THE EFFECTS OF RELATEDNESS AND SEX-BIASED DEMOGRAPHIC PROCESSES ON HUMAN GENETIC VARIATION

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF GENETICS
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Daniel Juetten Cotter

August 2023

This dissertation is online at: https://purl.stanford.edu/qb897xw6209

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Noah Rosenberg, Primary Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Stephen Montgomery

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Jonathan Pritchard

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Hua Tang

Approved for the Stanford University Committee on Graduate Studies.

Stacey F. Bent, Vice Provost for Graduate Education

This signature page was generated electronically upon submission of this dissertation in electronic format.

Abstract

Many evolutionary processes affect human genetic variation. One of these processes, consanguinity—mating between closely related individuals—increases the frequency at which identical genomic segments are inherited along separate paths of descent. This pairing of segments increases the fraction of the genome that is shared within or between individuals: the fraction that lies in runs of homozygosity (ROH) or that contains identical-by-descent (IBD) segments. Consanguinity is relatively common globally, with couples who are second cousins (or closer) as well as their offspring perhaps representing an estimated 10% of the human population. The types and degree of consanguinity vary widely across cultures and affect patterns of shared segments. This variation provides an opportunity to study how a cultural practice such as consanguinity can influence patterns of genetic variation that are observed in the genome.

Genomic sharing, both within and between individuals in a population, can be studied by noting that ROH and IBD at a genomic site are inversely proportional to its coalescence time: the time at which a pair of copies of the site find a common ancestor. First-cousin consanguinity can take one of four forms differing in the configuration of sexes in the pedigree of the male and female cousins who join in a consanguineous union: patrilateral parallel, patrilateral cross, matrilateral parallel, and matrilateral cross. Because of the different configurations of sexes in the pedigree, these four types of first-cousin consanguinity, which are equivalent in their effects on the autosomes, are expected to have differing effects on coalescence times—and therefore ROH and IBD—on the X chromosome. Over several chapters, this dissertation models each of these types of consanguinity to study the effect that each has on X-chromosomal genomic sharing relative to autosomal genomic sharing.

Chapter 1 explores the effect that consanguinity has on mean coalescence times. It demonstrates that the effect of consanguinity on the X chromosome differs from the autosomal pattern under matrilateral but not under patrilateral first-cousin mating: for matrilateral first cousins, the effect of consanguinity in reducing coalescence times is stronger on the X chromosome than on the autosomes. Chapter 2 solves for the limiting distribution of coalescence times under each type of consanguinity as well as in an arbitrary mixture of all four types. This chapter shows that for between-individual coalescence times, each of the first-cousin consanguinity types has a different coalescent effective size (given in terms of the consanguinity rate in a population). Chapter 3 then develops a theoretical model for the relationship between pairwise coalescence times on the X-chromosome and the expected fraction of the X that will be tiled by IBD or ROH, allowing for the calculation of a mathematical expectation of the X-to-autosomal ratio of ROH or IBD under first-cousin consanguinity. By comparing empirical IBD and ROH ratios in populations with known rates of consanguinity to these theoretical expectations, this chapter demonstrates the utility of the model in understanding the role of consanguinity in shaping ratios of genomic sharing between the X chromosome and the autosomes. Finally, chapter 4 shifts to explore the effect that sample size has on observed patterns of global allelic variation. It introduces a sample-size correction for comparing rare and common variation across multiple populations when the sample sizes from these populations differ; this method reveals subtle distinctions that are often not observed in analyses that use the full available sample sizes. In total, these chapters use a mixture of theoretical coalescent models and empirical data to explore the important role that demographic phenomena such as consanguinity have in shaping the genome, and they introduce new methods for refining our understanding of human genetic variation.

Acknowledgments

I would like to express my sincere gratitude to all those who have contributed to the completion of my degree. Foremost, I extend my deepest appreciation to my advisor, Noah Rosenberg, whose unwavering support and invaluable guidance have been instrumental in shaping me into the scientist I am today. I am also indebted to the entire Rosenberg lab for their contributions to my growth and development throughout my PhD journey.

Specifically, I wish to acknowledge the postdoctoral members of the lab—Gili Greenbaum, Airam Blancas, and Lily Tamir—for their scientific knowledge and mentorship, which played a pivotal role in shaping the direction of my research. Jaehee Kim, with her warm welcome and scientific advice, has been a true inspiration, even encouraging me to stay fit by joining workout classes on weekends. Although we were virtually connected during her postdoc, Jazlyn Mooney has become not only a colleague but also a dear friend.

I owe a debt of gratitude to my fellow PhD students, Xiran Liu and Kaleda Denton, whose unwavering support and advice have been indispensable in helping each other reach the finish line. Special appreciation goes to Alissa Severson, whose mentorship was a transformative experience, leading me to explore the more mathematical aspects of human genetics.

I extend my thanks to all other members of the lab—Egor, Maike, Chloe, Juan Estaban, Susan, Jonthan Kang, and all the undergraduate trainees—for their contributions that have enriched my time at Stanford and in the lab.

I am immensely thankful for the scientific advice and guidance I received from my dissertation committee members: Jonathan Pritchard, Stephen Montgomery, and Hua Tang. Their expertise and insights have been invaluable.

To my friends and colleagues in the Genetics Department at Stanford, you have been a constant source of support and camaraderie during the challenging times of graduate school. Special thanks go to Roshni Patel, Katie Hanson, and Dylan Maghini (and her cat Juniper) for being there through the ups and downs of the last couple of years. I also hold dear all of the other friendships I've cultivated during my time at Stanford, particularly those with Emily Greenwald and Lauren Pope.

Last but not least, I want to express my deepest appreciation to my family and friends. My partner Matt has been a steadfast pillar of support throughout this journey, and I am truly grateful for his patience and understanding. My mother has been my guiding light from the very beginning, and I owe my presence here today to her unwavering love and support. My father has always been a rock for me, helping me make important decisions and shaping my path. I am also thankful to my stepmom and stepdad for their constant support throughout this process. My twin, Austin, has been a key support system, and I am deeply grateful for our bond.

Without the contributions and support of all these remarkable individuals, this degree would not have been possible. I am immeasurably grateful to each and every one of them.

Contents

Tables

Figures

Introduction

There are many evolutionary and demographic processes that shape human genetic variation: population size, migration, assortative mating, and sex-biases, to name a few. The field of population genetics concerns itself with answering questions about these processes using genetic data collected from many individuals in a population. By studying features of genetic variation, such as allele frequencies or homozygosity, scientists can make inferences about the strength and effect of these demographic processes.

One set of demographic processes are those that concern sex-biases. These are processes that are shaped by sex differences between features of population history such as the number of mating males and females, rates of male and female migration, and male and female mutation and recombination rates. These processes are especially interesting because they differ in the signature they leave on genetic variation on the sex-linked chromosomes (i.e. the X chromosome, the Y chromosome, and the mitochondria) relative to the autosomes. One sex-specific process, which I will explore herein, is consanguinity, or matings between closely-related individuals.

Consanguinity has historically been common across human populations. One study estimates that couples related at a level of second-cousins or closer, as well as their offspring, could represent up to 10% of the global human population (Bittles and Black, 2010). Different types of consanguinity occur at differing rates across populations, often as a result of various social social structures and cultural norms (Bittles, 2012). Specifically, first-cousin consanguinity can occur in one of four configurations depending on the sexes in the pedigree of the male-female consanguineous pair. These four types are (1) patrilateral parallel, where a man marries his father's brother's daughter; (2) patrilateral cross, where a man marries his father's sister's daughter; (3) matrilateral parallel,

where a man marries his mother's sister's daughter; and (4) matrilateral cross, where a man marries his mother's brother's daughter. These four types differ in the pathway through which the X chromosome is inherited, each having a particular effect on genetic variation on the X chromosome relative to the autosomes.

The inheritance of a segment of the genome through both parents in a consanguineous union affects the variability of runs of homozygosity (ROH)—long segments of homozygous genotypes within a single individual—as well identical-by-descent (IBD) segments—segments of genomes that are shared in long runs between two individuals. Because both types of genomic sharing are inherently affected by the path that a lineage takes through a pedigree, the four different types of first-cousin consanguinity affect segments on the X-chromosome in several different ways. Past work on consanguinity has shown that ROH and IBD on the autosomes are inherently affected by these matings of closely related individuals (Severson *et al.*, 2019, 2021). This thesis demonstrates how consanguinity—specifically sex-biased forms of consanguinity—affects X-chromosomal genomic sharing.

In chapter 1, I study the effects that consanguinity has on *mean* coalescence times across the genome. By noting that the lengths of ROH and IBD at a site are inversely proportional to the number of generations since the most recent ancestor of two genetic lineages, we study this time to the most recent common ancestor (T_{MRCA}) directly instead of segment lengths. Using a first-step analysis on a coalescent model of each type of consanguinity, I demonstrate that there is an effect of matrilateral parallel and matrilateral-cross but not patrilateral consanguinity on shaping genomic sharing on the X chromosome. I also demonstrate that there is a stronger effect of matrilateral first cousin consanguinity in reducing the T_{MRCA} on the X chromosome relative to the autosomal genome.

Next, in chapter 2, I expand beyond this first-step analysis to calculate a full distribution of coalescence times for each type of first-cousin consanguinity as well as in a model that takes into account a mixture of all four types of first-cousin consanguinity. This chapter uses a separation-of-time scales approach to separately consider "fast" events that occur quickly in a coalescent process as well as "slow" events that take place over a much longer time period. By separating

these events from each other and analyzing the resulting Markov chains, I solve for the limiting-distributions for each type of consanguinity in their large-population limits. This computation demonstrates that each type of first-cousin consanguinity, in its large-population-size limit, is a standard coalescent process governed by a coalescent effective size that is in turn governed by the consanguinity rate.

Chapter 3 then uses theory on coalescent times and genomic sharing to establish a formal relationship between the coalescent models developed in Chapters 1 and 2 and the average length of ROH and IBD in a given population. In this chapter, I develop theory about the null ratio of IBD and ROH sharing on the X chromosome relative to the autosomes as well as the effect that each type of consanguinity has on shifting this ratio away from its null value. I then explore empirically calculated proportions of IBD and ROH on the X chromosome and the autosomes in a set of populations for which there exists historical data on rates of each type of first-cousin consanguinity.

Finally, in chapter 4, I explore a distinct but related topic in the study of population-genetic variation. Here, I discuss an approach to quantifying differences in allele frequencies across populations when sample sizes differ in magnitude. This chapter builds on a long history of work that explores the extent to which genetic variation differs within and between populations. I develop a rarefaction-based sample-size correction that allows for accurate comparison of differences in rare and common genetic variants across populations. By using a global genetic dataset, I show that correcting for sample-size differences can reveal subtle signals that are not apparent when sample sizes differ greatly or when a very-low frequency threshold is used to define a rare variant.

Collectively this body of work uses several theoretical and empirical techniques to explore the role of demographic phenomena in shaping human genetic variation. Chapters 1–3 offer insights into the effect of sex-biased consanguinity on runs of homozygosity and identity by descent. Recently, there has been renewed interest in long runs of homozygosity due to their connection to Mendelian (Bittles, 2001; Woods *et al.*, 2006) and complex (Bittles and Black, 2010; Ceballos *et al.*, 2018; Clark *et al.*, 2019) disease risk. By providing new insights into small populations and how specific types of relatedness affect these runs of homozygosity, this dissertation contributes a novel way of conceptualizing the connection between human demography—specifically, consanguinity

and population size—and rare disease. Similarly, chapter 4 offers a new perspective on the technical consequences of comparing genetic variation across groups when sample-sizes differ. Studies of ancient DNA often wish to make comparisons between ancient and modern populations (Witt *et al.*, 2022). Because sample sizes can potentially differ by many orders of magnitude in these comparisons, it is important to correct for the effect that these differences will have on their respective populations' allele frequencies. Similarly, outside of human genetics, conservation geneticists often study allele-frequency variation in endangered species for the purpose of actively intervening in the genetic diversity of those species. Tools such as the one I present in chapter 4 can correct for size differences in allele-frequency comparisons and help ensure that any interventions made are as informed and accurate as possible.

Human genetic variation represents an important element in our understanding of the history and demography of our species. By leveraging theoretical, statistical, and empirical approaches, this dissertation contributes to advancing this understanding of the forces that shape human evolutionary history.

Chapter 1

The effect of consanguinity on mean coalescence times on the X chromosome

The following chapter and figures were originally published as:

Cotter, D. J., A. L. Severson, and N. A. Rosenberg, 2021 The effect of consanguinity on coalescence times on the X chromosome. Theoretical Population Biology **140**: 32–43.

https://doi.org/10.1016/j.tpb.2021.03.004

Abstract

Consanguineous unions increase the frequency at which identical genomic segments are inherited along separate paths of descent, decreasing coalescence times for pairs of alleles drawn from an individual who is the offspring of a consanguineous pair. For an autosomal locus, it has recently been shown that the mean time to the most recent common ancestor (T_{MRCA}) for two alleles in the same individual and the mean T_{MRCA} for two alleles in two *separate* individuals both decrease with increasing consanguinity in a population. Here, we extend this analysis to the X chromosome, considering X-chromosomal coalescence times under a coalescent model with diploid, male–female

mating pairs. We examine four possible first-cousin mating schemes that are equivalent in their effects on autosomes, but that have differing effects on the X chromosome: patrilateral-parallel, patrilateral-cross, matrilateral-parallel, and matrilateral-cross. In each mating model, we calculate mean T_{MRCA} for X-chromosomal alleles sampled either within or between individuals. We describe a consanguinity effect on X-chromosomal T_{MRCA} that differs from the autosomal pattern under matrilateral but not under patrilateral first-cousin mating. For matrilateral first cousins, the effect of consanguinity in reducing T_{MRCA} is stronger on the X chromosome than on the autosomes, with an increased effect of parallel-cousin mating compared to cross-cousin mating. The theoretical computations support the utility of the model in understanding patterns of genomic sharing on the X chromosome.

1.1 Introduction

In consanguineous unions, parents are closely related, producing offspring who inherit identical genomic segments along both parental lines. Consanguinity is common in human populations; couples related at a level of second cousins or closer, together with their offspring, might represent as much as 10% of the global population (Bittles and Black, 2010).

Different forms of consanguinity vary in frequency across human populations, often as a result of cultural norms concerning preferred mate choices (Bittles, 2012). For example, first-cousin marriages can follow four distinct patterns, as classified by the sexes of the two sibling parents. These patterns are named from the perspective of the male in the consanguineous marriage: (1) patrilateral-parallel, a male marries his father's brother's daughter; (2) patrilateral-cross, he marries his father's sister's daughter; (3) matrilateral-parallel, he marries his mother's sister's daughter; and (4) matrilateral-cross, he marries his mother's brother's daughter. The "parallel" marriage patterns refer to same-sex sibling parents and the "cross" patterns refer to opposite-sex sibling parents.

The inheritance of a genomic segment via both parents due to consanguineous unions has consequences for genomic phenomena such as runs of homozygosity (ROH)—long homozygous segments in diploid individuals—and sharing between individuals of long identical-by-descent (IBD)

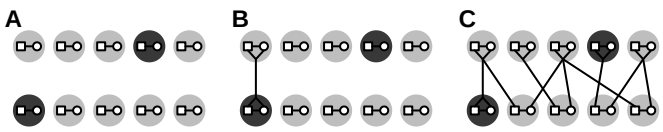

Figure 1.1: A diploid mating model for two successive generations. **(A)** Each generation has N monogamous mating pairs, a fraction c_0 of which are sib mating pairs, shaded ($N = 5$, $c_0 = 0.2$). **(B)** The sibs in each sib mating pair in the offspring generation (shaded pair in the offspring generation) share a parental pair from the previous generation. **(C)** Non-sib mating pairs are associated with two mating pairs from the previous generation. Note that a pair of parents chosen in the parental generation might itself be a sib pair (shaded pair in the parental generation). Small squares, males; small circles, females; large circles, mating pairs. The figure is modified from Severson *et al.* (2019).

segments. In particular, it is well-known that due to pairing of IBD segments within individuals, consanguinity increases the frequency and length of ROH (e.g. McQuillan *et al.*, 2008; Kirin *et al.*, 2010; Pemberton *et al.*, 2012; Kang *et al.*, 2016; Ceballos *et al.*, 2018).

Recently, we have argued that consanguinity also increases the frequency and length of IBD segments shared *between* pairs of individuals (Severson *et al.*, 2019). In a model of coalescence in a diploid population, the mean time to the most recent common ancestor (T_{MRCA}) for a pair of *alleles* in separate individuals was seen to decrease with increasing consanguinity. Because genomic sharing increases with decreasing T_{MRCA}, the reduction in T_{MRCA} in turn increases genomic sharing. The interpretation of Severson *et al.* (2019), established in further detail by Severson *et al.* (2021), is that consanguinity decreases effective population size, increasing genomic sharing both for within-individual ROH and for between-individual IBD segments.

When studying the autosomal genome, the four forms of first-cousin mating—patrilateral-parallel, patrilateral-cross, matrilateral-parallel, and matrilateral-cross—are indistinguishable in their effects, as males and females each contribute a copy of the autosomal genome in each generation. For the X chromosome, however, owing to the fact that the X chromosome is not transmitted from fathers to sons, the consanguineous pairs in these four types of unions have different levels of relatedness (Jacquard, 1974; Lange, 2002). Hence, populations with different forms of first-cousin mating potentially have different patterns for T_{MRCA}, ROH, and IBD sharing.

Here, we extend our autosomal model of monogamous diploid mating pairs to consider X-chromosomal inheritance. We focus our analysis on the four types of first-cousin mating. The results reveal effects of different types of consanguinity on autosomal and X-chromosomal T_{MRCA}.

1.2 Autosomal results

We extend the framework from Severson *et al.* (2019), which considers a diploid population of N monogamous mating pairs (Figure 1.1). This framework, which allows for a variety of forms of consanguinity, in turn generalizes a sib mating model from Campbell (2015).

We build on Severson *et al.* (2019) by explicitly including the sex of the individuals. For autosomal loci, this change adds states to our Markov chain, states that are needed for our computations with the X chromosome. The addition of the three new states produces results that reduce to those of Severson *et al.* (2019). We review these results, and we then examine the effect of consanguinity on X-chromosomal coalescence times.

The models of Campbell (2015) and Severson *et al.* (2019) consider alleles in three possible states: within an individual, in two individuals in a mating pair, and in two individuals in separate mating pairs. In incorporating sex, we decompose the three states into six (Figure 1.2). Enumerating these possible states, state 1 is two alleles in a male and state 2 is two alleles in a female. State 3 is two alleles in two different individuals in the same mating pair. States 4, 5, and 6 describe two alleles in two individuals in different mating pairs, where the two individuals are two males, a male and a female, and two females, respectively.

We define six random variables to denote coalescence times of two alleles in our six states (Figure 1.2). T_m is T_{MRCA} for two alleles in the same male (state 1) and T_f is T_{MRCA} for two alleles in

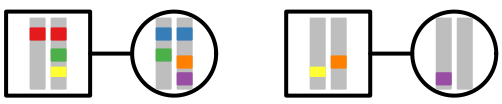

Figure 1.2: Six possible states for two sampled alleles at autosomal loci. Males are squares; females are circles. State 1: within a male (red). State 2: within a female (blue). State 3: in two individuals in a mating pair (green). State 4: in two males in different mating pairs (yellow). State 5: in a male and a female in different mating pairs (orange). State 6: in two females in different mating pairs (purple).

the same female (state 2). As in Severson *et al.* (2019), U is T_{MRCA} for two alleles in two individuals in the same mating pair (state 3). Finally, V_{mm} is T_{MRCA} for two males in separate mating pairs (state 4), V_{mf} is T_{MRCA} for a male and a female in separate mating pairs (state 5), and V_{ff} is T_{MRCA} for two females in separate mating pairs (state 6). We calculate the means $\mathbb{E}[T_m]$, $\mathbb{E}[T_f]$, $\mathbb{E}[U]$, $\mathbb{E}[V_{mm}]$, $\mathbb{E}[V_{mf}]$, and $\mathbb{E}[V_{ff}]$ for autosomal and X-chromosomal loci under sib and first-cousin mating. States 1 and 2 are "lumped" into one state by Severson *et al.* (2019), as are states 4, 5, and 6.

1.2.1 Siblings

We follow Campbell (2015) and Severson *et al.* (2019) to derive equations for mean coalescence times. In each generation, a constant fraction c_0 of the N mating pairs are sib mating pairs. Chance consanguinity is excluded, so that the remaining $N(1 - c_0)$ pairs are not sib mating pairs.

Beginning with T_m and T_f, two alleles within the same male (state 1) or female (state 2) must have been inherited from a single mating pair (state 3) one generation back. We have

$$\mathbb{E}[T_m] = \mathbb{E}[T_f] = \mathbb{E}[U] + 1. \tag{1.1}$$

For two alleles in separate individuals in a mating pair (state 3), the mating pair is a sib mating pair with probability c_0. Conditional on the individuals in the mating pair being sibs, the sampled alleles can be inherited from the previous generation in four ways. First, they could coalesce in the previous generation, an event that occurs with probability $\frac{1}{4}$. Next, they could both be inherited from the father and not coalesce, an event that has probability $\frac{1}{8}$, giving coalescence time $\mathbb{E}[T_m]+1$. Similarly, they could both be inherited from the mother and not coalesce, with probability $\frac{1}{8}$ and coalescence time $\mathbb{E}[T_f] + 1$. Finally, they could be inherited from two individuals in a mating pair in the previous generation, with probability $\frac{1}{2}$ and coalescence time $\mathbb{E}[U] + 1$.

The event that the mating pair is not a sib mating pair has probability $1 - c_0$. Conditional on this event, the alleles have three possibilities. They arise from two males in separate mating pairs with probability $\frac{1}{4}$, giving coalescence time $\mathbb{E}[V_{mm}] + 1$. They are inherited from a male and a female in separate mating pairs with probability $\frac{1}{2}$, giving coalescence time $\mathbb{E}[V_{mf}] + 1$.

Lastly, with probability $\frac{1}{4}$, they are inherited from two females in separate mating pairs, giving coalescence time $\mathbb{E}[V_{ff}] + 1$. Combining the cases, we have

$$\mathbb{E}[U] = c_0 \left[\frac{1}{4} + \frac{1}{8}(\mathbb{E}[T_m] + 1) + \frac{1}{8}(\mathbb{E}[T_f] + 1) + \frac{1}{2}(\mathbb{E}[U] + 1)\right] + (1 - c_0) \left[\frac{1}{4}(\mathbb{E}[V_{mm}] + 1) + \frac{1}{2}(\mathbb{E}[V_{mf}] + 1) + \frac{1}{4}(\mathbb{E}[V_{ff}] + 1)\right]. \tag{1.2}$$

In considering autosomal loci, coalescence patterns for alleles in two individuals in different mating pairs are the same irrespective of the sexes of the individuals. Hence, equations for $\mathbb{E}[V_{mm}]$, $\mathbb{E}[V_{mf}]$, and $\mathbb{E}[V_{ff}]$ are the same. Because we draw a parental pair uniformly with replacement for each individual in the current generation, the probability that, by chance, two individuals in separate mating pairs are siblings is $\frac{1}{N}$. Conditional on the event that we randomly sample two siblings, as in Eq. 1.2, the alleles have four possibilities: they coalesce, they come from the father in the previous generation and do not coalesce, they come from the mother in the previous generation and do not coalesce, or they come from separate individuals in the previous generation. The probabilities and coalescence times for these events follow from sib mating in Eq. 1.2.

The event that the two individuals are not siblings occurs with probability $1 - \frac{1}{N}$. Three events are then possible: the two alleles come from two males in separate mating pairs, from a male and a female, or from two females. The probabilities and coalescence times for these events follow from non-sib mating in Eq. 1.2. We have

$$\mathbb{E}[V_{mm}] = \mathbb{E}[V_{mf}] = \mathbb{E}[V_{ff}] = \frac{1}{N}\left[\frac{1}{4} + \frac{1}{8}(\mathbb{E}[T_m] + 1) + \frac{1}{8}(\mathbb{E}[T_f] + 1) + \frac{1}{2}(\mathbb{E}[U] + 1)\right] + \left(1 - \frac{1}{N}\right)\left[\frac{1}{4}(\mathbb{E}[V_{mm}] + 1) + \frac{1}{2}(\mathbb{E}[V_{mf}] + 1) + \frac{1}{4}(\mathbb{E}[V_{ff}] + 1)\right]. \tag{1.3}$$

Eqs. 1.1–1.3 form a linear system of equations in six variables, with solution

$$\mathbb{E}[T_m] = \mathbb{E}[T_f] = 4N(1 - c_0) + 6 \tag{1.4}$$

$$\mathbb{E}[U] = 4N(1 - c_0) + 5 \tag{1.5}$$

$$\mathbb{E}[V_{mm}] = \mathbb{E}[V_{mf}] = \mathbb{E}[V_{ff}] = 4N\left(1 - \frac{3}{4}c_0\right) + 4. \tag{1.6}$$

We reduce the model to a model without sexes by collapsing states 1 and 2 into a single state for two alleles from a single individual (male or female), and by collapsing states 4, 5, and 6 into a single state for two alleles from individuals of unspecified sex in different mating pairs.
Using T to represent T_{MRCA} for two alleles sampled in the combined states 1 and 2, Eq. 1.4 gives

$$\mathbb{E}[T] = 4N(1 - c_0) + 6. \tag{1.7}$$

We can use V to represent T_{MRCA} for the collapsed state for states 4, 5, and 6. Eq. 1.6 can be collapsed into

$$\mathbb{E}[V] = 4N\left(1 - \frac{3}{4}c_0\right) + 4. \tag{1.8}$$

Eqs. 1.7, 1.5, and 1.8 are equivalent to Eqs. 4, 5, and 6, respectively, of Severson *et al.* (2019). Breaking the three states of Severson *et al.* (2019) into six by consideration of the sexes of individuals provides further detail, but it does not change the quantitative results.

1.2.2 First cousins

Severson *et al.* (2019) considered first-cousin mating, assuming a constant fraction c_1 in each generation for the fraction of first-cousin mating pairs in a population, and again assuming that no chance consanguinity occurs. Rephrasing their results for mean autosomal coalescence times with a decomposition of states by sex, the derivation otherwise follows Severson *et al.* (2019), and the quantitative results are the same. The four types of cousin mating give the same recursions. For completeness, we include the derivation in Appendix A. We obtain

$$\mathbb{E}[T_m] = \mathbb{E}[T_f] = 4N\left(1 - \frac{1}{4}c_1\right) + 10 \tag{1.9}$$

$$\mathbb{E}[U] = 4N\left(1 - \frac{1}{4}c_1\right) + 9 \tag{1.10}$$

$$\mathbb{E}[V_{mm}] = \mathbb{E}[V_{mf}] = \mathbb{E}[V_{ff}] = 4N\left(1 - \frac{3}{16}c_1\right) + 7. \tag{1.11}$$

Severson *et al.* (2019) noted that $\mathbb{E}[V_{mm}] = \mathbb{E}[V_{mf}] = \mathbb{E}[V_{ff}]$ exceeds $\mathbb{E}[T_m] = \mathbb{E}[T_f]$ for nonzero c_1 and sufficiently large N. They also observed that in comparison with a non-consanguineous diploid population of size $2N$, the mean coalescence times are reduced by linear factors, $1 - \frac{1}{4}c_1$ in Eqs. 1.9 and 1.10, and $1 - \frac{3}{16}c_1$ in Eq. 1.11.

1.2.3 Double first cousins

We can similarly examine double-first-cousin mating, a case not considered by Severson *et al.* (2019). We use the six states, with a fraction c_1 of double-first-cousin mating pairs in each generation. The two types of bilateral cousin mating give the same recursions. The derivation appears in Appendix B. We obtain

$$\mathbb{E}[T_m] = \mathbb{E}[T_f] = 4N\left(1 - \frac{1}{2}c_1\right) + 10 \tag{1.12}$$

$$\mathbb{E}[U] = 4N\left(1 - \frac{1}{2}c_1\right) + 9 \tag{1.13}$$

$$\mathbb{E}[V_{mm}] = \mathbb{E}[V_{mf}] = \mathbb{E}[V_{ff}] = 4N\left(1 - \frac{3}{8}c_1\right) + 7. \tag{1.14}$$

The reduction factors due to consanguinity in comparison with a non-consanguineous population are $1 - \frac{1}{2}c_1$ for Eqs. 1.12 and 1.13, and $1 - \frac{3}{8}c_1$ for Eq. 1.14.

1.3 X-chromosomal results

We now derive new results for the X chromosome, comparing them to autosomal results. As before, we consider N mating pairs in each generation; the number of alleles in the population in each generation is $3N$, $2N$ in females and N in males. Because two sampled X-chromosomal alleles cannot be in the same male, we remove state 1.

1.3.1 Siblings

For sib mating, as before, let c_0 be the fraction of sib mating pairs each generation. Chance sib mating is forbidden. In the event that two X chromosomes are present within a female, in the

previous generation they must have been in a mating pair, and the mean coalescence time, $\mathbb{E}[T_f]$, remains the same as for autosomes:

$$\mathbb{E}[T_f] = \mathbb{E}[U] + 1. \tag{1.15}$$

For two X chromosomes within a mating pair (state 3), the individuals are siblings with probability c_0. Given that they are siblings, three possibilities exist for how two sampled alleles were inherited. First, with probability $\frac{1}{4}$, the alleles coalesce in the mother in the previous generation. Second, with probability $\frac{1}{4}$, they trace to the mother and do not coalesce, giving mean coalescence time $\mathbb{E}[T_f]+1$. Third, with probability $\frac{1}{2}$, they derive separately from the mother and father, giving mean coalescence time $\mathbb{E}[U] + 1$.

Suppose now that the individuals in the mating pair are not siblings, an event with probability $1 - c_0$. The X chromosome in the male is inherited from his mother, and the X chromosome in the female has probability $\frac{1}{2}$ of being inherited from her mother and probability $\frac{1}{2}$ of being inherited from her father. These events give coalescence times $\mathbb{E}[V_{ff}] + 1$ and $\mathbb{E}[V_{mf}] + 1$, respectively. Combining cases, we obtain

$$\begin{aligned}\mathbb{E}[U] = c_0 &\left[\frac{1}{4} + \frac{1}{4}(\mathbb{E}[T_f] + 1) + \frac{1}{2}(\mathbb{E}[U] + 1)\right] \\ &+ (1 - c_0)\left[\frac{1}{2}(\mathbb{E}[V_{mf}] + 1) + \frac{1}{2}(\mathbb{E}[V_{ff}] + 1)\right].\end{aligned} \tag{1.16}$$

Because X chromosomes are not inherited from father to son, states 4–6 no longer produce identical recursions, as they did for autosomes. In all three states, the probability continues to be $\frac{1}{N}$ that two alleles in separate individuals in separate mating pairs are in siblings. For V_{mm}, given that the two males are siblings, with probability $\frac{1}{2}$, the two alleles coalesce in one generation (in their mother). With probability $\frac{1}{2}$, they derive from the mother and do not coalesce, giving coalescence time $\mathbb{E}[T_f]+1$. If the two males are not siblings, an event with probability $1 - \frac{1}{N}$, then the X chromosomes trace to non-sib mothers, and the expected coalescence time is $\mathbb{E}[V_{ff}] + 1$.

Combining these terms, we have

$$\mathbb{E}[V_{mm}] = \frac{1}{N}\left[\frac{1}{2} + \frac{1}{2}(\mathbb{E}[T_f] + 1)\right] + \left(1 - \frac{1}{N}\right)(\mathbb{E}[V_{ff}] + 1). \tag{1.17}$$

For V_{mf}, if the male and female in separate mating pairs are siblings, then probabilities and coalescence times follow from Eq. 1.16. If the male and female are not siblings, then with probability $\frac{1}{2}$, the alleles have been inherited from a male and a female in separate mating pairs, and with probability $\frac{1}{2}$, they have been inherited from two females in separate mating pairs. The coalescence times follow similarly from Eq. 1.16. Combining terms, we get

$$\mathbb{E}[V_{mf}] = \frac{1}{N}\left[\frac{1}{4} + \frac{1}{4}(\mathbb{E}[T_f] + 1) + \frac{1}{2}(\mathbb{E}[U] + 1)\right] + \left(1 - \frac{1}{N}\right)\left[\frac{1}{2}(\mathbb{E}[V_{mf}] + 1) + \frac{1}{2}(\mathbb{E}[V_{ff}] + 1)\right]. \tag{1.18}$$

For V_{ff}, if the two females are siblings, then two sampled alleles coalesce in one generation in the father with probability $\frac{1}{4}$ and in the mother with probability $\frac{1}{8}$, giving total coalescence probability $\frac{3}{8}$. With probability $\frac{1}{8}$, the alleles derive from the mother and do not coalesce, giving coalescence time $\mathbb{E}[T_f] + 1$. With probability $\frac{1}{2}$, one allele derives from the mother and the other derives from the father, giving coalescence time $\mathbb{E}[U] + 1$. If the alleles do not come from siblings, then coalescence times and probabilities follow the pattern of autosomes, and the transition probabilities and coalescence times follow the non-sibling portion of Eq. 1.2. Combining terms, we have

$$\begin{aligned}\mathbb{E}[V_{ff}] = \frac{1}{N}&\left[\frac{3}{8} + \frac{1}{8}(\mathbb{E}[T_f] + 1) + \frac{1}{2}(\mathbb{E}[U] + 1)\right] \\ &+ \left(1 - \frac{1}{N}\right)\left[\frac{1}{4}(\mathbb{E}[V_{mm}] + 1) + \frac{1}{2}(\mathbb{E}[V_{mf}] + 1) + \frac{1}{4}(\mathbb{E}[V_{ff}] + 1)\right].\end{aligned} \tag{1.19}$$

Solving the system of five equations associated with the five states, Eqs. 1.15–1.19, we obtain

$$\mathbb{E}[T_f] = \frac{36N^3(1 - c_0) + 56N^2\left(1 - \frac{1}{28}c_0\right) - 6N(1 - c_0) - 6}{12N^2\left(1 - \frac{1}{4}c_0\right) - N(1 - c_0) - 1} \tag{1.20}$$

$$\mathbb{E}[U] = \frac{36N^3(1 - c_0) + 44N^2\left(1 + \frac{1}{44}c_0\right) - 5N(1 - c_0) - 5}{12N^2\left(1 - \frac{1}{4}c_0\right) - N(1 - c_0) - 1} \tag{1.21}$$

$$\mathbb{E}[V_{mm}] = \frac{36N^3\left(1-\frac{3}{4}c_0\right)+23N^2\left(1+\frac{9}{23}c_0\right)-9N}{12N^2\left(1-\frac{1}{4}c_0\right)-N(1-c_0)-1} \tag{1.22}$$

$$\mathbb{E}[V_{mf}] = \frac{36N^3\left(1-\frac{3}{4}c_0\right)+35N^2\left(1-\frac{3}{35}c_0\right)-N\left(1-5c_0\right)-5}{12N^2\left(1-\frac{1}{4}c_0\right)-N(1-c_0)-1} \tag{1.23}$$

$$\mathbb{E}[V_{ff}] = \frac{36N^3\left(1-\frac{3}{4}c_0\right)+29N^2\left(1+\frac{3}{29}c_0\right)-7N\left(1-\frac{3}{7}c_0\right)-3}{12N^2\left(1-\frac{1}{4}c_0\right)-N(1-c_0)-1}. \tag{1.24}$$

In examining these equations, we immediately notice that the highest-order terms and coefficients are the same for Eqs. 1.20 and 1.21 and for Eqs. 1.22–1.24. For large N,

$$\frac{\mathbb{E}[T_f]}{3N} = \frac{\mathbb{E}[U]}{3N} \approx \frac{1-c_0}{1-\frac{1}{4}c_0} \tag{1.25}$$

$$\frac{\mathbb{E}[V_{mm}]}{3N} = \frac{\mathbb{E}[V_{mf}]}{3N} = \frac{\mathbb{E}[V_{ff}]}{3N} \approx \frac{1-\frac{3}{4}c_0}{1-\frac{1}{4}c_0}. \tag{1.26}$$

Eqs. 1.20 and 1.23 divided by $3N$, the number of X chromosomes present each generation, are plotted in Figure 1.3. As N increases, $\mathbb{E}[T_f]/(3N)$ and $\mathbb{E}[V_{mf}]/(3N)$ both quickly approach their large-N limits. The limits, Eqs. 1.25 and 1.26, give reduction factors due to consanguinity. Note that $\mathbb{E}[V_{mf}] - \mathbb{E}[T_f] \approx 3Nc_0/(4-c_0)$, so for $c_0 > 0$, $\mathbb{E}[V_{mf}]/(3N) > \mathbb{E}[T_f]/(3N)$ in the large-N limit. When we take the limit as $N \to \infty$ of $\mathbb{E}[V_{mf}]/(3N)$, we observe that for $c_0 = 1$, the limit is $\frac{1}{3}$, so that consanguinity can reduce the mean coalescence time to $\frac{1}{3}$ of the value it attains without consanguinity.

Although Eqs. 1.22–1.24 have the same large-N approximation, they differ slightly. It is straightforward to verify from the equations that if $N \geq 2$ and $0 \leq c_0 \leq 1$, then $\mathbb{E}[V_{mm}] < \mathbb{E}[V_{ff}] < \mathbb{E}[V_{mf}]$. The pairwise differences among the three equations appear in Figure 1.4, where we can observe this result.

The sequence $\mathbb{E}[V_{mm}] < \mathbb{E}[V_{ff}] < \mathbb{E}[V_{mf}]$ can be explained by coalescence probabilities in a single generation. In each of states 4–6, when we sample parental mating pairs from the previous generation, two individuals are siblings with probability $\frac{1}{N}$. The coalescence probability in the previous generation for two alleles sampled in these siblings is $\frac{1}{2}$ for two males, $\frac{1}{4}$ for a male and a female, and $\frac{3}{8}$ for two females. The lower single-generation probability of coalescence for state 5, or $\frac{1}{4}$, contributes to $\mathbb{E}[V_{mf}]$ having the largest mean coalescence time among the three quantities.

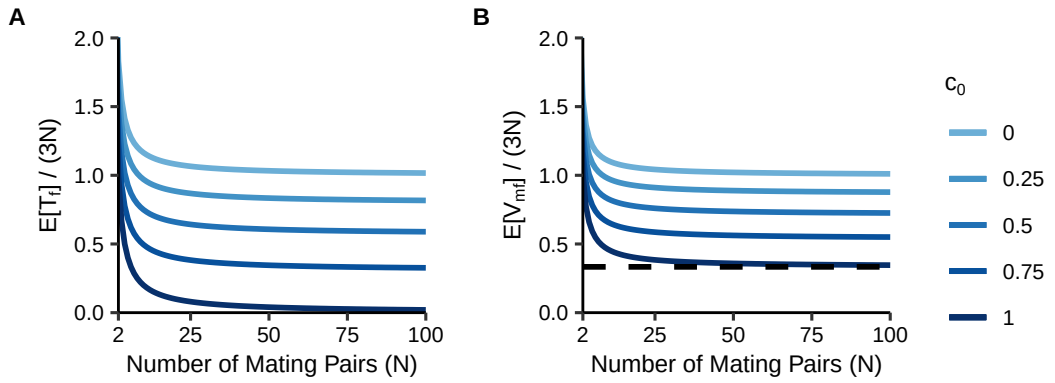

Figure 1.3: Normalized mean coalescence times on the X chromosome under a sib-mating model. Coalescence times are plotted as a function of the number of mating pairs (N) and the proportion of mating pairs that are sib mating pairs (c_0). **(A)** $\mathbb{E}[T_f]/(3N)$, Eq. 1.20. **(B)** $\mathbb{E}[V_{mf}]/(3N)$, Eq. 1.23. The dashed line at $\frac{1}{3}$ represents the minimal mean coalescence time under consanguinity, achieved at $c_0 = 1$.

The higher single-generation coalescence probability of $\frac{1}{2}$ for state 4 contributes to $\mathbb{E}[V_{mm}]$ having the lowest coalescence time, and the intermediate $\frac{3}{8}$ for state 6 places $\mathbb{E}[V_{ff}]$ between the other two values.

With the X-chromosomal mean coalescence times established for sib mating, we can compare them to corresponding autosomal values. Both for autosomes and for the X chromosome, the leading term in the mean coalescence time is a product of the number of alleles in the population, $4N$ or $3N$, and a reduction factor due to consanguinity. For within-individual coalescence times (Eqs. 1.7 and 1.25), the autosomes have reduction factor $1-c_0$ and the X chromosome has reduction factor $(1-c_0)/(1-\frac{1}{4}c_0)$. For $0 < c_0 < 1$, $1 - c_0 < (1-c_0)/(1-\frac{1}{4}c_0)$, and the autosomes have a smaller reduction factor. The reduction factor due to consanguinity has a stronger effect on the autosomal within-individual coalescence time than on the X-chromosomal value.

For the between-individual coalescence times for individuals in different mating pairs (Eqs. 1.8 and 1.26), the reduction factors are $1-\frac{3}{4}c_0$ for autosomes and $(1-\frac{3}{4}c_0)/(1-\frac{1}{4}c_0)$ for the X chromosome, so that again, the effect of consanguinity is stronger on the autosomes than on the X chromosome. Thus, both for pairs of alleles within individuals and for pairs of alleles in separate mating pairs, under sib mating, consanguinity reduces expected time to coalescence by a greater degree in the autosomal case compared to the X-chromosomal case.

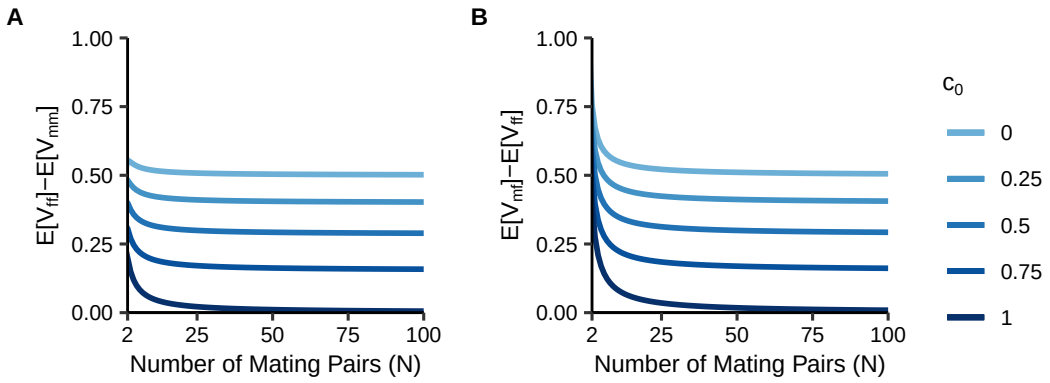

Figure 1.4: Differences between mean coalescence times of two X chromosomes sampled in separate individuals under sib mating. Differences are plotted as a function of the number of mating pairs (N) and the sib mating proportion (c_0). **(A)** $\mathbb{E}[V_{ff}] - \mathbb{E}[V_{mm}]$. **(B)** $\mathbb{E}[V_{mf}] - \mathbb{E}[V_{ff}]$. The panels visualize Eqs. 1.22–1.24.

1.3.2 First cousins

We move next to first-cousin mating. Whereas the four types of cousin mating have the same effect on autosomal coalescence times, for the X chromosome, their effects differ. We consider all four types of first-cousin mating, in addition to two double-first-cousin mating schemes: bilateral-parallel cousins, where a male mates with a female who is both his father's brother's daughter and his mother's sister's daughter, and bilateral-cross cousins, where a male mates with a female who is both his father's sister's daughter and his mother's brother's daughter (Figure 1.5).

In each scheme, we continue to consider N mating pairs, a fraction c_1 of which are consanguineous pairs of a specified cousin mating type. We continue to disallow chance consanguinity, and we permit only a single consanguinity regime at a time. As before, we exclude state 1, as two X chromosomes cannot be in the same male.

The mean coalescence time for two alleles in state 2 is the same irrespective of the type of consanguinity, and it follows from Eq. 1.15 for all of the cases. For states 4, 5, and 6, two individuals in separate mating pairs are siblings with probability $\frac{1}{N}$. Because chance first-cousin mating is not allowed, equations associated with these states follow from our X-chromosomal sib-mating model (Eqs. 1.17–1.19). Only for state 3 does the recursion differ across cases.

Patrilateral-parallel

For patrilateral-parallel cousin mating, it never occurs that X-chromosomal alleles are shared identically by descent in the consanguineous pair, as the male in the pair does not receive an X chromosome from his father (Figure 1.5A). Hence, the fact that their fathers are brothers does not reduce coalescence times of X-chromosomal loci in a consanguineous pair compared with a non-consanguineous pair, and we can disregard c_1.

The X chromosome in the male of the mating pair has probability $\frac{1}{2}$ of coming from a male two generations back, and probability $\frac{1}{2}$ of coming from a female two generations back. The X chromosome in the female of the mating pair has probability $\frac{1}{4}$ of coming from a male two generations back and probability $\frac{3}{4}$ of coming from a female two generations back. Combining these cases for the positions of the alleles two generations back, with probability $\frac{1}{8}$, the alleles in two individuals in a mating pair come from two males, producing coalescence time $\mathbb{E}[V_{mm}] + 2$. They come from a male and a female with probability $\frac{1}{2}$, producing coalescence time $\mathbb{E}[V_{mf}] + 2$. They come from

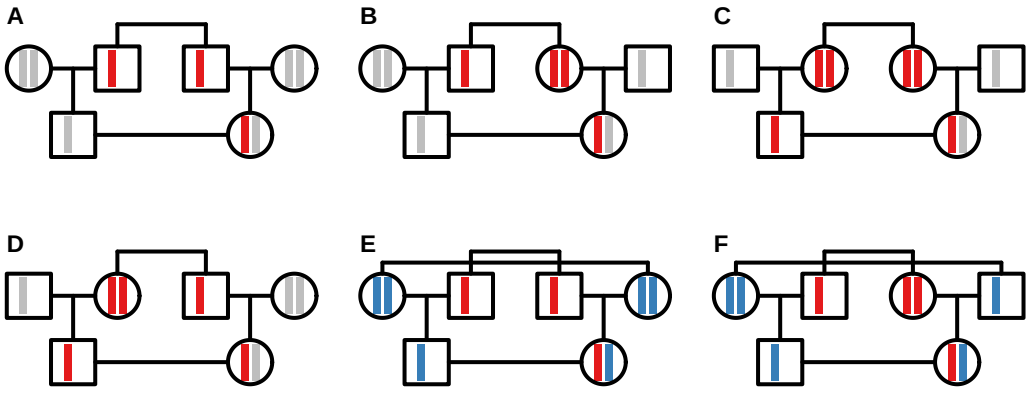

Figure 1.5: X chromosomes in first-cousin mating schemes. **(A)** Patrilateral-parallel. **(B)** Patrilateral-cross. **(C)** Matrilateral-parallel. **(D)** Matrilateral-cross. **(E)** Bilateral-parallel. **(F)** Bilateral-cross. X chromosomes are depicted in red for the sib parents and their offspring in the unilateral mating regimes. The two sets of X chromosomes in the two mating pairs are depicted in red and blue in the bilateral mating regimes, colored identically to associated X chromosomes in the offspring.

two females with probability $\frac{3}{8}$, producing coalescence time $\mathbb{E}[V_{ff}] + 2$. We have

$$\mathbb{E}[U] = \frac{1}{8}(\mathbb{E}[V_{mm}] + 2) + \frac{1}{2}(\mathbb{E}[V_{mf}] + 2) + \frac{3}{8}(\mathbb{E}[V_{ff}] + 2). \tag{1.27}$$

When we solve the linear system of equations formed by Eqs. 1.15, 1.27, and 1.17–1.19, we have

$$\mathbb{E}[T_f] = \frac{144N^3 + 368N^2 - 16N - 40}{48N^2 - N - 5} \tag{1.28}$$

$$\mathbb{E}[U] = \frac{144N^3 + 320N^2 - 15N - 35}{48N^2 - N - 5} \tag{1.29}$$

$$\mathbb{E}[V_{mm}] = \frac{144N^3 + 191N^2 - 65N}{48N^2 - N - 5} \tag{1.30}$$

$$\mathbb{E}[V_{mf}] = \frac{144N^3 + 239N^2 + 15N - 35}{48N^2 - N - 5} \tag{1.31}$$

$$\mathbb{E}[V_{ff}] = \frac{144N^3 + 215N^2 - 33N - 20}{48N^2 - N - 5}. \tag{1.32}$$

As $N \to \infty$, these equations have limit $3N$. The limiting mean coalescence time for each state depends only on the number of mating pairs N and not on the consanguinity rate c_1. For patrilateral-parallel first-cousin mating, consanguinity has no effect on the mean coalescence time of X-chromosomal alleles.

Patrilateral-cross

Patrilateral-cross first-cousin mating (Figure 1.5B) is similar to patrilateral-parallel first-cousin mating in that a pair of X-chromosomal alleles in the two individuals of a consanguineous pair cannot both derive from the sibling parents. As a result, if we proceed through the possible cases for where X chromosomes in a consanguineous pair could be located two generations in the past, we obtain exactly the same cases that underlie Eq. 1.27; for state 3, the mean coalescence time follows Eq. 1.27, and the consanguinity fraction c_1 has no effect on this coalescence time.

The resulting system of equations is the same as for patrilateral-parallel first-cousin mating, and therefore has the same solution. For each state, the mean T_{MRCA} approaches $3N$ as $N \to \infty$. As

was seen with our patrilateral-parallel solution, consanguinity has no effect on coalescence times of X-chromosomal alleles under patrilateral-cross first-cousin mating.

Matrilateral-parallel

In matrilateral-parallel first-cousin mating, the male in a mating pair mates with his mother's sister's daughter (Figure 1.5C). For two alleles sampled in state 3, the probability that the X chromosome in the male of a consanguineous pair originates from the sib parent is 1, and the corresponding probability is $\frac{1}{2}$ for the female. Hence, $c_1/2$ is the probability that a mating pair is consanguineous and both alleles chosen from the two individuals in the mating pair are from the sib parents. Given that the alleles trace to the sib parents, they coalesce two generations back with probability $\frac{3}{8}$. The event that the alleles come from a shared female ancestor two generations back and do not coalesce has probability $\frac{1}{8}$ and gives coalescence time $\mathbb{E}[T_f]+2$. The event that the two alleles come from separate individuals in a mating pair two generations back has probability $\frac{1}{2}$ and gives coalescence time $\mathbb{E}[U]+2$. The event the two alleles are sampled from a non-consanguineous pair or that they are sampled from a consanguineous pair and do not trace to the sib parents has probability $1-c_1/2$. In this case, the alleles follow the same pattern as in Eq. 1.27. Combining the cases, we have

$$\begin{aligned}\mathbb{E}[U] = \frac{c_1}{2}&\left[\frac{3}{8}\times 2+\frac{1}{8}(\mathbb{E}[T_f]+2)+\frac{1}{2}(\mathbb{E}[U]+2)\right]\\ &+\left(1-\frac{c_1}{2}\right)\left[\frac{1}{8}(\mathbb{E}[V_{mm}]+2)+\frac{1}{2}(\mathbb{E}[V_{mf}]+2)+\frac{3}{8}(\mathbb{E}[V_{ff}]+2)\right].\end{aligned} \tag{1.33}$$

Eqs. 1.15, 1.33, and 1.17–1.19 form a linear system, with solution

$$\mathbb{E}[T_f] = \frac{288N^3\left(1-\frac{1}{2}c_1\right)+736N^2\left(1-\frac{1}{92}c_1\right)-32N\left(1-\frac{1}{2}c_1\right)-80}{96N^2\left(1+\frac{1}{16}c_1\right)-2N\left(1-\frac{1}{2}c_1\right)-10\left(1+\frac{1}{10}c_1\right)} \tag{1.34}$$

$$\mathbb{E}[U] = \frac{288N^3\left(1-\frac{1}{2}c_1\right)+640N^2\left(1-\frac{7}{320}c_1\right)-30N\left(1-\frac{1}{2}c_1\right)-70\left(1-\frac{1}{70}c_1\right)}{96N^2\left(1+\frac{1}{16}c_1\right)-2N\left(1-\frac{1}{2}c_1\right)-10\left(1+\frac{1}{10}c_1\right)} \tag{1.35}$$

$$\mathbb{E}[V_{mm}] = \frac{288N^3\left(1-\frac{5}{16}c_1\right)+382N^2\left(1+\frac{25}{382}c_1\right)-130N\left(1-\frac{3}{130}c_1\right)}{96N^2\left(1+\frac{1}{16}c_1\right)-2N\left(1-\frac{1}{2}c_1\right)-10\left(1+\frac{1}{10}c_1\right)} \tag{1.36}$$

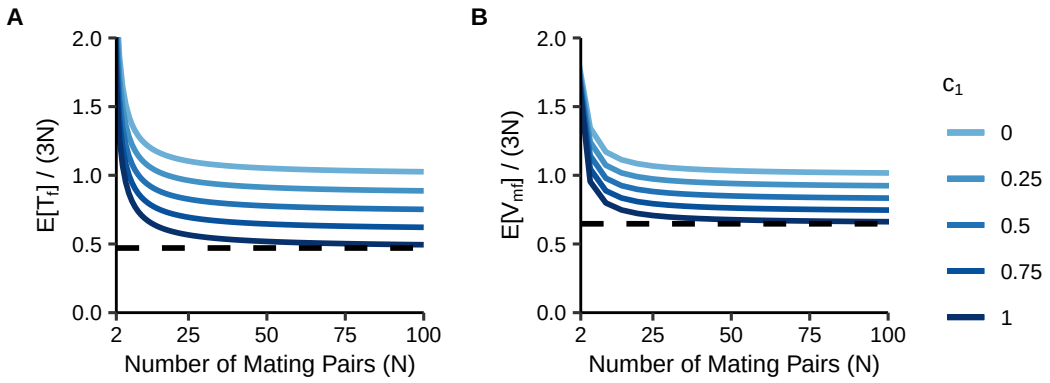

Figure 1.6: Normalized mean coalescence times on the X chromosome under matrilateral-parallel first-cousin mating. Coalescence times are plotted as a function of the number of mating pairs (N) and the proportion of mating pairs that are matrilateral-parallel pairs (c_1). **(A)** $\mathbb{E}[T_f]/(3N)$, Eq. 1.34. **(B)** $\mathbb{E}[V_{mf}]/(3N)$, Eq. 1.37. The dashed lines represent the maximal reduction due to consanguinity, obtained by setting $c_1 = 1$: $\frac{8}{17}$ in **(A)** and $\frac{11}{17}$ in **(B)**.

$$\mathbb{E}[V_{mf}] = \frac{288N^3\left(1-\frac{5}{16}c_1\right)+478N^2\left(1-\frac{23}{478}c_1\right)+30N\left(1+\frac{13}{30}c_1\right)-70\left(1-\frac{1}{70}c_1\right)}{96N^2\left(1+\frac{1}{16}c_1\right)-2N\left(1-\frac{1}{2}c_1\right)-10\left(1+\frac{1}{10}c_1\right)} \tag{1.37}$$

$$\mathbb{E}[V_{ff}] = \frac{288N^3\left(1-\frac{5}{16}c_1\right)+430N^2\left(1+\frac{1}{430}c_1\right)-66N\left(1-\frac{7}{66}c_1\right)-40}{96N^2\left(1+\frac{1}{16}c_1\right)-2N\left(1-\frac{1}{2}c_1\right)-10\left(1+\frac{1}{10}c_1\right)}. \tag{1.38}$$

The highest-order terms are the same for Eqs. 1.34 and 1.35 and for Eqs. 1.36–1.38. The $N \to \infty$ limits give

$$\lim_{N\to\infty}\frac{\mathbb{E}[T_f]}{3N} = \lim_{N\to\infty}\frac{\mathbb{E}[U]}{3N} = \frac{1-\frac{1}{2}c_1}{1+\frac{1}{16}c_1} \tag{1.39}$$

$$\lim_{N\to\infty}\frac{\mathbb{E}[V_{mm}]}{3N} = \lim_{N\to\infty}\frac{\mathbb{E}[V_{mf}]}{3N} = \lim_{N\to\infty}\frac{\mathbb{E}[V_{ff}]}{3N} = \frac{1-\frac{5}{16}c_1}{1+\frac{1}{16}c_1}. \tag{1.40}$$

Eqs. 1.34 and 1.37 normalized by $3N$ are plotted in Figure 1.6. As N increases, $\mathbb{E}[T_f]$ and $\mathbb{E}[V_{mf}]$ both quickly approach their limits, a product of the number of X chromosomes in the population and the reduction factor due to consanguinity. We have $\mathbb{E}[V_{mf}] - \mathbb{E}[T_f] \approx 9Nc_1/(16+c_1)$, so if $0 < c_1 \leq 1$, then $\mathbb{E}[V_{mf}] > \mathbb{E}[T_f]$ in the large-N limit. The lower bound on the limiting reduction factor for $\mathbb{E}[T_f]/(3N)$ due to consanguinity, achieved when $c_1 = 1$, is $\frac{8}{17}$. For $\mathbb{E}[V_{mf}]/(3N)$, the lower bound is $\frac{11}{17}$.

Matrilateral-cross

Next, we consider matrilateral-cross first-cousin mating, in which a male mates with his mother's brother's daughter (Figure 1.5D). We consider two alleles sampled in state 3. As seen in the matrilateral-parallel case, the probability is $c_1/2$ that a mating pair represents first cousins *and* that we sample the sib parent alleles. In this case, because a male is the sib parent of the female in the mating pair, the sampled alleles trace to the shared grandmother. These alleles have three possible origins. First, the two sampled alleles coalesce in two generations with probability $\frac{1}{4}$. Second, with probability $\frac{1}{4}$, the alleles derive from the shared grandmother but do not coalesce, giving coalescence time $\mathbb{E}[T_f] + 2$. Third, with probability $\frac{1}{2}$, the two alleles come separately from the male and female in the grandparental mating pair, giving coalescence time $\mathbb{E}[U] + 2$. As was seen with matrilateral-parallel cousins, if the mating pair is not a first-cousin pair or we do not sample chromosomes that trace to the sib parents, then the transition probabilities and coalescence times follow from Eq. 1.27. Combining cases, we have

$$\begin{aligned}\mathbb{E}[U] = \frac{c_1}{2}\left[\frac{1}{4}\times 2 + \frac{1}{4}(\mathbb{E}[T_f]+2) + \frac{1}{2}(\mathbb{E}[U]+2)\right] \\ + \left(1-\frac{1}{2}c_1\right)\left[\frac{1}{8}(\mathbb{E}[V_{mm}]+2) + \frac{1}{2}(\mathbb{E}[V_{mf}]+2) + \frac{3}{8}(\mathbb{E}[V_{ff}]+2)\right].\end{aligned} \tag{1.41}$$

Eqs. 1.15, 1.41, and 1.17–1.19 form a linear system with solution

$$\mathbb{E}[T_f] = \frac{288N^3\left(1-\frac{1}{2}c_1\right) + 736N^2\left(1-\frac{1}{92}c_1\right) - 32N\left(1-\frac{1}{2}c_1\right) - 80}{96N^2\left(1-\frac{1}{8}c_1\right) - 2N\left(1-\frac{1}{2}c_1\right) - 10\left(1-\frac{1}{10}c_1\right)} \tag{1.42}$$

$$\mathbb{E}[U] = \frac{288N^3\left(1-\frac{1}{2}c_1\right) + 640N^2\left(1+\frac{1}{160}c_1\right) - 30N\left(1-\frac{1}{2}c_1\right) - 70\left(1+\frac{1}{70}c_1\right)}{96N^2\left(1-\frac{1}{8}c_1\right) - 2N\left(1-\frac{1}{2}c_1\right) - 10\left(1-\frac{1}{10}c_1\right)} \tag{1.43}$$

$$\mathbb{E}[V_{mm}] = \frac{288N^3\left(1-\frac{3}{8}c_1\right) + 382N^2\left(1+\frac{33}{382}c_1\right) - 130N\left(1+\frac{3}{130}c_1\right)}{96N^2\left(1-\frac{1}{8}c_1\right) - 2N\left(1-\frac{1}{2}c_1\right) - 10\left(1-\frac{1}{10}c_1\right)} \tag{1.44}$$

$$\mathbb{E}[V_{mf}] = \frac{288N^3\left(1-\frac{3}{8}c_1\right) + 478N^2\left(1-\frac{15}{478}c_1\right) + 30N\left(1+\frac{17}{30}c_1\right) - 70\left(1+\frac{1}{70}c_1\right)}{96N^2\left(1-\frac{1}{8}c_1\right) - 2N\left(1-\frac{1}{2}c_1\right) - 10\left(1-\frac{1}{10}c_1\right)} \tag{1.45}$$

$$\mathbb{E}[V_{ff}] = \frac{288N^3\left(1-\frac{3}{8}c_1\right) + 430N^2\left(1+\frac{9}{430}c_1\right) - 66N\left(1-\frac{3}{22}c_1\right) - 40}{96N^2\left(1-\frac{1}{8}c_1\right) - 2N\left(1-\frac{1}{2}c_1\right) - 10\left(1-\frac{1}{10}c_1\right)}. \tag{1.46}$$

The highest-order terms agree for Eqs. 1.42,1.43 and 1.44–1.46. The $N \to \infty$ limits give

$$\lim_{N\to\infty} \frac{\mathbb{E}[T_f]}{3N} = \lim_{N\to\infty} \frac{\mathbb{E}[U]}{3N} = \frac{1-\frac{1}{2}c_1}{1-\frac{1}{8}c_1} \tag{1.47}$$

$$\lim_{N\to\infty} \frac{\mathbb{E}[V_{mm}]}{3N} = \lim_{N\to\infty} \frac{\mathbb{E}[V_{mf}]}{3N} = \lim_{N\to\infty} \frac{\mathbb{E}[V_{ff}]}{3N} = \frac{1-\frac{3}{8}c_1}{1-\frac{1}{8}c_1}. \tag{1.48}$$

Figure 1.7 plots Eqs. 1.42 and 1.45 divided by $3N$. As N increases, $\mathbb{E}[T_f]/(3N)$ and $\mathbb{E}[V_{mf}]/(3N)$ quickly approach their limits. $\mathbb{E}[V_{mf}] - \mathbb{E}[T_f] \approx 3Nc_1/(8-c_1)$, so for $0 < c_1 \leq 1$, $\mathbb{E}[V_{mf}] > \mathbb{E}[T_f]$ in the large-N limit. The lower bound on the large-N reduction factor for $\mathbb{E}[T_f]/(3N)$ is $\frac{4}{7}$, achieved at $c_1 = 1$. For $\mathbb{E}[V_{mf}]/(3N)$, the bound is $\frac{5}{7}$.

1.3.3 Double first cousins

Bilateral-parallel

In bilateral-parallel cousin mating, the male in a pair mates with a female who is his mother's sister's daughter and his father's brother's daughter (Figure 1.5E). The bilateral-parallel case contains both matrilateral-parallel and patrilateral-parallel cousin mating. Hence, for two alleles sampled in state 3, as in the matrilateral-parallel case, the event that the individuals are first cousins and the

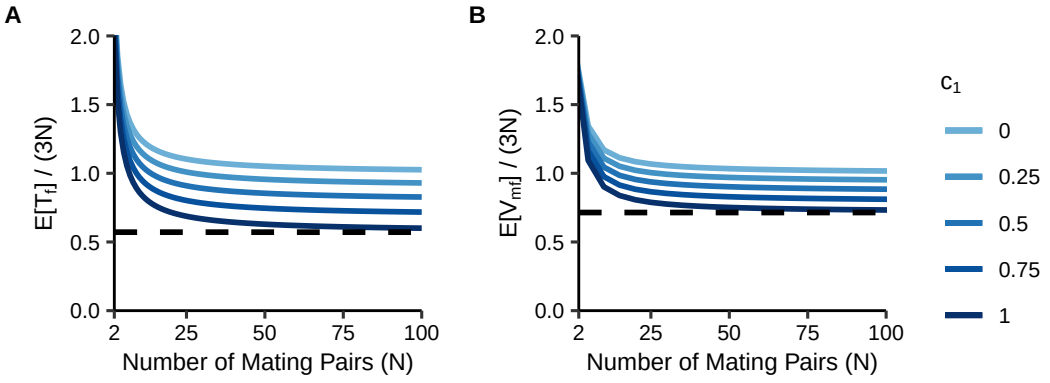

Figure 1.7: Normalized mean coalescence times on the X chromosome under matrilateral-cross first-cousin mating. Coalescence times are plotted as a function of the number of mating pairs (N) and the proportion of mating pairs that are matrilateral-cross pairs (c_1): **(A)** $\mathbb{E}[T_f]/(3N)$, Eq. 1.42. **(B)** $\mathbb{E}[V_{mf}]/(3N)$, Eq. 1.45. The dashed lines represent the maximal reduction due to consanguinity, obtained by setting $c_1 = 1$: $\frac{4}{7}$ in **(A)** and $\frac{5}{7}$ in **(B)**.

	Autosomal			X-chromosomal			
	Siblings	First cousins	Double first cousins	Siblings	First cousins		
					Patrilateral-parallel, Patrilateral-cross	*Matrilateral-parallel, Bilateral-parallel*	*Matrilateral-cross, Bilateral-cross*
Kinship coefficient	$\frac{1}{4}$	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{2}$	0	$\frac{3}{16}$	$\frac{1}{8}$
T_m, T_f, U	$1-c_0$	$1-\frac{1}{4}c_1$	$1-\frac{1}{2}c_1$	$\frac{1-c_0}{1-\frac{1}{4}c_0}$	1	$\frac{1-\frac{1}{2}c_1}{1+\frac{1}{16}c_1}$	$\frac{1-\frac{1}{2}c_1}{1-\frac{1}{8}c_1}$
V_{mm}, V_{mf}, V_{ff}	$1-\frac{3}{4}c_0$	$1-\frac{3}{16}c_1$	$1-\frac{3}{8}c_1$	$\frac{1-\frac{3}{4}c_0}{1-\frac{1}{4}c_0}$	1	$\frac{1-\frac{5}{16}c_1}{1+\frac{1}{16}c_1}$	$\frac{1-\frac{3}{8}c_1}{1-\frac{1}{8}c_1}$

Table 1.1: Large-N reduction factors of pairwise coalescence times due to various types of consanguinity. These reduction factors give multipliers for the mean T_{MRCA} for a non-consanguineous population of $2N$ individuals: $4N$ for autosomes and $3N$ for X chromosomes. The kinship coefficient for a mating pair refers to the probability that two alleles sampled from the two individuals of the pair are identical by descent. Kinship coefficients are computed separately for autosomal loci and X-chromosomal loci. Note that T_m is used only in the autosomal case; for the X chromosome, two alleles cannot be in the same male.

two alleles are sampled from the sib parents has probability $c_1/2$ to trace to the female sib parents and probability 0 to trace to the male sib parents. The patrilateral-parallel line of descent does not contribute to the possibility of identity by descent, so that the bilateral-parallel case behaves like the matrilateral parallel case. The equation associated with state 3 follows from Eq. 1.33. All equations for the bilateral-parallel case mirror the matrilateral-parallel case, and the same conclusions follow.

Bilateral-cross

In bilateral-cross cousin mating, the male in a pair mates with a female who is simultaneously his father's sister's daughter and his mother's brother's daughter (Figure 1.5F). The matrilateral-cross component of the bilateral-cross pedigree contributes to allele sharing in the mating pair, but the patrilateral-cross line of descent does not contribute to allele sharing in the mating pair. Coalescence times follow from the matrilateral-cross case.

1.3.4 Comparison of cousin-mating schemes

The six cousin mating regimes, four unilateral and two bilateral, produce three patterns in their effects on X-chromosomal coalescence times. The patrilateral-parallel and patrilateral-cross cases

are mathematically identical; for short, we refer to this pair simply as patrilateral. The matrilateral-parallel and bilateral-parallel cases are identical, as are the matrilateral-cross and bilateral-cross cases; we refer to these pairs as matrilateral-parallel and matrilateral-cross, respectively. Reduction factors for the various cases are summarized in Table 1.1.

As noted in Sections 1.3.2 and 1.3.2, consanguinity has no effect on patrilateral cousin mating because the male in the first-cousin mating pair never inherits an X chromosome from the sib parent (his father). We can, however, compare reduction factors due to consanguinity for the matrilateral-parallel and matrilateral-cross cases. For $\mathbb{E}[T_f]$ and $\mathbb{E}[U]$, the reduction factor for matrilateral-parallel cousin mating is $(1 - \frac{1}{2}c_1)/(1 + \frac{1}{16}c_1)$ (Eq. 1.39). The reduction factor for matrilateral-cross cousin mating is $(1 - \frac{1}{2}c_1)/(1 - \frac{1}{8}c_1)$ (Eq. 1.47).

To place the various X-chromosomal reduction factors in rank order, we observe that for $\mathbb{E}[T_f]$ and $\mathbb{E}[U]$, $(1 - \frac{1}{2}c_1)/(1 + \frac{1}{16}c_1) < (1 - \frac{1}{2}c_1)/(1 - \frac{1}{8}c_1) < 1$ for $0 < c_1 \leq 1$. The matrilateral-parallel case has the strongest reduction factor, followed by the matrilateral-cross case, followed by the patrilateral case, which has no reduction at all. For $\mathbb{E}[V_{mm}]$, $\mathbb{E}[V_{mf}]$, and $\mathbb{E}[V_{ff}]$, we can see that $(1 - \frac{5}{16}c_1)/(1 + \frac{1}{16}c_1) < (1 - \frac{3}{8}c_1)/(1 - \frac{1}{8}c_1) < 1$ for $0 < c_1 \leq 1$. As is true for $\mathbb{E}[T_f]$ and $\mathbb{E}[U]$, the matrilateral-parallel case has the strongest reduction factor, followed by the matrilateral-cross case, followed by the patrilateral case (Figure 1.8).

Additionally, $(1 - \frac{1}{2}c_1)/(1 - \frac{1}{8}c_1) < 1 - \frac{1}{4}c_1$ for $0 < c_1 \leq 1$, so that the reduction in $\mathbb{E}[T_f]$ and $\mathbb{E}[U]$ for the matrilateral-cross case is stronger than for the autosomal case. Similarly, $(1 - \frac{3}{8}c_1)/(1 - \frac{1}{8}c_1) < 1 - \frac{3}{16}c_1$, so that the same is true for $\mathbb{E}[V_{mm}]$, $\mathbb{E}[V_{mf}]$, and $\mathbb{E}[V_{ff}]$.

The reduction factors in Table 1.1 indicate that consanguinity has no effect on coalescence times for X-chromosomal loci under patrilateral cousin mating, and in increasing order of strength, the effect of consanguinity is greater on autosomal loci under first cousin mating, on X-chromosomal loci under matrilateral-cross mating, and on X-chromosomal loci under matrilateral-parallel mating.

When we also consider the autosomal double-first-cousin case, we see that $(1 - \frac{1}{2}c_1)/(1 + \frac{1}{16}c_1) < 1 - \frac{1}{2}c_1$ for $0 < c_1 \leq 1$, so that the reduction in $\mathbb{E}[T_f]$ and $\mathbb{E}[U]$ for the X-chromosomal matrilateral-parallel case is stronger than for the autosomal bilateral case. We also see that $1 - \frac{1}{2}c_1 < (1 - \frac{1}{2}c_1)/(1 - \frac{1}{8}c_1)$ for $0 < c_1 \leq 1$, so that the reduction in $\mathbb{E}[T_f]$ and $\mathbb{E}[U]$ is greater for

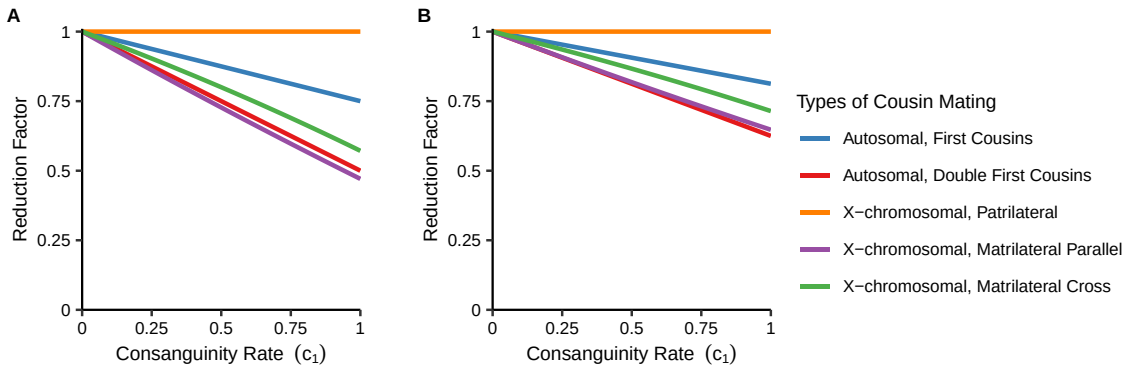

Figure 1.8: Reduction factors for mean autosomal and X-chromosomal pairwise coalescence times under various types of unilateral and bilateral first-cousin mating, plotted as a function of the consanguinity rate (c_1). **(A)** Reduction factors for $\mathbb{E}[T_m]$, $\mathbb{E}[T_f]$, and $\mathbb{E}[U]$ ($\mathbb{E}[T_m]$ for autosomes only). **(B)** Reduction factors for $\mathbb{E}[V_{mm}]$, $\mathbb{E}[V_{mf}]$, and $\mathbb{E}[V_{ff}]$. See Table 1.1 for formulae.

the autosomal bilateral case than for the X-chromosomal matrilateral-cross case. For $\mathbb{E}[T_f]$ and $\mathbb{E}[U]$, the strengths of reduction in increasing order are matrilateral-cross, autosomal bilateral, and matrilateral-parallel (Figure 1.8A).

We next examine the reduction factors for $\mathbb{E}[V_{mm}]$, $\mathbb{E}[V_{mf}]$, and $\mathbb{E}[V_{ff}]$ for double-first-cousin mating. We see that $1 - \frac{3}{8}c_1 < (1 - \frac{5}{16}c_1)/(1 + \frac{1}{16}c_1)$ for $0 < c_1 \leq 1$, so that the reduction in $\mathbb{E}[V_{mm}]$, $\mathbb{E}[V_{mf}]$, and $\mathbb{E}[V_{ff}]$ for the autosomal double-first-cousin case is stronger than for the X-chromosomal matrilateral-parallel case. For $\mathbb{E}[V_{mm}]$, $\mathbb{E}[V_{mf}]$, and $\mathbb{E}[V_{ff}]$, the effect of consanguinity in increasing order is matrilateral-cross, matrilateral-parallel, and autosomal bilateral (Figure 1.8B). Whereas the unilateral cases have the same rank order for the strength of reduction for within-individual states ($\mathbb{E}[T_f]$ and $\mathbb{E}[U]$) and between-individual states ($\mathbb{E}[V_{mm}]$, $\mathbb{E}[V_{mf}]$, and $\mathbb{E}[V_{ff}]$), these rank orders differ for the autosomal bilateral case.

1.4 Discussion

We have explored the effect of consanguinity on mean pairwise coalescence times on the X chromosome. We extended the model of Severson *et al.* (2019) to the X chromosome, comparing X-chromosomal and autosomal theoretical results. We found that patrilateral first-cousin mating

has no effect on X-chromosomal coalescence; however, matrilateral first-cousin mating—and especially matrilateral-parallel mating—decreases X-chromosomal mean pairwise coalescence times relative to autosomal coalescence times (Table 1.1).

In a coalescent model with first-cousin mating in a population of N diploid mating pairs, we have observed that in a large population, mean pairwise T_{MRCA} is the product of the mean coalescence time in the absence of consanguinity, $4N$ for autosomes and $3N$ for X chromosomes, and a reduction factor due to consanguinity (Table 1.1). Four types of unilateral first-cousin mating, which are equivalent for autosomal loci, differ in their effects on the X chromosome (Eqs. 1.9–1.11, 1.39 and 1.40, 1.47 and 1.48).

In all four types of unilateral first-cousin mating, mean coalescence time for two alleles in the same individual is reduced by consanguinity to a greater extent than mean coalescence time for two alleles in two individuals in separate mating pairs (Table 1.1). This effect for the X chromosome accords with similar autosomal results (Severson *et al.*, 2019). Because the X chromosome is never inherited from father to son, patrilateral-parallel and patrilateral-cross consanguinity do not affect X-chromosomal coalescence times. Matrilateral consanguinity, however, induces a stronger reduction in X-chromosomal coalescence times compared to autosomal coalescence times.

For bilateral consanguinity, corresponding to double-first-cousin matings, coalescence times follow the relationship of the mothers in the consanguineous pair: bilateral-parallel consanguinity has the same coalescence time as matrilateral-parallel consanguinity, and bilateral-cross consanguinity has the same coalescence time as matrilateral-cross consanguinity. Interestingly, however, whereas bilateral-parallel X-chromosomal coalescence times have a stronger reduction than that of bilateral autosomal coalescence times when considering within-individual coalescence ($\mathbb{E}[T_m]$ and $\mathbb{E}[T_f]$), this order is reversed for between-individual coalescence ($\mathbb{E}[V_{mm}]$, $\mathbb{E}[V_{mf}]$, and $\mathbb{E}[V_{ff}]$).

Genomic sharing around a site is inversely related to coalescence time at that site (Palamara *et al.*, 2012; Carmi *et al.*, 2014; Browning and Browning, 2015); specifically, ROH lengths are inversely related to within-individual coalescence times, and IBD lengths are inversely related to between-individual coalescence times. Thus, the model predicts that reduced coalescence times on the X chromosome—owing to both smaller population size and to a stronger effect of matrilateral consanguinity in reducing those times—give rise to longer ROH and IBD sharing on

the X chromosome than on the autosomes. Further, the smaller mean within-individual coalescence times ($\mathbb{E}[T_m]$ and $\mathbb{E}[T_f]$) than between-individual coalescence times ($\mathbb{E}[V_{mm}]$, $\mathbb{E}[V_{mf}]$, and $\mathbb{E}[V_{ff}]$) predict greater ROH sharing within individuals than IBD sharing between individuals. The smaller X-chromosomal population size compared to autosomal population size, together with the greater reduction from matrilateral consanguinity of X-chromosomal within- and between-individual T_{MRCA} compared to autosomal within- and between-individual T_{MRCA}, predicts greater ROH and IBD sharing on the X chromosome relative to autosomes.

Recent analyses have explored a variety of population-genetic features of the difference between the X chromosome and autosomes, tracing both to a difference in population size and also to effects of differing inheritance patterns. Such studies have included analyses focused on computations of nucleotide diversity (Arbiza *et al.*, 2014), coalescence times (Ramachandran *et al.*, 2008), and genomic sharing (Buffalo *et al.*, 2016), and on various consequences of sex-biased demography (Wilkins and Marlowe, 2006; Bustamante and Ramachandran, 2009; Goldberg and Rosenberg, 2015; Webster and Wilson Sayres, 2016). Our model adds to this work in its focus on effects of specific consanguinity models.

We note that our models are limited in that we have studied the effect of different types of first-cousin consanguinity separately, and we have not considered a population with a mixture of the various types. In actual populations, while one of the four types of unilateral consanguinity might be culturally preferred (Bittles, 2012), the appropriate model suited to a specific population might involve a superposition of two or more types. The separation-of-time-scales coalescent approach of Severson *et al.* (2021) successfully examined a superposition of consanguinity at different levels of relatedness; we expect that this method will be useful for superimposing multiple forms of first-cousin consanguinity. The separation-of-time-scales approach also has the benefit of producing asymptotic coalescence time distributions extending beyond mean coalescence times.

A second limitation is that our model formulation considers only the most recent shared ancestral pair for a consanguineous pair; that ancestral pair could itself be consanguineous, so that consanguineous pairs in the current generation might possess additional, more distant, shared ancestors. The approximation that this more distant consanguinity is ignored in computing mean

coalescence times is likely to be more problematic in cases in which multiple shared lines of descent are most probable, such as for small N or for large c_1 (Severson *et al.*, 2021).

This year marks the centennial of the pioneering studies of Sewall Wright (1921) on the effects of mating models on features of pointwise genotypic sharing—studies that have been central to the large volume of subsequent work on genetic consequences of consanguinity, inbreeding, and relatedness (Hill, 1996). In recent years, the study of consanguinity and its connections to runs of homozygosity and identity by descent has been much advanced by new models and genomic tools for data analysis (Bittles, 2012; Browning and Browning, 2012; Thompson, 2013; Romeo and Bittles, 2014; Cussens and Sheehan, 2016; Ceballos *et al.*, 2018). By exploring coalescent models that incorporate each of the various types of first-cousin consanguinity, we have determined the effects of first-cousin consanguinity in shaping X-chromosomal coalescence times, and by extension, genomic sharing. In addition to providing new coalescent theory for populations with consanguinity, the study further enhances the understanding of the effects of sex-biased processes on genomes, the factors that contribute to differences in genetic variation between X chromosomes and autosomes, and the determinants of patterns of genomic sharing.

1.5 Appendix A: Autosomal first cousins

We consider the two-sex autosomal model with a fraction, c_1, of first-cousin mating pairs. Each generation, N is the number of mating pairs. We forbid chance sib mating, chance first-cousin mating, and double-first-cousin mating. As in the autosomal sib mating case, alleles within an individual (states 1 and 2) must have come from a single mating pair one generation back. Hence $\mathbb{E}[T_m]$ and $\mathbb{E}[T_f]$ follow Eq. 1.1 from the sib mating case.

$\mathbb{E}[U]$ again represents the mean coalescence time for two alleles in two individuals in a mating pair (state 3). Two individuals in a mating pair are first cousins with probability c_1. Given that a mating pair represents first cousins, the probability that the sampled alleles both come from the sib parent is $\frac{1}{4}$. The sampled alleles then have four possible cases. First, the alleles coalesce two generations back with probability $\frac{1}{4}$. They both derive from the shared grandfather with probability $\frac{1}{8}$, giving coalescence time $\mathbb{E}[T_m] + 2$. They both derive from the shared grandmother with

probability $\frac{1}{8}$, giving coalescence time $\mathbb{E}[T_f] + 2$. With probability $\frac{1}{2}$, they derive separately from the two grandparents, coalescing with time $\mathbb{E}[U] + 2$.

If the sampled alleles are either not from a first-cousin mating pair or not from the sib parents of a first-cousin mating pair—an event with probability $1 - \frac{1}{4}c_1$— then the two alleles must be in state 4, 5, or 6 two generations back, and the transition probabilities follow the non-sib mating portion of Eq. 1.2. Combining cases, we have

$$\begin{aligned}\mathbb{E}[U] = \frac{c_1}{4}\left[\frac{1}{4}\times 2 + \frac{1}{8}(\mathbb{E}[T_m]+2) + \frac{1}{8}(\mathbb{E}[T_f]+2) + \frac{1}{2}(\mathbb{E}[U]+2)\right] \\ + \left(1-\frac{1}{4}c_1\right)\left[\frac{1}{4}(\mathbb{E}[V_{mm}]+2) + \frac{1}{2}(\mathbb{E}[V_{mf}]+2) + \frac{1}{4}(\mathbb{E}[V_{ff}]+2)\right].\end{aligned} \tag{1.49}$$

For two alleles in separate mating pairs, because parental mating pairs are chosen at random from the previous generation, the probability that the two individuals are siblings by chance is $\frac{1}{N}$. Eq. 1.3 continues to apply. Solving the linear system containing Eq. 1.1, 1.49, and 1.3, we obtain Eqs. 1.9–1.11. When sex is ignored, Eqs. 1.9–1.11 reduce to Eqs. 8-10 from Severson *et al.* (2019).

1.6 Appendix B: Autosomal double first cousins

We extend the first-cousin mating model of Appendix A to double first cousins. If two individuals are double first cousins, then each parent of one individual is a sibling of a parent of the other individual, and the individuals share two grandparental mating pairs (Figure 1.9). This scenario can occur with children of two brother-sister pairs or with children of two brothers and two sisters. For autosomes, the two categories are mathematically equivalent.

Each generation, the fraction of double-first-cousin mating pairs is a constant value c_1. Chance sibling mating, first-cousin mating, and double-first-cousin mating are forbidden among non-consanguineous mating pairs. If two alleles are present within one individual, then they must have been present in two individuals in a mating pair in the previous generation. $\mathbb{E}[T_m]$ and $\mathbb{E}[T_f]$ have the same recursions as before (Eq. 1.1).

For $\mathbb{E}[U]$, if two alleles are in two individuals of a mating pair, then with probability c_1, those individuals are double first cousins. The probability that two alleles in the double first cousins are

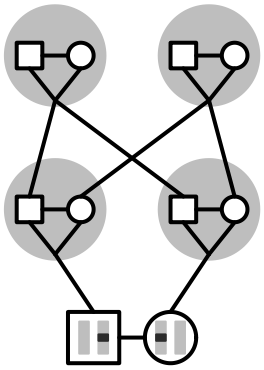

Figure 1.9: Double first cousins. If two individuals in a mating pair are double first cousins, then each parent of the female is a sibling of one of the parents of the male, and the male and female share two grandparental mating pairs.

inherited from one specific grandparental pair is $\frac{1}{4}$, and the probability that two alleles are inherited from the same grandparental pair is $\frac{1}{2}$. If the two alleles are inherited from a grandparental pair, then two generations ago they have four options. They coalesce with probability $\frac{1}{4}$, giving a coalescence time of 2. With probability $\frac{1}{8}$, they are inherited from the grandfather with mean coalescence time $\mathbb{E}[T_m] + 2$. With probability $\frac{1}{8}$, they are inherited from the grandmother with mean coalescence time $\mathbb{E}[T_f] + 2$. Finally, with probability $\frac{1}{2}$ they are two alleles in the two individuals in a grandparental mating pair, giving mean coalescence time $\mathbb{E}[U] + 2$. If the individuals are not double first cousins (and because chance sib mating, first-cousin mating, and double-first-cousin mating are forbidden), then the two alleles have probability $\frac{1}{4}$ of being in two males in separate mating pairs two generations ago, probability $\frac{1}{2}$ of being in a male and a female in separate mating pairs two generations ago, and probability $\frac{1}{4}$ of being in two females in separate mating pairs two generations ago. These cases have mean coalescence times $\mathbb{E}[V_{mm}] + 2$, $\mathbb{E}[V_{mf}] + 2$, and $\mathbb{E}[V_{ff}] + 2$, respectively. Combining cases gives

$$\begin{aligned}\mathbb{E}[U] = \frac{c_1}{2}\left[\frac{1}{4}\times 2 + \frac{1}{8}(\mathbb{E}[T_m]+2) + \frac{1}{8}(\mathbb{E}[T_f]+2) + \frac{1}{2}(\mathbb{E}[U]+2)\right] \\ + \left(1-\frac{1}{2}c_1\right)\left[\frac{1}{4}(\mathbb{E}[V_{mm}]+2) + \frac{1}{2}(\mathbb{E}[V_{mf}]+2) + \frac{1}{4}(\mathbb{E}[V_{ff}]+2)\right]. \qquad (1.50)\end{aligned}$$

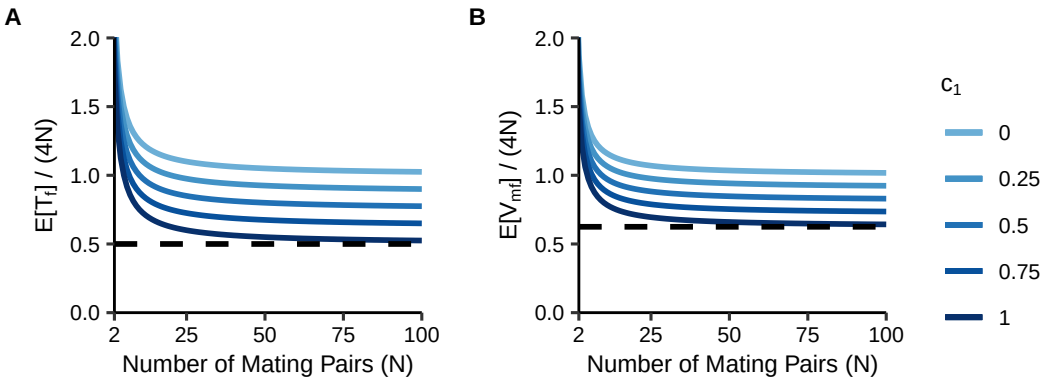

Figure 1.10: Normalized mean coalescence times for autosomal double-first-cousin mating. Coalescence times are plotted as a function of the number of mating pairs (N) and the proportion of mating pairs that are double-first-cousin pairs (c_1). **(A)** $\mathbb{E}[T_f]/(4N)$, Eq. 1.12. **(B)** $\mathbb{E}[V_{mf}]/(4N)$, Eq. 1.14. The dashed lines represent the maximal reduction due to consanguinity, obtained by setting $c_1 = 1$: $\frac{1}{2}$ in **(A)** and $\frac{5}{8}$ in **(B)**.

$\mathbb{E}[V_{mm}]$, $\mathbb{E}[V_{mf}]$, and $\mathbb{E}[V_{ff}]$ also have the same recursions as before; because parental mating pairs are chosen uniformly at random with replacement from the N possible pairs, if two alleles are in two individuals in separate mating pairs, then the individuals share the same parental mating pair with probability $\frac{1}{N}$, giving rise to Eq. 1.3. Eqs. 1.1, 1.50, and 1.3 form a system of equations, the solution to which is given in Eqs. 1.12–1.14.

Eqs. 1.12–1.14 have a similar form to the solutions for first cousin mating, the difference being that $c_1/4$ is replaced by $c_1/2$ in Eqs. 1.12 and 1.13 and $\frac{3}{16}c_1$ by $\frac{3}{8}c_1$ in Eq. 1.14; the kinship coefficient for double first cousins ($\frac{1}{8}$) is twice that of first cousins ($\frac{1}{16}$). Next, $\mathbb{E}[V_{mf}] - \mathbb{E}[T_m] = Nc_1/2 - 3$. Hence, if $c_1 > \frac{6}{N}$ (or the number of consanguineous mating pairs Nc_1 exceeds 6), then the mean coalescence time for two alleles in different mating pairs ($\mathbb{E}[V_{mm}]$, $\mathbb{E}[V_{mf}]$, and $\mathbb{E}[V_{ff}]$) exceeds the mean coalescence time for two alleles within an individual ($\mathbb{E}[T_m]$ and $\mathbb{E}[T_f]$). For c_1 near 0, the mean coalescence times are near $4N$, and as c_1 approaches 1, $\mathbb{E}[T_m] \approx 2N$ and $\mathbb{E}[V_{mf}] \approx \frac{5}{2}N$.

Eqs. 1.12 and 1.14 normalized by $4N$ are plotted in Figure 1.10. The means are bounded below by the maximal reduction due to consanguinity, $\frac{1}{2}$ for $\mathbb{E}[T_m]$ and $\frac{5}{8}$ for $\mathbb{E}[V_{mf}]$. As the number of mating pairs, N, increases, the mean coalescence times approach the product of $4N$, the mean coalescence time for a non-consanguineous diploid population of size $2N$, and the reduction factor due to consanguinity, $1 - \frac{1}{2}c_1$ for Eq. 1.12 and $1 - \frac{3}{8}c_1$ for Eq. 1.14.

Chapter 2

Limiting distribution of X-chromosomal coalescence times under first-cousin consanguineous mating

The following chapter and figures were originally published as:

Cotter, D. J., A. L. Severson, S. Carmi, and N. A. Rosenberg, 2022 Limiting distribution of X-chromosomal coalescence times under first-cousin consanguineous mating. Theoretical Population Biology **147**: 1–15.

https://doi.org/10.1016/j.tpb.2022.07.002

Abstract

By providing additional opportunities for coalescence within families, the presence of consanguineous unions in a population reduces coalescence times relative to non-consanguineous populations. First-cousin consanguinity can take one of six forms differing in the configuration of sexes in the pedigree of the male and female cousins who join in a consanguineous union: patrilateral

parallel, patrilateral cross, matrilateral parallel, matrilateral cross, bilateral parallel, and bilateral cross. Considering populations with each of the six types of first-cousin consanguinity individually and a population with a mixture of the four unilateral types, we examine coalescent models of consanguinity. We previously computed, for first-cousin consanguinity models, the mean coalescence time for X-chromosomal loci and the limiting distribution of coalescence times for autosomal loci. Here, we use the separation-of-time-scales approach to obtain the limiting distribution of coalescence times for X-chromosomal loci. This limiting distribution has an instantaneous coalescence probability that depends on the probability that a union is consanguineous; lineages that do not coalesce instantaneously coalesce according to an exponential distribution. We study the effects on the coalescence time distribution of the type of first-cousin consanguinity, showing that patrilateral-parallel and patrilateral-cross consanguinity have no effect on X-chromosomal coalescence time distributions and that matrilateral-parallel consanguinity decreases coalescence times to a greater extent than does matrilateral-cross consanguinity.

2.1 Introduction

The phenomenon of consanguinity, in which unions occur between closely related individuals, is a form of population structure that can dramatically affect properties of genetic variation (Crow and Kimura, 1970; Jacquard, 1974). By increasing the probability that deleterious recessive variants appear in homozygous form compared to the corresponding probability in a population in which it is absent, consanguinity contributes to the incidence of recessive disease (Bittles, 2001; Woods *et al.*, 2006); recent studies suggest that it contributes to incidence of complex disease as well (Bittles and Black, 2010; Yengo *et al.*, 2017; Ceballos *et al.*, 2018; Johnson *et al.*, 2018; Clark *et al.*, 2019). Consanguinity is common in human populations, with some populations promoting consanguineous marriages as a cultural preference (Bittles, 2012; Romeo and Bittles, 2014; Sahoo *et al.*, 2021).

The offspring of a consanguineous union are expected to possess large portions of their genomes shared between their two genomic copies, owing to the fact that an identical genomic segment can be inherited along both their maternal and paternal lines. For the loci contained in such segments,

the two copies coalesce at a common ancestor relatively few generations in the past. At other locations, neither copy or only one copy traces to a recent shared ancestor, so that coalescence occurs only much farther back in the past. Indeed, empirical genetic studies have identified multiple populations in which individuals carry long homozygous segments that indicate recent coalescence of the two genomic copies and that are attributable in large part to consanguinity practices (McQuillan *et al.*, 2008; Pemberton *et al.*, 2012; Ceballos *et al.*, 2018)

In typical coalescent-based models that investigate coalescence times for sets of lineages, diploid organisms are approximated by pairs of haploids independently drawn from a population (Hein *et al.*, 2005; Wakeley, 2009). This modeling choice is unsuited to the study of consanguineous families, in which the two lineages in an individual can be highly dependent. Hence, explicitly diploid coalescent models have been devised for the study of coalescence in a setting of consanguinity. The earliest studies focused on selfing in plants (Pollak, 1987; Nordborg and Donnelly, 1997; Nordborg and Krone, 2002), an extreme form of "consanguinity" in which both parents of a diploid offspring are the same individual. Campbell (2015) extended diploid coalescent models to consider a monogamous mating model with sibling mating, computing mean coalescence times under the model. This approach was then extended by Severson *et al.* (2019) to consider mean coalescence times in a diploid model with nth-cousin mating, for arbitrary values of n and for superpositions of multiple levels of nth-cousin mating.

In an extension of the work of Severson *et al.* (2019), Severson *et al.* (2021) advanced beyond mean coalescence times to derive a full limiting distribution of coalescence times under superposition models of autosomal consanguinity, considering the limit as the population size grows large. A limitation of the work of Severson *et al.* (2019) and Severson *et al.* (2021), however, is that it does not distinguish between males and females in the mating model; all individuals are exchangeable. Hence, it cannot accommodate the variety of scenarios in which differences between males and females are salient. We have recently extended the method of Severson *et al.* (2019) to distinguish between males and females, evaluating mean coalescence times in a two-sex model, to determine the effect of consanguinity on X-chromosomal coalescence times specifically (Cotter *et al.*, 2021).

Here, we use the advance from Severson *et al.* (2021) to compute the full distribution of coalescence times under a diploid, two-sex consanguinity model (Cotter *et al.*, 2021). Seeking to derive

distributions of X-chromosomal coalescence times, we consider each of the six types of first-cousin consanguinity and a model that includes all four unilateral types in a single population. For each model, we evaluate the distribution of coalescence times for two lineages sampled from the same individual and for two lineages sampled from members of different mating pairs.

2.2 Methods

We adapt the models of Severson *et al.* (2019, 2021) and Cotter *et al.* (2021). We consider a constant-sized population of N diploid mating pairs. Individuals are sex-specific, the X chromosome is considered, and specified forms of consanguinity are allowed. Using a Markov chain, we track lineage pairs back in time until they coalesce.

To analyze the large-N limit of the model, we make use of the separation-of-time-scales approach introduced by Möhle (1998). This approach was used by Severson *et al.* (2021) to obtain the limiting distribution of coalescence times under their autosomal diploid model of consanguinity. In the approach from Möhle (1998), the limiting distribution of a Markov process with transition matrix Π_N is obtained by writing

$$\Pi_N = \mathbf{A} + \frac{1}{N}\mathbf{B}. \tag{2.1}$$

Here, $\mathbf{A}$ describes "fast" transitions that have nontrivial probability in a single generation, and $\mathbf{B}$ describes "slow" transitions that have very small probabilities in a single generation. As $N \to \infty$, the fast transitions occur instantaneously, and the fast process can be described by an equilibrium distribution

$$\mathbf{P} = \lim_{r\to\infty} \mathbf{A}^r. \tag{2.2}$$

Rescaling t in units of N generations, as $N \to \infty$, Π_N converges to a continuous-time process

$$\Pi(t) = \lim_{N\to\infty} (\Pi_N)^{Nt} = \mathbf{P}e^{t\mathbf{G}}. \tag{2.3}$$

The rate matrix $\mathbf{G}$ satisfies $\mathbf{G} = \mathbf{PBP}$. Under Möhle's theorem, the process converges to a continuous-time process with an instantaneous jump at time 0 that corresponds to the "fast" transitions.

As Severson *et al.* (2021) did with autosomal models, we apply the separation-of-time-scales approach to our models of consanguinity on the X chromosome (Cotter *et al.*, 2021). We begin with the sib mating case and then consider each of the four types of unilateral first-cousin mating, the two cases of bilateral first-cousin mating, and a mixture of all four unilateral types in one model.

2.3 Results

2.3.1 Sibling mating

We consider N monogamous male–female mating pairs, a fraction c_0 of which are sib mating pairs. Pairs of X-chromosomal lineages can be in one of six states (Figure 2.1): two lineages have already coalesced (state 0); two lineages are in a female (state 1); two lineages are in opposite individuals of a mating pair (state 2); two lineages are in two individuals in different mating pairs, where the two individuals are two males (state 3), a male and a female (state 4), or two females (state 5). Note that for the X chromosome, there is no state for two lineages in a male, as males contain only one X chromosome. We track the state of the process backward in time until it reaches the most recent common ancestor for a pair of lineages (that is, until state 0 is reached). We denote by T_f, U, V_{mm}, V_{mf}, and V_{ff} the random coalescence time for pairs of lineages in states 1, 2, 3, 4, and 5, respectively.

If two lineages are in state 0 (coalesced), they remain in state 0 with probability 1; this state is absorbing. If two lineages are in a female (state 1), in the previous generation they must have been in separate individuals in a mating pair (state 2) with probability 1. If two lineages are in separate individuals in a mating pair (state 2), the pair is a sib mating pair with probability c_0. Given that the pair is a sib mating pair, the lineages transition to state 0 with probability $\frac{1}{4}$, state 1 with probability $\frac{1}{4}$, and state 2 with probability $\frac{1}{2}$. If the two lineages are not in a sib mating pair, an event with probability $1 - c_0$, then they transition to states 4 and 5 with equal probability $\frac{1}{2}$.

For each of the states 3–5, because we pick parental mating pairs with replacement from the previous generation, the probability is $\frac{1}{N}$ that the same mating pair is chosen. Thus, if two lineages are in state 3, and the pair are siblings (an event with probability $\frac{1}{N}$), then the lineages transition to state 0 or state 1, each with probability $\frac{1}{2}$. If the two lineages in state 3 do not have the same

Figure 2.1: Five states for two lineages. Males are squares; females are circles. State 1: within a female (blue). State 2: in two individuals in a mating pair (green). State 3: in two males in different mating pairs (yellow). State 4: in a male and a female in different mating pairs (orange). State 5: in two females in different mating pairs (purple).

parental pair (probability $1-\frac{1}{N}$), then they must transition to state 5 with probability 1. For state 4, if the two lineages are in siblings (probability $\frac{1}{N}$), then they transition to state 0 with probability $\frac{1}{4}$, state 1 with probability $\frac{1}{4}$, and state 2 with probability $\frac{1}{2}$. If the lineages are not from siblings (probability $1-\frac{1}{N}$), then they transition to state 4 or 5, each with probability $\frac{1}{2}$. Finally, two lineages in state 5, conditional on being in siblings (probability $\frac{1}{N}$), reach state 0 with probability $\frac{3}{8}$, state 1 with probability $\frac{1}{8}$, and state 2 with probability $\frac{1}{2}$. Conditional on not being in siblings (probability $1-\frac{1}{N}$), the lineages transition to state 3 with probability $\frac{1}{4}$, state 4 with probability $\frac{1}{2}$, and state 5 with probability $\frac{1}{4}$. Combining these transition probabilities, we can write the transition matrix as

$$\Pi_N = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ \frac{c_0}{4} & \frac{c_0}{4} & \frac{c_0}{2} & 0 & \frac{1-c_0}{2} & \frac{1-c_0}{2} \\ \frac{1}{2N} & \frac{1}{2N} & 0 & 0 & 0 & 1-\frac{1}{N} \\ \frac{1}{4N} & \frac{1}{4N} & \frac{1}{2N} & 0 & \frac{1-\frac{1}{N}}{2} & \frac{1-\frac{1}{N}}{2} \\ \frac{3}{8N} & \frac{1}{8N} & \frac{1}{2N} & \frac{1-\frac{1}{N}}{4} & \frac{1-\frac{1}{N}}{2} & \frac{1-\frac{1}{N}}{4} \end{pmatrix} \end{array}. \tag{2.4}$$

We can decompose Π_N (Eq. 2.4) into its fast and slow transitions, as in Eq. 2.1:

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ \frac{c_0}{4} & \frac{c_0}{4} & \frac{c_0}{2} & 0 & \frac{1-c_0}{2} & \frac{1-c_0}{2} \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & -1 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} & 0 & -\frac{1}{2} & -\frac{1}{2} \\ \frac{3}{8} & \frac{1}{8} & \frac{1}{2} & -\frac{1}{4} & -\frac{1}{2} & -\frac{1}{4} \end{pmatrix}. \tag{2.5}$$

We first find the equilibrium distribution of the "fast" process, obtained by iterating transition matrix $\mathbf{A}$. This calculation appears in Appendix A, producing

$$\mathbf{P} = \lim_{r\to\infty} \mathbf{A}^r = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ \frac{c_0}{4-3c_0} & 0 & 0 & \frac{1}{9}\left(\frac{4-4c_0}{4-3c_0}\right) & \frac{4}{9}\left(\frac{4-4c_0}{4-3c_0}\right) & \frac{4}{9}\left(\frac{4-4c_0}{4-3c_0}\right) \\ \frac{c_0}{4-3c_0} & 0 & 0 & \frac{1}{9}\left(\frac{4-4c_0}{4-3c_0}\right) & \frac{4}{9}\left(\frac{4-4c_0}{4-3c_0}\right) & \frac{4}{9}\left(\frac{4-4c_0}{4-3c_0}\right) \\ 0 & 0 & 0 & \frac{1}{9} & \frac{4}{9} & \frac{4}{9} \\ 0 & 0 & 0 & \frac{1}{9} & \frac{4}{9} & \frac{4}{9} \\ 0 & 0 & 0 & \frac{1}{9} & \frac{4}{9} & \frac{4}{9} \end{pmatrix}. \tag{2.6}$$

We then compute $\mathbf{G} = \mathbf{PBP}$ and solve for the limiting process $\Pi(t)$ using Eq. 2.3, obtaining the matrix exponential, $e^{t\mathbf{G}}$, as in Appendix B. Converting t back into units of N generations, this gives

$$\Pi(t) = \mathbf{P}e^{t\mathbf{G}} =$$

$$\begin{pmatrix} 1 & 0\ 0 & 0 & 0 & 0 \\ 1-\frac{1-c_0}{1-\frac{3}{4}c_0}e^{-\frac{t}{3N}\left(\frac{1-\frac{c_0}{4}}{1-\frac{3}{4}c_0}\right)} & 0\ 0 & \frac{1}{9}\cdot\frac{1-c_0}{1-\frac{3}{4}c_0}e^{-\frac{t}{3N}\left(\frac{1-\frac{c_0}{4}}{1-\frac{3}{4}c_0}\right)} & \frac{4}{9}\cdot\frac{1-c_0}{1-\frac{3}{4}c_0}e^{-\frac{t}{3N}\left(\frac{1-\frac{c_0}{4}}{1-\frac{3}{4}c_0}\right)} & \frac{4}{9}\cdot\frac{1-c_0}{1-\frac{3}{4}c_0}e^{-\frac{t}{3N}\left(\frac{1-\frac{c_0}{4}}{1-\frac{3}{4}c_0}\right)} \\ 1-\frac{1-c_0}{1-\frac{3}{4}c_0}e^{-\frac{t}{3N}\left(\frac{1-\frac{c_0}{4}}{1-\frac{3}{4}c_0}\right)} & 0\ 0 & \frac{1}{9}\cdot\frac{1-c_0}{1-\frac{3}{4}c_0}e^{-\frac{t}{3N}\left(\frac{1-\frac{c_0}{4}}{1-\frac{3}{4}c_0}\right)} & \frac{4}{9}\cdot\frac{1-c_0}{1-\frac{3}{4}c_0}e^{-\frac{t}{3N}\left(\frac{1-\frac{c_0}{4}}{1-\frac{3}{4}c_0}\right)} & \frac{4}{9}\cdot\frac{1-c_0}{1-\frac{3}{4}c_0}e^{-\frac{t}{3N}\left(\frac{1-\frac{c_0}{4}}{1-\frac{3}{4}c_0}\right)} \\ 1-e^{-\frac{t}{3N}\left(\frac{1-\frac{c_0}{4}}{1-\frac{3}{4}c_0}\right)} & 0\ 0 & \frac{1}{9}\cdot e^{-\frac{t}{3N}\left(\frac{1-\frac{c_0}{4}}{1-\frac{3}{4}c_0}\right)} & \frac{4}{9}\cdot e^{-\frac{t}{3N}\left(\frac{1-\frac{c_0}{4}}{1-\frac{3}{4}c_0}\right)} & \frac{4}{9}\cdot e^{-\frac{t}{3N}\left(\frac{1-\frac{c_0}{4}}{1-\frac{3}{4}c_0}\right)} \\ 1-e^{-\frac{t}{3N}\left(\frac{1-\frac{c_0}{4}}{1-\frac{3}{4}c_0}\right)} & 0\ 0 & \frac{1}{9}\cdot e^{-\frac{t}{3N}\left(\frac{1-\frac{c_0}{4}}{1-\frac{3}{4}c_0}\right)} & \frac{4}{9}\cdot e^{-\frac{t}{3N}\left(\frac{1-\frac{c_0}{4}}{1-\frac{3}{4}c_0}\right)} & \frac{4}{9}\cdot e^{-\frac{t}{3N}\left(\frac{1-\frac{c_0}{4}}{1-\frac{3}{4}c_0}\right)} \\ 1-e^{-\frac{t}{3N}\left(\frac{1-\frac{c_0}{4}}{1-\frac{3}{4}c_0}\right)} & 0\ 0 & \frac{1}{9}\cdot e^{-\frac{t}{3N}\left(\frac{1-\frac{c_0}{4}}{1-\frac{3}{4}c_0}\right)} & \frac{4}{9}\cdot e^{-\frac{t}{3N}\left(\frac{1-\frac{c_0}{4}}{1-\frac{3}{4}c_0}\right)} & \frac{4}{9}\cdot e^{-\frac{t}{3N}\left(\frac{1-\frac{c_0}{4}}{1-\frac{3}{4}c_0}\right)} \end{pmatrix}. \tag{2.7}$$

The first column of the matrix $\Pi(t)$ represents the cumulative probability of coalescence in time less than or equal to t generations. States 1 and 2 have the same cumulative distribution, representing the coalescence time for two lineages *within* a female (note that state 2, two lineages in the two individuals in a mating pair, is always reached from state 1 after one step). States 3–5 have the same cumulative distribution, representing the coalescence time for two lineages in two

distinct individuals. The cumulative distributions are

$$F_{T_f}(t) = F_U(t) = 1 - \frac{1 - c_0}{1 - \frac{3}{4}c_0} e^{-\frac{t}{3N}\left(\frac{1-\frac{c_0}{4}}{1-\frac{3}{4}c_0}\right)}, \tag{2.8}$$

$$F_{V_{mm}}(t) = F_{V_{mf}}(t) = F_{V_{ff}}(t) = 1 - e^{-\frac{t}{3N}\left(\frac{1-\frac{c_0}{4}}{1-\frac{3}{4}c_0}\right)}. \tag{2.9}$$

Computing the expectations of these distributions, recalling that for $X > 0$, $\mathbb{E}[X] = \int_0^\infty [1 - F_X(x)]\, dx$, we find

$$\mathbb{E}[T_f] = E[U] = 3N\left(\frac{1 - c_0}{1 - \frac{1}{4}c_0}\right), \tag{2.10}$$

$$\mathbb{E}[V_{mm}] = \mathbb{E}[V_{mf}] = \mathbb{E}[V_{ff}] = 3N\left(\frac{1 - \frac{3}{4}c_0}{1 - \frac{1}{4}c_0}\right). \tag{2.11}$$

where Eqs. 2.10 and 2.11 are the same as Eqs. 25 and 26 from Cotter *et al.* (2021), obtained by first-step analysis.

Eqs. 2.8 and 2.9 are plotted in Figure 2.2. In the figure, we observe that the cumulative probability of coalescence increases with the consanguinity probability c_0. For $c_0 = 0$, $\mathbb{E}[T_f] = \mathbb{E}[V_{mf}] = 3N$, the mean coalescence time for a haploid population with $3N$ chromosomes (the number of X

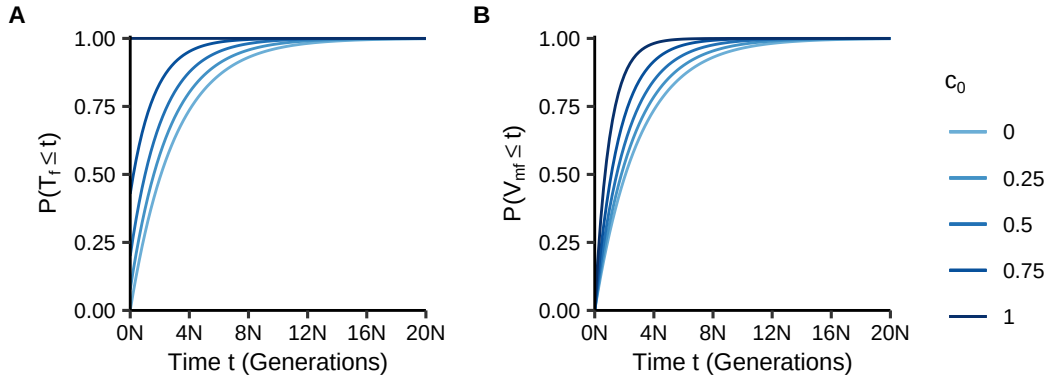

Figure 2.2: Cumulative distributions of coalescence times as functions of the number of generations t and the fraction of sib mating pairs c_0. **(A)** Coalescence time within individuals, $P(T_f \leq t)$, Eq. 2.8. **(B)** Coalescence time between individuals, $P(V_{mf} \leq t)$, Eq. 2.9.

chromosomes in our scenario). For $c_0 > 0, \mathbb{E}[T_f] < \mathbb{E}[V_{mf}]$ due to the probability of consanguinity whenever the two lineages are already in the same mating pair.

2.3.2 First cousins

We next consider first-cousin consanguinity on the X chromosome. We separately calculate the limiting distributions of coalescence times for each of the four types of first-cousin consanguinity: patrilateral parallel, a union of a male with his father's brother's daughter; patrilateral cross, a union of a male with his father's sister's daughter; matrilateral parallel, a union of a male with his mother's sister's daughter; and matrilateral cross, a union of a male with his mother's brother's daughter.

For each of these four types of first-cousin consanguinity, two lineages have seven possible states. State 0 is an absorbing state representing coalescence. State 1 is two lineages in a female. States 3–5 represent, as in the sibling case, two lineages that are in two individuals in *different* mating pairs, where the two individuals are two males (state 3), a male and a female (state 4), or two females (state 5).

Next, for pairs of lineages from the two individuals in a mating pair, we follow the model of a superposition of multiple mating levels from Severson *et al.* (2021), taking a special case of this approach. Under the superposition model, each state 2_i, $0 \leq i \leq n$, represents an ancestral state for two lineages from a mating pair. These ancestral states can be viewed as "holding states" that keep track of ancestral lineages of a mating pair in order to allow all possible ith-cousin levels of consanguinity up to nth cousins. As we restrict attention to first-cousin mating, we need only states 2_0 and 2_1 from Severson *et al.* (2021).

State 2_0 represents two lineages in the two individuals in a mating pair. State 2_1 represents two lineages in two individuals ancestral to the two individuals in a mating pair. Because, unlike Severson *et al.* (2021), we disallow sib mating, two lineages in state 2_0 cannot coalesce (state 0), they cannot transition to the same individual (state 1), nor can they transition to two individuals in a mating pair (state 2_0). Hence, lineages in 2_0 must transition to 2_1 (Figures 2.3 and 2.4).

In the absence of consanguinity, two lineages in state 2_1 can transition only to states 3, 4, and 5 (Figure 2.3). With first-cousin consanguinity present (Figure 2.4), two lineages in state 2_1 can

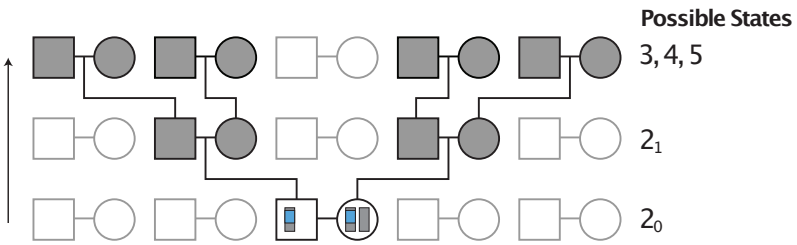

Figure 2.3: Example pedigree illustrating transitions from state 2_0 in the absence of consanguinity. Considering a pair of lineages in a mating pair, depicted in blue, the process always immediately transitions to the holding state 2_1 one generation in the past. From state 2_1, the lineages transition to two separate mating pairs, and hence, to states 3, 4, or 5.

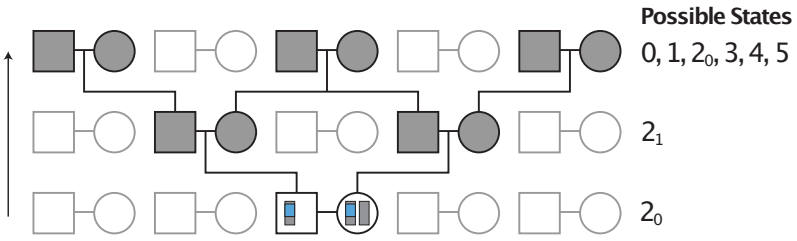

Figure 2.4: Example pedigree illustrating transitions from state 2_0 in the presence of first-cousin consanguinity. Considering a pair of lineages in a mating pair, depicted in blue, the process always immediately transitions to the holding state 2_1. From state 2_1, the lineages can potentially transition to any of states 0, 1, 2_0, 3, 4, or 5, depending on the type of first-cousin consanguinity. Matrilateral-cross consanguinity is depicted.

also coalesce (state 0) or transition to two lineages in the same female (state 1) or to two lineages in opposite individuals in a mating pair (state 2_0).

The transition matrix depends on the type of first-cousin consanguinity permitted. However, the type of consanguinity only affects transitions from state 2_1. For all types of consanguinity, state 0 is an absorbing state. State 1, two lineages in the same female, always transitions to state 2_0 because the two lineages must come from opposite individuals of the same mating pair. Because of the constraints we have placed on the process, state 2_0 always transitions to state 2_1. Finally, the transition probabilities from states 3, 4, and 5 follow the same pattern as given in the transition matrix in Eq. 2.4 (with state 2_0 in place of state 2).

Below, we consider each of the four different types of first-cousin mating, two cases of bilateral first-cousin mating, and a mixture of the four unilateral types. In each case, we define the transitions that the process makes from state 2_1, and we obtain the limiting distributions of coalescence times.

Patrilateral parallel

In patrilateral parallel first-cousin consanguinity, a union occurs between a male and his father's brother's daughter. There is no way for the X-chromosomal lineages in the first-cousin mating pair to have originated from the shared grandparental pair because X chromosomes are never transmitted from fathers to sons. Hence, irrespective of the fraction c_1 in the population, lineages in state 2_1 can only transition to states 3, 4, and 5.

In state 2_1, one X chromosome in one of the parental pairs is always in a female (the parent of the male in state 2_0). The probability is then $\frac{1}{2}$ that this X chromosome is in a male one generation ancestral to 2_1 and $\frac{1}{2}$ that it is in a female. The other X chromosome in state 2_1, located in a parent of the female in state 2_0, can be in a male or female, with equal probability. Hence, one generation ancestral to 2_1, this X chromosome is in a female with probability $\frac{3}{4}$ and in a male with probability $\frac{1}{4}$. We can multiply probabilities for the two separate X chromosomes to obtain transition probabilities from state 2_1. In particular, the two lineages will be in two separate males one generation previously (state 3) with probability $\frac{1}{8}$. They will be in a male and a female (state 4) with probability $\frac{1}{2}$. They will be in two separate females (state 5) with probability $\frac{3}{8}$. The transition matrix is:

$$\Pi_N = \begin{array}{c} \\ 0 \\ 1 \\ 2_0 \\ 2_1 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccccc} 0 & 1 & 2_0 & 2_1 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{8} & \frac{1}{2} & \frac{3}{8} \\ \frac{1}{2N} & \frac{1}{2N} & 0 & 0 & 0 & 0 & 1-\frac{1}{N} \\ \frac{1}{4N} & \frac{1}{4N} & \frac{1}{2N} & 0 & 0 & \frac{1-\frac{1}{N}}{2} & \frac{1-\frac{1}{N}}{2} \\ \frac{3}{8N} & \frac{1}{8N} & \frac{1}{2N} & 0 & \frac{1-\frac{1}{N}}{4} & \frac{1-\frac{1}{N}}{2} & \frac{1-\frac{1}{N}}{4} \end{pmatrix} \end{array}. \tag{2.12}$$

As with the sibling case, we can decompose the transitions into "fast" and "slow" transitions (Eq. 2.1):

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{8} & \frac{1}{2} & \frac{3}{8} \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 0 & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 & -1 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} & 0 & 0 & -\frac{1}{2} & -\frac{1}{2} \\ \frac{3}{8} & \frac{1}{8} & \frac{1}{2} & 0 & -\frac{1}{4} & -\frac{1}{2} & -\frac{1}{4} \end{pmatrix}. \tag{2.13}$$

We next solve for the limiting distribution of the fast transition matrix **A** using the method of Appendix A,

$$\mathbf{P} = \lim_{r\to\infty} \mathbf{A}^r = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{9} & \frac{4}{9} & \frac{4}{9} \\ 0 & 0 & 0 & 0 & \frac{1}{9} & \frac{4}{9} & \frac{4}{9} \\ 0 & 0 & 0 & 0 & \frac{1}{9} & \frac{4}{9} & \frac{4}{9} \\ 0 & 0 & 0 & 0 & \frac{1}{9} & \frac{4}{9} & \frac{4}{9} \\ 0 & 0 & 0 & 0 & \frac{1}{9} & \frac{4}{9} & \frac{4}{9} \\ 0 & 0 & 0 & 0 & \frac{1}{9} & \frac{4}{9} & \frac{4}{9} \end{pmatrix}. \tag{2.14}$$

Recalling $\mathbf{G} = \mathbf{PBP}$, we solve for the limit $\Pi(t)$ as in the sibling mating case, using Eq. 2.3, calculating the matrix exponential, $e^{t\mathbf{G}}$, as in Appendix B. We then convert t back into units of generations N. This step gives

$$\Pi(t) = \mathbf{P}e^{t\mathbf{G}} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 - e^{-\frac{t}{3N}} & 0 & 0 & 0 & \frac{1}{9}e^{-\frac{t}{3N}} & \frac{4}{9}e^{-\frac{t}{3N}} & \frac{4}{9}e^{-\frac{t}{3N}} \\ 1 - e^{-\frac{t}{3N}} & 0 & 0 & 0 & \frac{1}{9}e^{-\frac{t}{3N}} & \frac{4}{9}e^{-\frac{t}{3N}} & \frac{4}{9}e^{-\frac{t}{3N}} \\ 1 - e^{-\frac{t}{3N}} & 0 & 0 & 0 & \frac{1}{9}e^{-\frac{t}{3N}} & \frac{4}{9}e^{-\frac{t}{3N}} & \frac{4}{9}e^{-\frac{t}{3N}} \\ 1 - e^{-\frac{t}{3N}} & 0 & 0 & 0 & \frac{1}{9}e^{-\frac{t}{3N}} & \frac{4}{9}e^{-\frac{t}{3N}} & \frac{4}{9}e^{-\frac{t}{3N}} \\ 1 - e^{-\frac{t}{3N}} & 0 & 0 & 0 & \frac{1}{9}e^{-\frac{t}{3N}} & \frac{4}{9}e^{-\frac{t}{3N}} & \frac{4}{9}e^{-\frac{t}{3N}} \\ 1 - e^{-\frac{t}{3N}} & 0 & 0 & 0 & \frac{1}{9}e^{-\frac{t}{3N}} & \frac{4}{9}e^{-\frac{t}{3N}} & \frac{4}{9}e^{-\frac{t}{3N}} \end{pmatrix}. \tag{2.15}$$

Here, examining the first column of the matrix in Eq. 2.15—representing transitions to coalescence—we can see that two lineages within an individual (state 1), within a mating pair (state 2_0), or in

two separate mating pairs (states 3, 4, and 5) have equal coalescence times. In fact, as coalescence times are unaffected by patrilateral-parallel first-cousin consanguinity, they accord with the coalescence time distribution for a population of size $3N$ haploid individuals. Using the same random variables from the sibling case (where U now represents 2_0), we can extract the cumulative distribution functions of coalescence times from the first column of the matrix $\Pi(t)$:

$$F_{T_f}(t) = F_U(t) = 1 - e^{-\frac{t}{3N}}, \tag{2.16}$$

$$F_{V_{mm}}(t) = F_{V_{mf}}(t) = F_{V_{ff}}(t) = 1 - e^{-\frac{t}{3N}}. \tag{2.17}$$

For each of the five random random variables, the time to coalescence for two lineages is distributed as an exponential random variable with rate $1/(3N)$. The mean of these distributions—the reciprocal of the coalescence rate—is $3N$, matching the limiting means obtained by first-step analysis in Eqs. 28–32 of Cotter *et al.* (2021).

Patrilateral cross

For the patrilateral-cross case, a union occurs between a male and his father's sister's daughter. As with the parallel case, there is no way for the X-chromosomal lineages in the first-cousin mating pair to have originated from a shared ancestor. We obtain the exact same transition probabilities from state 2_1 and the same transition matrix (Eq. 2.12). The coalescence times for the patrilateral-cross case are the same as in the parallel case.

Matrilateral parallel

In the matrilateral parallel case, a union occurs between a male and his mother's sister's daughter. With probability $c_1/2$, two lineages in state 2_1 trace back to the shared grandparental pair. The lineages in state 2_1 coalesce with probability $\frac{3}{8}$ (state 0), they are in the shared grandmother with probability $\frac{1}{8}$ (state 1), and they are in opposite individuals of the grandparental mating pair with probability $\frac{1}{2}$ (state 2_0).

With probability $c_1/2$, two lineages in state 2_1 do not trace back to the shared grandparental pair. Conditional on not tracing to this pair, they are in a male and a female (state 4) or two females

(state 5), each with probability $\frac{1}{2}$. Finally, with probability $1-c_1$, the two lineages are not ancestral to a consanguineous mating pair; they then follow the same pattern as in the patrilateral-parallel case. Combining cases gives the transition matrix,

$$\Pi_N = \begin{array}{c} 0 \\ 1 \\ 2_0 \\ 2_1 \\ 3 \\ 4 \\ 5 \end{array} \overset{\begin{array}{ccccccc} 0 & 1 & 2_0 & 2_1 & 3 & 4 & 5 \end{array}}{\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ \frac{3c_1}{16} & \frac{c_1}{16} & \frac{c_1}{4} & 0 & \frac{1}{8}-\frac{c_1}{8} & \frac{1}{2}-\frac{c_1}{4} & \frac{3}{8}-\frac{c_1}{8} \\ \frac{1}{2N} & \frac{1}{2N} & 0 & 0 & 0 & 0 & 1-\frac{1}{N} \\ \frac{1}{4N} & \frac{1}{4N} & \frac{1}{2N} & 0 & 0 & \frac{1-\frac{1}{N}}{2} & \frac{1-\frac{1}{N}}{2} \\ \frac{3}{8N} & \frac{1}{8N} & \frac{1}{2N} & 0 & \frac{1-\frac{1}{N}}{4} & \frac{1-\frac{1}{N}}{2} & \frac{1-\frac{1}{N}}{4} \end{pmatrix}}. \tag{2.18}$$

As before, we decompose this matrix into "fast" and "slow" transitions (Eq. 2.1):

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ \frac{3c_1}{16} & \frac{c_1}{16} & \frac{c_1}{4} & 0 & \frac{1}{8}-\frac{c_1}{8} & \frac{1}{2}-\frac{c_1}{4} & \frac{3}{8}-\frac{c_1}{8} \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 0 & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 & -1 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} & 0 & 0 & -\frac{1}{2} & -\frac{1}{2} \\ \frac{3}{8} & \frac{1}{8} & \frac{1}{2} & 0 & -\frac{1}{4} & -\frac{1}{2} & -\frac{1}{4} \end{pmatrix}. \tag{2.19}$$

We next solve for the limiting distribution of the fast matrix $\mathbf{A}$ using the method of Appendix A:

$$\mathbf{P} = \lim_{r\to\infty} \mathbf{A}^r = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{3c_1}{16-5c_1} & 0 & 0 & 0 & \frac{1}{9}\left(\frac{16-8c_1}{16-5c_1}\right) & \frac{4}{9}\left(\frac{16-8c_1}{16-5c_1}\right) & \frac{4}{9}\left(\frac{16-8c_1}{16-5c_1}\right) \\ \frac{3c_1}{16-5c_1} & 0 & 0 & 0 & \frac{1}{9}\left(\frac{16-8c_1}{16-5c_1}\right) & \frac{4}{9}\left(\frac{16-8c_1}{16-5c_1}\right) & \frac{4}{9}\left(\frac{16-8c_1}{16-5c_1}\right) \\ \frac{3c_1}{16-5c_1} & 0 & 0 & 0 & \frac{1}{9}\left(\frac{16-8c_1}{16-5c_1}\right) & \frac{4}{9}\left(\frac{16-8c_1}{16-5c_1}\right) & \frac{4}{9}\left(\frac{16-8c_1}{16-5c_1}\right) \\ 0 & 0 & 0 & 0 & \frac{1}{9} & \frac{4}{9} & \frac{4}{9} \\ 0 & 0 & 0 & 0 & \frac{1}{9} & \frac{4}{9} & \frac{4}{9} \\ 0 & 0 & 0 & 0 & \frac{1}{9} & \frac{4}{9} & \frac{4}{9} \end{pmatrix}. \tag{2.20}$$

Finally, recalling $\mathbf{G} = \mathbf{PBP}$, we solve for the matrix exponential $e^{t\mathbf{G}}$ using the method of Appendix B. We then solve for the continuous-time process $\Pi(t)$ via Eq. 2.3, converting t back to units of N generations:

$$\Pi(t) = \mathbf{P}e^{t\mathbf{G}} =$$
$$\begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 - \frac{1-\frac{c_1}{2}}{1-\frac{5}{16}c_1} e^{-\frac{t}{3N}\left(\frac{1+\frac{c_1}{16}}{1-\frac{5}{16}c_1}\right)} & 0 & 0 & 0 & \frac{1}{9} \cdot \frac{1-\frac{c_1}{2}}{1-\frac{5}{16}c_1} e^{-\frac{t}{3N}\left(\frac{1+\frac{c_1}{16}}{1-\frac{5}{16}c_1}\right)} & \frac{4}{9} \cdot \frac{1-\frac{c_1}{2}}{1-\frac{5}{16}c_1} e^{-\frac{t}{3N}\left(\frac{1+\frac{c_1}{16}}{1-\frac{5}{16}c_1}\right)} & \frac{4}{9} \cdot \frac{1-\frac{c_1}{2}}{1-\frac{5}{16}c_1} e^{-\frac{t}{3N}\left(\frac{1+\frac{c_1}{16}}{1-\frac{5}{16}c_1}\right)} \\
1 - \frac{1-\frac{c_1}{2}}{1-\frac{5}{16}c_1} e^{-\frac{t}{3N}\left(\frac{1+\frac{c_1}{16}}{1-\frac{5}{16}c_1}\right)} & 0 & 0 & 0 & \frac{1}{9} \cdot \frac{1-\frac{c_1}{2}}{1-\frac{5}{16}c_1} e^{-\frac{t}{3N}\left(\frac{1+\frac{c_1}{16}}{1-\frac{5}{16}c_1}\right)} & \frac{4}{9} \cdot \frac{1-\frac{c_1}{2}}{1-\frac{5}{16}c_1} e^{-\frac{t}{3N}\left(\frac{1+\frac{c_1}{16}}{1-\frac{5}{16}c_1}\right)} & \frac{4}{9} \cdot \frac{1-\frac{c_1}{2}}{1-\frac{5}{16}c_1} e^{-\frac{t}{3N}\left(\frac{1+\frac{c_1}{16}}{1-\frac{5}{16}c_1}\right)} \\
1 - \frac{1-\frac{c_1}{2}}{1-\frac{5}{16}c_1} e^{-\frac{t}{3N}\left(\frac{1+\frac{c_1}{16}}{1-\frac{5}{16}c_1}\right)} & 0 & 0 & 0 & \frac{1}{9} \cdot \frac{1-\frac{c_1}{2}}{1-\frac{5}{16}c_1} e^{-\frac{t}{3N}\left(\frac{1+\frac{c_1}{16}}{1-\frac{5}{16}c_1}\right)} & \frac{4}{9} \cdot \frac{1-\frac{c_1}{2}}{1-\frac{5}{16}c_1} e^{-\frac{t}{3N}\left(\frac{1+\frac{c_1}{16}}{1-\frac{5}{16}c_1}\right)} & \frac{4}{9} \cdot \frac{1-\frac{c_1}{2}}{1-\frac{5}{16}c_1} e^{-\frac{t}{3N}\left(\frac{1+\frac{c_1}{16}}{1-\frac{5}{16}c_1}\right)} \\
1 - e^{-\frac{t}{3N}\left(\frac{1+\frac{c_1}{16}}{1-\frac{5}{16}c_1}\right)} & 0 & 0 & 0 & \frac{1}{9} e^{-\frac{t}{3N}\left(\frac{1+\frac{c_1}{16}}{1-\frac{5}{16}c_1}\right)} & \frac{4}{9} e^{-\frac{t}{3N}\left(\frac{1+\frac{c_1}{16}}{1-\frac{5}{16}c_1}\right)} & \frac{4}{9} e^{-\frac{t}{3N}\left(\frac{1+\frac{c_1}{16}}{1-\frac{5}{16}c_1}\right)} \\
1 - e^{-\frac{t}{3N}\left(\frac{1+\frac{c_1}{16}}{1-\frac{5}{16}c_1}\right)} & 0 & 0 & 0 & \frac{1}{9} e^{-\frac{t}{3N}\left(\frac{1+\frac{c_1}{16}}{1-\frac{5}{16}c_1}\right)} & \frac{4}{9} e^{-\frac{t}{3N}\left(\frac{1+\frac{c_1}{16}}{1-\frac{5}{16}c_1}\right)} & \frac{4}{9} e^{-\frac{t}{3N}\left(\frac{1+\frac{c_1}{16}}{1-\frac{5}{16}c_1}\right)} \\
1 - e^{-\frac{t}{3N}\left(\frac{1+\frac{c_1}{16}}{1-\frac{5}{16}c_1}\right)} & 0 & 0 & 0 & \frac{1}{9} e^{-\frac{t}{3N}\left(\frac{1+\frac{c_1}{16}}{1-\frac{5}{16}c_1}\right)} & \frac{4}{9} e^{-\frac{t}{3N}\left(\frac{1+\frac{c_1}{16}}{1-\frac{5}{16}c_1}\right)} & \frac{4}{9} e^{-\frac{t}{3N}\left(\frac{1+\frac{c_1}{16}}{1-\frac{5}{16}c_1}\right)}
\end{pmatrix}. \tag{2.21}$$

We are concerned with transitions from each of the various states to coalescence (state 0). The first column of $\Pi(t)$ gives the limiting cumulative distribution functions for the time to the most recent common ancestor for two lineages *within* an individual (state 1) and two lineages *between* individuals (states 3, 4 and 5):

$$F_{T_f}(t) = F_U(t) = 1 - \frac{1-\frac{c_1}{2}}{1-\frac{5}{16}c_1} e^{-\frac{t}{3N}\left(\frac{1+\frac{c_1}{16}}{1-\frac{5}{16}c_1}\right)}, \tag{2.22}$$

$$F_{V_{mm}}(t) = F_{V_{mf}}(t) = F_{V_{ff}}(t) = 1 - e^{-\frac{t}{3N}\left(\frac{1+\frac{c_1}{16}}{1-\frac{5}{16}c_1}\right)}. \tag{2.23}$$

To compute expectations, recalling that for $X > 0$, $\mathbb{E}[X] = \int_0^\infty [1 - F_X(x)]\, dx$, we find

$$\mathbb{E}[T_f] = E[U] = 3N\left(\frac{1-\frac{c_1}{2}}{1+\frac{c_1}{16}}\right), \tag{2.24}$$

$$\mathbb{E}[V_{mm}] = \mathbb{E}[V_{mf}] = \mathbb{E}[V_{ff}] = 3N\left(\frac{1-\frac{5}{16}c_1}{1+\frac{c_1}{16}}\right). \tag{2.25}$$

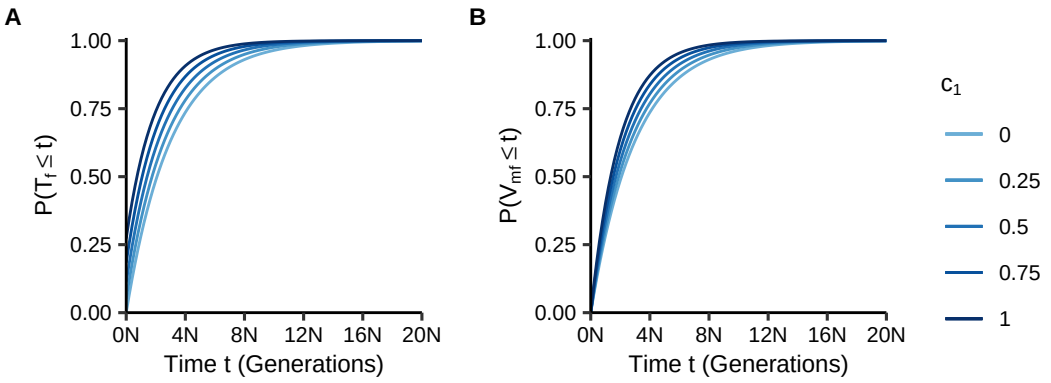

Figure 2.5: Cumulative distributions of coalescence times as functions of the number of generations t and the fraction of matrilateral-parallel mating pairs c_1. **(A)** Coalescence time within individuals, $P(T_f \leq t)$, Eq. 2.22. **(B)** Coalescence time between individuals, $P(V_{mf} \leq t)$, Eq. 2.23.

Eqs. 2.24 and 2.25 are the same as Eqs. 39 and 40 from Cotter *et al.* (2021). Eqs. 2.22 and 2.23 are plotted in Figure 2.5.

Matrilateral cross

In the matrilateral-cross case, a union occurs between a male and his mother's brother's daughter. This case is similar to the matrilateral-parallel case. With probability $c_1/2$, two lineages in state 2_1 trace to the shared grandparental pair. They coalesce with probability $\frac{1}{4}$ (state 0), they are in the shared grandmother with probability $\frac{1}{4}$ (state 1), and they are in opposite individuals of the grandparental mating pair with probability $\frac{1}{2}$ (state 2_0).

With probability $c_1/2$, two lineages in state 2_1 do not trace to the shared grandparental pair. Conditional on the lineages not both tracing to the shared grandparental pair, they are in two males (state 3), a male and a female (state 4) or two females (state 5), with probabilities $\frac{1}{4}$, $\frac{1}{2}$, and $\frac{1}{4}$, respectively. Finally, with probability $1 - c_1$, two lineages are not ancestral to a consanguineous mating pair.

In this case, they follow the same pattern as enumerated for the patrilateral-parallel case. The transition matrix is

$$\Pi_N = \begin{array}{c} \\ 0 \\ 1 \\ 2_0 \\ 2_1 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccccc} 0 & 1 & 2_0 & 2_1 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ \frac{c_1}{8} & \frac{c_1}{8} & \frac{c_1}{4} & 0 & \frac{1}{8} & \frac{1}{2}-\frac{c_1}{4} & \frac{3}{8}-\frac{c_1}{4} \\ \frac{1}{2N} & \frac{1}{2N} & 0 & 0 & 0 & 0 & 1-\frac{1}{N} \\ \frac{1}{4N} & \frac{1}{4N} & \frac{1}{2N} & 0 & 0 & \frac{1-\frac{1}{N}}{2} & \frac{1-\frac{1}{N}}{2} \\ \frac{3}{8N} & \frac{1}{8N} & \frac{1}{2N} & 0 & \frac{1-\frac{1}{N}}{4} & \frac{1-\frac{1}{N}}{2} & \frac{1-\frac{1}{N}}{4} \end{pmatrix} \end{array}. \tag{2.26}$$

We separate the "fast" and "slow" transitions as before (Eq. 2.1):

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ \frac{c_1}{8} & \frac{c_1}{8} & \frac{c_1}{4} & 0 & \frac{1}{8} & \frac{1}{2}-\frac{c_1}{4} & \frac{3}{8}-\frac{c_1}{4} \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 0 & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 & -1 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} & 0 & 0 & -\frac{1}{2} & -\frac{1}{2} \\ \frac{3}{8} & \frac{1}{8} & \frac{1}{2} & 0 & -\frac{1}{4} & -\frac{1}{2} & -\frac{1}{4} \end{pmatrix}. \tag{2.27}$$

Using the method of Appendix A, we solve for the stationary distribution of the "fast" process:

$$\mathbf{P} = \lim_{r\to\infty} \mathbf{A}^r = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{c_1}{8-3c_1} & 0 & 0 & 0 & \frac{1}{9}\left(\frac{8-4c_1}{8-3c_1}\right) & \frac{4}{9}\left(\frac{8-4c_1}{8-3c_1}\right) & \frac{4}{9}\left(\frac{8-4c_1}{8-3c_1}\right) \\ \frac{c_1}{8-3c_1} & 0 & 0 & 0 & \frac{1}{9}\left(\frac{8-4c_1}{8-3c_1}\right) & \frac{4}{9}\left(\frac{8-4c_1}{8-3c_1}\right) & \frac{4}{9}\left(\frac{8-4c_1}{8-3c_1}\right) \\ \frac{c_1}{8-3c_1} & 0 & 0 & 0 & \frac{1}{9}\left(\frac{8-4c_1}{8-3c_1}\right) & \frac{4}{9}\left(\frac{8-4c_1}{8-3c_1}\right) & \frac{4}{9}\left(\frac{8-4c_1}{8-3c_1}\right) \\ 0 & 0 & 0 & 0 & \frac{1}{9} & \frac{4}{9} & \frac{4}{9} \\ 0 & 0 & 0 & 0 & \frac{1}{9} & \frac{4}{9} & \frac{4}{9} \\ 0 & 0 & 0 & 0 & \frac{1}{9} & \frac{4}{9} & \frac{4}{9} \end{pmatrix}. \tag{2.28}$$

As before, using $\mathbf{G} = \mathbf{PBP}$, we calculate the matrix exponential, $e^{t\mathbf{G}}$, using the method of Appendix B. We then obtain $\Pi(t)$ from Eq. 2.3, converting t back to units of N generations:

$$\Pi(t) = \mathbf{P}e^{t\mathrm{G}} =$$

$$\begin{pmatrix} 1 & 0\,0\,0 & 0 & 0 & 0 \\ 1-\frac{1-\frac{c_1}{2}}{1-\frac{3}{8}c_1}e^{-\frac{t}{3N}\left(\frac{1-\frac{c_1}{8}}{1-\frac{3}{8}c_1}\right)} & 0\,0\,0 & \frac{1}{9}\cdot\frac{1-\frac{c_1}{2}}{1-\frac{3}{8}c_1}e^{-\frac{t}{3N}\left(\frac{1-\frac{c_1}{8}}{1-\frac{3}{8}c_1}\right)} & \frac{4}{9}\cdot\frac{1-\frac{c_1}{2}}{1-\frac{3}{8}c_1}e^{-\frac{t}{3N}\left(\frac{1-\frac{c_1}{8}}{1-\frac{3}{8}c_1}\right)} & \frac{4}{9}\cdot\frac{1-\frac{c_1}{2}}{1-\frac{3}{8}c_1}e^{-\frac{t}{3N}\left(\frac{1-\frac{c_1}{8}}{1-\frac{3}{8}c_1}\right)} \\ 1-\frac{1-\frac{c_1}{2}}{1-\frac{3}{8}c_1}e^{-\frac{t}{3N}\left(\frac{1-\frac{c_1}{8}}{1-\frac{3}{8}c_1}\right)} & 0\,0\,0 & \frac{1}{9}\cdot\frac{1-\frac{c_1}{2}}{1-\frac{3}{8}c_1}e^{-\frac{t}{3N}\left(\frac{1-\frac{c_1}{8}}{1-\frac{3}{8}c_1}\right)} & \frac{4}{9}\cdot\frac{1-\frac{c_1}{2}}{1-\frac{3}{8}c_1}e^{-\frac{t}{3N}\left(\frac{1-\frac{c_1}{8}}{1-\frac{3}{8}c_1}\right)} & \frac{4}{9}\cdot\frac{1-\frac{c_1}{2}}{1-\frac{3}{8}c_1}e^{-\frac{t}{3N}\left(\frac{1-\frac{c_1}{8}}{1-\frac{3}{8}c_1}\right)} \\ 1-\frac{1-\frac{c_1}{2}}{1-\frac{3}{8}c_1}e^{-\frac{t}{3N}\left(\frac{1-\frac{c_1}{8}}{1-\frac{3}{8}c_1}\right)} & 0\,0\,0 & \frac{1}{9}\cdot\frac{1-\frac{c_1}{2}}{1-\frac{3}{8}c_1}e^{-\frac{t}{3N}\left(\frac{1-\frac{c_1}{8}}{1-\frac{3}{8}c_1}\right)} & \frac{4}{9}\cdot\frac{1-\frac{c_1}{2}}{1-\frac{3}{8}c_1}e^{-\frac{t}{3N}\left(\frac{1-\frac{c_1}{8}}{1-\frac{3}{8}c_1}\right)} & \frac{4}{9}\cdot\frac{1-\frac{c_1}{2}}{1-\frac{3}{8}c_1}e^{-\frac{t}{3N}\left(\frac{1-\frac{c_1}{8}}{1-\frac{3}{8}c_1}\right)} \\ 1-e^{-\frac{t}{3N}\left(\frac{1-\frac{c_1}{8}}{1-\frac{3}{8}c_1}\right)} & 0\,0\,0 & \frac{1}{9}e^{-\frac{t}{3N}\left(\frac{1-\frac{c_1}{8}}{1-\frac{3}{8}c_1}\right)} & \frac{4}{9}e^{-\frac{t}{3N}\left(\frac{1-\frac{c_1}{8}}{1-\frac{3}{8}c_1}\right)} & \frac{4}{9}e^{-\frac{t}{3N}\left(\frac{1-\frac{c_1}{8}}{1-\frac{3}{8}c_1}\right)} \\ 1-e^{-\frac{t}{3N}\left(\frac{1-\frac{c_1}{8}}{1-\frac{3}{8}c_1}\right)} & 0\,0\,0 & \frac{1}{9}e^{-\frac{t}{3N}\left(\frac{1-\frac{c_1}{8}}{1-\frac{3}{8}c_1}\right)} & \frac{4}{9}e^{-\frac{t}{3N}\left(\frac{1-\frac{c_1}{8}}{1-\frac{3}{8}c_1}\right)} & \frac{4}{9}e^{-\frac{t}{3N}\left(\frac{1-\frac{c_1}{8}}{1-\frac{3}{8}c_1}\right)} \\ 1-e^{-\frac{t}{3N}\left(\frac{1-\frac{c_1}{8}}{1-\frac{3}{8}c_1}\right)} & 0\,0\,0 & \frac{1}{9}e^{-\frac{t}{3N}\left(\frac{1-\frac{c_1}{8}}{1-\frac{3}{8}c_1}\right)} & \frac{4}{9}e^{-\frac{t}{3N}\left(\frac{1-\frac{c_1}{8}}{1-\frac{3}{8}c_1}\right)} & \frac{4}{9}e^{-\frac{t}{3N}\left(\frac{1-\frac{c_1}{8}}{1-\frac{3}{8}c_1}\right)} \end{pmatrix}. \tag{2.29}$$

We extract the cumulative distribution functions from the first column of the matrix, finding

$$F_{T_f}(t) = F_U(t) = 1 - \frac{1-\frac{c_1}{2}}{1-\frac{3}{8}c_1}e^{-\frac{t}{3N}\left(\frac{1-\frac{c_1}{8}}{1-\frac{3}{8}c_1}\right)}, \tag{2.30}$$

$$F_{V_{mm}}(t) = F_{V_{mf}}(t) = F_{V_{ff}}(t) = 1 - e^{-\frac{t}{3N}\left(\frac{1-\frac{c_1}{8}}{1-\frac{3}{8}c_1}\right)}. \tag{2.31}$$

Solving for the expectations of these distributions, recalling that for $X > 0$, $\mathbb{E}[X] = \int_0^\infty [1 - F_X(x)]\,dx$, we find

$$\mathbb{E}[T_f] = E[U] = 3N\left(\frac{1-\frac{c_1}{2}}{1-\frac{c_1}{8}}\right),, \tag{2.32}$$

$$\mathbb{E}[V_{mm}] = \mathbb{E}[V_{mf}] = \mathbb{E}[V_{ff}] = 3N\left(\frac{1-\frac{3}{8}c_1}{1-\frac{c_1}{8}}\right). \tag{2.33}$$

Eqs. 2.32 and 2.33 are the same as Eqs. 47 and 48 from Cotter *et al.* (2021). Eqs. 2.30 and 2.31 are plotted in Figure 2.6.

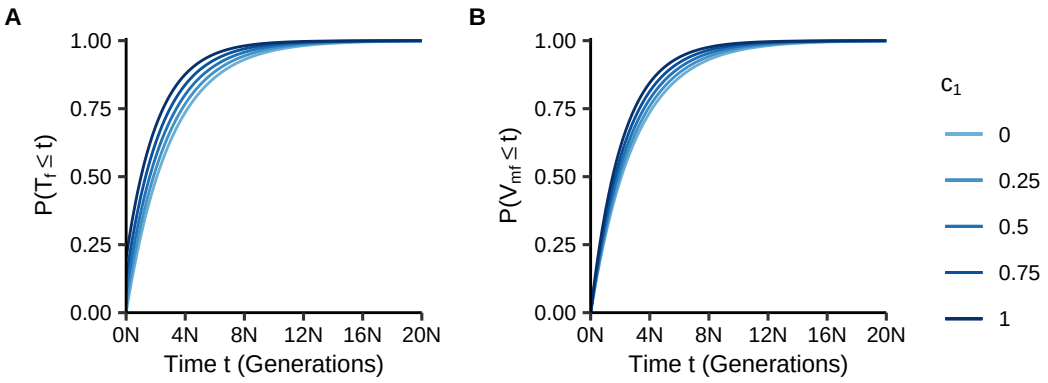

Figure 2.6: Cumulative distributions of coalescence times as functions of the number of generations t and the fraction of matrilateral-cross mating pairs c_1. **(A)** Coalescence time within individuals, $P(T_f \leq t)$, Eq. 2.30. **(B)** Coalescence time between individuals, $P(V_{mf} \leq t)$, Eq. 2.31.

Bilateral parallel

Having considered the four possible types of first-cousin consanguinity, we can also consider the two bilateral cases, in which a mating pair are cousins through both sets of grandparents. In bilateral-parallel first-cousin consanguinity, a union occurs between a male and a female who is both his mother's sister's daughter *and* his father's brother's daughter. We can consider this case to be a combination of the matrilateral-parallel case and the patrilateral-parallel case. In state 2_1, when the two lineages are ancestral to a bilateral-parallel mating pair, the male's lineage must transition through his mother because he cannot inherit an X chromosome from his father. Because there is no way for the lineages to transition through the patrilateral-parallel grandparental pair, the transitions in state 2_1 follow from the transitions for a matrilateral-parallel pair only. In the case of bilateral-parallel first-cousin consanguinity, the transition matrix thus has the form given for matrilateral-parallel first-cousin consanguinity in Eq. 2.18. The bilateral-parallel case then also shares the same cumulative distribution functions given in Eqs. 2.22 and 2.23.

Bilateral cross

Bilateral-cross first-cousin consanguinity occurs when a male shares a union with a female who is both his father's sister's daughter and his mother's brother's daughter. This case can be considered

to be a combination of matrilateral-cross and patrilateral-cross first-cousin consanguinity. The ancestral lineages cannot travel through the patrilateral-cross pair, and the transitions follow those for matrilateral-cross consanguinity. The transition matrix (Eq. 2.26) and cumulative distribution functions (Eqs. 2.30 and 2.31) follow similarly.

Mixture of first-cousin mating types

We next examine a population that possesses a mixture of all four unilateral first-cousin mating types. To determine the transition matrix, it suffices to determine the transition probabilities from state 2_1.

Recall that two lineages in state 2_1 are in two individuals ancestral to a mating pair that might or might not be consanguineous. With probability c_{pp}, this mating pair is a patrilateral-parallel first-cousin pair, with probability c_{pc} it is a patrilateral-cross first-cousin pair, with probability c_{mp} it is a matrilateral-parallel first-cousin pair, and with probability c_{mc} it is a matrilateral-cross first-cousin pair. If the mating pair is a first-cousin pair of a particular one of the four types, then transitions out of state 2_1 will match those derived for the associated case.

We can view the transition probabilities out of state 2_1 as a weighted combination of the transitions that each of these first-cousin cases makes when considered on its own. For example, in the case of coalescence (transition to state 0), two lineages in state 2_1 coalesce with probability $\frac{3}{16}$ for a matrilateral-parallel first-cousin pair (rate c_{mp}) and $\frac{1}{8}$ for a matrilateral-cross first-cousin pair (rate c_{mc}). Because patrilateral-parallel and -cross consanguinity do not affect transitions from state 2_1, corresponding rates c_{pp} and c_{pc} do not influence the transition probability to state 0. Combining all four cases, the transition probability from state 2_1 to state 0 is $\frac{3}{16}c_{mp} + \frac{1}{8}c_{mc}$. For transitions from state 2_1 to states 0, 1, and 2_0, the probabilities are obtained by summing corresponding terms in the matrices for the various types of unilateral first-cousin mating (Eqs. 2.12, 2.18, and 2.26).

For the transitions from state 2_1 to states 3, 4, and 5 (two lineages between individuals), consanguinity acts to reduce the probabilities. The probabilities in the case of patrilateral parallel consanguinity (Eq. 2.12) represent a null effect of no consanguinity. The c_{mp} and c_{mc} terms (Eqs. 2.18 and 2.26) reduce the probabilities of transitioning to states 3, 4, and 5 (while inflating the

0, 1, and 2_0 transitions). For state 3, for example, the null transition probability is $\frac{1}{8}$. Matrilateral-parallel consanguinity reduces this transition probability by $c_{mp}/8$, giving a combined transition probability of $\frac{1}{8} - c_{mp}/8$; matrilateral-cross consanguinity has no effect on this transition.

We proceed similarly to combine the remaining probabilities from the four unilateral first-cousin mating types to produce the transitions for state 2_1. The transition matrix is

$$\Pi_N = \begin{array}{c} \\ 0 \\ 1 \\ 2_0 \\ 2_1 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccccc} 0 & 1 & 2_0 & 2_1 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ \frac{3c_{mp}}{16} + \frac{c_{mc}}{8} & \frac{c_{mp}}{16} + \frac{c_{mc}}{8} & \frac{c_{mp}}{4} + \frac{c_{mc}}{4} & 0 & \frac{1}{8} - \frac{c_{mp}}{8} & \frac{1}{2} - \frac{c_{mp}}{4} - \frac{c_{mc}}{4} & \frac{3}{8} - \frac{c_{mp}}{8} - \frac{c_{mc}}{4} \\ \frac{1}{2N} & \frac{1}{2N} & 0 & 0 & 0 & 0 & 1 - \frac{1}{N} \\ \frac{1}{4N} & \frac{1}{4N} & \frac{1}{2N} & 0 & 0 & \frac{1-\frac{1}{N}}{2} & \frac{1-\frac{1}{N}}{2} \\ \frac{3}{8N} & \frac{1}{8N} & \frac{1}{2N} & 0 & \frac{1-\frac{1}{N}}{4} & \frac{1-\frac{1}{N}}{2} & \frac{1-\frac{1}{N}}{4} \end{pmatrix} \end{array}. \tag{2.34}$$

Matrices **A** and **B** follow from Eq. 2.1 and take the same form as those given for the matrilateral cases with state 2_1 in matrix **A** (Eqs. 2.19 and 2.27), now adopting the new combinations of transition probabilities. We solve for the stationary distribution of the "fast" transitions using the method of Appendix A:

$$\mathbf{P} = \lim_{r\to\infty} \mathbf{A}^r = \begin{pmatrix} 1 & 0\,0\,0 & 0 & 0 & 0 \\ \frac{\frac{3}{16}c_{mp}+\frac{c_{mc}}{8}}{1-\frac{5}{16}c_{mp}-\frac{3}{8}c_{mc}} & 0\,0\,0 & \frac{1}{9}\left(\frac{1-\frac{c_{mp}}{2}-\frac{c_{mc}}{2}}{1-\frac{5}{16}c_{mp}-\frac{3}{8}c_{mc}}\right) & \frac{4}{9}\left(\frac{1-\frac{c_{mp}}{2}-\frac{c_{mc}}{2}}{1-\frac{5}{16}c_{mp}-\frac{3}{8}c_{mc}}\right) & \frac{4}{9}\left(\frac{1-\frac{c_{mp}}{2}-\frac{c_{mc}}{2}}{1-\frac{5}{16}c_{mp}-\frac{3}{8}c_{mc}}\right) \\ \frac{\frac{3}{16}c_{mp}+\frac{c_{mc}}{8}}{1-\frac{5}{16}c_{mp}-\frac{3}{8}c_{mc}} & 0\,0\,0 & \frac{1}{9}\left(\frac{1-\frac{c_{mp}}{2}-\frac{c_{mc}}{2}}{1-\frac{5}{16}c_{mp}-\frac{3}{8}c_{mc}}\right) & \frac{4}{9}\left(\frac{1-\frac{c_{mp}}{2}-\frac{c_{mc}}{2}}{1-\frac{5}{16}c_{mp}-\frac{3}{8}c_{mc}}\right) & \frac{4}{9}\left(\frac{1-\frac{c_{mp}}{2}-\frac{c_{mc}}{2}}{1-\frac{5}{16}c_{mp}-\frac{3}{8}c_{mc}}\right) \\ \frac{\frac{3}{16}c_{mp}+\frac{c_{mc}}{8}}{1-\frac{5}{16}c_{mp}-\frac{3}{8}c_{mc}} & 0\,0\,0 & \frac{1}{9}\left(\frac{1-\frac{c_{mp}}{2}-\frac{c_{mc}}{2}}{1-\frac{5}{16}c_{mp}-\frac{3}{8}c_{mc}}\right) & \frac{4}{9}\left(\frac{1-\frac{c_{mp}}{2}-\frac{c_{mc}}{2}}{1-\frac{5}{16}c_{mp}-\frac{3}{8}c_{mc}}\right) & \frac{4}{9}\left(\frac{1-\frac{c_{mp}}{2}-\frac{c_{mc}}{2}}{1-\frac{5}{16}c_{mp}-\frac{3}{8}c_{mc}}\right) \\ 0 & 0\,0\,0 & \frac{1}{9} & \frac{4}{9} & \frac{4}{9} \\ 0 & 0\,0\,0 & \frac{1}{9} & \frac{4}{9} & \frac{4}{9} \\ 0 & 0\,0\,0 & \frac{1}{9} & \frac{4}{9} & \frac{4}{9} \end{pmatrix}. \tag{2.35}$$

Once again, using $\mathbf{G} = \mathbf{PBP}$, we obtain the matrix exponential, $e^{t\mathbf{G}}$, using the method of Appendix B. We then compute $\Pi(t)$ with Eq. 2.3, converting t back into units of N generations. The resulting matrix is structured in such a way that we can write:

$$\Pi(t) = \mathbf{P}e^{t\mathbf{G}} = \begin{pmatrix} 1 & 0\,0\,0 & 0 & 0 & 0 \\ 1-RE & 0\,0\,0 & \frac{1}{9}RE & \frac{4}{9}RE & \frac{4}{9}RE \\ 1-RE & 0\,0\,0 & \frac{1}{9}RE & \frac{4}{9}RE & \frac{4}{9}RE \\ 1-RE & 0\,0\,0 & \frac{1}{9}RE & \frac{4}{9}RE & \frac{4}{9}RE \\ 1-E & 0\,0\,0 & \frac{1}{9}E & \frac{4}{9}E & \frac{4}{9}E \\ 1-E & 0\,0\,0 & \frac{1}{9}E & \frac{4}{9}E & \frac{4}{9}E \\ 1-E & 0\,0\,0 & \frac{1}{9}E & \frac{4}{9}E & \frac{4}{9}E \end{pmatrix}, \tag{2.36}$$

where

$$R = \frac{1-\frac{c_{mp}}{2}-\frac{c_{mc}}{2}}{1-\frac{5}{16}c_{mp}-\frac{3}{8}c_{mc}}, \qquad E = e^{-\frac{t}{3N}\left(\frac{1+\frac{c_{mp}}{16}-\frac{c_{mc}}{8}}{1-\frac{5}{16}c_{mp}-\frac{3}{8}c_{mc}}\right)}.$$

In the matrix in Eq. 2.36, the first column represents transitions to coalescence. We extract from this column the cumulative distribution functions for time to coalescence for two lineages *within* an individual (state 1) and two lineages *between* individuals (states 3, 4, and 5):

$$F_{T_f}(t) = F_U(t) = 1 - \frac{1-\frac{c_{mp}}{2}-\frac{c_{mc}}{2}}{1-\frac{5}{16}c_{mp}-\frac{3}{8}c_{mc}} e^{-\frac{t}{3N}\left(\frac{1+\frac{c_{mp}}{16}-\frac{c_{mc}}{8}}{1-\frac{5}{16}c_{mp}-\frac{3}{8}c_{mc}}\right)}, \tag{2.37}$$

$$F_{V_{mm}}(t) = F_{V_{mf}}(t) = F_{V_{ff}}(t) = 1 - e^{-\frac{t}{3N}\left(\frac{1+\frac{c_{mp}}{16}-\frac{c_{mc}}{8}}{1-\frac{5}{16}c_{mp}-\frac{3}{8}c_{mc}}\right)}. \tag{2.38}$$

For the expectations of these distributions, recalling that for $X > 0$, $\mathbb{E}[X] = \int_0^\infty [1 - F_X(x)]\,dx$, we have

$$\mathbb{E}[T_f] = E[U] = 3N\left(\frac{1-\frac{c_{mp}}{2}-\frac{c_{mc}}{2}}{1+\frac{c_{mp}}{16}-\frac{c_{mc}}{8}}\right), \tag{2.39}$$

$$\mathbb{E}[V_{mm}] = \mathbb{E}[V_{mf}] = \mathbb{E}[V_{ff}] = 3N\left(\frac{1-\frac{5}{16}c_{mp}-\frac{3}{8}c_{mc}}{1+\frac{c_{mp}}{16}-\frac{c_{mc}}{8}}\right). \tag{2.40}$$

2.3.3 Comparisons

Limiting distribution versus exact distribution

Under the mixture model, to see how well the limiting distribution of coalescence times approximates the exact distribution, we perform simulations. In particular, for fixed values of the number of mating pairs N and rates of matrilateral-parallel (c_{mp}) and matrilateral-cross (c_{mc}) first-cousin mating, we simulate 10,000 realizations of the Markov chain in Eq. 2.34 to produce an empirical cumulative distribution function (CDF) of coalescence times for lineage pairs *within* and *between* individuals. This procedure amounts to simulating a distribution of the time to the most recent common ancestor (the time it takes to hit state 0) starting in either state 1 (within an individual) or state 4 (between individuals).

Figure 2.7 plots the simulated empirical CDFs alongside the limiting CDFs presented in Eqs. 2.37 and 2.38. Conducting these simulations for different values of the number of mating pairs N, we see that the limiting cumulative distribution functions tend to be slightly inflated compared to those simulated from the Markov chain; the limiting CDF tends to reach a specified probability before it is reached in the simulation. This effect is most visible for the smallest value of N, $N = 10$; as N increases, the limiting distribution functions (Eqs. 2.37 and 2.38) closely approximate the simulated, empirical distributions.

X chromosome versus autosomes

Each of the limiting distributions for coalescence times for lineages from separate mating pairs, both for single types of first-cousin consanguinity and for a superposition of multiple types, possesses a particular structure: an exponential CDF whose rate is the product of the population size and a reduction by a factor that accounts for consanguinity. We now examine these limiting CDFs for the X chromosome in relation to corresponding CDFs for autosomes. The autosomal coalescence time distributions under first-cousin consanguinity are obtained in Appendix C as a special case of the nth cousin mating model of Severson *et al.* (2019). Here, we calculate the ratio of the expected time to coalescence for the X chromosome (Eqs. 2.39 and 2.40) and for autosomes

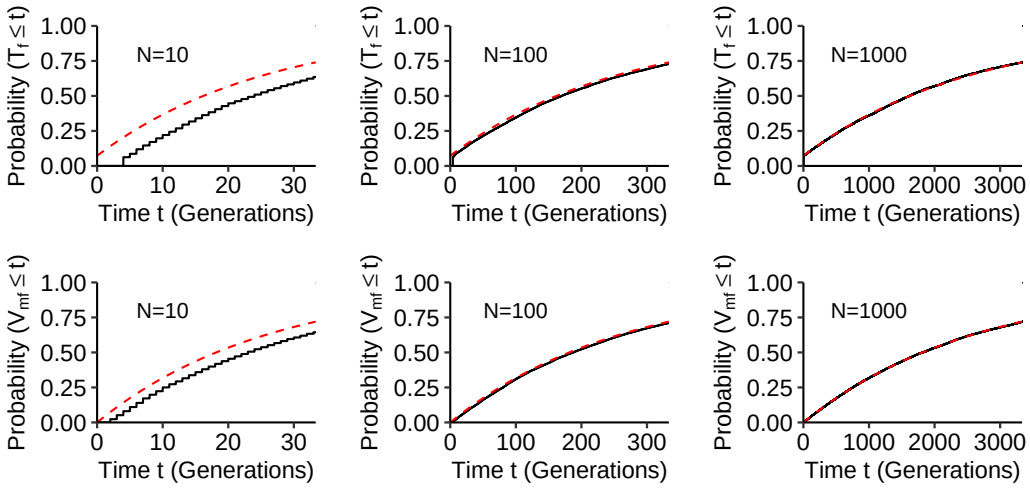

Figure 2.7: Cumulative distribution functions (CDFs) of coalescence times in a model with a mixture of types of consanguinity. The Markov chain is given in Eq. 2.34; we consider the case of $c_{mp} = 0.2$ and $c_{mc} = 0.2$ with each of three values for the number of mating pairs N. Dashed lines represent the limiting CDFs in Eqs. 2.37 and 2.38, and solid lines represent the simulated CDFs from 10,000 observations of the first-cousin mixture model (as described by the Markov chain in Eq. 2.34).

(Eqs. 2.47 and 2.48) within and between individuals, respectively, as we vary rates of matrilateral and patrilateral consanguinity (Figure 2.8).

We first consider the ratio of expected coalescence times on the X chromosome relative to the autosomes for pairs of lineages within individuals (Eq. 2.39/Eq. 2.47) as a function of patrilateral ($c_{pp} + c_{pc}$) and matrilateral-parallel (c_{mp}) consanguinity (Figure 2.8A). Because the expected coalescence time for two lineages on the X chromosome is a function of $3N$ and the corresponding autosomal mean depends on $4N$, in the absence of consanguinity, the null value of the ratio is $\frac{3}{4}$. The ratio achieves its minimum value of $\frac{8}{17}$, with a stronger effect of consanguinity in reducing X-chromosomal coalescence times relative to autosomal coalescence times, when we set c_{mp} to 1. It achieves its maximum value of 1, increasing X-chromosomal coalescence times compared to autosomal coalescence times, when instead we set $c_{pp} + c_{pc}$ to 1 (Figure 2.8A).

For the X:A ratio of between-individual expected coalescence times (Eq. 2.40/Eq. 2.48) as a function of patrilateral ($c_{pp} + c_{pc}$) and matrilateral-parallel (c_{mp}) consanguinity (Figure 2.8B), the minimum and maximum values differ less than for the within-individual case. The minimum exceeds $\frac{8}{17}$, equaling $\frac{132}{221}$, and is again reached at $c_{mp} = 1$. The maximum is less than 1, equaling $\frac{12}{13}$,

and is reached at $c_{pp} + c_{pc} = 1$. The minimum and maximum are less extreme than in the within-individual case, as consanguinity has less of an effect on reducing the expected coalescence times in the between-individual case, both for the X chromosome and for the autosomes.

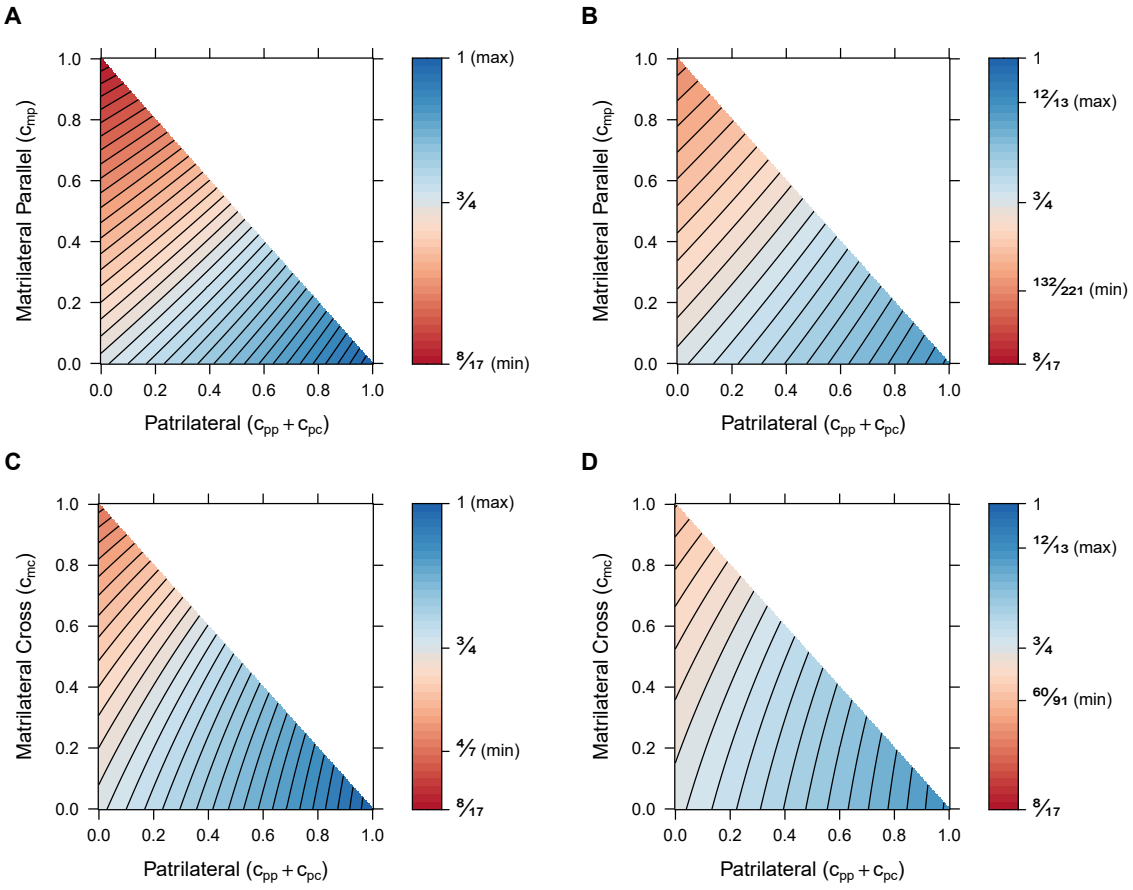

Figure 2.8: Ratios of X-chromosomal and autosomal mean coalescence times. Each point represents a ratio of coalescence times for a specified mixture of two types of consanguinity, depicted on the x and y axes. (A) Within individuals, matrilateral parallel and patrilateral consanguinity (Eq. 2.39/Eq. 2.47). (B) Between individuals, matrilateral parallel and patrilateral consanguinity (Eq. 2.40/Eq. 2.48). (C) Within individuals, matrilateral cross and patrilateral consanguinity (Eq. 2.39/Eq. 2.47). (D) Between individuals, matrilateral cross and patrilateral consanguinity (Eq. 2.40/Eq. 2.48). In each panel, the minimal ratio is indicated (obtained by setting matrilateral consanguinity to 1 and patrilateral consanguinity to 0), as is the maximum (obtained by setting matrilateral consanguinity to 0 and patrilateral consanguinity to 1). The value $\frac{3}{4}$ occurs with no consanguinity, located at the origin in each panel. Values *greater* than $\frac{3}{4}$ appear in blue, indicating combinations of parameter values that bring expected X chromosomal coalescence times closer to expected autosomal coalescence times. Values that reduce X chromosomal coalescence times to a greater extent than on autosomes, thereby shifting the ratio less than $\frac{3}{4}$, appear in red. Contour lines divide $[\frac{8}{17}, 1]$ into equal-sized intervals.

We next examine the X:A coalescence time ratio within individuals (Eq. 2.39/Eq. 2.47) as a function of patrilateral ($c_{pp} + c_{pc}$) and matrilateral-cross (c_{mc}) consanguinity (Figure 2.8C). The minimal ratio is slightly larger than in the matrilateral-parallel case, equaling $\frac{4}{7}$ at $c_{mc} = 1$. The maximum occurs at 1, the same value as the corresponding case with matrilateral-parallel in place of matrilateral-cross consanguinity, when $c_{pp} + c_{pc} = 1$. The slightly reduced range of values (i.e., the greater minimum) traces to the fact that the effect of matrilateral-cross consanguinity on X-chromosomal coalescence times is slightly weaker, producing a weaker reduction in coalescence times, than that of matrilateral-parallel consanguinity.

Finally, we analyze the X:A coalescence time ratio between individuals (Eq. 2.40/Eq. 2.48) as a function of patrilateral ($c_{pp} + c_{pc}$) and matrilateral-cross (c_{mc}) consanguinity (Figure 2.8D). The minimum occurs at $c_{mc} = 1$, equaling $\frac{60}{91}$. As in the corresponding matrilateral-parallel case, the maximum, achieved at $c_{pp} + c_{pc} = 1$, is $\frac{12}{13}$. As was seen within individuals, the range of permissible values is reduced relative to the matrilateral-parallel case, owing again to the weaker effect of matrilateral-cross consanguinity on X-chromosomal coalescence times.

2.4 Discussion

Extending our previous work on mean coalescence times on the X-chromosome in a consanguinity model, we have derived large-N limiting distributions for within-individual and between-individual X-chromosomal coalescence times under various types of first-cousin consanguinity. For between-individual coalescence times, each limiting distribution is exponential with a rate equal to the product of the number of X chromosomes and a reduction factor due to consanguinity (Eqs. 2.17, 2.23, and 2.31). Limiting distributions of within-individual coalescence times each have a point mass corresponding to instantaneous coalescence, and conditional on not coalescing instantaneously, are exponential (Eqs. 2.16, 2.22, and 2.30). These patterns also hold for limiting distributions of pairwise coalescence times for a model with a mixture of types of first-cousin consanguinity (Eqs. 2.37 and 2.38); in simulations, the limiting distributions under this superposition agree with exact distributions from the Markov chain (Eq. 2.34, Figure 2.7).

Our limiting distribution results can inform comparisons of the X chromosome with autosomes. The four types of first-cousin consanguinity have identical effects on the autosomes but vary in their effect on the X chromosome. Hence, a comparison of coalescence time distributions for the X chromosome and autosomes can be informative about features of consanguinity. Our results (Eqs. 2.37 and 2.38) directly show the effect of different rates and types of consanguinity on the distribution of X-chromosomal coalescence times. For example, increasing matrilateral-parallel and matrilateral-cross consanguinity decreases the ratio of X and autosomal mean coalescence times; increasing patrilateral-parallel and patrilateral-cross first-cousin consanguinity increases this ratio (Figure 2.8).

The results can be viewed in the setting of the idea of *coalescent effective size* (Nordborg and Krone, 2002; Sjödin *et al.*, 2005). As in other instances of the use of the separation-of-time-scales technique, the X-chromosomal consanguinity model behaves like a standard coalescent model, but with an altered effective size. Indeed, the model combines two phenomena for which the separation-of-time-scales approach has been separately used—consanguinity (Nordborg and Donnelly, 1997; Severson *et al.*, 2021) and a distinction between autosomes and the X chromosome (Ramachandran *et al.*, 2008). We have shown that even when combining multiple phenomena, the separation-of-time-scales approach can distill complicated demographic features into a standard coalescent with a rescaled coalescent effective size. Indeed, each of our consanguinity models both for the autosomes and for the X chromosome has a coalescent effective size that is a function of the number of chromosomes in the model ($4N$ or $3N$) and the rate and type of consanguinity in the population.

Consanguinity and other preferences for mate choice vary across human populations, often depending on cultural norms for certain types of consanguinity over others (Bittles, 2012). Because we have found that the different types of first-cousin consanguinity generate an observable effect on X chromosomal coalescence times, it is possible that features of coalescence times can be compared across populations to assess signatures of the different types of consanguinity. Such assessments can potentially capitalize on the inverse relationship between coalescence times and genomic sharing (Palamara *et al.*, 2012; Carmi *et al.*, 2014; Browning and Browning, 2015) to use genomic sharing patterns to uncover features of consanguinity (Arciero *et al.*, 2021).

We note that in our coalescent model, the consanguinity parameters are constant over a long-term. In human populations, features of consanguinity might change relatively rapidly, so that in data applications, it might not be appropriate to assume consanguinity parameters that persist over a large number of generations. If the relative ordering of the different types of consanguinity does not change, however, we expect that the model would continue to be informative.

In applications in which the model is sensible, a potential limitation is that exact rates of a given type of consanguinity might not be possible to infer from X-chromosomal data. For example, the effects of the matrilateral-cross first-cousin consanguinity parameter on the X-chromosomal coalescence times distributions are relatively small (Figure 2.6), so that given the difficulty in precisely estimating the coalescence times from data, the parameter might not be identifiable. By jointly considering X-chromosomal and autosomal data (Figure 2.8), however, more information will be available to conduct parameter inference.

Another limitation of our approach is that in formulating our model, we have disregarded higher-order consanguinity. While we have explicitly modeled first-cousin mating pairs, we have ignored the possibility that a pair has more distant consanguinity that is not captured in the model. It may be possible, however, to allow for such possibilities by incorporating into the nth cousin framework of Severson *et al.* (2021) sex-specific varieties of consanguinity at different levels of relationship.

Acknowledgments. We acknowledge support from United States–Israel Binational Science Foundation grant 2017024, NIH grant R01 HG005855, and NSF Graduate Research Fellowships to DJC and ALS.

2.5 Appendix A: Calculating the stationary distribution of the fast transition matrix

In this appendix, we solve for the stationary distribution of the "fast" transition matrix $\mathbf{A}$ in the case of sib mating on the X chromosome. This approach is also applied in the main text to obtain the stationary distribution of the fast transition matrix in other models.

First, we permute the states to rewrite matrix $\mathbf{A}$ in a canonical form. The matrix $\mathbf{A}$ in Eq. 2.5 has one absorbing state (state 0) and a closed communication class $C_1 = \{3, 4, 5\}$. We rearrange the matrix to take the form

$$\mathbf{D} = \begin{pmatrix} \mathbf{C} & \mathbf{0} \\ \mathbf{R} & \mathbf{Q} \end{pmatrix}, \tag{2.41}$$

listing the recurrent states before the transient states. Thus, square matrix $\mathbf{C}$ includes transitions between recurrent states (i.e., absorbing states and closed communication classes), and square matrix $\mathbf{Q}$ includes transitions between transient states. Matrix $\mathbf{R}$ includes transitions from the transient states to the recurrent states. For matrix $\mathbf{A}$ in Eq. 2.5, the recurrent states are state 0 (absorbing) and states 3, 4, and 5 (closed communication class C_1). The transient states are states 1 and 2. Permuting the matrix $\mathbf{A}$ to order the states 0, 3, 4, 5, 1, 2, we write

$$\mathbf{A}^* = \left(\begin{array}{cccc|cc} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 1 \\ \frac{c_0}{4} & 0 & \frac{1-c_0}{2} & \frac{1-c_0}{2} & \frac{c_0}{4} & \frac{c_0}{2} \end{array}\right).$$

We treat the closed communication class C_1 as a single absorbing state because any transitions made into C_1 transition infinitely often among the states it contains. We rewrite the transition matrix for the resulting Markov chain by collapsing the columns and rows corresponding to the

states in C_1. $\mathbf{A}^*$ becomes

$$\mathbf{A}^{**} = \left(\begin{array}{cc|cc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 \\ \frac{c_0}{4} & 1-c_0 & \frac{c_0}{4} & \frac{c_0}{2} \end{array}\right).$$

Matrix $\mathbf{A}^{**}$ now has the form in Eq. 2.41, with 2×2 submatrices and $\mathbf{C}$ as the identity matrix. Given a matrix in canonical form (Eq. 2.41 where $\mathbf{C}$ is the identity), the stationary distribution is given by

$$\lim_{r\to\infty} \mathbf{D}^{\mathbf{r}} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{NR} & \mathbf{0} \end{pmatrix},$$

where N is the fundamental matrix $\mathbf{N} = (\mathbf{I} - \mathbf{Q})^{-1}$ and $\mathbf{I}$ is the identity matrix (Kemeny and Snell, 1983, 3.3.7). The matrix $\mathbf{NR}$ defines for each pair consisting of a transient state and a recurrent state, the probability that from the transient state, the process reaches the recurrent state. For matrix $\mathbf{A}^{**}$, we have

$$\mathbf{P}^{**} = \lim_{r\to\infty} (\mathbf{A}^{**})^r = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \frac{c_0}{4-3c_0} & \frac{4-4c_0}{4-3c_0} & 0 & 0 \\ \frac{c_0}{4-3c_0} & \frac{4-4c_0}{4-3c_0} & 0 & 0 \end{pmatrix}.$$

To recover the stationary distribution of $\mathbf{A}^*$, we expand the absorbing state for the closed communication class C_1, replacing it with the stationary distribution for the irreducible 3×3 matrix associated with the class. We then weight the transient transition probabilities in $\mathbf{NR}$ by this stationary distribution. In other words, $\mathbf{NR}$ now gives, for each pair consisting of a transient and a recurrent state, the probability of the associated transition. Expanding the absorbing state for the closed communication class C_1, we get

$$\mathbf{P}^* = \lim_{r\to\infty} (\mathbf{A}^*)^r = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{9} & \frac{4}{9} & \frac{4}{9} & 0 & 0 \\ 0 & \frac{1}{9} & \frac{4}{9} & \frac{4}{9} & 0 & 0 \\ 0 & \frac{1}{9} & \frac{4}{9} & \frac{4}{9} & 0 & 0 \\ \frac{c_0}{4-3c_0} & \frac{1}{9}\left(\frac{4-4c_0}{4-3c_0}\right) & \frac{4}{9}\left(\frac{4-4c_0}{4-3c_0}\right) & \frac{4}{9}\left(\frac{4-4c_0}{4-3c_0}\right) & 0 & 0 \\ \frac{c_0}{4-3c_0} & \frac{1}{9}\left(\frac{4-4c_0}{4-3c_0}\right) & \frac{4}{9}\left(\frac{4-4c_0}{4-3c_0}\right) & \frac{4}{9}\left(\frac{4-4c_0}{4-3c_0}\right) & 0 & 0 \end{pmatrix}.$$

Finally, we permute $\mathbf{P}^*$ to recover $\mathbf{P}$ (Eq. 2.6).

2.6 Appendix B: The matrix exponential $e^{t\mathbf{G}}$

In this appendix, we obtain the matrix exponential, $e^{t\mathbf{G}}$, which is needed in calculating the large-N limit, $\Pi(t) = \mathbf{P}e^{t\mathbf{G}}$. The computations in this appendix are specific to sib mating on the X chromosome, but the same method can be applied to obtain the matrix exponential in the other models.

We first obtain the generator matrix from Eqs. 2.5 and 2.6:

$$\mathbf{G} = \mathbf{PBP} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{(4-4c_0)(4-c_0)}{3(4-3c_0)^2} & 0 & 0 & \frac{1}{9}\cdot\frac{(4-4c_0)(4-c_0)}{3(4-3c_0)^2} & \frac{4}{9}\cdot\frac{(4-4c_0)(4-c_0)}{3(4-3c_0)^2} & \frac{4}{9}\cdot\frac{(4-4c_0)(4-c_0)}{3(4-3c_0)^2} \\ \frac{(4-4c_0)(4-c_0)}{3(4-3c_0)^2} & 0 & 0 & \frac{1}{9}\cdot\frac{(4-4c_0)(4-c_0)}{3(4-3c_0)^2} & \frac{4}{9}\cdot\frac{(4-4c_0)(4-c_0)}{3(4-3c_0)^2} & \frac{4}{9}\cdot\frac{(4-4c_0)(4-c_0)}{3(4-3c_0)^2} \\ \frac{4-c_0}{3(4-3c_0)} & 0 & 0 & \frac{1}{9}\cdot\frac{4-c_0}{3(4-3c_0)} & \frac{4}{9}\cdot\frac{4-c_0}{3(4-3c_0)} & \frac{4}{9}\cdot\frac{4-c_0}{3(4-3c_0)} \\ \frac{4-c_0}{3(4-3c_0)} & 0 & 0 & \frac{1}{9}\cdot\frac{4-c_0}{3(4-3c_0)} & \frac{4}{9}\cdot\frac{4-c_0}{3(4-3c_0)} & \frac{4}{9}\cdot\frac{4-c_0}{3(4-3c_0)} \\ \frac{4-c_0}{3(4-3c_0)} & 0 & 0 & \frac{1}{9}\cdot\frac{4-c_0}{3(4-3c_0)} & \frac{4}{9}\cdot\frac{4-c_0}{3(4-3c_0)} & \frac{4}{9}\cdot\frac{4-c_0}{3(4-3c_0)} \end{pmatrix}. \tag{2.42}$$

The generator matrix, $\mathbf{G}$, has nonzero entries in the columns for state 0 and states 3, 4, and 5. It has the property

$$\mathbf{G}^2 = -\mathbf{G}\left[\frac{4-c_0}{3\,(4-3c_0)}\right].$$

For the constant $k = -(4-c_0)/[3\,(4-3c_0)]$, we can then recursively write

$$\mathbf{G}^n = k^{n-1}\mathbf{G}, \tag{2.43}$$

The matrix exponential, $e^{t\mathbf{G}} = \sum_{i=0}^{\infty} t^i \mathbf{G}^i / i!$, then equals

$$\begin{aligned} e^{t\mathbf{G}} &= \mathbf{I} + k^{-1}\mathbf{G}\sum_{i=1}^{\infty}\frac{t^i k^i}{i!} \\ &= \mathbf{I} - k^{-1}\left(1 - e^{kt}\right)\mathbf{G}. \end{aligned}$$

Converting t into units of N generations and multiplying by $\mathbf{P}$ (Eq. 2.6), we obtain $\mathbf{P}e^{t\mathbf{G}}$ as in Eq. 2.7. For each model studied, for the associated generator matrix $\mathbf{G}$, the corresponding quantity k that satisfies Eq. 2.43 appears in Table 2.1.

Table 2.1: Constants used in matrix exponentiation for consanguinity models.

Type of consanguineous mating	Chromosome	Section	Quantity k satisfying $G^n = k^{n-1}G$ for generator matrix G (Eq. 2.43)
Sibling	X	2.3.1	$-\frac{4-c_0}{3(4-3c_0)}$
Patrilateral-parallel first-cousin	X	2.3.2	$-\frac{1}{3}$
Patrilateral-cross first-cousin	X	2.3.2	$-\frac{1}{3}$
Matrilateral-parallel first-cousin	X	2.3.2	$-\frac{16+c_1}{3(16-5c_1)}$
Matrilateral-cross first-cousin	X	2.3.2	$-\frac{8-c_1}{3(8-3c_1)}$
Bilateral-parallel first-cousin	X	2.3.2	$-\frac{16+c_1}{3(16-5c_1)}$
Bilateral-cross first-cousin	X	2.3.2	$-\frac{8-c_1}{3(8-3c_1)}$
Superposition of first-cousin types	X	2.3.2	$-\frac{16+c_{mp}-2c_{mc}}{3(16-5c_{mp}-6c_{mc})}$
First-cousin	Autosomes	Appendix C	$-\frac{4}{16-3c_1}$

Note that c_{mp} and c_{mc} in Section 2.3.2 have the same meaning as c_1 in Sections 2.3.2 and 2.3.2, respectively.

2.7 Appendix C: Limiting distribution of autosomal coalescence times for first-cousin mating

Equation 46 of Severson *et al.* (2021) gives a limiting distribution of autosomal coalescence times for a model with a superposition of levels of cousin mating, up to nth cousins. In order to recover first-cousin mating on the autosomes to compare to our X-chromosomal results, we use the special case of this nth cousin model, where the rate of sibling mating c_0 is 0 and the rate of first-cousin mating is c_1, stopping at first cousins. This special case produces the following transition matrix, where state 0 is still coalescence, state 1 is two lineages in an individual, state 2_0 is two lineages in opposite individuals of a mating pair, state 2_1 is two lineages in two individuals one generation ancestral to a mating pair, and state 3 is two lineages in two individuals in different mating pairs:

$$\Pi_N = \begin{array}{c} \\ 0 \\ 1 \\ 2_0 \\ 2_1 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2_0 & 2_1 & 3 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ \frac{c_1}{16} & \frac{c_1}{16} & \frac{c_1}{8} & 0 & 1-\frac{c_1}{4} \\ \frac{1}{4N} & \frac{1}{4N} & \frac{1}{2N} & 0 & 1-\frac{1}{N} \end{pmatrix} \end{array}. \tag{2.44}$$

Note here that there is no need to use a two-sex model, as for autosomes, states referring to two males, a male and a female, and two females simply collapse into the combined state 3. No new information is gained for the autosomes when separating these states. Using Eq. 2.1, we split the transition matrix into fast and slow processes:

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ \frac{c_1}{16} & \frac{c_1}{16} & \frac{c_1}{8} & 0 & 1-\frac{c_1}{4} \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} & 0 & -1 \end{pmatrix}.$$

We solve for the stationary distribution of the fast matrix using the method in Appendix A (simpler here by a single absorbing state for two lineages between individuals rather than a closed communication class):

$$\mathbf{P} = \lim_{r\to\infty} \mathbf{A}^r = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ \frac{c_1}{16-3c_1} & 0 & 0 & 0 & \frac{16-4c_1}{16-3c_1} \\ \frac{c_1}{16-3c_1} & 0 & 0 & 0 & \frac{16-4c_1}{16-3c_1} \\ \frac{c_1}{16-3c_1} & 0 & 0 & 0 & \frac{16-4c_1}{16-3c_1} \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Using $\mathbf{G} = \mathbf{PBP}$, we obtain the matrix exponential $e^{t\mathbf{G}}$ using the method of Appendix B. We then compute $\Pi(t)$ via Eq. 2.3, converting t back into units of N generations:

$$\Pi(t) = \mathbf{P}e^{t\mathbf{G}} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 - \frac{1-\frac{c_1}{4}}{1-\frac{3}{16}c_1} e^{-\frac{t}{4N}\left(\frac{1}{1-\frac{3}{16}c_1}\right)} & 0 & 0 & 0 & \frac{1-\frac{c_1}{4}}{1-\frac{3}{16}c_1} e^{-\frac{t}{4N}\left(\frac{1}{1-\frac{3}{16}c_1}\right)} \\ 1 - \frac{1-\frac{c_1}{4}}{1-\frac{3}{16}c_1} e^{-\frac{t}{4N}\left(\frac{1}{1-\frac{3}{16}c_1}\right)} & 0 & 0 & 0 & \frac{1-\frac{c_1}{4}}{1-\frac{3}{16}c_1} e^{-\frac{t}{4N}\left(\frac{1}{1-\frac{3}{16}c_1}\right)} \\ 1 - \frac{1-\frac{c_1}{4}}{1-\frac{3}{16}c_1} e^{-\frac{t}{4N}\left(\frac{1}{1-\frac{3}{16}c_1}\right)} & 0 & 0 & 0 & \frac{1-\frac{c_1}{4}}{1-\frac{3}{16}c_1} e^{-\frac{t}{4N}\left(\frac{1}{1-\frac{3}{16}c_1}\right)} \\ 1 - e^{-\frac{t}{4N}\left(\frac{1}{1-\frac{3}{16}c_1}\right)} & 0 & 0 & 0 & e^{-\frac{t}{4N}\left(\frac{1}{1-\frac{3}{16}c_1}\right)} \end{pmatrix}.$$

We extract from the first column of this matrix the cumulative distribution functions for two lineages starting in state 1 (within an individual) and state 3 (between individuals):

$$F_T(t) = F_U(t) = 1 - \frac{1 - \frac{c_1}{4}}{1 - \frac{3}{16}c_1} e^{-\frac{t}{4N}\left(\frac{1}{1-\frac{3}{16}c_1}\right)}, \tag{2.45}$$

$$F_V(t) = 1 - e^{-\frac{t}{4N}\left(\frac{1}{1-\frac{3}{16}c_1}\right)}. \tag{2.46}$$

Severson *et al.* (2021) showed that the limiting distribution for nth cousin mating is given by their Eqs. 47 and 48:

$$F_T(t) = F_U(t) = 1 - \frac{1 - 4c}{1 - 3c} e^{-\frac{t}{4N}\left(\frac{1}{1-3c}\right)},$$
$$F_V(t) = 1 - e^{-\frac{t}{4N}\left(\frac{1}{1-3c}\right)}.$$

In the special case where we only have first-cousin mating, we replace their c term with $c_1/16$ and recover Eqs. 2.45 and 2.46, respectively.

For the expectations of these distributions, by $\mathbb{E}[X] = \int_0^\infty [1 - F_X(x)]\, dx$ for $X > 0$, we find

$$\mathbb{E}[T] = \mathbb{E}[U] = 4N\left(1 - \frac{c_1}{4}\right), \tag{2.47}$$

$$\mathbb{E}[V] = 4N\left(1 - \frac{3}{16}c_1\right). \tag{2.48}$$

Eqs. 2.47 and 2.48, obtained from the limiting distribution, accord with the large-N limit of Eqs. 8 and 10 from Severson *et al.* (2019), in which they were calculated via first-step analysis.

Chapter 3

Modeling the effects of consanguinity on autosomal and X-chromosomal runs of homozygosity and identity-by-descent sharing

The following chapters and figures are currently in preparation for publication: Daniel J. Cotter, Alissa L. Severson, Hormazd N. Godrej, Jonathan T. L. Kang, Shai Carmi, Noah A. Rosenberg.

Abstract

Runs of homozygosity (ROH) and identity-by-descent (IBD) sharing can be studied in diploid coalescent models by noting that ROH and IBD-sharing at a genomic site are predicted to be inversely related to coalescence times—which in turn can be mathematically obtained in terms of parameters describing consanguinity rates. Comparing autosomal and X-chromosomal coalescent models, we consider ROH and IBD-sharing in relation to consanguinity that proceeds via multiple forms of first-cousin mating. We predict that across populations with different levels of consanguinity, (1) in a manner that is qualitatively parallel to the increase of autosomal IBD-sharing with

autosomal ROH, X-chromosomal IBD-sharing increases with X-chromosomal ROH, owing to the dependence of both quantities on consanguinity levels; (2) even in the absence of consanguinity, X-chromosomal ROH and IBD-sharing levels exceed corresponding values for the autosomes, owing to the smaller population size and lower coalescence time for the X chromosome than for autosomes; (3) with matrilateral consanguinity, the relative increase in ROH and IBD-sharing on the X chromosome compared to the autosomes is greater than in the absence of consanguinity. Examining genome-wide SNPs in human populations for which consanguinity levels have been estimated, we find that autosomal and X-chromosomal ROH and IBD-sharing levels generally accord with the predictions. We find that each 1% increase in autosomal ROH is associated with an increase of 2.1% in X-chromosomal ROH, and each 1% increase in autosomal IBD-sharing is associated with an increase of 1.6% in X-chromosomal IBD-sharing. For each calculation, particularly for ROH, the estimate is reasonably close to the increase of 2% predicted by the population-size difference between autosomes and X chromosomes. The results support the utility of coalescent models for understanding patterns of genomic sharing and their dependence on sex-biased processes.

3.1 Introduction

Autosomes and the X chromosome carry different signatures of population-genetic processes, owing both to differences in their mode of transmission and to demographic differences between males and females. Comparisons of autosomes and X chromosomes can therefore contribute to understanding genomic consequences of the different modes of transmission and of sex-biased and sex-specific processes, and many studies of autosomes and X chromosomes have considered empirical aspects of their population genetics in seeking such understanding (Wilkins and Marlowe, 2006; Ramachandran *et al.*, 2008; Bustamante and Ramachandran, 2009; Ellegren, 2009; Arbiza *et al.*, 2014; Goldberg and Rosenberg, 2015; Buffalo *et al.*, 2016; Webster and Wilson Sayres, 2016).

One set of population-genetic signatures that has the potential to be informative about sex-specific phenomena concerns features of genomic sharing: patterns in runs of homozygosity (ROH) and identity-by-descent (IBD) sharing on autosomes and the X chromosome (Buffalo *et al.*, 2016; Cai *et al.*, 2022). Recently, we have studied the distribution of the time to the most recent

common ancestor (T_{MRCA}) for pairs of autosomal lineages and pairs of X-chromosomal lineages in diploid coalescent models under different types of consanguinity, considering coalescence of lineages within an individual and lineages in separate individuals (Severson *et al.*, 2019, 2021; Cotter *et al.*, 2021, 2022). This analysis finds that consanguinity decreases T_{MRCA} both for lineage pairs in the same individual *and* for lineage pairs in individuals in different mating pairs. Further, because genomic sharing at a locus increases with decreasing T_{MRCA}, consanguinity increases genomic sharing both within (ROH) and between individuals (IBD) (Severson *et al.*, 2019). Considering autosomal and X-chromosomal systems separately, relationships between consanguinity levels and T_{MRCA} values produce predictions about relative values of autosomal and X-chromosomal ROH and IBD—with consanguinity that proceeds via matrilateral first-cousin mating reducing X-chromosomal coalescence times to a greater extent than patrilateral first-cousin mating (Cotter *et al.*, 2021, 2022).

Here, we study the connections between autosomal and X-chromosomal T_{MRCA} and features of X-chromosomal and autosomal ROH and IBD. Adding consideration of recombination to our diploid coalescent models, we examine predictions that compare X-chromosomal ROH to X-chromosomal IBD-sharing, X-chromosomal ROH to autosomal ROH, and X-chromosomal IBD-sharing to autosomal IBD-sharing. We consider human population-genetic data on ROH and IBD in a set of populations with consanguinity rates documented from demographic studies, using the results to understand effects of different forms of consanguinity on genomic sharing.

3.2 Theory

3.2.1 No consanguinity

Model

To derive expectations about features of genomic sharing on the autosomes and the X chromosome, we first consider a diploid, constant-sized population with N male–female mating pairs. We assume that recombination is constant across the autosomes and occurs at a per-Morgan rate proportional to the number of generations, $2g$, separating two sampled alleles. To account for

differences between the X-chromosome and the autosomes, we assume $4N$ autosomes for every $3N$ X chromosomes and a scaled X-chromosomal recombination rate $\frac{2}{3}$ that of the autosomes—because recombination occurs only in females and X-chromosomes are in females two thirds of the time (Hedrick, 2007).

The calculations in this section derive from work on coalescent theory and its relationship to genomic sharing (Palamara *et al.*, 2012; Carmi *et al.*, 2014; Browning and Browning, 2015). In general, this type of theoretical computation combines the coalescence-time distribution and a random variable that describes the length distribution of a segment given a specified time to the most recent common ancestor. Below, we derive the ratio of the expectation of total sharing on the X chromosome to the expectation of total sharing on the autosomes.

Expected X-chromosomal:autosomal total genomic sharing

In the absence of consanguinity, we derive a prediction for the ratio of the expected fraction of the X chromosome that lies in IBD segments and the corresponding expected fraction of the autosomal genome that lies in IBD segments. For a population with a demographic model whose parameterization is abbreviated by a quantity θ and whose recombination process has parameterization ρ, Palamara *et al.* (2012) specified the probability density function $p(\ell \mid \theta, \rho)$ that a specific locus is spanned by an IBD segment of a specific genetic length ℓ. For the closed interval $R = [u, v]$, the probability that a locus is spanned by an IBD segment with length in R is

$$\mathbb{P}_R(\ell \mid \theta, \rho) = \int_u^v p(\ell \mid \theta, \rho)\, d\ell.$$

Palamara *et al.* (2012) separated $p(\ell \mid \theta, \rho)$ into two terms by marginalizing over the number of generations to the most recent common ancestor, measured in discrete time as a random variable g_{mrca}. Following their eqs. 1 and 2,

$$p(\ell \mid \theta, \rho) = \sum_{g=1}^{\infty} p(g_{mrca} = g \mid \theta)\, p(\ell \mid g_{mrca} = g, \rho). \tag{3.1}$$

The term $p(g_{mrca} = g \mid \theta)$ is the coalescence-time distribution, which for a constant-sized population (parameterizing θ with a population size of N_e lineages) is a geometric random variable with rate $1/N_e$. The term $p(\ell \mid g_{mrca} = g, \rho)$ is the probability density of the length of a segment around a randomly chosen locus with coalescence time $g_{mrca} = g$.

Treating the distance from the locus to a recombination event as exponentially distributed, so that the total length of a shared segment between two lineages is the sum of two exponential random variables—the distance to the next recombination on the left plus the distance to the next recombination on the right—and measuring $R = [u, v]$ in centimorgans, they obtained in their equation 4:

$$\mathbb{P}_R\left(\ell \,\middle|\, \theta = N_e,\ \rho = \frac{t}{50}\right) = \int_0^\infty \left[\frac{e^{-\frac{t}{N_e}}}{N_e} \int_u^v Erl_2\left(\ell; \frac{t}{50}\right) d\ell\right] dt. \tag{3.2}$$

The first term is $p(t_{mrca} = t \mid \theta)$ (note the switch to continuous time, substituting the discrete, geometric g_{mrca} by the continuous, exponential t_{mrca} still measured in units of generations). The second, $p(\ell \mid g_{mrca} = g, \rho)$, is an Erlang density $(t/50)^2 \ell e^{-\ell t/50}$ (Johnson *et al.*, 1994, pg. 552) with shape parameter 2 and rate parameter $\rho = \frac{t}{50}$ centimorgans. With $R = [u, \infty)$, representing segments of size u centimorgans or greater, the inner integral gives (Palamara *et al.*, 2012)

$$\mathbb{P}_R\left(\ell \,\middle|\, \theta = N_e,\ \rho = \frac{t}{50}\right) = \int_0^\infty \left[\frac{e^{-\frac{t}{N_e}}}{N_e}\left(1 + \frac{ut}{50}\right) e^{-ut/50}\right] dt.$$

For the autosomes, we set $N_e = 4N$ for a population size of $4N$ autosomal lineages:

$$\mathbb{P}_R^A\left(\ell \,\middle|\, \theta = 4N,\ \rho = \frac{t}{50}\right) = \int_0^\infty \left[\frac{e^{-\frac{t}{4N}}}{4N}\left(1 + \frac{ut}{50}\right) e^{-\frac{ut}{50}}\right] dt = \frac{25(25 + 4Nu)}{(25 + 2Nu)^2}. \tag{3.3}$$

Similarly, for the X chromosome, we set $N_e = 3N$ for the reduced number of X-chromosomal lineages. We rescale the $\rho = \frac{t}{50}$ centimorgans from eq. 3.2 by $\frac{2}{3}$, giving $\rho = \frac{t}{75}$, to account for the reduced recombination rate:

$$\mathbb{P}_R^X\left(\ell \,\middle|\, \theta = 3N,\ \rho = \frac{t}{75}\right) = \int_0^\infty \left[\frac{e^{-\frac{t}{3N}}}{3N}\left(1 + \frac{ut}{75}\right) e^{-\frac{ut}{75}}\right] dt = \frac{25(25 + 2Nu)}{(25 + Nu)^2}. \tag{3.4}$$

The expected fraction f of the genome that lies in IBD segments in length interval R is $\mathbb{E}_R[f \mid \theta, \rho] = \mathbb{P}_R(\ell \mid \theta, \rho)$ (Palamara *et al.*, 2012, equation 9). Using eqs. 3.3 and 3.4, we can express the ratio of the expected fraction of the X chromosome that lies in IBD segments with length in $R = [u, \infty)$ and the expected fraction of the autosomes that lies in IBD segments with length in $R = [u, \infty)$:

$$\frac{\mathbb{E}_R^X[f \mid \theta = 3N,\ \rho = \frac{t}{75}]}{\mathbb{E}_R^A[f \mid \theta = 4N,\ \rho = \frac{t}{50}]} = \frac{\mathbb{P}_R^X(\ell \mid \theta = 3N,\ \rho = \frac{t}{75})}{\mathbb{P}_R^A(\ell \mid \theta = 4N,\ \rho = \frac{t}{50})} = \frac{(25 + 2Nu)^3}{(25 + Nu)^2(25 + 4Nu)}.$$

Taking $N \longrightarrow \infty$, we obtain

$$\lim_{N \to \infty} \frac{\mathbb{E}_R^X[f \mid \theta = 3N,\ \rho = \frac{t}{75}]}{\mathbb{E}_R^A[f \mid \theta = 4N,\ \rho = \frac{t}{50}]} = 2. \tag{3.5}$$

Because this limit does not depend on the lower limit of interval R, the population-size difference for X chromosomes and autosomes gives rise to a prediction that, irrespective of the interval R, for large N, the fraction of the X chromosome that lies in IBD segments with lengths in R is twice the corresponding fraction for autosomes.

A similar argument holds for ROH. A pair of lineages in a single individual is inherited from two lineages in two separate individuals in the previous generation. In an infinite population without consanguinity, the two lineages in the parental generation represent two independent draws from the population. Hence, the genomic sharing of the parental lineages follows the behavior we have described for IBD-sharing. To produce two lineages in the offspring, one additional generation of recombination occurs; however, the probability that a recombination event changes the IBD status of two lineages in one generation is small, so that ROH behavior in the offspring closely follows the IBD behavior of the parents. We can conclude that, as we found for IBD segments, the fraction of the X chromosome that lies in ROH segments with lengths in R is equal to twice the corresponding fraction for autosomes.

3.2.2 Consanguinity

Model

We have previously studied the effects of first-cousin consanguinity on coalescence times (Cotter *et al.*, 2021, 2022). Under a coalescent model, extending work of Campbell (2015) and Severson *et al.* (2019, 2021), we considered a population of N diploid mating pairs, labeling individuals by sex. In each generation, a fraction c_1 of the mating pairs are consanguineous, with a specific mixture of different types of first-cousin consanguinity (c_{pp} for patrilateral-parallel, c_{pc} for patrilateral-cross, c_{mp} for matrilateral-parallel, c_{mc} for matrilateral-cross—see Figure 3.1). Under the model, we computed limiting distributions for pairwise values of the time to the most recent common ancestor (T_{MRCA}) for two autosomal lineages in the same individual, two X-chromosomal lineages in the same individual, two autosomal lineages in different individuals, and two X-chromosomal lineages in different individuals (Table 3.1). The results rely on $N \longrightarrow \infty$ limits via the separation-of-time-scales method of Möhle (1998), in which a "fast" process induces a nonzero probability of instantaneous coalescence; the remaining coalescence occurs by a "slow" process that takes a positive amount of time. They can be regarded as approximate for finite populations.

ROH lengths are inversely related to within-individual coalescence times, and IBD lengths are inversely related to between-individual coalescence times. Hence, the T_{MRCA} calculations in our model give rise to predictions about features of autosomal and X-chromosomal ROH and IBD. In general, because a population has fewer copies of an X-chromosomal locus than an autosomal locus, X-chromosomal coalescence times are smaller than autosomal coalescence times. We showed that in relation to values seen in a non-consanguineous population, X-chromosomal within-individual coalescence times are reduced by consanguinity to a greater extent than are X-chromosomal between-individual coalescence times (Cotter *et al.*, 2021, Table 1). Here, extending the results on genomic sharing from Palamara *et al.* (2012), we use the limiting coalescence-time distributions from Cotter *et al.* (2022) to derive theoretical predictions for features of ROH and IBD-sharing on the X-chromosome and the autosomes.

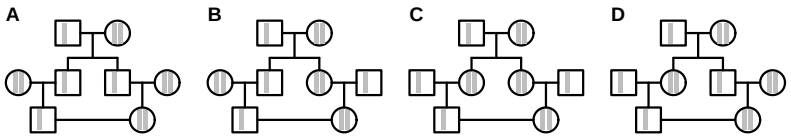

Figure 3.1: X chromosomes in first-cousin mating schemes. **(A)** Patrilateral-parallel. **(B)** Patrilateral-cross. **(C)** Matrilateral-parallel. **(D)** Matrilateral-cross.

	Chromosome	Cumulative distribution	Equation from Cotter *et al.* (2022)
Within (ROH)	Autosomes	$1 - \frac{1-\frac{c_1}{4}}{1-\frac{3}{16}c_1} e^{-\frac{t}{4N}\left(\frac{1}{1-\frac{3}{16}c_1}\right)}$	Eq. C2
	X	$1 - \frac{1-\frac{c_{mp}}{2}-\frac{c_{mc}}{2}}{1-\frac{5}{16}c_{mp}-\frac{3}{8}c_{mc}} e^{-\frac{t}{3N}\left(\frac{1+\frac{c_{mp}}{16}-\frac{c_{mc}}{8}}{1-\frac{5}{16}c_{mp}-\frac{3}{8}c_{mc}}\right)}$	Eq. 37
Between (IBD)	Autosomes	$1 - e^{-\frac{t}{4N}\left(\frac{1}{1-\frac{3}{16}c_1}\right)}.$	Eq. C3
	X	$1 - e^{-\frac{t}{3N}\left(\frac{1+\frac{c_{mp}}{16}-\frac{c_{mc}}{8}}{1-\frac{5}{16}c_{mp}-\frac{3}{8}c_{mc}}\right)}$	Eq. 38

Table 3.1: Limiting cumulative distribution functions for coalescence times for two X-chromosomal and two autosomal lineages sampled within- and between-individuals. Equations are taken from Cotter *et al.* (2022).

Expected X-chromosomal:autosomal total genomic sharing

To derive an expectation under our models of ROH and IBD-sharing with consanguinity, we begin by modifying eq. 3.1, once again switching to continuous time, t. Because only the coalescence-time distribution depends on the underlying demography—the population size and the rates of first-cousin consanguinity—it suffices to apply $p(t_{mrca} = t \mid \theta)$ and ρ in different versions of the demographic model.

It is convenient to begin with between-individual coalescence times and IBD-sharing. Using the coalescence-time distributions in Table 3.1, the time to the most recent common ancestor for two lineages in two separate individuals follows a coalescent with the population size scaled based on the rates for the different types of consanguinity. Converting the cumulative distributions in Table 3.1 to their probability density functions and annotating $\theta = \{4N, c_1\}$ and $\theta = \{3N, c_{mp}, c_{mc}\}$

for the autosomes and X chromosome, respectively, we have

$$p_A(t_{mrca} = t \mid \theta = \{4N, c_1\}) = \frac{1}{4N\left(1 - \frac{3}{16}c_1\right)} e^{-\frac{t}{4N}\left(\frac{1}{1-\frac{3}{16}c_1}\right)}, \tag{3.6}$$

$$p_X(t_{mrca} = t \mid \theta = \{3N, c_{mp}, c_{mc}\}) = \frac{1 + \frac{c_{mp}}{16} - \frac{c_{mc}}{8}}{3N\left(1 - \frac{5}{16}c_{mp} - \frac{3}{8}c_{mc}\right)} e^{-\frac{t}{3N}\left(\frac{1+\frac{c_{mp}}{16}-\frac{c_{mc}}{8}}{1-\frac{5}{16}c_{mp}-\frac{3}{8}c_{mc}}\right)}. \tag{3.7}$$

We solve for the expected fraction of the autosomes and the X chromosome appearing in IBD segments (using Palamara *et al.*, 2012, eq. 9). For the autosomes, using eq. 3.6 for the coalescence-time distribution and parameterizing recombination by $\rho = \frac{t}{50}$, the expected fraction of the autosomes shared identically by descent in a population with N mating pairs and proportion $c_1 = c_{pp} + c_{pc} + c_{mp} + c_{mc}$ of first-cousin mating per generation is

$$\begin{aligned}\mathbb{E}^A_{R,b}\left[f \,\middle|\, \theta = \{4N, c_1\}, \rho = \frac{t}{50}\right] &= \int_0^\infty p_A(t \mid \theta) \times \left[\left(1 + \frac{ut}{50}\right) e^{-\frac{ut}{50}}\right] dt \\ &= \frac{25\left[25 + 4N\left(1 - \frac{3}{16}c_1\right)u\right]}{\left[25 + 2N\left(1 - \frac{3}{16}c_1\right)u\right]^2}.\end{aligned} \tag{3.8}$$

Here, we have written $\mathbb{E}^A_{R,b}[f]$ for the expected fraction of the autosomal genome shared in $R \in [u, \infty)$ between individuals (with the subscript b differentiating this quantity from a corresponding expectation *within* individuals). For the X chromosome, using eq. 3.7 for coalescence times and $\rho = \frac{t}{75}$ for recombination, we have

$$\begin{aligned}\mathbb{E}^X_{R,b}\left[f \,\middle|\, \theta = \{3N, c_{mp}, c_{mc}\}, \rho = \frac{t}{75}\right] &= \int_0^\infty p_X(t \mid \theta) \times \left[\left(1 + \frac{ut}{75}\right) e^{-\frac{ut}{75}}\right] dt \\ &= \frac{25\left[25 + 2N\left(\frac{1-\frac{5}{16}c_{mp}-\frac{3}{8}c_{mc}}{1+\frac{c_{mp}}{16}-\frac{c_{mc}}{8}}\right)u\right]}{\left[25 + N\left(\frac{1-\frac{5}{16}c_{mp}-\frac{3}{8}c_{mc}}{1+\frac{c_{mp}}{16}-\frac{c_{mc}}{8}}\right)u\right]^2}.\end{aligned} \tag{3.9}$$

Next, relying on the within-individual coalescence-time distributions for two lineages, we use a similar framework to evaluate the expected fraction of the genome that lies in runs of homozygosity. A point mass exists for the probability of instantaneous coalescence at $t = 0$ in the cumulative

distributions in Table 3.1: $\left(\frac{c_1}{16}\right) / \left(1 - \frac{3}{16}c_1\right)$ for the autosomes and $\left(\frac{3}{16}c_{mp} + \frac{1}{8}c_{mc}\right) / \left(1 - \frac{5}{16}c_{mp} - \frac{3}{8}c_{mc}\right)$ for the X chromosome, obtained by substituting $t = 0$ in the cumulative distributions. We express the expected fractions of the autosomes and X chromosome that lie in ROH using the instantaneous coalescence probabilities; for non-instantaneous coalescence, we follow eqs. 3.6 and 3.7.

We write $\mathbb{E}^A_{R,w}[f]$ for the expected fraction of the genome shared within individuals in the length interval $R \in [u, \infty)$. For the autosomes, with recombination parameterized by $\rho = \frac{t}{50}$, we have

$$\begin{aligned}\mathbb{E}^A_{R,w}\left[f \,\middle|\, \theta = \{4N, c_1\}, \rho = \frac{t}{50}\right] &= \frac{\frac{c_1}{16}}{1 - \frac{3}{16}c_1} + \left(1 - \frac{\frac{c_1}{16}}{1 - \frac{3}{16}c_1}\right) \\ &\quad \times \int_0^\infty p_A(t \mid \theta) \times \left[\left(1 + \frac{ut}{50}\right) e^{-\frac{ut}{50}}\right] dt \\ &= \frac{\frac{c_1}{16}}{1 - \frac{3}{16}c_1} + \frac{1 - \frac{c_1}{4}}{1 - \frac{3}{16}c_1} \times \left(\frac{25\left[25 + 4N\left(1 - \frac{3}{16}c_1\right)u\right]}{\left[25 + 2N\left(1 - \frac{3}{16}c_1\right)u\right]^2}\right). \quad (3.10)\end{aligned}$$

Similarly, for the X chromosome, with $\rho = \frac{t}{75}$, we have

$$\begin{aligned}\mathbb{E}^X_{R,w}\left[f \,\middle|\, \theta = \{3N, c_{mp}, c_{mc}\}, \rho = \frac{t}{75}\right] &= \frac{\frac{3}{16}c_{mp} + \frac{c_{mc}}{8}}{1 - \frac{5}{16}c_{mp} - \frac{3}{8}c_{mc}} + \left(1 - \frac{\frac{3}{16}c_{mp} + \frac{c_{mc}}{8}}{1 - \frac{5}{16}c_{mp} - \frac{3}{8}c_{mc}}\right) \\ &\quad \times \int_0^\infty p_X(t \mid \theta) \times \left[\left(1 + \frac{ut}{75}\right) e^{-\frac{ut}{75}}\right] dt \\ &= \frac{\frac{3}{16}c_{mp} + \frac{c_{mc}}{8}}{1 - \frac{5}{16}c_{mp} - \frac{3}{8}c_{mc}} + \frac{1 - \frac{c_{mp}}{2} - \frac{c_{mc}}{2}}{1 - \frac{5}{16}c_{mp} - \frac{3}{8}c_{mc}} \\ &\quad \times \left(\frac{25\left[25 + 2N\left(\frac{1 - \frac{5}{16}c_{mp} - \frac{3}{8}c_{mc}}{1 + \frac{c_{mp}}{16} - \frac{c_{mc}}{8}}\right)u\right]}{\left[25 + N\left(\frac{1 - \frac{5}{16}c_{mp} - \frac{3}{8}c_{mc}}{1 + \frac{c_{mp}}{16} - \frac{c_{mc}}{8}}\right)u\right]^2}\right). \quad (3.11)\end{aligned}$$

In Figure 3.2, we explore the effects of the various types of first-cousin consanguinity on the ratio between X-chromosomal and autosomal ROH and IBD by plotting the ratio of eq. 3.11 to eq. 3.10 for ROH (Fig. 3.2A) and eq. 3.9 to eq. 3.8 for IBD (Fig. 3.2B). For illustration, we choose values $N = 500$ for the population size and $u = 5$ cM for the minimal segment length, varying

only one consanguinity rate at a time. Both for IBD and for ROH, increasing the first-cousin consanguinity shifts the X:autosomal ratio away from the expectation of 2 given in eq. 3.5. Patrilateral consanguinity decreases this ratio below 2, whereas matrilateral consanguinity increases it above 2, with matrilateral-parallel producing a greater increase than matrilateral-cross. The effect of consanguinity on the ROH ratios (Fig. 3.2A) has magnitude greater than the effect on corresponding IBD ratios (Fig. 3.2B).

These patterns accord with the large-N limits for the ROH and IBD X:autosomal ratios. For ROH, the $N \to \infty$ limit of the ratio of eq. 3.11 to eq. 3.10 is

$$\lim_{N\to\infty} \frac{\mathbb{E}^X_{R,w}\left[f \,\middle|\, \theta = \{3N, c_{mp}, c_{mc}\}, \rho = \frac{t}{75}\right]}{\mathbb{E}^A_{R,w}\left[f \,\middle|\, \theta = \{4N, c_1\}, \rho = \frac{t}{50}\right]} = \frac{\left(1 - \frac{3}{16}c_1\right)\left(\frac{3}{16}c_{mp} + \frac{c_{mc}}{8}\right)}{\frac{c_1}{16}\left(1 - \frac{5}{16}c_{mp} - \frac{3}{8}c_{mc}\right)}, \tag{3.12}$$

recalling that c_1 is the sum of the rates of all four types of first-cousin consanguinity, $c_{pp} + c_{pc} + c_{mp} + c_{mc}$. Varying $c_{pp} + c_{pc}$ in $(0, 1]$ and holding $c_{mp} = c_{mc} = 0$, the limiting ratio is 0: patrilateral consanguinity produces no ROH on the X chromosome but a positive level of ROH on the autosomes. For $c_{mp} \in (0, 1]$ and all other consanguinity rates set to 0, the limiting ratio varies from minimum 3 ($c_{mp} \to 0$) to maximum $\frac{39}{11} \approx 3.545$ ($c_{mp} = 1$). For $c_{mc} \in (0, 1]$ and all other consanguinity rates set to 0, the limiting ratio is 2 at the minimum ($c_{mc} \to 0$) and $\frac{13}{5} = 2.6$ at the maximum ($c_{mc} = 1$). Note that the limiting function is undefined for $c_1 = 0$.

Similarly for IBD, the $N \to \infty$ limit of the ratio of eq. 3.9 to eq. 3.8 is

$$\lim_{N\to\infty} \frac{\mathbb{E}^X_{R,b}\left[f \,\middle|\, \theta = \{3N, c_{mp}, c_{mc}\}, \rho = \frac{t}{75}\right]}{\mathbb{E}^A_{R,b}\left[f \,\middle|\, \theta = \{4N, c_1\}, \rho = \frac{t}{50}\right]} = 2\left[\frac{1 - \frac{3}{16}c_1}{\left(\frac{1 - \frac{5}{16}c_{mp} - \frac{3}{8}c_{mc}}{1 + \frac{c_{mp}}{16} - \frac{c_{mc}}{8}}\right)}\right]. \tag{3.13}$$

At $c_1 = 0$, this limit is 2, as in the case without consanguinity. If $c_{pp} + c_{pc} = 1$ and the other rates are held at 0, then the limiting ratio is $\frac{13}{8} = 1.625$. If $c_{mp} = 1$, then the limit is $\frac{221}{88} \approx 2.511$. If $c_{mc} = 1$, then it is $\frac{91}{40} = 2.275$.

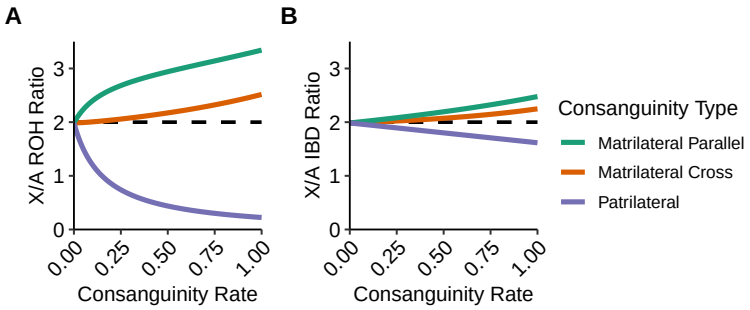

Figure 3.2: Expected ROH and IBD-sharing on the X chromosome relative to the autosomes as a function of consanguinity. (A) ROH. (B) IBD. For ROH, the ratio is calculated as eq. 3.11/eq. 3.10, and for IBD, it is calculated as eq. 3.9/eq. 3.8. In both cases, $N = 500$, $u = 5$ cM, and only one type of consanguinity is varied at a time while holding the others at 0. Patrilateral-parallel and patrilateral-cross consanguinity have the same effect.

3.3 Data analysis

3.3.1 Data

Demographic data

We consider a large demographic study that counted consanguineous pairs of various types—including first-cousin consanguineous pairs—among parents of newborns born in Israel 1955-1957 (Goldschmidt *et al.*, 1960). For each of a series of Jewish populations, among first-cousin mating pairs, Goldschmidt *et al.* (1960) tabulated numbers of patrilateral-parallel, patrilateral-cross, matrilateral-parallel, and matrilateral-cross cousin pairs. As a fraction of all mating pairs, we denote these quantities c_{pp}, c_{pc}, c_{mp}, and c_{mc}, respectively.

For nine populations that overlap between the demographic data of Goldschmidt *et al.* (1960) and genetic data used by Kang *et al.* (2016) and Severson *et al.* (2019), the rates c_{pp}, c_{pc}, c_{mp}, and c_{mc} appear in Table 3.2. In all nine populations, matrilateral consanguinity $c_{mp}+c_{mc}$ is nonzero, so that consanguinity influences X-chromosomal coalescence times, and hence ROH and IBD-sharing for both autosomes and X chromosomes.

Population	Frequency of first-cousin mating pairs (%)			
	Patrilateral parallel (c_{pp})	Patrilateral cross (c_{pc})	Matrilateral parallel (c_{mp})	Matrilateral cross (c_{mc})
Ashkenazi	0.507	0.296	0.465	0.084
Iranian	4.215	2.576	4.684	4.450
Iraqi	4.483	2.759	5.724	3.448
Libyan	2.013	2.685	0.671	0.671
Moroccan	0.794	0.794	1.984	1.587
Sephardi	0.329	0.494	0.988	1.318
Syrian	0.985	0.493	0.985	1.232
Tunisian	2.685	1.342	4.027	2.685
Yemenite	3.347	1.071	1.874	1.606

Table 3.2: Rates of the four different first-cousin mating types across 9 Jewish populations. Values are calculated from Tables 1 and 3 of Goldschmidt *et al.* (1960) as fractions of all mating pairs that are first-cousin pairs of particular types (omitting one double-first-cousin pair from both of its constituent categories of first-cousin pairs). As in Kang *et al.* (2016), the population listed as "Sephardi" corresponds to the "Turkey" population in Goldschmidt *et al.* (1960); the population listed as "Iranian" corresponds to the "Persia" population.

Autosomal genetic data

For the autosomes, we used genetic data from Kang *et al.* (2016), consisting of 202 Jewish individuals from 18 populations and 2,903 non-Jewish individuals from 123 populations, with 257,091 SNPs. These data are a merged data set constructed from data from Behar *et al.* (2013) and from the HGDP-CEPH and HapMap panels, as studied by Verdu *et al.* (2014). From these data, as in Severson *et al.* (2019), we consider the subset of 202 individuals from 18 Jewish populations, using the non-Jewish individuals only for phasing. These are the same individuals and same genotypes used by Kang *et al.* (2016) to call autosomal ROH segments and by Severson *et al.* (2019) to call autosomal IBD segments. We use the autosomal ROH segments directly from Kang *et al.* (2016), but we perform our own calls of autosomal IBD segments with updates of the method used by Severson *et al.* (2019).

X-chromosomal genetic data

For the X chromosome, we used genotypes from Behar *et al.* (2013). Beginning with 1,774 individuals and 32,823 SNPs, we first removed SNPs that were completely missing or monoallelic. Next, in individuals labeled as males, we verified the label by assessing heterozygosity of X-chromosomal

genotypes, converting the small number of heterozygous genotypes to missing data (Figure 3.3). We then removed, in sequence, SNPs missing in a large number of individuals (>200) and individuals missing a large number of SNPs (>2,500).

After processing, the data contained 1,647 individuals (1,227 males, 420 females) and 13,052 SNPs, comparable to the SNP density in the autosomal data (Figure 3.3). This collection contains 168 Jewish individuals from 18 populations (Table 3.3) and 1,479 non-Jewish individuals. We focus on the Jewish individuals for our analysis and include non-Jewish individuals only for phasing of both autosomal and X-chromosomal genotypes.

Population	**Females**		**Males**		**Demographic data?**
	Autosomes	**X**	**Autosomes**	**X**	
Ashkenazi	5	5	24	22	✓
Iranian	10	9	2	1	✓
Iraqi	5	5	8	5	✓
Libyan	0	0	6	6	✓
Moroccan	12	10	6	5	✓
Sephardi	5	5	17	14	✓
Syrian	0	0	2	2	✓
Tunisian	5	5	1	1	✓
Yemenite	12	11	6	4	✓
Algerian	1	1	4	4	
Azerbaijani	4	1	7	7	
Cochin	2	0	5	4	
Ethiopian	14	12	1	0	
Georgian	2	0	5	4	
Italian	3	3	7	7	
Kurdish	3	3	7	6	
Mumbai	0	0	6	4	
Uzbekistani	2	1	3	1	
Total	85	71	117	97	

Table 3.3: Numbers of sampled individuals in Jewish populations. Values for autosomes correspond to the 202 samples used by Kang *et al.* (2016), and values for the X chromosome correspond to the 168 individuals examined here. The "Demographic Data" column indicates the presence of rate data for the four types of first-cousin consanguinity in Goldschmidt *et al.* (1960)). As in Kang *et al.* (2016), the population listed as "Sephardi" in the table corresponds to the "Turkey" population in Goldschmidt *et al.* (1960), and the population listed as "Iranian" in the table corresponds to the "Persia" population in Goldschmidt *et al.* (1960). Note that quality control procedures differ on the X relative to the autosomes, so that fewer individuals are often available for the X chromosome.

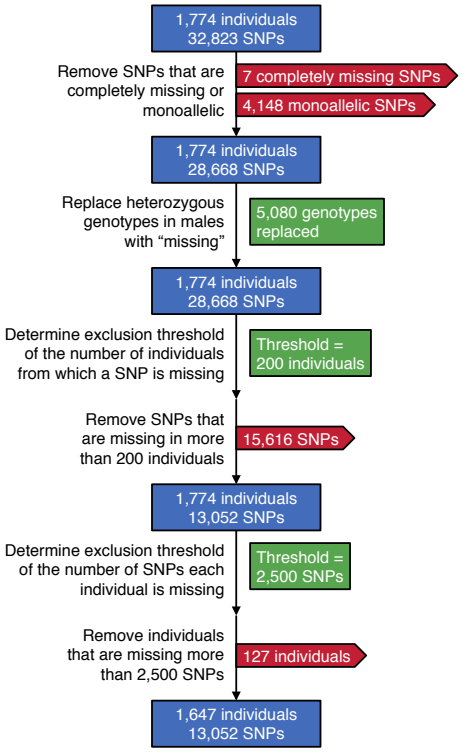

Figure 3.3: Pipeline for processing X-chromosomal data from Behar *et al.* (2013).

Data availability

For the autosomal data, see Kang *et al.* (2016); for the X-chromosomal data, see Behar *et al.* (2013). The demographic data on consanguinity are reported in Goldschmidt *et al.* (1960).

3.3.2 Methods

ROH

ROH lengths for the autosomes were taken directly from Kang *et al.* (2016). These ROH lengths were classified by Kang *et al.* (2016) into 3 length classes; for our analyses, we used the total length of all classes.

To measure ROH lengths for the X chromosome, we followed the procedure of Kang *et al.* (2016), with four modifications to account for differences between the X chromosome and autosomes. (1) In calculating sample allele frequencies for the X chromosome for each SNP in each

Population	LOD score cutoff
Algerian	0.2687779
Ashkenazi	2.193981
Azerbaijani	0.9832058
Ethiopian	3.118125
Iranian	0.9270845
Iraqi	1.302212
Italian	1.491206
Kurdish	1.101134
Moroccan	1.611013
Sephardi	2.135594
Tunisian	0.6681123
Uzbekistani	-0.0000279
Yemenite	1.710681

Table 3.4: Log-likelihood (LOD) score cutoffs from Kang *et al.* (2016) used for calling ROH in the 13 Jewish populations with female data.

population, we calculated the allele frequency with males contributing one allele and females contributing two. As in Kang *et al.* (2016), we performed 40 Bernoulli draws with this "true" allele frequency to obtain a sample allele frequency. This procedure reduces sample-size effects on ROH calls. (2) We used only females for identifying ROH, as males have only a single X chromosome. (3) For overlapping windows of 30 SNPs, Kang *et al.* (2016) calculated a log-likelihood (LOD) score to determine if windows were autozygous. The distribution of all LOD scores in a population was then used to set the threshold for calling ROH in the population. For consistency, and because identification of LOD score cutoffs for X-chromosomal data is more uncertain than for the autosomes due to a smaller number of X-chromosomal ROH available in our relatively small sample size, we used the autosomal LOD score cutoffs from Kang *et al.* (2016) rather than using X-chromosome-specific LOD scores (Table 3.4). (4) Due to the smaller amount of data available for subdividing ROH into length classes, we did not attempt to determine length classes for X-chromosomal ROH.

For each population, we summarized ROH lengths on the autosomes and X chromosome as the mean total proportion of the genome contained in ROH. First, we calculated the mean total ROH length as the sum of the lengths of ROH segments across all individuals in a population divided by the total number of individuals, considering only females for the X chromosome. For autosomes,

we normalized this quantity by 2,881.03 Mb for the combined length of chromosomes 1 through 22; for the X chromosome, we used 155.27 Mb. We base these lengths on human genome assembly GRCh37, as reported in the UCSC Genome Browser (Kent *et al.*, 2002).

IBD-sharing

We calculated autosomal IBD-sharing using the data from Kang *et al.* (2016). For each chromosome, we phased the full data set of 3,105 individuals using Beagle 5.1 (Browning and Browning, 2007) and default parameters (`burnin=6`, `iterations=12`, `phase-states=280`, `impute=false`, `ne=1,000,000`, `window=40.0`, `overlap=4.0`, `seed=-99,999`), with the GRCh37 genetic map for the map parameter (as provided with Beagle). We then considered the subset of 202 individuals in 18 Jewish populations, calling IBD segments using Refined IBD (Browning and Browning, 2013) with default parameters (`window=40.0`, `lod=3.0`, `length=1.5`, `trim=0.15`) and the map used for phasing. Our autosomal IBD calculations employed the method and data of Severson *et al.* (2019), except that we used a newer Beagle version and called IBD-sharing only on the subset of Jewish individuals rather than the whole sample.

For the X chromosome, we used data from the full 1,647 individuals (including the 168 Jewish individuals). We recoded alleles in males as pseudodiploid, as needed by Beagle 5.1 and Refined IBD. We then phased the 1,647 individuals with Beagle 5.1 using the same parameters and map as used for the autosomes. In the phased data, considering only the Jewish populations, we calculated IBD segments using Refined IBD in the same manner as for the autosomes. We then removed all duplicate IBD segments that resulted from pseudodiploid coding in males.

In each population, we summarized IBD-sharing as the mean total IBD proportion. That is, for each pair of individuals, we called IBD-sharing on the autosomes between four pairs of haplotypes, two in each individual in the pair. On the X chromosome, IBD comparisons considered one pair of haplotypes for pairs of males, two pairs for a male and a female, and four pairs for pairs of females. Thus, we divided the total IBD length between two individuals—summing across pairs of X chromosomes, one from one individual and one from the other—by one (two haplotypes), two (three haplotypes), or four (four haplotypes). We calculated mean total IBD length as the

mean across pairs of individuals after accounting for the number of pairwise haplotype comparisons. We then normalized this quantity, using the same genomic lengths as for ROH, to determine population-wise mean IBD proportions.

Population subsets

Because individuals with available X-chromosomal data represent a subset of the individuals with available autosomal data, in the following analyses, we used only a subset of the 18 populations. In particular, when comparing autosomal and X-chromosomal ROH, we considered only 13 populations, omitting 5 populations (Cochin, Georgian, Libyan, Mumbai, Syrian) for which no females and hence no X-chromosomal ROH calls were available (Table 3.3).

3.3.3 Results

Our theoretical results predict an increased proportion of ROH and IBD on the X chromosome relative to the autosomes as well as a positive relationship between IBD-sharing and ROH: increasing consanguinity decreases T_{MRCA} for two alleles within individuals as well as two alleles between individuals, in turn increasing both ROH and IBD-sharing (Severson *et al.*, 2019; Cotter *et al.*, 2021).

Empirical ROH levels and IBD levels are greater on the X chromosome than on the autosomes (Figure 3.5). The smaller total population size of the X chromosome, $3N$ compared to $4N$ in a population with equal sex ratio, produces lower coalescence times for the X chromosome, in turn giving rise to longer ROH and IBD segments.

We consider regressions of IBD proportions on ROH proportions, evaluating the coefficient of determination R^2 and the P-value for the null hypothesis of a regression slope of 0. In Figure 3.4, we plot the relationship between mean total IBD and ROH proportions in 13 populations, for both the autosomes and the X chromosome. Severson *et al.* (2019) previously performed this analysis for autosomes; here we compare autosomes and the X chromosome. In accord with the theoretical prediction, we see that IBD-sharing increases with ROH for the autosomes (Figure 3.4A; $R^2 = 0.27$), though not at the $P = 0.05$ significance level ($P = 0.07$). It also increases for the X chromosome (Figure 3.4B; $R^2 = 0.49$, $P = 0.008$), for which the relationship is stronger.

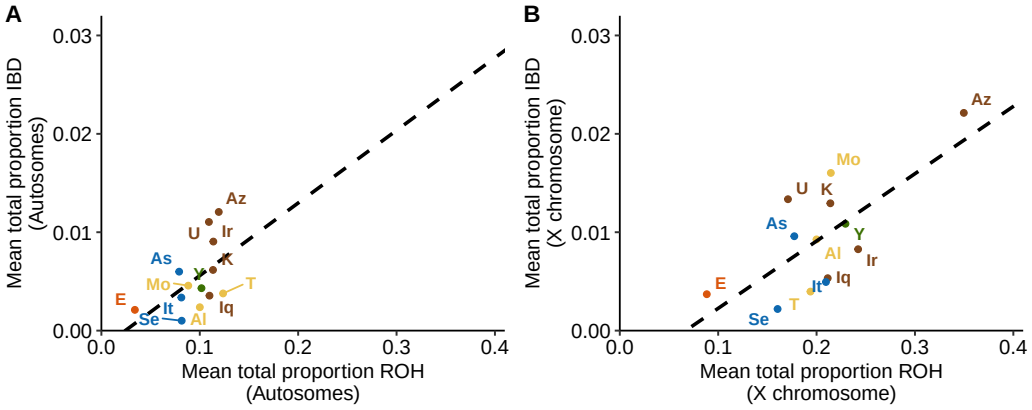

Figure 3.4: Mean genomic proportion contained in IBD segments versus mean genomic proportion contained in ROH segments. **(A)** Autosomes. **(B)** X chromosome. Thirteen populations are color-coded by regional group as in Kang *et al.* (2016) and Severson *et al.* (2019): Ethiopian, orange; European, blue; Middle Eastern, brown; North African, yellow; Yemenite, green. Population labels: Al, Algerian; As, Ashkenazi; Az, Azerbaijani; E, Ethiopian; Iq, Iraqi; Ir, Iranian; It, Italian; K, Kurdish; Mo, Moroccan; Se, Sephardi; T, Tunisian; U, Uzbekistani; Y, Yemenite. The regression equation is $y = 0.074x - 0.002$ ($R^2 = 0.27$, $P = 0.07$) for autosomes and $y = 0.068x - 0.005$ ($R^2 = 0.49$, $P = 0.008$) for the X chromosome. Both plots use the 13 Jewish populations with ROH data available for the X chromosome.

To explore the relationship between ROH patterns on the X chromosome and on autosomes, we next regress—with a fixed intercept of $y = 0$—the mean ROH genomic fraction on the autosomes onto the corresponding mean for the X chromosome. X-chromosomal and total autosomal ROH are positively related (Figure 3.6A; $R^2 = 0.96$, $P = 6.13 \times 10^{-10}$). The regression slope exceeds 2: for each 1% increase in total ROH on the autosomes, we see a 2.1% increase on the X chromosome. This greater increase for the X chromosome accords with the smaller X-chromosomal population size and reduced recombination rate—which inflate ROH for the X chromosome.

Next, having detected a relationship between total lengths in X-chromosomal and autosomal ROH, we compare genomic fractions of IBD-sharing. Fixing the regression intercept at $y = 0$, X-chromosomal IBD increases with autosomal IBD (Figure 3.6B; $R^2 = 0.87$, $P = 1.45 \times 10^{-6}$). A 1.6% increase in X-chromosomal IBD-sharing occurs for each 1% increase in autosomal IBD-sharing, consistent with the reduced population size of the X chromosome and its resulting reduction in coalescence times and increase in IBD segment length.

For the seven populations for which demographic estimates of consanguinity and genomic data are both available, we can compare the empirical ratio of the fractions of the X chromosome

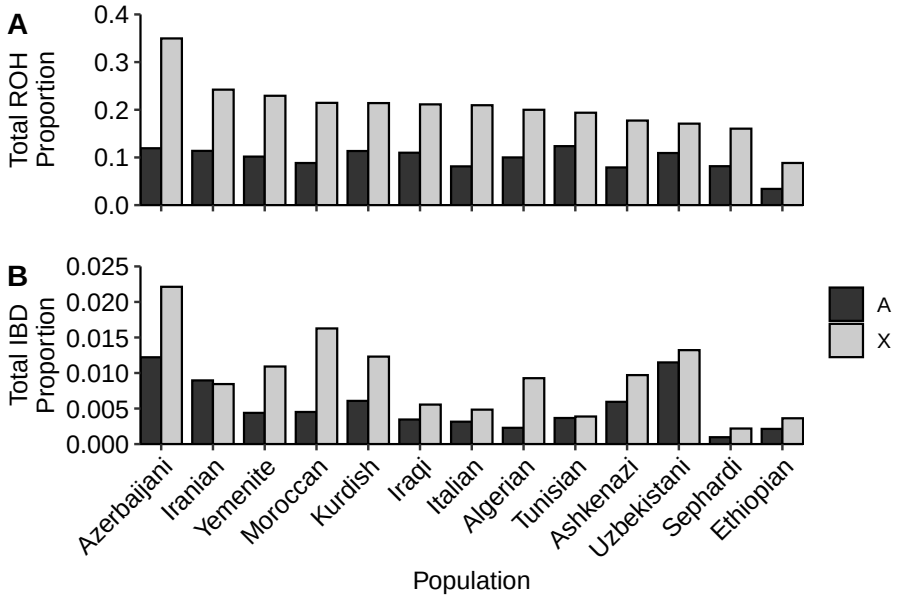

Figure 3.5: Proportion of autosomal and X-chromosomal ROH and IBD in each population. **(A)** ROH. **(B)** IBD. Populations are arranged in decreasing order by the proportion of the X-chromosomal genome lying in ROH.

and autosomal genome that lie in ROH to a theoretical prediction. Inserting the consanguinity rates from Table 3.2 and a range of values of the number of mating pairs N from 500 to 50000 into eqs. 3.11 and 3.10, we obtain predictions for the ratio of eqs. 3.11 and 3.10. The nontrivial patrilateral consanguinity in these populations, sometimes exceeding the matrilateral consanguinity, leads to predictions that lie below the ratio of 2 predicted from eq. 3.5 in the case of no consanguinity (Table 3.5). The empirical ratios tend to be near but somewhat greater than the theoretical range, suggesting that while the differing numbers of autosomal genomes and X chromosomes and the effects of consanguinity in part explain differences in autosomal and X-chromosomal ROH, other factors also contribute.

For IBD, a similar calculation of the theoretical ratio of X-chromosomal and autosomal ROH, using eqs. 3.9 and 3.8, places the seven populations into similar ranges. This similarity illustrates the lesser effect of differences in consanguinity rates on the predicted ratio of X-chromosomal and autosomal IBD compared to the corresponding ratio for ROH (Figure 3.2). Empirical IBD ratios tend to be farther from the predicted range than are empirical ROH ratios, indicating that the factors

we have considered—population-size differences between the X chromosome and autosomes, and consanguinity rates—may be less determinative of IBD patterns than of ROH patterns.

3.4 Discussion

This study has investigated the effect of consanguinity on X-chromosomal ROH and IBD-sharing. Under a coalescent model with consanguinity, we had previously obtained autosomal (Severson *et al.*, 2019, 2021) and X-chromosomal (Cotter *et al.*, 2021, 2022) distributions of coalescence times. Here, we have combined results on coalescence times with calculations based on properties of recombination to predict features of ROH and IBD-sharing under the model. We have also compared the predictions with empirical patterns in ROH and IBD-sharing in populations for which demographic measures of consanguinity have been reported.

For the coalescence times, we had previously observed that under the model, patrilateral first-cousin mating does not affect X-chromosomal coalescence times, and matrilateral first-cousin mating reduces X-chromosomal coalescence times relative to the non-consanguineous case; consanguinity produces a greater relative decrease in coalescence times for X chromosomes than for autosomes (Cotter *et al.*, 2021, 2022). Owing to the inverse relationship between genomic sharing around a site and the coalescence time at that site (Palamara *et al.*, 2012; Carmi *et al.*, 2014; Browning and Browning, 2015), corresponding results are reflected in ROH and IBD-sharing calculations under the model. The model predicts longer ROH and IBD-sharing on the X chromosome than on autosomes, owing to three factors: the smaller population size for X chromosomes produces a smaller coalescence time, the stronger effect of matrilateral consanguinity reduces coalescence times to a greater extent relative to the non-consanguineous model, and reduced recombination in X chromosomes increases ROH and IBD tract lengths.

In accord with this prediction, in data from Jewish populations, we observed that ROH and IBD-sharing did indeed cover a larger fraction of the X chromosome than the autosomes (Figure 3.5). Comparing X-chromosomal to autosomal ROH lengths, we observed an increased genomic fraction of ROH on the X-chromosome relative to the autosomes: a 1% increase in autosomal ROH

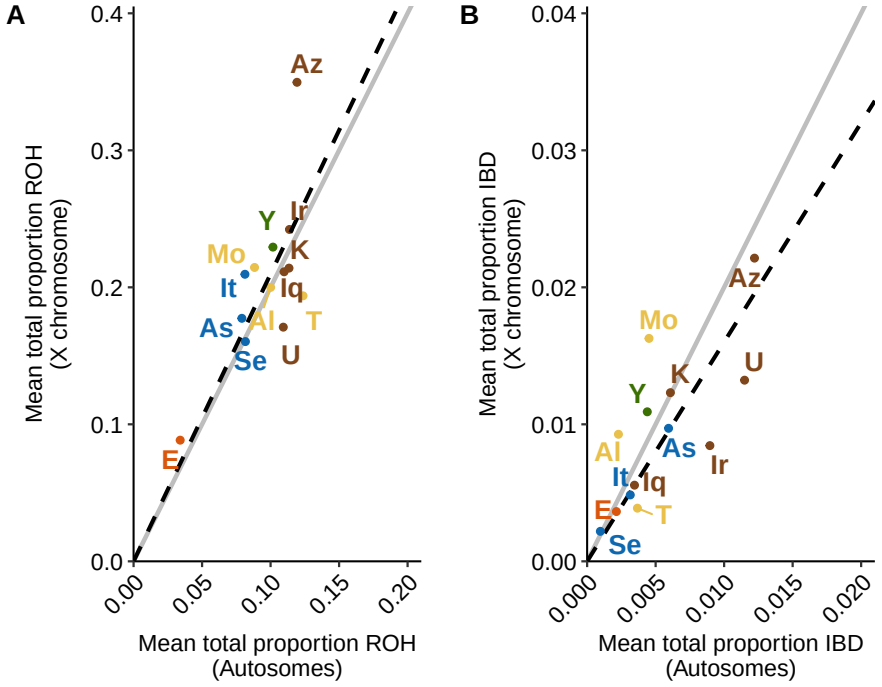

Figure 3.6: Mean genomic proportion contained in ROH and on the autosomes relative to the X chromosome. **(A)** ROH. **(B)** IBD. The solid line is the theoretical prediction $y = 2x$. The dashed line represents a regression with intercept fixed at 0: $y = 2.10x$ ($R^2 = 0.96$, $P = 6.13 \times 10^{-10}$) **(A)**, $y = 1.60x$ ($R^2 = 0.87$, $P = 1.45 \times 10^{-6}$) **(B)**.

gives rise to a 2.1% increase on the X chromosome (Figure 3.6A). For IBD-sharing, a 1% increase in autosomal IBD-sharing predicts a 1.6% increase on the X chromosome (Figure 3.6B).

The 2.1% and 1.6% increases on the X-chromosome generally align with model predictions. In a constant-sized population with no consanguinity, our model-based computations found that the ratio of the expected total fractions of the X chromosome and autosomes that lie in ROH or IBD segments approaches 2 for large N (eq. 3.5). In other words, for each 1% increase in the fraction of the autosomal genome covered by ROH or IBD segments, an increase of 2% is predicted for the corresponding coverage of the X chromosome.

We hypothesized that a portion of the increase in X-chromosomal ROH coverage for each 1% increase in autosomal ROH coverage (Figure 3.6A) differing from 2% and the corresponding difference from 2% for IBD was attributable to the effects of consanguinity—with matrilateral consanguinity increasing the prediction above 2% and patrilateral consanguinity decreasing it below

Population	ROH			IBD		
	Theoretical X:A ratio		Empirical X:A ratio	Theoretical X:A ratio		Empirical X:A ratio
	Minimum	Maximum		Minimum	Maximum	
Ashkenazi	1.541	1.935	2.247	1.542	1.991	1.633
Iranian	1.516	1.818	2.126	1.525	1.968	0.942
Iraqi	1.518	1.824	1.921	1.527	1.974	1.608
Moroccan	1.537	1.934	2.434	1.539	1.988	3.602
Sephardi	1.539	1.950	1.969	1.540	1.988	2.268
Tunisian	1.529	1.877	1.568	1.534	1.983	1.057

Table 3.5: Theoretical and empirical ratios of the proportion of the X-chromosome to the proportion of the autosomal genome lying in ROH and IBD segments. The theoretical ratio is the ratio of eqs. 3.11 and 3.10 for ROH and the ratio of eqs. 3.9 and 3.8 for IBD, inserting consanguinity rates from Table 3.2 and setting $u = 0.1$ cM for the the minimum size of ROH and IBD. We report the minimum and the maximum theoretical ratios achieved when varying N in the range $[500, 50000]$. The empirical ratio is calculated using the ROH and IBD proportions obtained via the Methods subsections on ROH and IBD, respectively.

2%. This potential attribution is compatible with the observation that the populations studied possess nonzero consanguinity rates, both matrilateral and patrilateral (Table 3.2). Using eqs. 3.8–3.11 to assess the effect of demographic consanguinity rates on ROH X:A ratios directly (Table 3.5), we see that agreement with predicted ranges is generally closer for ROH than for IBD.

That the empirical analysis generally follows model predictions, with greater sharing on the X chromosome than the autosomes in an amount close to the numerical prediction, supports the value of the model. However, many factors might contribute to deviations of the empirical X-chromosomal and autosomal ROH and IBD patterns from the predictions. First, processes not considered in the model influence differences in genetic variation between X chromosomes and autosomes. For example, differences in the numbers of mating males and females or differing male and female variance of reproductive success can alter effective population size for X chromosomes relative to autosomes (Webster and Wilson Sayres, 2016; Cai *et al.*, 2022). Recombination differences between X chromosomes and autosomes beyond the $\frac{2}{3}$ we have considered, with different autosomes having different rates per Mb (Kong *et al.*, 2002), can affect conversions of T_{MRCA} values to ROH and IBD lengths. These differences can also introduce differences in phasing and ROH and IBD detection; the detection problem is possibly also affected by our use of autosomal ROH cutoffs rather than X-chromosome-specific values in assigning X-chromosomal ROH. In particular, ROH

levels might be inflated by use of the autosomal LOD score cutoff for the higher-homozygosity X chromosome.

Beyond these concerns about ROH and IBD detection, a number of limitations may affect our empirical results. Our theoretical analysis relies on centimorgan measurements, whereas we analyze the data in megabases; a more precise comparison of X chromosomal and autosomal ROH and IBD could be performed by use of a genetic map. The comparison of theoretical and empirical ratios in Table 3.5 makes use of minimal genomic-sharing cutoffs; we used a cutoff standardized across all theory-based calculations, rather than adding complexity by choosing separate cutoffs for each component of the analysis (ROH vs. IBD, X-chromosomal vs. autosomal, and different populations).

We also note that consanguinity rates are unlikely to be stable over time in real populations, as the model assumes. For example, consanguinity rates from Goldschmidt *et al.* (1960), measured around the mean birth year of the sampled individuals (Kang *et al.*, 2016), represent births only at the single time point of 1955-1957; the number of generations over which they would have applied is unclear. Indeed, consanguinity rates have recently declined in some of the sampled populations (Tsafrir and Halbrecht, 1972; Cohen *et al.*, 2004).

Finally, the data set itself is also limited by a small number of females, so that few data points contribute to inferences on X-chromosomal ROH. We have used these data due to availability of demographic consanguinity rates measured for the four first-cousin types. Additional methodological choices could potentially be investigated in larger genomic data sets in consanguineous populations (e.g. Arciero *et al.*, 2021), and an ideal data set would include both large sample sizes as well as demographic estimates of consanguinity.

We have examined how coalescent models and ROH and IBD measurements on the X and the autosomes can provide information about sex-biased phenomena. Genomic effects of numerous sex-biased processes have been investigated extensively in theoretical models and data, particularly in relation to human populations (Wilkins and Marlowe, 2006; Ellegren, 2009; Arbiza *et al.*, 2014; Goldberg and Rosenberg, 2015; Webster and Wilson Sayres, 2016). Many organisms possess mating schemes that could induce different kinship levels for autosomes and sex chromosomes (e.g. sex-specific processes and the ZW system in birds, Pizzari *et al.* (2004); Schield *et al.* (2021)). As

genomic data on ROH and IBD data proliferate in diverse organisms (e.g. Florida scrub jays (Chen *et al.*, 2016), dogs (Mooney *et al.*, 2021)), our approach of examining coalescence times, ROH, and IBD-sharing can potentially contribute to understanding genomic effects of a variety of mating systems.

Acknowledgments. We acknowledge United States–Israel Binational Science Foundation grant 2017024 and NIH grant R01 HG005855 for support.

Chapter 4

A rarefaction approach for measuring population differences in rare and common variation

The following chapter and figures were originally published as:

Cotter, D. J., E. F. Hofgard, J. Novembre, Z. A. Szpiech, and N. A. Rosenberg, 2023 A rarefaction approach for measuring population differences in rare and common variation. GENETICS **224**: iyad070.

https://doi.org/10.1093/genetics/iyad070

Abstract

In studying allele-frequency variation across populations, it is often convenient to classify an allelic type as "rare," with nonzero frequency less than or equal to a specified threshold, "common," with frequency above the threshold, or entirely unobserved in a population. When sample sizes differ across populations, however, especially if the threshold separating "rare" and "common" corresponds to a small number of observed copies of an allelic type, discreteness effects can lead a

sample from one population to possess substantially more rare allelic types than a sample from another population, even if the two populations have extremely similar underlying allele-frequency distributions across loci. We introduce a rarefaction-based sample-size correction for use in comparing rare and common variation across multiple populations whose sample sizes potentially differ. We use our approach to examine rare and common variation in worldwide human populations, finding that the sample-size correction introduces subtle differences relative to analyses that use the full available sample sizes. We introduce several ways in which the rarefaction approach can be applied: we explore dependence of allele classifications on subsample sizes, we permit more than two classes of allelic types of nonzero frequency, and we analyze rare and common variation in sliding windows along the genome. The results can assist in clarifying similarities and differences in allele-frequency patterns across populations.

4.1 Introduction

The study of data on genetic variation often begins with simple questions. Which alleles are present? In which populations are they present, and where are they absent? Which alleles are common, and which are rare? Often, the first calculations that an analyst performs on a population-genetic dataset seek to answer such questions.

To take one example, a recent study of Witt *et al.* (2022) sought to characterize genetic variation in modern and archaic populations, with a particular interest in the sharing of alleles among groups. In their Figure 5, Witt *et al.* (2022) tabulated, for alleles classified as archaic, the fractions of those alleles that appear in modern Europeans, South Asians, and East Asians, in pairs among these three groups, and in all three groups.

In studies of presence and absence of alleles in populations, differing sample sizes among the groups can influence the resulting assessments. For example, an allele absent in a small sample might eventually be found in a larger sample, so that a population with a sample size that is small might appear to possess fewer alleles than a population with one that is large. This problem is addressed by the rarefaction method, borrowed for population genetics (e.g. Kalinowski, 2004) from

ecological work on species diversity (Hurlbert, 1971; Gotelli and Colwell, 2001). Using a combinatorial formula, given sample size N_j for population j and a fixed value of $g \leq N_j$, all possible subsamples of size g are considered, and the expected number of distinct alleles across random samples of size g is calculated. Multiple populations of different sample size can be compared by examining subsamples of equal size g.

Kalinowski (2004) devised a rarefaction-based calculation of "private allelic richness," a measure of the fraction of alleles that are private to a particular population—considering subsamples of size g from each of a series of populations. Generalizing this concept, Szpiech *et al.* (2008) introduced a calculation of the fraction of alleles that are private to a *set* of populations—that is, found in each of the populations—when subsamples of size g are taken in each population. Szpiech *et al.* (2008) examined geographic distributions of alleles in samples from multiple populations, all standardized with the same subsample size g. Thus, for example, for Populations 1, 2, and 3, with different sample sizes, the rarefaction-based calculation enables a comparison of the fraction of alleles found only in 1, only in 2, only in 3, in 1 and 2 but not 3, in 1 and 3 but not 2, in 2 and 3 but not 1, and in all three groups—assuming that all three groups have subsamples of equal size.

Recently, Biddanda *et al.* (2020) introduced a new computation and visualization to compare presence and absence of alleles across populations. Seeking to describe geographic distributions of alleles across multiple populations—as in Szpiech *et al.* (2008) and Witt *et al.* (2022)—Biddanda *et al.* (2020) made an additional distinction between alleles that are present and *rare* and those that are present and *common*. For each of several populations, they classified alleles into three categories: rare, common, and unobserved. For a population set, they tabulated fractions of alleles that possess particular classes, illustrating the classifications in new visualizations.

In the same way that sample size can affect presence and absence, sample size can also affect the classification of an allele as present and rare as opposed to present and common. Suppose a locus has the same allele frequencies in Populations 1 and 2, with sample sizes 39 and 40, respectively. Suppose a maximum of 5% is the largest allele frequency classified as rare. An allele A of frequency 5% that is regarded as rare in an infinite population will be regarded as rare in Population 1 when 1 copy is observed in the sample of size 39. The probability of observing exactly 1 copy is $\binom{39}{1}(0.05^1)(0.95^{38}) \approx 0.278$. The allele will be regarded as rare in

Population 2 if 1 copy is observed *or* if 2 copies are observed. The associated probability is $\binom{40}{1}(0.05^1)(0.95^{39}) + \binom{40}{2}(0.05^2)(0.95^{38}) \approx 0.548$. Hence, as a result of differing sample sizes, the two populations have the potential to differ dramatically in the number of their truly rare alleles (that is, rare at the population level) that are classified as rare in samples.

Here, we extend the geographic classification of alleles into categories of rare, common, and unobserved, as in Biddanda *et al.* (2020), but accounting for differences in sample size, as in Szpiech *et al.* (2008). In particular, we extend the rarefaction approach from Szpiech *et al.* (2008), which only considered presence and absence, to account for the three categories of Biddanda *et al.* (2020): unobserved, rare, and common. We examine whether the rarefaction correction to make use of equal sample sizes in the data of Biddanda *et al.* (2020) influences the interpretation of rare and common human variation. In the spirit of Biddanda *et al.* (2020), we also include a variety of visualizations for understanding sample-size-corrected patterns in the geographic distributions of rare and common alleles.

4.2 Statistical methods

Consider a single locus in an individual. We henceforth use "allelic type" to refer to one of a set of possible variants at a locus and "allele" to refer to an observation at a given locus in a single individual. Considering a locus with $I \geq 2$ allelic types, we denote by N_{ij} the number of copies of allelic type i observed in a sample from population j. By extension, $N_j = \sum_{i=1}^{I} N_{ij}$ is the sample size of population j at the locus. We consider $J \geq 2$ populations.

Biddanda *et al.* (2020) declare "rare" allelic types as those with nonzero frequency less than or equal to $100z\%$ in a population, where z is a specified numerical cutoff (they use $z = 0.05$). They then classify allelic types with frequency greater than $100z\%$ as "common." This classification gives rise to their three frequency categories of unobserved, rare, and common. Thus, considering all J populations, an allelic type takes on a "pattern" denoted by $\mathbf{x} = (x_1, x_2, \ldots, x_J)$, where each x_j has a value in $\{\text{unobserved, rare, common}\}$, herein shortened to $\{\text{U, R, C}\}$.

4.2.1 Three allelic classes: unobserved, rare, and common

For a sample with counts N_{ij} for the I allelic types in the J populations, we consider subsamples with specified sizes. Suppose that a sample of size g alleles is drawn in each of the J populations, for a total sample size of Jg. We calculate the probability that when we draw a sample of size Jg, an allelic type has pattern $\mathbf{x}$.

The probability U_{ijg} that allelic type i is *unobserved* in a subsample of size g from population j is

$$U_{ijg} = \frac{\binom{N_j - N_{ij}}{g}}{\binom{N_j}{g}}. \tag{4.1}$$

Here, the numerator is the number of ways to draw g alleles from among the alleles that *do not* have allelic type i. The denominator is the total number of ways to draw g alleles from among the N_j alleles in population j.

The probability R_{ijg} that allelic type i is *rare* in a subsample of size g is the probability of observing at least 1 and at most $\lfloor zN_j \rfloor$ copies of allelic type i in a subsample of size g. The floor function accounts for the classification of an allelic type with frequency exactly $100z\%$ as rare rather than common. The probability R_{ijg} satisfies

$$R_{ijg} = \frac{\sum_{k=1}^{\lfloor zN_j \rfloor} \left[\binom{N_{ij}}{k} \binom{N_j - N_{ij}}{g-k} \right]}{\binom{N_j}{g}}. \tag{4.2}$$

The numerator in eq. 4.2 sums over all possible ways to choose at least 1 and at most $\lfloor zN_j \rfloor$ copies of allelic type i. The denominator again gives the total number of ways to draw g alleles from the population sample size N_j.

Finally, the probability that allelic type i is *common* in a sample of size g taken from population j is simply

$$C_{ijg} = 1 - U_{ijg} - R_{ijg}. \tag{4.3}$$

Now that we have probabilities for an allelic type in a single population, we consider all J populations to determine the probability of a particular pattern $\mathbf{x}$. The probability that allelic type

i has pattern $\mathbf{x} = (x_1, x_2, \ldots, x_J)$ in a sample containing g alleles each from the J populations is

$$\prod_{j=1}^{J} f_{ijg}(x_j), \text{ where } f_{ijg}(x_j) = \begin{cases} U_{ijg}, & x_j = \text{U} \\ R_{ijg}, & x_j = \text{R} \\ C_{ijg}, & x_j = \text{C}. \end{cases} \tag{4.4}$$

At a locus, we sum across all I allelic types to give the expected fraction of allelic types that have pattern $\mathbf{x}$:

$$\frac{1}{I} \sum_{i=1}^{I} \prod_{j=1}^{J} f_{ijg}(x_j). \tag{4.5}$$

4.2.2 Extension to more than three classes

We can generalize the results describing unobserved, rare, and common allelic types to compute the probability P_{ijg} of finding an allelic type i in population j in a specified frequency window, where arbitrarily many windows are permitted. Define a window $(z_1, z_2]$, describing allelic types with frequency greater than z_1 and less than or equal to z_2. Eq. 4.2 for the probability that a sample of size g has its frequency for allelic type i in the window $(0, z]$ generalizes, and the probability that allelic type i has its frequency in $(z_1, z_2]$ is

$$P_{ijg} = \frac{\sum_{k=\lfloor z_1 N_j \rfloor + 1}^{\lfloor z_2 N_j \rfloor} \left[\binom{N_{ij}}{k} \binom{N_j - N_{ij}}{g-k} \right]}{\binom{N_j}{g}}. \tag{4.6}$$

Eq. 4.6 can consider arbitrary divisions of the unit interval for frequencies into disjoint intervals. Note that if we instead regard intervals as having a closed lower bound and an open upper bound, so that we consider the probability that an allelic type has frequency in $[z_1, z_2)$, then we simply change the limits of the sum to $\lceil z_1 N_j \rceil$ and $\lceil z_2 N_j \rceil - 1$.

4.2.3 Biallelic loci

For biallelic loci, $I = 2$, suppose we are interested in only one specific allelic type. We label this allelic type by 1 and the other allelic type by 2 and write simplified formulas for U_{ijg} and R_{ijg}. N_{1j}

is the count of allelic type 1 in population j and N_{2j} is the count of allelic type 2. Then

$$U_{ijg} = \frac{\binom{N_{2j}}{g}}{\binom{N_j}{g}} \tag{4.7}$$

$$R_{ijg} = \frac{\sum_{k=1}^{\lfloor zN_j \rfloor} \left[\binom{N_{1j}}{k} \binom{N_{2j}}{g-k} \right]}{\binom{N_j}{g}}. \tag{4.8}$$

Eq. 4.4 can then be used to calculate the probability that allelic type 1 has pattern $\mathbf{x} = (x_1, x_2, \ldots, x_J)$. With three frequency classes in each of J populations, allelic type 1 has 3^J possible patterns.

4.3 Data analysis

4.3.1 Biddanda *et al.* (2020) dataset

Biddanda *et al.* (2020) used data from the 2504 individuals in the 26 populations of the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2015; Byrska-Bishop *et al.*, 2022) to explore the relative abundances of different patterns $\mathbf{x}$, considering five "super-populations." They used the globally minor allele at each locus—the allelic type at global frequency less than 50%—to classify each locus as a pattern $\mathbf{x} = (x_1, x_2, \ldots, x_J)$, where $J = 5$.

They placed the 1000 Genomes populations, annotated here by three-letter abbreviations, into the five super-populations. From 1 to 5, vector entries correspond to African (ESN, GWD, LWK, MSL, YRI), European (FIN, GBR, IBS, TSI), South Asian (BEB, GIH, ITU, PJL, STU), East Asian (CDX, CHB, CHS, JPT, KHV), and American super-populations (ACB, ASW, CEU, CLM, MXL, PEL, PUR). Thus, for example, a locus rare in Africa, common in East Asia, and unobserved elsewhere has pattern $\mathbf{x} = \{\text{R, U, U, C, U}\}$ or RUUCU for short.

Biddanda *et al.* (2020) considered genome-wide biallelic SNPs, classifying each SNP into one of $3^5 - 1$ patterns based on the globally minor allele; because each locus is polymorphic by definition, the pattern UUUUU is omitted in their analysis.

We downloaded the data used by Biddanda *et al.* (2020) from the 1000 Genomes FTP server (see "Data availability"). We retained the same super-population groups used by Biddanda *et al.* (2020).

After filtering to consider only biallelic SNPs, we determined the globally minor allele for each SNP. Our definition of the minor allele is the allelic type that, when averaging relative frequencies across the five super-populations, has frequency below $\frac{1}{2}$; for 240 sites genome-wide with exactly 50% global frequency for each of the two allelic types, we chose one allelic type at random to be the "minor" allele. We then tabulated counts of the minor allele for the five super-populations, disregarding sites for which data were entirely missing in at least one of the five. This process left us with 95,563,258 SNPs in the 2504 individuals.

4.3.2 Pointwise rarefaction analysis

To evaluate the effect of sample-size correction on the geographic distribution of allelic types, we applied the rarefaction calculation (eq. 4.4) to the 1000 Genomes SNPs in the five super-populations. This calculation relies on the biallelic eqs. 4.7 and 4.8, along with eq. 4.3. For an illustrative analysis, we considered 1,226,225 SNPs on chromosome 22, ensuring that each SNP possessed a sample of size 500 or greater in each of the five super-populations (the equivalent of 250 diploid individuals).

Thus, for each of a series of values of g, for each SNP, focusing on the minor allele, we obtained probabilities for each of the $3^5 = 243$ patterns, treating 5% as the maximal frequency for allelic types treated as rare. For fixed g, for each of the 243 patterns, we averaged the SNP-specific probabilities across all SNPs to determine the mean probability that a randomly chosen locus in the SNP set has a specific pattern. To understand the effect of the subsample size on pattern probabilities, we modulated the sample size g in increments of 10, considering all multiples of 10 in $[10, 500]$.

Next, to study the numbers of super-populations in which variants are common and rare, we collapsed the 243 patterns into summaries that disregard the identities of the super-populations in which allelic types are unobserved, rare, and common. For these summaries, we track only the numbers of U's, R's, and C's for a given allelic type as an ordered triple $(|\mathrm{U}|, |\mathrm{R}|, |\mathrm{C}|)$. For example, if an allelic type has the pattern RUUUU, URUUU, UURUU, UUURU, or UUUUR, then it is summarized as $(4, 1, 0)$. The number of possible summaries is 21.

4.3.3 Sliding-window analysis

To examine the change in pattern probabilities along the genome, we calculated the probability distribution of patterns in sliding windows. We tiled the genome with non-overlapping 100-kb windows. Within each window, we averaged the 21 summaries across SNPs within the window, still focusing on the globally minor allele at each SNP. For this analysis, we focused on a single value of g, choosing $g = 500$, summarizing the patterns using the 21 ordered triples $(|U|, |R|, |C|)$.

4.3.4 Data availability

We downloaded publicly available data from the 1000 Genomes FTP site: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20190425_NYGC_GATK/. All code used for the analyses is available on GitHub: github.com/djcotter/rarefaction-rare-vs-common.

4.4 Results

4.4.1 Pointwise rarefaction analysis

Figure 4.1A shows the pointwise probabilities of the various patterns for 1,226,225 SNPs on chromosome 22. The figure visualizes the 11 patterns that have probability 1% or greater at $g = 250$, grouping the other 232 patterns into a single "other" category; this choice of the intermediate value of $g = 250$ facilitates visualization of patterns that are probable at high g or low g but not both.

The highest-frequency pattern for all sample sizes is UUUUU, the probability of observing no variation across the five super-populations; this pattern is the one most likely to be observed if an allelic type is present in the full data but extremely rare. Among the other high-frequency patterns, five of the next six represent allelic types that are rare in one super-population and absent in the other four; the sixth, RUUUR, is allelic types rare in both the African and American super-populations, likely a result of admixed African-descended populations in the American super-population. CCCCC is included; it is the only high-frequency pattern that includes any common variation.

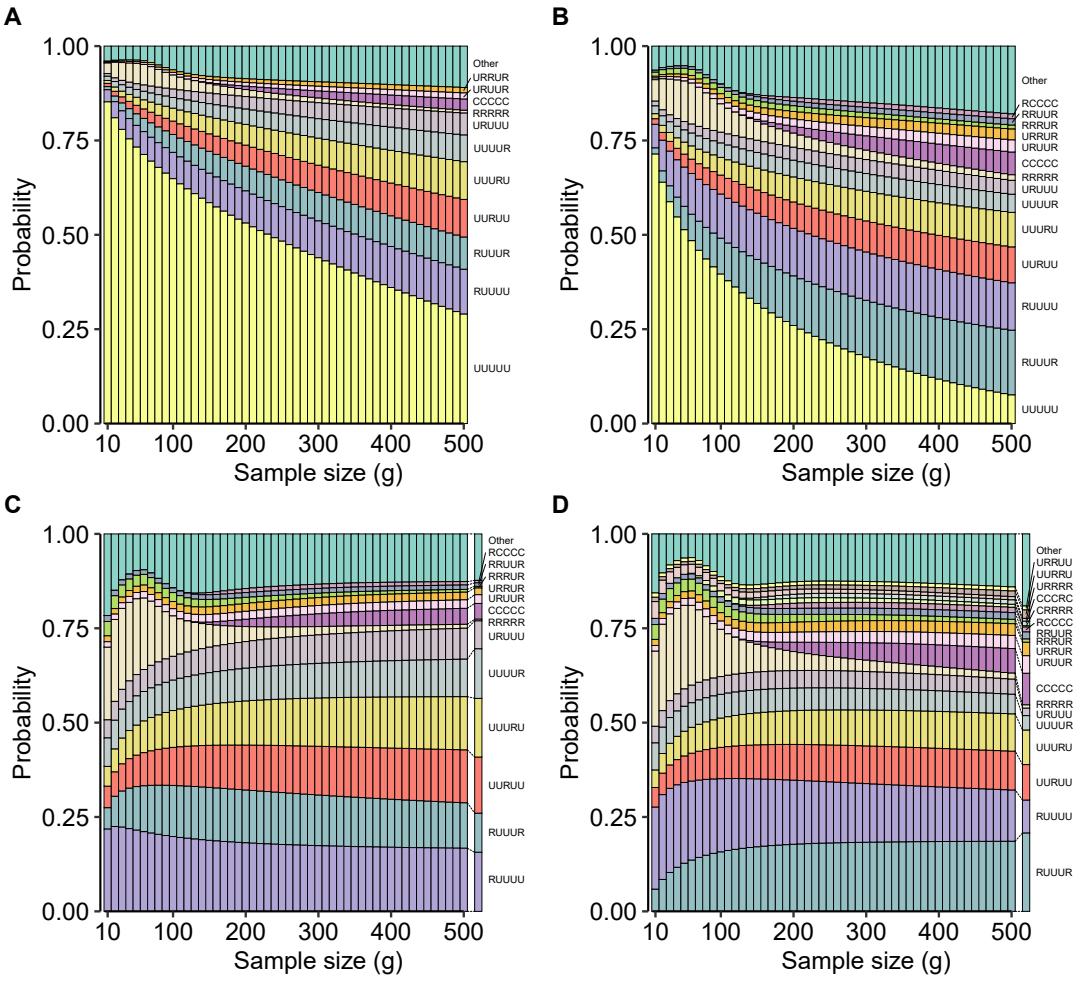

Figure 4.1: Probability that the globally minor allele at a locus has a given geographic distribution pattern as a function of g, the number of alleles sampled in each super-population (eq. 4.4). (A) All SNPs on chromosome 22. (B) All non-singleton SNPs on chromosome 22. (C) All SNPs on chromosome 22, normalizing by $1 - \mathbb{P}[\text{UUUUU}]$. (D) All non-singleton SNPs on chromosome 22, normalizing by $1 - \mathbb{P}[\text{UUUUU}]$. In a five-letter pattern, U is unobserved, R is rare (>0% and ≤ 5% population frequency), and C is common (>5%). The order in which super-populations are listed is Africa, Europe, South Asia, East Asia, and the Americas. For example, RUUUU refers to a minor allele that is rare in Africa and unobserved in each of the other four super-populations.

Increases in the sample size decrease the frequency of UUUUU and increase the frequencies of patterns containing rare allelic types. As the sample size increases, the probability increases that a rare variant is detected in a sample, so that previously unobserved allelic types are increasingly likely to be observed as rare.

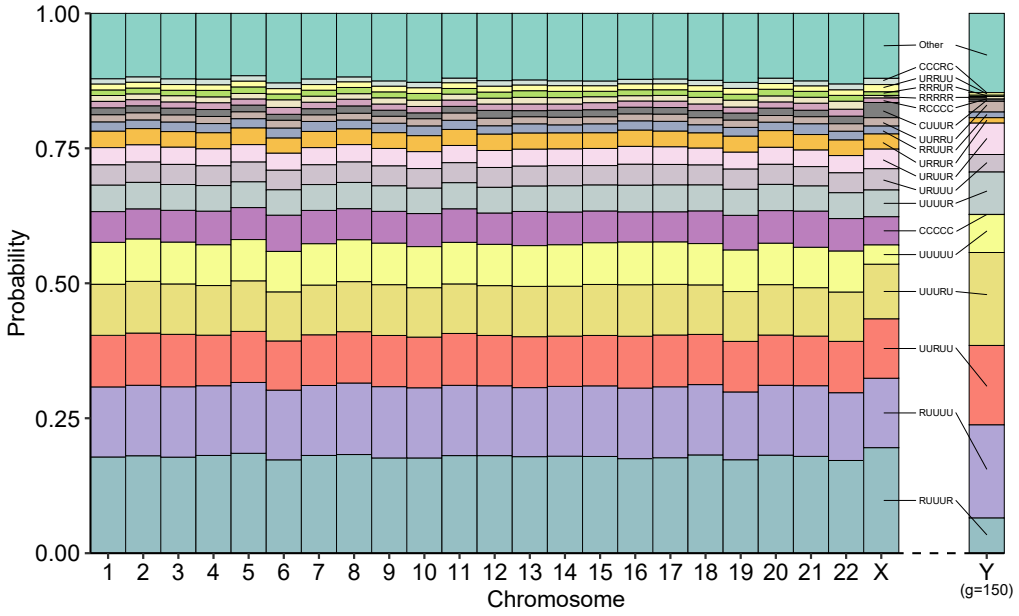

Figure 4.2: Probability that the globally minor allele at a locus has a given geographic distribution pattern, considering each of 22 autosomes and the two sex chromosomes. Probabilities are calculated using all non-singleton SNPs on each chromosome and $g = 500$ ($g = 150$ for the Y chromosome). Patterns that are present at greater than 1% frequency on any autosome are indicated, with all other patterns grouped into the "Other" category.

In Figure 4.1B, we analyze the effect of extremely rare allelic types on the patterns by discarding all 614,354 SNPs whose minor allele appears only once among the 2504 individuals, leaving 611,871 non-singleton SNPs on chromosome 22. By removing singletons, we deflate the UUUUU proportion, revealing patterns that previously grouped into the "other" category; the number of patterns with frequency at least 1% at $g = 250$ increases from 11 to 14. All 11 previous high-frequency patterns are observed, in addition to two in which allelic types are rare in multiple super-populations and unobserved in others (RRRUR, RRUUR) and one in which allelic types are common in some super-populations and rare in others (RCCCC). Patterns containing one R and four U's continue to be among the higher-frequency patterns, indicating that these patterns result from rare variation that is not limited to allelic types present in only a single copy. Similar observations hold genome-wide (Figure 4.2).

In Biddanda *et al.* (2020), without a sample-size correction, all loci are biallelic. Hence, no variant can be entirely unobserved, and Biddanda *et al.* (2020) did not consider the UUUUU pattern.

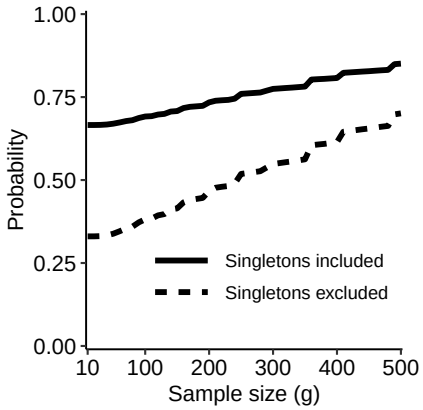

Figure 4.3: Probability as a function of the sample size g that across SNPs on chromosome 22, the highest-probability non-UUUUU pattern calculated using a sample-size correction (eq. 4.4) matches the empirically observed pattern without sample-size correction.

To facilitate a comparison of the relative probabilities of the remaining 242 patterns between our analysis and that of Biddanda *et al.* (2020), we remove the UUUUU pattern at each g and divide the remaining pattern frequencies by $1 - \mathbb{P}[\text{UUUUU}]$ (Figure 4.1C). With this normalization, three patterns with frequencies below 1% at $g = 250$ in Figure 4.1A now have frequencies greater than or equal to 1%: RRUUR, RCCCC, and RRRUR. For most patterns, the frequency is largely unaffected by changes in the sample size g. An interesting exception is RRRRR, for which a particularly strong effect of the discrete sample size is evident. At small g, this pattern is observed when exactly one copy of an allelic type is seen in each of the five super-populations; as described in the example in the Introduction, common allelic types with frequencies near the frequency cutoff between rare and common are mistakenly categorized as rare, and as the sample size g increases, it is possible to correctly determine that those allelic types are in fact common.

Figure 4.1C provides a comparison of the sample-size-corrected probabilities with the empirical pattern frequencies observed in the sample, the frequencies that correspond to the non-sample-size-corrected calculation of Biddanda *et al.* (2020) (rightmost column of Figure 4.1C). Although many of the corrected pattern frequencies differ from the uncorrected frequencies at small g, at $g = 500$, the sample-size-corrected probability of observing a given pattern is comparable to the

empirical frequency of that pattern in the full set of loci. A similar general agreement of the empirical frequency to sample-size-corrected probabilities at high g is observed in Figure 4.1D, with singletons excluded. As in the comparison of Figure 4.1B and Figure 4.1A, exclusion of singletons increases the number of patterns occurring at frequency ≥ 0.01 when $g = 250$, from 13 to 18. Exclusion of singletons reduces frequencies for patterns with one R and four U's, so that additional patterns cross the 1% threshold.

A further comparison of sample-size-corrected pattern frequencies with uncorrected frequencies appears in Figure 4.3. In this figure, we evaluate the fraction of loci for which the empirical pattern at a locus matches the (non-UUUUU) pattern with greatest sample-size-corrected probability. Performing this computation at each value of the sample size g, we observe that the probability that the empirical pattern is a match to the highest-probability pattern with sample-size correction increases with g (Figure 4.3). With singletons included, at $g = 10$, the probability of agreement is

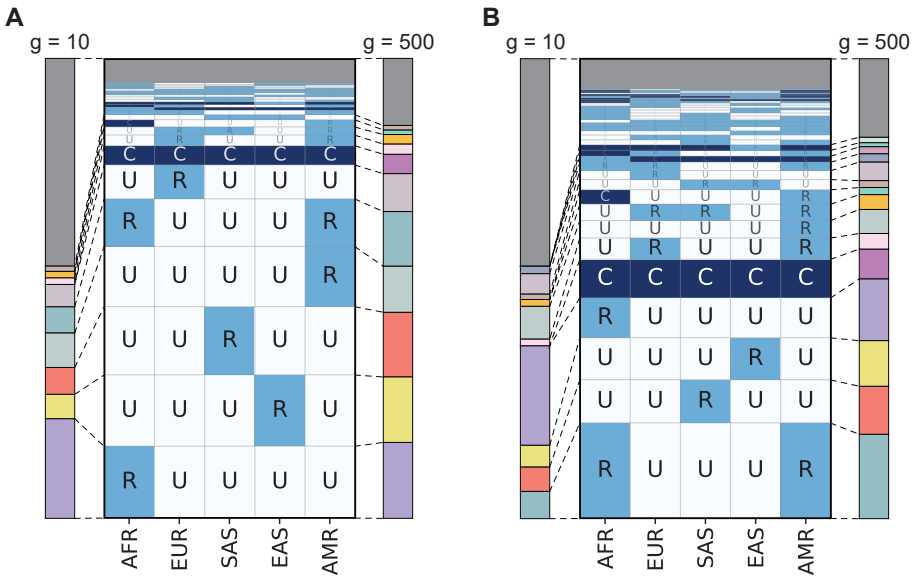

Figure 4.4: Pattern probabilities at $g = 10$ and $g = 500$ compared to non-sample-size-corrected pattern probabilities. The sample-size-corrected and non-sample-size-corrected probabilities are calculated on chromosome 22. (A) All SNPs on chromosome 22, as in Figure 4.1C, with non-sample-size-corrected pattern probabilities depicted analogously to Figure 3B of Biddanda *et al.* (2020). (B) Non-singleton SNPs on chromosome 22, as in Figure 4.1D, with non-sample-size-corrected pattern probabilities depicted analogously to Figure 3C of Biddanda *et al.* (2020). The colors used to depict pattern probabilities for $g = 10$ and $g = 500$ are the same as those used in Figure 4.1.

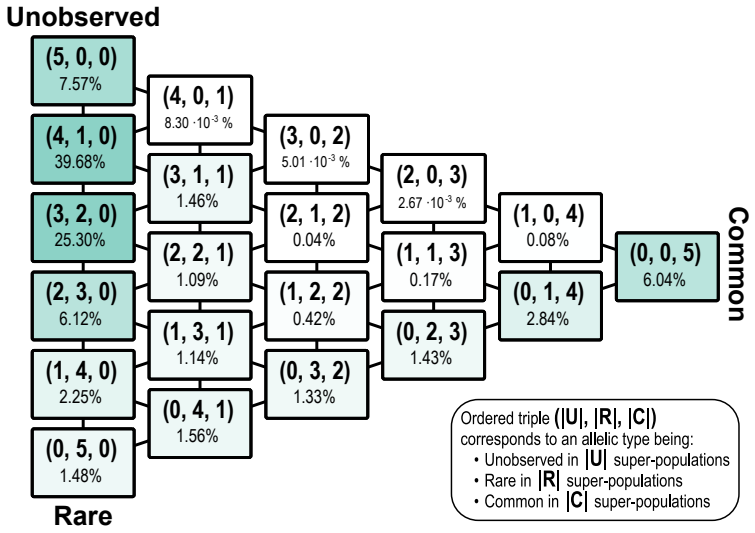

Figure 4.5: Probabilities for groups of patterns for a non-singleton minor allele on chromosome 22, in samples containing $g = 500$ alleles from each super-population. The figure summarizes the $g = 500$ column of Figure 4.1B, tabulating the numbers of super-populations in which allelic types are unobserved, rare, and common. An ordered triple is written $(|U|, |R|, |C|)$, so that, for example, 2.84% for the entry $(0, 1, 4)$ indicates that 2.84% of allelic types are unobserved in 0 super-populations, rare in 1 super-population, and common in 4 super-populations.

66.6%, and at $g = 500$, it is 85.1%. The probabilities are somewhat lower with singletons excluded; at $g = 10$, the agreement probability is 33.0%, and at $g = 500$, it is 70.1%. Because singleton loci can only take on one non-UUUUU pattern in the rarefaction calculation (rare in the one super-population where the allelic type is seen), given that they must be polymorphic in the empirical data, the empirical pattern necessarily agrees with the highest-probability sample-size-corrected non-UUUUU pattern.

We accentuate the comparison between sample-size-corrected and uncorrected pattern frequencies by depicting the non-UUUUU pattern frequencies at $g = 10$ and $g = 500$, alongside depictions of corresponding empirical pattern frequencies in the style of Biddanda *et al.* (2020) (Figure 4.4). At the smaller $g = 10$, common variation is unlikely: allelic types at the low end of the frequency interval for common variation are relatively unlikely to be sampled in such a small sample size, so that pattern CCCCC has a low probability. However, at $g = 500$, allelic types that are truly common are more likely to be detected as common. The pattern frequencies for large g generally agree with the empirical pattern frequencies without sample-size correction.

Figure 4.5 provides a summary of pattern frequencies at $g = 500$, collapsing the 243 patterns into 21 groups tabulating the numbers of super-populations in which allelic types are unobserved, rare, and common. Considering all 243 patterns and excluding singletons as in Figure 4.1B, we observe, as can be seen in Figure 4.1B, that the highest probabilities occur for groups

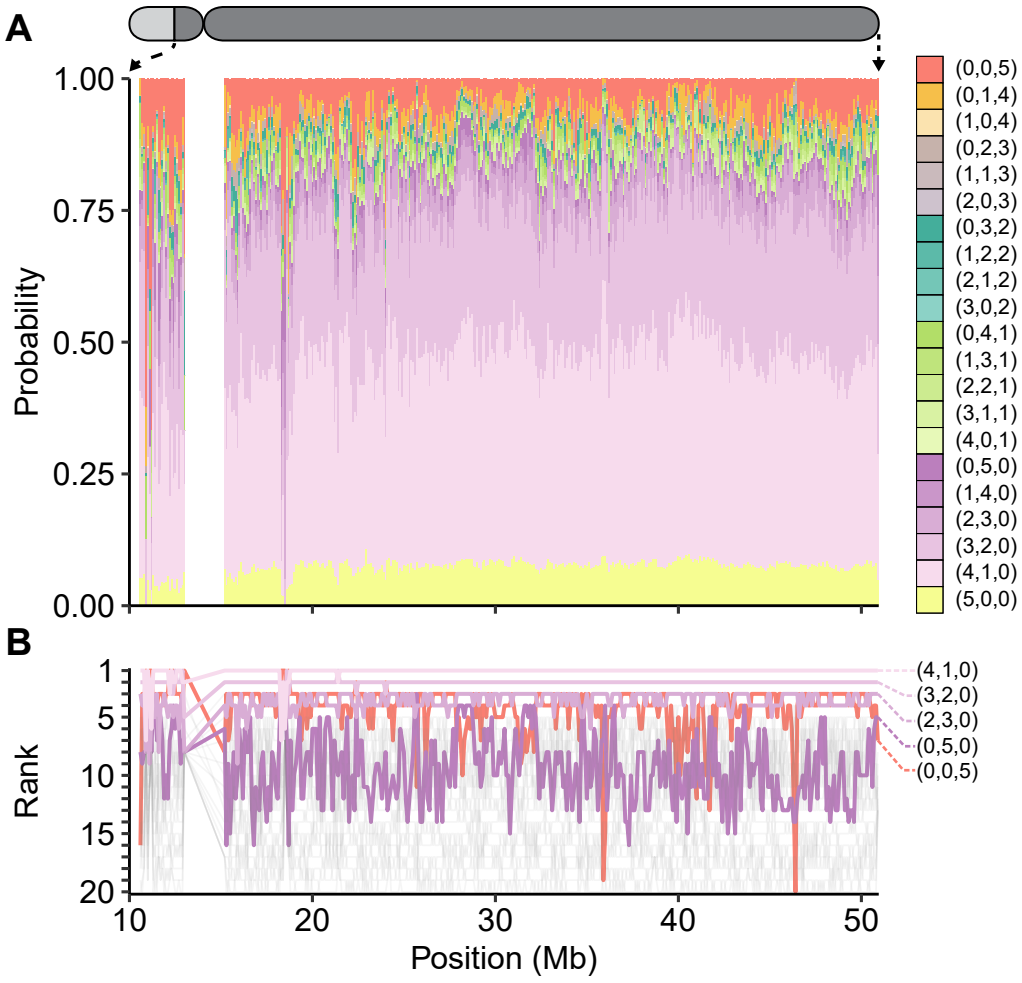

Figure 4.6: Probabilities for groups of patterns for minor alleles on chromosome 22, in samples containing $g = 500$ alleles from each super-population, averaged across all non-singleton loci in non-overlapping 100-kb sliding windows. Ordered triples are written $(|U|, |R|, |C|)$, with the entries representing the numbers of super-populations in which allelic types are unobserved, rare, and common, respectively. Triples are grouped by color, varying within classes with a given number of super-populations in which allelic types are common. (A) Probabilities for pattern groups. (B) Local frequency ranks of pattern groups, from 1 to 20 (the pattern in which allelic types are unobserved in all super-populations, $(5, 0, 0)$, is excluded). For simplicity, only those pattern groups that achieve frequency rank 1 or 2 in at least one window on the chromosome receive a color. The remaining pattern groups are shaded gray. Note that the first 10 Mb of chromosome 22 are excluded, as they do not appear in the 1000 Genomes dataset; the centromere is also excluded.

$(4, 1, 0)$, $(3, 2, 0)$, $(5, 0, 0)$, and $(2, 3, 0)$, representing allelic types that are rare or unobserved in all super-populations. Next in probability is $(0, 0, 5)$, representing allelic types that are common in all super-populations. Probabilities are particularly small for scenarios $(4, 0, 1)$, $(3, 0, 2)$, $(2, 0, 3)$, and $(1, 0, 4)$ representing variation that is common in some super-populations and unobserved in others.

4.4.2 Sliding-window analysis

Figure 4.6A shows the 21 groups of patterns as a function of genomic position in 100-kb, non-overlapping windows on chromosome 22, considering non-singleton loci and samples of size $g = 500$. In general, the probability distribution of the 21 groups shows little variation across the chromosome, mimicking the pointwise observations in Figure 4.5. The highest-probability pattern groups are generally those that represent allelic types that are rare in one or more super-populations and unobserved in the others. A relatively high probability also occurs for allelic types that are common in all five super-populations.

Figure 4.6B visualizes changes in rank for the groups of patterns as a function of position along the chromosome, highlighting the pattern groups that enter the top two ranks in at least one window. This visualization emphasizes that patterns in which allelic types are rare in one or two super-populations have the highest frequency in most windows. It also uncovers windows that show a difference from the chromosome-wide average. For example, between 18 and 19 Mb, a spike occurs in the probability that a minor allele is common in all five super-populations, and the group $(0, 0, 5)$, which often lies at rank 3, instead jumps to rank 1.

To illustrate one of many deviations from typical pattern probabilities that occur periodically across the genome (Figure 4.7), we consider an example. In particular, as local changes in the extent to which allelic types are globally common can reflect evolutionary processes such as balancing selection, we examine the local change in probabilities in the highly variable HLA region on chromosome 6 (Figure 4.8), where balancing selection is an important phenomenon (Meyer *et al.*, 2018). Interestingly, in the HLA region (28.5–33.5 Mb), the group $(0, 0, 5)$ has rank 1 in many windows, as might be expected for a region in which a balancing selection process maintains non-trivial frequencies for allelic types across many populations.

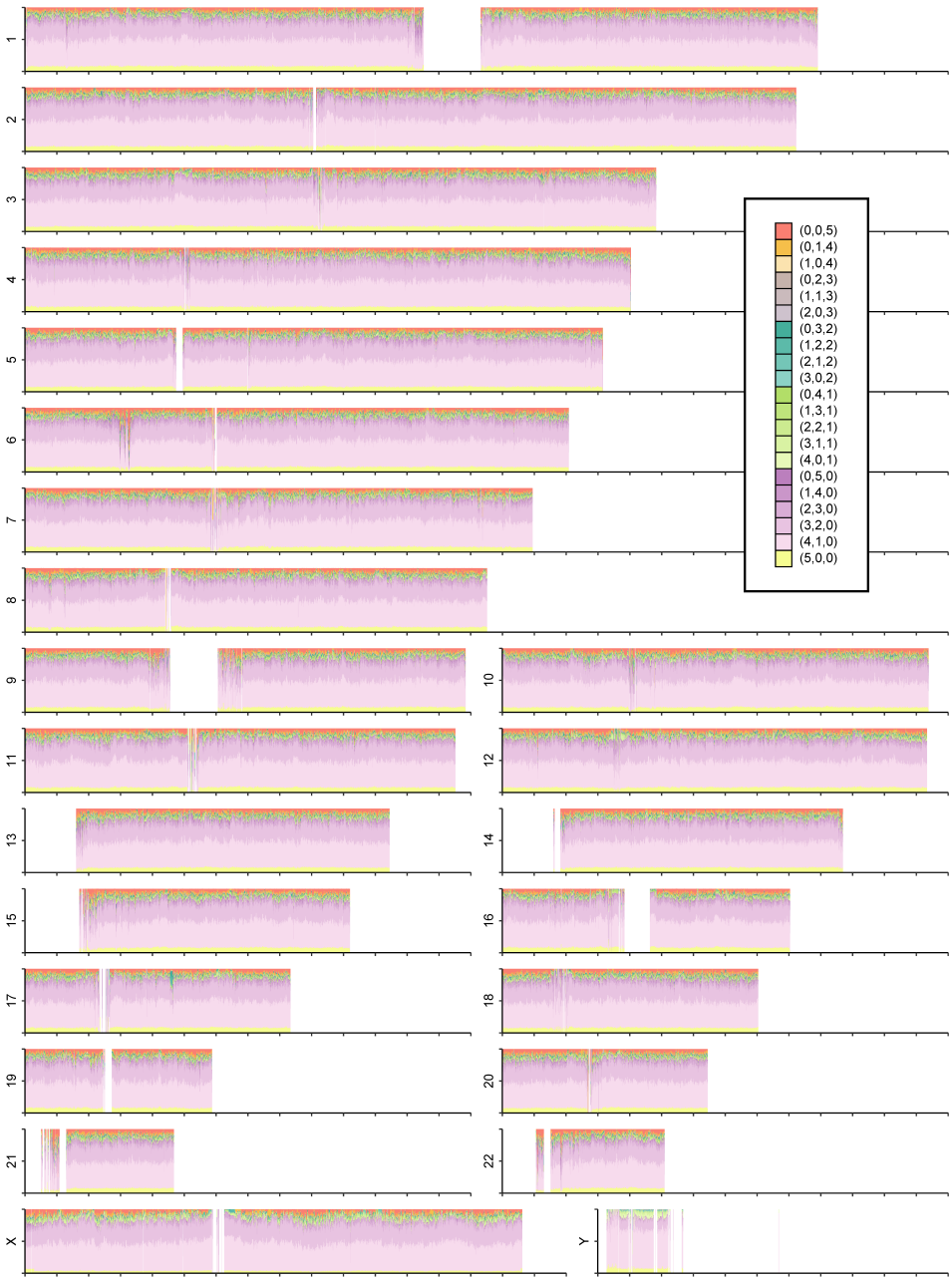

Figure 4.7: Probabilities for groups of patterns for minor alleles across all 22 autosomes and the two sex chromosomes, in samples containing $g = 500$ alleles from each super-population, averaged across all non-singleton loci in non-overlapping 100-kb sliding windows ($g = 150$ for the Y chromosome). Ordered triples are written $(|U|, |R|, |C|)$, with the entries representing the numbers of super-populations in which allelic types are unobserved, rare, and common, respectively. Triples are grouped by color, varying within classes with a given number of super-populations in which allelic types are common. Each X-axis tick mark corresponds to a distance of 10 Mb.

4.5 Discussion

We have introduced a method for obtaining sample-size-corrected pattern probabilities describing the geographic distribution of allelic types. The method combines the "Geovar" plots of Biddanda *et al.* (2020)—which describe the probabilities with which allelic types are unobserved, rare, or common in different population groups—with the rarefaction approach of Szpiech *et al.* (2008),

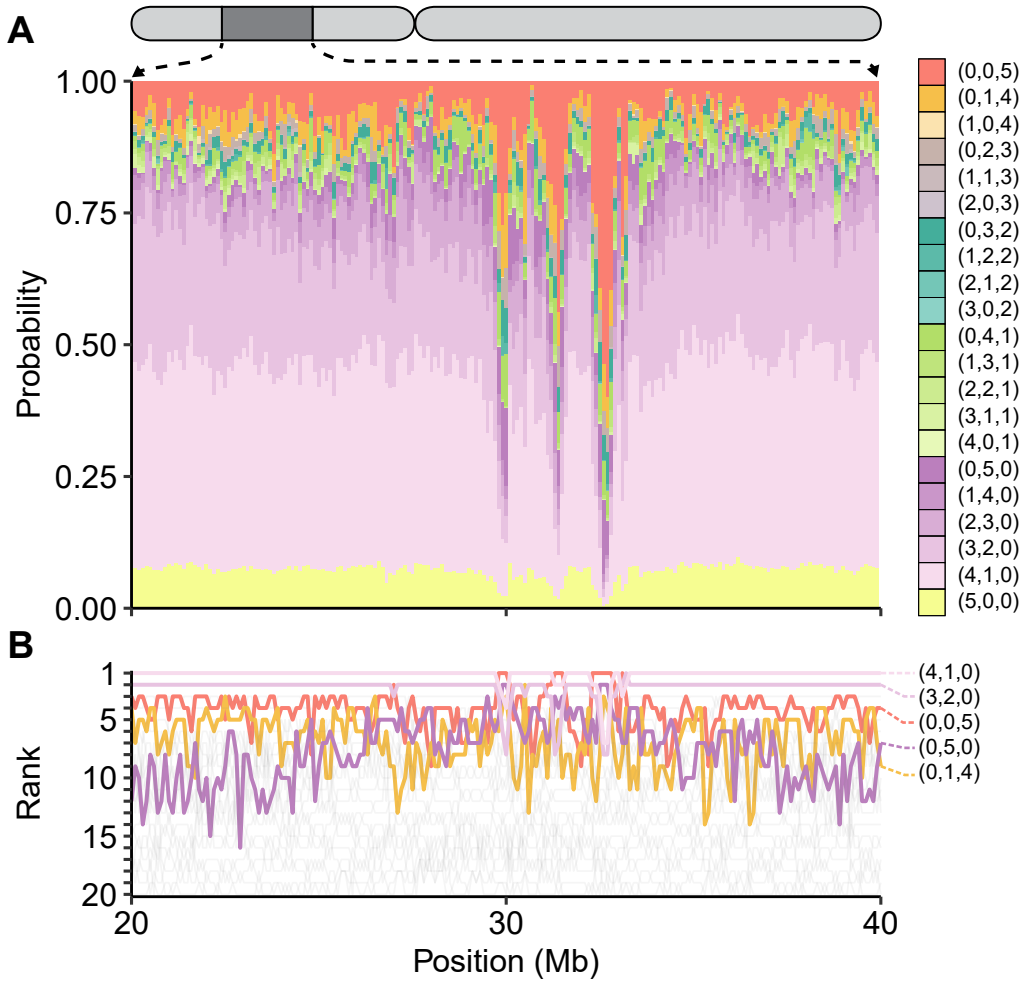

Figure 4.8: Probabilities for pattern groups for minor alleles of non-singleton loci appearing between 20–40 Mb on chromosome 6, covering the HLA region (approximately 28.5–33.5 Mb on reference build hg38). The data analysis and figure design follow Figure 4.6. (A) Probabilities for pattern groups. (B) Local frequency ranks of pattern groups.

which mathematically studies geographic distributions of allelic types in subsamples that have equal size in different groups.

Our analysis finds that with the use of a parameter g for the fixed sample size examined in each of the various groups, probabilities of allelic patterns do change somewhat (Figure 4.1). Most notably, as g increases, the probability of classifying an allelic type as entirely unobserved declines (Figure 4.1A,B). With this pattern omitted, pattern probabilities are relatively stable with g (Figures 4.1C,D). However, g must be sufficiently large before the stability emerges. In small samples, discreteness effects influence the probability that an allelic type is rare in all groups; in using the rarefaction approach to examine pattern probabilities, such effects can potentially be mitigated by increasing the maximal frequency regarded as rare in small samples. Such an approach might be warranted in cases in which some of the groups of interest have samples that are much smaller than those of other groups, such as in comparisons involving ancient and modern data; suitable choices of frequency thresholds will depend on the specific sample sizes in data sets and on the underlying distribution of true allele frequencies. Conversely, if all sample sizes are extremely large, it may be convenient to use eq. 4.6 to distinguish multiple tiers of rare allelic types, for example for separating frequency classes rare enough to be restricted to one group from higher-frequency classes whose allelic types are still rare but likely to be found in multiple populations.

With a large sample size, $g = 500$, our pattern probabilities with a sample-size correction closely match those observed without a sample-size correction in the manner of Biddanda *et al.* (2020) (Figure 4.4). This general agreement suggests that the sample sizes in the Biddanda *et al.* (2020) super-population assignment—504, 404, 489, 504, and 603 individuals for AFR, EUR, SAS, EAS, and AMR, respectively—are sufficiently large that differences among them likely had little effect on the non-sample-size-corrected pattern probabilities of Biddanda *et al.* (2020) when using 5% as the demarcation between rare and common allelic types. In particular, our pattern probability calculations with sample-size corrections recapitulate the finding that most allelic types are rare in one or a few super-populations and unobserved in the others, or common in all super-populations (Figure 4.5).

The work of Biddanda *et al.* (2020) is motivated by a goal not only of describing features of human genetic similarity and difference, it is also one of many examples of studies that place particular emphasis on new visualizations to capture those features (e.g. Mountain and Ramakrishnan, 2005; Conrad *et al.*, 2006; Pickrell *et al.*, 2009; Teo and Small, 2010; Rosenberg, 2011; San Lucas *et al.*, 2012; Petkova *et al.*, 2016; Marcus and Novembre, 2017; Diaz-Papkovich *et al.*, 2019; Greenbaum *et al.*, 2019; Peter *et al.*, 2020; Battey *et al.*, 2021). Such visualizations provide new representations of population-genetic statistics for use in understanding processes that affect genetic variation across populations. Emphases on visualization have been of increasing interest in light of ongoing misrepresentations of human population-genetic findings—particularly the misuse of graphical visualizations as apparent evidence of unsupportable views of human difference belied by the analyses that underlie the graphics (Carlson *et al.*, 2022). Pattern probabilities, such as those we have considered here and those of Biddanda *et al.* (2020), enable a variety of visualizations of human variation beyond the "Geovar" style. Our Figure 4.1, describing pattern probabilities in the categories "unobserved," "rare," and "common," updates visualizations of the sample-size-corrected pattern probabilities of Szpiech *et al.* (2008), which grouped rare and common allelic types in a single category of "observed" allelic types. Figure 4.5, summarizing pattern probabilities by the numbers of super-populations in which allelic types are unobserved, rare, and common, updates similar summaries that also did not distinguish between rare and common allelic types (Rosenberg *et al.*, 2002, Figure S1A; Jakobsson *et al.*, 2008, Figure 1A; Rosenberg, 2011, Figure 4A and Table 2; The 1000 Genomes Project Consortium, 2015, Figure 1A). Finally, Figure 4.6 illustrates that pattern probabilities can be considered locally as a function of genomic position; this form of analysis can also suggest signatures of population-genetic processes such as balancing selection in the HLA region (Figure 4.8).

Our analysis has made use of dense human genomic data. For genomes with a higher density of variants than the human genome, shorter window sizes may be convenient for measurement of pattern probabilities. For lower-density data, longer window sizes might be required for accumulating enough variable sites to accurately measure pattern probabilities. Even in the data we have examined, data quality might vary across windows; this problem might affect the HLA region, in which high variation levels can lead through technical artifacts to biased estimation of allele

frequencies (Brandt *et al.*, 2015). The window size can be tuned appropriately to the analysis of interest.

Our observation that allelic types are generally rare in some human groups and unobserved in others, or common in most or all groups—here seen with a rarefaction method—has been consistently observed across datasets and choices of population groups (Cavalli-Sforza *et al.*, 1994; The International HapMap 3 Consortium, 2010; Rosenberg, 2011; The 1000 Genomes Project Consortium, 2015; Biddanda *et al.*, 2020). Analyses enabled by a focus on pattern probabilities, with the improvements from the sample-size correction introduced here, provide new approaches to emphasizing and visualizing this fundamental result in human evolutionary genetics.

Acknowledgments. We thank two anonymous reviewers for their careful reading of the manuscript. We acknowledge NIH R01 HG005855, NIH R35 GM146926, and NSF BCS-2116322 for support.

Conclusion

By studying human genetic variation both theoretically and empirically, population geneticists are able to build an accurate understanding of the history of the human species. Here, I have contributed to advancing that understanding through the development of novel models of coalescence times as well as empirical calculations of genetic variation. In chapters 1–3, I have shown how consanguinity shapes runs of homozygosity and identity-by-descent sharing. These chapters advance the understanding of a process that has significant consequences for the study of rare disease in different populations. Simultaneously, in chapter 4, I have developed a framework for conceptualizing the effects that population-size differences have on the study of rare and common variation in the human genome.

Bibliography

Arbiza, L., S. Gottipati, A. Siepel, and A. Keinan, 2014 Contrasting X-linked and autosomal diversity across 14 human populations. American Journal of Human Genetics **94**: 827–844.

Arciero, E., S. A. Dogra, D. S. Malawsky, M. Mezzavilla, T. Tsismentzoglou, *et al.*, 2021 Fine-scale population structure and demographic history of British Pakistanis. Nature Communications **12**: 7189.

Battey, C. J., G. C. Coffing, and A. D. Kern, 2021 Visualizing population structure with variational autoencoders. G3 Genes|Genomes|Genetics **11**: jkaa036.

Behar, D. M., M. Metspalu, Y. Baran, N. M. Kopelman, B. Yunusbayev, *et al.*, 2013 No evidence from genome-wide data of a Khazar origin for the Ashkenazi Jews. Human Biology **85**: 859–900.

Biddanda, A., D. P. Rice, and J. Novembre, 2020 A variant-centric perspective on geographic patterns of human allele frequency variation. eLife **9**: e60107.

Bittles, A., 2001 Consanguinity and its relevance to clinical genetics. Clinical Genetics **60**: 89–98.

Bittles, A. H., 2012 *Consanguinity in context*. Cambridge University Press, Cambridge.

Bittles, A. H. and M. L. Black, 2010 Consanguinity, human evolution, and complex diseases. Proceedings of the National Academy of Sciences **107**: 1779–1786.

Brandt, D. Y. C., V. R. C. Aguiar, B. D. Bitarello, K. Nunes, J. Goudet, *et al.*, 2015 Mapping bias overestimates reference allele frequencies at the HLA genes in the 1000 Genomes Project phase I data. G3 Genes|Genomes|Genetics **5**: 931–941.

Browning, B. L. and S. R. Browning, 2013 Detecting identity by descent and estimating genotype error rates in sequence data. American Journal of Human Genetics **93**: 840–851.

Browning, S. R. and B. L. Browning, 2007 Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. American Journal of Human Genetics **81**: 1084–1097.

Browning, S. R. and B. L. Browning, 2012 Identity by descent between distant relatives: detection and applications. Annual Review of Genetics **46**: 617–633.

Browning, S. R. and B. L. Browning, 2015 Accurate non-parametric estimation of recent effective population size from segments of identity by descent. American Journal of Human Genetics **97**: 404–418.

Buffalo, V., S. M. Mount, and G. Coop, 2016 A genealogical look at shared ancestry on the X chromosome. Genetics **204**: 57–75.

Bustamante, C. D. and S. Ramachandran, 2009 Evaluating signatures of sex-specific processes in the human genome. Nature Genetics **41**: 8–10.

Byrska-Bishop, M., U. S. Evani, X. Zhao, A. O. Basile, H. J. Abel, *et al.*, 2022 High-coverage whole-genome sequencing of the expanded 1000 genomes project cohort including 602 trios. Cell **185**: 3426–3440.e19.

Cai, R., B. L. Browning, and S. R. Browning, 2022 IBD-based estimation of X chromosome effective population size with application to sex-specific demographic history. bioRxiv p. 10.1101/2022.07.06.499007.

Campbell, R., 2015 The effect of inbreeding constraints and offspring distribution on time to the most recent common ancestor. Journal of Theoretical Biology **382**: 74–80.

Carlson, J., B. M. Henn, D. R. Al-Hindi, and S. Ramachandran, 2022 Counter the weaponization of genetics research by extremists. Nature **610**: 444–447.

Carmi, S., P. R. Wilton, J. Wakeley, and I. Pe'er, 2014 A renewal theory approach to IBD sharing. Theoretical Population Biology **97**: 35–48.

Cavalli-Sforza, L. L., P. Menozzi, and A. Piazza, 1994 *The History and Geography of Human Genes*. Princeton University Press, Princeton, NJ.

Ceballos, F. C., P. K. Joshi, D. W. Clark, M. Ramsay, and J. F. Wilson, 2018 Runs of homozygosity: windows into population history and trait architecture. Nature Reviews Genetics **19**: 220–234.

Chen, N., E. J. Cosgrove, R. Bowman, J. W. Fitzpatrick, and A. G. Clark, 2016 Genomic Consequences of Population Decline in the Endangered Florida Scrub-Jay. Current Biology **26**: 2974–2979.

Clark, D. W., Y. Okada, K. H. S. Moore, D. Mason, N. Pirastu, *et al.*, 2019 Associations of autozygosity with a broad range of human phenotypes. Nature Communications **10**: 4957.

Cohen, T., R. Vardi-Saliternik, and Y. Friedlander, 2004 Consanguinity, intracommunity and intercommunity marriages in a population sample of Israeli Jews. Annals of Human Biology **31**: 38–48.

Conrad, D. F., M. Jakobsson, G. Coop, X. Wen, J. D. Wall, *et al.*, 2006 A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. Nature Genetics **38**: 1251–1260.

Cotter, D. J., E. F. Hofgard, J. Novembre, Z. A. Szpiech, and N. A. Rosenberg, 2023 A rarefaction approach for measuring population differences in rare and common variation. GENETICS **224**: iyad070.

Cotter, D. J., A. L. Severson, S. Carmi, and N. A. Rosenberg, 2022 Limiting distribution of X-chromosomal coalescence times under first-cousin consanguineous mating. Theoretical Population Biology **147**: 1–15.

Cotter, D. J., A. L. Severson, and N. A. Rosenberg, 2021 The effect of consanguinity on coalescence times on the X chromosome. Theoretical Population Biology **140**: 32–43.

Crow, J. and M. Kimura, 1970 *An Introduction to Population Genetics Theory*. Harper and Row, New York.

Cussens, J. and N. A. Sheehan, 2016 Special issue on new developments in relatedness and relationship estimation. Theoretical Population Biology **107**: 1–3.

Diaz-Papkovich, A., L. Anderson-Trocmé, C. Ben-Eghan, and S. Gravel, 2019 UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. PLoS Genetics **15**: e1008432.

Ellegren, H., 2009 The different levels of genetic diversity in sex chromosomes and autosomes. Trends in Genetics **25**: 278–284.

Goldberg, A. and N. A. Rosenberg, 2015 Beyond 2/3 and 1/3: The complex signatures of sex-biased admixture on the X chromosome. Genetics **201**: 263–279.

Goldschmidt, E., A. Ronen, and I. Ronen, 1960 Changing marriage systems in the Jewish communities of Israel. Annals of Human Genetics **24**: 191–204.

Gotelli, N. J. and R. K. Colwell, 2001 Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. Ecology Letters **4**: 379–391.

Greenbaum, G., A. Rubin, A. R. Templeton, and N. A. Rosenberg, 2019 Network-based hierarchical population structure analysis for large genomic data sets. Genome Research **29**: 2020–2033.

Hedrick, P. W., 2007 Sex: differences in mutation, recombination, selection, gene flow, and genetic drift. Evolution **61**: 2750–2771.

Hein, J., M. Schierup, and C. Wiuf, 2005 *Gene Genealogies, Variation and Evolution*. Oxford University Press, New York.

Hill, W. G., 1996 Sewall Wright's "Systems of Mating". Genetics **143**: 1499–1506.

Hurlbert, S. H., 1971 The nonconcept of species diversity: A critique and alternative parameters. Ecology **52**: 577–586.

Jacquard, A., 1974 *The Genetic Structure of Populations*. Springer, Berlin.

Jakobsson, M., S. W. Scholz, P. Scheet, J. R. Gibbs, J. M. VanLiere, *et al.*, 2008 Genotype, haplotype and copy-number variation in worldwide human populations. Nature **451**: 998–1003.

Johnson, E. C., L. M. Evans, and M. C. Keller, 2018 Relationships between estimated autozygosity and complex traits in the UK Biobank. PLoS Genetics **14**: e1007556.

Johnson, N. L., S. Kotz, and N. Balakrishnan, 1994 *Continuous Univariate Distributions, Volume 1*. Wiley Series in Probability and Statistics, John Wiley & Sons, Nashville, TN, second edition.

Kalinowski, S. T., 2004 Counting alleles with rarefaction: Private alleles and hierarchical sampling designs. Conservation Genetics **5**: 539–543.

Kang, J. T., A. Goldberg, M. D. Edge, D. M. Behar, and N. A. Rosenberg, 2016 Consanguinity rates predict long runs of homozygosity in Jewish populations. Human Heredity **82**: 87–102.

Kemeny, J. G. and J. L. Snell, 1983 *Finite Markov Chains*. Springer-Verlag, New York.

Kent, W. J., C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, *et al.*, 2002 The human genome browser at UCSC. Genome Research **12**: 996–1006.

Kirin, M., R. McQuillan, C. S. Franklin, H. Campbell, P. M. McKeigue, *et al.*, 2010 Genomic runs of homozygosity record population history and consanguinity. PLoS One **5**: e13996.

Kong, A., D. F. Gudbjartsson, J. Sainz, G. M. Jonsdottir, S. A. Gudjonsson, *et al.*, 2002 A high-resolution recombination map of the human genome. Nature Genetics **31**: 241–247.

Lange, K., 2002 *Mathematical and Statistical Methods for Genetic Analysis*. Statistics for Biology and Health, Springer, New York.

Marcus, J. H. and J. Novembre, 2017 Visualizing the geography of genetic variants. Bioinformatics **33**: 594–595.

McQuillan, R., A.-L. Leutenegger, R. Abdel-Rahman, C. S. Franklin, M. Pericic, *et al.*, 2008 Runs of homozygosity in European populations. American Journal of Human Genetics **83**: 359–372.

Meyer, D., V. R. C. Aguiar, B. D. Bitarello, D. Y. C. Brandt, and K. Nunes, 2018 A genomic perspective on HLA evolution. Immunogenetics **70**: 5–27.

Möhle, M., 1998 A convergence theorem for markov chains arising in population genetics and the coalescent with selfing. Advances in Applied Probability **30**: 493–512.

Mooney, J. A., A. Yohannes, and K. E. Lohmueller, 2021 The impact of identity by descent on fitness and disease in dogs. Proceedings of the National Academy of Sciences **118**: e2019116118.

Mountain, J. L. and U. Ramakrishnan, 2005 Impact of human population history on distributions of individual-level genetic distance. Human Genomics **2**: 4–19.

Nordborg, M. and P. Donnelly, 1997 The coalescent process with selfing. Genetics **146**: 1185–1195.

Nordborg, M. and S. M. Krone, 2002 Separation of time scales and convergence to the coalescent in structured populations. In *Modern Developments in Theoretical Population Genetics*, edited by M. Slatkin and M. Veuille, pp. 194–232, Oxford University Press, New York.

Palamara, P. F., T. Lencz, A. Darvasi, and I. Pe'er, 2012 Length distributions of identity by descent reveal fine-scale demographic history. American Journal of Human Genetics **91**: 809–822.

Pemberton, T. J., D. Absher, M. W. Feldman, R. M. Myers, N. A. Rosenberg, *et al.*, 2012 Genomic patterns of homozygosity in worldwide human populations. American Journal of Human Genetics **91**: 275–292.

Peter, B. M., D. Petkova, and J. Novembre, 2020 Genetic landscapes reveal how human genetic diversity aligns with geography. Molecular Biology and Evolution **37**: 943–951.

Petkova, D., J. Novembre, and M. Stephens, 2016 Visualizing spatial population structure with estimated effective migration surfaces. Nature Genetics **48**: 94–100.

Pickrell, J. K., G. Coop, J. Novembre, S. Kudaravalli, J. Z. Li, *et al.*, 2009 Signals of recent positive selection in a worldwide sample of human populations. Genome Research **19**: 826–837.

Pizzari, T., H. Løvlie, and C. K. Cornwallis, 2004 Sex–specific, counteracting responses to inbreeding in a bird. Proceedings of the Royal Society of London. Series B: Biological Sciences **271**: 2115–2121.

Pollak, E., 1987 On the theory of partially inbreeding finite populations. I. Partial selfing. Genetics **117**: 353–360.

Ramachandran, S., N. A. Rosenberg, M. W. Feldman, and J. Wakeley, 2008 Population differentiation and migration: Coalescence times in a two-sex island model for autosomal and X-linked loci. Theoretical Population Biology **74**: 291–301.

Romeo, G. and A. H. Bittles, 2014 Consanguinity in the contemporary world. Human Heredity **77**: 6–9.

Rosenberg, N. A., 2011 A population-genetic perspective on the similarities and differences among worldwide human populations. Human Biology **83**: 659–684.

Rosenberg, N. A., J. K. Pritchard, J. L. Weber, H. M. Cann, K. K. Kidd, *et al.*, 2002 Genetic structure of human populations. Science **298**: 2381–2385.

Sahoo, S. A., A. A. Zaidi, S. Anagol, and I. Mathieson, 2021 Long runs of homozygosity are correlated with marriage preferences across global population samples. Human Biology **93**: 201–216.

San Lucas, F. A., N. A. Rosenberg, and P. Scheet, 2012 Haploscope: a tool for the graphical display of haplotype structure in populations. Genetic Epidemiology **36**: 17–21.

Schield, D. R., E. S. C. Scordato, C. C. R. Smith, J. K. Carter, S. I. Cherkaoui, *et al.*, 2021 Sex-linked genetic diversity and differentiation in a globally distributed avian species complex. Molecular Ecology **30**: 2313–2332.

Severson, A. L., S. Carmi, and N. A. Rosenberg, 2019 The effect of consanguinity on between-individual identity-by-descent sharing. Genetics **212**: 305–316.

Severson, A. L., S. Carmi, and N. A. Rosenberg, 2021 Variance and limiting distribution of coalescence times in a diploid model of a consanguineous population. Theoretical Population Biology **139**: 50–65.

Sjödin, P., I. Kaj, S. Krone, M. Lascoux, and M. Nordborg, 2005 On the Meaning and Existence of an Effective Population Size. Genetics **169**: 1061–1070.

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.

Teo, Y. Y. and K. S. Small, 2010 A novel method for haplotype clustering and visualization. Genetic Epidemiology **34**: 34–41.

The 1000 Genomes Project Consortium, 2015 A global reference for human genetic variation. Nature **526**: 68–74.

The International HapMap 3 Consortium, 2010 Integrating common and rare genetic variation in diverse human populations. Nature **467**: 52–58.

Thompson, E. A., 2013 Identity by descent: variation in meiosis, across genomes, and in populations. Genetics **194**: 301–326.

Tsafrir, J. and I. Halbrecht, 1972 Consanguinity and marriage systems in the Jewish community in Israel. Annals of Human Genetics **35**: 343–347.

Verdu, P., T. J. Pemberton, R. Laurent, B. M. Kemp, A. Gonzalez-Oliver, *et al.*, 2014 Patterns of admixture and population structure in Native populations of northwest North America. PLoS Genetics **10**: e1004530.

Wakeley, J., 2009 *Coalescent Theory: an Introduction*. Roberts & Co., Greenwood Village, CO.

Webster, T. H. and M. A. Wilson Sayres, 2016 Genomic signatures of sex-biased demography: progress and prospects. Current Opinion in Genetics and Development **41**: 62–71.

Wilkins, J. F. and F. W. Marlowe, 2006 Sex-biased migration in humans: what should we expect from genetic data? BioEssays **28**: 290–300.

Witt, K. E., F. Villanea, E. Loughran, X. Zhang, and E. Huerta-Sanchez, 2022 Apportioning archaic variants among modern populations. Philosophical Transactions of the Royal Society B: Biological Sciences **377**: 20200411.

Woods, C. G., J. Cox, K. Springell, D. J. Hampshire, M. D. Mohamed, *et al.*, 2006 Quantification of homozygosity in consanguineous individuals with autosomal recessive disease. American Journal of Human Genetics **78**: 889–896.

Wright, S., 1921 Systems of mating. II. The effects of inbreeding on the genetic composition of a population. Genetics **6**: 124–143.

Yengo, L., Z. Zhu, N. R. Wray, B. S. Weir, J. Yang, *et al.*, 2017 Detection and quantification of inbreeding depression for complex traits from SNP data. Proceedings of the National Academy of Sciences of the United States of America **114**: 8602–8607.