

Human migration, population divergence, and the accumulation of deleterious alleles: insights from private genetic variation and whole-exome sequencing

by

Zachary A. Szpiech

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in the University of Michigan
2012

Doctoral Committee:

Associate Professor Noah A. Rosenberg, Co-Chair
Associate Professor Sebastian K. Zöllner, Co-Chair
Professor Michael L. Boehnke
Professor Margit Burmeister
Assistant Professor Jun Z. Li

© Zachary A. Szpiech 2012

All Rights Reserved

For Frances

ACKNOWLEDGEMENTS

I must of course thank my advisor Noah Rosenberg for his exquisite mentoring, endless patience, and meticulous attention to detail. Completing my PhD under him has been a pleasure. Additionally, I would like to thank my committee members Sebasitan Zöllner, Michael Boehnke, Margit Burmeister, and Jun Li for their constructive comments and criticisms.

I give much gratitude to my former cubicle mate Michael DeGiorgio with whom I shared countless academic and non-academic conversations. These were not only immeasurably valuable but also helped me retain my sanity. Paul Verdu has, similarly, kept me both sane and delighted with stimulating conversations and debates. These two are truly wonderful friends and colleagues.

Thank you always to Trevor Pemberton, who loves to remind me that simulating data is just “making it all up.” Trevor is there to give me perspective on the biological roots of my work.

To the entire rest of the Rosenberg Lab, I’m not sure I could have managed the last five years without your support and friendship.

Ekjyot Saini has provided me with the love and strength I needed to push through all the myriad phases of graduate school. Without her support, I couldn’t have succeeded.

To my parents Mary Alfano-Torres and Edward Szpiech, you have nurtured me into the scientist I am. You have always encouraged me to pursue my interests in math and science. Without you I would not be here (literally).

To my uncle Thomas Alfano and cousin Alexander Alfano, thank you for endless hours of comic relief.

I also want to thank John Plough, my high school physics teacher. His quirky humor kept me laughing, and his passion for science was infectious.

Finally I want to thank all of my co-authors for each of my chapters: Mattias Jakobsson (Chapter II), Noah Rosenberg (Chapter II, III, IV), Jishu Xu (Chapter IV), Trevor Pemberton (Chapter IV), Weiping Peng (Chapter IV), Sebastian Zöllner (Chapter IV), and Jun Li (Chapter IV).

TABLE OF CONTENTS

DEDICATION	ii
ACKNOWLEDGEMENTS	iii
LIST OF FIGURES	vii
LIST OF TABLES	xii
LIST OF APPENDICES	xiv
CHAPTER	
I. Introduction	1
II. ADZE: A rarefaction approach for counting alleles private to combinations of populations	7
2.1 Introduction	7
2.2 Theory	9
2.3 Example	11
2.3.1 Methods	11
2.3.2 Simulation	13
2.3.3 Results	14
2.3.4 Out of Africa and the Peopling of Oceania	15
2.4 Discussion	19
2.5 Acknowledgments	20
III. On the size distribution of private microsatellite alleles	32
3.1 Introduction	32
3.2 Theory	33
3.2.1 A naïve argument	35
3.2.2 The probability of microsatellite configurations	35

3.2.3	Mutations on a genealogical branch	36
3.2.4	Probability of the set of allele sizes on a genealogical tree	37
3.2.5	Assigning alleles the roles of A , B , and C	38
3.2.6	Summing over labeled histories	39
3.2.7	Integrating out the coalescence times	41
3.2.8	Implementing the computation	42
3.2.9	Small- θ approximation	44
3.3	Arbitrary divergence time	48
3.4	Properties of the probability that private alleles lie on the edges	49
3.5	Application to data	51
3.6	Discussion	53
3.7	Acknowledgements	56
IV. Long runs of homozygosity are enriched for deleterious variation		74
4.1	Introduction	74
4.2	Methods	78
4.2.1	Raw read processing and variant calling	78
4.2.2	Site-level quality control	79
4.2.3	Variant classification by predicted functional impact	80
4.2.4	Genotype-level quality control	82
4.2.5	Final data set for downstream analysis	83
4.2.6	ROH data analysis	83
4.3	Results	85
4.3.1	Heterozygous genotypes and ROH size classes	85
4.3.2	The number of deleterious homozygous genotypes occurring in ROH	87
4.3.3	Damaging versus non-damaging variation in any size ROH	88
4.3.4	Damaging versus non-damaging variation in small, medium, or long ROH	90
4.4	Discussion	93
V. Conclusion		110
APPENDICES		114
BIBLIOGRAPHY		128

LIST OF FIGURES

Figure

- 2.1 The distance between simulated and true values of the proportions of alleles with specific geographic distributions, summed across distributions and plotted as a function of standardized sample size g from 10 to 500. Results shown represent the mean across 100 replicates. 22
- 2.2 The number of alleles private to various combinations of geographic regions as a fraction of the total at a standardized sample size of 40. The outer circle corresponds to the entire dataset including known first and second degree relatives. The inner circle corresponds to the subset that excludes known relatives. Values are represented in proportion to the appropriate angles measured at the center of the circle. Geographic regions are abbreviated: Af – Africa; Eu – Eurasia; Ea – East Asia; Oc – Oceania; Am – Americas. 23
- 2.3 The number of alleles private to various combinations of geographic regions as a fraction of the total, using a subset of the data excluding relatives. The innermost circle corresponds to calculations uncorrected for sample size variation. Moving outward the next rings correspond to calculations at standardized sample sizes of 10, 20, 30, and 40. Values are represented in proportion to the appropriate angles measured at the center of the circle. Geographic regions are abbreviated: Af – Africa; Eu – Eurasia; Ea – East Asia; Oc – Oceania; Am – Americas. 24
- 2.4 The mean number of (A) distinct alleles per locus and (B) private alleles per locus, as functions of standardized sample size for five major geographic regions (excluding known relatives). 25

2.5	The mean number of alleles per locus private to combinations of k of five major geographic regions as a function of standardized sample size (excluding known relatives). (A) $k = 2$, (B) $k = 3$, (C) $k = 4$, and (D) $k = 5$. Geographic regions are abbreviated: Af – Africa; Eu – Eurasia; Ea – East Asia; Oc – Oceania; Am – Americas.	26
2.6	Three hypothetical migration patterns to Oceania out of Africa. Geographic regions are abbreviated: Af – Africa; Eu – Eurasia; Ea – East Asia; Oc – Oceania; Am – Americas.	27
2.7	The mean number of alleles per locus private to the combination of Oceania and another major geographic region as a function of standardized sample size (excluding known relatives). Error bars represent standard error of the mean across loci. Geographic regions are abbreviated: Af – Africa; Eu – Eurasia; Ea – East Asia; Oc – Oceania; Am – Americas.	28
3.1	An enumeration of all possible labeled histories relating four sampled alleles. Twelve histories have asymmetric topologies (1-12), and six histories have symmetric topologies (13-18).	57
3.2	Example labelings of the two possible unlabeled topologies for gene genealogies with four lineages. (A) The asymmetric topology and (B) the symmetric topology are parameterized by allele sizes at the nodes, $(x_1, x_2, x_3, x_4, n_3, n_2)$, and by coalescence times (t_2, t_3, t_4) indicating the lengths of certain segments of the branches.	58
3.3	The simulated frequency of occurrence of seven possible allele configurations as a function of scaled mutation rate (θ) on a log scale. 10^6 trees are simulated per θ step. These simulations utilize four alleles, two in each of two populations. Alleles are related by the coalescent, and they mutate according to the symmetric stepwise mutation model.	59
3.4	The probability that the private alleles lie on the edges of the size distribution conditional on production of an $\{AB/AC\}$ configuration, as a function of θ (log scale). This probability is plotted from simulations and for a range of truncations for the infinite sums in numerically approximating Equation 3.1. Simulation results are based on 100,000 $\{AB/AC\}$ trees simulated per θ step.	60

3.5 A small- θ “parsimony” approximation for calculating the probability that private alleles occur on the edges of the size distribution, for the case of $x_1 = x_4$ with history 8. Alleles x_1 and x_2 are in population 1, and alleles x_3 and x_4 are in population 2. In (A), the private alleles (± 1) lie on the edges of the size distribution ($-1 < 0 < 1$); however, in (B) and (C) the private alleles ($\pm 1, \pm 2$) are not on the two edges of the size distribution ($-2 < -1 < 0$ or $0 < 1 < 2$). 61

3.6 The four types of coalescent scenarios with $t_d > 0$, with their coalescent time parameterizations. In scenario E_{11} , t_3 is defined as the time to coalescence of the two lineages in population 1, and t_4 is defined as the time to coalescence of the two lineages in population 2. 62

3.7 Simulated probability that the private alleles lie on the edges of the size distribution, conditional on production of an $\{AB/AC\}$ configuration. The plot shows this probability as a function of θ (log scale) and t_d . 10^6 $\{AB/AC\}$ trees are simulated for each choice of θ and t_d 63

3.8 Simulated small- θ probabilities that the private alleles lie on the edges of the size distribution conditional on production of an $\{AB/AC\}$ configuration, and the corresponding limiting probability computed analytically for $\theta \rightarrow 0$ (Equation 3.18), as functions of t_d . The simulation approach follows that of Figure 3.7 and is described in Appendix B. 64

3.9 Relative difference between mean coalescence times conditional on obtaining configuration C_6 ($\bar{t}_i^{C_6}$) and unconditional mean coalescence times (\bar{t}_i), as a function of θ (log scale). The mean conditional coalescence times were calculated by taking the mean of 10^6 simulated coalescence times in scenarios that produced configuration C_6 65

3.10 The empirical probability that private alleles lie on the edges of the size distribution in a sample of size four from a pair of populations. Plotted are pairwise calculations of this frequency for all 53 worldwide populations from the Human Genome Diversity Panel, arranged in major geographic regions. African, Middle Eastern, European, Central/South Asian, East Asian, Oceanian, and American populations are arranged by color in the labels. Blue represents a lower probability, and red represents a higher probability. 66

3.11	F_{ST} vs. the empirical frequency with which private alleles in a sample of size four lie on the edges of the size distribution. Each point represents a pair among 53 worldwide populations from the Human Genome Diversity Panel, excluding comparisons involving Native American populations and comparisons of populations with themselves. Pearson's $r = 0.9333$	67
4.1	The total percentage of individual genomes covered by class A, class B, and class C ROH.	95
4.2	The exome-sequencing calling and quality-control pipeline.	96
4.3	A breakdown of variant sites by genomic location as determined by the MapSNPs program included with PolyPhen2.	97
4.4	A breakdown of variant sites by PolyPhen2 classification.	98
4.5	Alternate allele frequency versus read depth (DP) for heterozygous genotypes. Red points represent called exome heterozygotes that are homozygotes in the Illumina genotype data. Blue points represent called exome heterozygotes that are heterozygotes in the Illumina genotype data. Black points represent called exome heterozygotes that do not occur in the Illumina genotype data. Dotted red lines represent the quality control cutoff for heterozygous genotypes. Any heterozygotes that fall in between the two red dotted lines are retained. Others are removed from the dataset.	99
4.6	A breakdown of the final data set by PolyPhen2 classification and genome location after all filtering.	100
4.7	The number of damaging homozygotes versus the fraction of the genome covered by a called ROH for each individual. Colored points represent the number of damaging homozygotes falling within ROH regions, and black points represent the number of damaging homozygotes falling outside ROH regions.	101
4.8	The fraction of all genome-wide non-reference homozygotes falling in ROH regions versus the fraction of the genome covered by a called ROH for each individual. A) any ROH region, B) class A ROH regions, C) class B ROH regions, and D) class C ROH regions. Solid points represent damaging homozygotes, and open points represent non-damaging homozygotes.	102

4.9	The fraction of all genome-wide non-reference homozygotes falling in ROH regions versus the fraction of the genome covered by a called ROH for each individual. Solid points represent damaging homozygotes in class C ROH regions, and open points represent damaging homozygotes in class A ROH regions.	103
4.10	The fraction of all private variants that are synonymous or missense per population. Missense variants are further split into predicted benign, predicted possibly damaging, and predicted possibly damaging classes by PolyPhen2.	104

LIST OF TABLES

Table

2.1	Percentages of 8516 total alleles private to various combinations of geographic regions. These percentages are obtained both uncorrected and corrected for sample size, excluding known first and second degree relatives in the computation. The value of g gives the size of subsamples used in the sample size correction. Geographic regions are abbreviated: Af – Africa; Eu – Eurasia; Ea – East Asia; Oc – Oceania; Am – Americas.	29
2.2	Comparisons of numbers of private alleles for pairs of geographic regions that would support a given migration model if observed. Geographic regions are abbreviated: Af – Africa; Eu – Eurasia; Ea – East Asia; Oc – Oceania; Am – Americas.	31
3.1	The seven possible configurations of four alleles in two populations and the counts of shared, private, and total distinct alleles for each configuration.	68
3.2	The four allele size relationships possible for the $\{AB/AC\}$ allele configuration.	68
3.3	The probability contributions for a given set of values $(n_A, n_B, n_C, n_3, \Psi)$ for each of the 18 labeled histories. These probabilities occur in the sums in Equations 3.6 and 3.7.	69
3.4	Definitions for the 12 classes of probability in the small- θ “parsimony” approximation.	71
3.5	The small- θ approximation contributions to the numerator (Equation 3.12) and denominator (Equation 3.14) of the probability that the private alleles lie on the edges, for each labeled history.	72

4.1	Called exome genotype concordance with known Illumina genotypes per individual.	105
4.2	Mean coverage and percentage of sites with $\geq 20x$ coverage for the 96797 quality-controlled sites (Section 4.2.2) for each individual. . .	106
4.3	Genotype counts per individual for sites classified as damaging. . . .	107
4.4	Genotype counts per individual for sites classified as non-damaging. . . .	108
4.5	Percentage of all genotypes that are heterozygous in a given region of an individual's genome (genome-wide, non-ROH, any ROH, class A ROH, class B ROH, and class C ROH).	109
A.1	The reparameterizations of Ψ for the events E_{ij}	118
A.2	The probabilities of node allele sizes and the coalescence time densities conditional on events E_{ij}	119

LIST OF APPENDICES

Appendix

A.	Derivation for arbitrary divergence time	115
B.	Implementing the computation for arbitrary divergence time	120
C.	Small- θ approximation for arbitrary divergence time	123

CHAPTER I

Introduction

Population genetics is a field that interprets the empirical distribution of genetic variation within and between populations of individuals and mathematically models the underlying processes shaping this variation. Several processes influence the distribution of genetic variation. Mutation introduces novel variants into a population. Recombination re-assorts variation along otherwise physically linked stretches of chromosome, allowing alternate genetic forms to be inherited in novel combinations. Genetic drift stochastically changes variant frequencies, causing some variants to disappear from a population and other variants to fix in it. Changes in population size such as bottlenecks or population expansions can reduce or increase diversity within a population, respectively, exacerbating or diminishing the effect of drift. Individuals who migrate between populations introduce variation from one population into others. Finally, natural selection acts to purge variants that are less fit while bringing advantageous variants to high frequency. In humans, studies of these processes have been of great importance in diverse fields such as medical and clinical genetics (*Burchard et al.*, 2003; *Tishkoff and Kidd*, 2004; *Reiner et al.*, 2005; *Kumar et al.*, 2010; *Rosenberg et al.*, 2010), forensics (*Chakraborty and Kidd*, 1991; *Jefferys et al.*, 1991; *National Research Council of the USA*, 1992, 1996; *Evetts and Weir*, 1998), and anthropology (*Cavalli-Sforza and Feldman*, 2003; *Campbell and*

Tishkoff, 2010).

One interesting and important class of genetic variation is private genetic variation. A variant is deemed private if it is found only in a single population among a broader collection of sampled populations. Private genetic variation was first utilized in human evolutionary genetics by *Neel* (1973) in the analysis of electrophoretic protein motility polymorphisms across several native South American populations. In this first study, *Neel* (1973) made several assumptions about the nature of these private polymorphisms: that each was neutral and that each copy of the polymorphism descended from the same mutation. Also assuming that only one-third of these variants were electrophoretically detectable, he was able to use theoretical results on the expected time to extinction of a neutral mutation (*Kimura and Ohta*, 1969) to estimate a mutation rate for these variants. *Thompson and Neel* (1978), *Neel* (1978), and *Neel and Thompson* (1978) later revisited these theoretical results in the context of the native South American data and reached further refined mutation rate estimates.

Private genetic variation is also an important indicator of gene flow. *Slatkin* (1985) presented simulations of many demes under an island migration model and infinite alleles model. In this study, he considered “conditional average frequencies,” which are defined as the average frequency of all alleles found in exactly i demes. Whereas he showed that these statistics are robust to varying selection strength and mutation rate, he found that they are very sensitive to migration rate. In particular, he was able to provide a quantitative estimate of the number of effective migrants per generation (Nm) for the case of $i = 1$, corresponding to the average frequency of private alleles. *Barton and Slatkin* (1986) extended this result theoretically to give a more accurate estimate of Nm for private alleles sampled from populations of heterogeneous sample size.

Analyses of private alleles can also be informative about migration patterns

(*Calafell et al.*, 1998; *Schroeder et al.*, 2007; *Szpiech et al.*, 2008). For example, *Zhivotovsky et al.* (2003) identified a high-frequency microsatellite allele private to the Americas in their analysis of 377 microsatellite loci across 52 HGDP-CEPH populations. Subsequently, *Schroeder et al.* (2007) considered many possible explanations for this observation, including selection and admixture. They came to the conclusion that the most parsimonious explanation for this high-frequency private allele is that all modern Native American populations descended from the same founding population, in contrast to *Greenberg et al.* (1986) who proposed that three separate migrations were responsible for the peopling of the Americas.

Analyses of private genetic variation have been important in fields beyond human genetics as well. In molecular ecology and conservation genetics (e.g. *Petit et al.*, 1998; *Parker et al.*, 1999; *Fiumera et al.*, 2000; *Neel and Cummings*, 2003; *Torres et al.*, 2003; *Kalinowski*, 2004), private alleles are indicators of diversity and are thought to represent potential novel evolutionary pathways. By selecting individuals and populations for conservation on the basis of private allelic richness, conservation geneticists seek to preserve overall diversity and the potential for future adaptation and to decrease the risk of extinction.

In this dissertation I develop theoretical models and present empirical analyses of private genetic variation with results relevant for the inference of migration (Chapter II) and population divergence (Chapter III), and I further present an analysis of the accumulation of deleterious variation in the human genome (Chapter IV). Although the analyses in Chapters II and III are focused on human data, the theory is not specific to any particular species.

In Chapter II (*Szpiech et al.*, 2008), I introduce the concept of generalized private alleles and a method to count them while correcting for the heterogeneity of population sample sizes. The analysis of the distribution of alleles across populations is a fundamental tool for examining population diversity and relationships, and I

develop a generalized rarefaction approach for counting alleles private to combinations of populations. These generalized private alleles are distinct alleles found in each of a subset of populations and nowhere else, and my method evaluates the number of alleles found for each subset of populations but absent in all remaining populations. In order to make these calculations comparable with each other, the method considers equal-sized subsamples from each subset of populations. Combinatorics allows these computations to be made quickly. I apply this method to a human microsatellite dataset comprised of individuals from the HGDP-CEPH panel (*Rosenberg et al.*, 2005) and then interpret the results with respect to theories of early human migration out-of-Africa and into Oceania.

Next, I take a theoretical modeling approach to answer questions in the population genetics of private variation. The basic theoretical model I use is the coalescent (*Kingman*, 1982; *Wakeley*, 2009), a retrospective model of gene genealogies. The coalescent offers a robust and elegant mathematical framework in which to study the evolution of genetic lineages. Because it is a retrospective model, it only requires one to keep track of the evolution of the current sample. This allows simulations to be very efficient compared to forward-in-time models, which require one to track the evolution of the entire population.

In Chapter III (*Szpiech and Rosenberg*, 2011) I use the coalescent to model how population divergence and mutation rate affect the allele size distribution of private microsatellite alleles. It has been observed that private microsatellite alleles tend to be found in the tails rather than in the interior of the allele size distribution (*Wang et al.*, 2007). In order to explain this phenomenon, I investigate the size distribution of private alleles in a coalescent model of two populations, assuming the symmetric stepwise mutation model as the mode of microsatellite mutation. Under this model, I calculate the probability that the private alleles occupy the two tails of the size distribution. The model predicts that population divergence will greatly influence

the probability that private alleles are on the edge of the size distribution, and I test this prediction using the HGDP-CEPH microsatellite dataset (*Rosenberg et al.*, 2005).

In addition to studying the properties of private genetic variation, I also study the distribution of deleterious variation in human individuals. The study of deleterious variation has been of great importance to evolutionary genetics (*Muller*, 1950; *Morton et al.*, 1956; *Kondrashov*, 1995; *Charlesworth and Charlesworth*, 1998; *Eyre-Walker and Keightley*, 1999; *Fay et al.*, 2001; *Sunyaev et al.*, 2001; *Lohmueller et al.*, 2008; *Chun and Fay*, 2011; *Cooper and Shendure*, 2011; *Lohmueller et al.*, 2011; *Necşulea et al.*, 2011; *Lesecque et al.*, 2012; *Tennessen et al.*, 2012). Furthermore, although much progress has been made in identifying disease risk variants (*Rosenberg et al.*, 2010), much of the heritability of disease risk remains “missing” (*Lander*, 2011). The search for Mendelian disease genes has shifted to the discovery of rare variants, which are increasingly likely to be private (*Marth et al.*, 2011). This underscores the importance of understanding patterns of deleterious genetic variation, which provide a foundation on which we can build not only our understanding of genetic disease risk but also of evolutionary genetics.

In Chapter IV, I analyze the exome sequences of 27 individuals from 6 HGDP-CEPH populations. As changes in coding regions of the genome are likely to be disruptive, exome sequences provide a natural dataset to consider the distribution of deleterious changes. Exome sequencing offers the potential to study the genomic variables that underlie patterns of deleterious variants. Runs of homozygosity (ROH) are one such interesting genomic variable. These ROH are regions of the genome with long stretches of successive homozygous genotypes. ROH can be formed by both population demographics (isolation or bottlenecks) and cultural practices (consanguinity). These processes create different sizes of ROH, and in general ROH can be classified into three broad categories. Short ROH represent homozygosity for

ancient haplotypes (*International HapMap Consortium*, 2007); ROH of intermediate length represent homozygosity arising from background relatedness in a population (*Lencz et al.*, 2007; *Curtis et al.*, 2008; *Jakkula et al.*, 2008; *McQuillan et al.*, 2008); and long ROH result from recent parental relatedness (*Broman and Weber*, 1999; *Li et al.*, 2006; *Gibson et al.*, 2006; *International HapMap Consortium*, 2007; *Lencz et al.*, 2007; *Curtis et al.*, 2008; *Jakkula et al.*, 2008; *McQuillan et al.*, 2008; *Kirin et al.*, 2010). Here, I consider how ROH of different size classes correlate with deleterious variation.

This dissertation focuses on the theoretical modeling and empirical analysis of private genetic and coding variation with a specific focus on humans. I build a theoretical framework for analyzing the distribution of private genetic variation among populations and develop methodological tools to understand patterns of human demography. Furthermore, I analyze the distribution of deleterious genetic variation in humans. These results will be useful for disentangling the demographic histories of humans and non-humans alike and will provide results relevant to the study of population-specific variation in the search for disease genes.

CHAPTER II

ADZE: A rarefaction approach for counting alleles private to combinations of populations

2.1 Introduction

The analysis of the distributions of alleles across populations is important for elucidating genetic diversity and population relationships. Two fundamental quantities for a population at a given locus are the number of distinct alleles in the population and the number of alleles private to the population (that is, not found in other populations). These quantities are especially informative when populations are studied for highly variable multiallelic markers such as microsatellites.

The number of distinct alleles and the number of private alleles depend heavily on sample size, and they can be difficult to interpret when sample sizes differ across populations. The rarefaction approach has been an important strategy for producing estimates that are comparable in different populations (*Hurlbert*, 1971; *Petit et al.*, 1998; *Kalinowski*, 2004, 2005). The idea of the rarefaction method is to trim unequal samples to the same standardized sample size, a number less than or equal to the smallest sample size across populations. For a standardized size g , populations are compared by considering the estimates of “allelic richness” and “private allelic richness” that would be obtained when averaging across all subsamples of size g .

In the rarefaction framework, the estimated allelic richness of a population is the number of distinct alleles expected in a random subsample of size g drawn from the population (*Hurlbert, 1971; Petit et al., 1998*). The estimated private allelic richness is the number of private alleles expected in the population when random subsamples of size g are taken from each of J populations under consideration (*Kalinowski, 2004*). Combinatorial formulas make it possible to compute these statistics relatively quickly.

Often, as was noted by *Kalinowski (2004)*, especially if several populations in a sample are closely related, few alleles are private to individual populations. Instead, alleles may be private to groups of populations—that is, alleles may be found in each of several members of a larger set of populations. We therefore introduce a generalization of the private allelic richness concept of *Kalinowski (2004)*. We compute a generalized private allelic richness statistic that uses a rarefaction approach to measure the number of distinct alleles private to a *group* of populations and found in *all* populations in the group. This statistic makes it possible to evaluate the sample size-corrected number of distinct alleles private to any set of populations, and it reduces to private allelic richness when the group of populations consists of only a single population. We demonstrate the application of the new generalized private allelic richness statistic using microsatellite genotypes from human populations. By considering the sample size-corrected number of distinct alleles private to various combinations of major geographic regions, this analysis produces evidence in support of the hypothesis that an early human migration from Africa to Oceania did not have an appreciable effect on genetic variation in modern populations of Asia. We have implemented computations of allelic richness, private allelic richness, and our new measure of generalized private allelic richness in a computer program *ADZE*—a tool for “chopping” samples down to standardized sizes for data analysis.

2.2 Theory

Consider a locus with I distinct alleles, and define N_{ij} as the number of copies of allele type i in a sample from population j . $N_j = \sum_{i=1}^I N_{ij}$ is the sample size of population j at the locus. The probability of finding no copies of allele type i in a subsample of size g alleles from population j is

$$Q_{ijg} = \frac{\binom{N_j - N_{ij}}{g}}{\binom{N_j}{g}}. \quad (2.1)$$

Then the probability of finding at least one copy of allele type i in a sample of size g alleles from population j is $P_{ijg} = 1 - Q_{ijg}$, and

$$\hat{\alpha}_g^{(j)} = \sum_{i=1}^I P_{ijg} \quad (2.2)$$

is the estimated allelic richness of a sample of size g from population j (*Hurlbert*, 1971; *Petit et al.*, 1998; *Kalinowski*, 2004). Equation 2.2 estimates the expected number of distinct alleles that will be observed in population j in a sample of size g .

Using this notation, the estimated private allelic richness for a sample size g from population j can be written as

$$\hat{\pi}_g^{(j)} = \sum_{i=1}^I \left[P_{ijg} \left(\prod_{\substack{j'=1 \\ j' \neq j}}^J Q_{ij'g} \right) \right], \quad (2.3)$$

where J is the total number of populations (*Kalinowski*, 2004). This formula sums over distinct allele types, i , the probability that a random subsample of size g from population j contains allele type i and that subsamples of size g from the remaining populations do not contain i .

Generalizing the concept of private allelic richness, we can consider the number of distinct alleles private to some combination of k populations selected from

$\{1, 2, \dots, J\}$. Consider a set of J populations labeled 1 to J , and let $\mathcal{S} = \{1, 2, \dots, J\}$. Let \mathcal{C}_k be the set of all possible combinations of k elements from \mathcal{S} . There are $\binom{J}{k}$ possible combinations in \mathcal{C}_k . We label these combinations by \mathcal{C}_{km} , where m ranges from 1 to $\binom{J}{k}$. Using the following equation we can calculate $\hat{\pi}_{gk}^{(m)}$ —the estimated number of distinct alleles private to the m th combination of k populations, when samples of size g are drawn from each of the J populations:

$$\hat{\pi}_{gk}^{(m)} = \sum_{i=1}^I \left[\left(\prod_{j \in \mathcal{C}_{km}} P_{ijg} \right) \left(\prod_{j' \in \mathcal{S} \setminus \mathcal{C}_{km}} Q_{ij'tg} \right) \right]. \quad (2.4)$$

$\mathcal{S} \setminus \mathcal{C}_{km}$ denotes the set \mathcal{S} excluding the elements of \mathcal{C}_{km} . For $k = 1$, $\hat{\pi}_{gk}^{(m)}$ reduces to private allelic richness as in Equation 2.3. For $k = J - 1$, Equation 2.4 can be considered a measure for “missing allelic richness”, and it reduces to

$$\hat{\mu}_g^{(j)} = \sum_{i=1}^I \left[Q_{ijg} \left(\prod_{\substack{j'=1 \\ j' \neq j}}^J P_{ij'tg} \right) \right]. \quad (2.5)$$

In this equation, $\hat{\mu}_g^{(j)}$ gives a sample size-corrected measure of the number of distinct alleles found in all populations other than population j .

We note that the “alleles private to a combination of populations” that we define are different from the “regionally private alleles” described for hierarchical sampling schemes by *Kalinowski* (2004). This latter framework provides an approach for counting the number of alleles present in *at least one* of several “populations” contained within a larger “region,” and not contained in all other “regions.” Thus, the hierarchical method of *Kalinowski* (2004) obtains the number of regionally private alleles by “rarefacting” over populations and samples, considering all possible sets of samples from a region in which r populations within the region are represented, and in which each population sample has size g . Related computations also appear in *Kalinowski* (2005).

In contrast to the methods of *Kalinowski* (2004, 2005), our approach instead counts alleles required to be found in *each* of a set of populations, and required to be absent in all other populations. By considering all possible combinations of populations, this approach makes it sensible to use Equation 2.4 to obtain the proportion of alleles present in a particular combination, while adjusting for unequal sample sizes among populations. The resulting proportions can be viewed as a partition of the full set of alleles into categories defined by the populations in which they are found, whereas the hierarchical scheme of *Kalinowski* (2004, 2005) would count some alleles more than once if viewed in this manner.

For a given value of the sample size g , we first compute Equation 2.4 for each of the $2^J - 1$ nonempty combinations possible for a set of J populations. We then obtain the proportion of alleles for combination m by dividing the value from Equation 2.4 for this combination by the sum of all $2^J - 1$ values. Because sample size is controlled, this approach enables meaningful assessments of the proportions of alleles with particular geographic distributions.

2.3 Example

Since private alleles have proven useful in investigating population structure and migration patterns (e.g. *Neel*, 1978; *Calafell et al.*, 1998; *Schroeder et al.*, 2007), we now provide a detailed example to illustrate various ways in which our generalized private allelic richness approach can be used in data analysis.

2.3.1 Methods

We employ a dataset from human populations (*Rosenberg et al.*, 2005) containing genotypes of 1048 individuals—the H1048 collection of individuals (*Rosenberg*, 2006)—at 783 microsatellite loci. We also consider the genotypes for the H952 subset of the full H1048 dataset—a group of 952 individuals that contains no known

first or second degree relatives (*Rosenberg, 2006*). The individuals were classified as belonging to one of five major geographic regions—sub-saharan Africa, Eurasia (Europe, Central/South Asia, and the Middle East, including North Africa), East Asia, Oceania, and the Americas. We treat each of these regions as a “population” in the computations that follow.

We used Equations 2.2 and 2.3 to compute allelic richness and private allelic richness for each of the five geographic regions, and we used Equation 2.4 to compute generalized private allelic richness for various combinations of regions. The computation was performed for individual loci for values of g from 2 up to the maximum possible value for the dataset, and for each g the mean was taken across loci. For a given locus, the smallest number of individuals in one of the population groupings under consideration specifies the largest value of g possible to use for private allelic richness and generalized private allelic richness computations at that locus. Because missing data can reduce this maximal g , in our example we used the locus filtering feature in the *ADZE* computer program to restrict our attention to 721 loci for which each geographic region had a missing data rate less than or equal to 15% (similar results are obtained when using all 783 loci, with a lower maximal g). With this collection of loci in the H952 dataset, every locus had a sample size of at least 48 observations in each of the five geographic regions. The same collection of 721 loci was used in analyses that employed the full collection of 1048 individuals.

There are 31 combinations of one or more of the five geographic regions, and we computed generalized private allelic richness for each combination. For comparison, we also partitioned alleles among the 31 possible geographic distributions without correcting for sample size. Considering all loci, each distinct allele can be private to a single region, present in two regions, present in three regions, present in four regions, or present in all five regions. For each of the 31 geographic distributions, we determined the fraction of alleles in the dataset that had the specified distribution.

2.3.2 Simulation

We performed a simulation study to assess the extent to which our estimates of the proportions of alleles in various combinations of geographic regions reflect the true proportions. First, for each of the 783 loci in our dataset we considered the sample allele frequencies in each geographic region from the H952 subset of individuals as the true allele frequencies. For each locus and each geographic region, we sampled 250 diploid individuals (with replacement) to create a simulated dataset. Repeating this sampling, we produced 100 simulated datasets, each consisting of 250 diploid individuals per region at each of the 783 loci. For each simulated dataset, each locus, and each value of g from 10 to 500, we calculated the generalized private allelic richness for each of the 31 combinations of one or more of the five regions. We then divided each of the 31 values by their sum to determine the fraction of alleles present in each of these 31 categories. Similarly, continuing to treat the sample frequencies in the H952 dataset as true frequencies, we tallied the true number of distinct alleles in each of the 31 combinations of regions in the H952 dataset and divided by the total number of distinct alleles worldwide to obtain the true proportion of private alleles for each of the 31 combinations of regions. We then calculated $\sum_{i=1}^{31} (sim_i - true_i)^2$, where sim_i and $true_i$ denote the simulated and true proportions of alleles private to geographic combination i , respectively. The mean of this statistic was taken over the 100 replicate simulated datasets, and the resulting quantity was then plotted in Figure 2.1.

As the standardized sample size g increases, the distance between simulated and true values decreases considerably, so that for large g , our generalized private allelic richness measures provide a close approximation to the true values in the setting of the simulation (Figure 2.1). Because this simulation is based on our human microsatellite dataset, its results suggest that it is reasonable to make interpretations about allelic distributions in human populations using our method applied to the data in our

example.

2.3.3 Results

Figure 2.2 shows the generalized private allelic richness at $g = 40$ for each of the 31 combinations of geographic regions, as a fraction of the sum of the 31 values. Examining the percentages of alleles having a given geographic distribution at $g = 40$, the average absolute difference across geographic distributions is 0.12% between the computations including and excluding relatives. Because of the similarity in results including and excluding relatives, our subsequent analyses use only one of the two datasets (the H952 subset excluding relatives).

Figure 2.3 and Table 2.1 compare the fractions of alleles having each of the 31 geographic distributions, for four values of g (10, 20, 30, and 40) as well as uncorrected for sample size. Notable in the figure and table is the emergence of alleles that were found in various combinations of two, three, and four major regions when correcting for sample size, but that did not appear in the uncorrected calculations. Additionally, we see that the uncorrected analysis produces a rather different view of the allelic distribution compared to the analyses that correct for sample size. For example, considering the distribution of private alleles across the major geographic regions, the uncorrected calculations indicate that Eurasia contains the most private alleles, followed by Africa, East Asia, the Americas, and Oceania. However, when we correct for sample size differences using $g = 40$, Africa has the largest number of private alleles, followed by Eurasia, Oceania, East Asia, and the Americas. Similarly, in the uncorrected calculations the region with the largest number of missing alleles (alleles private to four of the five regions) is Oceania (AfEuEaAm) followed by the Americas (AfEuEaOc); in the corrected calculations (standardized sample size of $g = 40$) missing alleles are most numerous for the Americas (AfEuEaOc) followed by Oceania (AfEuEaAm).

For each geographic region, the mean number of distinct alleles per locus and the mean number of private alleles per locus are shown in Figures 2.4A and 2.4B as functions of standardized sample size g . From these plots we see that Africa has both the highest number of distinct alleles and the highest number of private alleles, and that the smallest values in both categories occur in the Americas.

The numbers of alleles private to combinations of regions are plotted in Figure 2.5. Figure 2.5A shows the mean number of alleles per locus private to pairs of major regions, demonstrating that the combination of Africa and Eurasia has the largest number of private alleles. The smallest number is observed in the combination of Oceania and the Americas. The highest number of alleles private to three regions is seen in the combination of Africa, Eurasia, and East Asia, followed closely by the combination of Africa, Eurasia, and Oceania (Figure 2.5B). In the plot for the number of missing alleles (Figure 2.5C), we see that the Americas have by far the largest number, followed by Oceania and Africa. Figure 2.5D, which shows the mean number of alleles simultaneously present in all regions, illustrates that the number of alleles found in all regions considerably exceeds the number private to any one region or any combination of two, three, or four regions.

2.3.4 Out of Africa and the Peopling of Oceania

We can interpret the patterns of private allelic richness in Figures 2.4 and 2.5 in relation to our expectations based on various perspectives about the history of human migrations. The larger numbers of alleles and private alleles in Africa, and the smaller numbers in the Americas, match the pattern expected for models of human evolution that begin from an African origin and reach the Americas only after a series of founder events (*Ramachandran et al.*, 2005). The pair of regions with the largest number of alleles is the combination of the geographically connected regions of Africa and Eurasia; the group of three regions with the largest number is the combination

of Africa, Eurasia, and East Asia; and the group of four regions with the largest number is the combination of Africa, Eurasia, East Asia, and Oceania. These results each fit the prediction of African-origin models that include serial founder effects during outward migrations, as many alleles in the founding population would only have migrated along part of the path outside of Africa.

One set of results that offers the potential to distinguish among competing hypotheses about human migrations concerns alleles found in combinations of geographic regions that include Oceania. The initial peopling of near Oceania (which includes the islands of New Guinea and Bougainville, from where our samples originate) involves the first demonstrable human sea-crossing (*Derricourt, 2005*). Fossil evidence of the presence of anatomically modern humans in Sahul—the ancient landmass of Australia and New Guinea separated by sea from Asia—dates to at least 42,000-45,000 years before the present (BP) (*O’Connell and Allen, 2004; Gillespie, 2002*), and earlier dates ($\sim 60,000$ BP) have also been proposed (*O’Connell and Allen, 2004; Thorne et al., 1999*). Several migration waves have entered Oceania since the initial colonization, creating a complex mixture of ancestries in many parts of the region (*Matisoo-Smith, 2007; Friedlaender et al., 2008*).

A theory of a single main migration out of Africa ultimately reaching Oceania proposes a recent dispersal of modern humans from sub-Saharan Africa into Eurasia, replacing earlier archaic humans. There are at least two plausible out-of-Africa routes of dispersal towards eastern Asia—a northern inland route through the Middle East and a southern coastal route via Arabia and India (*Cavalli-Sforza et al., 1994; Quintana-Murci et al., 1999; Bulbeck, 2007; Field et al., 2007*). Because the existence of multiple routes suggests the possibility that two or more major migrations taking different paths may have occurred, it is of interest to examine whether an additional main out-of-Africa event—distinct from the events responsible for most of the peopling of Asia and Europe—might have been responsible for the peopling of

Sahul.

To investigate the possibility of a separate migration wave from Africa to Oceania, we can consider three simplified scenarios concerning human dispersal from Africa to Oceania that have the potential to be distinguished based on multilocus population-genetic data (Figure 2.6). The first scenario, Model 1, corresponds to a single primary out-of-Africa migration through the Middle East and East Asia before reaching Oceania. This hypothesis predicts that variation in Oceania is largely a subset of East Asian variation. The second scenario, Model 2, postulates a peopling of Oceania from Eurasia. In this model, following a migration out of Africa into western Asia, the migration that carried human populations into Oceania was separate from the migration into East Asia and left a negligible genetic trace along the path to Oceania. Under this hypothesis, variation in Oceania would largely be a subset of variation in Eurasia. Finally, the third scenario, Model 3, suggests an early peopling of Oceania, perhaps by a southern route out of Africa via the Arabian peninsula, the Indian sub-continent, and Southeast Asia. In this scenario, populations in Asia along the migration path would have only a small or negligible fraction of ancestry from the time of the initial colonization of Oceania, and would descend largely from later out-of-Africa migrations. Variation in Oceania would then be a subset of variation in Africa but not of variation in Eurasia or East Asia.

Comparisons of the numbers of alleles with various geographic distributions can assist in distinguishing these alternative hypotheses (Table 2.2). Figure 2.7 displays the rarefaction curves for the four pairs of geographic regions that include Oceania, a subset of the pairs shown in Figure 2.5A. The Africa/Oceania combination has more private alleles than the other three pairs, and the East Asia/Oceania and Eurasia/Oceania pairs have nearly equal numbers of private alleles. These observations are compatible with Model 3, in which Oceania would retain many ancestrally African alleles not found elsewhere. They are also compatible with Model

1, as the relatively high number of alleles each non-African group shares with Africa could be a consequence of the particularly high level of African variation. The similar numbers of alleles private to Eurasia/Oceania and East Asia/Oceania would then result from the opposing effects of a higher level of variation in Eurasia than in East Asia and a higher degree of relationship with Oceania in East Asia than in Eurasia. The observations, however, are not compatible with Model 2, which would have been expected to produce an excess number of alleles private to the combination of Eurasia and Oceania compared to the number private to the combination of East Asia and Oceania.

Further support for Model 3 can be found in various additional comparisons in Figure 2.5A. Under Model 1, Oceania and the Americas both derive from East Asian ancestry, and therefore, the combinations Africa/Oceania and Africa/America, Eurasia/Oceania and Eurasia/America, and East Asia/Oceania and East Asia/America are directly comparable. In each of these three cases, the pair including Oceania has more alleles than the pair including the Americas, consistent with the higher allelic richness in Oceania compared to the Americas. However, the amount by which the number of alleles private to the combination of Africa and Oceania exceeds the number of alleles private to the combination of Africa and the Americas is considerably greater than the corresponding excess for the other two comparisons. Moreover, with the exception of Africa/Eurasia, the Africa/Oceania combination has more alleles than any other pair of regions—including the combination of Africa and East Asia. These observations, which are compatible with Model 3, are more difficult to reconcile with Model 1.

Examination of combinations of three regions in Figure 2.5B produces similar suggestive evidence of Model 3 to that obtained from combinations of two regions in Figures 2.5A and 2.7. Except for the combination of Africa, Eurasia, and East Asia, the combination of Africa, Eurasia, and Oceania has more private alleles than

any other three-region combination. Although this observation could potentially be explained by any of the three models, the amount by which the number of private alleles for the Africa/Eurasia/Oceania combination exceeds that of other combinations is least compatible with Model 1, which has several groups of three regions that might have been expected to have numbers of private alleles close to that of Africa/Eurasia/Oceania.

2.4 Discussion

Our analysis of human microsatellite data using rarefaction has yielded a variety of insights into the effects on the analysis of allelic distributions across populations of varying sizes among population samples. First, assessing alleles private to combinations of populations can assist in dealing with sampling designs that are uneven. It can also uncover additional information that analyses of populations one at a time may be unable to identify. As observed in Figure 2.3, failing to correct for sample size can produce results that are different from those obtained with a sample size correction.

Our computer program *ADZE* (Allelic Diversity AnalyZEr) was used in our analyses and is available for download at <http://rosenberglab.bioinformatics.med.umich.edu/adze.html>. *ADZE* is capable of performing computations on multilocus data sets with one or many populations and arbitrarily many alleles per locus. As the number of combinations of populations quickly grows large with the number of populations, *ADZE* offers the option of calculating the number of private alleles only for combinations of a particular size. *ADZE* also features a missing data filter that discards loci found to have at least one population with a missing data percentage greater than or equal to a specified value.

The use of *ADZE* with microsatellite data from human populations has produced an excess similarity of populations from Africa and Oceania, potentially suggestive

of a migration of modern humans from Africa to Oceania separate from the primary migrations responsible for human population ancestry in Europe and Asia. Of note, other analyses of overlapping data have found greater genetic similarity between Oceania and East Asia than between Oceania and Africa (*Rosenberg et al.*, 2002; *Zhivotovsky et al.*, 2003). However, in previous work, genetic cluster analyses with two or three clusters detected a greater signal of similarity between Africa and Oceania than between Africa and East Asia (*Rosenberg et al.*, 2002), likely reflecting the high number of alleles private to the combination of Africa and Oceania that we have observed here.

As the models of colonization that we have examined are highly simplified idealizations of a complex process, some mixture of Models 1 and 3 likely provides an explanation more compatible with the full collection of results with these data than does Model 1 or Model 3 alone. Model 1 might potentially receive greater support given more data from Southeast Asia, and the observed low level of private alleles for the combination of Oceania and East Asia might have resulted from limited sampling in these regions. Formal testing based on a more diverse sample from Oceania will be important for assessing the relative importance of early migrations from Africa and more recent interaction with populations from East Asia. In any case, the use of the rarefaction approach to examine alleles private to combinations of populations can produce novel observations that augment those obtained from other methods of data analysis, and that can form the basis for hypothesis tests with increasingly complex evolutionary models.

2.5 Acknowledgments

Support was provided by NIH grants R01 GM081441 and T32 GM070449, by a University of Michigan Center for Genetics in Health and Medicine postdoctoral fellowship (M.J.), by a Burroughs Wellcome Fund Career Award in the Biomedical

Sciences (N.A.R.), and by an Alfred P. Sloan Research Fellowship (N.A.R.).

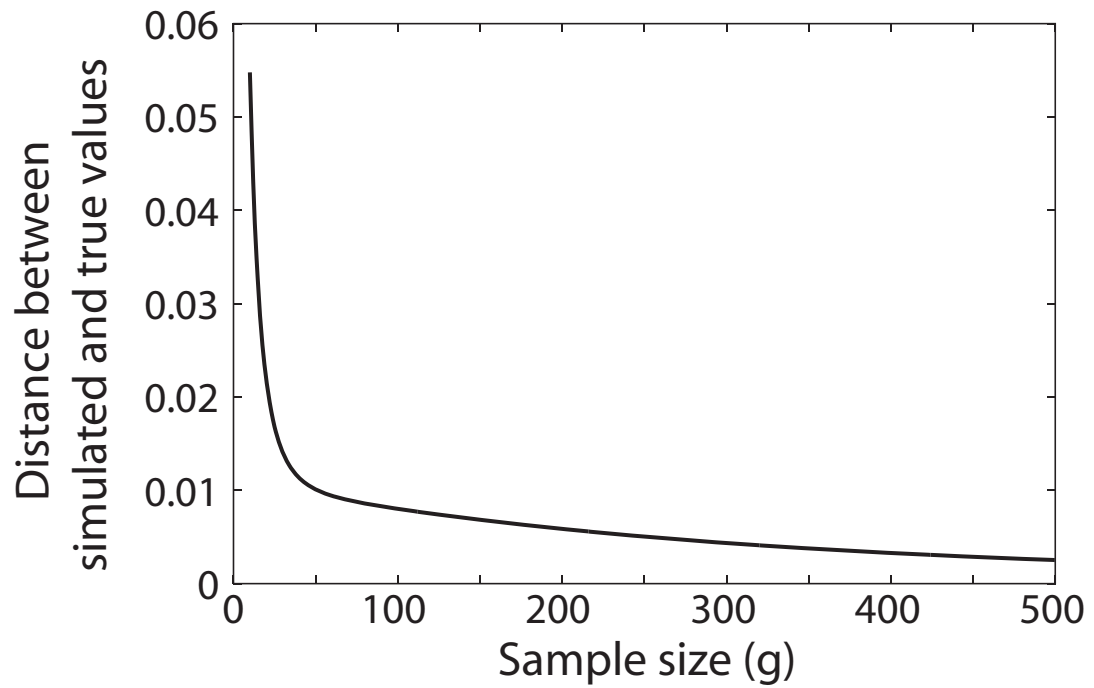


Figure 2.1: The distance between simulated and true values of the proportions of alleles with specific geographic distributions, summed across distributions and plotted as a function of standardized sample size g from 10 to 500. Results shown represent the mean across 100 replicates.

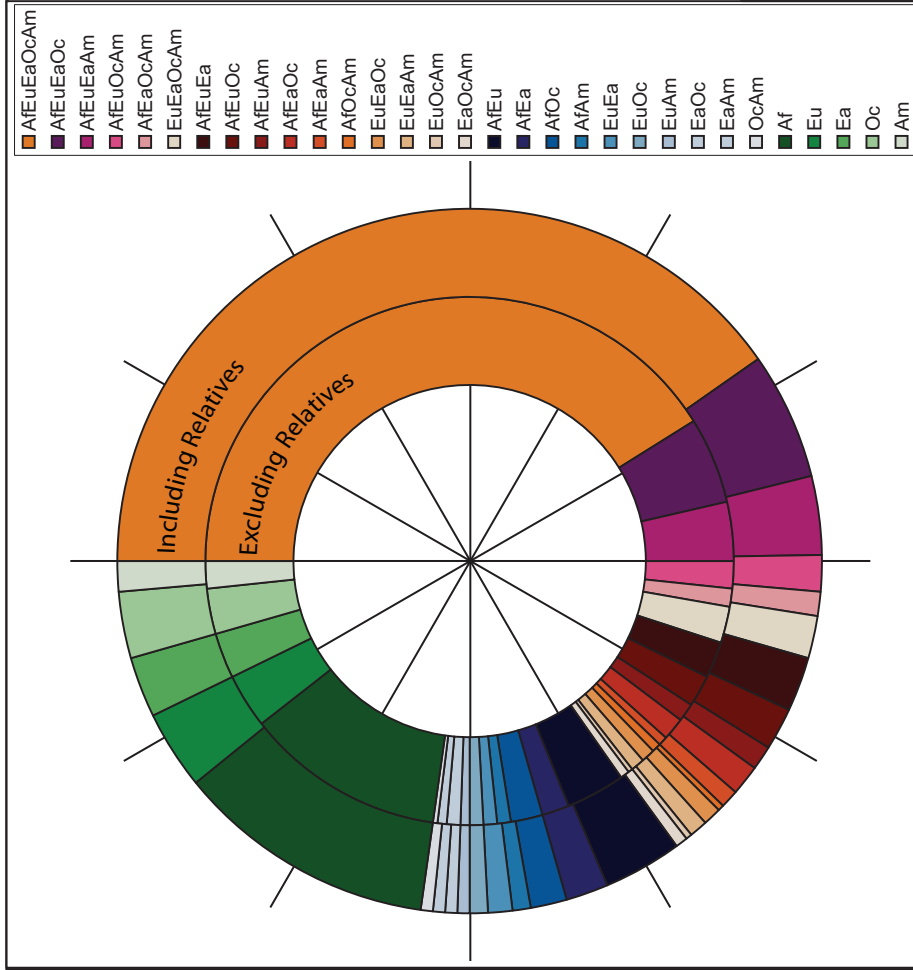


Figure 2.2: The number of alleles private to various combinations of geographic regions as a fraction of the total at a standardized sample size of 40. The outer circle corresponds to the entire dataset including known first and second degree relatives. The inner circle corresponds to the subset that excludes known relatives. Values are represented in proportion to the appropriate angles measured at the center of the circle. Geographic regions are abbreviated: Af – Africa; Eu – Eurasia; Ea – East Asia; Oc – Oceania; Am – Americas.

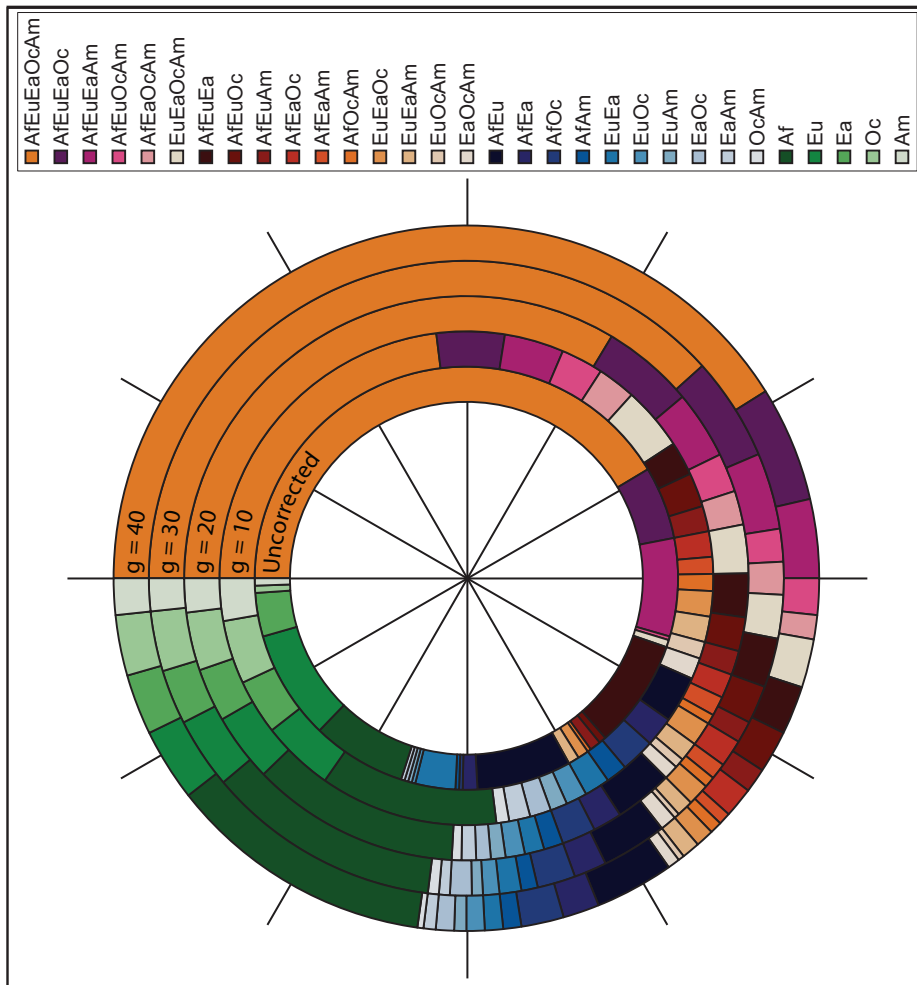


Figure 2.3: The number of alleles private to various combinations of geographic regions as a fraction of the total, using a subset of the data excluding relatives. The innermost circle corresponds to calculations uncorrected for sample size variation. Moving outward the next rings correspond to calculations at standardized sample sizes of 10, 20, 30, and 40. Values are represented in proportion to the appropriate angles measured at the center of the circle. Geographic regions are abbreviated: Af – Africa; Eu – Eurasia; Ea – East Asia; Oc – Oceania; Am – Americas.

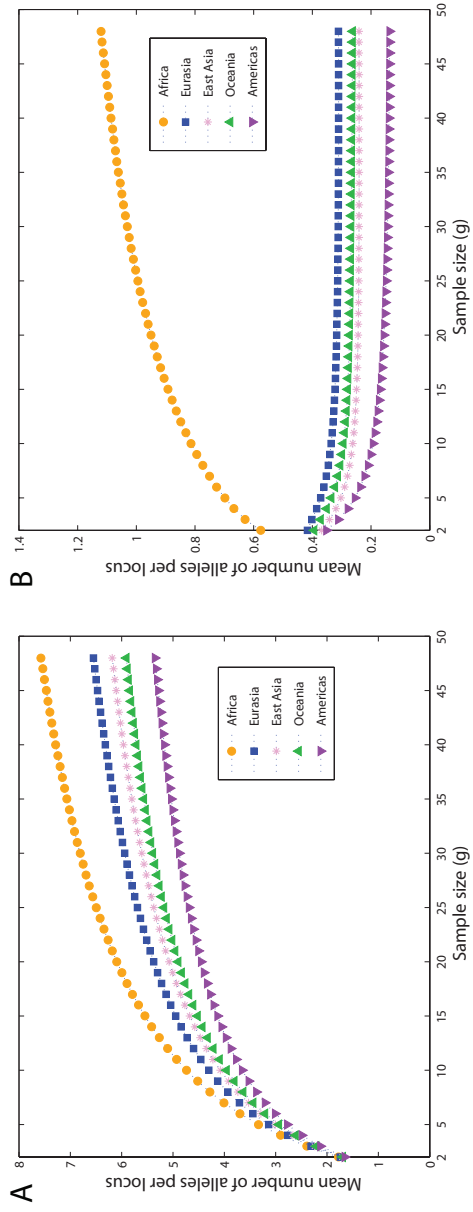


Figure 2.4: The mean number of (A) distinct alleles per locus and (B) private alleles per locus, as functions of standardized sample size for five major geographic regions (excluding known relatives).

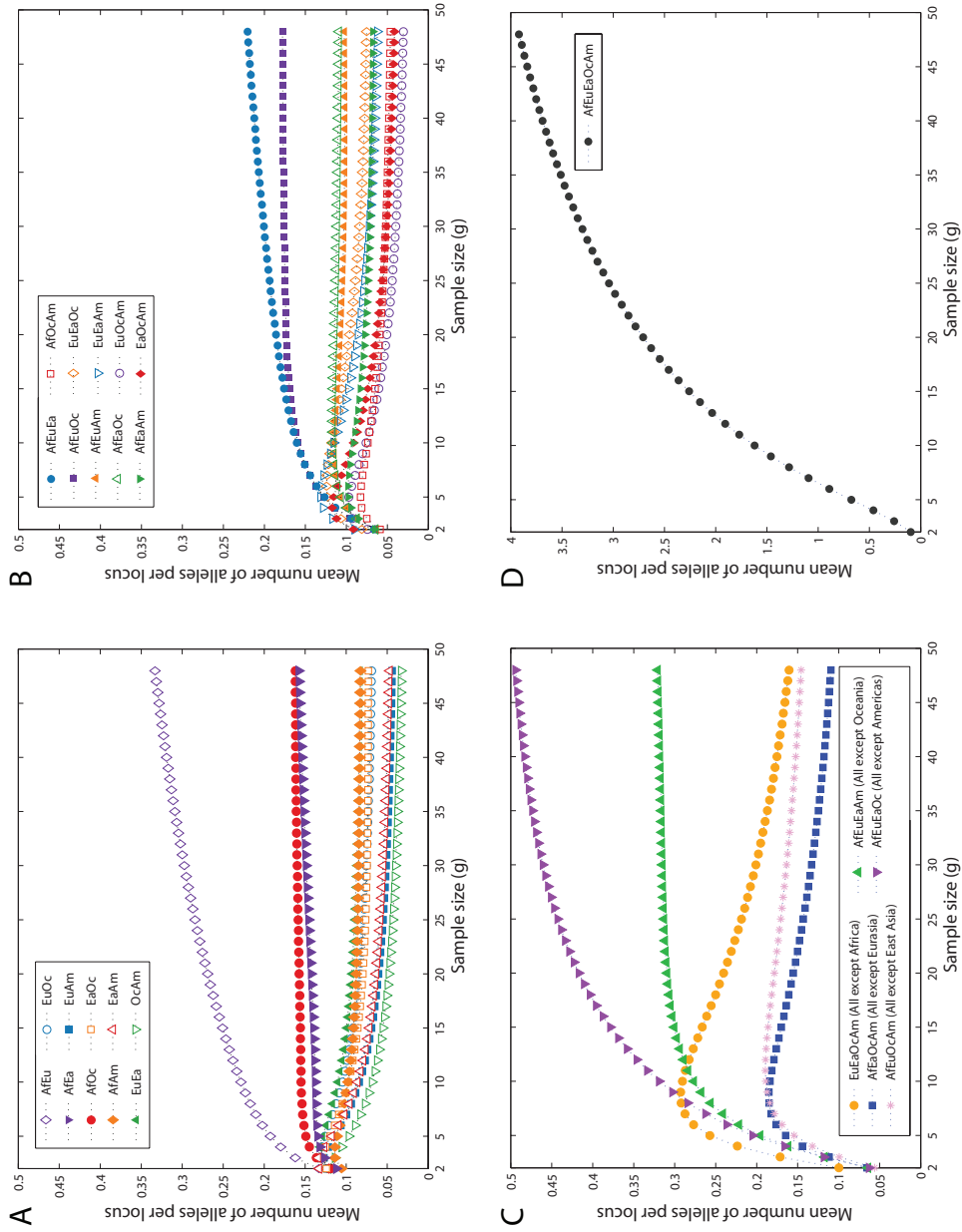


Figure 2.5: The mean number of alleles per locus private to combinations of k of five major geographic regions as a function of standardized sample size (excluding known relatives). (A) $k = 2$, (B) $k = 3$, (C) $k = 4$, and (D) $k = 5$. Geographic regions are abbreviated: Af – Africa; Eu – Eurasia; Ea – East Asia; Oc – Oceania; Am – Americas.

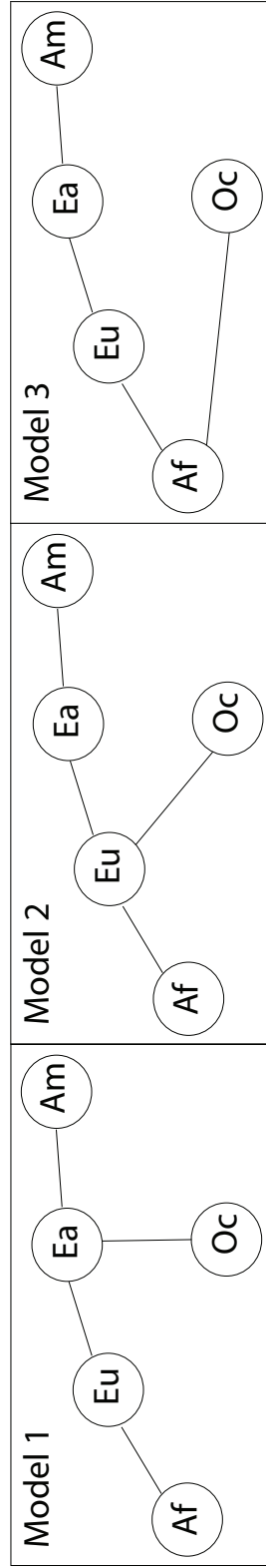


Figure 2.6: Three hypothetical migration patterns to Oceania out of Africa. Geographic regions are abbreviated: Af – Africa; Eu – Eurasia; Ea – East Asia; Oc – Oceania; Am – Americas.

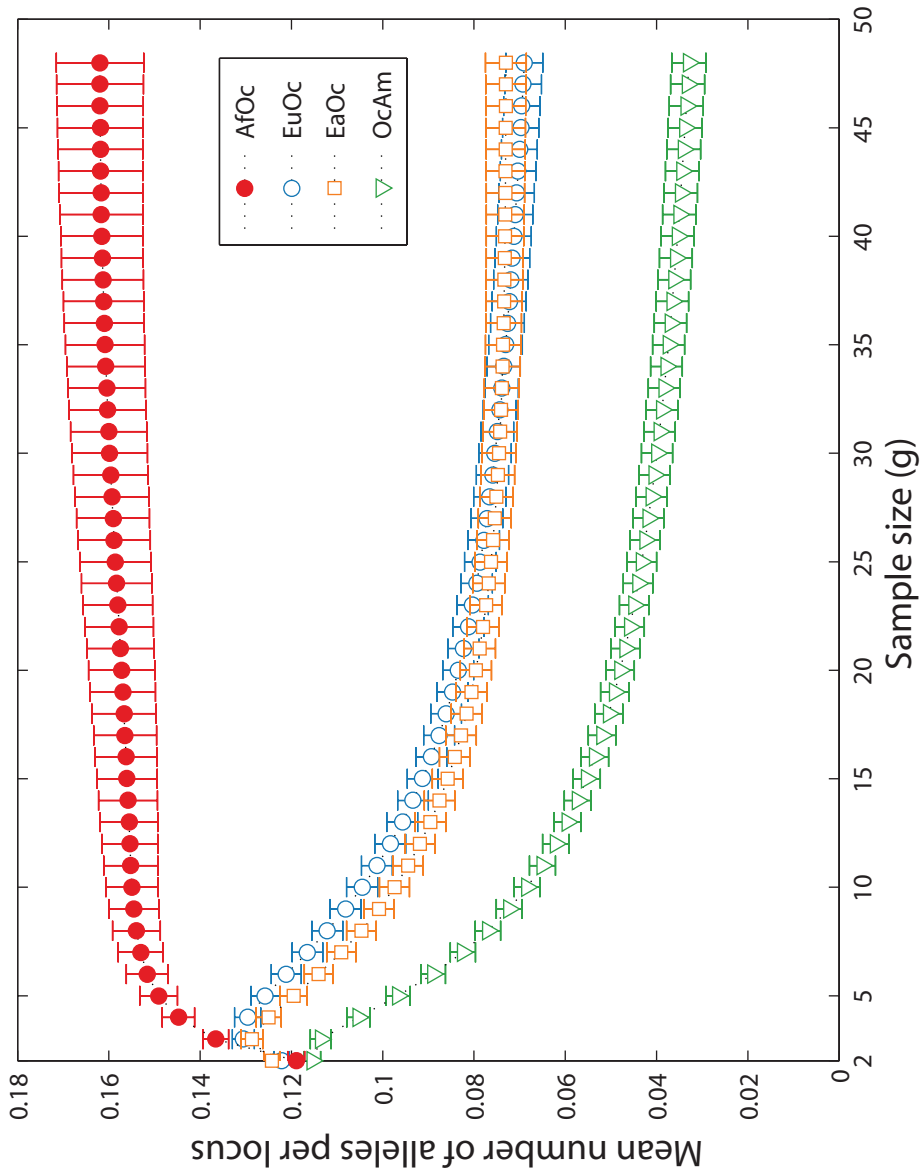


Figure 2.7: The mean number of alleles per locus private to the combination of Oceania and another major geographic region as a function of standardized sample size (excluding known relatives). Error bars represent standard error of the mean across loci. Geographic regions are abbreviated: Af – Africa; Eu – Eurasia; Ea – East Asia; Oc – Oceania; Am – Americas.

Table 2.1: Percentages of 8516 total alleles private to various combinations of geographic regions. These percentages are obtained both uncorrected and corrected for sample size, excluding known first and second degree relatives in the computation. The value of g gives the size of subsamples used in the sample size correction. Geographic regions are abbreviated: Af – Africa; Eu – Eurasia; Ea – East Asia; Oc – Oceania; Am – Americas.

Geographic Distribution	Uncorrected	$g = 10$	$g = 20$	$g = 30$	$g = 40$
AfEuEaOcAm	41.67	22.95	33.65	38.36	41.00
AfEuEaOc	5.70	4.51	5.18	5.35	5.37
AfEuEaAm	7.40	3.92	3.84	3.66	3.53
AfEuOcAm	0.18	2.67	2.22	1.90	1.69
AfEaOcAm	0.04	2.60	1.93	1.55	1.31
EuEaOcAm	0.54	4.11	2.99	2.34	1.95
AfEuEa	8.14	2.27	2.31	2.33	2.36
AfEuOc	0.59	2.25	2.15	2.04	1.97
AfEuAm	0.76	1.63	1.34	1.20	1.13
AfEaOc	0.08	1.65	1.43	1.30	1.22
AfEaAm	0.04	1.29	0.96	0.83	0.76
AfOcAm	0.02	1.04	0.73	0.60	0.54
EuEaOc	0.88	1.66	1.19	0.97	0.86
EuEaAm	0.96	1.63	1.07	0.84	0.73
EuOcAm	0.04	1.08	0.63	0.46	0.38
EaOcAm	0.02	1.29	0.79	0.59	0.50

Geographic Distribution	Uncorrected	$g = 10$	$g = 20$	$g = 30$	$g = 40$
AfEu	7.15	3.23	3.34	3.46	3.55
AfEa	1.17	1.95	1.77	1.73	1.71
AfOc	0.26	2.19	1.95	1.86	1.80
AfAm	0.18	1.38	1.09	0.99	0.93
EuEa	3.10	1.56	1.16	1.01	0.94
EuOc	0.38	1.48	1.04	0.88	0.79
EuAm	0.28	1.19	0.75	0.60	0.52
EaOc	0.15	1.38	0.99	0.87	0.81
EaAm	0.22	1.22	0.79	0.64	0.55
OcAm	0.01	0.97	0.60	0.46	0.39
Af	7.32	11.51	11.80	11.98	12.08
Eu	8.20	4.76	3.94	3.63	3.45
Ea	3.36	3.71	3.03	2.80	2.67
Oc	0.58	4.16	3.40	3.11	2.94
Am	0.62	2.73	1.94	1.67	1.54

Table 2.2: Comparisons of numbers of private alleles for pairs of geographic regions that would support a given migration model if observed. Geographic regions are abbreviated: Af – Africa; Eu – Eurasia; Ea – East Asia; Oc – Oceania; Am – Americas.

Model 1	Model 2	Model 3
$AfEa > AfOc$	$EuOc > EaOc$	$AfOc > EuOc$
		$AfOc > EaOc$

CHAPTER III

On the size distribution of private microsatellite alleles

3.1 Introduction

Private alleles are alleles that are found only in a single population among a broader collection of populations. These alleles have proven to be informative for diverse types of population-genetic studies, in such areas as molecular ecology and conservation genetics (e.g. *Petit et al.*, 1998; *Parker et al.*, 1999; *Fiumera et al.*, 2000; *Neel and Cummings*, 2003; *Torres et al.*, 2003; *Kalinowski*, 2004) and human evolutionary genetics (e.g. *Neel*, 1973, 1978; *Neel and Thompson*, 1978; *Calafell et al.*, 1998; *Schroeder et al.*, 2007; *Szpiech et al.*, 2008).

Some of the first investigations of private alleles trace to studies of private electrophoretic variants in Native American groups from South America (*Neel*, 1973, 1978; *Neel and Thompson*, 1978). Using private alleles, Neel and colleagues obtained estimates of mutation rates in these populations. *Slatkin* (1985) and *Barton and Slatkin* (1986) showed that private alleles can contribute to indicators of gene flow, finding in theoretical models of population structure that the occurrence of private alleles was related to with the mean number of migrants exchanged per generation between populations. Private alleles have also been used in empirical studies of human

migrations. *Calafell et al.* (1998) noted that in human populations, the mean number of private alleles is greater in Africa, providing support to models of human migration out of Africa. *Schroeder et al.* (2007) argued on the basis of a private allele ubiquitous in the Americas that all modern Native American populations are descended from the same founding population.

One recent study, which investigated 678 microsatellite markers in 29 Native American populations from North, Central, and South America (*Wang et al.*, 2007), identified a peculiar property of private alleles. *Wang et al.* (2007) characterized the distribution of private alleles across four subregions in the Americas, observing that private microsatellite alleles were found in the tails rather than in the interior of the allele size distribution more often than was expected by chance. In other words, private alleles at a locus frequently had very long or very short repeat lengths with respect to the other alleles at the locus.

Here we take a modeling approach to examine the reasons underlying the frequent occurrence of private alleles on the edges of the allele size distribution. Using a simple coalescent model, we assess the properties of microsatellite private alleles, thereby helping to explain patterns that exist in the relationship between privacy and allele size across human populations.

3.2 Theory

Let $\{x_1x_2/x_3x_4\}$ denote four sampled microsatellite alleles in two populations, where x_i indicates the allele size for sampled allele i , and the forward slash separates alleles from different populations. We restrict our attention to cases with four alleles; a scenario with two alleles each in two populations gives the smallest sample size useful for examining the phenomenon of interest, as we will explain below. Because the 4-allele case involves a tractable number of calculations, it is possible in this case to mathematically investigate the position of private alleles in the size distribution.

We map sets of four allele sizes in two populations to one of seven possible configurations of identity and nonidentity, using the letters A , B , C , and D to denote distinct allele sizes. Thus, if two sampled alleles are identical by state (IBS), we indicate this identity by assigning the alleles the same letter. For example, if all four sampled alleles are IBS, we represent the allele configuration by $\{AA/AA\}$. If one allele in population 1 is IBS to an allele in population 2 and the other allele in population 1 is IBS to the other allele in population 2 (and no alleles are IBS within populations), then we represent the allele configuration by $\{AB/AB\}$. We label the seven possible configurations by C_i for $i \in \{1, \dots, 7\}$, and we list them in Table 3.1.

We are interested in comparing private and shared alleles on the basis of size. In particular, we wish to examine whether alleles lie on the edges of the size distribution, that is, whether they have the longest or shortest lengths. To have a sensible definition of the “edges” and interior of the allele size distribution, we must have at least three distinct alleles among the four sampled alleles that we consider. Furthermore, because we are concerned with the location of private alleles with respect to shared alleles, we must have at least one shared allele and one private allele. The only one of the seven configurations of four alleles that satisfies both of these requirements—and that therefore enables a computation of the probability that private alleles lie on the edges of the allele size distribution—is C_6 (configuration $\{AB/AC\}$). This configuration, with sample size four, provides the smallest scenario that contains both private alleles and shared alleles and that contains both edges and an interior of the allele size distribution. We aim to compute the probability that B and C , the two private alleles in configuration $\{AB/AC\}$, both lie on the edges of the size distribution, conditional on this configuration being produced.

3.2.1 A naïve argument

If we disregard the genealogical relatedness of the alleles in our two-population four-allele model, what do we expect for the probability that the private alleles lie on the edges? There are six possible orderings of the three allele sizes A , B , and C ($A < B < C$, $A < C < B$, etc.), and, if no relationship exists between the size of an allele and its status as shared or private, we expect the six orderings to be equiprobable. Two of the six orderings place the private alleles B and C on the edges of the size distribution. Under this simple argument, we would expect the probability that both private alleles lie on the edges to be $1/3$.

This argument gives an initial sense of what might be predicted for the probability that the private alleles lie on the edges of the size distribution. However, it disregards the fact that the alleles are related through a common ancestor. We now turn to a genealogical argument that more directly models this relationship.

3.2.2 The probability of microsatellite configurations

To account for the genealogical relatedness of the four alleles in obtaining a prediction of the probability that private alleles lie on the edges of the allele size distribution, we use the coalescent with symmetric stepwise mutation. Initially, we consider the two populations to have instantaneously diverged zero coalescent time units in the past ($t_d = 0$). Later, we will consider arbitrary values of the divergence time t_d .

To calculate the desired probability, we first condition on the $\{AB/AC\}$ allele configuration (configuration C_6), the mutation rate, and the coalescence times of the genealogy. By considering the probability of a net change by d mutational steps along a genealogical branch, we construct the joint probability of an allele configuration and a particular labeled history for the four alleles, where the allele configuration refers to one of the seven scenarios in Table 3.1 and the labeled history refers to the sequence

of coalescences (Figure 3.1). We then calculate the total probability that the private alleles lie on the edges of the allele size distribution, summing across all labeled histories, and integrating over coalescence times to arrive at the desired probability, conditional only on a mutation rate θ .

Consider the events E_1 : $\text{size}(B) < \text{size}(A) < \text{size}(C)$, and E_2 : $\text{size}(C) < \text{size}(A) < \text{size}(B)$. These events are equiprobable, and we aim to calculate the probability

$$\mathbb{P}[E_1 \cup E_2 | C_6, \theta] = \frac{2\mathbb{P}[E_1, C_6 | \theta]}{\mathbb{P}[C_6 | \theta]}. \quad (3.1)$$

Under the symmetric single stepwise mutation model, a microsatellite allele can mutate by only one step at a time in either a positive or negative direction, and the probability of mutating +1 step is equal to the probability of mutating -1 step, independent of the size of the allele. We work with coalescent time units (units of $2N_e$ generations, where N_e is the effective size of each population, treated as containing diploid individuals) and with the population-scaled mutation rate $\theta = 4N_e\mu$, where μ is the per-locus per-generation mutation rate.

3.2.3 Mutations on a genealogical branch

The probability that a marker evolving according to the symmetric stepwise mutation model with population-scaled mutation rate θ has net change d units along a branch of length t coalescent time units is (*Wehrhahn, 1975; Wilson and Balding, 1998*)

$$f(|d|; t, \theta) = e^{-t\theta/2} I_{|d|}(t\theta/2), \quad (3.2)$$

where

$$I_{|d|}(x) = \sum_{k=0}^{\infty} \frac{(x/2)^{(2k+|d|)}}{k!(k+|d|)!}$$

is the modified Bessel function of the first kind (*Gradshteyn and Ryzhik, 2000*). Because positive and negative mutations are equally likely, we write f as a function

of $|d|$ rather than d , which can be positive, negative or zero.

3.2.4 Probability of the set of allele sizes on a genealogical tree

We can use Equation 3.2 to calculate the probability that changes along a coalescent tree ultimately give rise to a specified set of allele sizes. Considering that each branch evolves independently of the others, we calculate the probabilities of changes along individual branches and then multiply probabilities across branches to get the joint probability of all changes on the tree. There are two unlabeled topologies that we need to consider: an asymmetric topology (Figure 3.2A) and a symmetric topology (Figure 3.2B). Each topology is parameterized by a vector of allele sizes, $(n_2, n_3, x_1, x_2, x_3, x_4)$, and a vector of coalescence times, (t_2, t_3, t_4) . The x variables represent the sizes of alleles at the leaf nodes, and n_i represents the size of the allele at the interior node located at the reduction of the number of distinct lineages to i . The coalescence time t_i represents the length of time during which there exist i distinct lineages. Initially, we treat the coalescence times as fixed, and later we will integrate the probabilities against the density of coalescence times to obtain a probability unconditional on t_2 , t_3 , and t_4 . Because we assume that mutation probabilities do not depend on allele size, we can set the allele size of the most recent common ancestor of the four-allele sample (the root node) to 0 without loss of generality. However, following a choice similar to that of *Pritchard and Feldman* (1996) and *Zhang and Rosenberg* (2007), we instead choose to set $n_2 = 0$ rather than setting the root node to 0, and we treat the two branches that descend from the root as one branch with length equal to the sum of the lengths of its two constituent branches. This choice makes it possible to consider coalescent trees with five rather than six separate branches, thereby simplifying the computation.

Considering the asymmetric caterpillar topology (Figure 3.2A), we obtain the joint probability of $(n_2, n_3, x_1, x_2, x_3, x_4)$ given (t_2, t_3, t_4) by calculating the probability of

changing from n_2 to x_4 repeats along a branch of length $2t_2 + t_3 + t_4$, from n_2 to x_3 repeats along a branch of length $t_3 + t_4$, from n_2 to n_3 repeats along a branch of length t_3 , from n_3 to x_2 repeats along a branch of length t_4 , and from n_3 to x_1 repeats along a branch of length t_4 . Assuming $n_2 = 0$ and multiplying these five probabilities together gives

$$V^{cat}(x_1, x_2, x_3, x_4, n_3, \Psi) = f(|x_4|; 2t_2 + t_3 + t_4, \theta) \times f(|x_3|; t_3 + t_4, \theta) \times \\ f(|n_3|; t_3, \theta) \times f(|n_3 - x_2|; t_4, \theta) \times f(|n_3 - x_1|; t_4, \theta),$$

where $\Psi = (\tau, \theta)$ is a vector of parameters and $\tau = (t_2, t_3, t_4)$ is the vector of coalescence times. Similarly, for the symmetric topology, we calculate the probability of the set of allele sizes in Figure 3.2B to get

$$V^{sym}(x_1, x_2, x_3, x_4, n_3, \Psi) = f(|n_3|; 2t_2 + t_3, \theta) \times f(|n_3 - x_4|; t_4, \theta) \times \\ f(|n_3 - x_3|; t_4, \theta) \times f(|x_2|; t_3 + t_4, \theta) \times f(|x_1|; t_3 + t_4, \theta).$$

3.2.5 Assigning alleles the roles of A , B , and C

There are 18 labeled histories for the alleles $\{x_1, x_2, x_3, x_4\}$, which we denote by T_i for $i \in \{1, \dots, 18\}$ (Figure 3.1). We can then calculate $\mathbb{P}[C_6|T_i, \Psi]$ by considering all possible ways to get configuration C_6 with labeled history T_i . Because we have defined $\{x_1, x_2\}$ to be in population 1 and $\{x_3, x_4\}$ to be in population 2, we need to consider four cases for each history, reflecting the four possible assignments of the allele sizes x_1, x_2, x_3 , and x_4 to the roles of distinct alleles A, B , and C . These four cases are shown in Table 3.2.

If we represent the size of the shared allele (allele A) by n_A and the sizes of the two private alleles (B and C) by n_B and n_C , respectively, then we can calculate $\mathbb{P}[C_6|T_i, \Psi]$ by summing the individual probabilities of each of the four cases in Table

3.2. For example, consider T_1 :

$$\begin{aligned}
\mathbb{P}[C_6|T_1, \Psi] = & \sum_{n_3=-\infty}^{\infty} \sum_{n_A=-\infty}^{\infty} \sum_{n_B=-\infty}^{\infty} \sum_{n_C=-\infty}^{\infty} V^{cat}(n_A, n_B, n_A, n_C, n_3, \Psi) + \\
& V^{cat}(n_A, n_B, n_C, n_A, n_3, \Psi) + \\
& V^{cat}(n_B, n_A, n_A, n_C, n_3, \Psi) + \\
& V^{cat}(n_B, n_A, n_C, n_A, n_3, \Psi). \tag{3.3}
\end{aligned}$$

Here, without loss of generality, we treat the private allele in the first population as the B allele and the private allele in the second population as the C allele. Similar calculations can be performed for the 17 remaining labeled histories (Table 3.3).

3.2.6 Summing over labeled histories

In order to calculate $\mathbb{P}[E_1, C_6|\Psi]$, we proceed exactly as in Equation 3.3, conditioning on each history T_i , but we restrict the bounds of summation on n_B and n_C to $-\infty < n_B < n_A$ and $n_A < n_C < \infty$, respectively.

We now have

$$\mathbb{P}[C_6|\Psi] = \sum_{i=1}^{18} \mathbb{P}[C_6|T_i, \Psi] \mathbb{P}[T_i|\Psi] \tag{3.4}$$

and

$$\mathbb{P}[E_1, C_6|\Psi] = \sum_{i=1}^{18} \mathbb{P}[E_1, C_6|T_i, \Psi] \mathbb{P}[T_i|\Psi]. \tag{3.5}$$

Here, $\mathbb{P}[T_i|\Psi] = 1/18$ for all i because each labeled history of four lineages is equally likely under the assumption of the coalescent process that lineages join randomly going back in time. Note that symmetries exist in V^{cat} and V^{sym} as a result of exchangeability of certain nodes in the topologies that they consider. For asymmetric topologies,

$$V^{cat}(W, X, Y, Z, n_3, \Psi) = V^{cat}(X, W, Y, Z, n_3, \Psi).$$

For symmetric topologies,

$$V^{sym}(W, X, Y, Z, n_3, \Psi) = V^{sym}(X, W, Y, Z, n_3, \Psi) =$$

$$V^{sym}(W, X, Z, Y, n_3, \Psi) = V^{sym}(X, W, Z, Y, n_3, \Psi).$$

Using the list of probability contributions for each labeled history, as given in Table 3.3, we can exploit these symmetries and collect like terms across labeled histories to write Equation 3.4 as

$$\mathbb{P}[C_6|\Psi] = \sum_{n_3=-\infty}^{\infty} \sum_{n_A=-\infty}^{\infty} \sum_{n_B=-\infty}^{\infty} \sum_{n_C=-\infty}^{\infty} \left(4V^{cat}(n_A, n_A, n_B, n_C, n_3, \Psi) + \right.$$

$$4V^{cat}(n_A, n_A, n_C, n_B, n_3, \Psi) + 8V^{cat}(n_A, n_B, n_A, n_C, n_3, \Psi) +$$

$$8V^{cat}(n_A, n_B, n_C, n_A, n_3, \Psi) + 8V^{cat}(n_A, n_C, n_A, n_B, n_3, \Psi) +$$

$$8V^{cat}(n_A, n_C, n_B, n_A, n_3, \Psi) + 8V^{cat}(n_B, n_C, n_A, n_A, n_3, \Psi) +$$

$$8V^{sym}(n_A, n_B, n_A, n_C, n_3, \Psi) + 8V^{sym}(n_A, n_C, n_A, n_B, n_3, \Psi) +$$

$$\left. 4V^{sym}(n_A, n_A, n_B, n_C, n_3, \Psi) + 4V^{sym}(n_B, n_C, n_A, n_A, n_3, \Psi) \right) \times$$

$$\mathbb{P}[T_i|\Psi] \tag{3.6}$$

and Equation 3.5 as

$$\begin{aligned}
\mathbb{P}[E_1, C_6 | \Psi] = & \sum_{n_3=-\infty}^{\infty} \sum_{n_A=-\infty}^{\infty} \sum_{n_B=-\infty}^{n_A-1} \sum_{n_C=n_A+1}^{\infty} \left(4V^{cat}(n_A, n_A, n_B, n_C, n_3, \Psi) + \right. \\
& 4V^{cat}(n_A, n_A, n_C, n_B, n_3, \Psi) + 8V^{cat}(n_A, n_B, n_A, n_C, n_3, \Psi) + \\
& 8V^{cat}(n_A, n_B, n_C, n_A, n_3, \Psi) + 8V^{cat}(n_A, n_C, n_A, n_B, n_3, \Psi) + \\
& 8V^{cat}(n_A, n_C, n_B, n_A, n_3, \Psi) + 8V^{cat}(n_B, n_C, n_A, n_A, n_3, \Psi) + \\
& 8V^{sym}(n_A, n_B, n_A, n_C, n_3, \Psi) + 8V^{sym}(n_A, n_C, n_A, n_B, n_3, \Psi) + \\
& \left. 4V^{sym}(n_A, n_A, n_B, n_C, n_3, \Psi) + 4V^{sym}(n_B, n_C, n_A, n_A, n_3, \Psi) \right) \times \\
& \mathbb{P}[T_i | \Psi]. \tag{3.7}
\end{aligned}$$

3.2.7 Integrating out the coalescence times

Finally, we integrate over the density of coalescence times under the standard coalescent model. Under this model, the time in coalescent time units (units of $2N_e$ generations) for i lineages to coalesce to $i - 1$ lineages is exponentially distributed with rate $\binom{i}{2}$ (*Wakeley, 2009*). Separate coalescence times are independent, and we can write their joint distribution in the four-taxon case as $\rho(t_2, t_3, t_4) = \binom{2}{2} e^{-\binom{2}{2}t_2} \binom{3}{2} e^{-\binom{3}{2}t_3} \binom{4}{2} e^{-\binom{4}{2}t_4} = 18e^{-t_2-3t_3-6t_4}$. Using this density, we integrate to get

$$\mathbb{P}[C_6 | \theta] = \int_0^{\infty} \int_0^{\infty} \int_0^{\infty} \mathbb{P}[C_6 | \Psi] \rho(t_2, t_3, t_4) dt_2 dt_3 dt_4 \tag{3.8}$$

and

$$\mathbb{P}[E_1, C_6 | \theta] = \int_0^{\infty} \int_0^{\infty} \int_0^{\infty} \mathbb{P}[E_1, C_6 | \Psi] \rho(t_2, t_3, t_4) dt_2 dt_3 dt_4. \tag{3.9}$$

3.2.8 Implementing the computation

To calculate $\mathbb{P}[E_1 \cup E_2 | C_6, \theta]$ (Equation 3.1) in practice, we use two approaches, a numerical method and a simulation-based method.

3.2.8.1 Numerical computation

First, we employ Gaussian quadrature to numerically estimate the numerator ($2\mathbb{P}[E_1, C_6 | \theta]$, Equation 3.9) and denominator ($\mathbb{P}[C_6 | \theta]$, Equation 3.8) of $\mathbb{P}[E_1 \cup E_2 | C_6, \theta]$. In order to compute the integrals in finite time, we estimate the expression $e^{-t\theta/2} I_{|d|}(t\theta/2)$ using the GNU Scientific Library (GSL) function `gsl_sf_bessel_In_scaled(|d|, t\theta/2)`. Additionally, we truncate the bounds of the infinite sums embedded in $2\mathbb{P}[E_1, C_6 | \theta]$ and $\mathbb{P}[C_6 | \theta]$ to ± 10 instead of $\pm \infty$. These limits provide bounds on the size that an allele can have at any particular node. We additionally integrate all time parameters from 0 to 10 rather than from 0 to ∞ . For small values of θ , these approximations are very accurate, as it is unlikely that an allele will mutate more than a few steps away from its initial number of repeats. However, for large θ , the approximation will become less accurate, as large numbers of mutations are likely to occur. These mutations ultimately cause alleles to shift further from the initial base size and beyond the arbitrary truncation in our approximation, so that the calculation fails to account for a non-trivial portion of probability mass.

3.2.8.2 Simulation-based computation

In order to calculate $\mathbb{P}[E_1 \cup E_2 | C_6, \theta]$ accurately for large θ , we obtain the ratio in Equation 3.1 directly by simulating the coalescent and mutation processes and tabulating the outcomes of interest. The simulation proceeds as follows.

1. Beginning with $k = 4$ alleles, arbitrarily define two alleles to be in one population and the other two alleles to be in the other population.

2. Generate a random time to coalescence from an exponential($\binom{k}{2}$) distribution.
3. Randomly choose two alleles to coalesce; set $k = k - 1$.
4. If $k \neq 1$, go to 2.
5. For each branch of the genealogy, generate a random number of mutation events, x , from a Poisson distribution with rate $\theta t/2$, where t is the branch length.
6. Assign each mutation a value of $+1$ or -1 by sampling the number of $+1$ mutations from a binomial($x, 1/2$) distribution. Those mutations not chosen to be $+1$ are assigned a value of -1 .
7. Determine the allele size of each of the four sampled alleles by summing the net value of mutations from the root (allele size 0) down to the leaves.
8. Classify the collection of four alleles into one of the seven allele configurations (Table 3.1).
9. If the alleles are in the C_6 configuration, accept the simulation and determine if the sizes of the private alleles (B and C) are on the ends of the distribution ($n_B < n_A < n_C$ or $n_C < n_A < n_B$). If yes, count a success.

By repeating this algorithm until the number of accepted simulations reaches some pre-specified number (we choose 1,000,000), we can estimate the probability that the private alleles lie on the edges of the size distribution by simply dividing the number of successes by the number of accepted simulations.

Note that the proportion of simulations that have configuration C_6 provides an estimate of $\mathbb{P}[C_6|\theta]$. Through a separate application of 10^6 iterations of steps 1 to 8, we estimate the probabilities of all seven configurations as functions of θ . These estimates appear in Figure 3.3. At small values of θ , we see that most simulations produce configuration C_1 ($\{AA/AA\}$), a sensible result because mutations are unlikely

to happen for small θ . As θ grows larger, more mutations occur, and we see that configurations with two or more distinct alleles begin to rise in frequency. For large values of θ , mutations happen so often that most trees have configuration C_7 ($\{AB/CD\}$).

Figure 3.4 shows, as a function of θ , the probability of interest, $\mathbb{P}[E_1 \cup E_2 | C_6, \theta]$, calculated both by simulation and numerically. Because we must truncate the internal sums for the numerical computation, we plot several numerical calculations at varying truncation values. Most of the numerical computations are quite accurate at small θ : we expect few mutations in this case, and the approximation made by truncating the sums will reasonably cover most of the probability mass. We see that as θ becomes large, the numerical results differ from the simulation-based result; at large θ many mutations occur and the numerical approximation is poorer.

We note that the probability of interest appears to level off well above the naïve calculation of $1/3$ as the mutation rate grows large. Furthermore, as θ tends toward zero, we see that the probability remains above $1/3$ and appears to tend toward $1/2$. We can prove this small- θ limiting result by considering a parsimony-style approximation for our probability near $\theta = 0$.

3.2.9 Small- θ approximation

We can make some simplifications to approximate our calculation of $\mathbb{P}[E_1 \cup E_2 | C_6, \theta]$ (Equation 3.1) in the limit as θ becomes small. For small θ , we expect fewer superfluous mutations to occur along a branch with a change of d steps—that is, we expect fewer mutations in one direction to be canceled by mutations in the other direction. Therefore, for very small θ , we can approximate the probability of changing d steps along a branch length t by setting $k = 0$ in Equation 3.2 so that no extra mutations occur. Denoting the small- θ approximation to $f(|d|, t, \theta)$ by $f_s(|d|, t, \theta)$, we

then obtain

$$f_s(|d|; t, \theta) = e^{-t\theta/2} \frac{(t\theta/4)^{|d|}}{|d|!}.$$

Furthermore, for small θ , we also expect fewer mutations in total to occur on the whole genealogy. The minimum number of mutations needed to provide our pattern of interest, C_6 , is two (one mutation on each of two branches). Therefore, for sufficiently small θ , we expect to find no more than two mutations on the entire tree. The probability $f(|d|; t, \theta)$ in Equation 3.2 will take one of three forms:

$$f_s(|0|; t, \theta) = e^{-t\theta/2} \tag{3.10}$$

or

$$f_s(|-1|; t, \theta) = f_s(|1|; t, \theta) = e^{-t\theta/2} \frac{t\theta}{4}. \tag{3.11}$$

This situation is analogous to a problem in phylogenetics. When rates of change are low, likelihood calculations on trees that consider all possible changes among allelic states converge to calculations of a parsimony score, as only changes of a single unit along a branch have nontrivial likelihood (*Felsenstein*, 2004). Similarly, our calculation of the probability that the private alleles lie on the edges of the size distribution, considering all possible states for allele sizes, is reduced in the small- θ case to a parsimony-style approximation by replacing $f(|d|; t, \theta)$ with $f_s(|0|; t, \theta)$ and $f_s(|1|; t, \theta)$. This parsimony approximation further eliminates the sums over n_3 , n_A , n_B , and n_C , making $\mathbb{P}[E_1 \cup E_2 | C_6, \theta]$ (Equation 3.1) tractable to analytically compute.

Examining all the ways of placing two mutations on one of the 18 topologies such that the $\{AB/AC\}$ configuration is produced, each placement will contribute some probability to either the denominator in Equation 3.1 or to both the denominator and numerator in Equation 3.1. As an example, consider history 8 from Figure 3.1. We first examine the four ways of getting configuration C_6 by assignment of the roles of A , B , and C to the alleles x_1 , x_2 , x_3 , and x_4 . We then find all placements of two

mutations on the tree that are consistent with this configuration. Each placement will either place the private alleles on both ends of the size distribution, or the shared allele will be on one end. If the private alleles are on both ends, then the term contributes to both the numerator and the denominator. If the shared allele is on an end, then the term contributes to the denominator only. Figure 3.5 illustrates this approach for the case of $x_1 = x_4$.

We can substitute f_s for f in our definitions of V^{cat} and V^{sym} to get the following small- θ versions of the probability of an arbitrary set of allele sizes.

$$\begin{aligned} V_s^{cat}(x_1, x_2, x_3, x_4, n_3, \Psi) &= f_s(|x_4|; 2t_2 + t_3 + t_4, \theta) \times \\ & f_s(|x_3|; t_3 + t_4, \theta) \times f_s(|n_3|; t_3, \theta) \times \\ & f_s(|n_3 - x_2|; t_4, \theta) \times f_s(|n_3 - x_1|; t_4, \theta) \end{aligned}$$

$$\begin{aligned} V_s^{sym}(x_1, x_2, x_3, x_4, n_3, \Psi) &= f_s(|n_3|; 2t_2 + t_3, \theta) \times \\ & f_s(|n_3 - x_4|; t_4, \theta) \times f_s(|n_3 - x_3|; t_4, \theta) \times \\ & f_s(|x_2|; t_3 + t_4, \theta) \times f_s(|x_1|; t_3 + t_4, \theta). \end{aligned}$$

Each possible placement of two mutations on one of the 18 labeled histories has a probability that falls into one of 12 equivalence classes as a result of symmetries in V_s^{cat} and V_s^{sym} . We denote these classes by σ_i ($i \in \{1, \dots, 12\}$), as defined in Table 3.4.

By tabulating in Table 3.5 the contributions from each class to the numerator and denominator of the probability for each of the 18 labeled histories, we can now

compute the numerator, $2\mathbb{P}[E_1, C_6|\theta]$, in Equation 3.1 as

$$\int_0^\infty \int_0^\infty \int_0^\infty \rho(t_2, t_3, t_4) \frac{1}{18} \left(32\sigma_1(t_2, t_3, t_4, \theta) + 32\sigma_2(t_2, t_3, t_4, \theta) + 16\sigma_3(t_2, t_3, t_4, \theta) + 16\sigma_6(t_2, t_3, t_4, \theta) + 32\sigma_8(t_2, t_3, t_4, \theta) + 8\sigma_9(t_2, t_3, t_4, \theta) + 8\sigma_{11}(t_2, t_3, t_4, \theta) \right) dt_2 dt_3 dt_4, \quad (3.12)$$

which evaluates to

$$\frac{\theta^2(648 + 990\theta + 489\theta^2 + 79\theta^3)}{18(1 + \theta)^2(2 + \theta)^3(3 + \theta)^3}. \quad (3.13)$$

The denominator, $\mathbb{P}[C_6|\theta]$, of Equation 3.1 is

$$\int_0^\infty \int_0^\infty \int_0^\infty \rho(t_2, t_3, t_4) \frac{1}{18} \left(32\sigma_1(t_2, t_3, t_4, \theta) + 32\sigma_2(t_2, t_3, t_4, \theta) + 16\sigma_3(t_2, t_3, t_4, \theta) + 16\sigma_4(t_2, t_3, t_4, \theta) + 16\sigma_5(t_2, t_3, t_4, \theta) + 16\sigma_6(t_2, t_3, t_4, \theta) + 32\sigma_7(t_2, t_3, t_4, \theta) + 32\sigma_8(t_2, t_3, t_4, \theta) + 8\sigma_9(t_2, t_3, t_4, \theta) + 16\sigma_{10}(t_2, t_3, t_4, \theta) + 8\sigma_{11}(t_2, t_3, t_4, \theta) + 16\sigma_{12}(t_2, t_3, t_4, \theta) \right) dt_2 dt_3 dt_4, \quad (3.14)$$

which evaluates to

$$\frac{\theta^2(432 + 630\theta + 295\theta^2 + 45\theta^3)}{6(1 + \theta)^2(2 + \theta)^3(3 + \theta)^3}. \quad (3.15)$$

Taking the ratio of expressions 3.13 and 3.15 and evaluating the limit as θ tends to 0 gives us

$$\lim_{\theta \rightarrow 0} \mathbb{P}[E_1 \cup E_2 | C_6, \theta] = \lim_{\theta \rightarrow 0} \frac{(648 + 990\theta + 489\theta^2 + 79\theta^3)}{3(432 + 630\theta + 295\theta^2 + 45\theta^3)} = \frac{1}{2}. \quad (3.16)$$

This result shows that, for low mutation rates, we expect the private alleles in an $\{AB/AC\}$ sample of size four to be on the ends of the size distribution approximately 1/2 of the time. This is substantially more often than the value of 1/3 predicted when the relatedness of the alleles was not taken into account.

3.3 Arbitrary divergence time

Extending our two-population model, we now consider two populations separated by arbitrary divergence time t_d (Figure 3.6). Note that as shown in Figure 3.6, the definitions of t_2 , t_3 , and t_4 differ slightly from those used in the calculations for the $t_d = 0$ case in Figure 3.2. We can formulate Equation 3.1 for arbitrary divergence time t_d and compute

$$\mathbb{P}[E_1 \cup E_2 | C_6, \theta, t_d] = \frac{2\mathbb{P}[E_1, C_6 | \theta, t_d]}{\mathbb{P}[C_6 | \theta, t_d]}. \quad (3.17)$$

Detailed derivations appear in Appendix A. We calculate Equation 3.17 numerically by Gaussian quadrature and by simulation using methods similar to those used for the $t_d = 0$ case (Appendix B).

Figure 3.7 shows, as a function of θ and t_d , the probability that the private alleles lie on the edges of the size distribution, as obtained using the simulation in Appendix B. We see that throughout the parameter space, the probability always exceeds the naïve expectation of 1/3. For all values of θ , we observe that increasing the divergence time between the populations increases the probability of finding the private alleles on the edges of the size distribution. Furthermore, we see that for small θ , the probability that private alleles in a sample of size four are found on the edges of the size distribution quickly tends toward 1 as t_d increases. By applying the small- θ approximation of Equations 3.10 and 3.11, we can show that this probability does indeed converge to 1 as t_d tends to infinity.

Conditioning on each of the four possible scenarios depicted in Figure 3.6, we follow an approach similar to the $t_d = 0$ small- θ derivation to obtain a small- θ approximation for the case of arbitrary divergence time (Appendix C). The resulting limiting expression for this approximation as θ tends to 0 is

$$\lim_{\theta \rightarrow 0} \mathbb{P}[E_1 \cup E_2 | C_6, \theta_{small}, t_d] = \frac{3e^{2t_d} - 2t_d - 2}{3e^{2t_d} - 1}. \quad (3.18)$$

Equation 3.18 is sensible in that it agrees with the small- θ result of $1/2$ at $t_d = 0$ (Equation 3.16), and it approaches the conditional result $\mathbb{P}[E_1 \cup E_2 | C_6, \theta_{small}, t_d, E_{11}] = 1$ as t_d increases without bound (Equation C.6). In Figure 3.8, we plot the function of t_d in Equation 3.18 along with simulated results at increasingly small θ . We see that for each θ , the probability that the private alleles lie on the edges of the size distribution increases monotonically as a function of the divergence time, and that the simulated probability approaches the limiting expression as θ approaches 0.

3.4 Properties of the probability that private alleles lie on the edges

In order to investigate the probability that private alleles lie on the edges of the size distribution, we started with a naïve argument that suggests that this should happen $1/3$ of the time in a sample of four alleles, two from each of two populations. However, this naïve argument ignored the relatedness of the four alleles. We have presented a calculation of the desired probability using a coalescent framework for gene relatedness, together with the symmetric stepwise mutation model. When fixing t_d , we see a monotonic decrease in the probability that the private alleles lie on the edges as θ grows, but for every collection of parameter values evaluated (scaled mutation rate θ and divergence time t_d between the two populations), the probability

remains greater than $1/3$.

Furthermore, the probability appears to stay well above $1/3$ even for very large θ . For large θ , we might expect so many mutations to occur on the tree that the allele sizes would not be correlated, effectively “erasing” the genealogical relatedness. In this case, we would expect the naïve prediction of $1/3$ to hold. However, in order to observe a C_6 configuration, two alleles must be identical by state. Thus, when conditioning on configuration C_6 , the distribution of branch lengths is biased toward shorter branches compared to the unconditional distribution, and even for large θ , the number of mutations tends to be small enough that genealogical relatedness remains important.

Holding t_d fixed at 0, Figure 3.9 plots $(\bar{t}_i^{C_6} - \bar{t}_i)/\bar{t}_i$ versus θ , where \bar{t}_i is the unconditional expectation of t_i under the coalescent and $\bar{t}_i^{C_6}$ is the conditional expectation given configuration C_6 , as obtained in 10^6 simulations that produced this configuration. We see that as θ increases, the relative difference between the conditional mean coalescence times given configuration C_6 and the unconditional mean coalescence times becomes increasingly negative. Most notably, t_4 becomes particularly short, reflecting the observation that for large θ , scenarios with configuration C_6 often have a “cherry” with short external branches of length t_4 on which no mutations occur.

In the small- θ case, we find that for $t_d = 0$, the probability that the private alleles lie on the edges in a sample of size four approaches $1/2$ as θ tends to zero. By letting the divergence time between the two populations exceed zero ($t_d > 0$), we see a monotonic increase in this probability. In fact, in the small- θ limit, the probability that the private alleles lie on the edges in a sample of size four tends to 1 as t_d tends to infinity.

These results show that the genealogical history of a set of microsatellite alleles is an important factor in determining the prevalence of private alleles in the ends

of the allele size distribution, even under circumstances in which we might expect the genealogy to be relatively unimportant. Our calculations also predict that the probability that private alleles lie on the edges of the allele size distribution grows as the divergence time between populations increases.

3.5 Application to data

To test the prediction that the probability that private alleles lie on the edges of the allele size distribution grows as the divergence time between populations grows, we analyzed data on microsatellites at 783 loci covering 1048 individuals in 53 worldwide populations from the Human Genome Diversity Panel (*Rosenberg et al.*, 2005). Computations with these microsatellites have established a general increase of genetic differentiation (and hence, divergence time) with increasing geographical distance between a pair of populations (*Ramachandran et al.*, 2005). Thus, although a strict divergence model is only an approximation to the population histories, we can consider the pairwise comparisons of populations that are geographically near each other to represent populations that diverged recently. Similarly, we can consider the pairwise comparisons of populations that are geographically distant from each other to represent populations that diverged relatively farther in the past. Pairwise comparisons of a population with itself can be interpreted as the case in which a population divergence happened at time $t_d = 0$ in the past. Based on the theory we have developed, we expect that pairs of geographically separated populations will produce a higher probability that the private alleles will lie on the edges of the size distribution. Similarly, we expect smaller probabilities for pairs of geographically proximate populations and the smallest probabilities for comparisons of populations with themselves. We further expect that measures of genetic differentiation such as F_{ST} will correlate with this probability as well, since these measures can be taken as a loose proxy for the divergence time between two populations.

To estimate the empirical frequency that the private alleles in a sample of size four lie on the edges of the size distribution, we perform the following analysis. For each population at each locus, we estimate the allele frequency distribution by counting the total number of observations of each distinct allele size and dividing by the total number of observations in the population. For a pair of populations, we then draw two alleles from the empirical allele frequency distribution in each population. If the set of four alleles has an $\{AB/AC\}$ configuration, we accept the draw and determine if the private alleles lie on the edges of the size distribution. If so, then the draw is counted as a success. We repeatedly draw sets of four alleles until 100,000 draws are accepted. Finally, we calculate the empirical frequency that the private alleles lie on the edges of the size distribution for a locus by dividing the number of successes by the number of acceptances, and we calculate the mean of this empirical frequency across loci. By performing this analysis, we get an estimate for the mean frequency that private alleles lie on the edges of the size distribution.

The results of this analysis are plotted in Figure 3.10, and we find that real populations do indeed follow the expected theoretical trend. The probabilities that private alleles lie on the edges range from 0.3759 to 0.4595. African populations paired with each other have lower probabilities, and a trend towards higher probabilities occurs as African populations are paired with other populations that are more geographically distant. The pairings of African populations with Native American populations (representing the most genetically distant pairs) have the highest probabilities. Furthermore, pairings close to the diagonal in Figure 3.10 tend to be more closely related than pairings farther away from the diagonal, and for these pairs, we see mostly low probabilities. Finally, the main diagonal represents the analysis of a population paired with itself; this is interpreted as comparing two populations with a divergence time of $t_d = 0$. We find that probabilities along the diagonal are the lowest among all pairs considered.

Because we also expect the frequency of private alleles on the edges to correlate with measures of genetic differentiation, we calculate pairwise F_{ST} between populations using Equation 5.3 from *Weir* (1996). In Figure 3.11, F_{ST} values are plotted against the frequency with which private alleles occur on the edges of the size distribution, and we find a very tight correlation ($r = 0.9333$). Thus, our empirical calculations show that our model for explaining the size distribution of private microsatellite alleles is able to predict phenomena observed in real data.

3.6 Discussion

We have modeled the phenomenon of private microsatellite alleles lying on the edges of the allele size distribution in order to explain an observation by *Wang et al.* (2007) that they occupy these locations more often than is expected by chance. Using a simple two-population model with sample size four, we have provided a naïve argument, in which we expect the probability that private microsatellite alleles lie on the edges of the size distribution to be $1/3$. Using a coalescent model with symmetric stepwise mutation to explicitly calculate this probability as a function of two parameters (mutation rate θ and divergence time t_d), we find that this probability appears to always exceed $1/3$. Furthermore, the model predicts that the probability that private alleles lie on the edges of the size distribution grows larger as the divergence time between populations increases. We have found that this prediction holds in an analysis of worldwide microsatellite data in humans.

Intuitively, we can understand why $\mathbb{P}[E_1 \cup E_2 | C_6, \theta, t_d]$ might be expected to exceed the naïve expectation by considering the process by which private alleles are generated. When an ancestral population splits into two groups, all allele sizes present in the population become shared alleles in the descendant populations, and these shared alleles define the center of the allele size distribution. As allele sizes diffuse away from the center in the separate descendant populations, mutations in either

population toward the edges of the size distribution are likely to generate alleles that are novel and therefore private. Conversely, mutations that push alleles toward the center of the size distribution are likely to produce sizes that already exist in both populations, as a result of the shared descent of central allele sizes. Furthermore, to produce shared alleles on an edge of the size distribution, unless the edge allele size is inherited by descent from the ancestral population in both descendant groups, alleles from each population must separately mutate to the same size on the edge. Because more mutations in total are required for producing such a shared allele on the edge compared to the number required in one population to produce a private allele on the edge, we expect private alleles to lie on the edges of the size distribution more often than is predicted under the assumption that there is no relationship between privacy and allele size.

This work augments the coalescent theory of microsatellite markers by providing predictions about the properties of private alleles in a simple model with sample size four. Previous work has examined additional quantities in the case of a four-allele sample. For example, *Kimmel and Chakraborty (1996)* and *Pritchard and Feldman (1996)* studied the expectation $E[(X_i - X_k)^2(X_j - X_\ell)^2]$ for random allele sizes X_i , X_j , X_k , and X_ℓ in a stepwise mutation model. *Zhang and Rosenberg (2007)* studied the genealogies of duplicated microsatellites in a model with four sampled alleles, two each for two paralogous microsatellite loci. Together with these other efforts, our work demonstrates that analytical formulas can sometimes be obtained in coalescent-based microsatellite models of non-trivial size.

While our main goal has been to explore the properties of our simple model, the model may potentially enable the inference of θ and t_d . For each of a collection of loci whose mutational characteristics are assumed to be identical, the probability that private alleles lie on the edges of the size distribution could be estimated from data by repeatedly sampling alleles from the observed allele frequency distributions for

pairs of populations. Using this empirical estimate, a likelihood surface could then be constructed to jointly estimate θ and t_d . This approach might not produce identifiable estimates; however, if θ has already been estimated by another method or if additional summary statistics are combined with a private allele statistic, a potentially viable method for estimating t_d might be constructed, considering the dramatic effect that this parameter has on the probability that private alleles lie on the edges of the size distribution.

We conclude with a discussion of model limitations. Because of the complexity of the probability calculations, we have restricted our attention to a sample of size four. We have assumed a simple demographic model of two populations, in which population sizes are equal and no migration occurs after the populations diverge. The simple stepwise mutation model assumes symmetry in the direction of mutation and independence of the mutation rate with allele size, and both the demographic model and the mutation model likely reflect conditions that are not strictly met in the human population example that we consider. Indeed, more complex mutation models, allowing for directional bias, multistep mutations, length-dependent mutation rates, or a combination of these factors could potentially be considered (e.g. *Calabrese and Durrett*, 2003; *Whittaker et al.*, 2003; *Watkins*, 2007). In general, however, we did not need a more complex model to explain the core observation that private alleles frequently lie on the edges of the size distribution. While the true demographic and mutational phenomena are undoubtedly more complicated than our model captures, we are still able to observe that as predicted, the probability that private microsatellite alleles lie on the edges of the size distribution in a sample of four alleles correlates with the genetic differentiation between pairs of populations.

3.7 Acknowledgements

Support for this work was provided by NIH grants R01 GM081441 and T32 HG000040, NSF grant DEB-0716904, and a grant from the Burroughs Wellcome Fund.

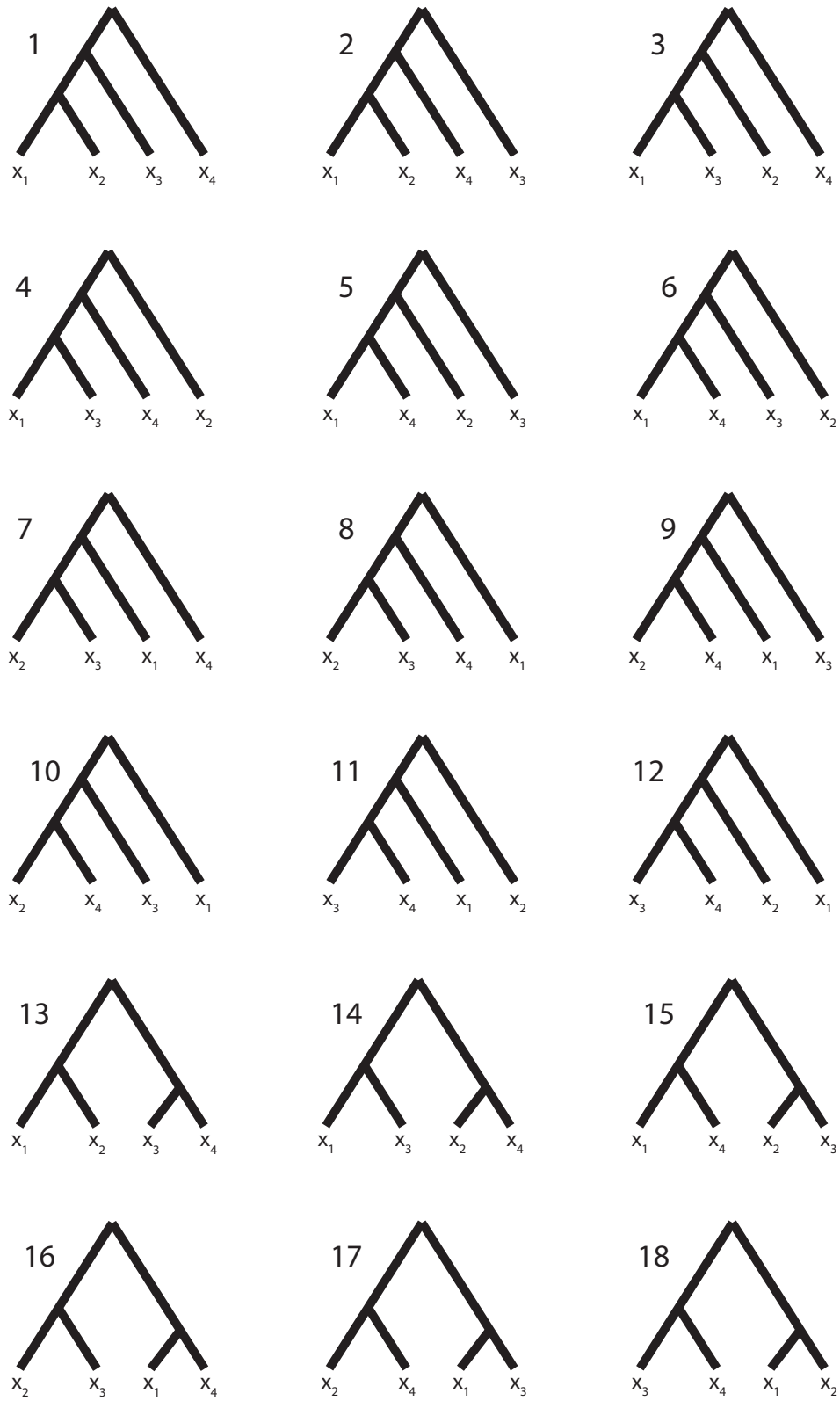


Figure 3.1: An enumeration of all possible labeled histories relating four sampled alleles. Twelve histories have asymmetric topologies (1-12), and six histories have symmetric topologies (13-18).

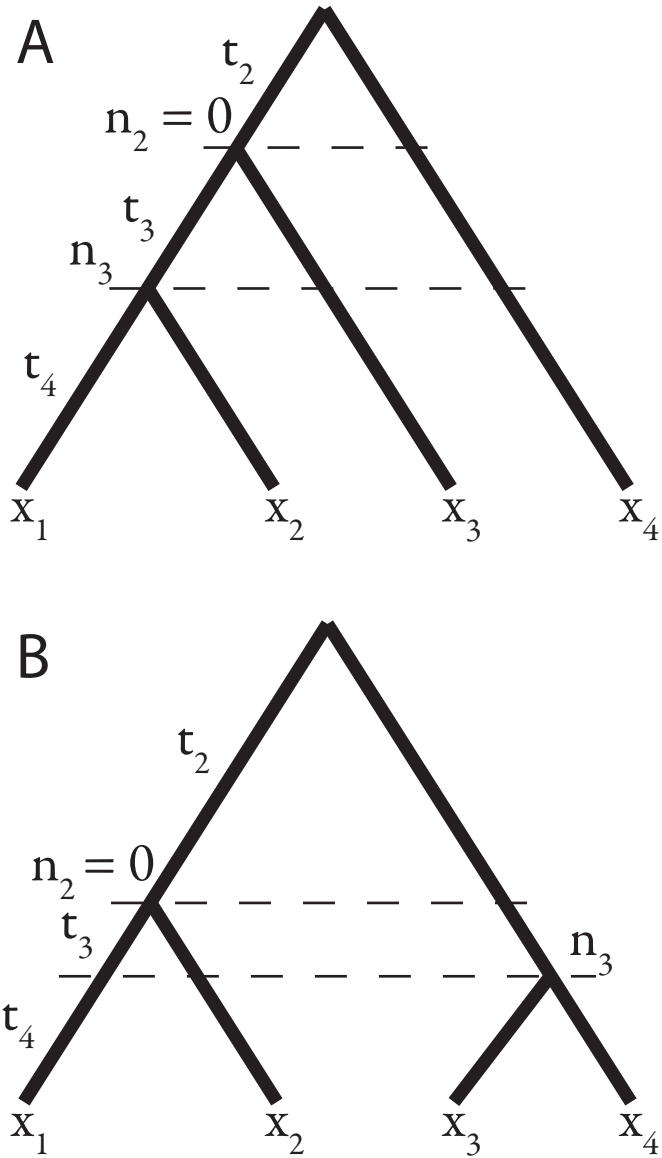


Figure 3.2: Example labelings of the two possible unlabeled topologies for gene genealogies with four lineages. (A) The asymmetric topology and (B) the symmetric topology are parameterized by allele sizes at the nodes, $(x_1, x_2, x_3, x_4, n_3, n_2)$, and by coalescence times (t_2, t_3, t_4) indicating the lengths of certain segments of the branches.

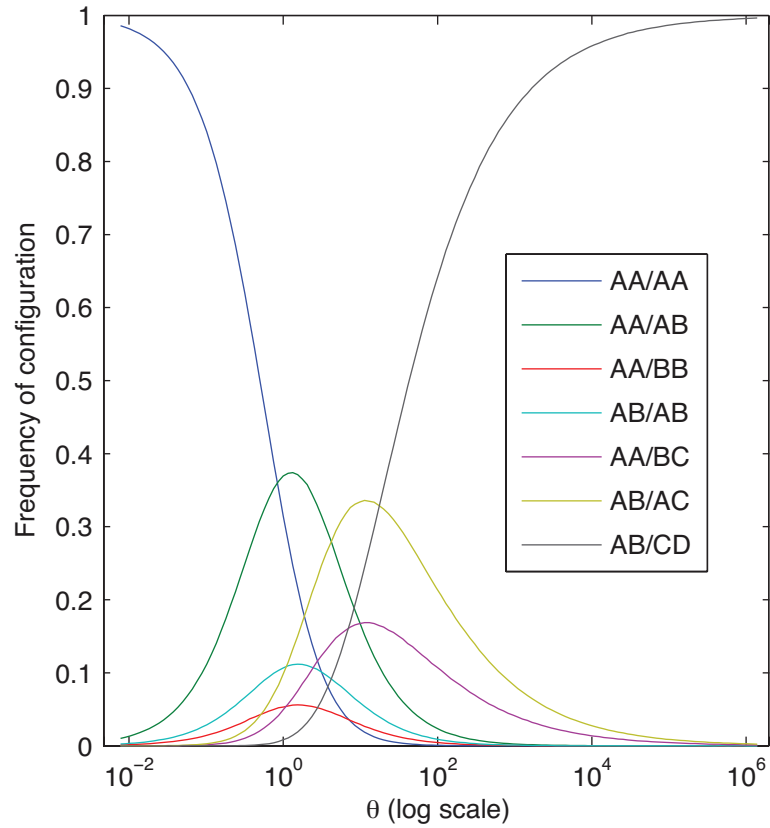


Figure 3.3: The simulated frequency of occurrence of seven possible allele configurations as a function of scaled mutation rate (θ) on a log scale. 10^6 trees are simulated per θ step. These simulations utilize four alleles, two in each of two populations. Alleles are related by the coalescent, and they mutate according to the symmetric stepwise mutation model.

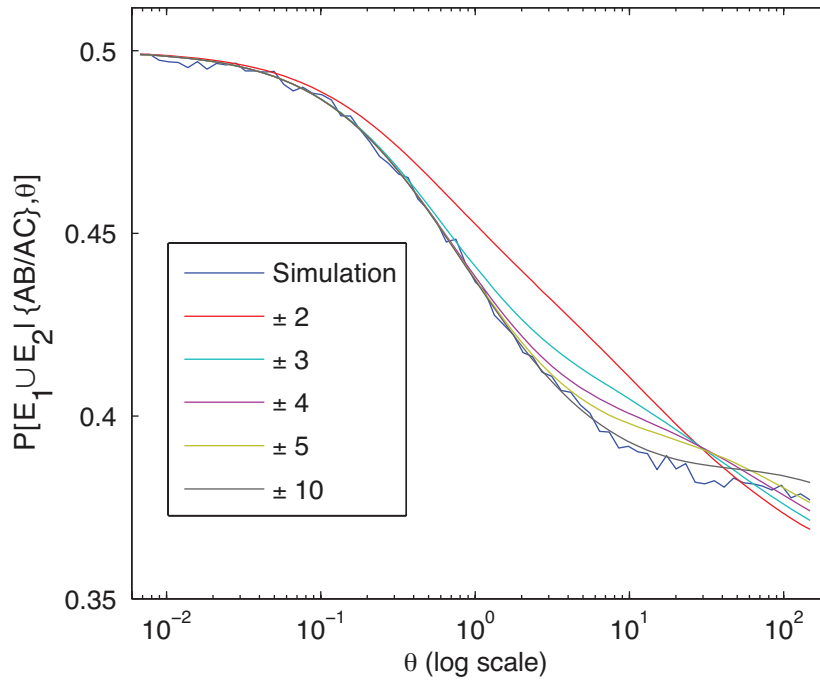


Figure 3.4: The probability that the private alleles lie on the edges of the size distribution conditional on production of an $\{AB/AC\}$ configuration, as a function of θ (log scale). This probability is plotted from simulations and for a range of truncations for the infinite sums in numerically approximating Equation 3.1. Simulation results are based on 100,000 $\{AB/AC\}$ trees simulated per θ step.

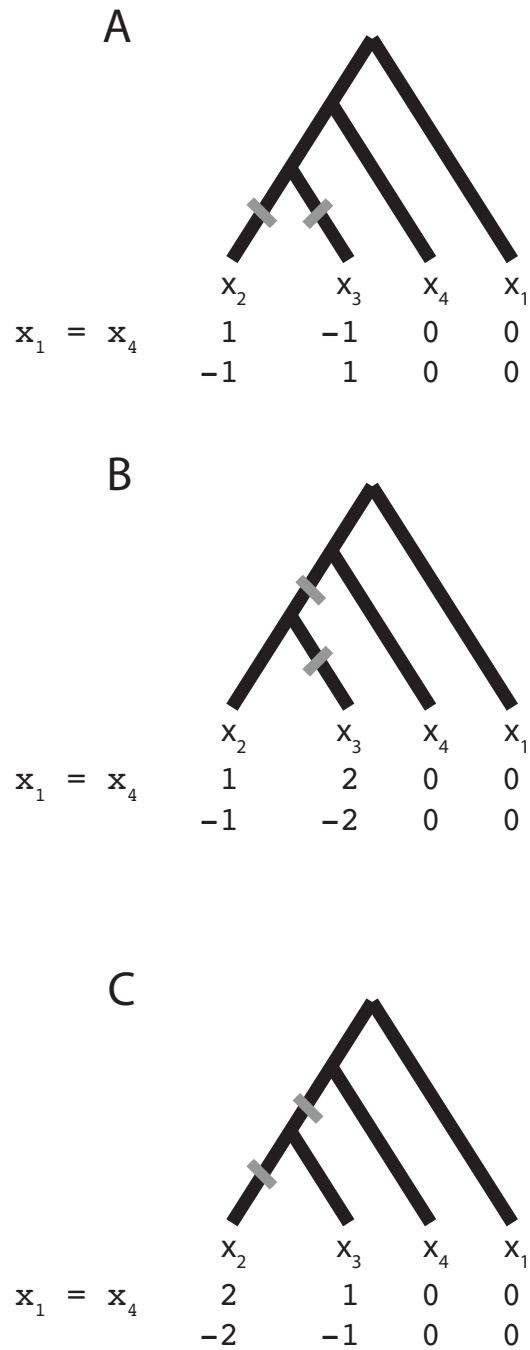


Figure 3.5: A small- θ “parsimony” approximation for calculating the probability that private alleles occur on the edges of the size distribution, for the case of $x_1 = x_4$ with history 8. Alleles x_1 and x_2 are in population 1, and alleles x_3 and x_4 are in population 2. In (A), the private alleles (± 1) lie on the edges of the size distribution ($-1 < 0 < 1$); however, in (B) and (C) the private alleles ($\pm 1, \pm 2$) are not on the two edges of the size distribution ($-2 < -1 < 0$ or $0 < 1 < 2$).

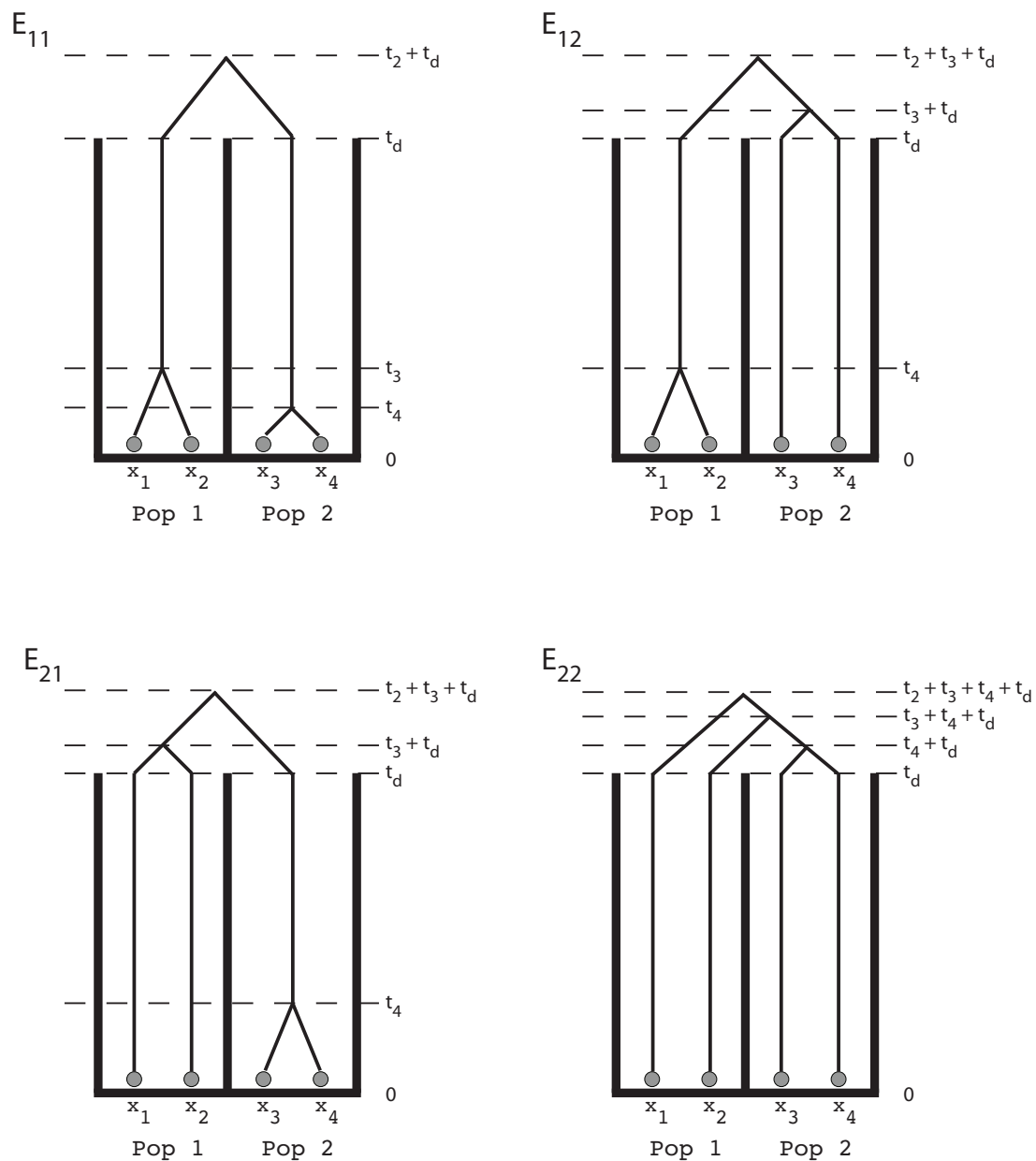


Figure 3.6: The four types of coalescent scenarios with $t_d > 0$, with their coalescent time parameterizations. In scenario E_{11} , t_3 is defined as the time to coalescence of the two lineages in population 1, and t_4 is defined as the time to coalescence of the two lineages in population 2.

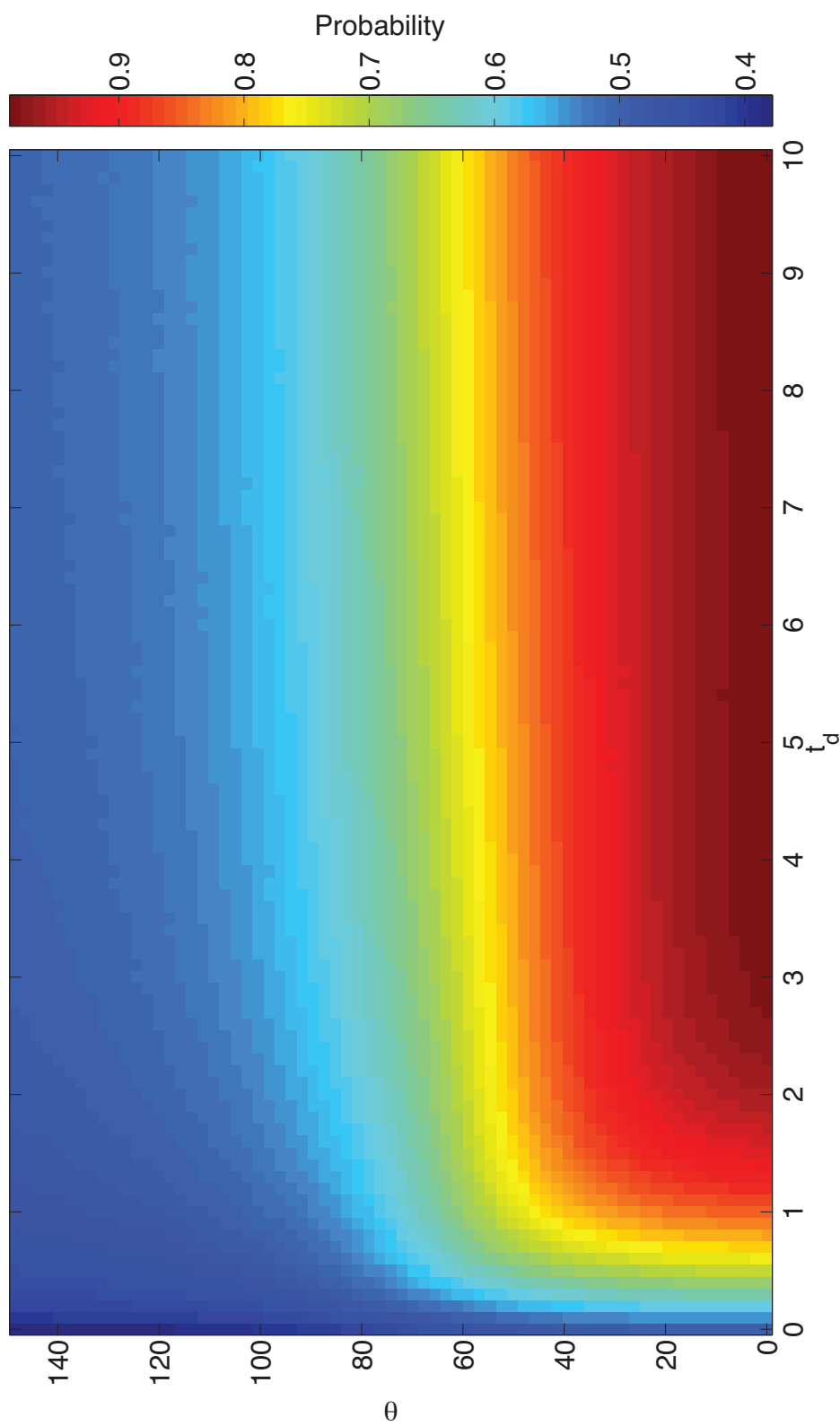


Figure 3.7: Simulated probability that the private alleles lie on the edges of the size distribution, conditional on production of an $\{AB/AC\}$ configuration. The plot shows this probability as a function of θ (log scale) and t_d . 10^6 $\{AB/AC\}$ trees are simulated for each choice of θ and t_d .

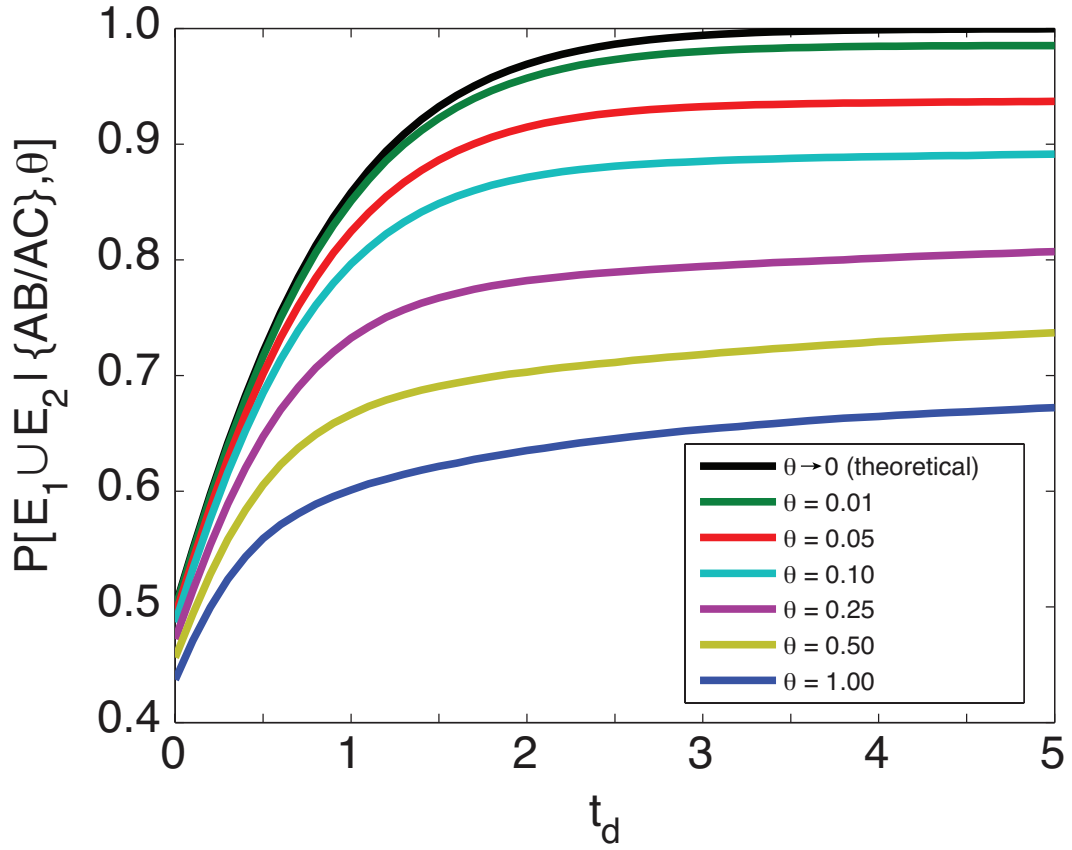


Figure 3.8: Simulated small- θ probabilities that the private alleles lie on the edges of the size distribution conditional on production of an $\{AB/AC\}$ configuration, and the corresponding limiting probability computed analytically for $\theta \rightarrow 0$ (Equation 3.18), as functions of t_d . The simulation approach follows that of Figure 3.7 and is described in Appendix B.

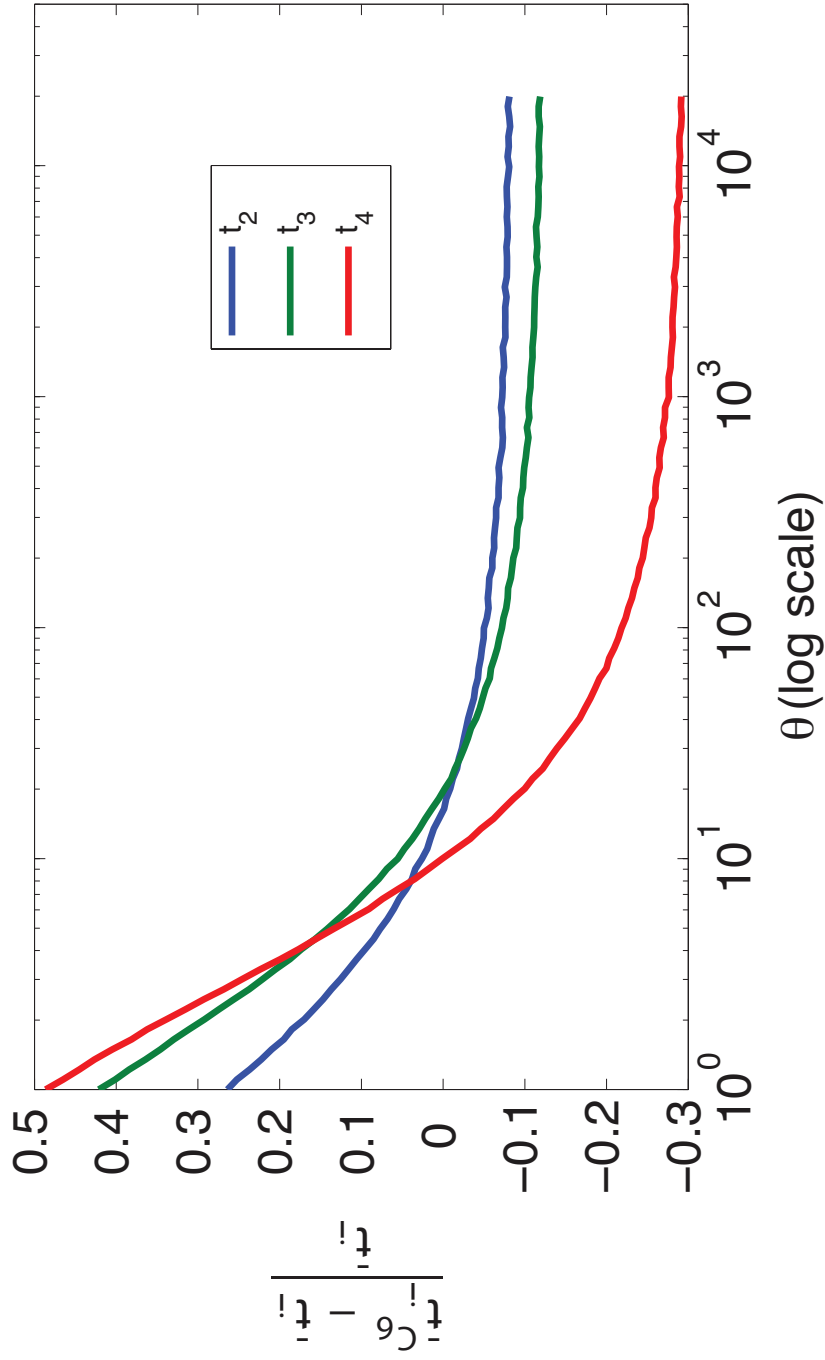


Figure 3.9: Relative difference between mean coalescence times conditional on obtaining configuration C_6 ($\bar{t}_i^{C_6}$) and unconditional mean coalescence times (\bar{t}_i), as a function of θ (log scale). The mean conditional coalescence times were calculated by taking the mean of 10^6 simulated coalescence times in scenarios that produced configuration C_6 .

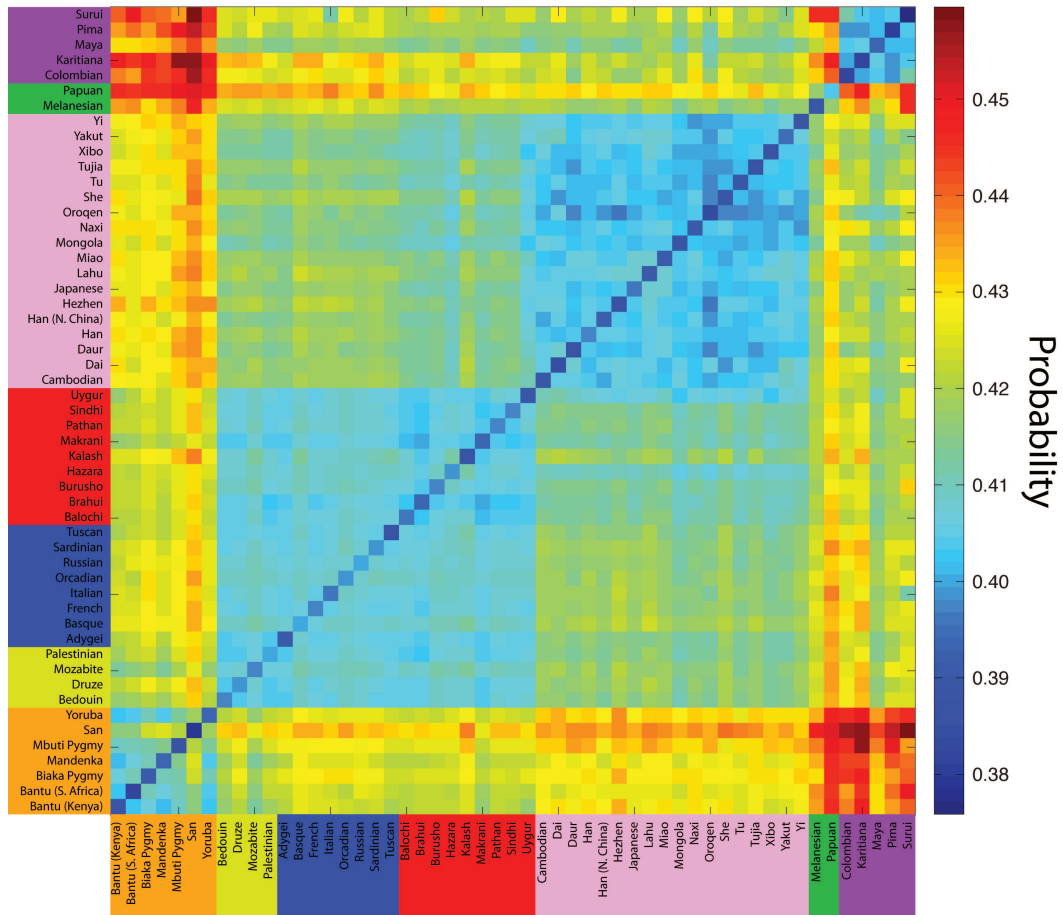


Figure 3.10: The empirical probability that private alleles lie on the edges of the size distribution in a sample of size four from a pair of populations. Plotted are pairwise calculations of this frequency for all 53 worldwide populations from the Human Genome Diversity Panel, arranged in major geographic regions. African, Middle Eastern, European, Central/South Asian, East Asian, Oceanian, and American populations are arranged by color in the labels. Blue represents a lower probability, and red represents a higher probability.

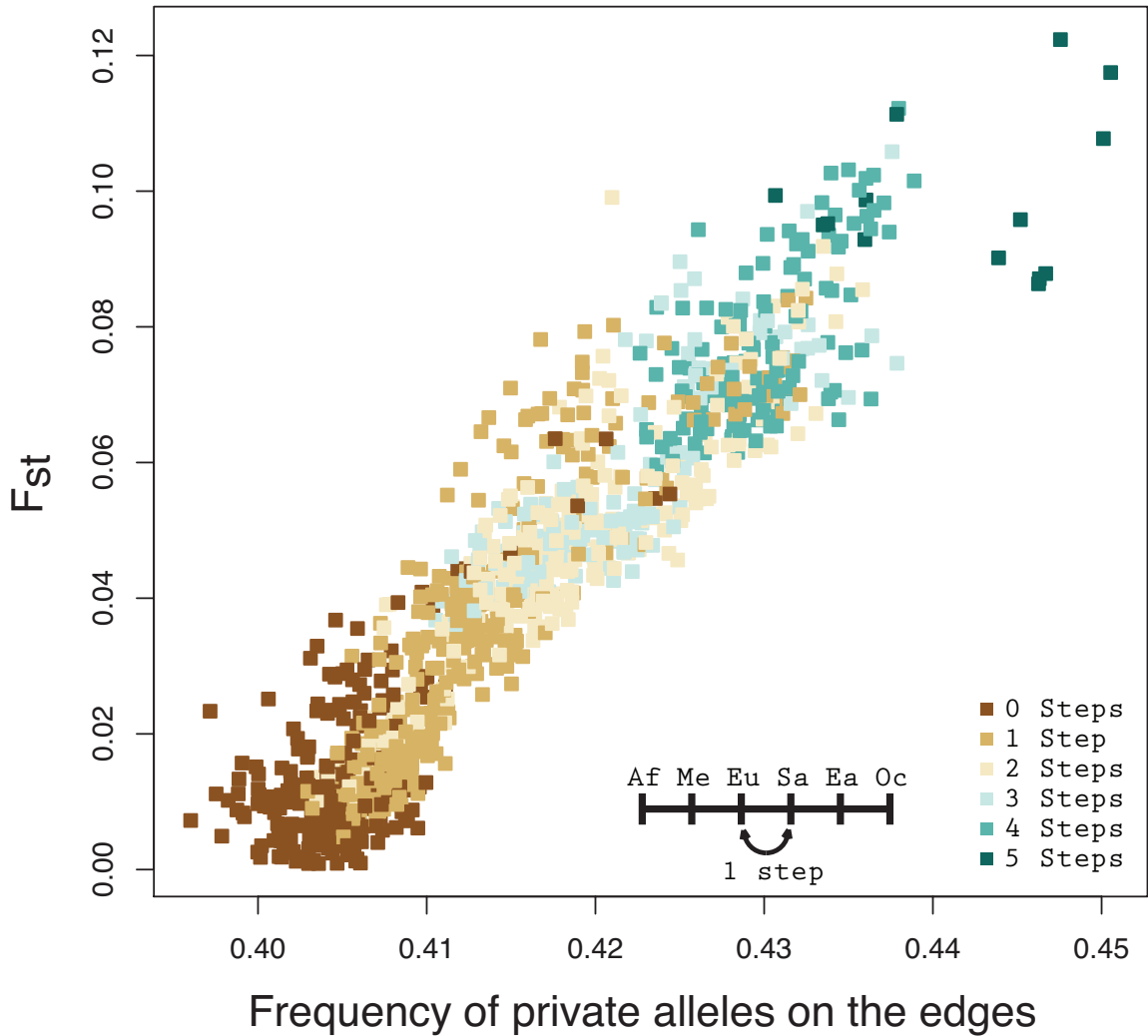


Figure 3.11: F_{ST} vs. the empirical frequency with which private alleles in a sample of size four lie on the edges of the size distribution. Each point represents a pair among 53 worldwide populations from the Human Genome Diversity Panel, excluding comparisons involving Native American populations and comparisons of populations with themselves. Pearson's $r = 0.9333$.










Table 3.1: The seven possible configurations of four alleles in two populations and the counts of shared, private, and total distinct alleles for each configuration.

Event	Configuration	Number of shared alleles	Number of private alleles	Total number of distinct alleles
C_1	$\{AA/AA\}$	1	0	1
C_2	$\{AA/AB\}$	1	1	2
C_3	$\{AA/BB\}$	0	2	2
C_4	$\{AB/AB\}$	2	0	2
C_5	$\{AA/BC\}$	0	3	3
C_6	$\{AB/AC\}$	1	2	3
C_7	$\{AB/CD\}$	0	4	4

Table 3.2: The four allele size relationships possible for the $\{AB/AC\}$ allele configuration.

Case	Allele size relationship				Allele roles in $\{AB/AC\}$			
	x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4
1	$x_1 = x_3$	$x_2 \neq x_4$	$x_1 \neq x_2$	$x_1 \neq x_4$	A	B	A	C
2	$x_1 = x_4$	$x_2 \neq x_3$	$x_1 \neq x_2$	$x_1 \neq x_3$	A	B	C	A
3	$x_2 = x_3$	$x_1 \neq x_4$	$x_2 \neq x_1$	$x_2 \neq x_4$	B	A	A	C
4	$x_2 = x_4$	$x_1 \neq x_3$	$x_2 \neq x_1$	$x_2 \neq x_3$	B	A	C	A

Table 3.3: The probability contributions for a given set of values $(n_A, n_B, n_C, n_3, \Psi)$ for each of the 18 labeled histories. These probabilities occur in the sums in Equations 3.6 and 3.7.

History number	History	Contribution
1		$V^{cat}(n_A, n_B, n_A, n_C, n_3, \Psi) + V^{cat}(n_A, n_B, n_C, n_A, n_3, \Psi) + V^{cat}(n_B, n_A, n_A, n_C, n_3, \Psi) + V^{cat}(n_B, n_A, n_C, n_A, n_3, \Psi)$
2		$V^{cat}(n_A, n_B, n_C, n_A, n_3, \Psi) + V^{cat}(n_A, n_B, n_A, n_C, n_3, \Psi) + V^{cat}(n_B, n_A, n_C, n_A, n_3, \Psi) + V^{cat}(n_B, n_A, n_C, n_A, n_3, \Psi)$
3		$V^{cat}(n_A, n_A, n_B, n_C, n_3, \Psi) + V^{cat}(n_A, n_C, n_B, n_A, n_3, \Psi) + V^{cat}(n_B, n_A, n_A, n_C, n_3, \Psi) + V^{cat}(n_B, n_C, n_A, n_A, n_3, \Psi)$
4		$V^{cat}(n_A, n_A, n_C, n_B, n_3, \Psi) + V^{cat}(n_A, n_C, n_A, n_B, n_3, \Psi) + V^{cat}(n_B, n_A, n_C, n_A, n_3, \Psi) + V^{cat}(n_B, n_C, n_A, n_A, n_3, \Psi)$
5		$V^{cat}(n_A, n_C, n_B, n_A, n_3, \Psi) + V^{cat}(n_A, n_A, n_B, n_C, n_3, \Psi) + V^{cat}(n_B, n_C, n_A, n_A, n_3, \Psi) + V^{cat}(n_B, n_A, n_A, n_C, n_3, \Psi)$
6		$V^{cat}(n_A, n_C, n_A, n_B, n_3, \Psi) + V^{cat}(n_A, n_A, n_C, n_B, n_3, \Psi) + V^{cat}(n_B, n_C, n_A, n_A, n_3, \Psi) + V^{cat}(n_B, n_A, n_A, n_C, n_3, \Psi)$
7		$V^{cat}(n_B, n_A, n_A, n_C, n_3, \Psi) + V^{cat}(n_B, n_C, n_A, n_A, n_3, \Psi) + V^{cat}(n_A, n_A, n_B, n_C, n_3, \Psi) + V^{cat}(n_A, n_C, n_B, n_A, n_3, \Psi)$
8		$V^{cat}(n_B, n_A, n_C, n_A, n_3, \Psi) + V^{cat}(n_B, n_C, n_A, n_A, n_3, \Psi) + V^{cat}(n_A, n_A, n_C, n_B, n_3, \Psi) + V^{cat}(n_A, n_C, n_B, n_A, n_3, \Psi)$
9		$V^{cat}(n_B, n_C, n_A, n_A, n_3, \Psi) + V^{cat}(n_B, n_A, n_A, n_C, n_3, \Psi) + V^{cat}(n_A, n_C, n_B, n_A, n_3, \Psi) + V^{cat}(n_A, n_A, n_B, n_C, n_3, \Psi)$




























History number	History	Contribution
10		$V^{cat}(n_B, n_C, n_A, n_A, n_3, \Psi) + V^{cat}(n_B, n_A, n_C, n_A, n_3, \Psi) + V^{cat}(n_A, n_C, n_A, n_B, n_3, \Psi) + V^{cat}(n_A, n_A, n_C, n_B, n_3, \Psi)$
11		$V^{cat}(n_A, n_C, n_A, n_B, n_3, \Psi) + V^{cat}(n_C, n_A, n_A, n_B, n_3, \Psi) + V^{cat}(n_A, n_C, n_A, n_B, n_3, \Psi) + V^{cat}(n_A, n_C, n_A, n_B, n_3, \Psi)$
12		$V^{cat}(n_A, n_C, n_B, n_A, n_3, \Psi) + V^{cat}(n_C, n_A, n_B, n_A, n_3, \Psi) + V^{cat}(n_A, n_C, n_A, n_B, n_3, \Psi) + V^{cat}(n_C, n_A, n_A, n_B, n_3, \Psi)$
13		$V^{sym}(n_A, n_B, n_A, n_C, n_3, \Psi) + V^{sym}(n_A, n_B, n_C, n_A, n_3, \Psi) + V^{sym}(n_B, n_A, n_A, n_C, n_3, \Psi) + V^{sym}(n_B, n_A, n_C, n_A, n_3, \Psi)$
14		$V^{sym}(n_A, n_A, n_B, n_C, n_3, \Psi) + V^{sym}(n_A, n_C, n_B, n_A, n_3, \Psi) + V^{sym}(n_B, n_A, n_A, n_C, n_3, \Psi) + V^{sym}(n_B, n_C, n_A, n_A, n_3, \Psi)$
15		$V^{sym}(n_A, n_C, n_B, n_A, n_3, \Psi) + V^{sym}(n_A, n_A, n_B, n_C, n_3, \Psi) + V^{sym}(n_B, n_C, n_A, n_A, n_3, \Psi) + V^{sym}(n_B, n_C, n_A, n_A, n_3, \Psi)$
16		$V^{sym}(n_B, n_A, n_A, n_C, n_3, \Psi) + V^{sym}(n_B, n_C, n_A, n_A, n_3, \Psi) + V^{sym}(n_A, n_A, n_B, n_C, n_3, \Psi) + V^{sym}(n_A, n_C, n_B, n_A, n_3, \Psi)$
17		$V^{sym}(n_B, n_C, n_A, n_A, n_3, \Psi) + V^{sym}(n_B, n_A, n_A, n_C, n_3, \Psi) + V^{sym}(n_A, n_C, n_B, n_A, n_3, \Psi) + V^{sym}(n_A, n_A, n_B, n_C, n_3, \Psi)$
18		$V^{sym}(n_A, n_C, n_A, n_B, n_3, \Psi) + V^{sym}(n_C, n_A, n_A, n_B, n_3, \Psi) + V^{sym}(n_A, n_C, n_B, n_A, n_3, \Psi) + V^{sym}(n_C, n_A, n_B, n_A, n_3, \Psi)$

Table 3.4: Definitions for the 12 classes of probability in the small- θ “parsimony” approximation.

Class	Defined probability
1	$\sigma_1(\Psi) = V_s^{cat}(1, 0, 0, 0, 1, \Psi)$
2	$\sigma_2(\Psi) = V_s^{cat}(1, 0, 0, 1, 0, \Psi)$
3	$\sigma_3(\Psi) = V_s^{cat}(0, 0, 0, 1, 1, \Psi)$
4	$\sigma_4(\Psi) = V_s^{cat}(0, 0, 1, 1, 0, \Psi)$
5	$\sigma_5(\Psi) = V_s^{cat}(0, 0, 1, 0, 1, \Psi)$
6	$\sigma_6(\Psi) = V_s^{cat}(1, 1, 0, 0, 0, \Psi)$
7	$\sigma_7(\Psi) = V_s^{cat}(1, 0, 1, 0, 0, \Psi)$
8	$\sigma_8(\Psi) = V_s^{sym}(1, 0, 0, 1, 0, \Psi)$
9	$\sigma_9(\Psi) = V_s^{sym}(0, 0, 0, 1, 1, \Psi)$
10	$\sigma_{10}(\Psi) = V_s^{sym}(0, 0, 1, 1, 0, \Psi)$
11	$\sigma_{11}(\Psi) = V_s^{sym}(1, 1, 0, 0, 0, \Psi)$
12	$\sigma_{12}(\Psi) = V_s^{sym}(1, 0, 1, 0, 0, \Psi)$

Table 3.5: The small- θ approximation contributions to the numerator (Equation 3.12) and denominator (Equation 3.14) of the probability that the private alleles lie on the edges, for each labeled history.

History number	History	Contribution to numerator	Contribution to denominator
1		$4\sigma_1 + 4\sigma_2$	$4\sigma_1 + 4\sigma_2$
2		$4\sigma_1 + 4\sigma_2$	$4\sigma_1 + 4\sigma_2$
3		$2\sigma_1 + 2\sigma_2 + 2\sigma_3 + 2\sigma_6$	$2\sigma_1 + 2\sigma_2 + 2\sigma_3 + 2\sigma_4 + 2\sigma_5 + 2\sigma_6 + 4\sigma_7$
4		$2\sigma_1 + 2\sigma_2 + 2\sigma_3 + 2\sigma_6$	$2\sigma_1 + 2\sigma_2 + 2\sigma_3 + 2\sigma_4 + 2\sigma_5 + 2\sigma_6 + 4\sigma_7$
5		$2\sigma_1 + 2\sigma_2 + 2\sigma_3 + 2\sigma_6$	$2\sigma_1 + 2\sigma_2 + 2\sigma_3 + 2\sigma_4 + 2\sigma_5 + 2\sigma_6 + 4\sigma_7$
6		$2\sigma_1 + 2\sigma_2 + 2\sigma_3 + 2\sigma_6$	$2\sigma_1 + 2\sigma_2 + 2\sigma_3 + 2\sigma_4 + 2\sigma_5 + 2\sigma_6 + 4\sigma_7$
7		$2\sigma_1 + 2\sigma_2 + 2\sigma_3 + 2\sigma_6$	$2\sigma_1 + 2\sigma_2 + 2\sigma_3 + 2\sigma_4 + 2\sigma_5 + 2\sigma_6 + 4\sigma_7$
8		$2\sigma_1 + 2\sigma_2 + 2\sigma_3 + 2\sigma_6$	$2\sigma_1 + 2\sigma_2 + 2\sigma_3 + 2\sigma_4 + 2\sigma_5 + 2\sigma_6 + 4\sigma_7$
9		$2\sigma_1 + 2\sigma_2 + 2\sigma_3 + 2\sigma_6$	$2\sigma_1 + 2\sigma_2 + 2\sigma_3 + 2\sigma_4 + 2\sigma_5 + 2\sigma_6 + 4\sigma_7$

History number	History	Contribution to numerator	Contribution to denominator
10		$2\sigma_1 + 2\sigma_2 + 2\sigma_3 + 2\sigma_6$	$2\sigma_1 + 2\sigma_2 + 2\sigma_3 + 2\sigma_4 + 2\sigma_5 + 2\sigma_6 + 4\sigma_7$
11		$4\sigma_1 + 4\sigma_2$	$4\sigma_1 + 4\sigma_2$
12		$4\sigma_1 + 4\sigma_2$	$4\sigma_1 + 4\sigma_2$
13		$8\sigma_8$	$8\sigma_8$
14		$4\sigma_8 + 2\sigma_9 + 2\sigma_{11}$	$4\sigma_8 + 2\sigma_9 + 4\sigma_{10} + 2\sigma_{11} + 4\sigma_{12}$
15		$4\sigma_8 + 2\sigma_9 + 2\sigma_{11}$	$4\sigma_8 + 2\sigma_9 + 4\sigma_{10} + 2\sigma_{11} + 4\sigma_{12}$
16		$4\sigma_8 + 2\sigma_9 + 2\sigma_{11}$	$4\sigma_8 + 2\sigma_9 + 4\sigma_{10} + 2\sigma_{11} + 4\sigma_{12}$
17		$4\sigma_8 + 2\sigma_9 + 2\sigma_{11}$	$4\sigma_8 + 2\sigma_9 + 4\sigma_{10} + 2\sigma_{11} + 4\sigma_{12}$
18		$8\sigma_8$	$8\sigma_8$

CHAPTER IV

Long runs of homozygosity are enriched for deleterious variation

4.1 Introduction

The study of deleterious variation in the genome has been of fundamental importance to evolutionary genetics (*Muller, 1950; Morton et al., 1956; Kondrashov, 1995; Charlesworth and Charlesworth, 1998; Eyre-Walker and Keightley, 1999; Fay et al., 2001; Sunyaev et al., 2001; Lohmueller et al., 2008; Chun and Fay, 2011; Cooper and Shendure, 2011; Lohmueller et al., 2011; Necşulea et al., 2011; Lesecque et al., 2012; Tennessen et al., 2012*). An individual human genome may contain from tens to hundreds of variants that would be lethal in homozygous form (*Morton et al., 1956; Kondrashov, 1995*) and may also contain hundreds to thousands of mildly deleterious variants (*Fay et al., 2001; Sunyaev et al., 2001; Lohmueller et al., 2008; Tennessen et al., 2012*). *Crow (1997)* suggested that the accumulation of many deleterious variants could have long-term health consequences, while *Charlesworth and Charlesworth (1998)* have even hypothesized that the evolution of the Y chromosome could be the result of the accumulation of deleterious variation. In order to gain insights into human adaptation, evolution, and genetic disease, It is therefore of critical importance to understand the distribution of this variation within and

between individuals and populations, as well as to understand the effect that natural selection and other genomic variables have on shaping patterns of diversity.

Early empirical studies have used data on a limited number of genes to make inferences about the distribution, accumulation, and effects of deleterious variation. *Eyre-Walker and Keightley* (1999) used GenBank sequences of human, chimpanzee, and the closest available primate species to estimate the deleterious mutation rate in humans. *Fay et al.* (2001) used single nucleotide polymorphism and divergence data from ~ 180 genes to estimate that 80% of amino acid mutations are deleterious and that there are approximately 300 deleterious variants per diploid genome.

With the widespread availability of next-generation sequencing technology, exome sequencing allows for the targeted study of all known protein-coding regions. As mutations within protein-coding regions are particularly likely to be disruptive—by altering the encoded amino acid sequence—relative to non-coding regions, exome sequencing offers natural data sets with which to study the genomic distribution of potentially deleterious variation. Furthermore, when an encoded amino acid is changed, the functional impact can be predicted reasonably accurately (*Cooper and Shendure*, 2011). *Lohmueller et al.* (2008) studied the exomes of 20 European Americans and 15 African Americans and found an excess of homozygous deleterious variation in the European American samples. Similarly, *Tennesen et al.* (2012) showed through deep exome sequencing that a large fraction of coding variation is recent, rare, and deleterious.

Examining the patterns of deleterious variation with respect to various genomic variables is of particular interest, allowing the further characterization of fundamental processes that shape genetic variation across the genome. Using whole-genome sequences, *Lohmueller et al.* (2011) investigated how natural selection on non-neutral variants affects patterns of neutral variation, finding that the correlation between neutral diversity and recombination rate is the result of negative selection acting on

large numbers of weakly deleterious variants.

Runs of homozygosity (ROH) are another interesting genomic variable to study with respect to the distribution of deleterious alleles. ROH are long stretches of consecutive homozygous genotypes that can occur in the genome as the result of several processes: population demographics, cultural practices, and natural selection. Bottlenecks or isolation can lead to a small effective population size and ultimately reduce diversity across the whole population, creating an increased pairing of common haplotypes, thereby generating ROH. Endogamy and consanguineous marriage can also generate ROH by increasing the likelihood of pairing identical-by-descent stretches of chromosome. Finally, natural selection can reduce genetic diversity around a non-neutral part of the genome, leading to a higher likelihood of pairing for identical-by-state haplotypes and thus generating ROH.

ROH regions have been the central focus of homozygosity mapping studies of inbred individuals (*Lander and Botstein, 1987; Broman and Weber, 1999*). By searching for long stretches of identical-by-descent regions, many efforts to identify recessive Mendelian disease genes have been successful (*Botstein and Risch, 2003*). These efforts, however, have focused primarily on affected inbred individuals, although efforts are being made to extend these methods to non-inbred individuals (*Gibbs and Singleton, 2006; Hildebrandt et al., 2009; Collin et al., 2011; Hagiwara et al., 2011; Schuurs-Hoeijmakers et al., 2011*). Because of this transition toward the use of homozygosity mapping methods in outbred populations, there has been interest in characterizing the worldwide patterns of ROH to better understand how these regions are distributed within individuals and across populations (*Kirin et al., 2010; Leutenegger et al., 2011; Pemberton et al., 2012*). In light of the importance of ROH regions for disease-gene studies and the multitude of processes that create these regions, we aim to provide a characterization of the patterns of deleterious variation occurring inside and outside ROH regions.

We form two possible alternative hypotheses for how deleterious variants might correlate with ROH. First, consider that many deleterious variants act recessively and therefore only reduce an individual's fitness when occurring as homozygotes. A genome with an ROH region containing many deleterious variants would carry these variants as homozygotes, and especially if the variants interact synergistically, the individual would then have reduced fitness. As a result of this reduction in fitness, this genome would be less likely to be viable as compared to a genome with fewer ROH regions carrying fewer deleterious homozygotes. Thus, we form hypothesis 1: if we take the genome-wide frequency of neutral homozygotes across ROH and non-ROH regions as a baseline, we should expect to observe a smaller fraction of all genome-wide deleterious homozygotes to occur in ROH regions as compared to the fraction of neutral homozygotes occurring in ROH regions. While deleterious homozygotes occurring outside of ROH regions will also incur a fitness cost, under this hypothesis we expect that selection would more effectively purge long homozygous regions with many deleterious genotypes.

On the other hand, low-frequency variants are more likely to be deleterious than are common variants (*Marth et al.*, 2011), and inbred regions of the genome can present low-frequency variants in homozygous form at a higher rate than non-inbred regions. For example, consider a variant that has allele frequency p . If this variant is in a non-inbred region, then it would occur in homozygous form with probability p^2 . If instead it occurred in an inbred region, because of the identity-by-descent of an individual's two haplotypes, it would be homozygous with the greater probability p . Because of this higher probability, when the homozygous deleterious variants are not lethal and inbreeding is recent, we might then expect that selection has not had enough time to eliminate deleterious variants in these regions. Therefore we form hypothesis 2: taking the genome-wide frequency of neutral homozygotes across ROH and non-ROH regions as a baseline, we should expect to observe a *larger* fraction of

all genome-wide deleterious homozygotes to occur in ROH regions as compared to the fraction of neutral homozygotes occurring in ROH regions. We further expect that longer (newer) ROH might have a higher fraction of deleterious homozygotes than shorter (older) ROH.

To test these alternative hypotheses, we perform whole-exome sequencing and computational prediction of deleterious alleles and analyze these data in conjunction with the worldwide atlas of genomic patterns of ROH from *Pemberton et al.* (2012). We select 27 individuals from *Pemberton et al.* (2012) across 6 populations, including San ($n = 2$), Biaka Pygmy ($n = 5$), Mozabite ($n = 3$), Maya ($n = 5$), Surui ($n = 6$), Karitiana ($n = 6$). We choose these individuals to represent the extreme ends of the distribution of ROH across the genome (4%-46%, Figure 4.1, *Pemberton et al.* (2012)). To predict whether a variant allele is deleterious, we use the PolyPhen2 program (*Adzhubei et al.*, 2010). As a surrogate for neutral variation, we consider both synonymous sites and missense sites predicted to be benign. Next, using the coordinates of the called ROH regions from *Pemberton et al.* (2012), we assess the number of predicted deleterious variants in each individual's runs of homozygosity. Finally, we determine whether deleterious homozygotes occur within ROH more frequently than expected from the pattern of occurrence of neutral homozygotes, and whether this pattern differs for different classes of ROH that are believed to result from different population-genetic processes.

4.2 Methods

4.2.1 Raw read processing and variant calling

Figure 4.2 parts A and B give an overview of the workflow for processing the raw sequencing reads and variant calls. We sequenced the 27 individuals using Nimblegen SeqCap EZ v1 (Roche Nimblegen, Madison, WI) exome capture followed

by sequencing using the Illumina HiSeq2000 system. We aligned raw reads to the HG18 reference sequence with BWA (*Li and Durbin, 2009*). So that variants would be called with unbiased read counts, duplicate reads were marked with Picard (*Picard, 2011*). We used the Genome Analysis Tool Kit (GATK) (*McKenna et al., 2010*) for lane-level local realignment around known and possible indels and for lane-level recalibration of base quality scores. Finally, we called variants using all samples jointly with the UnifiedGenotyper module of GATK with a minimum phred-scaled confidence score of 30, which gives us a set of raw variant sites. In these analyses we considered only single nucleotide variant sites and excluded any insertion-deletions and multi-allelic sites.

4.2.2 Site-level quality control

The raw set of variant sites is expected to be very sensitive to true positive variant sites but also to contain many false positives as a result of machine artifacts. We further filtered the initial set of variant calls to reduce false positives (outlined in Figure 4.2 part C). First, we removed variant sites that fell outside the targeted regions. To separate the true positives from the false positives, we utilized the variant quality score recalibrator module of GATK (*DePristo et al., 2011*) to build an adaptive error model using known variant sites and their various quality score annotations (i.e. RMS Mapping Quality, Fisher’s exact test for strand bias, etc.). This utilized variant site quality measure from the joint variant calling step above and allowed us to estimate the probability that a variant is a true genetic variant versus a machine artifact.

The variant quality score recalibrator requires a set of likely true variant sites to train its error model. Here we consider two sets of likely true variant sites, jointly: called exome variant sites that have been previously identified as HapMap 3.3 variant sites are taken to be true with a phred-scaled prior of 15 (96.84%) and

called exome variant sites that have previously identified as Omni 2.5M variant sites are taken to be true with a phred-scaled prior of 12 (93.69%), as recommended by *DePristo et al.* (2011). Given this set of likely true variant sites, we trained the error model with the HaplotypeScore, HRun, MQRankSum, MQ, and FS quality score annotations. Other quality score annotations were not informative for distinguishing true versus false positives. HaplotypeScore measures the consistency of the site with exactly two segregating haplotypes, and higher scores are indicative of bad alignments. HRun gives the largest contiguous homopolymer run of the variant allele in either direction on the reference sequence. MQRankSum is the phred-scaled p-value from a Mann-Whitney rank sum test for mapping qualities. MQ is the root mean square of the mapping quality of the reads across all samples. FS is the phred-scaled p-value from a Fisher’s exact test to detect strand bias. These choices were also informed by *DePristo et al.* (2011).

After training the error model, all called variant sites in the original data site were annotated with the variant quality score log-of-odds (VQSLOD), which represents the log odds of a site being a true variant versus a false positive. Rather than choosing an arbitrary VQSLOD cut-off, we considered the distribution of VQSLOD scores for called variant sites also found in HapMap 3.3 and chose a cut-off that returns 99% of these sites, as recommended by *DePristo et al.* (2011). After filtering sites below this cutoff, 96,797 remained. Using dbSNP build 132 (minus sites added after build 129) 53,285 were known ($T_i/T_v = 3.1017$), and 43,511 were novel ($T_i/T_v = 2.8356$).

4.2.3 Variant classification by predicted functional impact

Some of our variant sites may not be in the coding regions of the targeted genes because the NimbleGen platform pads the capture target by 100 bps on each side, so we annotate the genomic location of each called variable site using the MapSNPs algorithm provided with the PolyPhen2 (*Adzhubei et al.*, 2010) program. MapSNPs

determined the genomic location of each site with respect to the Consensus CDS (CCDS) set of high-quality coding regions (*Pruitt et al.*, 2009) and successfully annotated 91,770 sites (Figure 4.3). There were 701 sites that mapped to two CCDS regions. Any site that had a mutation classified as missense and something else (i.e. missense in one CCDS and synonymous in another CCDS) was considered only as a missense mutation for downstream analyses. Furthermore, if a site had a mutation classified as synonymous in one CCDS and nonsense or UTR mutation in another CCDS, it was removed. If a site had a mutation classified as missense in more than one CCDS, it was retained for further classification by PolyPhen2 (*Adzhubei et al.*, 2010). If a missense mutation was classified by PolyPhen2 with respect to more than one CCDS, it was retained if the classifications were identical, and it was removed if the classifications differed. After reconciling these double hits, we were left with 26776 missense sites and 29914 synonymous sites.

Since we are interested in potentially deleterious variants in these data, we utilized the PolyPhen2 program to classify the non-reference alleles that are missense changes. Given a set of missense mutations, PolyPhen2 predicts the potential disruption that the non-reference allele has on the encoded protein by incorporating knowledge of amino acid biochemistry, folded structure (if known), and conservation score. Using a naïve Bayes classifier, PolyPhen2 puts missense mutations into “probably damaging,” “possibly damaging,” and “benign” categories based on the estimated false discovery rate. The final set of missense mutations successfully classified by PolyPhen2 is given in Figure 4.4, and since we are interested in both deleterious and non-deleterious variation, the final dataset of coding variation used in downstream analyses consists of these PolyPhen2-classified missense sites and the synonymous sites.

4.2.4 Genotype-level quality control

While the site-level quality control above generates a set of sites that are very likely to be truly variant in our sample, there may be specific genotypes at individuals that are of poor quality. Therefore, we performed a final round of quality control per individual genotype on the remaining 29914 synonymous and 26776 missense sites. We assessed concordance with known genotypes by comparing the called genotypes with those found on the Illumina HumanHap 650Y chip. All of the 27 individuals have been genotyped with this chip by *Li et al.* (2008). There are 6,180 variant sites that overlapped between the called variant sites and the genotyped SNP sites. There is an overall genotype concordance (percentage of called exome genotypes that agree with the Illumina genotypes) of 99.2%. This can be broken down by genotype, where we find 99.3% concordance for called non-reference homozygotes and 98.7% concordance for called heterozygotes. A breakdown of genotype concordance per individual is given in Table 4.1. Considering these concordance levels, we choose a filter for homozygous genotypes of $DP < 3$, where DP is the read depth for the sample at that site. Applying this filter gives a new concordance rate of 99.6% for non-reference homozygous genotypes, while removing 42.6% of mismatches and only 1.1% of matches. We now seek to filter heterozygous genotypes to achieve a similar rate of concordance. In order to do this, we consider the distribution of called heterozygotes as a function of both DP and non-reference allele frequency (Figure 4.5). Here we choose, by hand, a progressive filter based on non-reference allele frequency as a function of DP. The cut-off is more permissive at lower DP and more restrictive at higher DP (denoted by red dotted line in Figure 4.5). Applying this filter gives a new concordance rate of 99.6% for heterozygous genotypes, while removing 71.7% of mismatches and 0.5% of matches. After filtering, 64 former variant sites did not have variant calls for any individual, and 1288 were monomorphic in the sample. These were removed from the dataset. After genotype filtering, the data set consists

of 54,359 sites (Figure 4.6). This data set has, for each individual, mean coverage between 38x and 81x, and the percentage of sites with $\geq 20x$ coverage ranges between 62% and 90% across individuals (Table 4.2).

4.2.5 Final data set for downstream analysis

The final data set for further downstream analysis consists of 54,359 sites (Figure 4.6) across the 27 individuals. The mean per-individual missing data rate is 3.3% with a maximum individual missing data rate of 10.6%. In some downstream analyses, we combine the synonymous sites with the benign sites into a super-class denoted “non-damaging” and combine the possibly damaging and probably damaging sites into a super-class denoted “damaging” for the sake of simplicity. While there may be truly damaging variants in the non-damaging class and vice versa, we create these super-classes to represent sets of sites enriched for damaging and non-damaging variants. In aggregate, these super-classes will be useful to observe genome-wide trends in deleterious variation. Note that while MapSNPs and PolyPhen2 classify individual mutations, we refer to a *site* as “synonymous,” “probably damaging,” “possibly damaging,” or “benign,” if the alternate allele at that site has been classified the same. It is important to note, then, that reference alleles at “damaging” loci are not expected to be damaging.

4.2.6 ROH data analysis

Pemberton et al. (2012) have characterized worldwide patterns of runs of homozygosity in 1839 human individuals across 64 populations using an autozygosity based LOD score method. They further classified these ROH into three broad categories: short ROH (denoted class A ROH) tens of kb in size, medium ROH (denoted class B ROH) hundreds of kb to a few Mb in size, and long ROH (denoted class C ROH) several Mb in size. For the 27 individuals in our exome sequencing data

set, we took the coordinates defining the ROH regions as well as the ROH size class boundary values so that we could identify a given ROH segment as of a particular size class. Using this information, we calculate

$$R_i = \frac{\text{total length of all ROH regions in individual } i}{\text{total length of the genome}}. \quad (4.1)$$

This represents the total fraction of the genome of individual i covered by an ROH region.

With this information, we map each variant site from Figure 4.6 to a specific ROH segment for each individual. An individual at a given site will be either homozygous for the reference allele (0/0), heterozygous (0/1), or homozygous for the alternate allele (1/1). Since it is the alternate allele that has been classified as damaging or non-damaging, it is useful to make counts of specific genotypes, and we calculate $g_i^{n,k}$, $g_i^{d,k}$, $g_{i,j}^{n,k}$ and $g_{i,j}^{d,k}$. $g_i^{n,k}$ and $g_i^{d,k}$ represent the number of genotypes with $k \in \{0, 1, 2\}$ alternate alleles in individual i at non-damaging and damaging sites, respectively. $g_{i,j}^{n,k}$ and $g_{i,j}^{d,k}$ represent the number of genotypes with $k \in \{0, 1, 2\}$ alternate alleles in individual i falling in ROH class $j \in \{A, B, C\}$ at non-damaging and damaging sites, respectively. This means that the number of genotypes in individual i with $k \in \{0, 1, 2\}$ alternate alleles that fall into a non-ROH region is given by

$$g_i^{n,k} - \sum_{j \in \{A, B, C\}} g_{i,j}^{n,k} \quad (4.2)$$

for non-damaging sites and

$$g_i^{d,k} - \sum_{j \in \{A, B, C\}} g_{i,j}^{d,k} \quad (4.3)$$

for damaging sites. These genotype counts are summarized for damaging sites in Table 4.3 and for non-damaging sites in Table 4.4.

4.3 Results

4.3.1 Heterozygous genotypes and ROH size classes

The short, medium, and long ROH size classes that *Pemberton et al.* (2012) identified correspond to ROH lengths previously described. Short ROH (denoted class A ROH) are tens of kb in size and thought to reflect the homozygosity of old haplotypes (*International HapMap Consortium*, 2007). Medium ROH (denoted class B ROH) are hundreds of kb to a few Mb in size and thought to arise from background relatedness (*Lencz et al.*, 2007; *Curtis et al.*, 2008; *Jakkula et al.*, 2008; *McQuillan et al.*, 2008). Finally, long ROH (denoted class C ROH) are several Mb in size and likely result from recent parental relatedness (*Broman and Weber*, 1999; *Li et al.*, 2006; *Gibson et al.*, 2006; *International HapMap Consortium*, 2007; *Lencz et al.*, 2007; *Curtis et al.*, 2008; *Jakkula et al.*, 2008; *McQuillan et al.*, 2008; *Kirin et al.*, 2010). If this is truly the case with the *Pemberton et al.* (2012) ROH regions, then we should see a substantially lower fraction of all an individual's genotypes that are heterozygotes occurring in ROH regions. In particular, we should expect the lowest fraction of all an individual's genotypes that are heterozygotes in their long class C ROH with a slightly higher fraction in medium class B and short class A ROH. We should further see an enrichment of heterozygotes relative to the genome-wide fraction in non-ROH regions. We calculate the genome-wide fraction of genotypes that are heterozygotes in individual i as

$$\frac{g_i^{d,1} + g_i^{n,1}}{\sum_{k=0}^2 (g_i^{d,k} + g_i^{n,k})}, \quad (4.4)$$

and the results are in Table 4.5 column 3. Similarly, we calculate:

$$\frac{g_i^{d,1} - \sum_{j \in \{A,B,C\}} g_{i,j}^{d,1} + g_i^{n,1} - \sum_{j \in \{A,B,C\}} g_{i,j}^{n,1}}{\sum_{k=0}^2 (g_i^{d,k} + g_i^{n,k})}, \quad (4.5)$$

representing the fraction of all genome-wide genotypes that are heterozygotes in individual i that do not occur in an ROH region (Table 4.5 column 4);

$$\frac{\sum_{j \in \{A,B,C\}} g_{i,j}^{d,1} + \sum_{j \in \{A,B,C\}} g_{i,j}^{n,1}}{\sum_{k=0}^2 (g_i^{d,k} + g_i^{n,k})} \quad (4.6)$$

representing the fraction of all genome-wide genotypes that are heterozygotes in individual i that occur in any ROH region (Table 4.5 column 5);

$$\frac{g_{i,A}^{d,1} + g_{i,A}^{n,1}}{\sum_{k=0}^2 (g_i^{d,k} + g_i^{n,k})} \quad (4.7)$$

representing the fraction of all genome-wide genotypes that are heterozygotes in individual i that occur in class A ROH region (Table 4.5 column 6);

$$\frac{g_{i,B}^{d,1} + g_{i,B}^{n,1}}{\sum_{k=0}^2 (g_i^{d,k} + g_i^{n,k})} \quad (4.8)$$

representing the fraction of all genome-wide genotypes that are heterozygotes in individual i that occur in class B ROH region (Table 4.5 column 7); and

$$\frac{g_{i,C}^{d,1} + g_{i,C}^{n,1}}{\sum_{k=0}^2 (g_i^{d,k} + g_i^{n,k})} \quad (4.9)$$

representing the fraction of all genome-wide genotypes that are heterozygotes in individual i that occur in class C ROH region (Table 4.5 column 8).

We observe, as expected, that the percentage of heterozygotes in any ROH region is substantially lower than in the genome-wide and non-ROH regions. We also note that as we move from small to long ROH the percentage of heterozygotes in these regions drops off dramatically. This is consistent with the view that short class A ROH are older (and therefore have accumulated more mutations) and long class C ROH are younger (and therefore have not accumulated many mutations). However it

is also possible that some of this pattern is the result of small ROH being called with less confidence and long ROH being called with greater confidence, so there may be more false positives in the smaller ROH calls. It is important to note that the set of sites that we consider in these analyses are only sites that are polymorphic in our sample of 27 individuals, and so the percentage of all genotypes that are heterozygotes does not include homozygous reference allele genotypes at sites that are fixed in our sample.

4.3.2 The number of deleterious homozygous genotypes occurring in ROH

Table 4.3 tabulates the counts for each reference homozygote (0/0), heterozygote (0/1), and non-reference homozygote (1/1) at damaging sites that fall into ROH regions and non-ROH regions (all $g_{i,j}^{n,k}$ and $g_{i,j}^{d,k}$). These results underscore the substantial mutational burden many individuals are carrying, particularly the individuals with a very high genomic ROH content. For instance individual 837 of the Surui has the highest ROH coverage (46.4% of the genome) and carries a total of 357 predicted damaging (189 probably damaging and 168 possibly damaging) variants in homozygous form. This contrasts strongly with individual 459 of the Biaka who has the lowest ROH coverage (4.0% of the genome) and has 212 predicted damaging (109 probably damaging and 103 possibly damaging) variants in homozygous form.

Next, we compare the distribution of variants of different predicted function. Figure 4.7 shows the total number of damaging non-reference homozygotes (1/1) per individual as a function of total fraction of the genome covered in ROH (R_i). The colored points represent damaging homozygotes that occur within ROH ($\sum_{j \in \{A,B,C\}} g_{i,j}^{d,2}$) while the black points represent the damaging homozygotes that occur outside of ROH ($g_i^{d,2} - \sum_{j \in \{A,B,C\}} g_{i,j}^{d,2}$). As the genome is covered in more and longer ROH (high values of R_i), we naturally expect a greater raw number

of homozygotes (damaging or not) to fall within ROH, and we see a strong linear correlation between the number of damaging homozygotes and genomic ROH fraction (Pearson $\rho = 0.9897$, slope = 584.3, intercept = -9.2). Similarly, the raw number of homozygotes occurring outside ROH should decrease with genomic ROH fraction, as there are simply fewer ROH-free regions (Pearson $\rho = -0.8378$, slope = -139.0 , intercept = 181.3). We note in Figure 4.7 that the decreasing slope of the homozygotes in non-ROH regions (slope = -139.0) is shallower than the increasing slope of the homozygotes in ROH regions (slope = 584.3), indicating that the rise in damaging homozygotes in ROH regions outpaces the decline of damaging homozygotes in non-ROH regions. The fitted lines predict that an average non-inbred individual carries approximately 181 damaging variants in homozygous form. With every 10% increase in genomic ROH coverage, non-ROH regions decrease by 10%, and there is a loss of 18 damaging homozygotes. However, the increase in ROH regions increases the number of damaging homozygotes by 58, for a net gain of 40.

4.3.3 Damaging versus non-damaging variation in any size ROH

Next, we turn to testing the two alternative hypotheses proposed above. Recall that our hypothetical expectations are centered around comparing the numbers of damaging homozygotes inside and outside of ROH regions to the numbers of non-damaging homozygotes inside and outside of ROH regions. In particular under hypothesis 1, damaging homozygotes occur more often in non-ROH regions relative to the proportion of genome-wide non-damaging homozygotes occurring in these regions. Under hypothesis 2, damaging homozygotes will occur more often in ROH regions relative to the proportion of genome-wide non-damaging homozygotes occurring in ROH regions. Additionally, hypothesis 2 posits that there should be a differential effect based on the size class of ROH, with long ROH having the largest number of damaging homozygotes.

To test these hypotheses, we compute

$$f_i^n = \frac{\sum_{j \in \{A, B, C\}} g_{i,j}^{n,2}}{g_i^{n,2}}, \quad (4.10)$$

where f_i^n is the genome-wide fraction of non-damaging 1/1 homozygotes in individual i that fall in either class A, B or C ROH. These numbers represent the neutral baseline with which we compare the distribution of damaging homozygotes. Similarly, we compute

$$f_i^d = \frac{\sum_{j \in \{A, B, C\}} g_{i,j}^{d,2}}{g_i^{d,2}}, \quad (4.11)$$

where f_i^d is the genome-wide fraction of damaging 1/1 homozygotes in individual i that fall in either class A, B or C ROH. If hypothesis 1 is correct, then we should expect to see f_i^n systematically higher than f_i^d .

Figure 4.8A plots f_i^d (solid symbols) and f_i^n (open symbols) versus total genomic ROH coverage (R_i). We find that both the genome-wide fraction of non-damaging (presumed neutral) and the genome-wide fraction of damaging (presumed deleterious) homozygous genotypes are positively correlated with total genomic ROH coverage (non-damaging Pearson $\rho = 0.9983$ and damaging Pearson $\rho = 0.9938$). A linear correlation is expected in general, given that we expect a larger fraction of homozygous genotypes to occur in ROH as ROH comprise increasingly more of the genome. However, we further observe that the fraction of genome-wide damaging homozygotes f_i^d consistently exceeds the fraction of genome-wide non-damaging homozygotes f_i^n .

To assess the statistical significance of the two linear regressions for the damaging and non-damaging genotypes, we create the following linear model:

$$f_i^{\{n,d\}} = \beta_0 + \beta_1 R_i + \beta_2 D_i + \beta_3 R_i D_i + \epsilon, \quad (4.12)$$

where $f_i^{\{n,d\}}$ is a vector across individuals containing the fraction of genome-wide

damaging homozygotes in ROH regions and the fraction of genome-wide non-damaging homozygotes in ROH regions. R_i is the fraction of the genome covered by any size ROH for individual i , and D_i is an indicator variable for individual i , taking a value of 1 if the observed response is of damaging homozygotes and taking a value of 0 if the observed response is of non-damaging homozygotes. In this framework, a statistically significant β_2 indicates a difference in the intercepts of the two regressions, and a statistically significant β_3 indicates a difference in the slopes of the two regressions. We find $\beta_2 = 0.5340$ ($p = 8.59 \times 10^{-6}$) and $\beta_3 = 0.0965$ ($p = 0.00839$), indicating a significantly different intercept and slope between the two regressions plotted in Figure 4.8A. These results show that damaging homozygotes are biased toward occurring in ROH relative to the baseline of non-damaging homozygotes. These results are not compatible with hypothesis 1 but are consistent with hypothesis 2. If hypothesis 2 is correct, we should further be able to observe an excess of damaging homozygotes specifically in class C ROH regions versus non-damaging homozygotes in class C ROH regions, and furthermore an excess of damaging homozygotes in class C ROH versus damaging homozygotes in class A ROH.

4.3.4 Damaging versus non-damaging variation in small, medium, or long ROH

To further refine our support for hypothesis 2, we separately consider the data for each ROH size. For homozygous genotypes falling in either small class A, medium class B, or long class C ROH, we calculate

$$f_{i,j}^d = \frac{g_{i,j}^{d,2}}{g_i^{d,2}} \quad (4.13)$$

and

$$f_{i,j}^n = \frac{g_{i,j}^{n,2}}{g_i^{n,2}} \quad (4.14)$$

for damaging and non-damaging 1/1 homozygotes, respectively, where $j \in A, B, C$ denotes the ROH class. Figure 4.8B plots $f_{i,A}^d$ (solid symbols) and $f_{i,A}^n$ (open symbols) versus total genomic coverage by class A ROH. We find that both the genome-wide fraction of non-damaging (presumed neutral) and the genome-wide fraction of damaging (presumed deleterious) homozygous genotypes are positively correlated with total genomic coverage by class A ROH (non-damaging Pearson $\rho = 0.9829$ and damaging Pearson $\rho = 0.9365$). However, when we repeat the significance test of the differences in the two regressions, we find no significant difference in either the intercept ($p = 0.303$) or slope ($p = 0.647$). Figure 4.8C plots $f_{i,B}^d$ (solid symbols) and $f_{i,B}^n$ (open symbols) versus total genomic coverage by class B ROH and the two linear regressions for non-damaging (Pearson $\rho = 0.9892$) and damaging (Pearson $\rho = 0.9629$) homozygotes. Again, we find no significant difference in either the intercept ($p = 0.131$) or slope ($p = 0.142$). Figure 4.8D plots $f_{i,C}^d$ (solid symbols) and $f_{i,C}^n$ (open symbols) versus total genomic coverage by class C ROH and the two linear regressions for non-damaging (Pearson $\rho = 0.9921$) and damaging (Pearson $\rho = 0.9727$) homozygotes. However, we now find a significant difference in both the intercept ($p = 0.0368$) and slope ($p = 0.0186$). These results are consistent with hypothesis 2, suggesting that inbreeding is driving the differences in the proportion of damaging homozygotes in ROH regions versus non-damaging homozygotes in ROH regions.

Finally, under hypothesis 2, we should expect damaging homozygotes to occur more frequently in class C ROH versus class A ROH. We compare the fraction of damaging homozygotes falling in class C ROH ($f_{i,C}^d$) to the fraction of damaging homozygotes falling in class A ROH ($f_{i,A}^d$, Figure 4.9). While we find that the intercepts of the regressions are not significantly different ($p = 0.7278$), the slope

differences are significantly different ($p = 0.0139$). This suggests that the increase in fraction of damaging homozygotes is higher per unit increase in ROH coverage for class C ROH versus class A ROH. Indeed, we can see in Figure 4.9 that the high ROH coverage individuals have a substantially higher fraction of genome-wide damaging homozygotes occurring in class C versus class A, consistent with hypothesis 2.

Since class C ROH are thought to be the result of recent inbreeding and since inbred regions of the genome will present low-frequency variants as homozygotes at a higher rate than will non-inbred regions, we might expect the large divergence in slopes to be caused by a proportional excess of damaging variants in the American populations versus the African populations. Indeed, in a study of African American and European American individuals *Lohmueller et al.* (2008) have observed that the proportion of variants private to African Americans that are non-synonymous (47.0%) was significantly lower than the proportion of variants private to European Americans that are non-synonymous (55.4%). We calculate for each of our populations the proportion of all private alleles that are of a particular functional class by computing

$$F_{p,c} = \frac{N_{p,c}}{N_p}, \quad (4.15)$$

where $F_{p,c}$ is the fraction of private alleles in population p that have predicted functional class $c \in \{\text{synonymous, benign, probably damaging, possibly damaging}\}$, $N_{p,c}$ is the number of private alleles in population p with predicted functional class c , and N_p is the total number of private alleles in population p . In agreement with *Lohmueller et al.* (2008), we find that (Figure 4.10) the proportion of private variants that are non-synonymous is lowest in African populations (< 50% non-synonymous, $\sim 20\%$ predicted damaging) and highest in the American populations ($\sim 60\%$ non-synonymous, $> 30\%$ predicted damaging).

4.4 Discussion

Through sequencing-based variant discovery efforts it has been widely recognized that each human carries numerous deleterious variants (*Lohmueller et al.*, 2008; *MacArthur et al.*, 2012; *Tennessen et al.*, 2012). Our data set extends this observation by directly showing that many individuals carry at least 212 and upwards to 357 of such damaging variants in homozygous form. The fact that the combined presence of this many homozygous variants is compatible with life supports the view that most observed deleterious variants must have relatively small fitness effects.

Our analysis of deleterious variation with respect to ROH was framed by two alternative hypotheses. Under hypothesis 1, we might have expected the fraction of genome-wide damaging homozygotes occurring in ROH to be less than the fraction of genome-wide non-damaging homozygotes because of more effective selection against deleterious variants. In this case, the result would have been driven by the expectation that selection would purge these homozygote-rich regions of damaging genotypes. On the other hand, under hypothesis 2, we might have expected inbreeding to present an excess of low-frequency (and likely damaging) variants in homozygous form, with selection not having had sufficient time to eliminate them. In this case, ROH would contain a higher fraction of all genome-wide damaging homozygotes with respect to the fraction of all genome-wide non-damaging homozygotes. In particular, class C ROH would potentially drive this difference, as they are expected to be both recent and the result of recent inbreeding.

As we see in Figure 4.8A, the genome-wide fraction of damaging homozygotes is significantly higher in ROH regions than would be predicted for non-damaging homozygotes based on the total fraction of ROH in the genome. This result disagrees with hypothesis 1 and lends support to hypothesis 2. We are able to further refine support for hypothesis 2 by examining the genome-wide fraction of damaging homozygotes in ROH regions separately for each size class of ROH (Figures

4.8B-4.8D), finding that only for class C ROH is there a significant difference between the fraction of damaging homozygotes and the fraction non-damaging homozygotes. Finally, we are able to show that the fraction of genome-wide damaging homozygotes occurring in class C ROH is greater than the fraction of genome-wide damaging homozygotes occurring in class A ROH (Figure 4.9). This suggests that inbreeding is largely responsible for the accumulation of damaging homozygous genotypes in individual genomes.

These results are consistent with the notion that the human genome contains a spectrum of variants with a rich gradation of functional impact. As a consequence of negative selection, truly lethal variants are rare and most of the observed variants are either neutral or mildly deleterious, even though many of these could be predicted to impact molecular function. However, our results suggest that inbreeding amplifies deleterious homozygotes in genomes, with potentially important health consequences. Indeed, inbreeding has long been known to be deleterious to the health of offspring (*Darwin*, 1876; *Garrod*, 1902; *Morton et al.*, 1956; *Bittles and Neel*, 1994; *Bittles*, 2001; *Jorde*, 2001). If a variant in a population is homozygous lethal, inbreeding will greatly increase the chance of generating a genome with this lethal genotype. When the deleterious variants have less dramatic effects, however, they have the opportunity to accumulate in the genome as homozygotes, where they can be brought to high frequency.

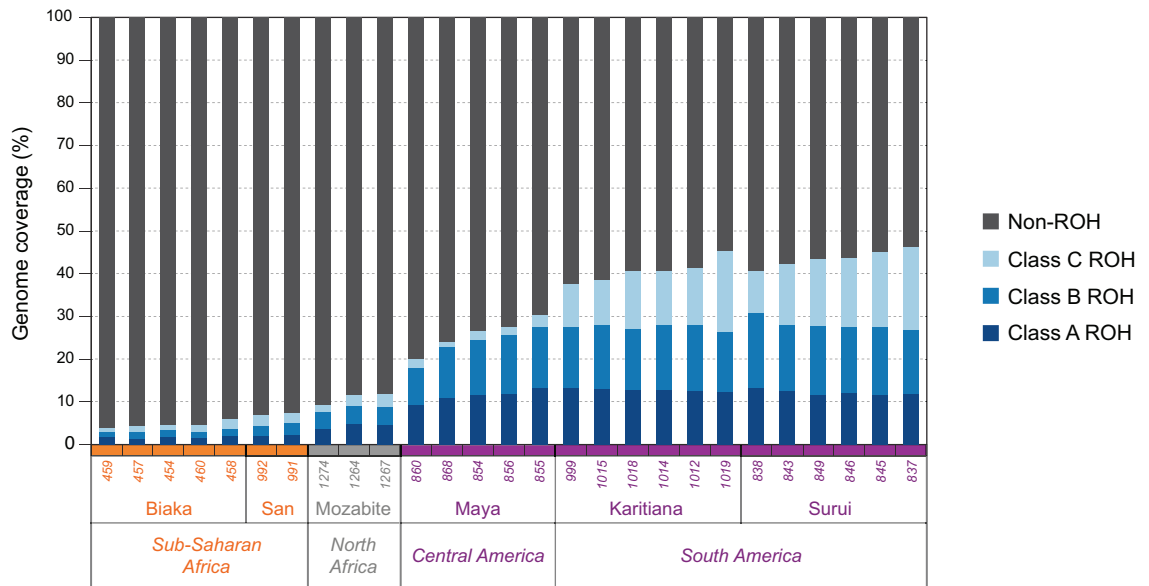


Figure 4.1: The total percentage of individual genomes covered by class A, class B, and class C ROH.

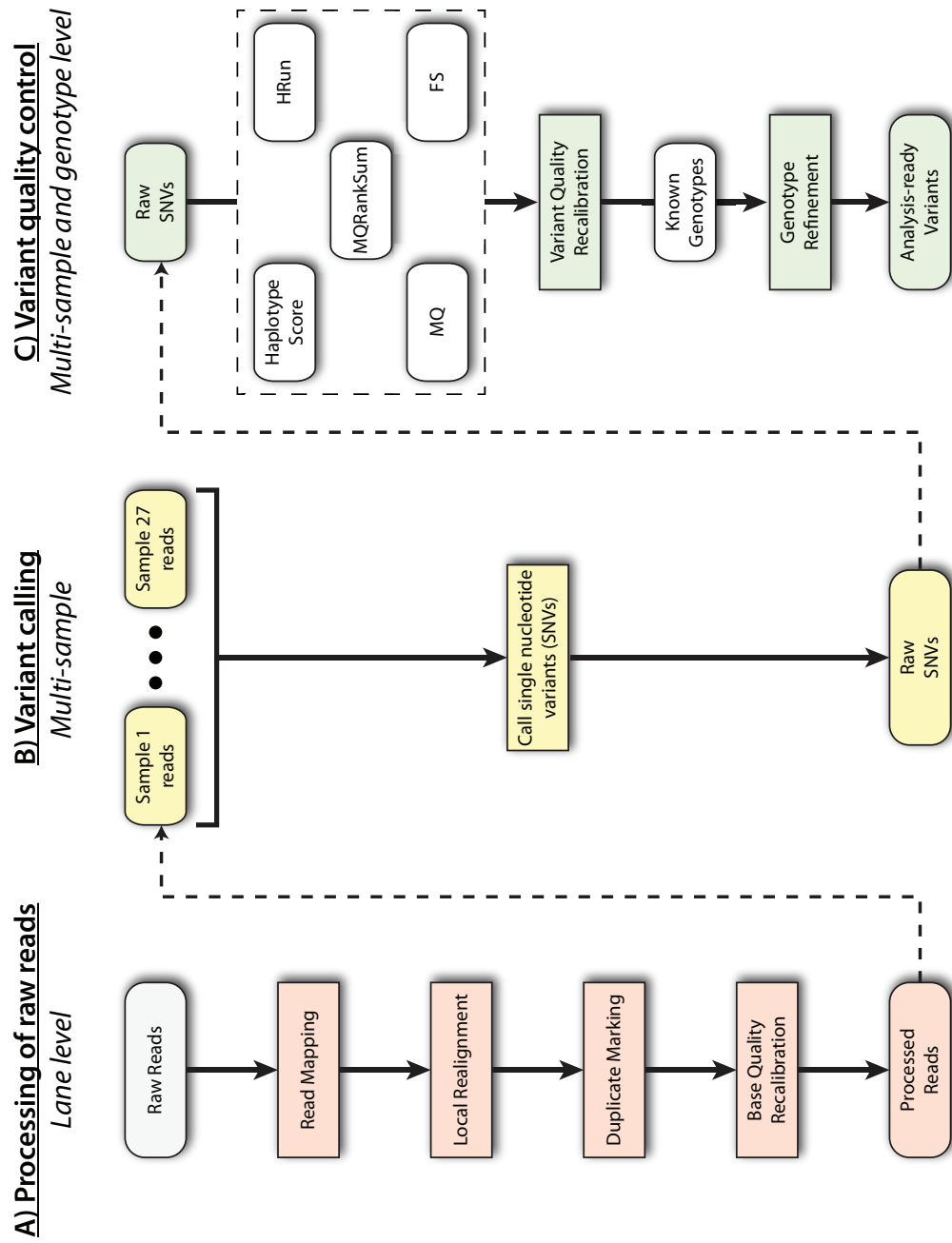


Figure 4.2: The exome-sequencing calling and quality-control pipeline.

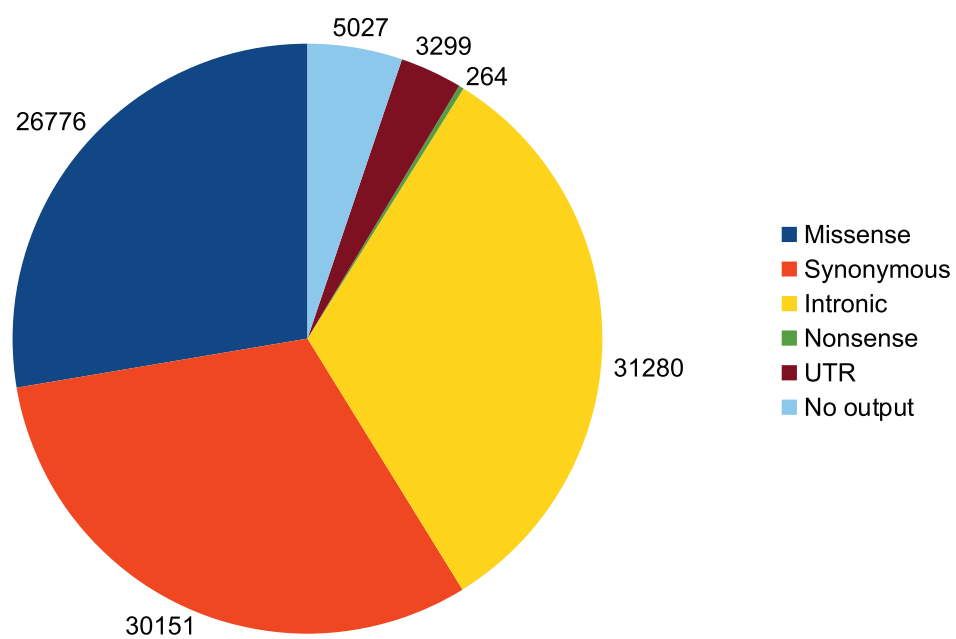


Figure 4.3: A breakdown of variant sites by genomic location as determined by the MapSNPs program included with PolyPhen2.

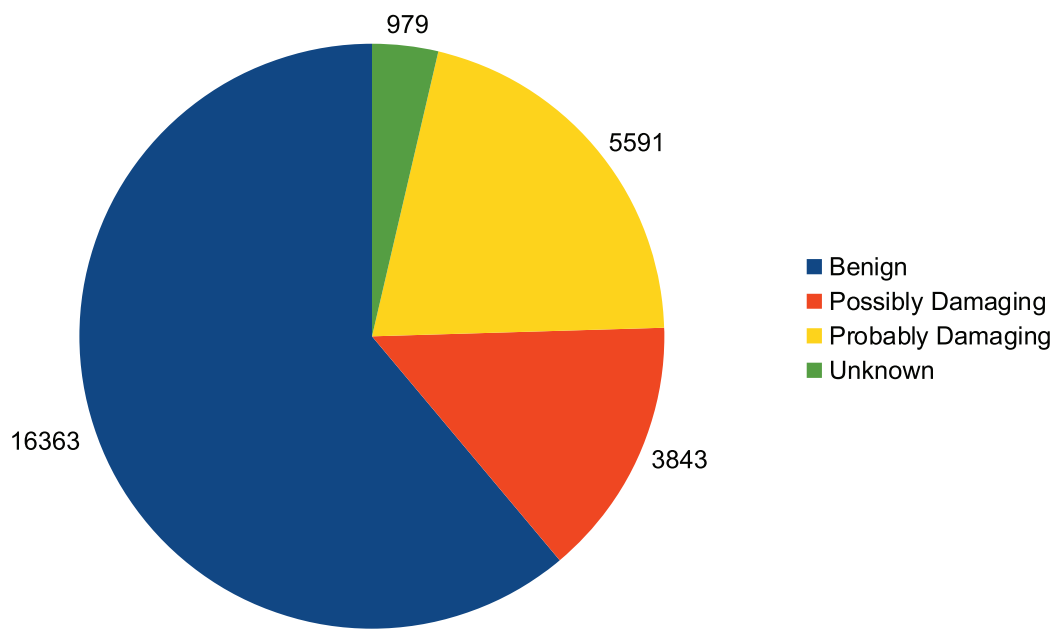


Figure 4.4: A breakdown of variant sites by PolyPhen2 classification.

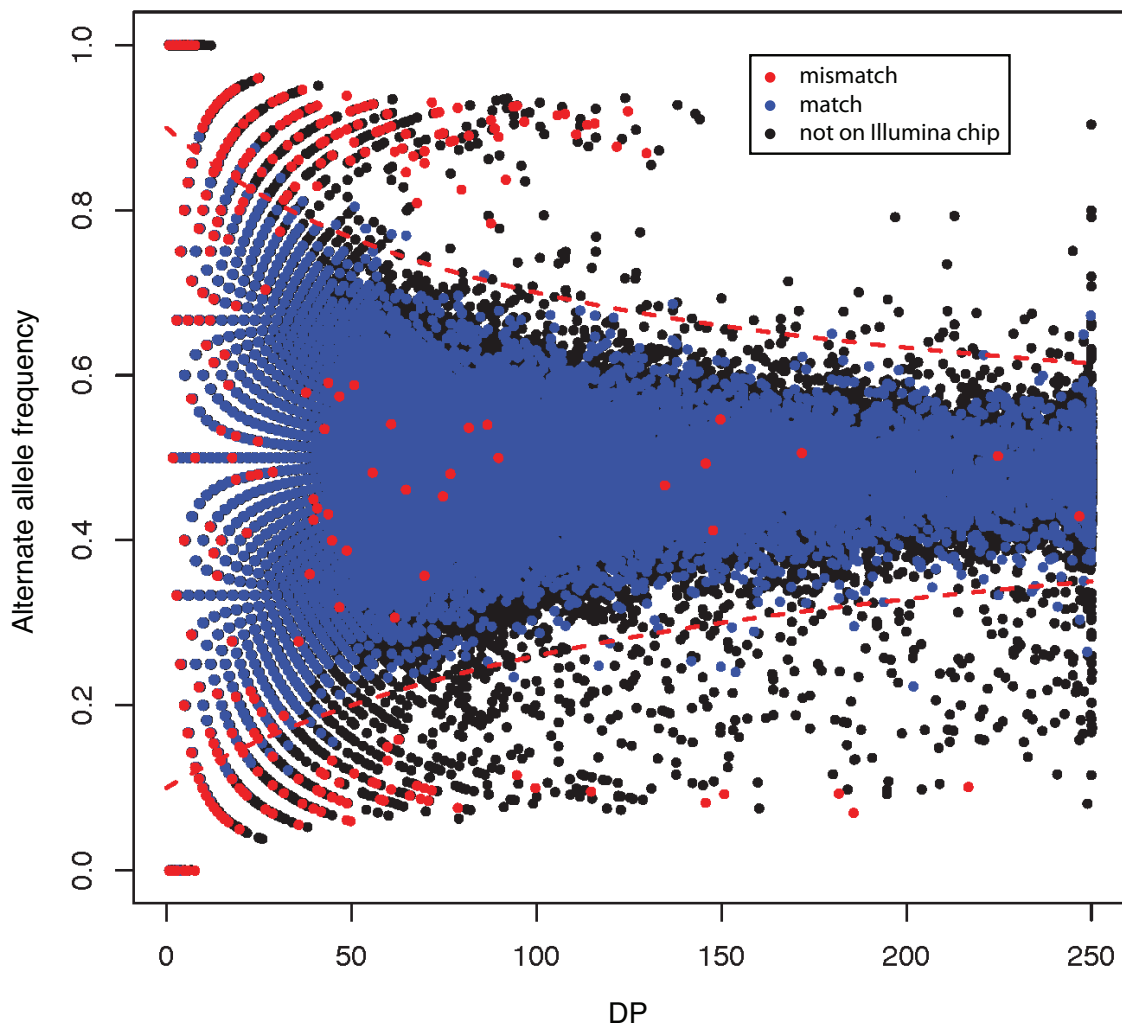


Figure 4.5: Alternate allele frequency versus read depth (DP) for heterozygous genotypes. Red points represent called exome heterozygotes that are homozygotes in the Illumina genotype data. Blue points represent called exome heterozygotes that are heterozygotes in the Illumina genotype data. Black points represent called exome heterozygotes that do not occur in the Illumina genotype data. Dotted red lines represent the quality control cutoff for heterozygous genotypes. Any heterozygotes that fall in between the two red dotted lines are retained. Others are removed from the dataset.

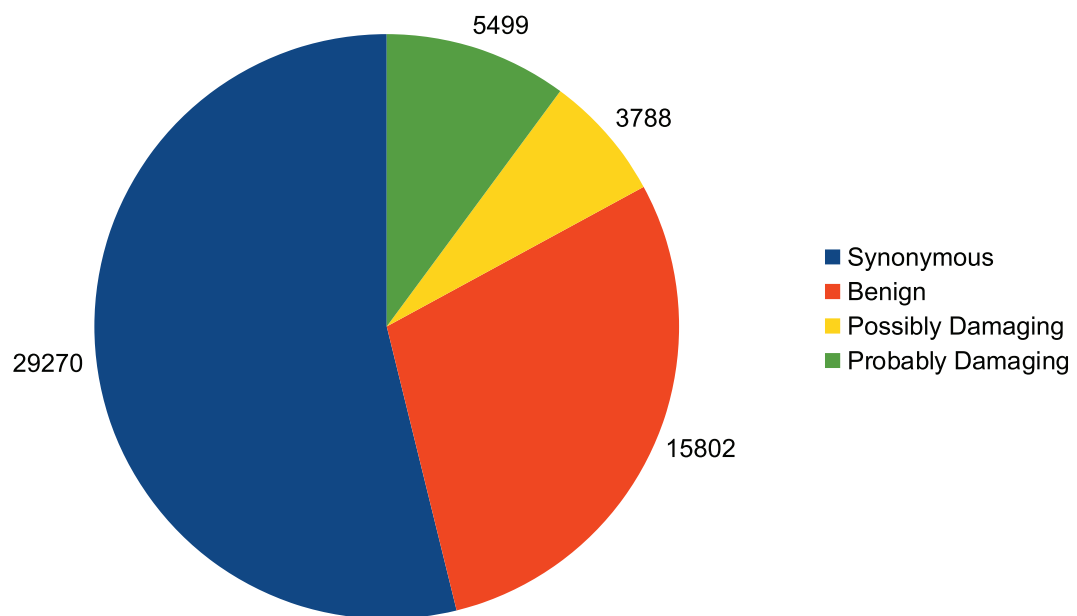


Figure 4.6: A breakdown of the final data set by PolyPhen2 classification and genome location after all filtering.

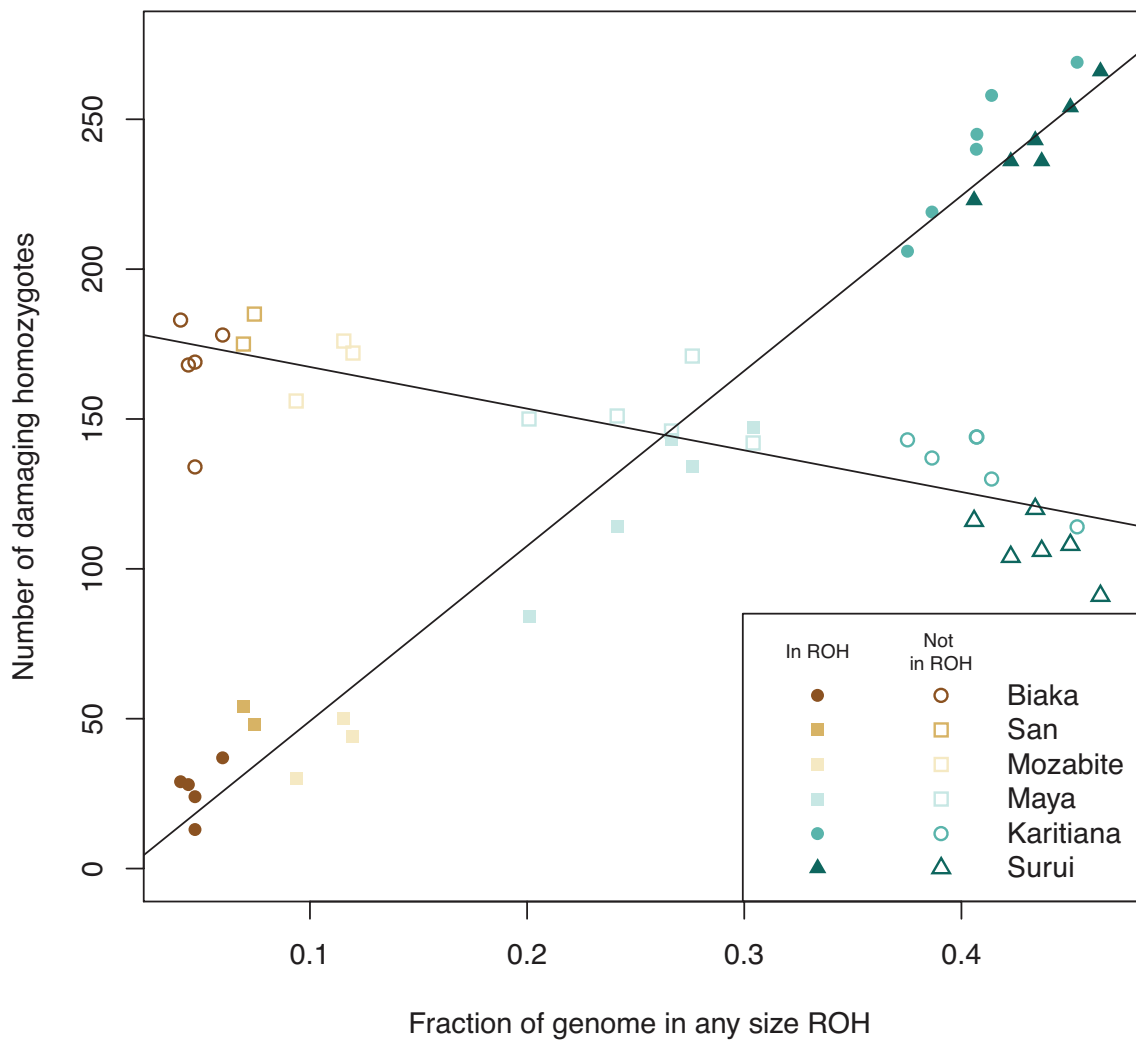


Figure 4.7: The number of damaging homozygotes versus the fraction of the genome covered by a called ROH for each individual. Colored points represent the number of damaging homozygotes falling within ROH regions, and block points represent the number of damaging homozygotes falling outside ROH regions.

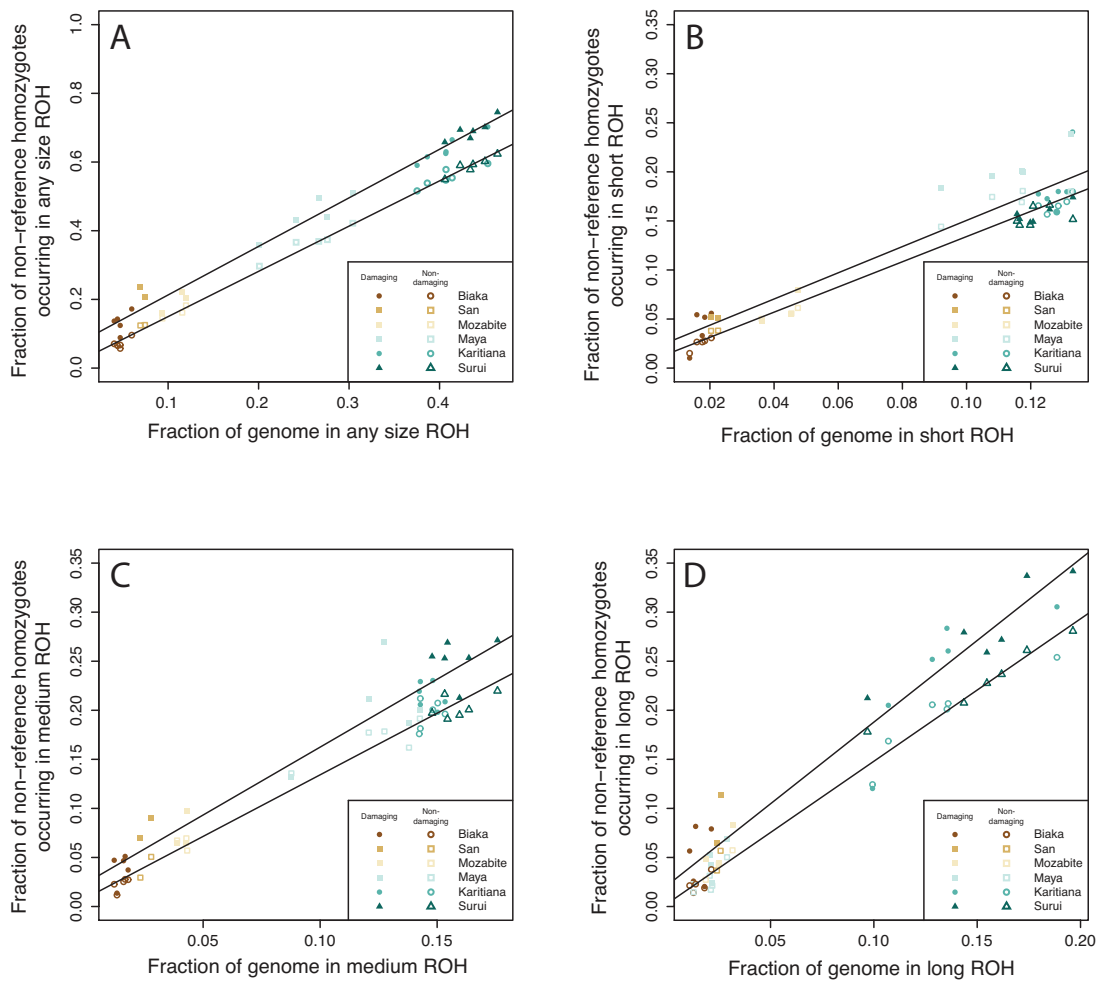


Figure 4.8: The fraction of all genome-wide non-reference homozygotes falling in ROH regions versus the fraction of the genome covered by a called ROH for each individual. A) any ROH region, B) class A ROH regions, C) class B ROH regions, and D) class C ROH regions. Solid points represent damaging homozygotes, and open points represent non-damaging homozygotes.

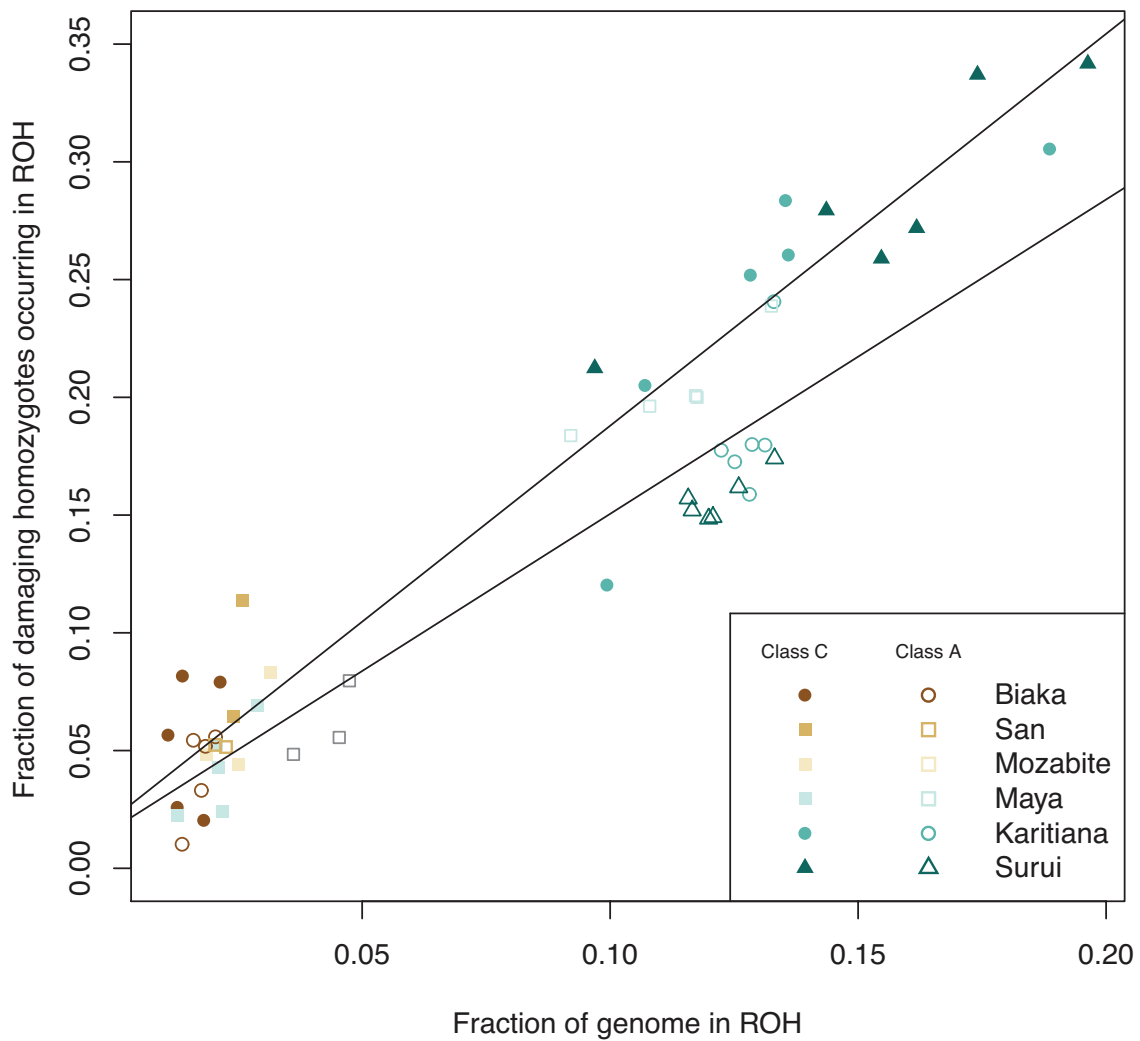


Figure 4.9: The fraction of all genome-wide non-reference homozygotes falling in ROH regions versus the fraction of the genome covered by a called ROH for each individual. Solid points represent damaging homozygotes in class C ROH regions, and open points represent damaging homozygotes in class A ROH regions.

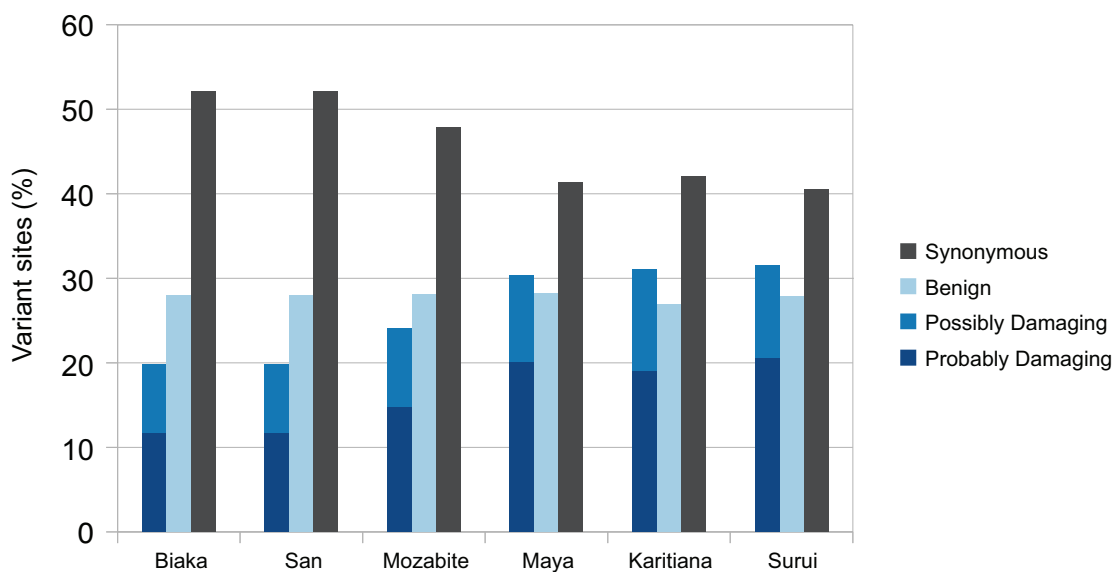


Figure 4.10: The fraction of all private variants that are synonymous or missense per population. Missense variants are further split into predicted benign, predicted possibly damaging, and predicted possibly damaging classes by PolyPhen2.

Table 4.1: Called exome genotype concordance with known Illumina genotypes per individual.

Population	Individual ID	Reference homozygote concordance (%)	Heterozygote concordance (%)	Non-reference homozygote concordance (%)
San	991	99.7	99.6	99.9
	992	99.7	99.2	99.5
Biaka	454	99.6	99.5	99.4
	457	99.7	99.4	99.6
	458	99.5	98.7	98.9
	459	99.6	99.4	99.4
	460	98.6	86.0	97.5
Mozabite	1264	99.7	99.3	99.5
	1267	99.8	99.3	99.8
	1274	99.9	99.7	99.8
Maya	854	99.4	99.8	99.4
	855	99.6	99.7	99.5
	856	99.8	99.8	99.5
	860	99.0	99.5	98.1
	868	99.4	99.6	98.6
	999	99.6	99.4	99.7
Karitiana	1012	99.9	99.6	99.7
	1014	99.6	99.1	99.5
	1015	99.6	99.4	99.7
	1018	99.8	99.9	99.7
	1019	99.6	99.2	99.6
Surui	837	99.3	98.5	99.0
	838	99.4	98.9	99.3
	843	98.7	97.1	98.9
	845	99.3	98.6	99.6
	846	99.4	99.1	99.0
	849	99.4	99.6	99.5

Table 4.2: Mean coverage and percentage of sites with $\geq 20x$ coverage for the 96797 quality-controlled sites (Section 4.2.2) for each individual.

Population	Individual ID	Mean coverage (number of copies)	Percentage of sites with $\geq 20x$ coverage
San	991	44	75
	992	38	70
Biaka	454	70	81
	457	58	82
	458	56	75
	459	42	75
	460	49	70
Mozabite	1264	36	81
	1267	38	84
	1274	44	90
Maya	854	62	79
	855	68	82
	856	72	83
	860	55	72
	868	48	71
Karitiana	999	57	77
	1012	74	83
	1014	66	78
	1015	53	73
	1018	75	84
Surui	1019	60	76
	837	47	62
	838	81	79
	843	66	69
	845	83	78
	846	64	72
	849	74	79

Table 4.3: Genotype counts per individual for sites classified as damaging.

Population	Individual ID	class A ROH		class B ROH		class C ROH		not-ROH					
		1/1	0/1	0/0	1/1	0/1	0/0	1/1	0/1	0/0			
San	991	12	12	194	21	13	231	15	1	136	185	1237	7137
	992	12	8	179	16	3	139	26	2	266	175	1168	7189
Biaka	454	10	4	124	9	9	118	5	1	68	169	1225	7384
	457	2	4	83	10	1	143	16	2	107	168	1224	7396
	458	12	5	144	8	1	89	17	1	230	178	1100	7216
	459	7	5	129	10	1	142	12	0	119	183	1171	7379
Mozabite	460	8	10	134	2	4	68	3	1	97	134	1032	7127
	1264	18	8	373	22	8	280	10	2	206	176	923	7211
Maya	1267	12	13	260	14	9	355	18	8	309	172	946	7128
	1274	9	19	299	12	5	331	9	0	169	156	1005	7246
	854	58	30	935	78	33	1137	7	4	159	146	715	5853
	855	69	31	1049	58	27	1268	20	7	374	142	678	5415
Karitiana	856	61	23	965	57	32	1021	16	9	242	171	690	5885
	860	43	30	674	31	20	688	10	5	118	150	746	6156
	868	52	26	876	56	37	1040	6	3	76	151	676	6018
	999	84	40	1096	80	30	1249	42	4	692	143	583	5111
Surui	1012	67	34	938	81	50	1358	110	4	1166	130	549	4687
	1014	70	46	1067	77	28	1412	98	6	1202	144	495	4462
	1015	64	36	1079	82	40	1180	73	4	1000	137	547	4842
	1018	61	36	955	79	33	1219	100	9	1298	144	568	4671
Surui	1019	68	28	1043	84	25	1161	117	3	1665	114	480	4272
	837	53	29	789	91	35	1365	122	5	1782	91	461	3873
	838	59	44	1072	92	40	1358	72	7	1028	116	518	4528
	843	55	30	910	86	18	1254	95	6	1180	104	493	4247
Surui	845	55	27	912	77	41	1308	122	18	1492	108	452	4299
	846	51	36	932	92	29	1238	93	5	1360	106	521	4453
	849	57	34	944	92	29	1391	94	12	1364	120	492	4264

Table 4.4: Genotype counts per individual for sites classified as non-damaging.

Population	Individual ID	class A ROH			class B ROH			class C ROH			non-ROH		
		1/1	0/1	0/0	1/1	0/1	0/0	1/1	0/1	0/0	1/1	0/1	0/0
San	991	181	74	876	240	64	1056	174	5	674	4149	8110	28986
	992	179	59	720	139	24	631	268	5	1203	4134	8025	28997
Biaka	454	114	27	547	104	12	533	58	0	306	3831	8212	30253
	457	65	25	348	119	19	466	98	12	389	4008	8344	30416
	458	127	39	600	113	11	433	156	1	816	3739	7877	29494
	459	115	25	592	98	7	449	92	0	403	4009	8359	30172
	460	89	42	635	39	27	269	63	5	430	3151	7633	28678
Mozabite	1264	214	49	1536	199	45	1282	151	15	993	2927	6812	30447
	1267	201	59	1228	251	37	1392	207	7	1262	2952	6863	30355
	1274	184	40	1189	247	45	1357	100	1	555	3142	7344	30739
Maya	854	676	133	4211	713	156	5033	85	4	549	2522	5009	25091
	855	766	160	4843	815	126	5177	214	14	1332	2461	4701	23549
	856	747	154	4190	670	88	4589	130	13	868	2591	4956	25279
	860	499	117	2963	470	73	2951	59	8	489	2435	5408	26106
	868	671	134	3877	682	114	4523	56	4	326	2438	5264	25259
	999	814	172	4867	960	151	5510	563	14	2899	2192	4269	21757
Karitiana	1012	730	150	4488	914	175	5989	937	21	5367	2077	3877	19656
	1014	781	159	4645	980	162	6065	972	20	5186	1993	3772	19201
	1015	776	138	4646	918	133	5386	772	22	4182	2112	4068	20704
	1018	740	130	4388	844	126	5488	962	36	5183	2105	3960	20398
	1019	789	159	4445	839	111	5134	1211	25	7417	1929	3417	18197
	837	654	129	3895	885	135	5423	1260	23	7396	1690	3385	16831
	838	654	195	4537	948	157	5779	768	38	4290	1945	3978	19707
Surui	843	691	152	4000	900	106	5184	863	36	5074	1705	3371	18235
	845	649	161	4035	869	144	5609	1163	51	6437	1774	3598	18242
	846	732	170	4322	847	139	5364	1049	30	6072	1805	3792	18424
	849	673	148	4358	901	119	5664	1022	33	5909	1896	3681	18358

Table 4.5: Percentage of all genotypes that are heterozygous in a given region of an individual's genome (genome-wide, non-ROH, any ROH, class A ROH, class B ROH, and class C ROH).

Population	Individual ID	Genome-wide (%)	non-ROH (%)	any ROH (%)	class A (%)	class B (%)	class C (%)
San	991	17.7	18.8	4.2	6.4	4.7	0.6
	992	17.4	18.5	2.6	5.8	2.8	0.4
Biaka	454	17.9	18.5	2.6	3.8	2.7	0.2
	457	18.0	18.6	3.3	5.5	2.6	2.2
	458	17.2	18.1	2.1	4.7	1.8	0.2
	459	17.9	18.6	1.7	3.4	1.1	0.0
	460	17.6	18.1	4.6	5.7	7.6	1.0
Mozabite	1264	14.6	15.9	2.3	2.6	2.9	1.2
	1267	14.7	16.1	2.4	4.1	2.2	0.8
	1274	15.6	16.8	2.4	3.4	2.5	0.1
Maya	854	11.4	14.6	2.6	2.7	2.6	1.0
	855	10.8	14.6	2.2	2.8	2.0	1.1
	856	11.2	14.3	2.3	2.9	1.9	1.7
	860	12.8	15.0	2.7	3.4	2.2	1.9
	868	12.0	14.9	2.5	2.8	2.3	1.5
Karitiana	999	9.1	14.3	1.9	2.9	2.3	0.4
	1012	8.8	14.2	1.8	3.0	2.6	0.3
	1014	9.4	14.2	1.8	2.6	2.2	0.3
	1015	9.1	14.2	1.7	2.6	2.2	0.4
	1018	8.1	13.7	1.4	2.9	2.0	0.6
	1019	9.9	14.2	2.1	3.0	1.8	0.3
	837	8.3	14.6	1.5	2.8	2.1	0.3
Surui	838	9.6	14.6	2.3	3.6	2.4	0.7
	843	8.6	13.7	1.7	3.1	1.6	0.6
	845	8.7	14.2	1.9	3.2	2.3	0.7
	846	9.1	14.8	1.8	3.3	2.2	0.4
	849	8.8	14.5	1.6	2.9	1.8	0.5

CHAPTER V

Conclusion

In this dissertation I have presented theoretical and empirical analyses of two important classes of genetic variation important in evolutionary genetics: private variants and coding variants. I have developed theoretical and empirical population-genetic techniques to analyze genomic data from diverse worldwide human populations. The analyses involving population-specific genetic variation have provided valuable information on the evolutionary relationships between populations, while my analyses of coding variation have provided insights into the distribution of deleterious variation with respect to an important genomic variable. I have shown that stretches of homozygous genotypes in the genome that are the result of recent inbreeding are enriched for deleterious variants and that individual genomes carry a large number of these variants in homozygous form. These results help to clarify the allelic architecture of human diseases and provide parameter boundaries for modeling evolutionary processes.

In Chapter II, I have developed the concept of generalized private alleles and introduced a method to analyze their distribution across populations while correcting for heterogeneity in population sample sizes. I have applied this method to a dataset of 1048 human individuals from 52 human HGDP-CEPH populations typed at 783 microsatellite markers. I observed an excess number of alleles private to the

combination of African and Oceanian populations, as compared to the number of alleles private to all other combinations including Oceania. This result supports the theory of a migration out of Africa into Oceania separate from the migrations responsible for the majority of the ancestry of the modern populations of Eurasia and East Asia (*Quintana-Murci et al.*, 1999; *Bulbeck*, 2007).

In Chapter III, I have taken a theoretical approach to modeling a peculiar population-genetic phenomenon. Using the coalescent model of gene genealogies and the symmetric stepwise mutation model, I investigated the observation that private microsatellite alleles often lie in the extreme tails of the allele size distribution. For the case in which four alleles are sampled, two from each population, I conditioned on the configuration in which three distinct allele sizes are present, one of which is common to both populations, one of which is private to one population, and the third of which is private to the other population. I calculated the probability that the private alleles in this scenario occur on the edges of the size distribution as a function of two population-genetic variables, mutation rate and population divergence. This probability, which increases as a function of mutation rate and divergence time between the two populations, is seen to be greater than the value that would be predicted if there was no relationship between privacy and location in the allele size distribution. I further found that, based on this model, increased population divergence is positively correlated with the probability that private alleles lie on the edge of the size distribution.

To test the prediction that population divergence has a strong influence on the occurrence of private microsatellite alleles on the edge of the size distribution, I empirically estimated the frequency that private alleles occur in the edge of the size distribution for all pairs among the 52 HGDP-CEPH populations. In accordance with the prediction of the model, the frequency with which private microsatellite alleles occur in the tails of the allele size distribution increases as a function of genetic

differentiation between populations. I concluded that the model accurately describes patterns of private microsatellites in diverged populations, and that the frequency of occurrence of private microsatellite alleles on the edge of the size distribution could potentially be used to make inferences about population divergence times.

Finally in Chapter IV, I turned to the analysis of coding variation. In the analysis of the exome sequences of 27 individuals across 6 HGDP-CEPH populations, I collected a set of predicted deleterious variants and examined the patterns of deleterious variation with respect to runs of homozygosity (ROH). ROH are generated by numerous processes, including changes in population demography, consanguinity, and natural selection. Because of the range of causes for ROH in the genome, I formulated two hypotheses for how deleterious variants could correlate with ROH. First, because many deleterious variants act recessively, they might be more effectively purged by selection in ROH-rich regions than in non-ROH regions. Thus, ROH could contain a lower fraction of all genome-wide damaging homozygotes relative to the fraction of all genome-wide non-damaging homozygotes. On the other hand, inbreeding can present low-frequency variants in homozygous form, and many of these variants are more likely to be deleterious than are common variants. When the homozygous deleterious variants are not lethal, one might expect ROH to contain a higher fraction of genome-wide deleterious homozygotes compared to the fraction of all genome-wide neutral homozygotes, and longer (newer) ROH might have a higher fraction than shorter (older) ROH.

I found that, for individual genomes, long ROH are enriched for deleterious variation. Specifically, the fraction of all genome-wide homozygotes lying in ROH is positively correlated with the total length of ROH in the genome. Damaging homozygotes show a significantly higher fraction falling into any size ROH than non-damaging homozygotes. This trend is also significant for long ROH but not for intermediate or short ROH. Furthermore, damaging homozygotes show a

significantly higher fraction falling into long ROH than short ROH. In this study, I provided a demonstration that long ROH harbor disproportionately more deleterious homozygotes than would be predicted solely by the ROH coverage of the genome. I further showed that human individuals can carry a large number of deleterious homozygotes, consistent with other studies (*Lohmueller et al.*, 2008) and with the idea that the human genome contains many mildly deleterious variants. This knowledge will be useful for understanding the basis of human diseases.

In this thesis, I have described and explained patterns of private genetic variation and the evolutionary processes that influence them, thus further expanding our knowledge of human history and evolution. I have also analyzed the distribution of deleterious variation with respect to runs of homozygosity, an important genomic variable associated with population-genetic processes such as natural selection and cultural practices such as consanguineous marriage. This work builds on existing theoretical and empirical results and advances our understanding of human population and evolutionary genetics.

APPENDICES

APPENDIX A

Derivation for arbitrary divergence time

The expression that must be calculated in order to obtain the probability that the private alleles lie on the edges of the size distribution for arbitrary t_d appears in Equation 3.17. To perform the calculation in Equation 3.17, we must utilize the probability that two lineages reduce to one lineage during time t_d as well as the probability that two lineages survive until t_d . Under the coalescent (*Wakeley, 2009*), these probabilities are $g_{21}(t_d) = 1 - e^{-t_d}$ and $g_{22}(t_d) = e^{-t_d}$, where $g_{ij}(t_d)$ denotes the probability under the coalescent that i lineages reduce to j lineages during time t_d .

We can partition our probability calculation into four pieces corresponding to the four coalescent scenarios possible by time t_d (Figure 3.6). First, in each population, the two lineages could coalesce more recently than t_d (event E_{11}). Second, the two lineages in population 1 could coalesce more recently than t_d , and the two lineages in population 2 could survive to t_d (event E_{12}). Third, the two lineages in population 1 could survive to t_d , and the two lineages in population 2 could coalesce more recently than t_d (event E_{21}). Finally, in each population, the two lineages could survive to t_d (event E_{22}). These four events happen with the following probabilities:

$$\mathbb{P}[E_{11}|t_d] = (g_{21}(t_d))^2 = (1 - e^{-t_d})^2, \quad (\text{A.1})$$

$$\mathbb{P}[E_{12}|t_d] = g_{21}(t_d)g_{22}(t_d) = (1 - e^{-t_d})e^{-t_d}, \quad (\text{A.2})$$

$$\mathbb{P}[E_{21}|t_d] = g_{22}(t_d)g_{21}(t_d) = e^{-t_d}(1 - e^{-t_d}), \quad (\text{A.3})$$

$$\mathbb{P}[E_{22}|t_d] = (g_{22}(t_d))^2 = e^{-2t_d}. \quad (\text{A.4})$$

We then calculate $\mathbb{P}[E_1 \cup E_2|C_6, \theta, t_d]$ by separately conditioning on E_{11} , E_{12} , E_{21} , and E_{22} to get

$$\mathbb{P}[E_1 \cup E_2|C_6, \theta, t_d] = \frac{2 \sum_{i=1}^2 \sum_{j=1}^2 \mathbb{P}[E_1, C_6|\theta, t_d, E_{ij}]\mathbb{P}[E_{ij}|\theta, t_d]}{\sum_{i=1}^2 \sum_{j=1}^2 \mathbb{P}[C_6|\theta, t_d, E_{ij}]\mathbb{P}[E_{ij}|\theta, t_d]}, \quad (\text{A.5})$$

in which

$$\mathbb{P}[E_1, C_6|\theta, t_d, E_{ij}] = \int_0^\infty \int_0^\infty \int_0^\infty \mathbb{P}[E_1, C_6|t_2, t_3, t_4, \theta, t_d, E_{ij}]\rho_{ij}(t_2, t_3, t_4)dt_2 dt_3 dt_4, \quad (\text{A.6})$$

$$\mathbb{P}[C_6|\theta, t_d, E_{ij}] = \int_0^\infty \int_0^\infty \int_0^\infty \mathbb{P}[C_6|t_2, t_3, t_4, \theta, t_d, E_{ij}]\rho_{ij}(t_2, t_3, t_4)dt_2 dt_3 dt_4, \quad (\text{A.7})$$

$$\mathbb{P}[E_1, C_6|t_2, t_3, t_4, \theta, t_d, E_{ij}] = \sum_{n_3=-\infty}^\infty \sum_{n_A=-\infty}^\infty \sum_{n_B=-\infty}^{n_A-1} \sum_{n_C=n_A+1}^\infty V_{E_{ij}}(n_A, n_B, n_C, n_3, \Psi), \quad (\text{A.8})$$

and

$$\mathbb{P}[C_6|t_2, t_3, t_4, \theta, t_d, E_{ij}] = \sum_{n_3=-\infty}^\infty \sum_{n_A=-\infty}^\infty \sum_{n_B=-\infty}^\infty \sum_{n_C=-\infty}^\infty V_{E_{ij}}(n_A, n_B, n_C, n_3, \Psi). \quad (\text{A.9})$$

We can determine the values of the conditional probability $V_{E_{ij}}$ of the node allele sizes and the conditional coalescence time density ρ_{ij} by examining which labeled histories are possible for each E_{ij} . For example, for event E_{11} both pairs of lineages coalesce more recently than time t_d , and only symmetric histories are possible. Furthermore, x_1 will always coalesce with x_2 and x_3 will always coalesce with x_4 in this scenario, leaving only two possible equiprobable histories (histories 13 and 18 in Figure 3.1). Therefore, we only sum over the V^{sym} terms that are associated with these histories.

In addition, for each event, compared to the case of $t_d = 0$, we must reparameterize the branch lengths of the histories to account for changes due to forced survival of lineages to time t_d . For event E_{11} , we reparameterize by setting $\Psi = (\tau, \theta)$ with $\tau = (t_2 + (t_d - \max(t_3, t_4)), \max(t_3, t_4) - \min(t_3, t_4), \min(t_3, t_4))$, as illustrated in Figure 3.6 and tabulated in Table A.1. By conditioning on one of the four events E_{11} , E_{12} , E_{21} , or E_{22} , the density of coalescence times differs from the corresponding density $\rho(t_2, t_3, t_4)$ defined in the $t_d = 0$ case.

For event E_{11} , the distribution of coalescence times is $\rho_{11}(t_2, t_3, t_4) = \rho_{11t_2}(t_2)\rho_{11t_3}(t_3)\rho_{11t_4}(t_4)$, where $\rho_{11t_2}(t) = e^{-t}$ and $\rho_{11t_3}(t) = \rho_{11t_4}(t) = \mathbf{1}_{t < t_d} e^{-t} / (1 - e^{-t_d})$. We can then write

$$\begin{aligned}
V_{E_{11}} = & \frac{1}{2} \left(V^{sym}(n_A, n_B, n_A, n_C, n_3, \Psi) + V^{sym}(n_A, n_B, n_C, n_A, n_3, \Psi) + \right. \\
& V^{sym}(n_B, n_A, n_A, n_C, n_3, \Psi) + V^{sym}(n_B, n_A, n_C, n_A, n_3, \Psi) + \\
& V^{sym}(n_A, n_A, n_B, n_C, n_3, \Psi) + V^{sym}(n_A, n_C, n_B, n_A, n_3, \Psi) + \\
& \left. V^{sym}(n_B, n_A, n_A, n_C, n_3, \Psi) + V^{sym}(n_B, n_C, n_A, n_A, n_3, \Psi) \right). \quad (\text{A.10})
\end{aligned}$$

We proceed with similar arguments for events E_{12} , E_{21} , and E_{22} . The corresponding values for Ψ are tabulated in Table A.1, and the values for $V_{E_{ij}}$ and ρ_{ij} are tabulated in Table A.2.

Table A.1: The reparameterizations of Ψ for the events E_{ij} .

Event	$\Psi = (\tau, \theta)$
E_{11}	$\tau = (t_2 + t_d - \max(t_3, t_4), \max(t_3, t_4) - \min(t_3, t_4), \min(t_3, t_4))$
E_{12}	$\tau = (t_2, t_3 + t_d - t_4, t_4)$
E_{21}	$\tau = (t_2, t_3 + t_d - t_4, t_4)$
E_{22}	$\tau = (t_2, t_3, t_4 + t_d)$

Table A.2: The probabilities of node allele sizes and the coalescence time densities conditional on events E_{ij} .

Event	$V_{E_{ij}}$	$\rho_{ij}(t_2, t_3, t_4, t_d) = \rho_{ij t_2}(t_2, t_d) \rho_{ij t_3}(t_3, t_d) \rho_{ij t_4}(t_4, t_d)$	
		$\rho_{ij t_2}(t, t_d)$	$\rho_{ij t_3}(t, t_d)$
E_{11}	$\frac{1}{2} \left(4V^{sym}(n_A, n_B, n_A, n_C, n_3, \Psi) + 4V^{sym}(n_A, n_C, n_A, n_B, n_3, \Psi) \right)$	e^{-t}	$\mathbf{1}_{t < t_d} e^{-t} / (1 - e^{-t_d})$
			$\mathbf{1}_{t < t_d} e^{-t} / (1 - e^{-t_d})$
E_{12}	$\frac{1}{3} \left(4V^{cat}(n_A, n_B, n_A, n_C, n_3, \Psi) + 4V^{cat}(n_A, n_B, n_C, n_A, n_3, \Psi) + 4V^{sym}(n_A, n_C, n_A, n_B, n_3, \Psi) \right)$	e^{-t}	$3e^{-3t}$
			$\mathbf{1}_{t < t_d} e^{-t} / (1 - e^{-t_d})$
E_{21}	$\frac{1}{3} \left(4V^{cat}(n_A, n_B, n_A, n_C, n_3, \Psi) + 4V^{cat}(n_A, n_B, n_C, n_A, n_3, \Psi) + 4V^{sym}(n_A, n_C, n_A, n_B, n_3, \Psi) \right)$	e^{-t}	$3e^{-3t}$
			$\mathbf{1}_{t < t_d} e^{-t} / (1 - e^{-t_d})$
E_{22}	$\frac{1}{18} \left(4V^{cat}(n_A, n_A, n_B, n_C, n_3, \Psi) + 4V^{cat}(n_A, n_A, n_C, n_B, n_3, \Psi) + 8V^{cat}(n_A, n_B, n_A, n_C, n_3, \Psi) + 8V^{cat}(n_A, n_B, n_C, n_A, n_3, \Psi) + 8V^{cat}(n_A, n_C, n_A, n_B, n_3, \Psi) + 8V^{cat}(n_A, n_C, n_B, n_A, n_3, \Psi) + 8V^{cat}(n_B, n_C, n_A, n_A, n_3, \Psi) + 8V^{sym}(n_A, n_B, n_A, n_C, n_3, \Psi) + 8V^{sym}(n_A, n_C, n_A, n_B, n_3, \Psi) + 4V^{sym}(n_A, n_A, n_B, n_C, n_3, \Psi) + 4V^{sym}(n_B, n_C, n_A, n_A, n_3, \Psi) \right)$	e^{-t}	$3e^{-3t}$
			$6e^{-6t}$

APPENDIX B

Implementing the computation for arbitrary divergence time

To implement the calculation of $\mathbb{P}[E_1 \cup E_2 | C_6, \theta, t_d]$ (Equation 3.17) derived in Appendix A, we use Gaussian quadrature and a simulation-based approach. These approaches are analogous to the approaches that we used in the case of $t_d = 0$.

As in the $t_d = 0$ case, we use Gaussian quadrature to numerically evaluate $\mathbb{P}[E_1, C_6 | \theta, t_d, E_{ij}]$ (Equation A.6) and $\mathbb{P}[C_6 | \theta, t_d, E_{ij}]$ (Equation A.7), once again estimating the expression $e^{-t\theta/2} I_{|d|}(t\theta/2)$ using the GNU Scientific Library (GSL) function `gsl_sf_bessel_In_scaled(|d|, t\theta/2)`. We use the same value as in the $t_d = 0$ case (± 10) to truncate the infinite sums in Equations A.8 and A.9. Additionally, we again integrate all time dimensions in Equations A.6 and A.7 from 0 to 10 rather than from 0 to ∞ . As in the case of $t_d = 0$, these calculations are very accurate for small values of θ and less accurate for large values of θ (not shown).

As in the $t_d = 0$ case, we are able to accurately estimate the quantity $\mathbb{P}[E_1 \cup E_2 | C_6, \theta, t_d]$ (Equation 3.17), directly obtaining the ratio $2\mathbb{P}[E_1, C_6 | \theta, t_d] / \mathbb{P}[C_6 | \theta, t_d]$ by simulating the coalescent and mutation processes and counting the outcomes of interest. The simulation proceeds as follows.

1. Beginning with $k = 4$ alleles, arbitrarily define two alleles to be in one population and the other two alleles to be in the other population.
2. Randomly choose an event E_{11} , E_{12} , E_{21} , or E_{22} based on their relative probabilities conditional on t_d (Equations A.1-A.4).
3. If event E_{11} is chosen:
 - (a) Generate a random time to coalescence from an exponential($\binom{2}{2}$) distribution conditional on being less than t_d .
 - (b) Coalesce the pair of lineages in population 1; set $k = k - 1$.
 - (c) Generate a random time to coalescence from an exponential($\binom{2}{2}$) distribution conditional on being less than t_d .
 - (d) Coalesce the pair of lineages in population 2; set $k = k - 1$.
4. If event E_{12} or E_{21} is chosen:
 - (a) Generate a random time to coalescence from an exponential($\binom{2}{2}$) distribution conditional on being less than t_d .
 - (b) Coalesce a pair of lineages in population 1 (if event E_{12}) or population 2 (if event E_{21}); set $k = k - 1$.
5. Extend all remaining lineages up to t_d .
6. Generate a random time to coalescence from an exponential($\binom{k}{2}$) distribution.
7. Randomly choose two lineages to coalesce; set $k = k - 1$.
8. If $k \neq 1$, go to 6.
9. For each branch of the genealogy, generate a random number of mutation events, x , from a Poisson distribution with rate $\theta t/2$, where t is the branch length.

10. Assign each mutation a value of $+1$ or -1 by sampling the number of $+1$ mutations from a binomial($x, 1/2$) distribution. Those mutations not chosen to be $+1$ are assigned a value of -1 .
11. Determine the allele size of each of the four sampled alleles by summing the net value of mutations from the root (allele size 0) down to the leaves.
12. Classify the collection of four alleles into one of the seven allele configurations (Table 3.1).
13. If the alleles are in the C_6 configuration, accept the simulation and determine if the sizes of the private alleles (B and C) are on the ends of the distribution ($n_C < n_A < n_B$ or $n_B < n_A < n_C$). If yes, count a success.

As in the $t_d = 0$ case, after the number of accepted simulations reaches some pre-specified number (we choose 1,000,000), we estimate the probability of the private alleles occurring on the edges of the size distribution by dividing the number of successes by the number of accepted simulations.

APPENDIX C

Small- θ approximation for arbitrary divergence time

With $t_d > 0$, we can consider a small- θ approximation to the probability that the private alleles lie on the edges in a similar way to the corresponding calculation with $t_d = 0$. By considering a fixed t_d , we proceed as before, counting the contributions of each labeled history to the numerator and denominator in Equation 3.17. The probability distribution of labeled histories depends on t_d , and the 18 histories are no longer equiprobable when $t_d > 0$. Conditional on one of the events $\{E_{11}, E_{12}, E_{21}, E_{22}\}$, however, we can determine the possible histories and weight the probability contributions of these histories to the numerator and denominator as before.

Thus, following Equation A.5 for the small- θ case, we wish to calculate

$$\mathbb{P}[E_1 \cup E_2 | C_6, \theta_{small}, t_d] = \frac{\sum_{i=1}^2 \sum_{j=1}^2 \mathbb{P}[E_1 \cup E_2, C_6 | \theta_{small}, t_d, E_{ij}] \mathbb{P}[E_{ij} | \theta_{small}, t_d]}{\sum_{i=1}^2 \sum_{j=1}^2 \mathbb{P}[C_6 | \theta_{small}, t_d, E_{ij}] \mathbb{P}[E_{ij} | \theta_{small}, t_d]} \quad (\text{C.1})$$

Note that although E_1 and E_2 have the same probability, in this calculation it is convenient to calculate $E_1 \cup E_2$ directly. We do this by tabulating contributions to the numerator and denominator conditional on each event E_{ij} (Table 3.5),

reparameterizing Ψ to augment certain branch lengths by amounts dependent on t_d (Table A.1).

First, consider event E_{11} . If both pairs of lineages coalesce more recently than the population divergence time, then the only possible histories are 13 and 18, and the conditional contribution to the denominator of Equation C.1 is

$$\mathbb{P}[C_6|\theta_{small}, t_d, E_{11}] = \int_0^\infty \int_0^\infty \int_0^\infty \frac{1}{2} 16\sigma_8(t_2 + t_d - t_{\max}, t_{\max} - t_{\min}, t_{\min}, \theta) \times \\ \rho_{E_{11}}(t_2, t_3, t_4, t_d) dt_2 dt_3 dt_4, \quad (\text{C.2})$$

where $t_{\max} = \max(t_3, t_4)$ and $t_{\min} = \min(t_3, t_4)$. Here, we obtain the coefficients for each σ_i by referencing histories 13 and 18 in Table 3.5, and we use the conditional density of coalescence times $\rho_{E_{11}}(t_2, t_3, t_4, t_d)$ from Table A.2. Equation C.2 also provides the $\mathbb{P}[E_1 \cup E_2, C_6|\theta_{small}, t_d, E_{11}]$ term in the numerator, because for histories 13 and 18, at small θ , the private alleles always lie on the edges of the size distribution.

Next, consider event E_{ij} ($i \neq j$). If the two lineages in one population coalesce more recently than the divergence time, and the two lineages in the other population survive to the divergence time, then the only possible histories are 1, 2, and 18 for E_{12} or 11, 12, and 13 for E_{21} . Because E_{12} and E_{21} differ only in which population contains the coalescence more recent than the population divergence, they have the same probability. The conditional contribution to the denominator for either event is then

$$\mathbb{P}[C_6|\theta_{small}, t_d, E_{ij}] = \int_0^\infty \int_0^\infty \int_0^\infty \frac{1}{3} \left(8\sigma_1(t_2, t_3 + t_d - t_4, t_4, \theta) + 8\sigma_2(t_2, t_3 + t_d - t_4, t_4, \theta) + \right. \\ \left. 8\sigma_8(t_2, t_3 + t_d - t_4, t_4, \theta) \right) \rho_{E_{ij}}(t_2, t_3, t_4, t_d) dt_2 dt_3 dt_4, \quad (\text{C.3})$$

where the σ_i coefficients are taken from Table 3.5 using either set of histories (1, 2,

and 18 for E_{12} or 11, 12, and 13 for E_{21}) and $\rho_{E_{ij}}(t_2, t_3, t_4, t_d)$ is taken from Table A.2. Equation C.3 is also equal to the $\mathbb{P}[E_1 \cup E_2, C_6 | \theta_{small}, t_d, E_{ij}]$ term in the numerator, because for either set of histories, at small θ , the private alleles always lie on the edges of the size distribution.

For event E_{22} , if in both populations the two lineages survive to the divergence time, then all 18 histories are possible. The conditional contribution to the denominator is

$$\begin{aligned} \mathbb{P}[C_6 | \theta_{small}, t_d, E_{22}] = & \int_0^\infty \int_0^\infty \int_0^\infty \frac{1}{18} \left(32\sigma_1(t_2, t_3, t_4 + t_d, \theta) + 32\sigma_2(t_2, t_3, t_4 + t_d, \theta) + \right. \\ & 16\sigma_3(t_2, t_3, t_4 + t_d, \theta) + 16\sigma_4(t_2, t_3, t_4 + t_d, \theta) + \\ & 16\sigma_5(t_2, t_3, t_4 + t_d, \theta) + 16\sigma_6(t_2, t_3, t_4 + t_d, \theta) + \\ & 32\sigma_7(t_2, t_3, t_4 + t_d, \theta) + 32\sigma_8(t_2, t_3, t_4 + t_d, \theta) + \\ & 8\sigma_9(t_2, t_3, t_4 + t_d, \theta) + 16\sigma_{10}(t_2, t_3, t_4 + t_d, \theta) + \\ & \left. 8\sigma_{11}(t_2, t_3, t_4 + t_d, \theta) + 16\sigma_{12}(t_2, t_3, t_4 + t_d, \theta) \right) \times \\ & \rho_{E_{22}}(t_2, t_3, t_4) dt_2 dt_3 dt_4 \end{aligned} \quad (\text{C.4})$$

and the conditional contribution to the numerator is

$$\begin{aligned}
\mathbb{P}[E_1 \cup E_2, C_6 | \theta_{small}, t_d, E_{22}] = & \int_0^\infty \int_0^\infty \int_0^\infty \frac{1}{18} \left(32\sigma_1(t_2, t_3, t_4 + t_d, \theta) + \right. \\
& 32\sigma_2(t_2, t_3, t_4 + t_d, \theta) + \\
& 16\sigma_3(t_2, t_3, t_4 + t_d, \theta) + \\
& 16\sigma_6(t_2, t_3, t_4 + t_d, \theta) + \\
& 32\sigma_8(t_2, t_3, t_4 + t_d, \theta) + \\
& 8\sigma_9(t_2, t_3, t_4 + t_d, \theta) + \\
& \left. 8\sigma_{11}(t_2, t_3, t_4 + t_d, \theta) \right) \times \\
& \rho_{E_{22}}(t_2, t_3, t_4) dt_2 dt_3 dt_4, \quad (C.5)
\end{aligned}$$

where the σ_i coefficients are from Table 3.5 and $\rho_{E_{22}}(t_2, t_3, t_4)$ is from Table A.2.

We can understand how Equation C.1 will behave for large values of t_d by considering the behavior of $\mathbb{P}[E_{ij}|t_d]$ (Equations A.1-A.4) as t_d tends toward ∞ . Independently of the value of θ , when the divergence time between populations grows very large, we expect each pair of lineages to always coalesce before the population divergence (event E_{11}). Taking the limits of Equations A.1-A.4, $\lim_{t_d \rightarrow \infty} \mathbb{P}[E_{11}|t_d] = 1$ and $\lim_{t_d \rightarrow \infty} \mathbb{P}[E_{12}|t_d] = \lim_{t_d \rightarrow \infty} \mathbb{P}[E_{21}|t_d] = \lim_{t_d \rightarrow \infty} \mathbb{P}[E_{22}|t_d] = 0$. Thus as t_d tends to ∞ , Equation C.1 reduces to

$$\mathbb{P}[E_1 \cup E_2 | C_6, \theta_{small}, t_d] = \frac{\mathbb{P}[E_1 \cup E_2, C_6 | \theta_{small}, t_d, E_{11}]}{\mathbb{P}[C_6 | \theta_{small}, t_d, E_{11}]} = 1. \quad (C.6)$$

Therefore, for large t_d , we intuitively expect the small- θ probability that the private alleles lie on the edges of the size distribution to tend to unity.

Note that Equations C.4 and C.5 differ from Equations 3.14 and 3.12 only in the definitions of the time parameters and densities of coalescence times. Using the

conditional contributions in Equations C.2-C.5 together with $\mathbb{P}[E_{ij}|t_d]$ in Equations A.1-A.4, we can calculate Equation C.1. The resulting expression is unwieldy (not shown), but taking its limit as θ tends to 0, we obtain Equation 3.18.

BIBLIOGRAPHY

BIBLIOGRAPHY

- Adzhubei, I. A., S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev (2010), A method and server for predicting damaging missense mutations, *Nature Methods*, 7, 248–249.
- Barton, N. H., and M. Slatkin (1986), A quasi-equilibrium theory of the distribution of rare alleles in a subdivided population, *Heredity*, 56, 409–415.
- Bittles, A. H. (2001), Consanguinity and its relevance to clinical genetics, *Clinical Genetics*, 60, 89–98.
- Bittles, A. H., and J. V. Neel (1994), The costs of human inbreeding and their implications for variations at the DNA level, *Nature Genetics*, 8, 117–121.
- Botstein, D., and N. Risch (2003), Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease, *Nature Genetics*, 33, 228–237.
- Broman, K. W., and J. L. Weber (1999), Long homozygous chromosomal segments in reference families from the Centre d'étude du Polymorphisme Humain, *American Journal of Human Genetics*, 65, 1493–1500.
- Bulbeck, D. (2007), Where river meets sea: A parsimonious model for *Homo sapiens* colonization of the Indian Ocean rim and Sahul, *Current Anthropology*, 48, 315–321.
- Burchard, E. G., et al. (2003), The importance of race and ethnic background in biomedical research and clinical practice, *New England Journal of Medicine*, 348, 1170–1175.
- Calabrese, P., and R. Durrett (2003), Dinucleotide repeats in the *Drosophila* and human genomes have complex, length-dependent mutation processes, *Molecular Biology and Evolution*, 20, 715–725.
- Calafell, F., A. Shuster, W. C. Speed, J. R. Kidd, and K. K. Kidd (1998), Short tandem repeat polymorphism evolution in humans, *European Journal of Human Genetics*, 6, 38–49.
- Campbell, M. C., and S. A. Tishkoff (2010), The evolution of human genetic and phenotypic variation in africa, *Current Biology*, 20, R166–R173.

- Cavalli-Sforza, L. L., and M. W. Feldman (2003), The application of molecular genetic approaches to the study of human evolution, *Nature Genetics*, *33*, 266–275.
- Cavalli-Sforza, L. L., P. Menozzi, and A. Piazza (1994), *The History and Geography of Human Genes*, Princeton University Press, Princeton.
- Chakraborty, R., and K. K. Kidd (1991), The utility of DNA typing in forensic work, *Science*, *254*, 1735–1739.
- Charlesworth, B., and D. Charlesworth (1998), Some evolutionary consequences of deleterious mutations, *Genetica*, *102-103*, 3–19.
- Chun, S., and J. C. Fay (2011), Evidence for hitchhiking of deleterious mutations within the human genome, *PLoS Genetics*, *7*, e1002,240.
- Collin, R. W. J., et al. (2011), High-resolution homozygosity mapping is a powerful tool to detect novel mutations causative of autosomal recessive RP in the Dutch population, *Investigative Ophthalmology and Visual Science*, *52*, 2227–2239.
- Cooper, G. M., and J. Shendure (2011), Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data, *Nature Reviews Genetics*, *12*, 628–640.
- Crow, J. F. (1997), The high spontaneous mutation rate: Is it a health risk?, *Proceedings of the National Academy of Sciences USA*, *94*, 8380–8386.
- Curtis, D., A. E. Vine, and J. Knight (2008), Study of regions of extended homozygosity provides a powerful method to explore haplotype structure of human populations, *Annals of Human Genetics*, *72*, 261–278.
- Darwin, C. (1876), *The effects of cross and self fertilisation in the vegetable kingdom*, John Murray, London, UK.
- DePristo, M. A., et al. (2011), A framework for variation discovery and genotyping using next-generation DNA sequencing data, *Nature Genetics*, *43*, 491–498.
- Derricourt, R. (2005), Getting “Out of Africa”: Sea crossings, land crossings and culture in the Hominin migrations, *Journal of World Prehistory*, *19*, 119–132.
- Evett, I. W., and B. S. Weir (1998), *Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists*, Sinauer Associates, Inc., Sunderland, MA, USA.
- Eyre-Walker, A., and P. D. Keightley (1999), High genomic deleterious mutation rates in hominids, *Nature*, *397*, 344–347.
- Fay, J. C., G. J. Wyckoff, and C. Wu (2001), Positive and negative selection on the human genome, *Genetics*, *158*, 1227–1234.
- Felsenstein, J. (2004), *Inferring Phylogenies*, Sinauer Associates, Inc., Sunderland, MA, USA.

- Field, J. S., M. D. Petraglia, and M. M. Lahr (2007), The southern dispersal hypothesis and the South Asian archaeological record: Examination of dispersal routes through GIS analysis, *Journal of Anthropological Archaeology*, *26*, 88–108.
- Fiumera, A. C., P. G. Parker, and P. A. Fuerst (2000), Effective population size and maintenance of genetic diversity in captive-bred populations of a Lake Victoria cichlid, *Conservation Biology*, *14*, 886–892.
- Friedlaender, J. S., et al. (2008), The genetic structure of Pacific Islanders, *PLoS Genetics*, *4*, e19.
- Garrod, A. E. (1902), The incidence of Alkaptonuria: a study in chemical individuality, *Lancet*, *2*, 1616–1620.
- Gibbs, J. R., and A. Singleton (2006), Application of genome-wide single nucleotide polymorphism typing: simple association and beyond, *PLoS Genetics*, *2*, e150.
- Gibson, J., N. E. Morton, and A. Collins (2006), Extended tracts of homozygosity in outbred human populations, *Human Molecular Genetics*, *15*, 789–795.
- Gillespie, R. (2002), Dating the first Australians, *Radiocarbon*, *44*, 455–472.
- Gradshteyn, I. S., and I. M. Ryzhik (2000), *Table of Integrals, Series, and Products*, 6th ed., Academic Press, London.
- Greenberg, J. H., C. G. Turner II, and S. L. Zegura (1986), The settlement of the Americas: a comparison of the linguistic, dental, and genetic evidence, *Current Anthropology*, *27*, 477–497.
- Hagiwara, K., et al. (2011), Homozygosity mapping on homozygosity haplotype analysis to detect recessive disease-causing genes from a small number of unrelated, outbred patients, *PLoS ONE*, *6*, e25,059.
- Hildebrandt, F., et al. (2009), A systematic approach to mapping recessive disease genes in individuals from outbred populations, *PLoS Genetics*, *5*, e1000,353.
- Hurlbert, S. H. (1971), Nonconcept of species diversity: A critique and alternative parameters, *Ecology*, *4*, 577–586.
- International HapMap Consortium (2007), A second generation human haplotype map of over 3.1 million SNPs, *Nature*, *449*, 851–861.
- Jakkula, E., et al. (2008), The genome-wide patterns of variation expose significant substructure in a founder population, *American Journal of Human Genetics*, *83*, 787 – 794.
- Jefferys, A. J., M. Turner, and P. Debenham (1991), The efficiency of multilocus DNA fingerprint probes for individualization and establishment of family relationships, determined from extensive casework, *American Journal of Human Genetics*, *48*, 824–840.

- Jorde, L. B. (2001), Consanguinity and prereproductive mortality in the Utah Mormon population, *Human Heredity*, 52, 61–65.
- Kalinowski, S. T. (2004), Counting alleles with rarefaction: private alleles and hierarchical sampling designs, *Conservation Genetics*, 5, 539–543.
- Kalinowski, S. T. (2005), HP-RARE 1.0: a computer program for performing rarefaction on measures of allelic richness, *Molecular Ecology Notes*, 5, 187–189.
- Kimmel, M., and R. Chakraborty (1996), Measures of variation at DNA repeat loci under a general stepwise mutation model, *Theoretical Population Biology*, 50, 345–367.
- Kimura, M., and T. Ohta (1969), The average number of generations until extinction of an individual mutant gene in a finite population, *Genetics*, 63, 701–709.
- Kingman, J. (1982), The coalescent, *Stochastic Processes and their Applications*, 13, 235 – 248.
- Kirin, M., R. McQuillan, C. S. Franklin, H. Campbell, P. M. McKeigue, and J. F. Wilson (2010), Genomic runs of homozygosity record population history and consanguinity, *PLoS ONE*, 5, e13,996.
- Kondrashov, A. S. (1995), Contamination of the genome by very slightly deleterious mutations: why have we not died 100 times over?, *Journal of Theoretical Biology*, 175, 583–594.
- Kumar, R., et al. (2010), Genetic ancestry in lung-function predictions, *New England Journal of Medicine*, 363, 321–330.
- Lander, E. S. (2011), Initial impact of the sequencing of the human genome, *Nature*, 470, 187–197.
- Lander, E. S., and D. Botstein (1987), Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children, *Science*, 236, 1567–1570.
- Lencz, T., C. Lambert, P. DeRosse, K. E. Burdick, T. V. Morgan, J. M. Kane, R. Kucherlapati, and A. K. Malhotra (2007), Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia, *Proceedings of the National Academy of Sciences of the USA*, 104, 19,942–19,947.
- Lesecque, Y., P. D. Keightley, and A. Eyre-Walker (2012), A resolution of the mutation load paradox in humans, *Genetics*, 191, 1321–1330.
- Leutenegger, A. L., M. Sahbatou, S. Gazal, H. Cann, and E. Genin (2011), Consanguinity around the world: what do the genomic data of the HGDP-CEPH diversity panel tell us?, *European Journal of Human Genetics*, 19, 583–587.
- Li, H., and R. Durbin (2009), Fast and accurate short read alignment with Burrows-Wheeler transform, *Bioinformatics*, 25, 1754–1760.

- Li, J. Z., et al. (2008), Worldwide human relationships inferred from genome-wide patterns of variation, *Science*, *22*, 1100–1104.
- Li, L.-H., et al. (2006), Long contiguous stretches of homozygosity in the human genome, *Human Mutation*, *27*, 1115–1121.
- Lohmueller, K. E., et al. (2008), Proportionally more deleterious genetic variation in european than in african populations, *Nature*, *451*, 994–998.
- Lohmueller, K. E., et al. (2011), Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome, *PLoS Genetics*, *7*, e1002326.
- MacArthur, D. G., et al. (2012), A systematic survey of loss-of-function variants in human protein-coding genes, *Science*, *335*, 823–828.
- Marth, G. T., et al. (2011), The functional spectrum of low-frequency coding variation, *Genome Biology*, *12*, R84.
- Matisoo-Smith, E. (2007), The peopling of Oceania, in *Anthropological Genetics*, edited by M. Crawford, pp. 408–432, Cambridge University Press, Cambridge, UK.
- McKenna, A., et al. (2010), The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, *Genome Research*, *20*, 1297–1303.
- McQuillan, R., et al. (2008), Runs of homozygosity in European populations, *The American Journal of Human Genetics*, *83*, 359 – 372.
- Morton, N. E., J. F. Crow, and H. J. Muller (1956), An estimate of the mutational damage in man from data on consanguineous marriages, *Proceedings of the National Academy of Sciences of the United States of America*, *42*, 855–863.
- Muller, H. J. (1950), Our load of mutations, *American Journal of Human Genetics*, *2*, 111–176.
- National Research Council of the USA (1992), *DNA Technology in Forensic Science*, National Academy Press, Washington, DC, USA.
- National Research Council of the USA (1996), *The Evaluation of Forensic DNA Evidence*, National Academy Press, Washington, DC, USA.
- Necşulea, A., A. Popa, D. N. Cooper, P. D. Stenson, D. Mouchiroud, C. Gautier, and L. Duret (2011), Meiotic recombination favors the spreading of deleterious mutations in human populations, *Human Mutation*, *32*, 198–206.
- Neel, J. V. (1973), “Private” genetic variants and the frequency of mutation among South American Indians, *Proceedings of the National Academy of Sciences of the United States of America*, *70*, 3311–3315.

- Neel, J. V. (1978), Rare variants, private polymorphisms, and locus heterozygosity in Amerindian populations, *American Journal of Human Genetics*, *30*, 465–490.
- Neel, J. V., and E. A. Thompson (1978), Founder effect and number of private polymorphisms observed in Amerindian tribes, *Proceedings of the National Academy of Sciences of the United States of America*, *75*, 1904–1908.
- Neel, M. C., and M. P. Cummings (2003), Effectiveness of conservation targets in capturing genetic diversity, *Conservation Biology*, *17*, 219–229.
- O’Connell, J. F., and J. Allen (2004), Dating the colonization of Sahul (Pleistocene Australia-New Guinea): A review of recent research, *Journal of Archaeological Science*, *31*, 835–853.
- Parker, K. M., R. J. Sheffer, and P. W. Hedrick (1999), Molecular variation and evolutionarily significant units in the endangered Gila topminnow, *Conservation Biology*, *13*, 108–116.
- Pemberton, T. J., D. Absher, M. W. Feldman, R. M. Myers, N. A. Rosenberg, and J. Z. Li (2012), Genomic patterns of homozygosity in worldwide human populations, *American Journal of Human Genetics*, *91*, 275–292.
- Petit, R., A. E. Mousadik, and O. Pons (1998), Identifying populations for conservation on the basis of genetic markers, *Conservation Biology*, *12*, 844–855.
- Picard (2011), Picard Tools, <http://picard.sourceforge.net/>.
- Pritchard, J. K., and M. W. Feldman (1996), Statistics for microsatellite variation based on coalescence, *Theoretical Population Biology*, *50*, 325–344.
- Pruitt, K. D., et al. (2009), The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes, *Genome Research*, *19*, 1316–1323.
- Quintana-Murci, L., O. Semino, H. J. Bandelt, G. Passarino, K. McElreavy, and A. S. Santachiara-Benerecetti (1999), Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa, *Nature Genetics*, *23*, 437–441.
- Ramachandran, S., O. Deshpande, C. C. Roseman, N. A. Rosenberg, M. W. Feldman, and L. L. Cavalli-Sforza (2005), Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa, *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 15,942–15,947.
- Reiner, A. P., et al. (2005), Population structure, admixture, and aging-related phenotypes in African American adults: The cardiovascular health study, *American Journal of Human Genetics*, *76*, 463–477.

- Rosenberg, N. A. (2006), Standardized subsets of the HGDP-CEPH Human Genome Diversity Panel, accounting for atypical and duplicated samples and pairs of close relatives, *Annals of Human Genetics*, *70*, 841–847.
- Rosenberg, N. A., J. K. Pritchard, J. L. Weber, H. M. Cann, K. K. Kidd, L. A. Zhivotovsky, and M. W. Feldman (2002), Genetic structure of human populations, *Science*, *298*, 2381–2385.
- Rosenberg, N. A., S. Mahajan, S. Ramachandran, C. Zhao, J. K. Pritchard, and M. W. Feldman (2005), Clines, clusters, and the effect of study design on the inference of human population structure, *PLoS Genetics*, *1*, 660–671.
- Rosenberg, N. A., L. Huang, E. M. Jewett, Z. A. Szpiech, I. Jankovic, and M. Boehnke (2010), Genome-wide association studies in diverse populations, *Nature Reviews Genetics*, *11*, 356–366.
- Schroeder, K. B., T. G. Schurr, J. C. Long, N. A. Rosenberg, M. H. Crawford, L. A. Tarskaia, L. P. Osipova, S. I. Zhadanov, and D. G. Smith (2007), A private allele ubiquitous in the Americas, *Biology Letters*, *3*, 218–223.
- Schuurs-Hoeijmakers, J. H. M., et al. (2011), Homozygosity mapping in outbred families with mental retardation, *European Journal of Human Genetics*, *19*, 597–601.
- Slatkin, M. (1985), Rare alleles as indicators of gene flow, *Evolution*, *39*, 53–65.
- Sunyaev, S., V. Ramensky, I. Koch, W. Lathe III, A. S. Kondrashov, and P. Bork (2001), Prediction of deleterious human alleles, *Human Molecular Genetics*, *10*, 591–597.
- Szpiech, Z. A., and N. A. Rosenberg (2011), On the size distribution of private microsatellite alleles, *Theoretical Population Biology*, *80*, 100–113.
- Szpiech, Z. A., M. Jakobsson, and N. A. Rosenberg (2008), ADZE: a rarefaction approach for counting alleles private to combinations of populations, *Bioinformatics*, *24*, 2498–2504.
- Tennessen, J. A., et al. (2012), Evolution and functional impact of rare coding variation from deep sequencing of human exomes, *Science*, *337*, 64–69.
- Thompson, E. A., and J. V. Neel (1978), Probability of founder effect in a tribal population, *Proceedings of the National Academy of Sciences of the USA*, *75*, 1442–1445.
- Thorne, A., R. Grun, G. Mortimer, N. A. Spooner, J. J. Simpson, M. McCulloch, L. Taylor, and D. Curnoe (1999), Australia’s oldest human remains: Age of the Lake Mungo 3 skeleton, *Journal of Human Evolution*, *36*, 591–612.

- Tishkoff, S. A., and K. K. Kidd (2004), Implications of biogeography of human populations for 'race' and medicine, *Nature Genetics*, *36*, S21–S27.
- Torres, E., J. M. Iriondo, and C. Pérez (2003), Genetic structure of an endangered plant, *Antirrhinum microphyllum* (Scrophulariaceae): allozyme and RAPD analysis, *American Journal of Botany*, *90*, 85–92.
- Wakeley, J. (2009), *Coalescent Theory: An Introduction*, Roberts and Company Publishers, Greenwood Village, CO, USA.
- Wang, S., et al. (2007), Genetic variation and population structure in Native Americans, *PLoS Genetics*, *3*, e185.
- Watkins, J. C. (2007), Microsatellite evolution: Markov transition functions for a suite of models, *Theoretical Population Biology*, *71*, 147–159.
- Wehrhahn, C. F. (1975), The evolution of selectively similar electrophoretically detectable alleles in finite natural populations, *Genetics*, *80*, 375–394.
- Weir, B. S. (1996), *Genetic Data Analysis II*, Sinauer, Sunderland, Massachusetts.
- Whittaker, J. C., R. M. Harbord, N. Boxall, I. Mackay, G. Dawson, and R. M. Sibly (2003), Likelihood-based estimation of microsatellite mutation rates, *Genetics*, *164*, 781–787.
- Wilson, I. J., and D. J. Balding (1998), Genealogical inference from microsatellite data, *Genetics*, *150*, 499–510.
- Zhang, K., and N. A. Rosenberg (2007), On the genealogy of a duplicated microsatellite, *Genetics*, *177*, 2109–2122.
- Zhivotovsky, L. A., N. A. Rosenberg, and M. W. Feldman (2003), Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers, *American Journal of Human Genetics*, *72*, 1171–1186.