploids. **(A)** For branches with 3, 6, or 9 descendants in the final sample we calculate the two-dimensional probability distribution for branch number and summed branch length using our method and coalescent simulations. **(B)** The differences between the distributions presented in **A**, with the probability of observing a given variable pair in our method subtracted from the probability in the coalescent simulations. Blue represents a higher probability using our method, pink a higher probability using the coalescent.

length $>$ 7 coalescent units, and $< 0.05\%$ have a summed length $>$ 4 coalescent units. With coalescent simulations 0.02% of simulations have summed lengths $>$ 7 coalescent units, and 0.34% are $>$ 4 coalescent units. In these extreme cases, the variance between branch lengths is often great, and we observe a variance $>$ 1 in 0.52% of coalescent simulations compared with $< 0.002\%$ of analytic calculations. While this pattern is real, it is important to note how small these values are. Outside of the tails of the distributions, the two methods produce very similar results. For example, using our calculations 14.99% of realizations have a summed length of branches with 3 descendants between 0.8 and 1.2 coalescent units, while 14.95% of coalescent simulations fall in this range.

4.4 Discussion

To create a computationally efficient method for sampling internal branch lengths from the Kingman coalescent we derived a series of analytic calculations. In the case of a constant size population, we can explicitly write the probability distribution function for branch lengths, and sample from it using an accept-reject algorithm. When we compare the output of this algorithm to coalescent simulations we find that the methods result in the same distributions of branch lengths. For populations with variable sizes we estimate expected branch lengths in place of an exact length distribution. When we compare summary statistics between this approach and coalescent simulations we find they are generally very similar. The number of branches with a specific number of descendants is nearly identical between methods. However, coalescent simulations have a greater probability of generating extremely long branches, and correspondingly we find more extremely long summed branch lengths and inter-branch variances in coalescent simulations.

Thinking about the use of our method in an inference setting, the lack of extreme realizations with our analytic approach should not alter conclusions, at least for inference based exclusively on the frequency spectrum. Outlying realizations add noise to likelihood estimation, but inference is performed using hundreds or thousands of realizations, and averaged across multiple loci, each with their own genealogical history. Inference is concerned with finding the model which on average looks the closest to the observed data, and that should remain the same between both our method and coalescent simulations. In the context of including linkage disequilibrium measures into inference, a goal of future research, the variance between branches is important, and the impact of using expected values is still an open question.

At the core of our analytic method is an assumption that branch lengths are independent. While branches generated using the coalescent violate this assumption, for branches during periods where many lines remain on the coalescent tree the correlation between branches is negligible. It is the portions of the coalescent tree with the smallest amounts of correlation between branch lengths along with we are applying our method to sample branches. Consequently, for most cases our assumption of independence should not alter results. However, where correlation between branches rises above negligible levels in the coalescent, it will result in smaller variances between branch lengths. As with our use of expected waiting times discussed above, inference based on the frequency spectrum is based on expected values and should not be effected by correlation between branch lengths, but inference using pairwise r^2 may be altered by the differences in variance. It is interesting to note that in our model, the use of expected waiting times and the assumption of independence change inter-branch variance in different directions; how these effects interact is another topic for further investigation.

We designed our method to be more efficient than full coalescent simulations in two key ways. The first results from the fact that coalescent topology is independent of demography, and for a given sample size, with our method the probabilities of different topology structures can be calculated, stored, and then re-used repeatedly. This limits the computational cost of performing likelihood estimation over many demographic models. The second efficiency gain from our method is the ability to summarize ancestry above a MAC threshold with an easily calculated value. This calculation would remove the need to model a large portion of each genealogy, a significant advantage, particularly as sample size expands. Currently, due to the large variances on times between coalescent events when few ancestral lines remain, we have been unable to directly generate the length above our threshold. Our workaround, simulating full sample genealogies to calculate a distribution of times remaining above the MAC threshold, is less optimal. Developing a way to directly sample from the distribution of these times would be a substantial improvement, however, in the case of models with variable past sizes this is particularly challenging due to the lack of independence between inter-coalescence times.

The applicability of our method to demographic inference is contingent on scaling it to large enough samples that power exists to detect recent growth. Critical to our sampling approach is calculation of the probability that branches start and end at specific coalescent events. As mentioned in the methods section, the calculation of the probability a branch with j descendants starts at a given coalescent event requires summing over all possible histories preceding the event. This value increases incredibly quickly with sample size. For event $n - k + 1$ the number of possible histories is the total way to partition $n - k$ items, at $n = 250$ and $k = 50$ this value is over 2.1×10^{14}, and this sum needs to calculated at every event. Even with an importance sampling approach that limits the space of considered histories to only those with a non-zero probability of giving rise to a branch with j descendants, calculation of $P(start)$ is too computationally burdensome for sample sizes above ≈ 100. In the future, an efficient method to calculate or estimate $P(start)$ would greatly expand the capabilities of the work presented here.

A driving motivation for our development of the sampling method presented in this chapter was to facilitate incorporation of pairwise r^2 into demographic estimation. In appendix C.0.12 we present a novel likelihood the includes the probability of a sample's observed pattern of pairwise r^2 under a given demographic likelihood. The likelihood can be calculated using either coalescent simulations or the method presented here, and research is ongoing into the potential power gains it can provide versus likelihood estimation made using the frequency spectrum exclusively.

A global picture of the varying rates at which human populations have grown is going to require inference to be made on many large genetic samples. The current widely applied methods, performing multitudinous coalescent simulations or using diffusion approximations, can accomplish such inference, but leave substantial room for improvement. Here we have developed a novel method designed to be more efficient than coalescent simulations, while not limiting the possible summary statistics used for likelihood estimation. While our method has issues with scalability, with further work it has the potential to be an important building block in our effort to understand the history of our species.

CHAPTER 5

Discussion

From the search for a population level understanding of disease and demography to the personal level of clinical diagnoses, genetic sequencing has become an integral tool in all areas of human genetics. One major discovery made possible through the sequencing of large samples is the great quantity of rare genetic variation present in human populations. In this dissertation, we have presented three novel methods aimed at leveraging that abundance to further our knowledge of both complex disease and human history.

In chapter 2 we evaluated the robustness of group-based association tests to population stratification using the joint site frequency spectrum of samples from several European populations. We found that the tests clustered into two classes which differed in their susceptibility to p-value inflation caused by population structure. Using the statistics of allele sharing and weighted symmetry, we quantified two types of rare variant population structure, and showed that each is correlated with inflation of p-values in one of the group-based test classes.

Genomic control (Devlin and Roeder, 1999) and principal components analysis (Price *et al.*, 2006) are broadly applicable methods for correcting p-value inflation due to population stratification in single variant association tests. Such general remedies have not been forthcoming for group-based association tests. Correction is much more challenging for group-based statistics, which are composed of variants that each have their own unique population history. While correction of population stratification remains problematic for group-based tests, its existence in genome-wide data should be easily recognized using QQ-plots. For candidate gene studies, or other studies with insufficient data to detect p-value inflation using a QQ-plot, our work suggests an additional method for detecting stratification. To minimize the number of null variants in a statistic, annotation based filters are used to select variants for inclusion. Variants excluded by these filters will still contain signal for underlying population structure, however, we expected them to be mostly unassociated with phenotypes of interest. Thus, performing association testing using group-based statistics comprised of excluded variants should detect inflation due to structure, and is applicable in datasets of any size.

Generally, the application of group-based association statistics to real datasets has resulted in too few rather than too many associations. The most compelling explanation for the absence of group-based findings in current studies is a lack of power. Even though group-based tests often have substantially more power than single variant tests to detect genotype-phenotype associations in the context of rare variants, in reality that power rests on a precarious balance between including enough variants, without including too many non-causal variants. Recent studies are using several approaches to address the lack of power and findings in the context of rare variants. New methods, building on the statistics we reviewed in chapter 2, have been proposed to optimize the power of group-based tests (Chen *et al.*, 2012; Lee *et al.*, 2012). For many complex diseases, large meta-analyses that combined the results from multiple smaller studies yielded extensive novel single variant associations (Teslovich *et al.*, 2010; Peden *et al.*, 2011). Methods to allow meta-analysis of group-based tests is an active area of research (Lee *et al.*, 2013; Feng *et al.*, 2014), likely to be widely adopted by the consortia that successfully collaborated to find common variants associated with disease. Study design is another area where efforts are being made to increase the power to detect rare variant associations. Family based studies were critical for early genetic discoveries (Tsui *et al.*, 1985; MacDonald *et al.*, 1993), and with their potential for enrichment of rare variants (Peng *et al.*, 2013), interest has been renewed in studies of related samples. For example, as part of the T2D-GENES consortium the San Antonio Family Heart Study (SAFHS) (Mitchell *et al.*, 1996) and San Antonio Family Diabetes/Gall Bladder Study (Hunt *et al.*, 2005) have strategically selected samples from 20 large pedigrees and performed whole genome sequencing with the goal of maximizing their power to detect rare variant associations with type 2 diabetes and related metabolic phenotypes (Matt Zawistowski, personal communication, July 2, 2014).

While it is possible that some of the initial excitement over rare variant's potential to explain the heritability of complex traits was misplaced (Simons *et al.*, 2014), there have been some exciting discoveries made (Guerreiro *et al.*, 2012; Cruchaga *et al.*, 2014; Ortega *et al.*, 2014). With the growing quantity of data available, and with more powerful statistics and study designs being used, it is nearly certain that additional significant findings are on the horizon. As studies grow more powerful, distinguishing true phenotype associations from spurious findings caused by population structure will be critical, and the methods introduced in chapter 2 to quantify population structure and predict its impact on testing results will likewise become increasingly valuable.

In chapter 3 we moved from disease genetics to population demography and introduced a two-parameter model of accelerating, faster than exponential population growth. We showed that the additional "acceleration" parameter, β, in our models made possible the generation of samples containing large quantities of very rare variants without inflating the quantities of more common variants, and thus overcame the inadequacies of simple exponential growth models as observed in Coventry *et al.* (2010).

While we showed the potential of our models in chapter 3, an important next step is showing that they can accurately reflect human history. This is a challenging problem, and is likely to require very large samples. While α and β are theoretically identifiable, there are an infinite number of pairs for every possible current size and growth duration, making fine scale distinction more difficult than with a single exponential growth parameter. A complicating factor is that accurate inference of the rate of population growth requires genetic regions as free of selection as possible. Current practice holds that the regions least likely to be under selection are those containing few genes or known regulatory elements, regions of little interest to functional and disease association studies generating the majority of publicly available human sequence data. While Gazave *et al.* (2014) were able to secure funding for the sequencing of a sample specifically for inference of population growth rates, it was modest in size and scope, limiting their potential findings. The best possibility for the creation of large datasets with coverage of putatively neutral regions of the genome comes from the continuing decline in the price of whole genome sequencing. The sequencing of whole genomes in large samples by well funded disease association studies will have the added benefit of making accurate inferences about the rate of population growth possible.

Efforts to expand on the findings of chapter 3 will need to be cognizant of the limitations of the coalescent model with respect to large samples. A key assumption of the coalescent model is that the sample whose ancestry is being simulated is significantly smaller than the overall population it is drawn from. We show in appendix B.0.5 that for the sample and population sizes considered in this dissertation, any distortions to the frequency spectrum caused by violations of this assumption are likely to be mild. However, if in the future larger samples are available for inference, it will be important to verify that the coalescent assumptions remain reasonable. In the event that the coalescent is unable to accurately model a genealogy due to the sample's size, a strategy like that of Bhaskar *et al.* (2014) can be employed. Bhaskar *et al.* (2014) presents a hybrid model, making use of a computationally intensive discrete time process, one that allows more than two samples to coalesce in a generation and more than two samples to coalesce to the same ancestor, during the period when the ratio of sample size to population size is too great for accurate approximation via the coalescent.

In chapter 4 we proposed a series of analytic equations that allowed us to sample the lengths of internal and external branches from realizations of a sample's genealogy. We showed that the results of our method were comparable with those of full coalescent simulations, with room for substantial improvements in computational efficiency. As currently implemented, these efficiency gains remain largely theoretical, especially for large sample sizes. However, with further research, hopefully uncovering a way to quickly estimate the conditional probability of a branch with a given number of descendants starting at a specific coalescent event (equation 4.5) and an accurate way of sampling from the distribution of time remaining in a genealogy above a given threshold, our calculations have the potential to prove very useful in future inference endeavors.

In addition to our analytic calculations, deriving a method for incorporating linkage disequilibrium information into demographic inference is also a major goal of our current research. In the appendix for chapter 4 we develop a framework for likelihood estimation that includes the probability of observed pairwise r^2. Work is ongoing to determine how much additional power our likelihood provides for distinguishing between the increasingly complicated demographic models necessary for capturing a recent acceleration in population growth.

The recent rate of discovery in human genetics has been staggering, and has generated a technology fueled avalanche of data. Pulling further findings from this data will require robust and powerful methods of analysis. In this dissertation we have presented three new statistical approaches, each aimed at harnessing the abundance of rare variation in the human genome to expand our understanding of complex disease and human demography.

APPENDIX A

Appendix for Chapter 2

A.0.1 An analytic model of the JSFS

In our research we also develop a way to generate analytic JSFS, and perform simulations with them parallel to our empirical JSFS. The qualitative results from these models are the same as those shown in Figures 2.1, 2.3, 2.4, 2.5.

Analytic JSFS are created by calculating the expected JSFS, $\Phi = \{\phi(i,j)|f_1, f_2\}$, between two populations with divergence parameters f_1 and f_2. Motivated by the work of Balding and Nichols (1995) we use a hierarchical beta model to define the marginal probabilities of allele frequencies for two populations (labeled 1 and 2). For a given single nucleotide variant, let $p_A \sim Beta(\alpha, \beta)$ be a prior distribution for the frequency of a non-reference allele in the ancestral population of population 1 and 2. Let p_1 be the frequency of the non-reference allele in population 1, and p_2 the frequency in population 2. To induce a correlation between the frequencies, let

$$p_1|p_A \sim Beta(\frac{1-f_1}{f_1}p_A, \frac{1-f_1}{f_1}(1-p_A)) \text{ and } p_2|p_A \sim Beta(\frac{1-f_2}{f_2}p_A, \frac{1-f_2}{f_2}(1-p_A)) \tag{A.1}$$

such that f_1 and f_2 are controlling the extent of correlation. An interpretation of this model is two daughter populations diverging from an ancestral population A with allele frequencies p_A, and under these conditions f_1 and f_2 are the F_{ST} values between repeated draws from $Beta(\frac{1-f_1}{f_1}p_A, \frac{1-f_1}{f_1}(1-p_A))$ and $Beta(\frac{1-f_2}{f_2}p_A, \frac{1-f_2}{f_2}(1-p_A))$ respectively.

For a given variant site, let $0 \leq X_1 \leq N_1$ be the number of non-reference alleles observed in N_1 population 1 haplotypes, and $0 \leq X_2 \leq N_2$ be the number of non-reference alleles observed in N_2 population 2 haplotypes. Then $X_1|f_1, p_A$ follows $Beta-Binomial(N_1, \theta_1 p_A, \theta_1(1-p_A))$ and $X_2|f_2, p_A$ follows $Beta-Binomial(N_2, \theta_2 p_A, \theta_2(1-p_A))$ where $\theta_z = \frac{1-f_z}{f_z}$. Then, conditional on p_A, the joint distribution of allele counts in a dataset is

$$\phi(i,j|f_1,f_2,p_A) = P(X_1 = i, X_2 = j|f_1,f_2,p_A)$$
$$= \binom{N_1}{i}\frac{B(i+\theta_1 p_A, N_1 - i\theta_1(1-p_A))}{B(\theta_1 p_A, \theta_1(1-p_A))}\binom{N_2}{j}\frac{B(j+\theta_2 p_A, N_2 - j\theta_2(1-p_A))}{B(\theta_2 p_A, \theta_2(1-p_A))} \tag{A.2}$$

where $B(x,y)$ is the β function. The unconditional distribution of X_1 and X_2 for populations with divergence parameter f_1 and f_2 is therefore

$$\begin{aligned}\phi(i,j|f_1,f_2,p_A) &= P(X_1 = i, X_2 = j|f_1,f_2,p_A)\\ &= \int_0^1 P(X_1 = i, X_2 = j|f_1,f_2,p_A)g(p_A|\alpha,\beta)dp_A\end{aligned} \tag{A.3}$$

where $g(p_A|\alpha,\beta)$ is the density of a $Beta(\alpha,\beta)$ random variable. The integral in A.3 can be computed either numerically or stochastically.

A.0.2 Using genomic control to correct for stratification

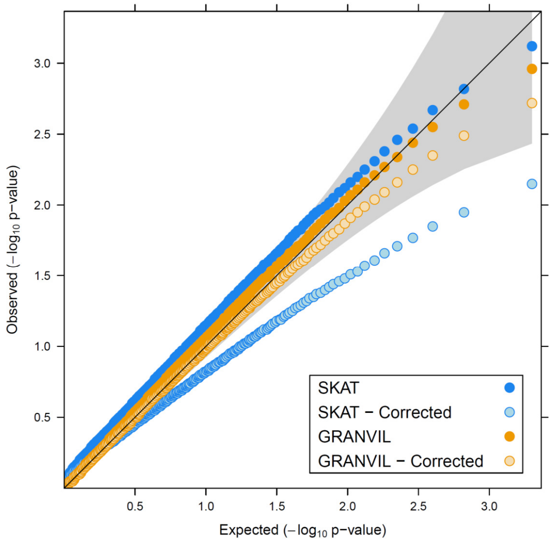

Figure A.1: A QQ-plot showing p-values for the dispersion test SKAT (blue) and burden test GRANVIL (orange) before (solid circles) and after (hollow circles) applying a correction based on genomic control (λ_{50}). The correction leads to overly conservative p-values for group-based rare variant test. The overcorrection is more severe in dispersion tests such as SKAT. The pictured scenario is Central and Northern Europeans at a mixing proportion of $r = 0.8$.

A.0.3 Summary statistics by functional annotation

	Allele Sharing			Weighted Symmetry		
Populations	Nonsynonymous	Fourfold Degenerate	Intronic	Nonsynonymous	Fourfold Degenerate	Intronic
Central-Northern	0.665	0.692	0.714	0.955	0.963	0.961
Central-Northwestern	0.817	0.868	0.872	0.999	0.996	0.999
Central-Western	0.866	0.903	0.908	0.983	0.972	0.978
Northern-Northwestern	0.684	0.688	0.735	0.952	0.954	0.958
Northern-Western	0.618	0.621	0.678	0.937	0.936	0.943
Northwestern-Western	0.841	0.843	0.863	0.985	0.980	0.987

Table A.1: We computed allele sharing and weighted symmetry values for the JSFS of rare (MAF $< 1\%$) nonsynonymous, fourfold degenerate and intronic sites within four European populations. Differences in selective pressures and allele frequencies produce unique JSFS for each class of variants, and we therefore do not expect identical summary statistics. However, similar patterns of population structure captured by nonsynonymous variants are also seen for fourfold degenerate and intronic sites. As a result, group-based analyses of either fourfold degenerate or intronic sites would produce differential stratification between the dispersion and burden tests, and provide a method for observing population stratification in candidate gene studies without genome-wide data.

APPENDIX B

Appendix for Chapter 3

B.0.4 Achieving infinite population size in finite time

Our two parameter formulation of growth has the form

$$P_t = \begin{cases} [\frac{P_0^{\beta-1}}{1+P_0^{\beta-1}(\beta-1)\alpha t}]^{\frac{1}{\beta-1}} & \text{for } \beta \neq 1 \\ P_0 e^{-\alpha t} & \text{for } \beta = 1 \end{cases} \tag{B.1}$$

P is the initial population size in haplotypes, t is time, and the model parameters α and β are constants. For $\beta \neq 1$ population size approaches infinity as the denominator of the solution approaches 0 which occurs as

$$P_0^{\beta-1}(\beta-1)\alpha t \to 1. \tag{B.2}$$

This can be rewritten as

$$t \to \frac{1}{\alpha P_0^{\beta-1}(\beta-1)}, \tag{B.3}$$

implying there is a finite time at which a population growing according to this model approaches infinite size. Assuming t, α, and P_0 are all positive, $\beta > 1$ results in infinite size in finite time. In coalescent simulation, the parameters α and β can be selected with equation (B.3) to make sure population size in the present is finite.

B.0.5 A comparison of sample size and current effective population size

An underlying assumption of the coalescent model is that the sample size is much smaller than the overall population size. When this assumption does not hold, the simplifying assumptions that there is at most one coalescent event per generation and that no more than two samples coalesce to the same common ancestor in a given event become unrealistic.

Wakeley and Takahashi (2003) performed a thorough study of how violations of the assumption n $<<$ N alter a coalescent sample's frequency spectrum and conclude that while n $\leq$ N the effects are "surprisingly mild". Of concern in our work is that FTE growth is so rapid that looking backwards in time the underlying population shrinks at a faster rate than our sample, resulting in a situation where the sample size and population size are of comparable magnitude and consequently the simplifying assumptions of the basic coalescent no longer hold.

To assess this possibility, we perform a series of simulations with a sample size of 20,000 haplotypes, a current population size of 8,000,000 haplotypes and an ancestral size of 20,000 haplotypes during which we calculate the ratio of sample size to average overall population size during growth. We find that for β between 0.1 and 3.5, the range investigated for this study, the average sample size to population size ratio does not exceed 0.177 (Figure B.1). In particular, for β values close to exponential growth, those between 0.5 and 1.5, the ratio does not exceed 0.05. In light of this and the values observed in our models, our work does not appear to be violating simplifying assumptions of the basic coalescent too substantially, and using our growth model within this framework is reasonable.

In our paper, we also present a constant population size model and a model of instantaneous growth as bounds on many of the results. While they are not present on figure B.1, the constant population size model begins with a ratio of sample size to population size of 1, and spends a good portion of its history close to this value. Likewise, at the time where the transition between current and ancestral size occurs in the instantaneous growth model sample size is still very large relative to population size, achieving a maximum ratio of 0.61. In both these cases the simplifying assumptions of the basic coalescent are likely more substantially violated than in our FTE models. Wakeley and Takahashi (2003) report that the major effect of violating this assumption is a deficiency in singleton variation, so results for these models may under-represent singletons. We include these models in this study for comparison only. If drawing accurate inferences from these models were of real interest, it would be important to correct for this issue, and it would likely require simulations using a more complex coalescent model.

B.0.6 Measuring linkage disequilibrium decay

We measure the decay of linkage disequilibrium in our samples with two commonly used pairwise statistics: r^2 and D'. For two variants, A and B, let p_A and p_B be their respective minor allele frequencies and p_{AB} the frequency of the haplotype with both minor alleles, then we can define the disequilibrium coefficient $D_{AB} = p_{AB} - p_A p_B$. Using D, r^2 is the

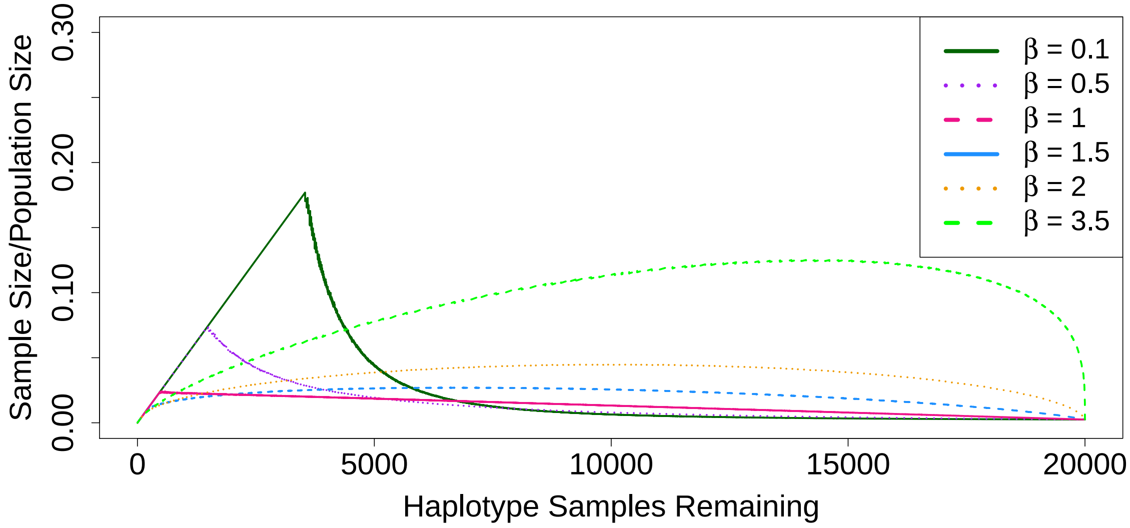

Figure B.1: Ratio of sample size to average population size across a range of β values. Results are based on average values from 1,000 independent samples, each of 20,000 haplotypes simulated from an ancestral population of 20,000 haplotypes growing to 8,000,000 haplotypes over 500 generations. The figure also gives a sense of the average number of haplotypes remaining at the end of growth for each model, corresponding to the beginning of linear decline and particularly noticeable for β = 0.1 and 0.5.

correlation coefficient between the loci and is defined as

$$r_{AB}^2 = \frac{D_{AB}^2}{p_A(1-p_A)p_B(1-p_B)}. \tag{B.4}$$

D' is defined as

$$D'_{AB} = \begin{cases} \frac{D_{AB}}{min(p_Ap_B,(1-p_A)(1-p_B))} & D_{AB} \leq 0 \\ \frac{D_{AB}}{min(p_A(1-p_B),(1-p_A)p_B)} & D_{AB} > 0 \end{cases}. \tag{B.5}$$

In our research we are interested in the magnitude but not the sign of D', and consequently, for the sake of comparison, we use the absolute value of this statistic in our results.

B.0.7 Non-singleton rare variation as growth accelerates

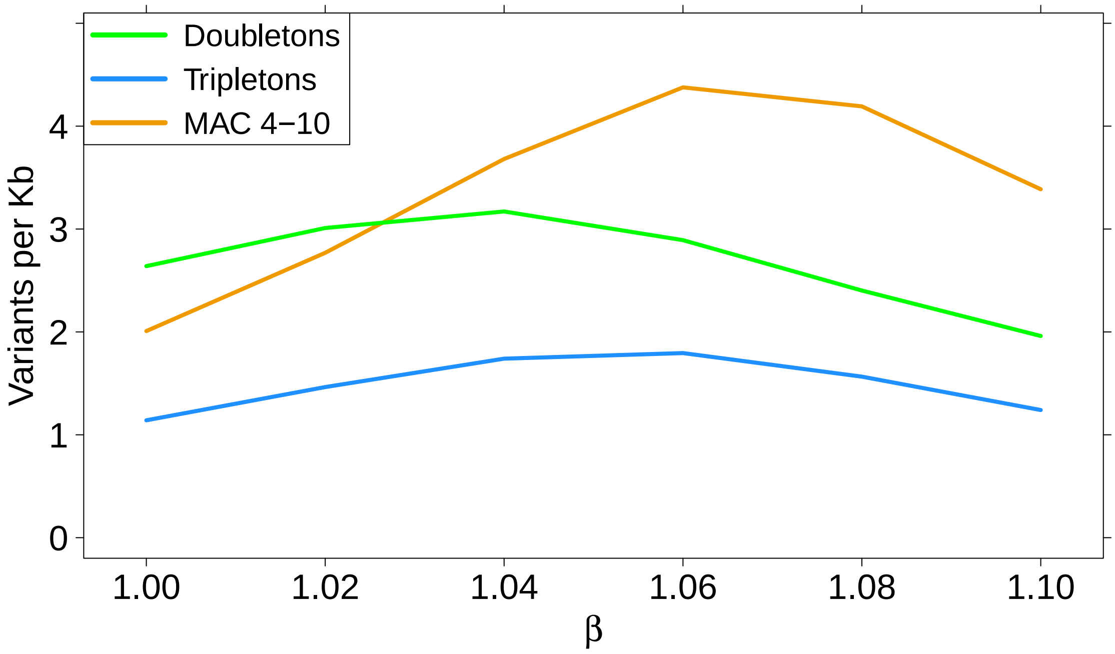

Figure B.2: The amount of very rare non-singleton variation in samples of 10,000 haplotypes with an initial growth rate of $\alpha = 100$ and 500 generations of growth.

B.0.8 The quantity of variation changes with the duration of growth and the rate of acceleration

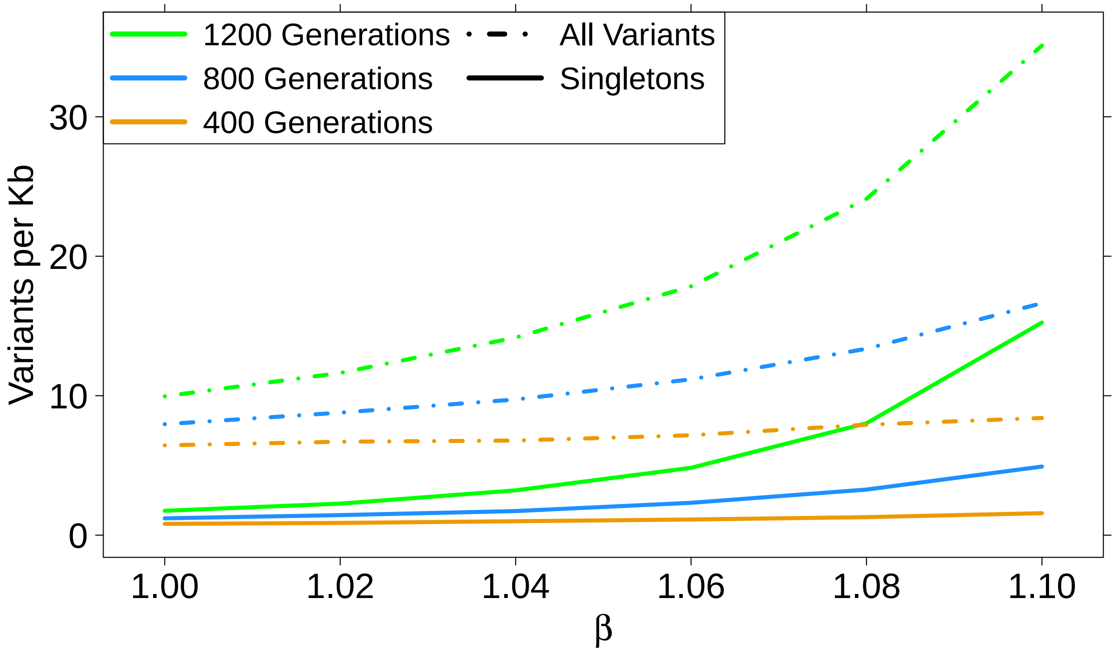

Figure B.3: The average number of variants per kilobase (kb) in samples of 10,000 haplotypes with an initial growth rate of $\alpha = 20$. Each colored line represents a different growth duration, with dashed lines giving the values for all variants per kb, and solid lines only the values for singleton variants per kb. The shorter the duration of growth, the less impact accelerating growth has on the values. And for any given acceleration value β the longer the duration the greater the number of variants and singletons present in the samples.

B.0.9 The quantity of variation changes with the duration of growth and the rate of acceleration

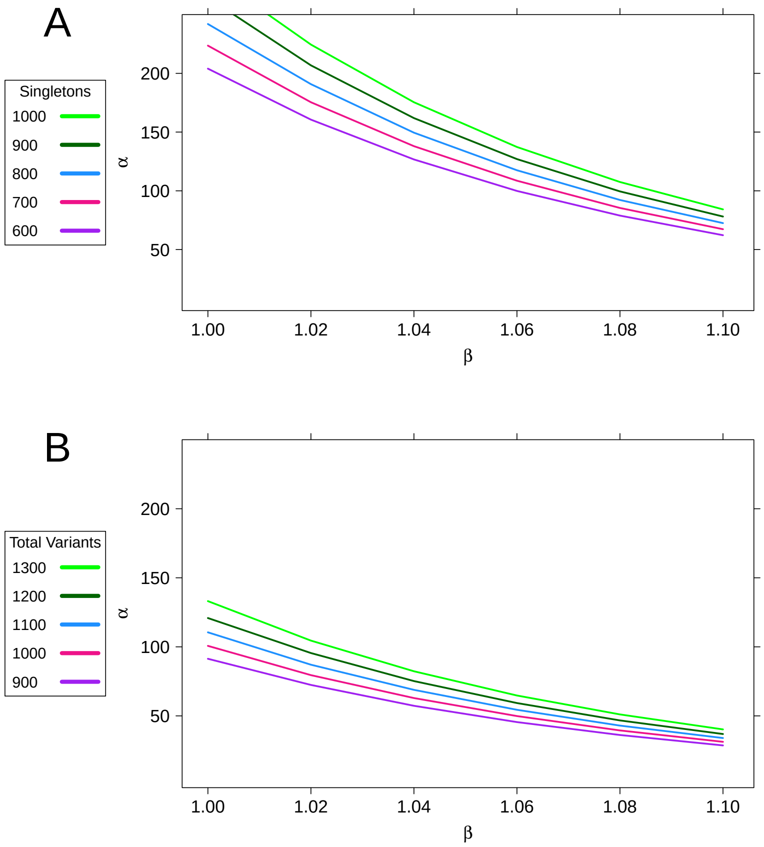

Figure B.4: The current population sizes presented in figure 3.4 correspond to distinct α, β pairs, plotted here. These results are based on the average number of **(A)** singletons and **(B)** all variants in 30 kb of simulated sequence in 10,000 haplotype samples drawn from a population growing from an ancestral size of 20,000 haplotypes over 500 generations. Each colored line represents a different number of singletons or total variants. Contour lines were generated using linear interpolation over a fine-scale range of α determined using a grid search.

B.0.10 Linkage disequilibrium decay in models with the same initial growth rate

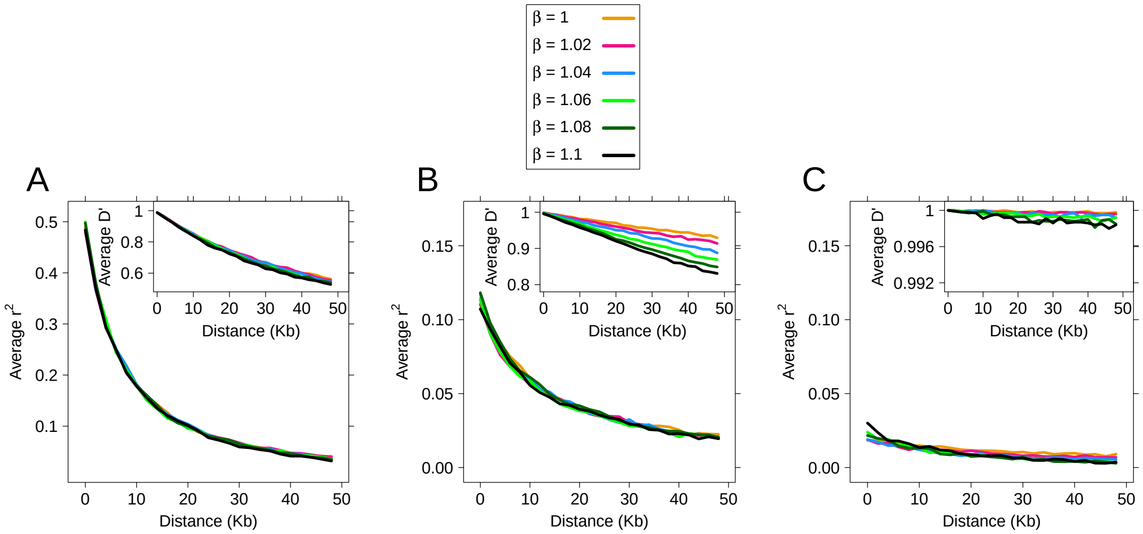

Figure B.5: Pairwise linkage disequilibrium decay measured by r^2 and D' in samples of 10,000 haplotypes from populations with an initial growth rate of $\alpha = 100$ and 500 generations of growth. The panels show LD decay in variants with a sample minor allele frequency of **(A)** 10% to 15%; **(B)** 1% to 2%; **(C)** a fixed allele count of 10. Each colored line represents a different growth model.

APPENDIX C

Appendix for Chapter 4

C.0.11 Pairwise r^2 and recent demography

Existing inference frameworks that summarize data using the frequency spectrum rely on a function of the mean aggregate length of branches with the same number of descendants in the final sample. This suggests that power to distinguish between demographic models can be increased by incorporating statistics that capture additional information about the distribution of branches in the ARG. Pairwise r^2 (Appendix B.0.6), a measure of LD between two variants, is one such statistic. Given two ARGs with identical mean lengths for branches with j descendants, the expected number of variants observed in the final sample with minor allele count j will be the same. However, if the variances in individual branch lengths differ between the ARGs then the probability of observing multiple mutations on the same branch will be different. When mutation events occur on the same branch of an ARG, they are perfectly correlated and have $r^2 = 1$ until they are split by a recombination event. Similarly, mutations arising on different ARG branches have $r^2 = 0$ until they are combined on a haplotype via recombination. Hence, for pairs of genetically close, putatively young genetic variants, unlikely to have been split by recombination during the ancestry of a sample, pairwise r^2 takes on a binary nature, and indicates whether mutations arose on the same branch of the ARG. As a result, pairwise r^2 gives information about the variance in individual branch lengths along a sample's ARG, which is influenced by population demography but not captured by the frequency spectrum.

C.0.12 A demographic likelihood incorporating pairwise r^2

Given an observed frequency spectrum and the pattern of pairwise r^2 at a locus we can write the likelihood of the observations under a given demographic model in the following way. Let Ψ denote the space of all possible genealogies G at the locus. Then for a model where the population size is $2N$ haplotypes, the per base mutation rate is μ, and growth occurs according to a pair of growth parameters (α, β) (Reppell *et al.*, 2014), the likelihood of the model based on the observations can be written as

$$L(N, \alpha, \beta, \mu | S, LD) = \sum_{G \in \Psi} P(LD|S, G, \mu) P(S|G, \mu) P(G|N, \alpha, \beta, \mu), \tag{C.1}$$

where S represents the observed frequency spectrum and LD the observed pattern of pairwise r^2. The parameter space Ψ is too large for summation over its entirety, so we approximate the likelihood via Monte-Carlo estimation using sampled realizations from Ψ, indexed by z:

$$\widehat{L}(N, \alpha, \beta, \mu | S, LD) = \frac{1}{Z} \sum_{z=1}^{Z} P(LD|S, G_z, \mu) P(S|G_z, \mu). \tag{C.2}$$

Our likelihood is comprised of three components. $P(S|G_z, \mu)$ encompasses both a "rate" component calculated using the total amount a variation observed at the locus and a "shape" component calculated using the allele counts of the observed variation. The third component is the probability of the observed pattern of pairwise r^2, written as $P(LD|S, G_z, \mu)$. For a sample of n haplotypes, at a locus with m sites, including non-variable sites, we use $j \in [1, n-1]$ to index variant minor allele counts (MACs) in the sample and i to index individual branches along a genealogy (Table C.1). Note, mutations that occur on a branch of the genealogy with exactly j descendants in the final sample appear as variants with j minor alleles.

Define $S_{tot} = \sum_{j=1}^{n-1} S_j$, where S_{tot} is the total number of variants observed at the locus and S_j is the number of observed variants with with MAC j. Correspondingly, $S_j = \sum_{i=1}^{w_j} s_{i,j}$ where $s_{i,j}$ is the number of mutations which occur along genealogy branch i and i runs from 1 to w_j, with w_j the total number of branches in the genealogy with j descendants in the final sample. The branches of the genealogy are measured in units of $2N$ generations, and the total length of the genealogy is written as L_{tot}, with $L_{tot} = \sum_{j=1}^{n-1} L_j = \sum_{j=1}^{n-1} \sum_{i=1}^{w_j} l_{i,j}$ for the individual branches i.

Following the approach of Boyko *et al.* (2008), subsequently employed in both Coventry *et al.* (2010) and Nelson *et al.* (2012), the rate likelihood component is specified by a Poisson distribution for S_{tot}, of the form

$$P(S_{tot}|G_z, \mu) = e^{-m\mu L_{tot}} \frac{(m\mu L_{tot})^{S_{tot}}}{S_{tot}!}, \tag{C.3}$$

Parameter	Definition
G_z	genealogy z
μ	mutation rate (mut/base/gen)
n	number of haplotypes in sample
m	total number of sites (both variant and monomorphic) at locus
S_{tot}	Total observed variants at locus
S_j	number observed variants with MAC j at locus
$s_{i,j}$	number of mutation events along branch $l_{i,j}$
L_{tot}	total length of entire ancestral tree
L_j	total length of branches in tree with MAC j
$l_{i,j}$	individual tree branch with j descendants, indexed by i
w_j	the number of branches in a genealogy with j descendants

Table C.1: Parameters used in our proposed likelihood, with indices $i \in [1, w_j]$ and $j \in [1, n-1]$.

and the shape likelihood follows a multinomial distribution:

$$P(S_j, j = 1, ..., n-1 | G_z, S_{tot}, \mu) = \frac{S_{tot}!}{\prod_{j=1}^{n-1} S_j!} \prod_{j=1}^{n-1} (\frac{L_j}{L_{tot}})^{S_j} \tag{C.4}$$

Like the shape likelihood component, the LD likelihood component follows a multinomial distribution. For each MAC j, with S_j total observations and success probabilities determined by individual branch lengths, the probability of the observed pairwise r^2 is calculated as:

$$P(LD|S, G_z, \mu) = \prod_{j=1}^{n-1} \sum_{\underline{S}_j} \frac{S_j!}{\prod_{i=1}^{w_j} s_{i,j}!} \prod_{i=1}^{w_j} (\frac{l_{i,j}}{L_j})^{s_{i,j}} = \prod_{j=1}^{n-1} \sum_{\underline{S}_j} P(s_j | l_j, S_j) \tag{C.5}$$

$\underline{S}_j$ denotes the possible configurations of mutations along the $\underline{l}_j = (l_{1,j}, l_{2,j}, ..., l_{w_j,j})$ branches that result in the observed number of variant pairs with $r^2 = 1$. For example, if we observe three variants with a given MAC, and two have $r^2 = 1$, we sum over all possible configurations where two variants occur on the same branch and the third occurs on a different branch.

To calculate this probability we make the assumption that variants with $r^2 = 1$ arose on the same branch, while those with $r^2 < 1$ did not. This assumption only makes sense for very rare and genetically close variants. To address this, instead of calculating over

$j \in [1, n-1]$ we specify a MAC cutoff C_{LD} above which we do not calculate the probability of observed pairwise r^2. Likewise, for the shape likelihood, there are relatively few observations of variants with larger MACs in the size of loci we perform inference over. Subsequently, a MAC cutoff is also applied to the shape likelihood, with all variants above the threshold grouped into a single category with success probability $(1 - \sum_{j=1}^{C_{shape}} L_j)/L_{tot}$. Note, C_{LD} need not equal C_{shape}.

Using formulas C.3, C.4, and C.5, we estimate the likelihood for a single locus. We then extend our inference across multiple loci $(1, ..., p)$ by taking the product:

$$\widehat{L}(N, \alpha, \beta, \mu) = \prod_{p=1}^{P} \widehat{L}_p(N, \alpha, \beta, \mu) \tag{C.6}$$

C.0.13 Likelihood estimation using importance sampling

In equation C.5 the number of terms in $\underline{S}_j$ is often so large that direct summation is prohibitive. $\underline{S}_j$ is composed of individual configurations of mutations $\underline{s}_j = (s_{1,j}, s_{2,j}, ..., s_{w_j,j})$ which can be grouped by sorting their entries in descending order (Figure C.1). We label these ordered configurations $\underline{s}_j^*$, then the sum from equation C.5 can be written as

$$\sum_{\underline{S}_j} P(\underline{s}_j | \underline{l}_j, S_j) = \sum_{\underline{s}_j^* \in \underline{S}_j} \sum_{\underline{s}_j \in \underline{s}_j^*} P(\underline{s}_j | \underline{l}_j, S_j). \tag{C.7}$$

After ordering, K of the configurations, $\underline{s}_j$ have the same form $\underline{s}_j^*$, then we have the useful identity

$$\sum_{\underline{s}_j \in \underline{s}_j^*} P(\underline{s}_j | \underline{l}_j, S_j) = K \widehat{P}(\underline{s}_j | \underline{l}_j, S_j) \tag{C.8}$$

where $\bar{P}(\underline{s}_j | l_j, S_j)$ is the average value of $P(\underline{s}_j | \underline{l}_j, S_j)$ for $\underline{s}_j \in \underline{s}_j^*$. For each $\underline{s}_j^*$ we estimate $\bar{P}(\underline{s}_j | \underline{l}_j, S_j)$ using importance sampling, and through equations C.7 and C.8 this allows us to calculate the likelihood $P(LD | S, G_z, \mu)$ across $\underline{S}_j$ of any size.

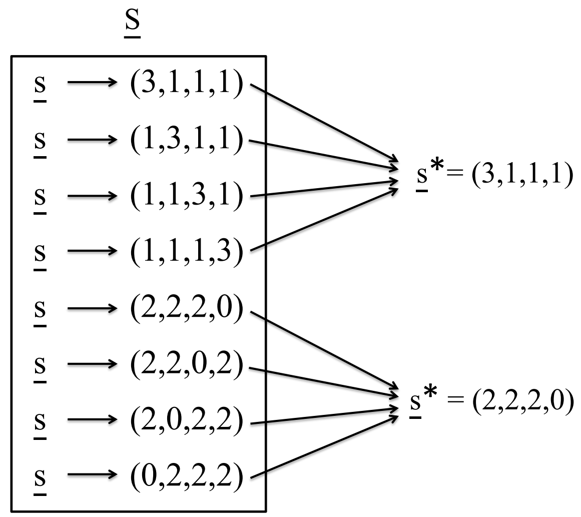

Figure C.1: Visual representation of $\underline{\text{S}}$, $\underline{\text{s}}$, and $\underline{\text{s}}^*$. For an arbitrary minor allele count, six variants are observed at a locus in a sample, with three of them in $r^2 = 1$. A realization of the sample's ancestry contains four branches with the corresponding number of descendants. $\underline{\text{S}}$ is then comprised of eight possible mutation configurations, $\underline{\text{s}}$, across the four branches. By ordering the mutation counts the eight $\underline{\text{s}}$ configurations are grouped into two vectors $\underline{\text{s}}^*$.

The importance sampling algorithm works as follows. we begin with an $\underline{s}^*$, comprised of w_j ordered counts $s_{t,j}$, $t \in [1, w_j]$ is used here to make the distinction that these mutation counts are not associated with a specific genealogy branch i before sampling. Then, for each $s_{t,j} > 0$ we sample a branch i with length $l_{i,j}$ to place the $s_{t,j}$ mutations on according to

$$P(l_{i,j}, s_{t,j}) = \frac{(l_{i,j})^{s_{t,j}}}{\sum\limits_{i=1}^{w_j}(l_{i,j})^{s_{t,j}}} \text{ for } j = 1, ..., C_{LD}. \tag{C.9}$$

Branches can only be selected once, so the probability of a complete configuration, where a branch has been sampled for every $s_{t,j} > 0$, is written as:

$$P(\underline{l}_j, \underline{s}_j^*) = \frac{w_j!}{(w_j - \sum\limits_{k=1}^{n-w_j+1}\lambda_k)! \prod\limits_{k=1}^{n-w_j+1}\lambda_k!} \prod_{i,t=1}^{w_j} P(l_{i,j}, s_{t,j}) \tag{C.10}$$

Where n is the sample size and λ_k is the number of branches with k mutation events. Using this sampling algorithm with U iterations, we estimate

$$\widehat{P}(\underline{s}_j|\underline{l}_j, S_j) = \frac{1}{U}\sum_{u=1}^{U}\left[\frac{S_j!(w_j - \sum\limits_{k=1}^{n-w_j+1}\lambda_k)!\prod\limits_{k=1}^{n-w_j+1}\lambda_k!}{w_j!\prod\limits_{t=1}^{w_j}s_{t,j}!}\prod_{i,t=1}^{w_j}\frac{(\frac{l_{i,j,u}}{L_j})^{s_{t,j,u}}}{P(l_{i,j,u}, s_{t,j,u})}\right]. \quad \text{(C.11)}$$

Which is substituted into equation C.8 to estimate the likelihood.

BIBLIOGRAPHY

1000 Genomes Project Consortium, Abecasis, G., Altshuler, D., Auton, A., Brooks, L., Durbin, R., Gibbs, R., Hurles, M., and McVean, G. (2010). A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–73.

Adams, A. and Hudson, R. (2004). Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. *Genetics*, **168**(3), 1699–712.

Balding, D., , and Nichols, R. (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica*, **96**, 3–12.

Bentley, A., Chen, G., Shriner, D., Doumatey, A., Zhou, J., Huang, H., Mullikin, J., Blakesley, R., Hansen, N., Bouffard, G., Cherukuri, P., Maskeri, B., Young, A., Adeyemo, A., and Rotimi, C. (2014). Gene-based sequencing identifies lipid-influencing variants with ethnicity-specific effects in african americans. *PLoS Genet.*, **10**, e1004190.

Bhaskar, A., Clark, A., and Song, Y. (2014). Distortion of genealogical properties when the sample is very large. *Proc. Natl. Acad. Sci. USA*, **111**, 2385–90.

Boyko, A., Williamson, S., Indap, A., Degenhardt, J., Hernandez, R., Lohmueller, K., Adams, M., Schmidt, S., Sninsky, J., Sunyaev, S., White, T., Nielsen, R., Clark, A., and Bustamante, C. (2008). Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet.*, **4**, e1000083.

Campbell, C., Chong, J., Malig, M., Ko, A., Dumont, B., Han, L., Vives, L., O'Roak, B., Sudamant, P., Shendure, J., Abney, M., Ober, C., and Eichler, E. (2012). Estimating the human mutation rate using autozygosity in a founder population. *Nat. Genet.*, **44**, 1277–81.

Chen, L., Hsu, L., Gamazon, E., Cox, N., and D.L., N. (2012). An exponential combination procedure for set-based association tests in sequencing studies. *Am. J. Hum. Genet.*, **91**, 977–86.

Coventry, A., Bull-Otterson, L., Liu, X., Clark, A., Maxwell, T., Crosby, J., Hixson, J., Rea, T., Muzny, D., Lewis, L., Wheeler, D., Sabo, A., Lusk, C., Weiss, K., Akbar, H., Cree, A., Hawes, A., Newsham, I., Varghese, R., Villasana, D., Gross, S., Joshi, V., Santibanez, J., Morgan, M., Chang, K., Hale IV, W., Templeton, A., Boerwinkle, E., Gibbs, R., and

Sing, C. (2010). Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nature Communications*, **1**, 131.

Cruchaga, C., Karch, C., Jin, S., Benitez, B., Cai, Y., Guerreiro, R., Harari, O., Norton, J., Budde, J., Bertelsen, S., Jeng, A., Cooper, B., Skorupa, T., Carrell, D., Levitch, D., Hsu, S., Choi, J., Ryten, M., UK Brain Expression Consortium, Hardy, J., Ryten, M., Trabzuni, D., Weale, M., Ramasamy, A., Smith, C., Sassi, C., Bras, J., Gibbs, J., Hernandez, D., Lupton, M., Powell, J., Forabosco, P., Ridge, P., Corcoran, C., Tschanz, J., Norton, M., Munger, R., Schmutz, C., Leary, M., Demirci, F., Bamne, M., Wang, X., Lopez, O., Ganguli, M., Medway, C., Turton, J., Lord, J., Braae, A., Barber, I., Brown, K., Alzheimer's Research UK Consortium, Passmore, P., Craig, D., Johnston, J., McGuinness, B., Todd, S., Heun, R., Kölsch, H., P.G., K., Hooper, N., Vardy, E., Mann, D., Pickering-Brown, S., Brown, K., Kalsheker, N., Lowe, J., Morgan, K., David Smith, A., Wilcock, G., Warden, D., Holmes, C., Pastor, P., Lorenzo-Betancor, O., Brkanac, Z., Scott, E., Topol, E., Morgan, K., Rogaeva, E., Singleton, A., Hardy, J., Kamboh, M., St George-Hyslop, P., Cairns, N., Morris, J., Kauwe, J., and Goate, A. (2014). Rare coding variants in the phospholipase d3 gene confer risk for alzheimer's disease. *Nature*, **505**, 550–4.

Devlin, B. and Risch, N. (1995). A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics*, **29**, 311–22.

Devlin, B. and Roeder, K. (1999). Genomic control for association studies. *Biometrics*, **55**, 997–1004.

Donnelly, P. and Tavaré, S. (1995). Coalescents and genealogical structure under neutrality. *Annu. Rev. Genet.*, **29**, 401–21.

Feenstra, B., Geller, F., Carstensen, L., Romitti, P., Körberg, I., Bedell, B., Krogh, C., Fan, R., Svenningsson, A., Caggana, M., Nordenskjöld, A., Mills, J., Murray, J., and Melbye, M. (2013). Plasma lipids, genetic variants near apoa1, and the risk of infantile hypertrophic pyloric stenosis. *JAMA*, **310**, 714–21.

Feng, S., Dajiang, L., Zhan, X., Wing, M., and Abecasis, G. (2014). Raremetal: fast and powerful meta-analysis for rare variants. *Bioinformatics*, **Epub ahead of print**.

Fu, Y. and Li, W. (1993). Statistical tests of neutrality of mutations. *Genetics*, **133**, 693–709.

Gazave, E., Ma, L., Chang, D., Coventry, A., Gao, F., Muzny, D., Boerwinkle, E., Gibbs, R., Sing, C., Clark, A., and Keinan, A. (2014). Neutral genomic regions refine models of recent rapid human population growth. *Proc. Natl. Acad. Sci. USA*, **111**, 757–62.

Gravel, S., Henn, B., Gutenkunst, R., Indap, A., Marth, G., Clark, A., Yu, F., Gibbs, R., 1000 Genomes Project, and Bustamante, C. (2011). Demographic history and rare allele sharing among human populations. *Proc. Natl. Acad. Sci. USA*, **108**(29), 11983–8.

Guerreiro, R., Lohmann, E., Brás, J., Gibbs, J., Rohrer, J., Gurunlian, N., Dursun, B., Bilgic, B., Hanagasi, H., Gurvit, H., Emre, M., Singleton, A., and Hardy, J. (2012). Using exome sequencing to reveal mutations in trem2 presenting as a frontotemporal dementia-like syndrome without bone involvement. *JAMA Neurol.*, **70**, 78–84.

Gutenkunst, R., Hernandez, R., Williamson, S., and Bustamante, C. (2009). Inferring the joint demographic history of multiple populations from multidimensional snp frequency data. *PLoS Genet.*, **5**, e1000695.

He, H., Li, W., Wu, D., Nagy, R., Liyanarachchi, S., Akagi, K., Jendrzejewski, J., Jiao, H., Hoag, K., Wen, B., Srinivas, M., Waidyaratne, G., Wang, R., Wojcicka, A., Lattimer, I., Stachlewska, E., Czetwertynska, M., Dlugosinska, J., Gierlikowski, W., Ploski, R., Krawczyk, M., Jazdzewski, K., Kere, J., Symer, D., Jin, V., Wang, Q., and de la Chapelle, A. (2013). Ultra-rare mutation in long-range enhancer predisposes to thyroid carcinoma with high penetrance. *PLoS One*, **8**, e61920.

Hudson, R. (1983). Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol.*, **23**(2), 183–201.

Hunt, K., Lehman, D., Arya, R., Fowler, S., Leach, R., Göring, H., Almasy, L., Blangero, J., Dyer, T., Duggirala, R., and Stern, M. (2005). Genome-wide linkage analyses of type 2 diabetes in mexican americans: the san antonio family diabetes/gallbladder study. *Diabetes*, **54**, 2655–62.

Huyghe, J., Jackson, A., Fogarty, M., Buchkovich, M., Stanlokov, A., Stringham, H., Sim, X., Yang, L., Fuchsberger, C., Cederberg, H., Chines, P., Teslovich, T., Romm, J., Ling, H., McMullen, I., Ingersoll, R., Pugh, E., Doheny, K., Neale, B., Daly, M., Kuusisto, J., Scott, L., Kang, H., Collins, F., Abecasis, G., Watanabe, R., Boehnke, M., Laakso, M., and Mohlke, K. (2013). Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. *Nat. Genet.*, **45**, 197–201.

Keinan, A. and Clark, A. (2012). Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science*, **336**, 740–43.

Keinan, A., Mullikin, J., Patterson, N., and Reich, D. (2007). Measurement of the human allele frequency spectrum demonstrates greater genetic drift in east asians than in europeans. *Nat. Genet.*, **39**, 1251–55.

Kersting, G. and Stanciu, I. (2013). The internal branch lengths of the kingman coalescent. *arXiv*, page 1303.4562 [math.PR].

Kiezun, A., Garimella, K., Do, R., Stitziel, N., Neale, B., McLaren, P., Gupta, N., Sklar, P., Sullivan, P., Moran, J., Hultman, C., Lichtenstein, P., Magnusson, P., Lehner, T., Shugart, Y., Price, A., de Bakker, P., Purcell, S., and Sunyaev, S. (2012). Exome sequencing and the genetic basis of complex traits. *Nat. Genet.*, **44**, 623–30.

Kingman, J. (1982a). The coalescent. *Stoch. Process Appl.*, **13**, 235–48.

Kingman, J. (1982b). On the genealogy of large populations. *Essays Stat. Sci.*, **19**, 27–43.

Kong, A., Gudbjartsson, D., Sainz, J., Jonsdottir, G., Gudjonsson, S., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., Shlien, A., Palsson, S., Frigge, M., Thorgeirsson, T., Gulcher, J., and Stefansson, K. (2002). A high-resolution recombination map of the human genome. *Nat. Genet.*, **31**, 241–7.

Ladouceur, M., Dastani, Z., Aulchenko, Y., Greenwood, C., and Richards, J. (2012). The empirical power of rare variant association methods: results from sanger sequencing in 1,998 individuals. *PLoS Genet.*, **8**, e1002496.

Lee, S., Emond, M., Bamshad, M., Barnes, K., Rieder, M., Nickerson, D., NHLBI GO Exome Sequencing Project Lung Project Team, Christiani, D., Wurfel, M., and Lin, X. (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet.*, **91**, 224–37.

Lee, S., Teslovich, T., Boehnke, M., and Lin, X. (2013). General framework for meta-analysis of rare variants in sequencing association studies. *Am. J. Hum. Genet.*, **93**, 42–53.

Li, B. and Leal, S. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.*, **83**, 311–21.

Li, C. (1969). Population subdivision with respect to multiple alleles. *Ann. Hum. Genet.*, **33**, 23–29.

Liu, K., Fast, S., Zawistowski, M., and Tintle, N. (2013a). A geometric framework for evaluating rare variant tests of association. *Genet. Epidemiol.*, **37**, 345–57.

Liu, L., Sabo, A., Neale, B., Nagaswamy, U., Stevens, C., Lim, E., and Bodea, C. (2013b). Analysis of rare, exonic variation amongst subjects with autism spectrum disorders and population controls. *PLoS Genet.*, **9**, Epub.

Liu, Q., Nicolae, D., and Chen, L. (2013c). Marbled inflation from population structure in gene-based association studies with rare variants. *Genet. Epidemiol.*, **37**, 286–92.

Lohmueller, K., Sparsø, T., Li, Q., Andersson, E., Korneliussen, T., Albrechtsen, A., Banasik, K., Grarup, N., Hallgrimsdottir, I., Kiil, K., Kilpeläinen, T., Krarup, N., Pers, T., Sanchez, G., Hu, Y., Degiorgio, M., Jø rgensen, T., Sandbaek, A., Lauritzen, T., Brunak, S., Kristiansen, K., Li, Y., Hansen, T., Wang, J., Nielsen, R., and Pedersen, O. (2013). Whole-exome sequencing of 2,000 danish individuals and the role of rare coding variants in type 2 diabetes. *Am. J. Hum. Genet.*, **93**, 1072–86.

MacDonald, M., Ambrose, C., Duyao, M., Myers, R., Lin, C., Srinidhi, L., Barnes, G., Taylor, S., James, M., Groot, N., MacFarlane, H., Jenkins, B., Anderson, M., Wexler, N., Gusella, J., Bates, G., Baxendale, S., Hummerich, H., Kirby, S., North, M., Youngman, S., Mott, R., Zehetner, G., Sedlacek, Z., Poustka, A., Frischauf, A., Lehrach, H., Buckler, A., Church, D., Doucette-Stamm, L., O'Donovan, M., Riba-Ramirez, L., Shah, M., Stanton, V., Strobel, S., Draths, K., Wales, J., Dervan, P., Housman, D., Altherr, M., Shiang, R., Thompson, L., Fielder, T., Wasmuth, J., Tagle, D., Valdes, J., Elmer, L., Allard, M., Castilla, L., Swaroop, M., Blanchard, K., Collins, F., Snell, R., Holloway, T., Gillespie, K., Datson, N., and Shaw, D. Harper, P. (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on huntington's disease chromosomes. *Cell*, **72**, 971–83.

Madsen, B. and Browning, S. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.*, **5**, e1000384.

Maher, M., Uricchio, L., Torgerson, D., and Hernandez, R. (2012). Population genetics of rare variants and complex diseases. *Hum. Hered.*, **74**, 118–28.

Marth, G., Czabarka, E., Murvai, J., and Sherry, S. (2004). The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics*, **166**, 351–72.

Mathieson, I. and McVean, G. (2012). Differential confounding of rare and common variants in spatially structured populations. *Nat. Genet.*, **44**, 243–6.

Mitchell, B., Kammerer, C., Blangero, J., Mahaney, M., Rainwater, D., Dyke, B., Hixson, J., Henkel, R., Sharp, R., Comuzzie, A., VandeBerg, J., Stern, M., and MacCluer, J. (1996). Genetic and environmental contributions to cardiovascular risk factors in mexican americans. the san antonio family heart study. *Circulation*, **94**, 2159–70.

Morris, A. and Zeggini, E. (2010). An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet. Epidemiol.*, **34**, 188–93.

N.A., R. (2006). The mean and variance of the numbers of r-pronged nodes and r-caterpillars in yule-generated genealogical trees. *Ann. Combinatorics*, **10**, 129–46.

Neale, B., Rivas, M., Voight, B., Altshuler, D., Devlin, B., Orho-Melander, M., Kathiresan, S., Purcell, S., Roeder, K., and Daly, M. (2011). Testing for an unusual distribution of rare variants. *PLoS Genet.*, **7**, e1001322.

Nelson, M., Wegmann, D., Ehm, M., Kessner, D., St Jean, P., Verzilli, C., Shen, J., Tang, Z., Bacanu, S., Fraser, D., Warren, L., Aponte, J., Zawistowski, M., Liu, X., Zhang, H., Zhang, Y., Li, J., Li, Y., Li, L., Woollard, P., Topp, S., Hall, M., Nangle, K., Wang, J., Abecasis, G., Cardon, L., Zöllner, S., Whittaker, J., Chissoe, S., Novembre, J., and Mooser, V. (2012). An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science*, **337**(6090), 100–4.

Ortega, V., Hawkins, G., Moore, W., Hastie, A., Ampleford, E., Busse, W., Castro, M., Chardon, D., Erzurum, S., Israel, E., Montealegre, F., Wenzel, S., Peters, S., Meyers, D., and Bleecker, E. (2014). Effect of rare variants in adrb2 on risk of severe exacerbations and symptom control during longacting agonist treatment in a multiethnic asthma population: a genetic study. *Lancet Respir. Med.*, **3**, 204–13.

Peden, J., Hopewell, J., Saleheen, D., Chambers, J., Hager, J., Soranzo, N., Collins, R., Danesh, J., Elliott, P., Farrall, M., Stirrups, K., Zhang, W., Hamsten, A., Parish, S., Lathrop, M., Watkins, H., Clarke, R., Deloukas, P., Kooner, J., Goel, A., Ongen, H., Strawbridge, R., Heath, S., Mälarstig, A., Helgadottir, A., Öhrvik, J., Murtaza, M., Potter, S., Hunt, S., Delepine, M., Jalilzadeh, S., Axelsson, T., Syvanen, A., Gwilliam, R., Bumpstead, S., Gray, E., Edkins, S., Folkersen, L., Kyriakou, T., Franco-Cereceda, A., Gabrielsen, A., Seedorf, U., MuTHER Consortium, Eriksson, P., Offer, A., Bowman, L., Sleight, P., Armitage, J., Peto, R., Abecasis, G., Ahmed, N., Caulfield, M., Donnelly, P., Froguel, P., Kooner, A., McCarthy, M., Samani, N., Scott, J., Sehmi, J., Silveira, A., Hellénius, M., van't Hooft, F., Olsson, G., Rust, S., Assman, G., Barlera, S., Tognoni, G., Franzosi, M., Linksted, P., Green, F., Rasheed, A., Zaidi, M., Shah, N., Samuel, M., Mallick, N., Azhar, M., Zaman, K., Samad, A., Ishaq, M., Gardezi, A., Fazal-ur Rehman, M., Frossard, P., Spector, T., Peltonen, L., Nieminen, M., Sinisalo, J., Salomaa, V., Ripatti, S., Bennett, D., Leander, K., Gigante, B., de Faire, U., Pietri, S., Gori, F., Marchioli, R., Sivapalaratnam, S., Kastelein, J., Trip, M., Theodoraki, E., Dedoussis, G., Engert, J., Yusuf, S., and Anand, S. (2011). A genome-wide association study in europeans and south asians identifies five new loci for coronary artery disease. *Nat. Genet.*, **43**, 339–44.

Peng, G., Fan, Y., Palculict, T., Shen, P., Ruteshouser, E., Chi, A., Davis, R., Huff, V., Scharfe, C., and Wang, W. (2013). Rare variant detection using family-based sequencing analysis. *Proc. Natl. Acad. Sci. USA*, **110**, 3985–90.

Price, A., Patterson, N., Plenge, R., Weinblatt, M., Shadick, N., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, **38**, 904–9.

Pritchard, J. (2001). Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.*, **69**, 124–37.

Reppell, M., Boehnke, M., and Zöllner, S. (2012). Ftec: a coalescent simulator for modeling faster than exponential growth. *Bioinformatics*, **28**, 1282–3.

Reppell, M., Boehnke, M., and Zöllner, S. (2014). The impact of accelerating faster than exponential population growth on genetic variation. *Genetics*, **196**, 819–28.

Robert, C. and Casella, G. (2004). Random variable generation. In *Monte Carlo Statistical Methods*. Springer Verlag, New York.

Schaffner, S., Foo, C., Gabriel, S., Reich, D., Daly, M., and Altshuler, D. (2005). Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.*, **15**, 1576–83.

Simons, Y., Turchin, M., Pritchard, J., and Sella, G. (2014). The deleterious mutation load is insensitive to recent population history. *Nat. Genet.*, **46**, 220–4.

Spouge, J. (2014). Within a sample from a population, the distribution of the number of descendants of a subsample's most recent common ancestor. *Theor. Popul. Biol.*, **92**, 51–4.

Stahl, E., Wegmann, D., Trynka, G., Gutierrez-Achury, J., Do, R., Voight, B., Kraft, P., Chen, R., Kallberg, H., Kurreeman, F., Diabetes Genetics Replication and Meta-analysis Consortium, Myocardial Infarction Genetics Consortium, Kathiresan, S., Wijmenga, C., Gregersen, P., Alfredsson, L., Siminovitch, K., Worthington, J., de Bakker, P., Raychaudhuri, S., and Plenge, R. (2012). Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nat. Genet.*, **44**, 483–9.

Tajima, F. (1989). The effect of change in population size on dna polymorphism. *Genetics*, **123**, 597–601.

Tennessen, J., Bigham, A., O'Connor, T., Fu, W., Kenny, E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., Kang, H., Jordan, D., Leal, S., Gabriel, S., Rieder, M., Abecasis, G., Altshuler, D., Nickerson, D., Boerwinkle, E., Sunyaev, S., Bustamante, C., Bamshad, M., Akey, J., Broad GO, Seattle GO, and NHLBI Exome Sequencing Project (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, **337**(6090), 64–9.

Teslovich, T., Musunuru, K., Smith, A., Edmondson, A., Stylianou, I., Koseki, M., Pirruccello, J., Ripatti, S., Chasman, D., Willer, C., Johansen, C., Fouchier, S., Isaacs, A., Peloso, G., Barbalic, M., Ricketts, S., Bis, J., Aulchenko, Y., Thorleifsson, G., Feitosa, M., Chambers, J., Orho-Melander, M., Melander, O., Johnson, T., Li, X., Guo, X., Li, M., Shin Cho, Y., Jin Go, M., Jin Kim, Y., Lee, J., Park, T., Kim, K., Sim, X., Twee-Hee Ong, R., Croteau-Chonka, D., Lange, L., Smith, J., Song, K., Hua Zhao, J., Yuan, X., Luan, J., Lamina, C., Ziegler, A., Zhang, W., Zee, R., Wright, A., Witteman, J., Wilson, J., Willemsen, G., Wichmann, H., Whitfield, J., Waterworth, D., Wareham, N., Waeber, G., Vollenweider, P., Voight, B., Vitart, V., Uitterlinden, A., Uda, M., Tuomilehto, J., Thompson, J., Tanaka, T., Surakka, I., Stringham, H., Spector, T., Soranzo, N., Smit, J., Sinisalo, J., Silander, K., Sijbrands, E., Scuteri, A., Scott, J., Schlessinger, D., Sanna, S., Salomaa, V., Saharinen, J., Sabatti, C., Ruokonen, A., Rudan, I., Rose, L., Roberts, R., Rieder, M., Psaty, B., Pramstaller, P., Pichler, I., Perola, M., Penninx, B., Pedersen, N., Pattaro, C., Parker, A., Pare, G., Oostra, B., O'Donnell, C., Nieminen, M., Nickerson, D., Montgomery, G., Meitinger, T., McPherson, R., McCarthy, M., McArdle, W., Masson, D., Martin, N., Marroni, F., Mangino, M., Magnusson, P., Lucas, G., Luben, R., Loos, R., Lokki, M., Lettre, G., Langenberg, C., Launer, L., Lakatta, E., Laaksonen, R., Kyvik, K., Kronenberg, F., König, I., Khaw, K., Kaprio, J., Kaplan, L., Johansson, A., Jarvelin, M., Janssens, A., Ingelsson, E., Igl, W., Kees Hovingh, G., Hottenga, J., Hofman, A., Hicks, A., Hengstenberg, C., Heid, I., Hayward, C., Havulinna, A., Hastie, N., Harris, T., Haritunians, T., Hall, A., Gyllensten, U., Guiducci, C., Groop, L., Gonzalez, E., Gieger, C., Freimer, N., Ferrucci, L., Erdmann, J., Elliott, P., Ejebe, K., Döring, A., Dominiczak, A., Demissie, S., Deloukas, P., de Geus, E., de Faire, U., Crawford, G., Collins, F., Chen, Y., Caulfield, M., Campbell, H., Burtt, N., Bonnycastle, L., Boomsma, D., Boekholdt, S., Bergman, R., Barroso, I., Bandinelli, S., Ballantyne, C., Assimes, T., Quertermous, T., Altshuler, D., Seielstad, M., Wong, T., Tai, E., Feranil, A., Kuzawa, C., Adair, L., Taylor, H., Borecki, I., Gabriel, S., Wilson, J., Holm, H., Thorsteinsdottir, U., Gudnason, V., Krauss, R., Mohlke, K., Ordovas, J., Munroe, P., Kooner, J., Tall, A., Hegele, R., Kastelein, J., Schadt, E., Rotter, J., Boerwinkle, E., Strachan, D., Mooser, V., Stefansson, K., Reilly, M., Samani, N., Schunkert, H., Cupples, L., Sandhu, M., Ridker, P., Rader, D., van Duijn, C., Peltonen, L., Abecasis, G., Boehnke, M., and Kathiresan, S. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, **466**, 707–13.

Tolle, J. (2003). Can growth be faster than exponential, and just how slow is the logarithm? *The Mathematical Gazette*, **87**, 522–525.

Tsui, L., Buchwald, M., Barker, D., Braman, J., Knowlton, R., Schumm, J., Eiberg, H., Mohr, J., Kennedy, D., Plavsic, N., and et al (1985). Cystic fibrosis locus defined by a genetically linked polymorphic dna marker. *Science*, **230**, 1054–7.

United Nations Department of Economic and Social Affairs Population Division (2011). World population prospects.

VanLiere, J. and Rosenberg, N. (2008). Mathematical properties of the r2 measure of linkage disequilibrium. *Theor. Popul. Biol.*, **74**, 130–7.

Voight, B., Adams, A., Frisse, L., Qian, Y., Hudson, R., and Di Rienzo, A. (2005). Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proc. Natl. Acad. Sci. USA*, **102**(51), 18508–13.

Wakeley, J. and Takahashi, T. (2003). Gene genealogies when the sample size exceeds the effective size of the population. *Mol. Biol. Evol.*, **20**, 208–13.

Weir, B., Cardon, L., Anderson, A., Nielsen, D., and Hill, W. (2005). Measures of human population structure show heterogeneity among genomic regions. *Genome Research*, **15**, 1468–76.

Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–78.

Williamson, S., Hernandez, R., Fledel-Alon, A., Zhu, L., Nielsen, R., and Bustamante, C. (2005). Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proc. Natl. Acad. Sci. USA*, **102**(22), 7882–7.

Wu, M., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.*, **89**, 82–93.

Zawistowski, M., Gopalakrishnan, S., Ding, J., Li, Y., Grimm, S., and Zollner, S. (2010). Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes. *Am. J. Hum. Genet.*, **87**, 604–17.