Applications of Identity-By-Descent Analysis in Population Genetics Research: Methods for Demographic History Inference and Genetic Association Analysis

Ruoyi Cai

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2025

Reading Committee:

Sharon R. Browning, Chair

Elizabeth Blue

Guanghao Qi

Program Authorized to Offer Degree:

Biostatistics

University of Washington

Abstract

Applications of Identity-By-Descent Analysis in Population Genetics Research: Methods for Demographic History Inference and Genetic Association Analysis

Ruoyi Cai

Chair of the Supervisory Committee:

Sharon R. Browning

Department of Biostatistics

The analysis of identity-by-descent (IBD) segments is an important tool in population genetics research and has led to important discoveries about human genetic history and population structures. While previous research has made significant progress in utilizing IBD segments to study population genetics, ongoing methodological advancements and increasing availability of genomic data create new opportunities for further exploration of the power of IBD segments in population genetics research. In this dissertation, we further explore the potential of IBD information in two key areas of genetics research: inferring demographic history and performing genetic association analysis (IBD mapping) for complex traits. First, we present a method to estimate the X chromosome effective population size using X chromosome IBD segments, and we demonstrate how the X chromosome effective population size can be combined with

autosomal effective population size to inform sex-specific demographic history. Second, we introduce an IBD mapping approach for association analysis between genome-wide loci and complex traits, along with a novel multiple testing adjustment strategy that accounts for the correlation structure among test statistics in genome-wide IBD scans. Our research contributes new statistical and computational tools that enhance the use of IBD information in demographic inference and genetic association studies, improving our understanding of human evolutionary history and the genetic architecture of complex traits.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

As I reflect on the past 10 years since I left my hometown of Beijing to pursue higher education in the U.S., I realize how challenging this journey has been—especially with the turbulence outside and the pandemic. There were many moments when I felt overwhelmed, frustrated, and questioned whether I could make it to the end. It is the love and support of my mentors, family, and friends that have carried me through these difficult times and helped me reach this point.

First and foremost, I am sincerely grateful to my Ph.D. advisor, Dr. Sharon Browning, for her invaluable guidance in both my research and professional development. Sharon's kindness, patience, understanding, and unwavering support have been instrumental not only in my academic progress but also in my personal life and career. As both a brilliant scholar and a wonderful mentor, she introduced me to the field of statistical genetics and imparted her knowledge and expertise. More importantly, she nurtured me into an independent researcher with a dedicated, rigorous altitude to science.

I also wish to express my gratitude to my dissertation committee members—Dr. Brian Browning, Dr. Elizabeth Blue, Dr. Ellen Wijsman, and Dr. Guanghao Qi—for their unique perspective and thoughtful feedback on my research. Their insights and encouragement have been invaluable. I am particularly grateful to Dr. Brian Browning and Dr. Elizabeth Blue for their support during my internship and job search. I would also like to thank my peers and colleagues in the Browning Lab, especially Seth Temple and Nobu Masaki, for the inspiring discussions I have had with them about my research.

Outside of my committee, I have received tremendous mentorship and support from the faculty and staff in the Biostatistics Department. I am thankful to Dr. Ken Rice and Dr. Lurdes Inoue for their help and guidance while I served as a teaching assistant in their courses, as well as their dedication to the Ph.D. program. I also appreciate the warm mentorship of Gitana Garofalo, whose support helped me navigate the early years of my Ph.D. study. I am also grateful for the indispensable assistance that Deb Nelson, Minh Vo, and Maggie Tarnawa have provided for administrative procedures, particularly as I approached the final steps toward graduation.

I also want to extend my heartfelt thanks to mentors outside of UW. I appreciate the guidance from Dr. Prateek Gundannavar and Dr. Anil Raj during my internship. I also feel compelled to

thank Dr. Audrey Gasch, Dr. Cécile Ané, and Dr. Hyunseung Kang, who mentored me during my undergraduate years. They helped me find my passion for statistics and genetics, and their encouragement motivated my decision to pursue a Ph.D. degree in biostatistics.

Lastly, I am incredibly grateful to my family, friends, and peers who made the non-academic parts of my life so enriching and memorable. My deepest thanks go to my parents for their unconditional love and support at every stage of my life. Without their support, I would never have had the opportunity to study in the U.S. or reach where I am today. I am also thankful for my peers in the Biostatistics and Statistics departments for all the study groups, random chats, and fun activities over the past six years. I would like to thank my friends Zeyu Wei and Ellen Xing, along with their adorable kittens, for bringing joy and relaxation into my life outside of school.

Most importantly, I want to thank my husband, Yichen Yang, whose love, support, and encouragement have been my anchor through the toughest times. As both a talented statistician and a wonderful partner, his unwavering emotional support helped me believe in myself during difficult moments, motivated me to keep moving forward, and helped me grow into a better version of myself.

There is still much ahead as I continue to pursue my passion for statistical genetics in the biomedical field. I will carry forward the love, guidance, and support I have received throughout my journey thus far to remain brave in the face of new challenges.

DEDICATION

to my family, with special thanks to my husband,
for their unwavering love, support,
and encouragement throughout this journey

Chapter 1

INTRODUCTION

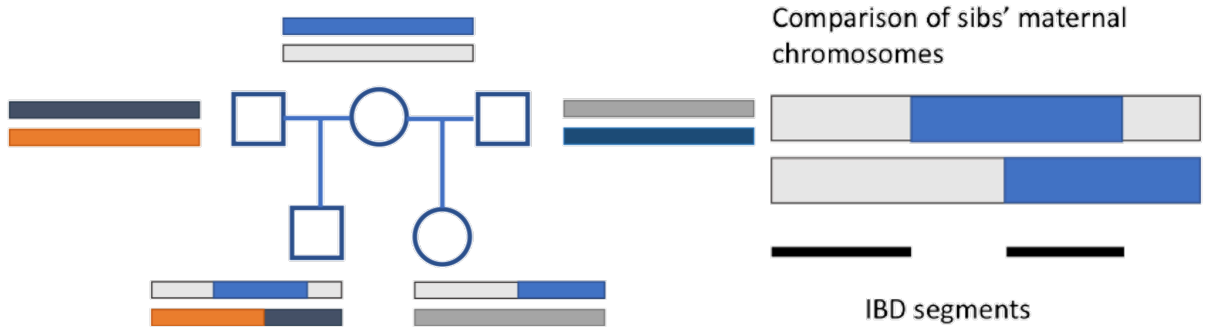
1.1 Identity by descent

The analysis of identity-by-descent (IBD) segments is an important tool in population genetics research and has led to important discoveries about human genetic history and the relationships among populations. The concept of identity by descent (IBD) originated with the idea that alleles inherited from a common ancestor are considered as identical.[1] This notion was later formalized into mathematical frameworks that quantify the probabilities of alleles being identical-by-descent in pedigrees.[2-4] Subsequent research expanded the focus from individual genes to multiple loci, leading to the recognition of haplotypes as the fundamental units for identifying IBD segments.[5,6]

Haplotypes are contiguous sequences of co-inherited alleles on the same chromosome, while IBD segments refer to those haplotypes that two or more individuals have inherited from a common ancestor. The definition of IBD depends on a relative ancestral reference point, either a specified founder in pedigree studies, or more generally, an ancestral population at a specific time point in the past, with very remote ancestry implicitly ignored.[7] Because recombination events during meiosis disrupt the continuous inheritance of genomic segments, IBD segments are delimited by the crossover breakpoints that occur between a common ancestor and its descendants. A simple example of IBD segments shared between a pair of half siblings relative to their common parent is presented in Fig 1.1. According to Haldane's model of recombination[8], lengths of IBD segments inherited from a common ancestor g generations ago follow an

exponential distribution with a mean of $50/g$ centiMorgans (cM). The number and length of shared IBD segments serve as quantitative measures of genetic relatedness between individuals.

Figure 1.1. Illustration of shared IBD segments between two half siblings.
The left panel shows a pedigree where the siblings share the same mother but have different fathers. Each parent's diploid chromosome copies are represented as two segments of distinct colors. For each offspring, the segments representing the diploid chromosome copies consist regions of different colors, indicating the genetic material inherited from the corresponding chromosome copy of each parent. The right panel zooms into the maternal chromosome copies of the half siblings, highlighting regions where they have inherited identical segments from their mother.



The advent of SNP array genotype data and the development of statistical haplotype inference methods enabled modern IBD detection methods through comparisons of haplotypes across individuals from population-level genotype data.[9-16] Methods for IBD detection builds on the idea that a haplotype shared by a group of individuals that is unlikely to occur by chance given its population frequency is strong evidence of IBD.[7,17] There are two primary classes of IBD detection approaches: probabilistic-based methods and length-threshold methods. Probabilistic-based methods use statistical models to estimate the likelihood that a given haplotype segment is IBD, whereas length-threshold approaches identify segments containing mostly identical alleles that exceed a predetermined genetic length, under the assumption that longer segments are less likely to occur by chance.

The progression of IBD detection methods reflects both the increasing density of genetic markers and advances in computational tools. Early probabilistic methods employed hidden Markov models to estimate IBD probabilities across markers in a pointwise manner, often ignoring linkage disequilibrium (LD) given sparse genotype data or through marker pruning.[18,19] Later methods improved the detection of IBD segments in high-density SNP datasets by incorporating LD either by computing IBD probabilities conditional on nearby markers or by jointly modeling IBD status with haplotype structures.[20-23] However, probabilistic-based approaches for IBD detection are inherently slow as the computation burden scales quadratically with sample size, leading to a shift towards length-based methods.

GERMLINE is widely recognized as one of the first length-based methods, employing a hashing algorithm to create a hash table of short haplotype segments, which facilitates fast matching across individuals and identification of long contiguous IBD tracts.[24] This approach has been extended by methods that combine both probabilistic and length-based strategies by first searching for shared haplotype tracts and then calculating the probability of IBD on the candidate segments.[25-29] Later, length-based IBD detection approaches evolved further with the adoption of the positional Burrows-Wheeler transform (PBWT) algorithm[30], which leverages the natural ordering of haplotypes to efficiently identify long shared segments, outperforming traditional fixed-window hashing methods in large-scale datasets.[31-33] Hap-IBD[32] is a state-of-the-art approach for length-based IBD detection that is both accurate and memory-efficient. It first identifies short “seed” haplotype segments and then extended seed segments into full IBD tracts by dynamically refining the boundaries based on local haplotype patterns while allowing for minor regions of discordant alleles.

Several factors can affect IBD detection and lead to false positives or false negatives in reported IBD segments, especially for shorter segments that are more likely to occur by chance. Genotyping errors, phasing errors, mutation, and gene conversion events during meiosis may all complicate IBD inference by introducing sporadic differences that fragment true IBD segments into shorter pieces and confound segment boundaries.[16,17,22,23,25-27,32,33] Probabilistic-based approaches rely on precise estimates of population allele and haplotype frequencies, but these estimates can be skewed by factors such as population stratification or technical artifacts in genotyping and phasing.[7,17,23,25] For length-based approaches, selecting an appropriate segment length threshold for IBD detection is crucial, as a threshold set too low may yield excessive false positives, especially in lower-density data, while a threshold set too high may overlook short IBD tracts.[31-33] Due to false positive and false negative results, pairwise IBD segments may not be jointly consistent when comparing among multiple individuals, which may complicate downstream analyses.[7,17,26]

Aggregating shared ancestry information across multiple individuals has emerged as a powerful strategy to reduce false positives and recover missed signals of pairwise IBD detection, especially for small IBD regions.[17,34] One approach to extend IBD detection beyond pairwise level is to directly model the states of potential IBD tracts across multiple individuals using probabilistic techniques, as demonstrated by Moltke et al. (2011) with a Markov Chain Monte Carlo framework.[34] However, like other probabilistic methods, this approach is computationally intensive and is not feasible for analyses of whole-genome data from more than a few hundred individuals. An alternative strategy is to first identify pairwise IBD segments and then infer

multi-individual sharing by aggregating overlapping segments through either graph-based clustering or hidden Markov models.[35-37] Nevertheless, as the number of pairwise IBD segments increases quadratically with sample size, such approaches would require an impractically high amount of memory and computation time on biobank-scale dataset. In comparison, Browning and Browning (2024) applies PBWT sorting to directly merge haplotypes that share the same allele sequence over a minimum length threshold in a targeted region among all individuals in the sample, avoiding exhaustive pairwise comparisons.[38] This approach reliably identifies clusters of shared ancestry and scales linearly with sample size, making it well-suited for biobank-scale analyses.

With the development of state-of-art software to efficiently identify IBD segments from large-scale genotyping or sequencing data, application of IBD segments has been explored in a range of topics, including estimation of population genetics parameters such as kinship[25,39,40], effective population size[41-43], mutation rates[44-46] and recombination rates[47], detection of recent selection[48-51], and association mapping of disease-related genes (IBD mapping)[19,21,28,34,35,37,52,53]. While previous research has made significant progress in utilizing IBD segments to study population genetics, ongoing methodological advancements and increasing availability of genomic data provide opportunities for further exploration of the power of IBD segments in population genetics research. In this dissertation, we further explore the potential of IBD information in effective population size estimation (Chapter 2) and in IBD mapping (Chapter 3-5).

1.2 IBD-based Effective Population Size Estimation

Effective population size (N_e) is a fundamental parameter in population genetics that represents the size of an idealized, random mating population experiencing the same level of genetic variation as the observed population.[54] Understanding how effective population size has varied over different historical periods can uncover critical demographic events that have shaped a population's evolution history and provide insights into the genetic variation observed today.[41,42,55-60] Estimating effective population size is fundamentally an analysis of the coalescent history of a population, since N_e reflects the rate at which genetic lineages converge to a common ancestor. This coalescent process leaves detectable signatures in genetic data. In small populations, genetic drift accelerates coalescence, resulting in a reduced level of genetic diversity, longer and more continuous IBD segments due to fewer recombination events, and extended runs of homozygosity. In large populations, longer coalescence times and more frequent recombination lead to shorter IBD tracts, more diverse allele frequencies, and lower overall levels of homozygosity.

Various methods have been developed to infer effective population size by leveraging genetic signatures that reflect a population's coalescent history. Patterns of linkage disequilibrium[55,56,61] and site frequency spectrums[58,62,63] can be used to reconstruct historical demographic events and fluctuations in effective population size that have shaped allele frequencies and correlations over time. Alternatively, effective population size can be inferred by reconstructing the history since the most recent common ancestor through the distribution and lengths of IBD segments shared among individuals.[41,42,64] IBD-based estimates of N_e tend to reflect recent demographic history

rather than ancient population dynamics, since short IBD segments linked to more ancient coalescence events are more difficult to detect reliably.

In Chapter 2, we propose a method to estimate the X chromosome effective population size from X chromosome IBD segments building on the software IBDNe[42], which implements a robust non-parametric model to estimate recent effective population size from the length distribution of IBD segments. We show how to use the estimated autosome effective population size and X chromosome effective population size to estimate female and male effective population sizes. We applied our method to analyze the demographic history of several human populations represented by samples of White British and Indian individuals from the UK Biobank cohort and samples of African American individuals in the HyperGEN study from the TOPMed project.

1.3 IBD mapping

Before the advent of modern sequencing techniques, disease mapping leveraging identity-by-descent information was primarily performed through linkage analysis in pedigrees, through tracking the co-segregation of genetic markers and disease phenotypes in family studies. The traditional approach for pedigree-based linkage analysis in this setting involves comparing the observed allele sharing between individuals to the expected IBD probabilities derived from pedigree structures, thereby identifying disease loci where affected individuals share alleles more frequently than anticipated by chance. An alternative strategy for linkage analysis uses variance component models to partition observed phenotypic variance into components attributable to a specific quantitative trait locus (QTL) and to other genetic and environmental factors.[65-71] In this framework, a QTL's effect is assessed by testing whether including a locus-specific covariance

component significantly improves the model fit compared to a model without it. This covariance component is derived from a genetic relatedness matrix calculated from the expected IBD probabilities at the candidate QTL, which quantify the extent of allele sharing between individuals based on their known lineages.

Pedigree-based linkage analysis has limited resolution due to the few meioses and recombination events available among close relatives, and its applicability is constrained by the need for extensive family-based data. With the advent of high-density genotyping and whole-genome sequencing, IBD mapping has emerged as a population-based linkage mapping approach suited for analyzing large cohorts and rich genomic datasets using IBD sharing among individuals in an outbred population. IBD mapping works on the principle that individuals exhibiting a specific trait or disease are more likely to share alleles inherited from a common ancestor in regions that influence the phenotype. Thus, when a significant correlation is observed between levels of IBD sharing and phenotypic similarity, it suggests that the shared genomic regions may harbor causal variants, making these regions candidates for further investigation in disease or trait mapping.

In previous IBD mapping studies utilizing pairwise IBD segments, a common strategy is comparing pairwise IBD sharing rates among individuals in case-control cohorts.[24,52] In these approaches, the frequency of IBD segments shared between two affected individuals is contrasted with that observed in case-control or control-control pairs across the genome, under the premise that regions with elevated sharing among cases are more likely to harbor disease-associated variants. However, these methods are generally designed for binary disease

phenotypes, and their reliance on simple pairwise comparisons may limit their applicability to complex traits in broader cohorts.

Another approach for IBD mapping involves identifying clusters of haplotypes that are IBD at a specific locus and then testing the association between the frequencies of IBD clusters and the phenotype of interest.[34,35,37] By analyzing IBD haplotypes shared among groups rather than relying solely on pairwise comparisons, this strategy can improve power through more accurate IBD calls and by capturing additional group-specific effects that are ignored in pairwise IBD analysis. However, multi-individual IBD mapping has not yet been widely adopted for studying genetic variants underlying complex traits due to challenges in efficiently detecting IBD clusters in biobank-scale datasets and in developing versatile models to incorporate IBD cluster information for association testing.

IBD mapping can also be conducted using the traditional variance component model in pedigree-based linkage analysis, where the genetic relatedness matrix is directly derived from the proportion of inferred IBD segments across the genome.[53,72,73] By treating the genetic effect at a locus as a random effect, this approach captures the full covariance structure of local genetic sharing, making it flexible in accommodating complex relatedness patterns. In Chapter 3 of this dissertation, we present an IBD mapping approach that build on the traditional variance component model in linkage analysis to test the association between local genetic similarity and complex trait variation. Our method evaluates whether levels of genetic similarities at specific genomic locations, captured by local relatedness matrices derived from multi-individual IBD sharing, are associated with phenotypic variation in complex traits. A similar approach has been

recently proposed by Chen et al. (2023), but their method relies on pairwise IBD and employs a different statistical testing framework. In Chapter 4, we propose an approach to adjust for multiple testing in genome-wide IBD mapping scans based on the correlation structure between test statistics across the genome. Through simulation studies, we demonstrate that our test has a well-controlled genome-wide type-I error rate and superior power to detect rare and untyped variants compared to standard single-variant tests. In Chapter 5, we applied our IBD mapping test to systolic blood pressure data from White British individuals in the UK Biobank.

Chapter 2

IBD-BASED ESTIMATION OF X CHROMOSOME EFFECTIVE POPULATION SIZE WITH APPLICATION TO SEX-SPECIFIC DEMOGRAPHIC HISTORY

This chapter contains material published in:
Cai R, Browning BL, Browning SR. Identity-by-descent-based estimation of the X chromosome effective population size with application to sex-specific demographic history. G3: Genes, Genomes, Genetics. 2023 Oct;13(10):jkad165.

2.1 Introduction

The effective size of a population (N_e) is defined as the number of breeding individuals in an idealized randomly mating population that has the same expected value of a parameter of interest as the actual population under consideration.[54] The effective population size is a fundamental parameter in population genetics because it determines the strength of genetic drift and the efficacy of evolutionary forces such as mutation, selection, and migration.[74] Previous studies have demonstrated that estimates of recent effective population size can reveal aspects of a population's demographic history, such as past population growth or bottleneck events.[42,43]

Identity-by-descent (IBD) segments can be used to estimate effective population size in the recent past. IBD segments are haplotypes which two or more individuals have inherited from a common ancestor. IBD segments end at positions where crossovers have occurred in the meioses between the common ancestor and the descendant individuals. IBD segments for which the common ancestor is in the distant past tend to be shorter than IBD segments from a recent

common ancestor because there are more meioses since the common ancestor on which crossovers can occur. The autosomes and the X chromosome are both subject to recombination, making them both amenable to IBD segment analysis.[75,76] Previous studies have developed methods for estimating recent effective population size from autosomal IBD segments.[41,77,78] However, no such effort has been made with X chromosome IBD segments.

The autosomes are equally influenced by female and male demographic processes. In contrast, the X chromosome is influenced more strongly by female demographic processes than by male demographic processes. This is because females have two copies of the X chromosome, while males have only one. Thus, comparison of statistics from the X chromosome and from autosome data can be used to estimate sex-specific parameters such as female and male effective population sizes.[76,79-87]

The standard Wright-Fisher model used to define the effective population size assumes equal numbers of breeding females and males. To define female and male effective population sizes, we consider an idealized Wright-Fisher population modified to allow for different numbers of females and males. The female and male effective sizes (N_e^f and N_e^m) of a non-idealized population are the values that would give the same rates of IBD on autosomes and sex chromosomes as the Wright-Fisher population with the corresponding numbers of females and males.[88]

In this work, we develop an IBD-based method to estimate the trajectory of X chromosome effective population size in the recent history. We show that estimated X chromosome N_e can be

combined with estimated autosome N_e to estimate the trajectory of female and male effective population sizes over time. This application of the X chromosome N_e provides a useful complement to previous methods that give only a single time-averaged estimate for the sex-specific effective sizes in human populations.[84,86,89] Through simulation studies, we validate the theoretical relationship between the autosome and X chromosome N_e, and we show that our method can accurately estimate the autosome and X chromosome N_e in simulated populations. We examine the application of X chromosome N_e to estimate sex-specific N_e and show that long-term pronounced differences in female and male effective population sizes can be detected, but that short-term observed differences in the estimated effective population sizes may not represent true differences.

2.2 Estimation of X chromosome effective population size

2.2.1 Probability modelling for the X chromosome

All meioses from mothers transmit an X chromosome, while half of meioses from fathers transmit an X chromosome. Over many generations, approximately two-thirds of meioses from parents to offspring that transmit an X chromosome will be from females, as we show below. However, the proportion of females in a sample of individuals affects the actual proportion of meioses transmitting an X chromosome in recent generations that are from females. For example, if the sample is solely made up of females (with two X chromosomes, one from each parent), half of the meioses in the most recent generation that transmit an X chromosome are from females. If the sample is made up solely of males (with only one X chromosome, inherited from the mother), all the meioses that directly contribute the X chromosomes in the sample are from females. If half the sampled individuals are female and half are male, then 2/3 of the

meioses in the previous generation that contribute the X chromosomes in the sample will be from females. We show in the next paragraph that when half of the samples are female this 2/3 ratio will apply in expectation in all prior generations, regardless of sex-specific demographic forces.

Consider the lineage of a randomly selected haplotype at a point in the genome. Let p_g be the probability that the ancestral haplotype at generation g before present is carried by a female, where $g = 0$ corresponds to the generation of the sampled individuals, $g = 1$ to the generation of their parents, and so on. A given haplotype carried by a female has a 50% probability that its parent haplotype is carried by a female, while a given haplotype carried by a male always has its parent haplotype carried by a female. Thus, for $g \geq 0$,

$$p_{g+1} = 0.5p_g + (1 - p_g) = 1 - 0.5p_g.$$

If $p_g = 2/3$, then $p_{g+1} = 2/3$. If the sampled haplotype is randomly chosen from a set of individuals with equal numbers of females and males, then the sampled haplotype has probability $p_0 = 2/3$ of being carried by a female, and as a result $p_g = 2/3$ for all g. More generally, it can be shown by mathematical induction that

$$p_g = \frac{2}{3}\left(1 - \left(-\frac{1}{2}\right)^g\right) + \left(-\frac{1}{2}\right)^g p_0$$

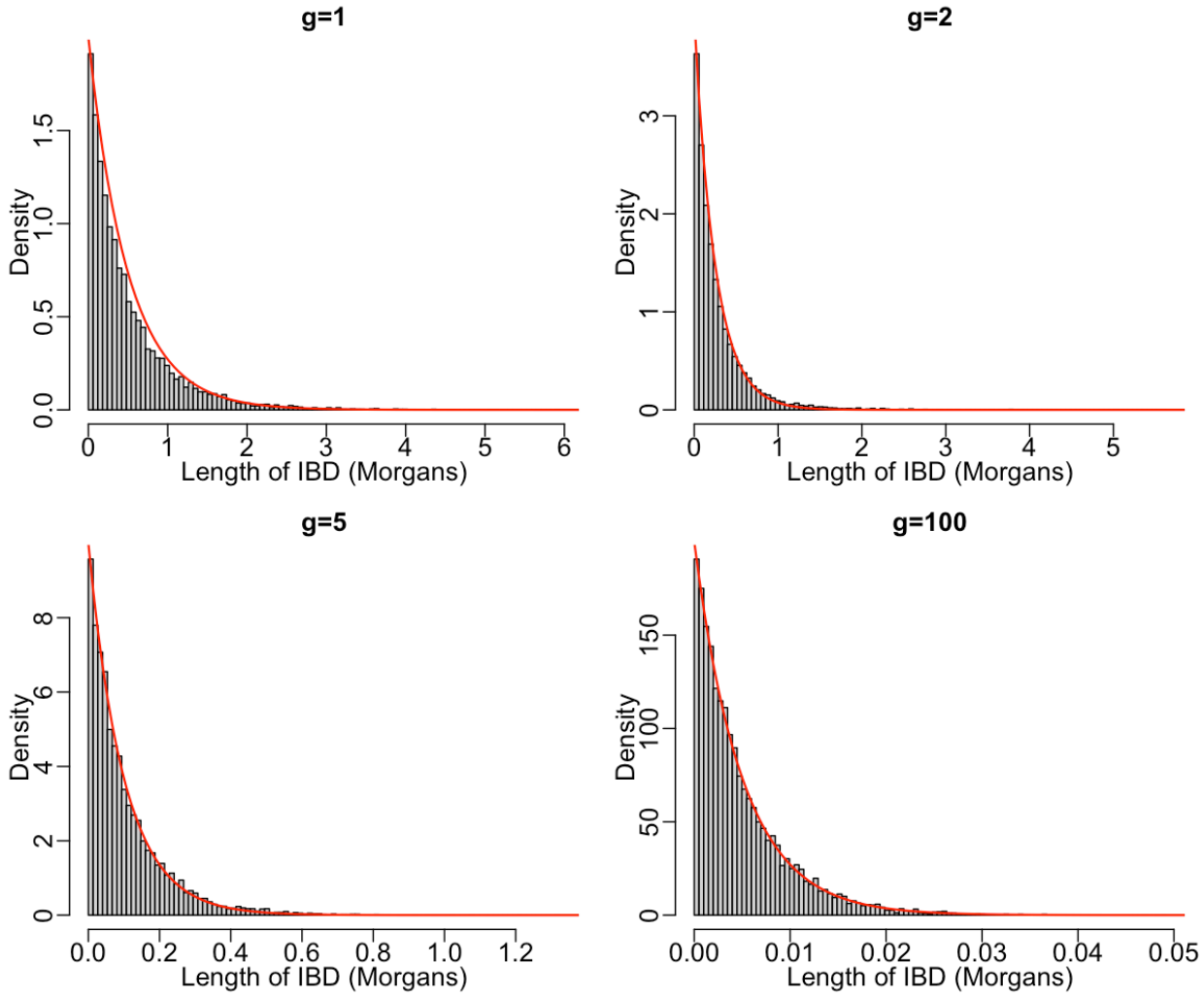
which converges to $2/3$ for large g. This equation is similar to equations for X chromosome admixture proportions and for X chromosome allele frequencies in a particular breeding systems.[82,90] In what follows, we assume that $p_g = 2/3$ for all g.

Consider IBD sharing on the X chromosome resulting from a common ancestor living g generations ago. Let F be the number of female meioses out of the $2g$ meioses in the path of

inheritance. Then F follows a binomial$(2g, 2/3)$ distribution. Assuming Haldane's model[8], the length of the IBD segment is exponentially distributed with rate $3F/2$ per Morgan, as crossovers that end an IBD segment happen only in female meiosis on the X chromosome (which comprise 2/3 of the meioses) and the female recombination rate is 3/2 per Morgan on the X chromosome when using sex-averaged genetic distances. As an approximation, we model the probability distribution of the length of an X chromosome IBD segment resulting from a coalescence event occurring g generations ago as an exponential random variable with rate $2g$ per Morgan, which is the same model that we use for estimating IBD-based effective population size in autosomal data.[77] In Figure 2.1, we compared the simulated distribution of draws from an exponential$(3F/2)$ distribution with F drawn from a binomial$(2g, 2/3)$ distribution to the exponential$(2g)$ distribution for different values of g. The result shows that the exponential$(2g)$ model approximates the distribution of IBD length on the X chromosome very well for $g > 1$.

Figure 2.1. Simulated distribution of the length of an X chromosome IBD segment measured in sex-averaged genetic distance given the number of generations to the shared ancestor.

The distribution of 10,000 simulated exponential$(3F/2)$ observations, where $F\sim$binomial$(2g, 2/3)$, is compared to the Exponential$(2g)$ density, shown as the red curve in each plot, for $g = 1, 2, 5$ and 100 generations. When $F = 0$, the simulated length of IBD segment is set to 100 Morgans. Although not shown in the figure, the observations corresponding to $F = 0$ were accounted for when calculating the simulated density.



2.2.2 IBD-based estimation of X chromosome effective population size

Our method for IBD-based estimation of X chromosome effective population size history is based on IBDNe which was designed to estimate recent effective population size from autosomal IBD segments.[77] The IBDNe method calculates the expected length distribution of IBD segments exceeding a given length threshold (2 cM by default) for a given effective population size history. It finds the effective population size history that equates the observed and expected IBD length distributions using an iterative scheme. IBDNe applies smoothing over intervals of eight generations to avoid overfitting. Although IBDNe was designed for autosomal data, we show that it can also be used with X chromosome data, with some adjustments to the analysis procedure that we describe in the following paragraphs.

The first adjustment ensures proper inference of IBD segments on the X chromosome by encoding male X chromosome genotypes as haploid. This coding conforms to the VCF specification.[91] Male X chromosome genotypes are frequently coded as homozygous diploid genotypes rather than haploid genotypes, which typically results in duplicate reported IBD segments when using IBD detection methods designed for autosomal data. The hap-ibd[32] program can correctly analyze chromosome X data with haploid male genotypes. We exclude the pseudo-autosomal regions from analysis.

The second adjustment is to use a sex-averaged genetic map, as we also do for the autosomes. X chromosomes transmitted from females are subject to recombination, while X chromosomes transmitted from males are not (except for the pseudo-autosomal regions which we exclude from all analyses). Genetic maps, such as the HapMap map[92] and the deCODE map[93], typically report

the female-specific recombination map for the X chromosome. Since an average of 2/3 of meioses transmitting an X chromosome are from females, the sex-averaged X chromosome recombination map can be obtained by multiplying female-specific genetic distances by 2/3. For example, a region with length 3 cM on the female-specific map has length 2 cM on the sex-averaged map. Equivalently, the sex-averaged recombination rates can be obtained by multiplying female-specific recombination rates by 2/3. For example, an X chromosome region with female-specific recombination rate of 3×10^{-8} per base pair per generation has a sex-averaged recombination rate of 2×10^{-8} per base pair per generation.

The third adjustment ensures equal numbers of sampled females and males. If the sample is unbalanced, we remove some randomly selected females or males to obtain equal numbers of females and males. Consequently, p_0, the proportion of sampled X chromosome haplotypes carried by females, is 2/3, and hence p_g, the probability that the ancestral haplotype of a sampled X chromosome haplotype g generations before the present is carried by a female is always 2/3 (see the preceding "Probability Modelling for the X chromosome" section).

The fourth adjustment modifies the IBDNe "npairs" parameter to be equal to the number of analyzed haplotype pairs. By default, IBDNe assumes that each individual contributes two haplotypes to the analysis, and that all cross-individual pairs are analyzed, resulting in $(2n)(2n-2)/2$ haplotype pairs when there are n individuals. On the X chromosome, with n_f females and n_m males, the number of haplotype pairs is

$$2n_f(2n_f - 2)/2 \; + \; n_m(n_m - 1)/2 \; + \; 2n_f n_m.$$

Equation 2.1

We set the IBDNe “npairs” parameter to the value in Equation 2.1 when analyzing X chromosome data, after adjusting it for removal of close relative pairs as described below.

The fifth adjustment is to manually remove detected IBD segments corresponding to close relatives (parent-offspring and siblings). By default, IBDNe identifies the close relatives using the input IBD segments, removes the IBD segments between them, and adjusts the “npairs” parameter to account for the removed sample pairs. However, this strategy does not work for the X chromosome because one cannot reliably detect close relatives using only X chromosome data. We thus turn off IBDNe’s filtering of close relatives by setting “filtersamples=false”. We can identify close relatives based on autosomal data or from a pedigree file if available, and then remove IBD segments for these pairs and update the “npairs” parameter accordingly in the chromosome X analysis. Removing only IBD segments between the related pairs rather than completely removing one individual from each pair of relatives reduces the loss of information from the data.

The sixth adjustment enables calculation of confidence intervals. IBDNe obtains confidence intervals for the estimated effective population sizes by bootstrapping over chromosomes. We thus divide the X chromosome into six pieces of equal cM length and treat these as separate “chromosomes” in the analysis with IBDNe.

2.3 From X chromosome effective population size to sex-specific effective population size

We describe how the estimated X chromosome effective population size can be used in conjunction with the estimated autosome effective population size to estimate female and male effective population sizes. We will write N_g^X and N_g^A for the X chromosome and autosomal effective population sizes at generation g. And we will write N_g^f and N_g^m for the female and male effective population sizes at generation g, which can be derived from the X chromosome and autosomal effective population sizes as described below.

The IBD-based effective population size (for autosomes or X chromosome) is defined in terms of the conditional coalescence probability for a Wright-Fisher population. We first consider autosomes. For a randomly selected pair of haplotypes, let G be the number of generations before present that the haplotypes coalesce. Conditional on the haplotypes not coalescing by generation $g-1$ before present, the ancestral haplotypes are distinct at that generation. For them to coalesce at generation g before present, their two parental haplotypes at generation g must be the same haplotype. If the diploid autosomal effective population size is N_g^A at g generations before present, there are $2N_g^A$ autosomal haplotypes available, and the probability that the two parental autosomal haplotypes are the same is thus $1/(2N_g^A)$. That is, $P_A(G = g|G > g-1) = 1/(2N_g^A)$. Thus, if we know the value of $P_A(G = g|G > g-1)$, then we can obtain the effective population size g generations before present:

$$N_g^A = 1/\big(2P_A(G = g|G > g-1)\big).$$

On the X chromosome, the conditional coalescence probability can be obtained by considering that 2/3 of meioses are from female parents, while 1/3 are from male parents. For coalescence to occur, both haplotypes' parent haplotypes must be the same. This means both haplotypes must be inherited from parents that have the same sex, and the second haplotype must have the same parental haplotype as that of the first (within that sex). Thus,

$$P_X(G = g|G > g-1) = \frac{(2/3)^2}{2N_g^f} + \frac{(1/3)^2}{N_g^m}.$$

And hence,

$$N_g^X = 1/\left(2P_X(G = g|G > g-1)\right)$$

$$= \frac{9N_g^f N_g^m}{2N_g^f + 4N_g^m}$$

Equation 2.2

For comparison, on the autosomes, by the same reasoning but with half of the meioses from each sex and with diploid males,

$$P_A(G = g|G > g-1) = \frac{(1/2)^2}{2N_g^f} + \frac{(1/2)^2}{2N_g^m}.$$

And hence,

$$N_g^A = \frac{1}{2}\Big/\left(\frac{(1/2)^2}{2N_g^f} + \frac{(1/2)^2}{2N_g^m}\right) = \frac{4N_g^f N_g^m}{N_g^f + N_g^m}.$$

Equation 2.3

From Equations 2.2 and 2.3, the ratio of X to autosomal effective population size, which we denote as α, is

$$\alpha = \frac{N_g^X}{N_g^A}$$

$$= \frac{9\left(N_g^f + N_g^m\right)}{8\left(N_g^f + 2N_g^m\right)}$$

$$= \frac{9\left(1 + \left(N_g^m / N_g^f\right)\right)}{8\left(1 + 2\left(N_g^m / N_g^f\right)\right)}.$$

Equation 2.4

Thus $N_g^m / N_g^f \to 0$ as $\alpha \to \frac{9}{8}$, and $N_g^m / N_g^f \to \infty$ as $\alpha \to \frac{9}{16}$. Since α is a decreasing function of N_g^m / N_g^f, the ratio of X to autosomal effective population size satisfies $\frac{9}{16} < \alpha < \frac{9}{8}$.

With algebra, it can be shown that Equations 2.2 and 2.3 imply that

$$N_g^f = \frac{2N_g^X N_g^A}{9N_g^A - 8N_g^X}$$

Equation 2.5

and

$$N_g^m = \frac{2N_g^X N_g^A}{16N_g^X - 9N_g^A}.$$

Equation 2.6

Thus, given estimates of N_g^X and N_g^A, one can use Equations 2.5 and 2.6 to obtain estimates for N_g^f and N_g^m. These are standard equations for estimating sex-specific effective population sizes based on N_g^X and N_g^A,[94,95] although usually presented in the context of constant effective population sizes across time.

According to Equation 2.5, the allowable range of X chromosome N_e is between $\frac{9}{16}$ and $\frac{9}{8}$ of the autosome N_e at each generation. When the X chromosome N_e is overestimated or the autosome

N_e is underestimated, the estimated female effective population size can be negative. On the other hand, underestimation of the X chromosome N_e or overestimation of the autosome N_e can result in a negative estimate of the male effective population size (Equation 2.6).

2.4 Analysis pipeline

We start with phased sequence or SNP array data (using true phase for the simulated data, and inferred phase for real data), with males coded as haploid on the X chromosome. We use hap-ibd[32] to infer IBD segments. For hap-ibd analysis on sequence data, we set the minimum seed length to 0.5 cM and the minimum extension length to 0.2 cM. The relatively small minimum seed length and minimum extension length increase power to detect short IBD segments. We exclude rare variants by setting the minimum minor allele count filter to 100 because these lower frequency variants are less informative, have less accurate phasing, and may be recent mutations. This corresponds to a minor allele frequency (MAF) of 0.1% in the simulated sequence data and a MAF of 4% in the TOPMed sequence data we analyzed in this study. For SNP-array data, due to the lower marker density, we set the minimum seed length to 1 cM, the minimum extension length to 0.1 cM, and the minimum number of markers in a seed IBS segment to 50. Other parameters are left at their default values. The IBD analysis parameters are the same for the autosomes and X chromosome.

We then run IBDNe on the detected IBD segments, with one analysis for autosomes and a separate analysis for the X chromosome. The genetic map file for the analysis is assumed to be a sex-averaged map. For the X chromosome, this means multiplying cM positions in the female-specific map by 2/3.

When applying IBDNe on the simulated data (autosomes or X), we set “filtersamples=false” and “gmin=1” because the chromosomes are simulated independently so that a pair of individuals can share ancestry one generation back (i.e. be siblings) on one chromosome without such sharing occurring on other chromosomes. These settings tell IBDNe not to look for and remove close relatives, and to model IBD from shared ancestry starting from one generation before present.

Real data often has an excess of close relatives due to the sampling scheme. Thus, in the analysis of real autosomal data we allow IBDNe to detect and remove close relatives, which is the default behavior. The X chromosome on its own is not sufficient to detect close relatives, so we use either available pedigree information or the close relative pairs identified by IBDNe in the autosomal analysis to manually remove IBD segments from close relatives in the X chromosome data. We then update the “npairs” parameter accordingly (see below), and set “filtersamples=false” in the X chromosome IBDNe analysis.

In the IBDNe analysis of the X chromosome data, we set the “npairs” parameter equal to the number of haplotype pairs for which IBD segments could be present (i.e., all pairs except for those for which we have explicitly removed IBD segments). We first calculate the number of haplotype pairs based on the numbers of females and males in the sample, using Equation 2.1. We then adjust the number of haplotype pairs to account for the number of close relative pairs that were removed. Removing a male-male pair decreases the count by 1. Removing a male-female pair decreases the count by 2. Removing a female-female pair decreases the count by 4.

In order for IBDNe to obtain bootstrap confidence intervals for the X chromosome effective population size estimates, we divide the X chromosome into six pieces of equal cM length. We recode the chromosome field of the IBD segment file using integer values between 1 and 6 according to the location of the IBD segments. IBD segments that cross more than one of these "chromosomes" are split into subsegments at the boundaries of the corresponding "chromosomes". We remove the centromere region from IBD segments, so that each IBD segment that crosses the centromere region is replaced with two IBD segments: a segment before and a segment after the centromere.

After obtaining the X chromosome and autosomal effective population sizes, we estimate sex-specific effective population sizes using Equations 2.5 and 2.6. We obtain bootstrap values for these estimates by taking pairs of bootstrap values from the X and autosomal analyses. For example, for the n-th bootstrap value of the female effective population size at generation g, we take the n-th bootstrap value for the X chromosome effective population size at generation g and the n-th bootstrap value for the autosomal effective population size at generation g, and apply Equation 2.5. After obtaining all the bootstrap values, we use the 2.5^{th} and 97.5^{th} percentile to obtain an approximate 95% confidence interval for the female (for example) effective population size at generation g.

2.5 Simulation study

2.5.1 Data simulation pipeline

We conducted a simulation study to evaluate the performance of our method. We used SLiM, a forward simulator,[96] to simulate the demographic history for the most recent 5000 generations, and we used msprime, a coalescent simulator, to complete the simulation back to full coalescence of the sample.[97,98] For all scenarios, we simulated data for 30 autosomes of length 100 Mb and an X chromosome of length 180 Mb.

For the simulation in SLiM, we used a mutation rate of 10^{-8} per base pair per generation, and a recombination rate of 10^{-8} per base pair per meiosis on the autosomes and for female meioses on the X chromosome. We set the gene conversion initiation rate to 2×10^{-8} per base pair per meiosis on the autosomes and per female meiosis on the X chromosome, with mean gene conversion tract length of 300 bp. We simulated populations with equal sex ratio, and we also simulated populations with 20% females, 40% females, 60% females, and 80% females. We used the same total effective population size ($N_g^f + N_g^m$) in each simulation. We sampled 50,000 individuals comprising 25,000 females and 25,000 males for each analysis.

We simulated data under two different demographic models. The first model is a four-stage exponentially growing population with increased growth rate over time, which we call the "UK-like" scenario because it approximates the demographic history of the UK population.[23] The forward simulation of this model in SLiM starts from 5000 generations ago with an initial size of 3000. At 300 generations ago, this population starts to grow at an exponential rate of 1.4% per generation. At 60 generations ago, the rate of the exponential growth increases to 6%. In the

most recent 10 generations, the growth rate further increases to 25%, and the population size reaches around 21 million at the time of sampling.

The second demographic model was modified from the "UK-like" model by adding a bottleneck 10 generations before present, which approximates the immigration of Europeans to America. This bottleneck reduces the population size to 100,000. The population then continues to grow at an exponential rate of 25% and reaches a final size around 1.2 million. We call this simulated population the "US-like" scenario.

For both the UK-like model and the US-like model, the part of the simulation conducted in msprime corresponds to the demographic history earlier than 5,000 generations ago that has a constant population size of 3,000. When completing the simulations with msprime, we did not apply gene conversion or sex-specific population size. In the msprime simulations, mutation occurred at a rate of 10^{-8} per base pair per generation, and recombination at a rate of 10^{-8} per base pair per meiosis. There is no differentiation between female and male meioses in the generations prior to the starting generation of the forward simulation because msprime does not have sex-specific functionality. This lack of sex-specific treatment in the period more than 5000 generations ago will affect the level of variation in the data but will not affect the distribution of IBD segments of length > 2 cM (the segments analyzed by IBDNe) because such segments typically have ancestry within the past three hundred generations.

2.5.2 Results from simulation study

We first compared the observed distribution of X chromosome IBD segments between males and females. We grouped observed X chromosome IBD segments from female-female haplotype pairs, female-male haplotype pairs, and male-male haplotype pairs by lengths into consecutive bins spaced at 1 cM intervals. We calculated the rate of IBD segments in each length bin for each sex combination as the number of IBD segments from pairs of individuals with the corresponding sex in that bin divided by the total number of haplotype pairs of the corresponding sex combination. We observed that the distribution of X chromosome IBD segments in female-female haplotype pairs, that in female-male haplotype pairs, and that in male-male haplotype pairs are consistent in all simulations (Figure 2.2 and Figure 2.3).

Figure 2.2. The distribution of IBD segments between sexes in UK-like simulation with various sex ratio.
The red line, green line and blue line show the rate of IBD segments over a series of consecutive length bins in female-female haplotype pairs, male-female haplotype pairs, and male-male haplotype pairs, respectively. IBD lengths are measured in sex-averaged units.

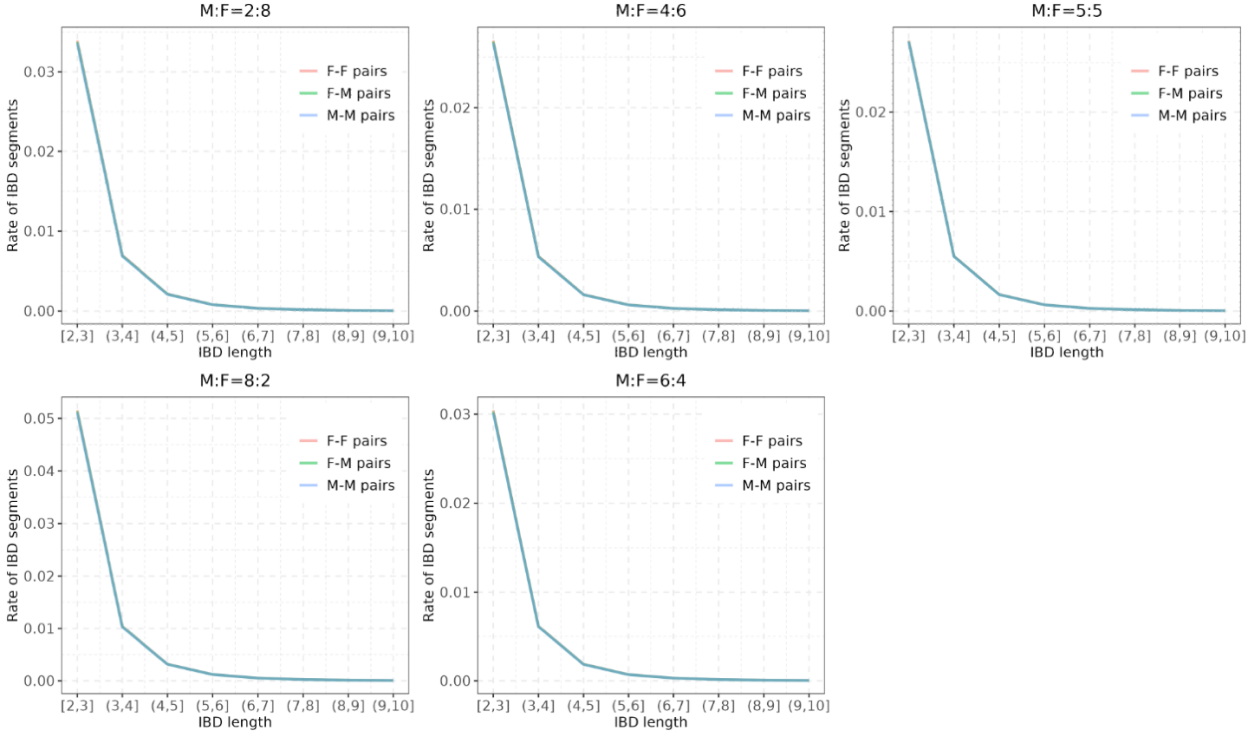
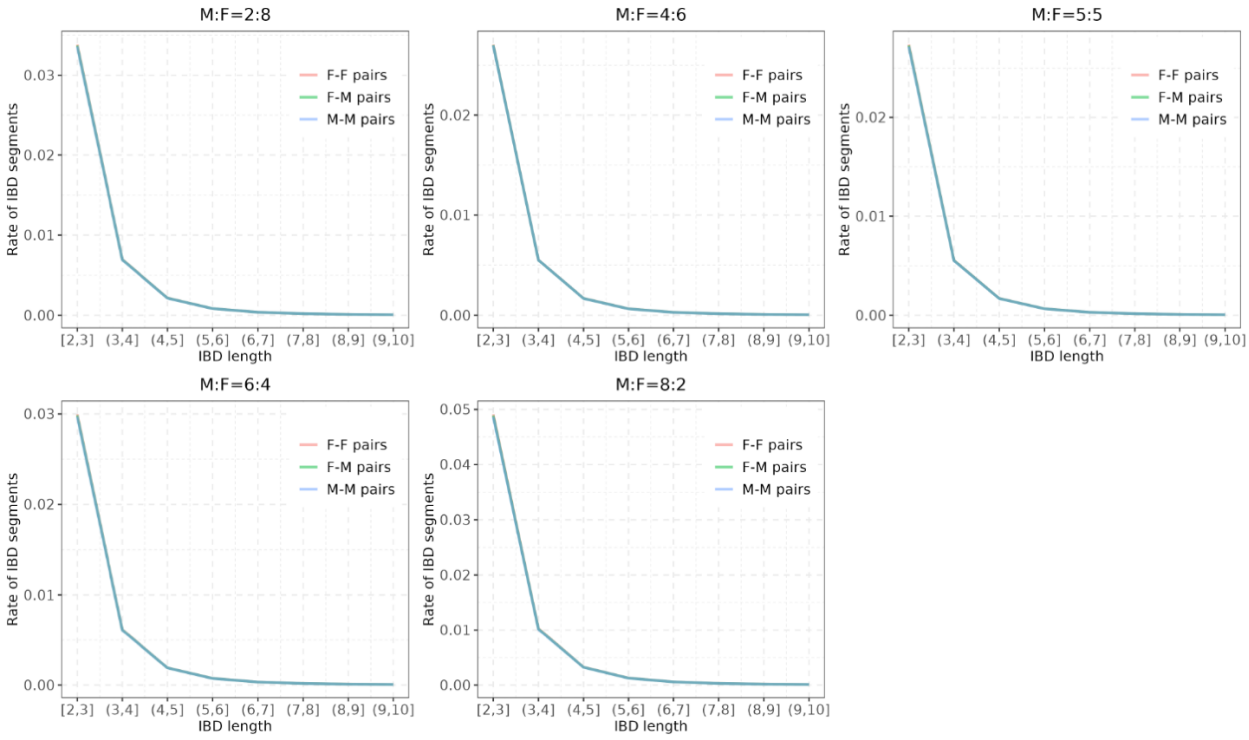


Figure 2.3. The distribution of IBD segments between sexes in US-like simulation with various sex ratio.
The red line, green line and blue line show the rate of IBD segments over a series of consecutive length bins in female-female haplotype pairs, male-female haplotype pairs, and male-male haplotype pairs, respectively. IBD lengths are measured in sex-averaged units.



We compared N_e estimated from the simulated autosome and X chromosome data to the true N_e for the UK-like and US-like demographic models. The true autosome or X chromosome N_e can be obtained from the true simulated effective population sizes given the simulated sex ratios using formulas in Table 2.1, derived from Equations 2.2 and 2.3. The estimated N_e generally matches the true N_e closely (Figure 2.4-2.5), and the estimated X chromosome N_e obtained by splitting the X chromosome into six pieces to enable bootstrapping are consistent with the N_e estimated on the undivided X chromosome (Figure S1-S2). However, some discrepancies exist between the estimated and actual N_e because IBDNe cannot localize sharp changes in the population size to one exact generation and it tends to over-smooth corners of the population-size

trajectory.[77] For the US-like simulations with a sharp bottleneck event that happened at a single generation, the estimated duration of bottleneck is spread over multiple generations (Figure 2.4).

Table 2.1. Theoretical autosome or X chromosome effective population size in terms of the true simulated effective population size at various sex ratios in the population.
The true simulated size is represented by N. Here we used N instead of N_g for simplicity as the same formula applies for every generation g, under the assumption that $p_g = 2/3$ for all g (see Section 2.2.1 and Section 2.3). N^A represents the theoretical autosome effective population size and N^X represents the theoretical X chromosome effective population size.

Sex ratio	Autosome coalescence rate	X chromosome coalescence rate	N^A	N^X	$\frac{N^X}{N^A}$
$\mathrm{F:M} = 2:8$	$\frac{1}{2 \times 0.64 \times N}$	$\frac{1}{2 \times 0.40 \times N}$	$0.64 \times N$	$0.40 \times N$	0.6250
$\mathrm{F:M} = 4:6$	$\frac{1}{2 \times 0.96 \times N}$	$\frac{1}{2 \times 0.675 \times N}$	$0.96 \times N$	$0.675 \times N$	0.7031
$\mathrm{F:M} = 5:5$	$\frac{1}{2N}$	$\frac{1}{2 \times 0.75 \times N}$	N	$0.75 \times N$	0.7500
$\mathrm{F:M} = 6:4$	$\frac{1}{2 \times 0.96 \times N}$	$\frac{1}{2 \times 0.7714 \times N}$	$0.96 \times N$	$0.7714 \times N$	0.8035
$\mathrm{F:M} = 8:2$	$\frac{1}{2 \times 0.64 \times N}$	$\frac{1}{2 \times 0.60 \times N}$	$0.64 \times N$	$0.60 \times N$	0.9375

We next used the estimated X chromosome N_e and autosome N_e to estimate the sex-specific effective population sizes. We find that the formulas for the sex-specific N_e as functions of the autosome and X chromosome N_e (Equation 2.5-2.6) are sensitive to errors in N_e estimation. In UK-like scenarios, with more accurate estimates, the estimated sex-specific N_e are similar to the actual values, except at around 20 generations ago, where there was a large change in the population growth rate (Figure 2.4). For the US-like scenarios, around the time of the bottleneck event where there is greater inaccuracy in the estimated N_e, the estimated sex-specific N_e differs significantly from the actual sex-specific N_e (Figure 2.5).

Figure 2.4. Estimates of autosome, X chromosome, and sex-specific N_e in UK-like simulation studies.

Autosomal N_e is shown in the left column, X chromosome N_e in the middle column, and sex-specific N_e in the right column. From top to bottom, the first row displays results from a UK-like simulation with equal sex ratio, and the following rows show results with 80%, 60%, 40%, and 20% females. For the N_e plots the Y-axes are on a log scale. Y-axes show N_e plotted on a log scale. In cases where the estimated sex-specific N_e is negative (see Methods), it is not shown.

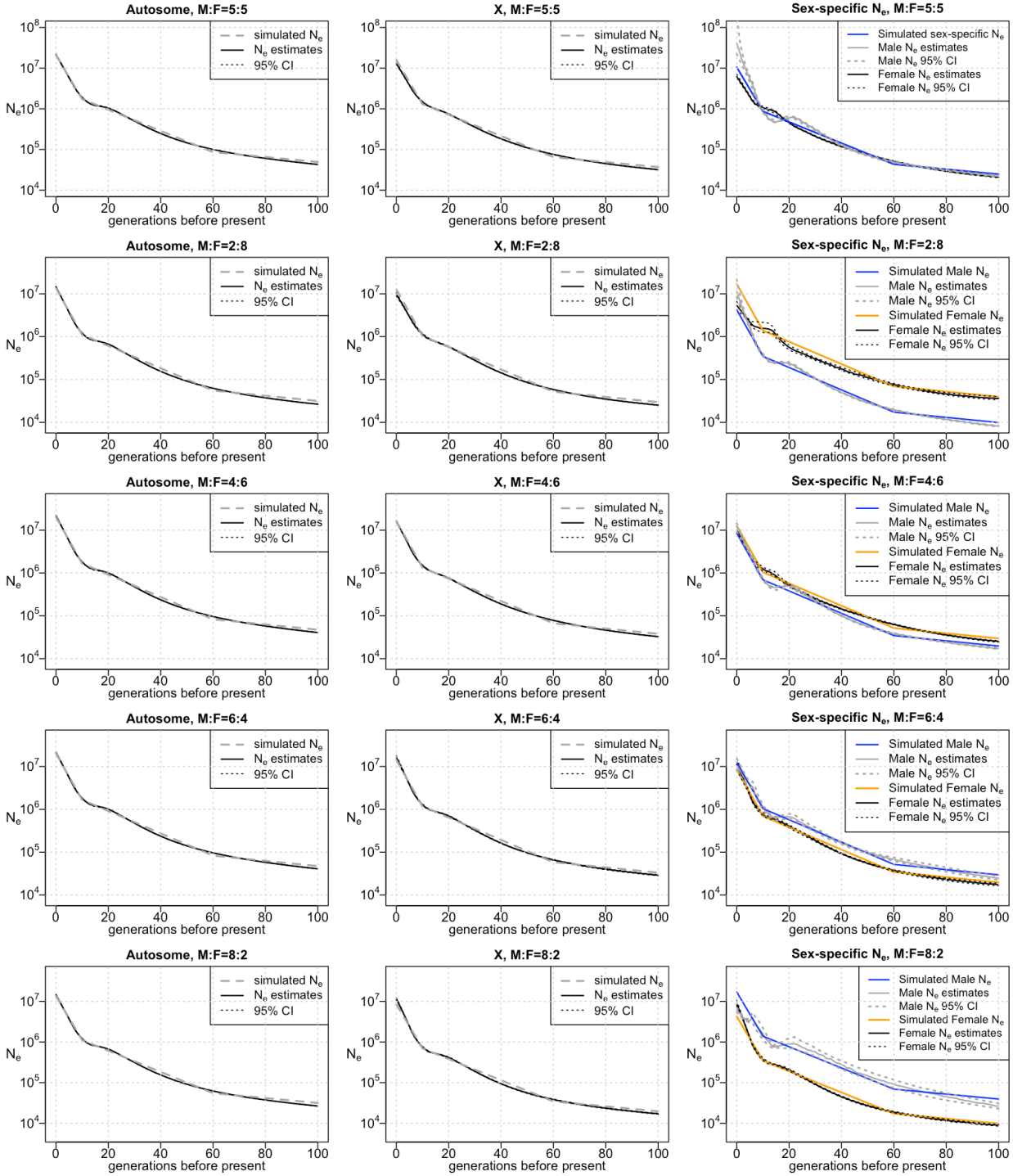
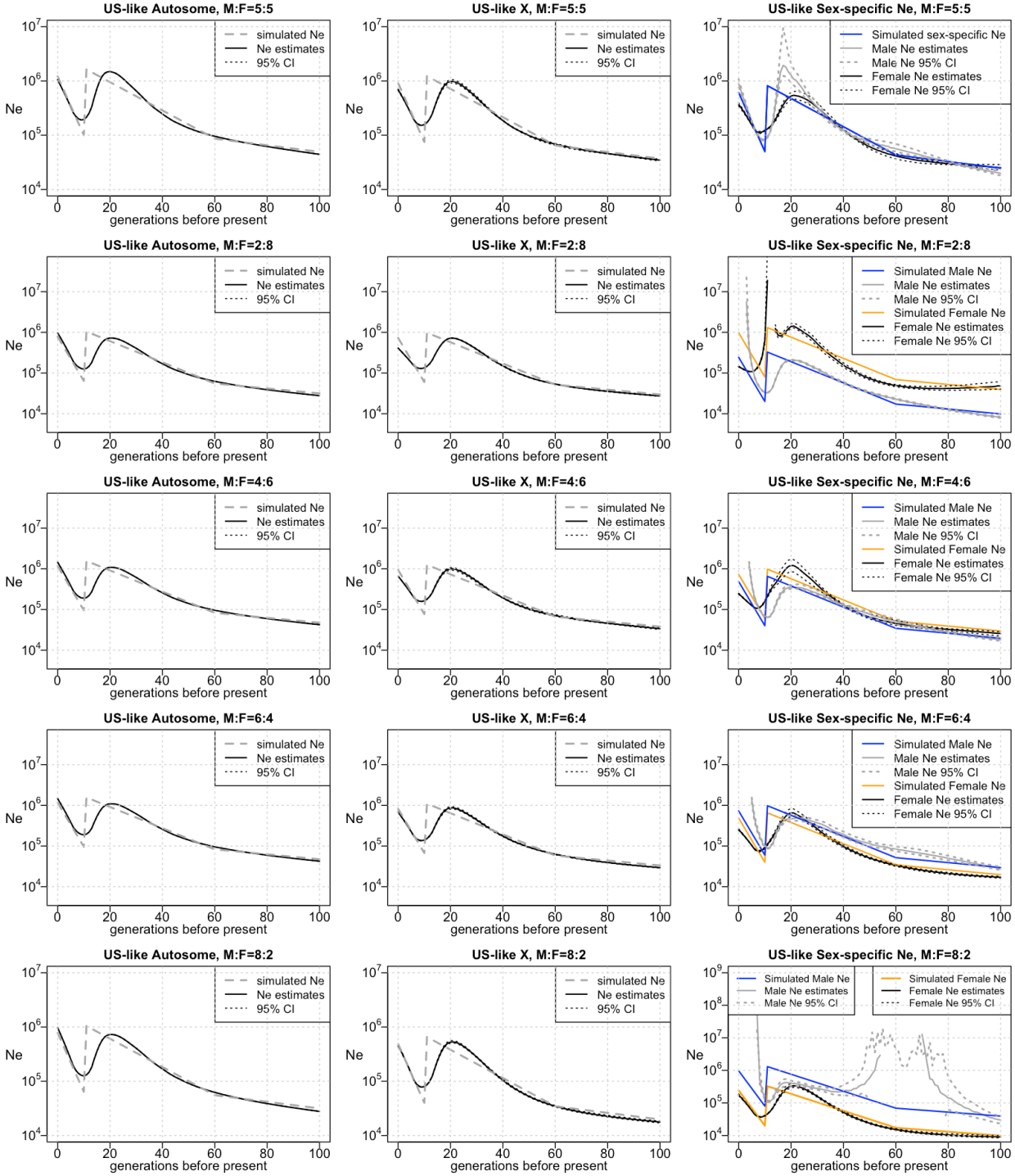


Figure 2.5. Estimates of autosome, X chromosome, and sex-specific N_e in US-like simulation studies.

Autosomal N_e is shown in the left column, X chromosome N_e in the middle column, and sex-specific N_e in the right column. From top to bottom, the first row displays results from a UK-like simulation with equal sex ratio, and the following rows show results with 80%, 60%, 40%, and 20% females. For the N_e plots the Y-axes are on a log scale. Y-axes show N_e plotted on a log scale. In cases where the estimated sex-specific N_e is negative (see Methods), it is not shown.



2.6 Analysis of real data

We applied the above analysis pipeline on SNP array data from the UK Biobank[99], whole genome sequence data from UK Biobank, and whole genome sequence data from the TOPMed project[100] to infer the autosomal and X chromosome effective population size and investigate the sex-specific population history of several human populations.

2.6.1 UK Biobank data

The UK Biobank is a large-scale biomedical database that contains in-depth genetic, physical and health data collected between 2006 and 2010 on half a million UK participants aged between 40 and 69.[101] Genome-wide genetic SNP-array data were collected on every participant using the UK Biobank Axiom array that assays approximately 850,000 genetic variants across genome.[99] Recently, the UK Biobank whole-genome sequencing (WGS) consortium released high-coverage whole genome sequence data for 200,031 study participants.[102] The whole genome sequence data was phased using Beagle 5.4[103] and the previously-described phasing pipeline[104] on the UK Biobank Research Analysis Platform. We compared the estimation of effective population size obtained using different types of genetic data.

We first analyzed the population history of the White British participants, which is the largest ethnic group in the UK Biobank cohort. The UK Biobank genotype data include 221,141 White British females and 187,802 White British males. The UK Biobank sequence data include 91,532 White British females and 75,298 White British males. Since IBDNe has limits on the number of IBD segments that it can process, we randomly selected 5,000 White British females and 5,000

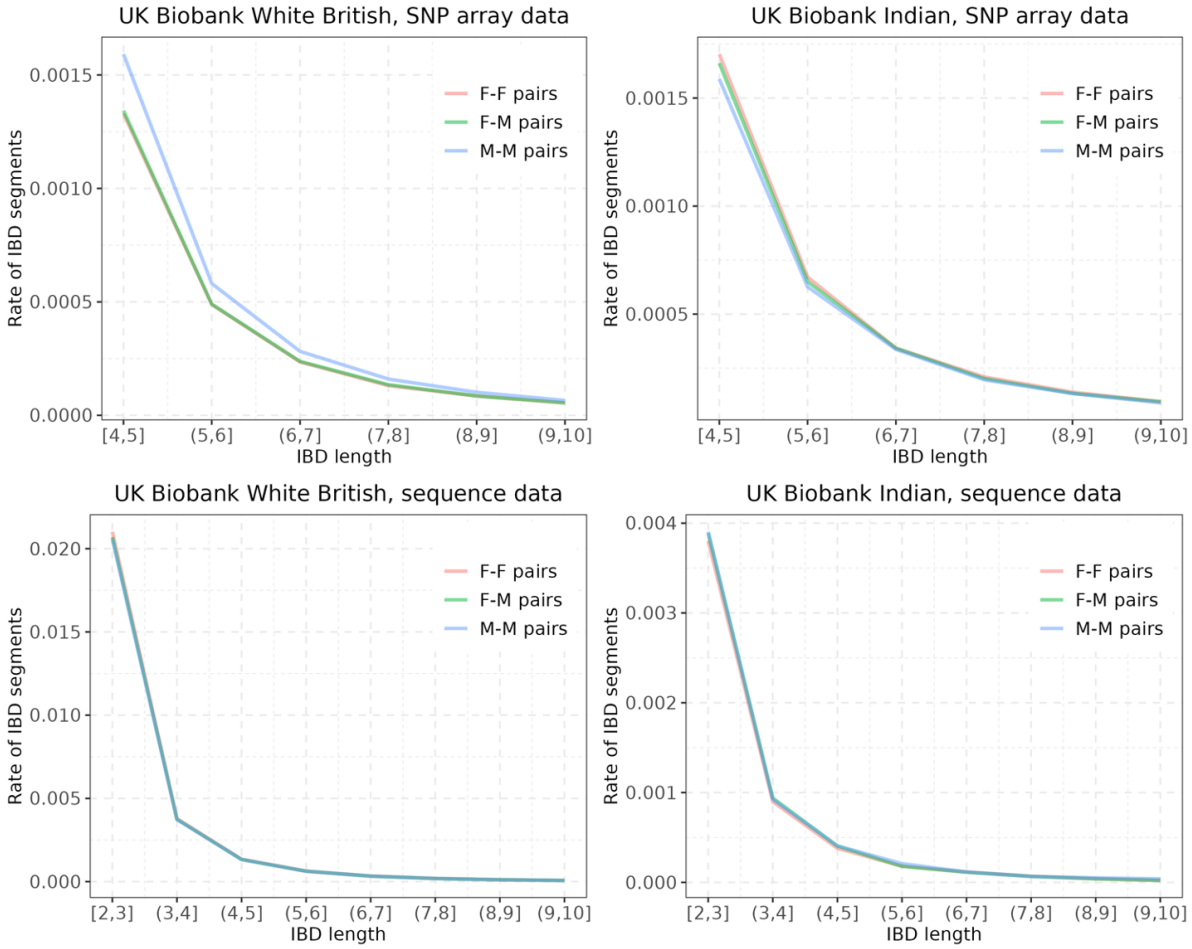
White British males for estimation of the autosome N_e. For analysis of the X chromosome N_e using UKB genotype data, we removed 33,339 randomly selected females to ensure equal numbers of females and males in the sample. For analysis of the X chromosome N_e using UKB sequence data, we removed 16,234 randomly selected females to ensure equal numbers of females and males in the sample. On both the sequence data and the genotype data, close relatives are removed using default settings of IBDNe for analysis of the autosomes. For analysis on the X chromosome, we used the UK Biobank's kinship estimates to identify and remove IBD segments from sibling pairs and parent-offspring pairs.[99]

The UK Biobank cohort also includes participants from several ethnic minority groups including Black British, Indian, Pakistani, Asian, and Bangladeshi. Among these, we chose to analyze the effective population size of the Indian group, which is the largest minority group in the UK Biobank, although this group itself contains considerable diversity. There are genotype data for 2885 males and 2775 females with Indian ancestry in UK Biobank. We removed 110 randomly selected males to achieve equal numbers of females and males for the subsequent analysis. There are sequence data for 1258 males and 1293 females with Indian ancestry in the UK Biobank. We removed 35 randomly selected females to achieve equal numbers of females and males for the subsequent analysis. By default, IBDNe automatically removes IBD segments from pairs of related individuals and generates a list of these related pairs. We manually removed X chromosome IBD segments for these pairs of related individuals prior to estimation of X chromosome N_e using IBDNe.

For analysis of the sequence data, we considered 2 cM as the IBD length threshold and estimated effective population size over the past 100 generations. Previous IBDNe analyses of SNP array data have found that effective population size estimates are less precise when estimating the effective population size more than 50 or so generations before the present due to uncertainty in the IBD-segment endpoints.[77] A previous study on UK Biobank data also showed lower false positive rates in estimated IBD segments that were at least 4 cM compared to estimated IBD segments with length below 4 cM.[32] Therefore, when analyzing UK Biobank SNP array data, we used IBD segments that were at least 4 cM to estimate population history in the past 40 generations. We used the deCODE map[93] in analyses of the sequence data and the GRCh37 European recombination map developed by Bherer et al. (2017) in analyses of the SNP array data.[105]

Using the UK Biobank SNP array data, we observe different rates of X-chromosome IBD in male-male haplotype pairs than in female-male or female-female haplotypes for both the White British group and the Indian group (Figure 2.6). This difference may be due to phasing errors in females. Males are haploid on the X chromosome so are not subject to phasing errors. Phase errors that reduce the observed IBD rate will lead in over-estimation of the effective population size. In contrast, when using the UK Biobank sequence data, the distribution of X-chromosome IBD lengths are mostly consistent in male-male, female-male, or female-female haplotypes for both the White British group and the Indian group (Figure 2.6).

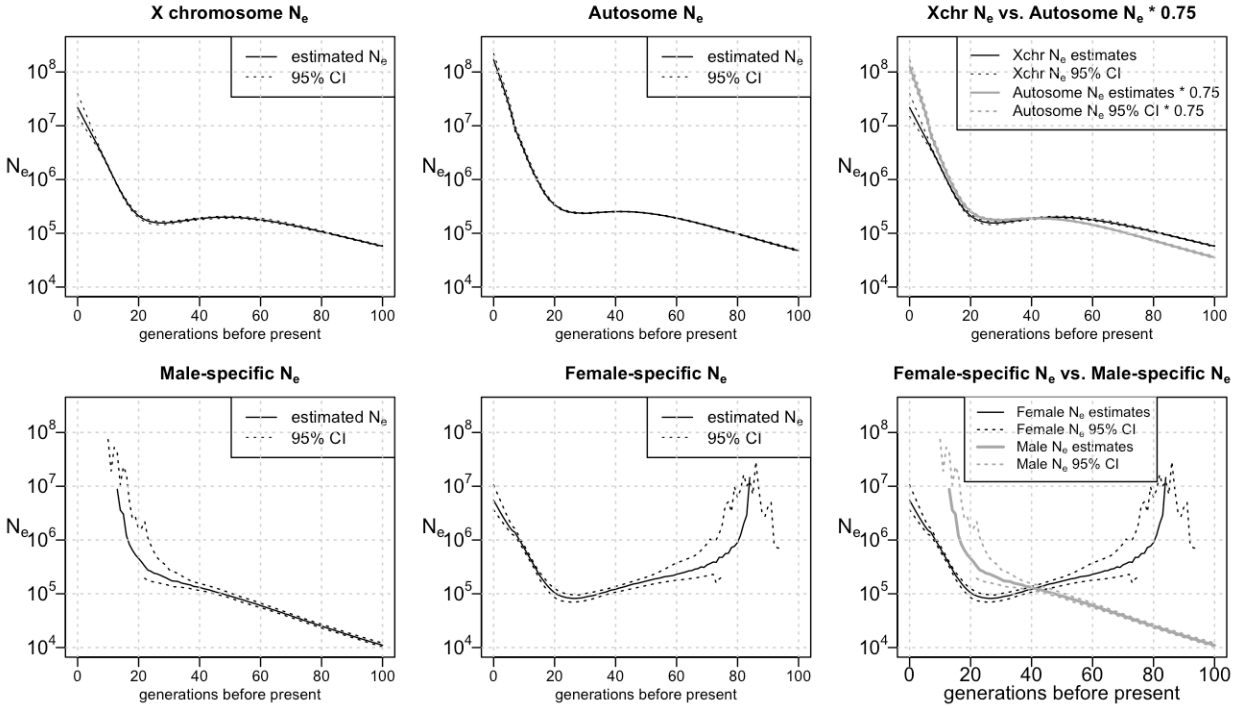
Figure 2.6. The distribution of IBD segments between sexes in UK Biobank White British group and UK Biobank Indian group.
The red line, green line and blue line show the rate of IBD segments over a series of consecutive length bins in female-female haplotype pairs, male-female haplotype pairs, and male-male haplotype pairs, respectively. IBD rates are calculated in the same way as in the simulation study. IBD lengths are measured in sex-averaged units.



For both the White British group and the Indian group, the estimated trajectory of autosome N_e using the sequence data showed a similar trend to that using the SNP array data in the past 40 generations. For the UK Biobank's White British population, the estimated autosomal N_e trajectory showed a high rate of growth in the most recent 20 generations. Before that, this population had no noticeable change in its effective population size (Figure 2.7 and Figure 2.8). For the UK Biobank Indian population, the estimated autosome N_e trajectory shows slow growth

until around 10 generations ago and a higher rate of growth in the past 10 generations (Figure 2.9 and Figure 2.10).

Figure 2.7. Effective population size of the UK Biobank White British group using sequence data.
From left to right, the top row shows the estimated X chromosome N_e, the estimated autosome N_e, and a comparison of the estimated X chromosome N_e with 75% of the estimated autosome N_e. The bottom row displays the male-specific N_e, the female-specific N_e, and a comparison between them. For the N_e plots, the Y-axes show N_e on a log scale. In cases where the estimated sex-specific N_e or its confidence band is negative (see Methods), the negative values are not shown.



However, the estimated effective population sizes at the current generation using sequence data are substantially higher compared to the results estimated using SNP array data. Using the SNP array data, the estimated current effective population size is 40 million (95% confidence interval = 35-47 million) for the White British group (Figure 2.8) and is 4.2 million (95% confidence interval = 3.7-5.0 million) for the Indian group (Figure 2.10). In contrast, using the sequence data, the estimated current autosome effective population size is 169 million (95% confidence

interval = 139-221 million) for the White British group (Figure 2.7) and is 89 million (95% confidence interval = 47-221 million) for the Indian group (Figure 2.9). IBDNe estimates the most recent generations by extrapolating the growth rate of earlier generations and doesn't account for a possible recent decrease in population growth rate, hence the N_e for generation 0 may be overestimated.[77] Furthermore, differences the accuracy and power of IBD detection between SNP array and sequence data can lead to variations in the length distributions of observed IBD segment, resulting in discrepancies in effective population size estimates derived using different data types.

Figure 2.8. Effective population size of the UK Biobank White British group using SNP array data.
From left to right, the top row shows the estimated X chromosome N_e, the estimated autosome N_e, and a comparison of the estimated X chromosome N_e with 75% of the estimated autosome N_e. The bottom row displays the male-specific N_e, the female-specific N_e, and a comparison between them. For the N_e plots, the Y-axes show N_e on a log scale. In cases where the estimated sex-specific N_e or its confidence band is negative (see Methods), the negative values are not shown.

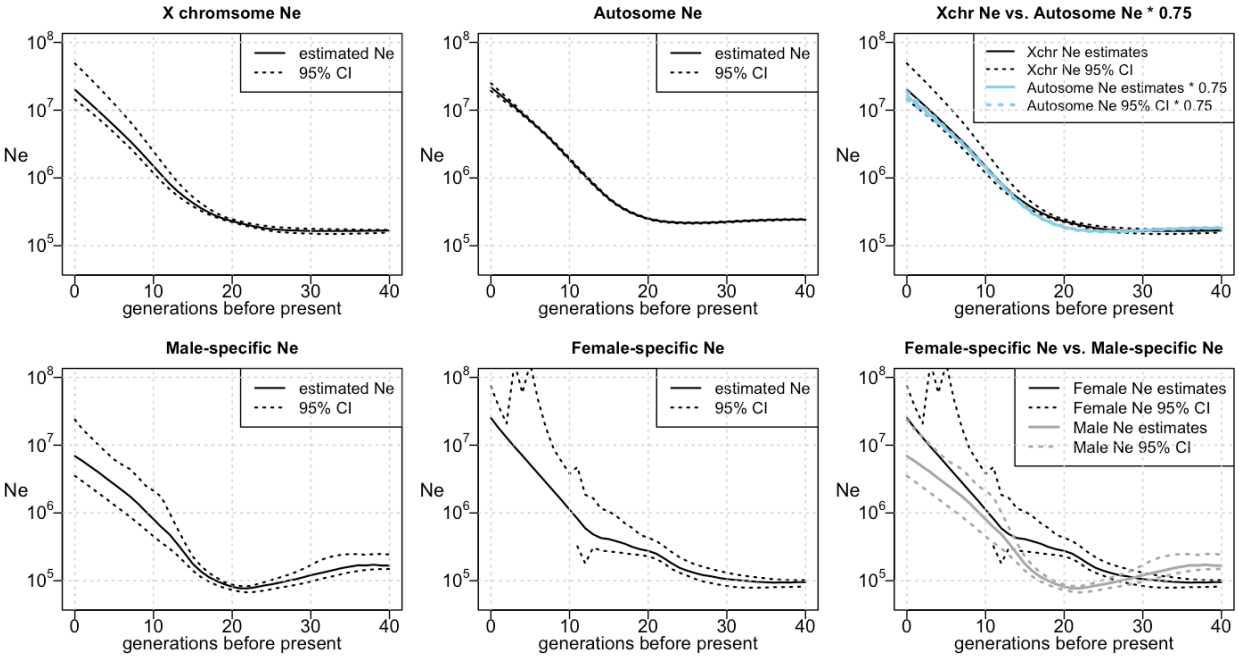
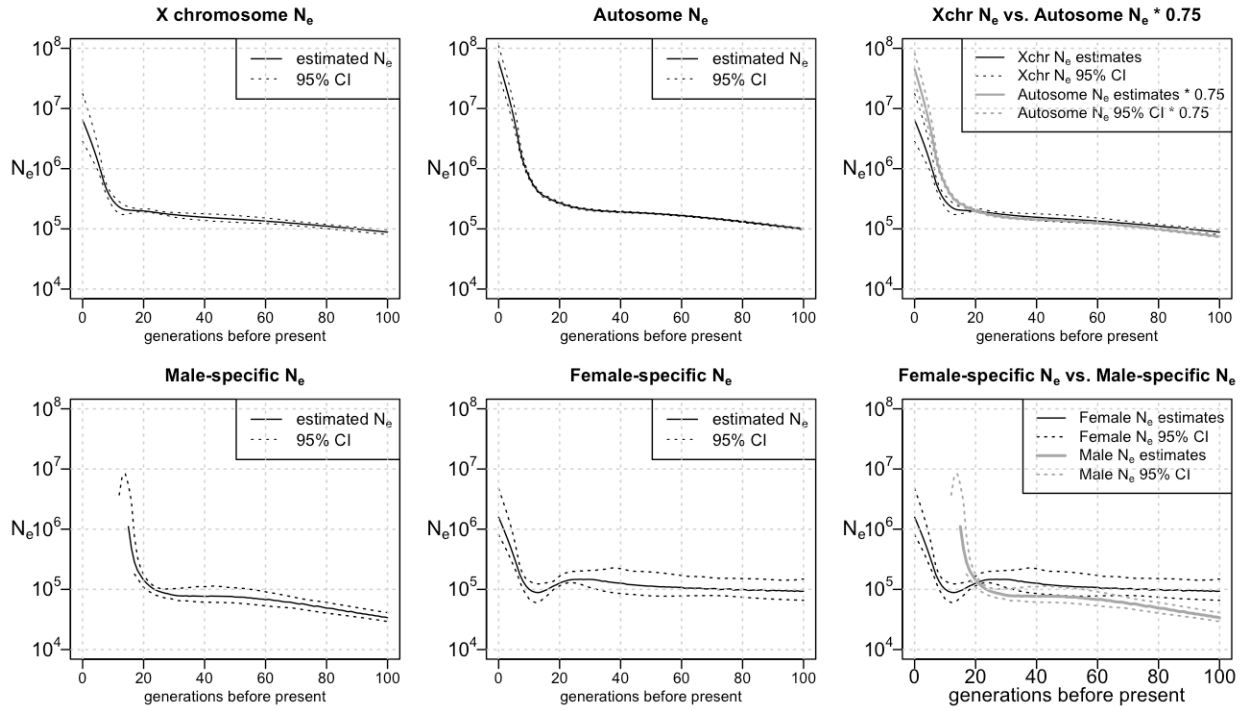


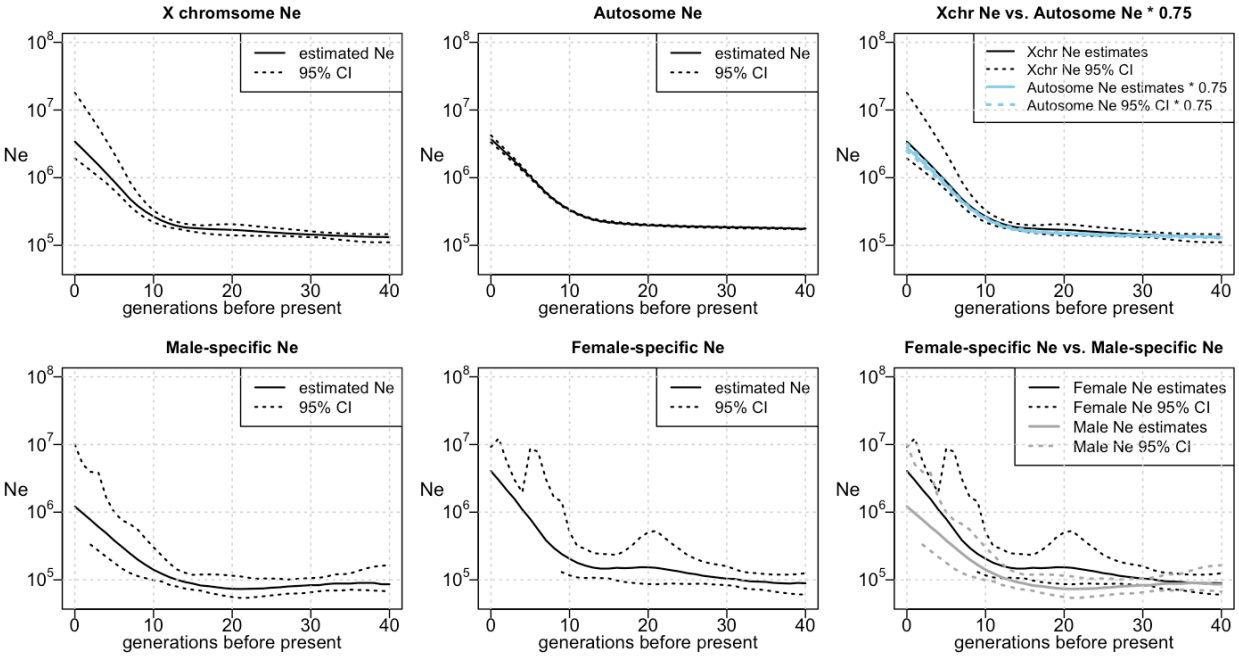
Figure 2.9. Effective population size of the UK Biobank Indian group using sequence data. From left to right, the top row shows the estimated X chromosome N_e, the estimated autosome N_e, and a comparison of the estimated X chromosome N_e with 75% of the estimated autosome N_e. The bottom row displays the male-specific N_e, the female-specific N_e, and a comparison between them. For the N_e plots, the Y-axes show N_e on a log scale. In cases where the estimated sex-specific N_e or its confidence band is negative (see Methods), the negative values are not shown.



The inferred X chromosome N_e has a similar shape as the inferred autosome N_e for both groups using either sequence data or SNP array data. The estimated X chromosome N_e obtained by splitting the X chromosome into six pieces to enable bootstrapping is consistent with the N_e estimated on the undivided X chromosome (Figure S4). The estimated X chromosome N_e is fairly close to 75% of the inferred autosomal effective population size, which is the expected effective population size that would be obtained from the X chromosome data if the female and male effective population sizes have been equal in these populations. Notably, in the analyses of SNP array data, the confidence band for the X chromosome N_e overlaps that of 75% of the

autosome N_e over the entire 40 generation period for the UK Indian population (Figure 2.8), as well as in the most recent 10 generations for the UK White British population (Figure 2.10).

Figure 2.10. Effective population size of the UK Biobank Indian group using SNP array data.
From left to right, the top row shows the estimated X chromosome N_e, the estimated autosome N_e, and a comparison of the estimated X chromosome N_e with 75% of the estimated autosome N_e. The bottom row displays the male-specific N_e, the female-specific N_e, and a comparison between them. For the N_e plots, the Y-axes show N_e on a log scale. In cases where the estimated sex-specific N_e or its confidence band is negative (see Methods), the negative values are not shown.



In comparison, the estimated X chromosome effective size is more divergent from the estimated 75% autosome effective size in analyses using the sequence data. The estimated 75% autosome effective size is lower than the X chromosome effective size between 42 and 100 generations ago for the UK Biobank White British, while it is higher than the X chromosome effective size in the past 41 generations, although the confidence intervals for the two estimates overlap in parts of this range (Figure 2.7). In the UK Biobank Indian group, we see a similar pattern, with higher estimated X chromosome size between 22 and 100 generations ago and higher estimated 75%

autosome effective size in the past 21 generations, although in the period between 22 and 100 generations ago the estimates are very close, and the confidence intervals largely overlap (Figure 2.9).

Next, we investigate the historical sex-specific effective population sizes of the UK Biobank White British population and the UK Biobank Indian population using the estimated X chromosome and autosome N_e. In results obtained from SNP array data, for both populations, the sex-specific N_e has a similar overall trend to that of the entire population (Figures 2.8 and 2.10). Moreover, the confidence bands for the female-specific N_e and the male-specific N_e for the UK Indian population overlap during the past 40 generations, and in the UK White British population their confidence bands overlap during the most recent 10 generations.

In contrast, in analyses using the sequence data, we find that in the more recent past (up to 41 generations ago for the UK Biobank White British group and up to 21 generations ago for the UK Biobank Indian group) the estimated male effective population size is higher than the female, while this is reversed in the more distant past, ignoring generations for which estimates are negative (Figures 2.7 and 2.9). Additionally, for the UK Biobank White British, there is an apparent dip in female effective size around 27 generations ago (Figure 2.7, bottom center), but this may be an artifact similar to the dip in estimated effective size in the simulated UK-like population around 15 generations ago (Figure 2.4, top right); both of these dips occur around the timing of an increase in population growth rate. For the Indian group, a similar dip appears in estimated female effective size around 13 generations ago (Figure 2.9, bottom center).

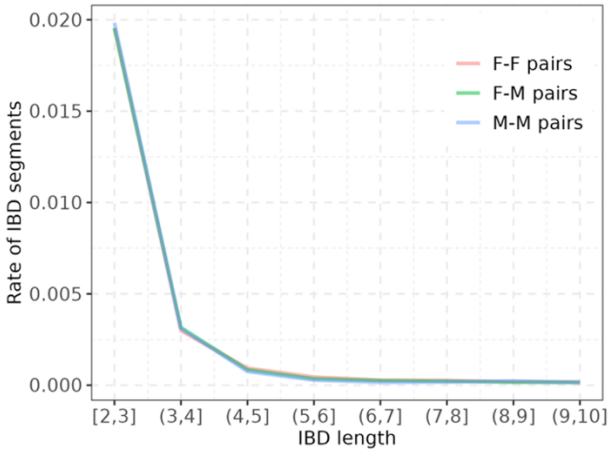
The sex-specific N_e estimated with sequence data for both the White British group and the Indian group contain negative values. The estimated female effective population size of the White British group is negative earlier than 86 generations before the present due to overly high X chromosome N_e estimates compared to autosome N_e estimates. Likewise, the estimated male effective population size of this group is negative from 13 generations ago to the present as the X chromosome N_e estimates is low compared to autosome N_e estimates. In addition, the estimated male effective population size of the Indian group is negative from 13 generations ago to the present as a result of excessively low X chromosome N_e estimates compared to autosome N_e estimates. The negative N_e estimates over these periods are not scientifically meaningful and are generally accompanied by large confidence intervals.

2.6.2 Trans-Omics for Precision Medicine (TOPMed) study data

The Hypertension Genetic Epidemiology Network Study (HyperGEN) study includes individuals with hypertension from Birmingham, AL and Winston-Salem, NC. Whole genome sequencing was performed as part of the TOPMed project (dbGaP: phs001293.v2.p1). Haplotype phasing was performed as part of the phasing of a larger set of Freeze 8 TOPMed data.[103] We estimated the effective population size from the 1,586 Black non-Hispanic participants from this study. Prior to analysis, we randomly removed 416 Black non-Hispanic female participants to ensure equal numbers of females and males in the sample. IBD segments from closely related pairs of individuals were excluded in the IBDNe analysis on the autosome data. X chromosome IBD segments from close relatives were manually removed using the list of relative pairs identified by IBDNe in the autosomal analysis. We used 2 cM as the IBD length threshold and estimated

effective population size over the past 100 generations. We used the GRCh38 deCODE map developed by Halldorsson et al. (2019) in the analyses.[93]

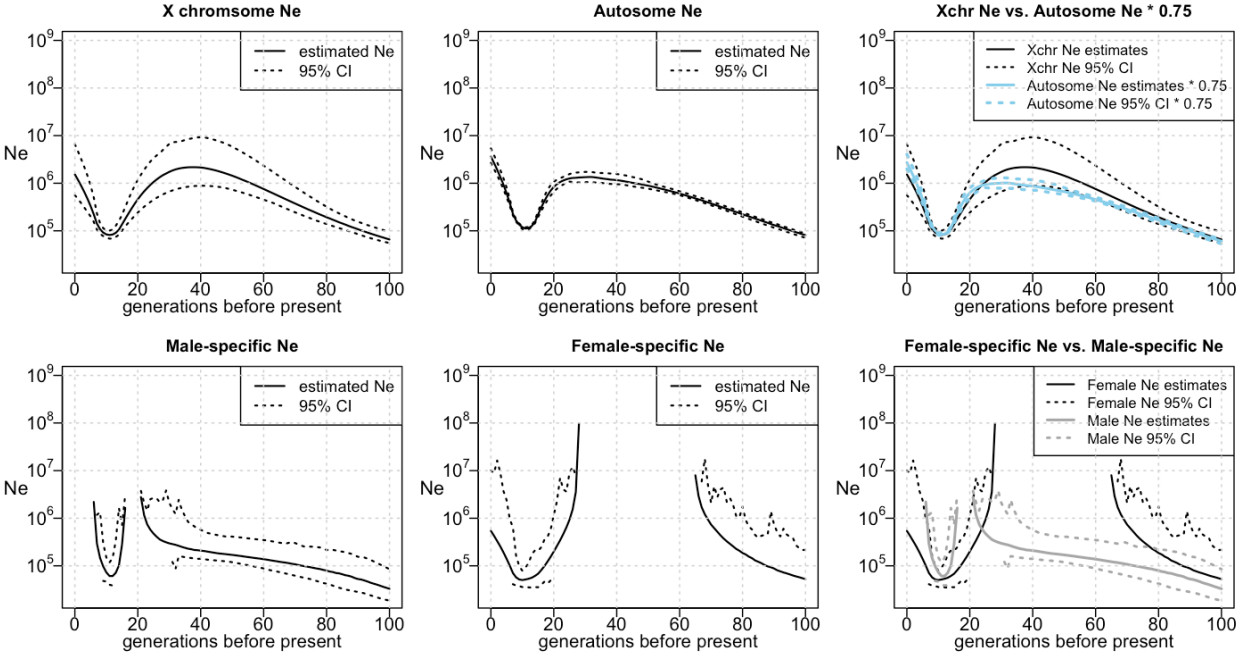
Figure 2.11. The distribution of IBD segments between sexes in the HyperGEN Black non-Hispanic group.
The red line, green line and blue line show the rate of IBD segments over a series of consecutive length bins in female-female haplotype pairs, male-female haplotype pairs, and male-male haplotype pairs, respectively. IBD rates are calculated in the same way as in the simulation study. IBD lengths are measured in sex-averaged units.



In the TOPMed HyperGEN data, the distributions of X chromosome IBD segments are consistent between female-female pairs, female-male pairs, and male-male pairs (Figure 2.11). The inferred effective population size history of the TOPMed HyperGEN Black non-Hispanic cohort shows a period of fast growth from 100 generations before present until around 40 generations ago. The effective population size then stayed relatively stable for around 10 generations. A bottleneck event from 32 to 11 generations ago reduced the N_e of this population from 1.4 million (95% CI: 1.1-1.7 million) to 112 thousand (95% CI: 105-119 thousand). A period of high growth since then has restored the N_e to 3.6 million at the present time (95% CI: 2.7-5.4 million) (Figure 2.12). The recent bottleneck event in the estimated effective population size trajectory of the HyperGEN Black non-Hispanic participants is consistent with an

immigration bottleneck resulting from the transatlantic slave trade in the history of the African American population.

Figure 2.12. Effective population size of the Black non-Hispanic group in the HyperGEN cohort.
From left to right, the top row shows the estimated X chromosome N_e, the estimated autosome N_e, and a comparison of the estimated X chromosome N_e with 75% of the estimated autosome N_e. The bottom row displays the male-specific N_e, the female-specific N_e, and a comparison between them. For the N_e plots, the Y-axes show N_e on a log scale. In cases where the estimated sex-specific N_e or its confidence band is negative (see Methods), the estimated values are not shown.



As in the analysis of UK Biobank data, the X chromosome N_e inferred from the HyperGEN group follows the same trend as the autosome N_e of this group (Figure 2.12). However, we observed more discrepancy between the estimated X chromosome N_e and the estimated autosomal N_e than was seen in the UK Biobank data, matching the comparison of results found in the US-like and UK-like simulations. Compared to the estimated autosome N_e, the X chromosome N_e estimate was high between 65 to 30 generations ago, which produced negative estimates of female-specific N_e during this period. The estimated X chromosome N_e was also

low compared to the estimated autosome N_e both during the most recent 6 generations and 21 to 18 generations ago, which led to negative male-specific N_e. The negative estimates of the sex-specific effective population size are not scientifically meaningful. The confidence band of the X chromosome N_e overlaps that of 75% of the autosome N_e across the whole trajectory, which is consistent with an approximately balanced sex ratio in the recent history of this population. Nevertheless, the large confidence band of the X chromosome N_e indicates that there is significant uncertainty involved in this result.

2.7 Discussion

Previous studies have shown that the X chromosome can provide information about demographic processes that cannot be revealed by the analysis of autosomes alone.[80,82,83,85,106,107] In this work, we focused on utilizing IBD segments on the X chromosome to infer the X chromosome effective population size. We developed a framework to model X chromosome N_e and derived the relationship between X chromosome and autosome N_e by considering the different coalescence rates among X chromosomes and autosomes. We also showed how to apply this information to calculate the female and male effective population sizes in a population as functions of the X chromosome and autosome N_e. We applied our method to estimate the X chromosome effective population size for the UK British and UK Indian populations in the UK Biobank, as well as a US Black non-Hispanic population from the TOPMed HyperGEN study.

We validated the performance of our method in simulated populations with similar histories to the UK and the US populations. We found that pronounced differences between female and male effective population sizes can be detected. Such differences are easier to detect in populations

that have not undergone recent bottlenecks, because recent bottlenecks distort the estimates of female and male effective population sizes. In the simulations, however, we observe significant discrepancies between the estimated sex-specific N_e and the true sex-specific N_e across timescales of tens of generations. This indicates that the estimated sex-specific N_e calculated from the estimated autosomal and X chromosome N_e cannot be used to test hypotheses about differences in effective population size between the sexes, particularly when those hypotheses involve differences that occur for limited time periods rather than across the full estimation timescale.

In our analyses of the sex-specific effective population size for the UK Biobank White British individuals and the UK Biobank Indian individuals, we observe a time point within the past decades of generations when male estimated N_e intersected with and continued to exceed female estimated N_e. We speculate that demographic trends such as polygamy that favored a lower male effective population size may have been more prevalent in the more distant past, and that migration rates, which have increased in recent times, may have increased more in males than in females which would increase the male effective population size relative to the female effective population size. However, it is also possible that artifacts such as better detection of long IBD segments on the X chromosome due to more accurate phasing resulting from the smaller effective population size and haploid males could be responsible for these trends.

In addition, we observed that estimates of effective population size derived from IBD segments can vary between SNP array and whole-genome sequence data. One plausible explanation for this discrepancy is the variation in marker density, which affects the resolution of IBD detection

across these data types. Sequence data, with its higher marker density and improved phasing accuracy, enables the detection of shorter IBD segments with greater confidence, thereby facilitating the reconstruction of finer-scale demographic changes over longer historical periods. However, this increased sensitivity may also amplify noise from very short segments, which can complicate the interpretation of effective population size estimates.

When analyzing the SNP array data from UK Biobank, we also observed different distributions of IBD segments by length between female-female, female-male, and male-male haplotype pairs (Figure S2), indicating that the accuracy of X chromosome IBD detection may vary between the sexes. The discrepancy in the IBD distribution between sexes is likely a consequence of phase accuracy, where the male X chromosome does not need to be phased, but the estimated phase of female X chromosome haplotypes can contain error, especially when using SNP array data. This may potentially distort the inference of the X chromosome effective population size and affect downstream analysis on sex-specific effective population sizes.

Although X chromosome IBD information has been used by previous studies for the estimation of genealogical relations between individuals, especially for kinship estimation in forensic settings[75,76,108], there has been a lack of studies that use X chromosome IBD segments to estimate recent effective population size. Our work thus fills a gap that existed in the application of X chromosome IBD information in population genetic studies.

Previous methods for estimating sex-specific population history or sex bias in human populations have relied on comparisons of genetic diversity between autosomes and the X chromosome using

allele frequency differentiation, patterns of neutral polymorphism, and the site frequency spectrum.[84,86,87,89,107,109] Most of these methods considered only a single estimate of the effective sex ratio over the entire history of a population, although this ratio can vary over time.[84,86,89,109] Some of these methods also focused on a constant overall effective population size across time, although changes in effective population size can distort these analyses.[107,109] Recently, Musharoff et al. developed a likelihood ratio test for population sex bias that considered populations of non-constant size and changing sex ratios using site frequency spectrum data.[87] However, this method requires demographic parameters to be constant within time epochs. In comparison, our approach for estimating the X chromosome effective population size and the sex-specific effective population sizes requires minimal assumptions and allows the effective population sizes to vary independently over time. The ability of our IBD-based analyses to infer effective population sizes in the past hundred generations distinguishes our approach from other methods.

There are several limitations to our approach. First, IBD-based estimation of effective population size requires a large sample of individuals from the population, although this may be less of an issue given the continuing increase in the size of genetic studies, particularly in humans. Second, our method for estimating N_e is less accurate around generations where the population experienced a drastic change in population size trajectory such as a bottleneck event. In addition, the performance of IBDNe is affected by the number and accuracy of detected IBD segments. For example, we observe wider confidence bands for the estimated N_e in the most recent generations since there tend to be fewer very long IBD segments in the sample. Similarly, we observe wider confidence bands for the X chromosome N_e compared to the autosomal N_e due to

the smaller amount of data in the X chromosome. Moreover, the estimated autosome and X chromosome effective population size tend to deviate more from the simulated effective population size in the more distant past due to the lower relative accuracy when estimating the lengths of shorter IBD segments.

Given these limitations, we note that although our method provides reliable estimates of the X chromosome effective population size, its application to estimate sex-specific population history is not suitable for rigorously testing hypotheses on sex-specific N_e in a population, because small inaccuracies in estimation of autosome and X chromosome effective population sizes are magnified when transforming these to estimates of sex-specific effective population size. We thus recommend only considering the estimated sex-specific effective population size as a tool to explore the overall pattern of sex-specific past demographic events.

Chapter 3

IDENTITY-BY-DESCENT MAPPING USING MULTI-INDIVIDUAL IBD SHARING IN OUTBRED POPULATIONS

This chapter contains material published in:
Cai R, Browning S. Identity-by-descent mapping using multi-individual IBD with genome-wide multiple testing adjustment. *bioRxiv.* 2025:2025-01.

3.1 Introduction

With the development of sequencing technology, there has been a surge in understanding the association between genetic variants and complex traits over the past two decades. Historically, there are two major classes of methods for association analysis: linkage mapping and genome-wide association studies. Linkage mapping identifies regions of the genome that are co-inherited with a trait within families, which is particularly powerful for detecting rare variants.[67,69-72] However, linkage mapping is often limited by its low resolution and requirement for family data. In contrast, as a more commonly used method at present, genome-wide association studies (GWAS) can analyze common genetic variants across large, unrelated populations, offering high resolution and the ability to detect associations with complex traits at a genome-wide level.[110-112]

While single-variant tests in GWAS are powerful for detecting common variants, they often struggle to detect structural variants or rare variants that are more likely to be population-specific, untyped in the genotype data, and have small individual effect sizes.[113-115] Variant-set tests improve upon single-variant tests by aggregating effects across multiple variants within a genomic region or a set of related genes, thus increasing power to detect rare variants.[116-119]

However, these methods often rely on pre-defined models about the genetic architecture underlying complex traits, the sparsity of causal variants, and the choice of weights on the effects of variants. In comparison, population-based identity-by-descent (IBD) mapping offers a complementary approach for linkage mapping, GWAS, and variants-set tests, providing a more flexible and comprehensive framework for association testing.

IBD mapping identifies genomic regions containing potential causal variants by searching for genomic positions where levels of IBD sharing are associated with phenotypic variation in a group of individuals for a trait of interest.[17,19,35,37,52,53,120-124] By leveraging long stretches of IBD segments shared between individuals, IBD mapping can capture the combined effects of co-occurring proximate rare alleles, even when individual variants have small or modest effects.[17,52,125] In addition, IBD mapping approaches can indirectly recover signals of untyped rare variants, structural variants, or population-specific variants tagged by IBD haplotypes.[17,19,35,52,53] Furthermore, IBD mapping is not reliant on assumptions about the underlying genetic structure or models on the effects of causal variants, which contributes to its applicability and robustness in analysis of complex traits with unknown or intricate genetic architectures.[116-119]

Previous IBD mapping studies have mostly focused on using IBD segments shared between pairs of individuals.[19,52,53,120-124] However, IBD information can extend beyond the pairwise level by considering clusters of haplotypes that are all IBD with each other.[35-38] Using IBD haplotypes shared among groups of individuals may enhance power for IBD mapping due to more accurate IBD calls and the additional information provided by specific IBD group effects that are ignored

in pairwise IBD analysis. Nevertheless, multi-individual IBD mapping has not been widely used to study variants underlying complex traits due to challenges in identifying IBD clusters in biobank-scale studies and in designing versatile models to incorporate IBD cluster information for association testing.[35,37]

In this work, we develop an IBD mapping test leveraging multi-individual IBD sharing between distantly related individuals in large, outbred populations. Motivated by variance component models in linkage analysis, we construct local relatedness matrices from multi-individual IBD sharing to quantify genetic similarities at consecutive genomic locations. Using a likelihood ratio framework, we test for associations between genomic regions and a complex trait of interest by evaluating whether local genetic similarities significantly contribute to the phenotypic variation of the trait.

3.2 Variance component model of quantitative traits

We use a variance component model for quantitative traits that is commonly used in linkage analysis.[67,69,70,72,126] Denote $\boldsymbol{Y}$ as the vector of quantitative trait values observed from a population sample. We model the trait values with the following linear mixed-effects model:

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{G} + \boldsymbol{Q}_l + \boldsymbol{\varepsilon}. \tag{3.1}$$

In the above formula, $\mathbf{X}$ is a matrix of fixed-effects covariates with effect sizes $\boldsymbol{\beta}$. $\boldsymbol{G}$ and $\boldsymbol{Q_l}$ are random-effects terms that represent the genome-wide additive effect and the location-specific effect at position l respectively, and $\boldsymbol{\varepsilon}$ is the environmental effect. The genome-wide additive effect, $\boldsymbol{G}$, is assumed to follow a multivariate normal distribution $N(\mathbf{0}, \boldsymbol{\Psi}\sigma_a^2)$, where $\boldsymbol{\Psi}$ is the

genome-wide relatedness matrix that can be estimated from genome-wide IBD sharing. We refer to $\boldsymbol{\Psi}$ as the global IBD matrix. The location-specific effect $\boldsymbol{Q}_l$ at position l is assumed to follow the multivariate normal distribution $N(\mathbf{0},\ \boldsymbol{\Phi}_l\sigma_{Q_l}^2)$, where $\boldsymbol{\Phi}_l$ is the location-specific relatedness matrix that characterizes the proportion of alleles shared IBD between each pair of individuals at a specific genomic location l. We refer to $\boldsymbol{\Phi}_l$ as the local IBD matrix at position l. The environmental effect is assumed to be independent between individuals and follow a normal distribution with mean 0 and variance σ_ε^2.

Given the mixed-effects model in (3.1), the phenotypic variance of the quantitative trait Y is represented by the sum of three variance components:

$$\mathbf{V} \coloneqq Var(\boldsymbol{Y}) = \boldsymbol{\Psi}\sigma_a^2 + \boldsymbol{\Phi}_l\sigma_{Q_l}^2 + \mathbf{I}\,\sigma_\varepsilon^2. \tag{3.2}$$

The parameters σ_a^2, $\sigma_{Q_l}^2$ and σ_ε^2 represent the effect sizes of the three variance components. The structuring matrices $\boldsymbol{\Psi}$, $\boldsymbol{\Phi}_l$, and $\mathbf{I}$ predict the covariances among individuals attributable to the effect of each variance component, respectively.

Testing whether a locus at position l is associated with the quantitative trait is equivalent to testing the variance component hypothesis $H_0\colon \sigma_{Q_l}^2 = 0$ vs. $H_1\colon \sigma_{Q_l}^2 > 0$. Many different forms of test statistics have been proposed for variance component tests. In this study, we use the log-of-odds (LOD) score[70,127] as the test statistic to take advantage of the maximal information considered under the likelihood ratio framework. To test $H_0\colon \sigma_{Q_l}^2 = 0$ vs. $H_1\colon \sigma_{Q_l}^2 > 0$, we consider the test statistic W_l defined as:

$$W_l = 2\ln(10) \times \text{LOD score}$$

$$= 2\ln(10) \times \log_{10}\left(\frac{\max_{\boldsymbol{\beta},\sigma_a^2,\sigma_{Q_l}^2,\sigma_\varepsilon^2} L_{\boldsymbol{Y}}\left(\boldsymbol{\beta},\sigma_a^2,\sigma_{Q_l}^2,\sigma_\varepsilon^2;\boldsymbol{\Phi_l},\boldsymbol{\Psi}\right)}{\max_{\boldsymbol{\beta},\sigma_a^2,\sigma_\varepsilon^2} L_{\boldsymbol{Y}}\left(\boldsymbol{\beta},\sigma_a^2,\sigma_{Q_l}^2=0,\sigma_\varepsilon^2;\boldsymbol{\Phi_l},\boldsymbol{\Psi}\right)}\right)$$

$$= 2\left(\max_{\boldsymbol{\beta},\sigma_a^2,\sigma_{Q_l}^2,\sigma_\varepsilon^2} l_{\boldsymbol{Y}}\left(\boldsymbol{\beta},\sigma_a^2,\sigma_{Q_l}^2,\sigma_\varepsilon^2;\boldsymbol{\Phi_l},\boldsymbol{\Psi}\right) - \max_{\boldsymbol{\beta},\sigma_a^2,\sigma_\varepsilon^2} l_{\boldsymbol{Y}}\left(\boldsymbol{\beta},\sigma_a^2,\sigma_{Q_l}^2=0,\sigma_\varepsilon^2;\boldsymbol{\Phi_l},\boldsymbol{\Psi}\right)\right)$$

(3.3)

where $l_{\boldsymbol{Y}}\left(\boldsymbol{\beta},\sigma_a^2,\sigma_{Q_l}^2,\sigma_\varepsilon^2;\boldsymbol{\Phi_l},\boldsymbol{\Psi}\right) = \ln L_{\boldsymbol{Y}}\left(\boldsymbol{\beta},\sigma_a^2,\sigma_{Q_l}^2,\sigma_\varepsilon^2;\boldsymbol{\Phi_l},\boldsymbol{\Psi}\right)$ is the restricted model log-likelihood that takes the following form, with the variance matrix $\mathbf{V}$ defined in (2):

$$l_{\boldsymbol{Y}}\left(\boldsymbol{\beta},\sigma_a^2,\sigma_{Q_l}^2,\sigma_\varepsilon^2;\boldsymbol{\Phi_l},\boldsymbol{\Psi}\right) = constant - \frac{1}{2}\ln|\mathbf{V}| - \frac{1}{2}\ln|\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}| - \frac{1}{2}(\boldsymbol{Y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\boldsymbol{Y}-\mathbf{X}\boldsymbol{\beta}).$$

(3.4)

To calculate the test statistic, we use $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\boldsymbol{Y}$ as the estimator for the fixed effect and find restricted maximum likelihood estimators (REML) of the variance effects $\sigma_a^2, \sigma_{Q_l}^2$ and σ_ε^2.

Under the null hypothesis, the distribution of W_l is a $\frac{1}{2}:\frac{1}{2}$ mixture of a point mass at 0 and a χ^2 distribution with one degree of freedom.[70] A proof of this result is outlined in the following. Consider re-parametrize the model with $\boldsymbol{\gamma} \coloneqq (\boldsymbol{\beta},\sigma_a^2,\sigma_\varepsilon^2) \in \Gamma \subset \mathbb{R}^{p+2}$ (suppose $\boldsymbol{\beta}$ is a vector of length p) and $\lambda \coloneqq \sigma_{Q_l}^2 \in [0,\infty)$. The test hypothesis is then expressed as $H_0\colon \lambda = 0$ vs. $H_1\colon \lambda > 0$. Note that $\lambda = 0$ is a boundary point of its parameter space. For each fixed $\lambda \geq 0$, define

$$l^*(\lambda) \coloneqq \sup_{\boldsymbol{\gamma}\in\Gamma} l_{\boldsymbol{Y}}(\boldsymbol{\gamma},\lambda).$$

In terms of $l^*(\lambda)$, the test statistic W_l can be written as:

$$W_l = 2\left[\max_{\lambda\geq 0} l^*(\lambda) - l^*(0)\right].$$

Assume $l_{\boldsymbol{Y}}(\gamma, \lambda)$ is sufficiently smooth and satisfies usual regularity and identifiability conditions. The solution of $\frac{d}{d\lambda} l^*(\lambda) = 0$, denoted as $\tilde{\lambda}$, is the unconstrained maximum likelihood estimator of $l^*(\lambda)$, i.e., $\tilde{\lambda} = \arg\max_{\lambda \in \mathbb{R}} \ l^*(\lambda)$. Thus,

$$\hat{\lambda} \coloneqq \arg\max_{\lambda \geq 0} \ l^*(\lambda) = \max\{0, \tilde{\lambda}\}.$$

Denote $U(\lambda) \coloneqq \frac{d}{d\lambda} l^*(\lambda)$ and $I(\lambda) \coloneqq -\frac{d^2}{d\lambda^2} l^*(\lambda)$. If $U(0) = \frac{d}{d\lambda} l^*(\lambda)|_{\lambda=0} < 0$, then we have that $\tilde{\lambda} < 0$, so $\hat{\lambda} = 0$ and $W_l = 0$. If $U(0) > 0$, then $\tilde{\lambda} > 0$, and by asymptotic properties of the maximum likelihood estimator, we have that $W_l \xrightarrow{d} \chi_1^2$. Since $U(0) \xrightarrow{d} N\big(0, I(0)\big)$, $\Pr\big[\tilde{\lambda} < 0\big] = \Pr[U(0) < 0] = \frac{1}{2}$. Therefore, the distribution of W_l is a 50:50 mixture of a point mass at 0 and a χ_1^2 under H_0. The same result has also been proved in Chernoff (1954), Miller (1977), and Self and Liang (1987).[128-130]

The application of the likelihood ratio statistic in variance component tests faces two major challenges. First, without closed-form solutions for REML estimates of variance components, an efficient and accurate numerical optimization algorithm is essential. Second, calculating the log-likelihood for multiple iterations during optimization can become computationally expensive with large sample sizes (N), as matrix inversion typically has a complexity of $O(N^3)$. To address these issues, we use the BFGS algorithm[131-135] from the SciPy optimize[136] library in Python due to its robustness and efficiency in handling large-scale optimization problems. Additionally, we stored the global and local IBD matrices as sparse matrices using the data structures and functions provided by the SciPy sparse library, ensuring efficient storage and manipulation of the variance matrix. Furthermore, to expedite calculation while ensuring numerical stability, we

avoid direct matrix inversion by first performing sparse matrix Cholesky decomposition and then conducting forward substitution, with utilities implemented in the scikit-sparse library.[137,138]

3.3 Multi-individual identity-by-descent inference

In this study, we use the recently developed software ibd-cluster to infer IBD sharing among multiple individuals.[38] Comparing to previous approaches that identify multi-individual IBD by clustering pairwise IBD haplotypes[35-37], the ibd-cluster software scans locations across the genome to directly cluster haplotypes that are IBD at each specific genomic location, ensuring memory- and time-efficient performance on biobank-scale dataset. This algorithm first assigns each haplotype to an individual cluster and then merges clusters when a pair of haplotypes, one from each cluster, share the same allele sequence in a genomic region of at least L cM that contains the markers of interest and extends at least T cM ($T < L/2$) in both direction from the markers of interest. We refer to L as the haplotype length threshold and T as the trimming threshold. Due to IBD transitivity, only clusters for pairs of haplotypes that are adjacent in the positional Burrows-Wheeler Transform sorting need to be merged, so this approach scales linearly with the number of individuals in the dataset in terms of computation time, memory requirements, and output size.

At a given genomic location, a group identifier representing the IBD state is assigned to each of the two haplotypes for each individual in the sample. Haplotypes sharing the same group identifier are identical by descent with each other at the given location. This format is highly efficient for computing local relatedness matrices, as it avoids iterating through a potentially quadratic number of IBD segments relative to the sample size. Instead, computations are

performed directly on a vector of haplotype cluster indices, which scales linearly with the sample size.

Using smaller haplotype length thresholds L or smaller trimming thresholds T to detect multi-individual IBD will lead to inclusion of shorter shared haplotypes in clustering, which will potentially lead to larger multi-individual IBD clusters. As a result, the local IBD matrix will tend to be less sparse, and also more likely to include false-positive IBD. In contrast, using larger L and T to detect multi-individual IBD imposes a more stringent standard when clustering IBD haplotypes, and thus the resulting multi-individual clusters will tend to contain longer but fewer segments, leading to a sparser local IBD matrix. Through simulation studies, we evaluate how different choices of L and T for multi-individual IBD inference affect the performance of our IBD mapping test.

3.4 Global and local IBD matrices

The global IBD matrix $\mathbf{\Psi}$ is a symmetric matrix, with 1's on the diagonal, and the (i, j)-th entry being the coefficient of relatedness between the i-th and the j-th individuals, which equals twice the kinship coefficient of this pair of individuals. It reflects the correlation between individuals due to shared genetic background. To construct the global IBD matrix, we infer pairwise IBD segments using the software hap-ibd[32] from phased genotype or sequence data, and estimate kinship coefficients using the software IBDkin.[40] IBDkin calculates kinship coefficients based on the proportion of the genome shared between pairs of individuals with 1 or 2 IBD haplotypes. Unlike averaging local relatedness matrices across genomic locations, which can be computationally intensive and susceptible to noise from sparse or unevenly distributed local IBD

data, IBDkin leverages genome-wide information to yield a robust and efficient estimate of relatedness. To maintain the sparsity of $\mathbf{\Psi}$, we set kinship coefficients below 0.044 (corresponding to relationships beyond third-degree relatives) to zero.

The local IBD matrix $\mathbf{\Phi}_l$ characterizes the genetic similarities between individuals at a specific position l on the genome. Similar to the global IBD matrix, $\mathbf{\Phi}_l$ is a symmetric matrix with diagonal of 1's and the (i, j)-th entry as the localized relatedness coefficient at position l between individuals i and j, which is the proportion of alleles shared IBD by these two individuals at the given position. For a pair of individuals who share 0, 1, or 2 alleles IBD, their corresponding entry in the local IBD matrix $\mathbf{\Phi}_l$ is 0, 0.5, or 1. Based on multi-individual IBD inferred using the ibd-cluster software, the localized relatedness coefficient in the local IBD matrix at a given genomic position for each pair of individuals can be obtained as below. As the first step, we summarize the proportion of each individual's chromosome in each multi-individual IBD group at the given location l with a matrix $\mathbf{A}_l$. We define $\mathbf{A}_l$ as a $N \times K_l$ matrix, where N is the number of individuals in the sample and K_l is the total number of IBD clusters at the given location l. The (i, k)-th entry of $\mathbf{A}_l$ is the proportion of individual i's chromosomes (0, ½, or 1) that are assigned to IBD group k at locus l.

The local IBD matrix at genomic location l is then obtained as $\mathbf{\Phi}_l = 2 \times \mathbf{A}_l \mathbf{A}_l^T$. This means that we sum the proportion of IBD haplotypes over all clusters at l for each pair of individuals. As a result, $\mathbf{\Phi}_l$ is a symmetric matrix with diagonal of 1's and the (i, j)-th entry as the proportion of alleles shared IBD by individuals i and j at position l, which would be non-zero only if individuals i and j have haplotypes in the same IBD cluster. Local IBD matrices constructed in

this way are guaranteed to be positive (semi)definite, which ensures that Cholesky decomposition can be performed on the variance matrix in our optimization algorithm to efficiently optimize the model likelihood.

3.5 Analysis pipeline

The required input data for our IBD mapping test include phased genotype sequence or array data, a quantitative trait of interest, fixed-effect covariates if applicable, and a genetic map.

As the first step, we construct a global IBD matrix using kinship coefficients that are estimated by the IBDkin program based on pairwise IBD segments, which are identified from the sample using the hap-ibd program.[32,40] For the hap-ibd analysis on sequence data, we set the minimum seed length to 0.5 cM, the minimum extension length to 0.2 cM, and the minimum output segments length to 2 cM. We also exclude rare variants by setting the minimum minor allele count filter to 100. For the hap-ibd analysis on SNP array data, due to the lower marker density, we set the minimum seed length to 1 cM, the minimum extension length to 0.1 cM, the minimum number of markers in a seed IBS segment to 50, and the minimum output segments length to 3 cM. Other parameters are left at their default values. We use all default parameters of the IBDkin program.

Next, we run the ibd-cluster software to obtain multi-individual IBD clusters. In our simulation studies, we applied different haplotype length thresholds L at 2 cM or 3 cM and different trimming thresholds T at 0.25 cM, 0.5 cM, or 1 cM to investigate how different IBD clustering thresholds affect the performance of our test.

We then select a collection of locations across the genome as candidates to compute the local IBD matrices and perform the IBD mapping test. The test locations do not need to be directly available in the array data or in the IBD cluster output, as we can simply use the nearest location to the testing position in the IBD cluster output to perform the test. As IBD states do not change immediately across nearby markers, we recommend running tests at 0.1 cM intervals when considering IBD clusters based on haplotypes that are at least 2 cM long. For the simulation studies to estimate genome-wide type I error rates, we ran IBD mapping tests at 0.1 cM intervals.

Our method is also applicable to genomic regions in addition to single genomic locations. For testing a given region on the genome, we compute the local IBD matrix over this region by averaging the local IBD matrices computed at a few selected locations within this region. In the simulation studies for power estimation, we tested each 0.05 cM region containing simulated causal variants using the local IBD matrix computed by averaging the IBD matrices at the start and the end positions of the region.

We developed our IBD mapping software as a Python package (https://github.com/RC0515/IBDmap). The program accepts as input the quantitative trait of interest, optional fixed-effect covariates, an estimated global IBD matrix, and designated testing locations. It then performs IBD mapping tests at all specified genomic positions in parallel to maximize efficiency. For each location, the software calculates a local IBD matrix, optimizes the likelihood function to compute the test statistic, and outputs a p-value that can be compared

against a significance threshold to assess the association between that genomic region and the trait of interest.

3.6 Benchmark

We benchmark the performance of our IBD mapping test with existing approaches for genetic association analysis. We first considered the sequencing Kernel Association Test (SKAT) developed by Wu et al. (2011)[117], a widely used variant-set test that forms the basis for this class of methods. SKAT employs a variance component model with a score statistic and uses a kernel matrix to aggregate the effects of genetic variants within a specific genomic region of interest. To benchmark the performance of our IBD mapping test on sequence data, we applied SKAT using default parameter settings, calculating power at a genome-wide significance threshold 10^{-6} as recommended in Wu et al. (2011) for genome-wide testing.[117]

On the SNP array data, we compare the performance of our test to the traditional single-variant test used in GWAS. For each simulated genomic region containing causal variants, we test every marker in the SNP data using simple linear regression between the phenotype value and the copy number of minor alleles. The minimum p-value across all tests within a region is used as the overall single-variant test p-value for that region. For the single-variant test, power is calculated using the traditional GWAS significance threshold 5×10^{-8}, which is approximately 0.05 divided by the total number of markers in our simulated SNP array data.

We also compare our method with FiMAP, a recently published IBD mapping test that employs the same variance component model as in our study but utilizes a score-type statistic for testing

the variance component hypothesis with pairwise IBD segments.[125] FiMAP approximates the test statistic using random matrix-based algorithm, allowing for computational efficiency in large-scale datasets. We investigated whether this approximation approach differs in performance from the full variance component likelihood approach used in our method. Following the analysis pipeline in Chen et al. (2023)[125], we utilize functions from the R package FiMAP to construct global and local IBD matrices using pairwise IBD segments inferred with hap-ibd[32], applying the same parameter settings as described in Section 3.5. We evaluated p-values of the FiMAP test against a significance threshold adjusted using the Bonferroni correction.

3.7 Simulation studies

We used msprime[139,140] to simulate whole-genome sequence data for 5,000 individuals, with each genome consisting of 30 chromosomes, all 100 cM in length. The demographic model for the simulation, adopted from previous studies, resembles the demographic history of the UK population.[23] In the msprime simulations, mutation occurred at a rate of 10^{-8} per base pair per generation, and recombination at a rate of 10^{-8} per base pair per meiosis.

We estimated the genome-wide type I error rate in the simulated sequence data by simulating 1000 replicates of phenotypes without genetic associations on random samples of 1000 individuals from the original dataset. On the simulated sequence data, we generated phenotypes following $\boldsymbol{Y} = \boldsymbol{g} + \boldsymbol{\varepsilon}$, where $\boldsymbol{g}$ follows a multivariate $N(\mathbf{0}, \widehat{\boldsymbol{\Psi}})$ distribution, with $\widehat{\boldsymbol{\Psi}}$ being the genome-wide relatedness matrix constructed from estimated kinship coefficients of all pairs of individuals in the sample, and ε is a vector of standard normal observations. For each replicate, we performed genome-wide IBD mapping at 0.1 cM intervals and obtained a genome-wide

significance threshold based on the observed test scores. We calculated the genome-wide type I error rate as the proportion of replicates that contain at least one test score achieving the Bonferroni threshold $0.05/30000 = 1.7 \times 10^{-6}$.

We estimate the power of our test by simulating phenotypes associated with four types of causal variants using simulated sequence data. Variants in the sequence data are classified based on minor allele frequencies (MAF) as common (MAF > 10%), low-frequency (1% < MAF < 10%), rare (0.05% < MAF < 1%), and ultra-rare (MAF < 0.05% but appearing at least once in the data). We simulate 1000 replicates of phenotypes associated with a randomly selected 0.05 cM region on the simulated genome that contains at least k causal variants in each of these four classes. For phenotypes associated with common or low-frequency variants, we consider $k = 4$ and randomly designate half of the targeted variants as causal. For rare and ultra-rare variants, we consider $k = 8$ and randomly select a quarter of the targeted variants as causal. For each replicate, the simulated phenotype for the i^{th} individual is constructed as:

$$Y_i = g_i + \sum_{l=1}^{k} \theta_l \, q_l^{(i)} + \varepsilon_i,$$

where g_i and ε_i are the i^{th} observations of the vectors $\boldsymbol{g}$ and $\boldsymbol{\varepsilon}$ that are defined as above, $q_l^{(i)} = 0, 1$ or 2 is the number of copies of the minor allele at marker l, and the effect size of each causal marker depends on its minor allele frequency as $\theta_l = \sqrt{0.05/(2MAF_l(1 - MAF_l))}$.

From the simulated sequence data, we generate a simulated SNP array dataset for each chromosome by excluding all variants with MAF below 1% and all variants associated with the simulated phenotypes, and then randomly selecting 30,000 variants per chromosome. We

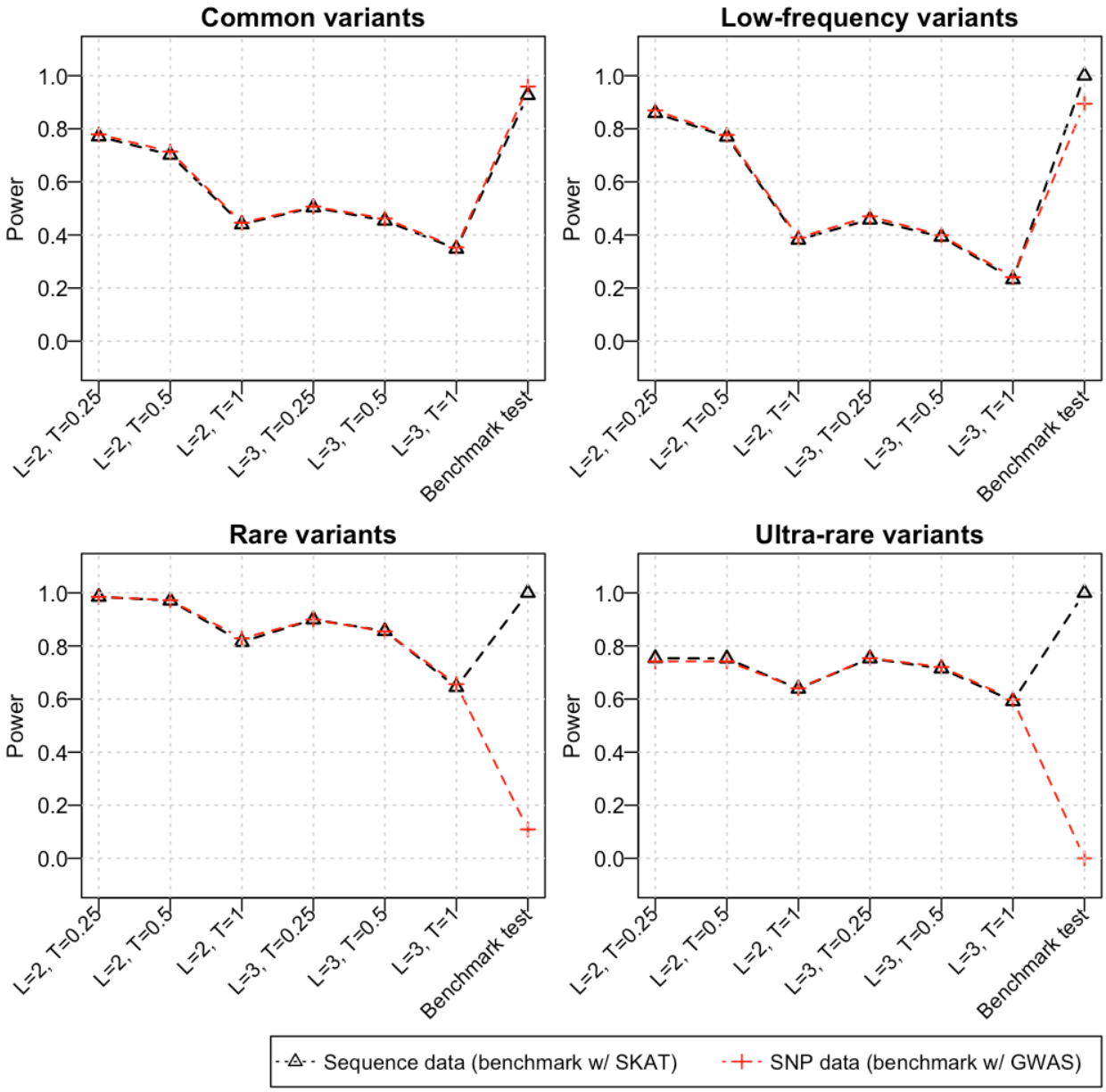
evaluate the power of our test to detect variants associated with the simulated phenotypes using both the full simulated sequence data, where all genetic variants are directly available, and the simulated SNP array data, where none of the causal variants are directly present. We evaluated the power of our IBD mapping test against the Bonferroni threshold $0.05/30000 = 1.7 \times 10^{-6}$.

Table 3.1 summarizes the estimated genome-wide type I error rates of our IBD mapping test using the Bonferroni thresholds on both the simulated sequence data and the simulated SNP array data at a genome-wide significance level of 0.05. We compared the results when different haplotype length and trimming thresholds are used for multi-individual IBD detection. Across all parameter settings, the type I error rates were lower than the nominal level of 0.05, indicating that the Bonferroni correction is conservative, especially when using a 3 cM haplotype length threshold or a 0.25 cM trimming threshold for multi-individual IBD clustering. This result aligns with expectations, as a larger haplotype length threshold or a smaller trimming threshold leads to the inclusion of longer IBD segments in multi-individual clustering, increasing the correlation between test statistics. Since the Bonferroni correction assumes independence between tests, it becomes more stringent as the correlation among test statistics grows.

Table 3.1. Genome-wide type I error rates for the IBD mapping tests using a Bonferroni adjustment at a genome-wide significance level of 0.05.
Results using different haplotype length thresholds (***L***) and trimming thresholds (***T***) for multi-individual IBD detection are shown on the first row for the simulated sequence data and on the second row for the simulated SNP array data.

L	**2cM**			**3cM**		
T	**0.25 cM**	**0.5 cM**	**1 cM**	**0.25 cM**	**0.5 cM**	**1 cM**
Simulated sequence data	0.03	0.04	0.046	0.019	0.026	0.032
Simulated SNP array data	0.033	0.035	0.045	0.022	0.036	0.013

Figure 3.1. Power of IBD mapping test for detecting common (>10% MAF), low-frequency (1-10% MAF), rare (0.05-1% MAF), and ultra-rare (<0.05% MAF) causal variants using the Bonferroni correction.
On the simulated sequence data, we compared the IBD mapping test to the sequence kernel association test (SKAT). On the simulated SNP array data, we compared the IBD mapping test to the traditional single-variant test used in GWAS. The tick labels on the x-axis indicate the haplotype length threshold (***L***) and the trimming threshold (***T***) used for multi-individual IBD detection in the corresponding IBD mapping test.



The power of our IBD mapping test to detect different types of causal variants is compared in Figure 3.1. Our IBD mapping test demonstrates similar power to detect all classes of causal variants when applied to both the simulated sequence data and the simulated SNP array data. Across all parameter settings for multi-individual IBD clustering, the highest power to detect causal variants is achieved for rare variants with MAFs between 0.05% and 1%. When using multi-individual IBD detected from haplotypes of length at least 2 cM, trimmed by 0.25 cM or 0.5 cM, the IBD mapping test reaches over 97% power to detect rare variants associated with the simulated phenotypes in both the simulated sequence and SNP-array data. These settings also yield good power (70%-80%) to detect common, low-frequency, or ultra-rare variants. Additionally, the IBD mapping test using a minimum haplotype length of 3 cM with trimming thresholds of 0.25 cM or 0.5 cM for multi-individual IBD clustering provides good power for detecting rare and ultra-rare variants.

When using the simulated sequence data as the testing dataset, the IBD mapping test does not outperform the SKAT test, which achieved 90% power for detecting common causal variants and 100% power for detecting the other three types of causal variants. When performing tests on the simulated SNP-array data, where none of the causal variants are directly available, the single-variant test achieved higher power for detecting common variants (96%). However, the single-variant test showed similar power to detect low-frequency variants as the IBD mapping test, and it achieved only 11% power for detecting rare variants and completely failed to detect ultra-rare variants. In contrast, the IBD mapping test significantly outperformed the single-variant test in detecting rare and ultra-rare causal variants across all parameter settings. The highest power for detecting rare causal variants reaches 98.5% on both the simulated sequence data and the

simulated SNP data when using a haplotype threshold of 2 cM and a trimming threshold of 0.25 cM for multi-individual IBD detection. The highest power for detecting ultra-rare causal variants also reaches around 75% in both types of data for several parameter settings, including when using a haplotype threshold of 2 cM combined with trimming threshold of 0.25 cM or 0.5 cM, or using a haplotype threshold of 3 cM combined a trimming threshold of 0.25 cM.

As a comparison, we also examined the performance of FiMAP[125] using our simulated sequence data. We used functions from its R package with pairwise IBD segments inferred by hap-ibd to construct local IBD matrices at 0.05 cM intervals across the simulated genome and test all simulated genomic regions harboring causal variants in each of the four classes. However, nearly half of the tests failed during the initial fitting of the linear mixed-effects model under the null hypothesis, due to an error that the estimated residual variance was zero (Table 3.2). This issue likely arises from the non-identifiability of the residual variance in our simulated dataset, where the global IBD matrix is extremely sparse (only four non-zero off-diagonal entries), which is nearly identical to the identity matrix. In such cases, the model may struggle to distinguish variance attributable to global IBD from residual variance, leading to scenarios where the estimated residual variance collapses to zero.

In contrast, our likelihood ratio test is robust to this non-identifiability issue because it directly evaluates the contribution of local IBD effects by comparing models with and without the local variance component, rather than requiring a precise partitioning of variance into global and residual components. Even when the global IBD matrix is nearly identical to the identity matrix

and the residual variance is difficult to separate, our likelihood ratio test can still reliably capture the effect of the variance component attributable to local IBD sharing.

Table 3.2. Performance of FiMAP in detecting different types of simulated causal variants on simulated sequence data.
Row names indicate the corresponding class of causal variants in each simulation scenario. In the first column, we summarized the proportion of failed FiMAP tests out of 1000 replicates. In the second column, for replicates where FiMAP did output valid p-values, we calculated the proportion of p-values achieved the same Bonferroni threshold 1.7×10^{-6} as we used in the power analysis of our IBD mapping test. As a comparison, in the third column, we calculated the proportion of significant IBD mapping test (using length threshold of 2 cM and trimming threshold of 0.25 cM for multi-individual IBD detection) over replicates with valid FiMAP results, and we presented the overall estimated power of our IBD mapping test on the simulated dataset in the last column.

	Proportion of failed FiMAP tests	Power of FiMAP test over replicates with valid results	Power of IBD mapping over replicates with valid FiMAP results	Power of IBD mapping test over all replicates
Common variants	0.44	0.586	0.690	0.684
Low-freq. variants	0.44	0.627	0.771	0.766
Rare variants	0.44	0.987	0.985	0.986
Ultrarare variants	0.51	0.660	0.633	0.636

Another limitation of FiMAP is that it requires precomputed global and local IBD matrices, necessitating the storage of each local matrix prior to testing. This requirement limits flexibility in selecting testing regions and imposes substantial demand on data storage, especially in fine-scale analyses of large datasets. In contrast, our method dynamically computes the local IBD matrix at each test location, reducing memory burden and allowing for more flexible, on-the-fly analyses.

3.8 Discussion

In this chapter, we present an IBD mapping approach that leverages multi-individual IBD sharing among distantly related individuals in large, outbred populations. Our method constructs local relatedness matrices from multi-individual IBD clusters and uses a likelihood ratio test to determine whether variation in local genetic similarity is significantly associated with phenotypic variation in the trait of interest. To address computational challenges in optimizing the model likelihood, we exploit the sparse structure of the data and perform fast matrix inversion using Cholesky decomposition and forward substitution, leveraging efficient optimization algorithms available in existing Python libraries. These improvements enable us to run thousands of genome-wide IBD mapping simulations within a few hours.

Through simulation studies, we evaluated the performance of our IBD mapping test across different types of testing datasets and a range of parameter settings for multi-individual IBD clustering. Our results demonstrated that our IBD mapping test maintains a conservative genome-wide type I error rate across various parameter settings for multi-individual IBD detection. In addition, our test exhibits robust power to detect various types of causal variants when appropriate haplotype length thresholds and trimming thresholds are chosen for multi-individual IBD detection, with particular strength in detecting rare or untyped variants compared to traditional GWAS single-variant tests. The power of IBD mapping was similar when applied to both SNP array and sequence data, making it a versatile approach that is suitable for use with either type of dataset.

However, in our simulation study, the IBD mapping test did not outperform the SKAT test when applied to sequence data in which the causal variants were genotyped. It was also less powerful than single-variant GWAS for variants with MAF>10% on SNP array data. These findings emphasize that the strengths of our method lie in its complementary nature rather than as a replacement for existing approaches. Nevertheless, the IBD mapping test may still provide additional perspectives in scenarios involving structural variants or other genetic features that are not directly captured in the sequence data.

We also noted that the performance of our IBD mapping approach is inherently dependent on the quality of the inferred multi-individual IBD sharing. The accuracy and resolution of multi-individual IBD detection, influenced by parameters such as haplotype length and trimming thresholds, directly impact both the type I error rate and the power of the mapping tests. Suboptimal parameter choices or inaccuracies in IBD clustering can lead to loss of power. This suggests a need for further refinement of selection strategies for parameters used for multi-individual IBD detection to balance sensitivity, error control and computational efficiency in our IBD mapping approach.

In addition, the commonly used Bonferroni correction for genome-wide multiple testing yields conservative type I error rates because it fails to account for the correlation between IBD mapping tests at nearby loci, which are often covered by the same IBD segment. This conservativeness can reduce the power of the IBD mapping test. Therefore, an appropriate multiple testing threshold that adjusts for the correlation among test statistics is required, as we demonstrate in the next chapter.

Chapter 4

MULTIPLE TESTING ADJUSTMENT FOR GENOME-WIDE IBD MAPPING SCANS

This chapter contains material published in:
Cai R, Browning S. Identity-by-descent mapping using multi-individual IBD with genome-wide multiple testing adjustment. *bioRxiv.* 2025:2025-01.

4.1 Introduction

To determine an appropriate genome-wide significance threshold, we model the joint distribution of test statistics across genomic locations in IBD mapping. Our approach adjusts for multiple testing using an analytical formula for the correlation structure of test statistics, derived from stochastic process theory. Through simulation studies, we demonstrate that this genome-wide multiple-testing adjustment effectively controls the family-wise type I error rate at the nominal level while improving the power of the IBD mapping test.

4.2 Modified Ornstein-Uhlenbeck process

Consider standardized test statistics $\{Z_l\}$ that are approximately normally distributed under the null hypothesis of no genetic associations, such that $Z_l \sim N(0,1)$ under H_0. Suppose these test statistics are computed at p equally spaced genomic locations $l_1, l_2, \ldots, l_p$ along the genome. Then, $\{Z_l\}_{l_1:l_p}$ forms a Gaussian process because the joint distribution of any subset of these statistics is multivariate normal under H_0. Additionally, the process exhibits the Markov property, as the correlation between test statistics decays with increasing distance between loci, indicating

that the process is memoryless. More specifically, $\{Z_l\}_{l_1:l_p}$ is an instance of an Ornstein-Uhlenbeck (OU) process, where the mean $E[Z_l] = 0$, and the covariance between any pair of test statistics Z_{l_i} and Z_{l_j} is given by $cov\left(Z_{l_i}, Z_{l_j}\right) = \exp\left(-\alpha\left|l_i - l_j\right|\right)$.[141-143] The parameter $\alpha > 0$ is the rate of correlation decay as genomic distances increases between test statistics, and we refer to it as the decay parameter of the OU process.

The IBD mapping statistics W_l for testing genomic locations $l_1, l_2, \ldots, l_p$ can be define in terms of standard normal $\{Z_l\}$ as:

$$W_l = \begin{cases} Z_l^2 & if\ Z_l > 0 \\ 0 & if\ Z_l \leq 0 \end{cases}, \tag{4.1}$$

If the $\{Z_l\}_{l_1:l_p}$ form an OU process, then we refer to $\{W_l\}_{l_1:l_p}$ as the modified OU process. We expect that the IBD mapping statistics will be approximately memoryless as a process along the genome, so that the modified OU process may provide a good approximation to the joint distribution of the test statistics. In what follows, we assume that the IBD mapping statistics do indeed follow a modified OU process.

4.3 Correlation of IBD mapping test statistics

Our goal is to derive $Corr\left(W_{l_i}, W_{l_j}\right)$ at any locations l_i and l_j that are spaced at intervals of length $d = \left|l_i - l_j\right|$. First, we derive $Cov\left(W_{l_i}, W_{l_j}\right) = E\left[W_{l_i} W_{l_j}\right] - E\left[W_{l_i}\right] E\left[W_{l_j}\right]$. Since the marginal distribution of Z_{l_i} is standard normal, we have:

$$E\left[W_{l_i}\right] = E\left[Z_{l_i}^2\right] * P\left(Z_{l_i} > 0\right) = \frac{1}{2}$$

$$Var[W_{l_i}] = E[W_{l_i}^2] - (E[W_{l_i}])^2 = E[Z_{l_i}^4]P(Z_{l_i} > 0) - \frac{1}{4} = \frac{3}{2} - \frac{1}{4} = \frac{5}{4}.$$

Furthermore,

$$E\left[W_{l_i}W_{l_j}\right] = E\left\{\left(Z_{l_i}Z_{l_j}\right)^2 \middle| Z_{l_i} > 0 \; and \; Z_{l_j} > 0\right\}$$

$$= \iint Z_{l_i}^2 Z_{l_j}^2 f\left(Z_{l_i}, Z_{l_j}\right) I_{[Z_{l_i}>0, Z_{l_j}>0]} \, dz_{l_i} dz_{l_j}$$

where $f\left(Z_{l_i}, Z_{l_j}\right)$ is the probability density function of the joint binormal distribution of Z_{l_i} and Z_{l_i}. We evaluated this integral using Mathematica[144], and we obtained that

$$E\left[W_{l_i}W_{l_j}\right] = \frac{1}{4} + \frac{2\pi\rho^2(d) + 6\rho(d)\sqrt{1-\rho^2(d)} + (2 + 4\rho^2(d))\sin^{-1}\rho(d)}{4\pi},$$

where we define $\rho(d) \coloneqq \exp(-\alpha d)$. Hence,

$$Cov\left(W_{l_i}, W_{l_j}\right) = \frac{2\pi\rho^2(d) + 6\rho(d)\sqrt{1-\rho^2(d)} + (2 + 4\rho^2(d))\sin^{-1}\rho(d)}{4\pi}.$$

And so

$$Corr\left(W_{l_i}W_{l_j}\right) = \frac{1}{5}\left\{\frac{2\pi\rho^2(d) + 6\rho(d)\sqrt{1-\rho^2(d)} + (2 + 4\rho^2(d))\sin^{-1}\rho(d)}{\pi}\right\}. \tag{4.2}$$

The correlation between W_{l_i} and W_{l_j} depends only on the decay parameter α and the distance between the testing locations l_i and l_j. The correlation between test statistics spaced at a fixed distance decreases as the value of α increases.

4.4 Derivation of genome-wide significance threshold

Based on (4.2), we can estimate α from observed correlations between LOD scores under the null hypothesis. To do so, we simulate phenotype values under the null hypothesis and use pairs

of IBD mapping statistics obtained at any two locations spaced at distance d to calculate the empirical correlation $\hat{f}(\alpha, d)$, and then estimate $\hat{\rho}(\alpha, d)$ by solving $f(\alpha, d) - \hat{f}(\alpha, d) = 0$ for a given d. Since $\rho(\alpha, d) = \exp(-\alpha d)$, we have $-\log\big(\rho(\alpha, d)\big) = \alpha d$. Hence, α can be estimated by the slope of a linear regression line without intercept between $-\log\big(\hat{\rho}(\alpha, d)\big)$ and d at a series of different values of d.

Given $\hat{\alpha}$, we implemented a Monte Carlo approach find the significance threshold for genome-wide multiple testing based on the distribution for $\max_l\{W_l\}$. We simulate observations $\{W_l'\}$ from the modified OU process, starting from $l = 0$ and incrementing l by the same space between test locations selected in genome-wide IBD mapping, until l reaches the end position on the genome. The detailed algorithm to simulate the modified OU process and obtain $\max_l\{W_l'\}$ is presented in Section 4.5. The simulated process $\{W_l'\}$ represents a single replicate of the genome-wide IBD mapping test under the null hypothesis. An empirical distribution for $\max_l\{W_l\}$ can be obtained by repeating the simulation thousands of times and recording $\max_l\{W_l'\}$ from each replicate. We take the 95% quantile from this empirical distribution as the 95% significance threshold of genome-wide IBD mapping on the given data. We note that the genome-wide multiple testing correction depends on the specific settings used to detect multi-individual IBD, because the length distribution of IBD haplotypes considered in the algorithm will affect the covariance between test statistics.

4.5. Simulation algorithm of the modified OU process

Input:

- `chrlens`: A vector of chromosome lengths in centiMorgans
- `alpha`: value of the decay parameter α
- `stepsize`: size of the interval between consecutive observations
- `n_iter`: total number of replicates generated

Output:

- `maxval_final`: a vector of length `n_iter` containing the maximum simulated observations across all chromosomes from all replicates

Procedure:

1. Initialize parallel processing:
   - Run the following steps in parallel for each $\texttt{chrlen}_k \in \texttt{chrlens}$ $(k = 1, \dots, K)$.
2. For each chromosome length chrlen_k, execute:
   - Generate $\texttt{x}_k$ as an array of `n_iter` random $N(0,1)$ observations.
   - Construct an array $\texttt{modified\_x}_k$ such that:

$$\text{modified\_x}_k[i] = \begin{cases} 0, & \text{if } \text{x}_k[i] < 0 \\ \text{x}_k[i]^2, & \text{if } \text{x}_k[i] \geq 0 \end{cases}$$

   - Initialize an array $\texttt{maxval}_k = \texttt{modified\_x}_k$.
3. Iterative process for $i = 2$ to $\left\lfloor \frac{\text{chrlen}_k}{\text{stepsize}} \right\rfloor$:
   - Generate an array $\texttt{newx}_k$ of `n_iter` random observations from the normal distribution with:

$$\text{mean} = \text{x}_\text{k} \cdot \exp(-\text{stepsize} \cdot \text{alpha})$$

$$\text{sd} = \sqrt{1 - \exp(-2 \cdot \text{stepsize} \cdot \text{alpha})}$$

- Construct an array `new_modified_x`$_k$ such that:

$$\text{new\_modified\_x}_\text{k}[j] = \begin{cases} 0, & \text{if newx}_\text{k}[j] < 0 \\ \text{newx}_\text{k}[j]^2, & \text{if newx}_\text{k}[j] \geq 0 \end{cases}$$

- Update `maxval`$_k$ element-wise:

$$\text{maxval}_\text{k} = \max\left(\text{maxval}_\text{k}, \text{new\_modified\_x}_\text{k}\right)$$

- Set x_k=`newx`$_k$.

4. End parallel execution
   - Collect results for all K chromosome lengths into a matrix `maxval` of size n_iter $\times K$.
5. Compute final maximum across all chromosomes

$$\text{maxval\_final}[i] = \max(\text{maxval}_1[i], \text{maxval}_2[i], \dots, \text{maxval}_K[i])$$

   for each $i \in \{1,2, \dots, \text{n\_iter}\}$.

6. Return `maxval_final`.

4.6 Analysis pipeline

To derive the multiple testing adjustment for genome-wide IBD mapping, we first obtain IBD mapping statistics under the null hypothesis. We simulate null phenotypes as the sum of independent random variables drawn from the standard normal distribution and random variables drawn from the multivariate normal distribution with mean 0 and covariance matrix $\widehat{\mathbf{\Psi}}$, where $\widehat{\mathbf{\Psi}}$ is the global IBD matrix estimated from the data. We then conduct IBD mapping tests at 0.1 cM intervals across the genome and calculate the empirical correlation $\hat{f}(\alpha, d)$ between pairs of test statistics spaced at d cM for a sequence of d from 0.1 cM to 1 cM, spacing at 0.1 cM intervals.

We then estimated $\hat{\rho}(d)$ from $\hat{f}(\alpha, d)$ based on equation (6) for each d. Finally, we fit a regression on $-\log(\hat{\rho}(d))$ against d and use the slope of the fitted line as the estimate $\hat{\alpha}$ for the decay parameter α of the modified OU process formed by IBD mapping statistics across the genome under the null hypothesis.

We perform a bootstrap analysis across chromosomes to obtain confidence intervals for $\hat{\alpha}$. Specifically, we set the number of bootstrap replicates to 10,000, and we resample chromosomes with replacement in each replicate. We take all pairs of test statistics spaced at d cM within each chromosome, and combine pairs over all sampled chromosome to calculate the sample correlations and estimate the decay parameter $\hat{\alpha}$. We calculate the 2.5^{th} and 97.5^{th} percentile of $\hat{\alpha}$ across all bootstrap replicates to construct the 95% confidence interval for the estimated decay parameter $\hat{\alpha}$.

We then use $\hat{\alpha}$ as the decay parameter to simulate observations at 0.1 cM intervals from a modified OU process that spans the same distance as the total size of the genome for 10,000 replicates, and we take the 95% quantile of maximum test statistics from all simulations as the genome-wide significance threshold.

4.7 Simulation studies

On the simulated sequence data and the simulated SNP array data described in Chapter 3, we generated a single replicate of phenotypes on all 5,000 samples as $\boldsymbol{Y} = \boldsymbol{g} + \boldsymbol{\varepsilon}$, where $\boldsymbol{g}$ follows a multivariate $N(\mathbf{0}, \widehat{\boldsymbol{\Psi}})$ distribution, with $\widehat{\boldsymbol{\Psi}}$ being the genome-wide relatedness matrix constructed from estimated kinship coefficients of all pairs of individuals in the sample, and $\boldsymbol{\varepsilon}$ is a vector of standard normal observations. We then performed IBD mapping test at 0.1 cM intervals using different haplotype length and trimming thresholds for multi-individual IBD clustering to derive the genome-wide multiple testing threshold for IBD mapping tests following the analysis pipeline described above. For each test scenario, the estimated decay parameter $\hat{\alpha}$ of the modified OU process is shown in Fig 4.1 along with its 95% bootstrap confidence interval. The 95% genome-wide multiple testing thresholds for test p-values are summarized in Table 4.1.

We calculated the theoretical correlations between random variables in each modified OU process using equation (4.2) and the estimated $\hat{\alpha}$ for each test scenario. Across all test scenarios, the observed correlations between test statistics were consistent with the theoretical correlations of the modified OU process for both the simulated sequence data (Fig 4.2) and the simulated SNP data (Fig 4.3), especially for pairs of test statistics less than 1 cM apart. These results confirm that the modified OU process provides a reliable approximation to the correlation structure of the IBD mapping test statistics.

Figure 4.1. Estimated decay parameter $\hat{\alpha}$ for genome-wide IBD mapping tests on simulated data under different algorithm parameters for multi-individual IBD detection.
The left panel shows results using a haplotype length threshold of 2 cM, while the right panel shows results for a 3 cM threshold. The y-axis represents the estimated $\hat{\alpha}$, plotted against the corresponding trimming thresholds used in each test. Results from tests on the simulated sequence data are shown in black, and results from tests on the simulated SNP array data are shown in red.

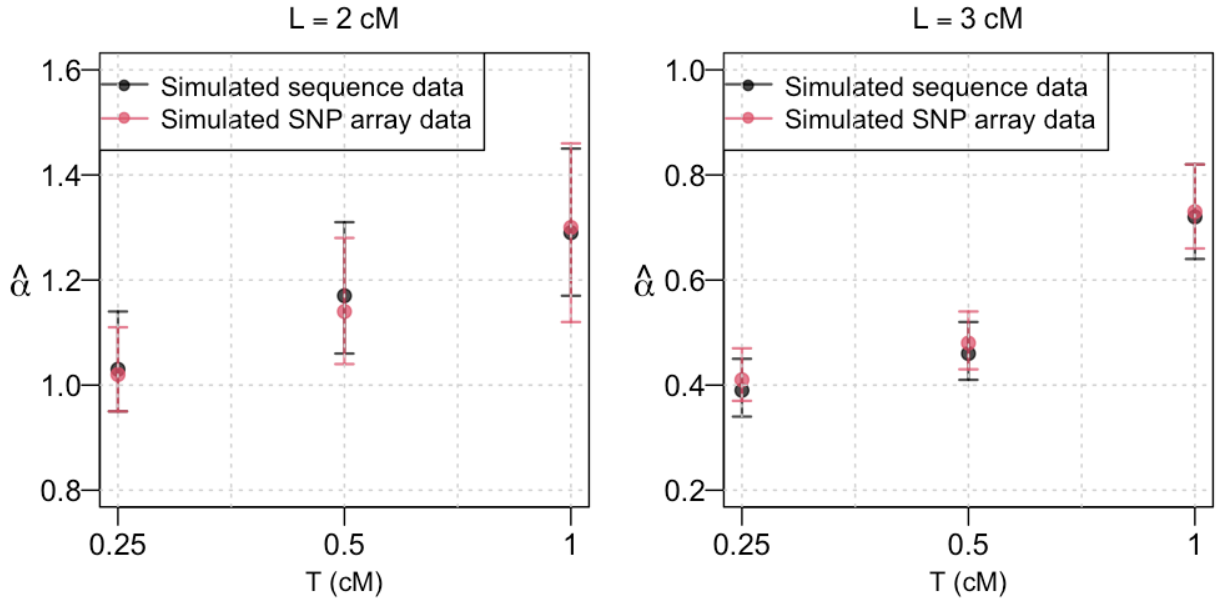


Table 4.1. Estimated decay parameters $\hat{\alpha}$ and the corresponding genome-wide 95% significance threshold for test p-values.
Results using different haplotype length thresholds (***L***) and trimming thresholds (***T***) for multi-individual IBD detection are shown on the first two row for the simulated sequence data and the following two row for the simulated SNP array data.

L		**2cM**			**3cM**		
T		**0.25 cM**	**0.5 cM**	**1 cM**	**0.25 cM**	**0.5 cM**	**1 cM**
Simulated sequence data	$\hat{\alpha}$	1.03	1.17	1.29	0.39	0.46	0.72
	Threshold	2.9×10^{-6}	2.8×10^{-6}	2.7×10^{-6}	5.2×10^{-6}	4.7×10^{-6}	3.5×10^{-6}
Simulated SNP array data	$\hat{\alpha}$	1.02	1.14	1.3	0.41	0.48	0.73
	Threshold	3.0×10^{-6}	2.8×10^{-6}	2.7×10^{-6}	5.0×10^{-6}	4.5×10^{-6}	3.5×10^{-6}

Figure 4.2. Empirical correlation between test statistics from genome-wide IBD mapping tests with phenotypes simulated under the null hypothesis on simulated sequence data, using different haplotype length thresholds (*L*) and trimming thresholds (*T*) for multi-individual IBD detection.

The 95% confidence interval for the correlation between test statistics spaced at d cM is constructed by taking the 2.5th and 97.5th percentile of the empirical correlation of test statistics spaced at d cM across all bootstrap samples. The theoretical correlation is calculated using the corresponding estimated decay parameter $\hat{\alpha}$.

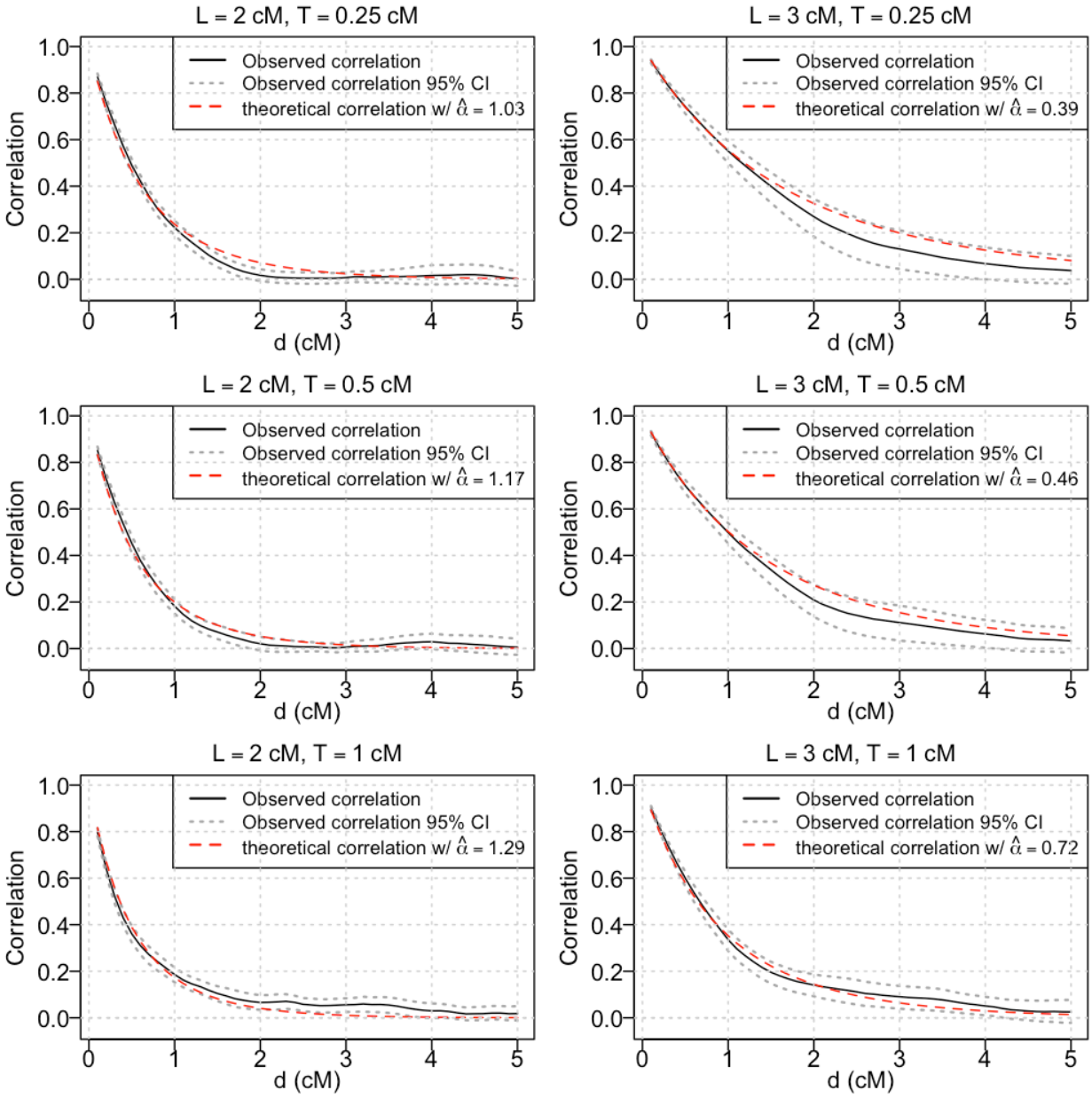


Figure 4.3. Empirical correlation between test statistics from genome-wide IBD mapping tests with phenotypes simulated under the null hypothesis on simulated SNP array data, using different haplotype length thresholds (*L*) and trimming thresholds (*T*) for multi-individual IBD detection.

The 95% confidence interval for the correlation between test statistics spaced at d cM is constructed by taking the 2.5th and 97.5th percentile of the empirical correlation of test statistics spaced at d cM across all bootstrap samples. The theoretical correlation is calculated using the corresponding estimated decay parameter $\hat{\alpha}$.

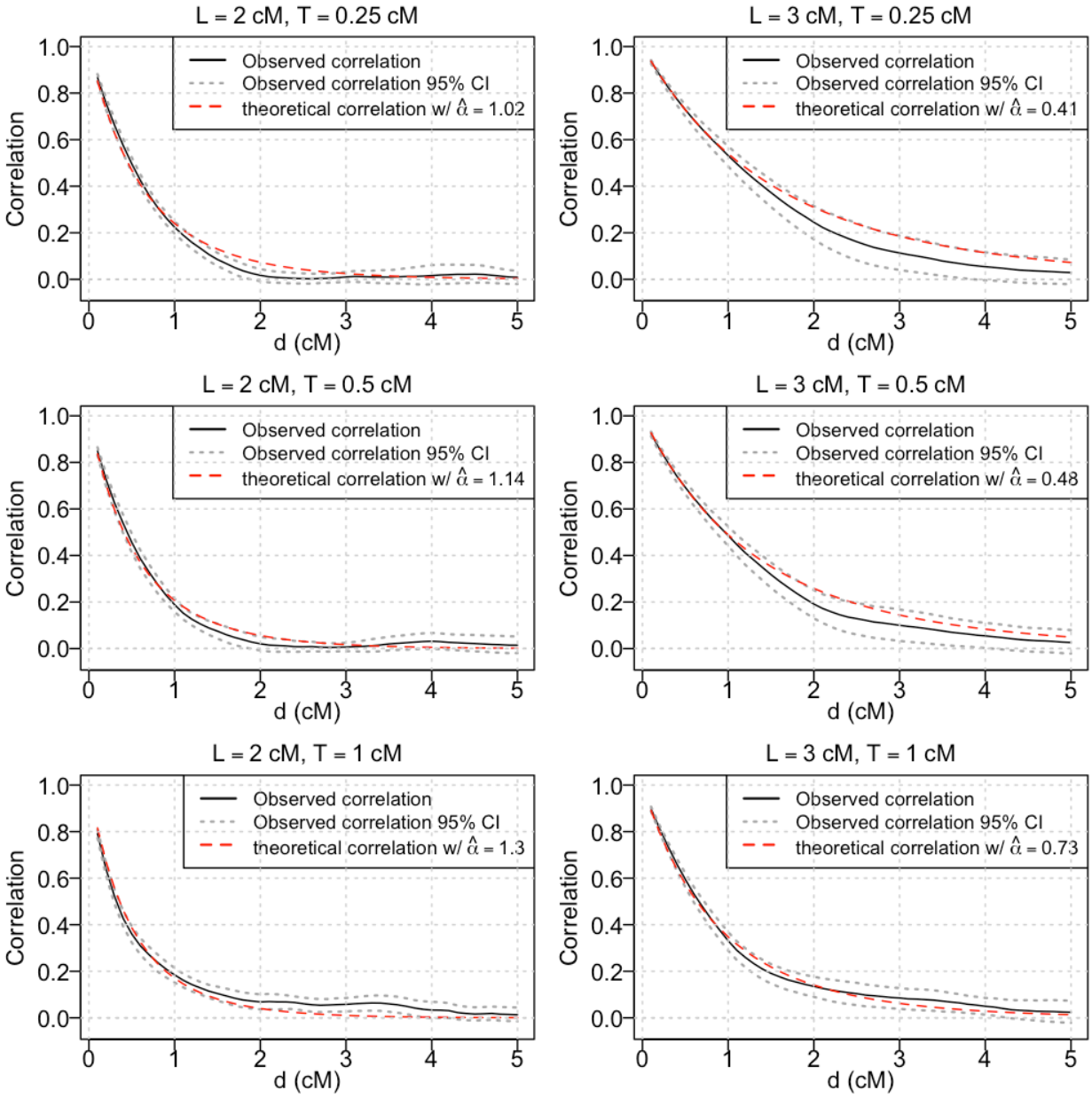


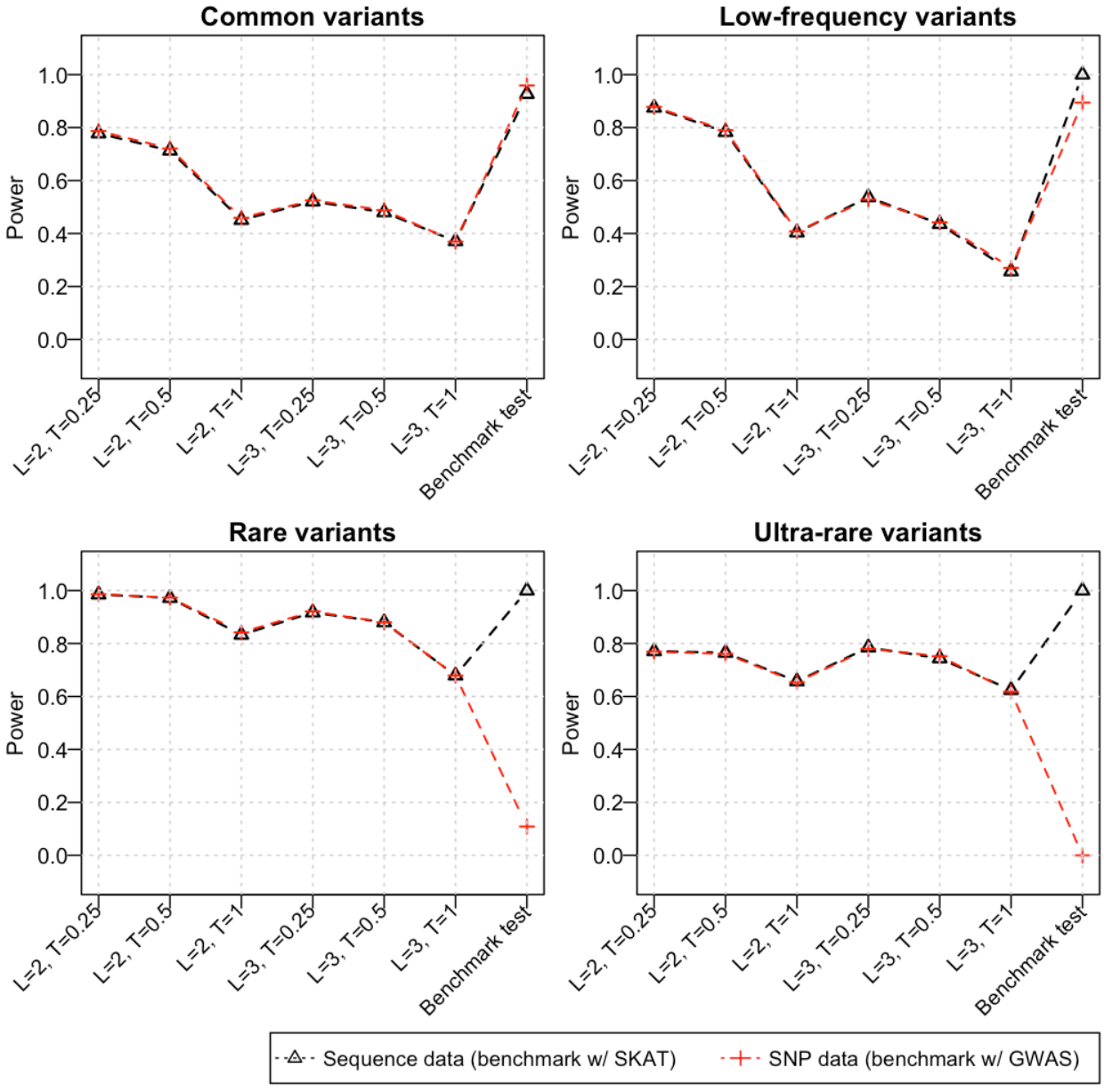
Table 4.2 summarizes the estimated genome-wide type I error rates of our IBD mapping test using the proposed genome-wide multiple testing adjustment on both the simulated sequence data and the simulated SNP array data when different haplotype length and trimming thresholds are used for multi-individual IBD detection. In most test scenarios, our proposed genome-wide multiple-testing adjustment maintains a well-controlled genome-wide type I error rate close to the nominal level of 5%.

Table 4.2. Genome-wide type I error rates for the IBD mapping tests using the proposed genome-wide multiple testing adjustment at a genome-wide significance level of 0.05. Results using different haplotype length thresholds (***L***) and trimming thresholds (***T***) for multi-individual IBD detection are shown on the first row for the simulated sequence data and on the second row for the simulated SNP array data.

L	**2cM**			**3cM**		
T	**0.25 cM**	**0.5 cM**	**1 cM**	**0.25 cM**	**0.5 cM**	**1 cM**
Simulated sequence data	0.046	0.057	0.065	0.049	0.064	0.059
Simulated SNP array data	0.051	0.053	0.067	0.052	0.056	0.032

The power of our IBD mapping test to detect different types of causal variants, calculated using the corresponding genome-wide significance threshold from Table 4.2 under various parameter settings for multi-individual IBD clustering, is compared in Figure 4.4. Our IBD mapping test still demonstrates similar power to detect all classes of causal variants when applied to both the simulated sequence data and the simulated SNP array data, but we observed power gain achieved by using the proposed genome-wide multiple-testing adjustment compared to the naïve Bonferroni threshold $0.05/30000 = 1.7 \times 10^{-6}$.

Figure 4.4. The power of IBD mapping test for detecting common (>10% MAF), low-frequency (1-10% MAF), rare (0.05-1% MAF), and ultra-rare (<0.05% MAF) causal variants using the proposed genome-wide multiple testing threshold.
On the simulated sequence data, we compared the IBD mapping test to the sequence kernel association test (SKAT). On the simulated SNP array data, we compared the IBD mapping test to the traditional single-variant test used in GWAS. The tick labels on the x-axis indicates the haplotype length threshold (***L***) and the trimming threshold (***T***) used for multi-individual IBD detection in the corresponding IBD mapping test.



In most test scenarios, we observed an absolute increase of at least 2-3% in power using our proposed multiple-testing adjustment compared to using the Bonferroni correction on both the simulated sequence data and the simulated SNP data. The power gain is more notable for tests using a larger haplotype length threshold or a smaller trimming threshold in multi-individual IBD clustering. Notably, when using a haplotype threshold of 3 cM and a trimming threshold of 0.25 cM for multi-individual IBD clustering, the power to detect low-frequency causal variants is improved by about 8% (from 45.8% to 53.7%) on simulated sequence data and by about 6% (from 47.0% to 52.8%) on simulated SNP array data using our proposed multiple-testing adjustment.

4.8 Discussion

To address genome-wide multiple testing, we propose a p-value adjustment that models the correlation between test statistics across the genome using stochastic process theory, providing an alternative to the commonly used Bonferroni correction. Through simulation studies, we demonstrate that our proposed approach for genome-wide multiple testing adjustment effectively controls the genome-wide type I error rate at the nominal significance level and enhance the power of our IBD mapping test.

In previous studies, resampling methods such as permutation tests have been employed to derive multiple-testing adjustments that account for the correlations between test statistics.[52,145,146] However, these approaches are computationally intensive, especially in large-scale genome-wide association studies. In contrast, our proposed analytical adjustment is more computationally efficient while still ensuring reliable control of the family-wise type I error rate by addressing the

inherent correlation structure in test statistics. Additionally, whereas permutation testing need to be repeated for each new trait or different model specification used in the analysis, our approach requires estimating the correlation structure only once, making it particularly advantageous for large-scale analyses where computational efficiency is critical.

The efficiency of our genome-wide multiple-testing adjustment could be further optimized. The current approach requires an additional genome-wide IBD mapping scan using phenotypes simulated under the null model at 0.1 cM intervals, in addition to the scan on the trait of interest. This increases the computational burden, particularly in large-scale datasets such as biobank cohorts. Future research could explore alternative strategies to estimate the decay parameter, such as leveraging subsets of samples or chromosomes or utilizing IBD mapping results from real traits in genomic regions without significant associations.

Chapter 5

APPLICATION OF IBD MAPPING IN UK BIOBANK

This chapter contains material published in:
Cai R, Browning S. Identity-by-descent mapping using multi-individual IBD with genome-wide multiple testing adjustment. *bioRxiv.* 2025:2025-01.

5.1 Introduction

We analyzed the systolic blood pressure of 124,376 White British UK Biobank individuals with no missing data in age, sex, the top 10 genetic PCs, the history of medication, and two measurements of systolic blood pressures at the initial assessment visit. The final phenotypic values for systolic blood pressures are obtained as the mean of the two measurements at initial assessment visit and adjusted for medication record. For individuals with a history of taking medications to control blood pressure, we increase their systolic blood pressure by 15, as suggested by previous genetic association studies on blood pressure in the UK Biobank White British cohort.[147]

5.2 Analysis pipeline

When analyzing UK Biobank data, we set $L = 2$ cM and $T = 0.5$ cM, which is the setting that we found through our simulation studies to give good power while produce reasonably sparse local IBD matrices that are computationally feasible for large-scale data. We used the GRCh37 deCODE map developed by Bherer et al. (2017) in the analyses of the UK Biobank data.[105]

We first estimated the genome-wide type I error rate when applying our IBD mapping test on the UK Biobank dataset. We selected random subsets of 1000 individuals and simulated 1000 replicates of phenotypes with no genetic associations using the model $\boldsymbol{Y} = 0.05 * age + 0.5 * sex + \boldsymbol{g} + \boldsymbol{\varepsilon}$, where $\boldsymbol{g}$ and $\boldsymbol{\varepsilon}$ are defined in the same way as the studies on the simulated data, and observed age and sex are included as fixed effects. For each replicate, based on the random subset of 1000 individuals, we conducted IBD mapping at 33,498 positions that are equally spaced at 0.1 cM intervals across the genome, and derived the genome-wide significance threshold following the pipeline described in Section 5.4. We calculated the genome-wide type I error rates based on results of 1000 replicates as described previously. As a comparison, we also calculated the proportion of replicates where at least one test exceeded the significance threshold based on the Bonferroni correction $0.05/33498 = 1.5 \times 10^{-6}$.

We then derive the genome-wide multiple testing adjustment from the genotype data of all 124,376 individuals in the analysis. We simulated a single replicate of phenotypes with no genetic associations using the model $\boldsymbol{Y} = \boldsymbol{g} + \boldsymbol{\varepsilon}$, where $\boldsymbol{g}$ and $\boldsymbol{\varepsilon}$ are defined in the same way as above. On the simulated phenotypes, we conducted IBD mapping test at 0.1 cM intervals and derived the genome-wide multiple testing threshold following the pipeline as above.

Next, to analyze genetic variants associated with systolic blood pressure from White British individuals in UK Biobank, we conducted IBD mapping tests adjusted for age, sex, age squared, their interactions, and the first 10 genetic principal components as fixed effects in the model. We assessed the test results against the genome-wide multiple testing threshold derived from the genotype data of all 124,376 individuals in the analysis. When performing a genome-wide IBD

mapping scan, to save computational resources given the large sample size, we chose a two-step approach. On each chromosome, we first ran IBD mapping test at locations spaced 1 cM apart across the genome, and we identified the top 10 locations ranked by ascending test p-values. We then zoomed into 1 cM regions centered around these locations by conducting finer-scale tests at 0.1 cM intervals. To estimate the genome-wide multiple testing threshold in the UK Biobank data, we ran IBD mapping tests at 0.1 cM intervals.

We also analyzed the data using FiMAP.[53] To balance power and computational efficiency, FiMAP recommends an IBD threshold of 3 cM on the UK Biobank data because using a shorter threshold (such as 2 cM) would substantially increase the number of IBD segments in large cohorts, resulting in denser local IBD matrices and significantly higher computational demands.[53] Following the analysis procedure outlined in Chen et al. (2023), we used hap-ibd to identify IBD segments of at least 3 cM in length and to construct local IBD matrices at 3,363 consecutive, non-overlapping 1 cM regions across the genome. We performed the FiMAP test at each 1 cM interval, adjusting the same set of fixed effects as in our IBD mapping test. FiMAP p-values are evaluated against a Bonferroni-adjusted significance threshold of $0.05/3363 = 1.5 \times 10^{-5}$.

5.3 Results

Table 5.1 summarizes the estimated genome-wide type I error rates of our IBD mapping test using the proposed genome-wide multiple testing adjustment when different haplotype length and trimming thresholds are used for multi-individual IBD detection. Similar to results on simulated data, we observed genome-wide type I error rate close to the nominal level of 5%

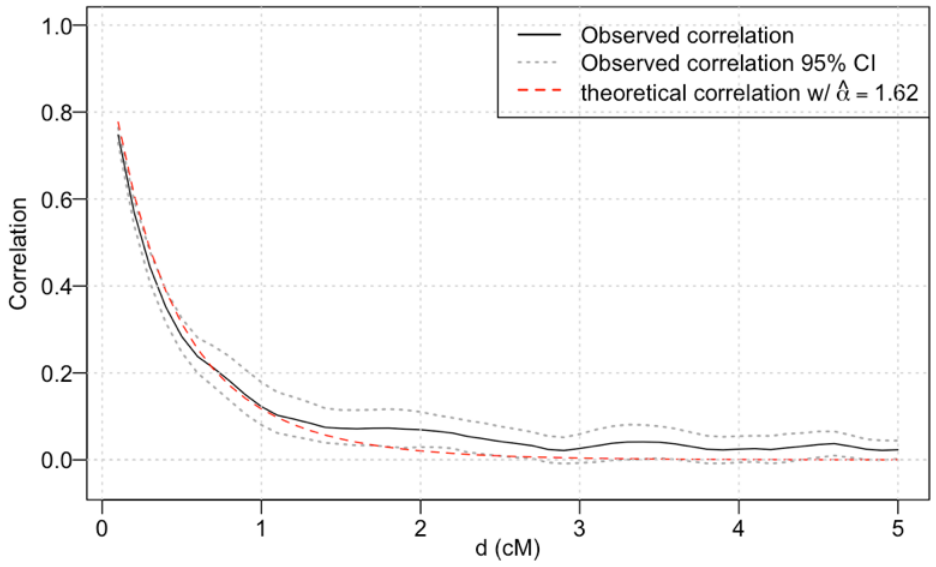
using our proposed genome-wide multiple-testing adjustment in most test scenarios, while the Bonferroni correction leads to overly conservative observed genome-wide type I error rate.

Table 5.1. Genome-wide type I error rates for the proposed genome-wide multiple-testing adjustment and the Bonferroni adjustment for IBD mapping tests at a genome-wide significance level of 0.05, with different haplotype length thresholds (L) and trimming thresholds (T) for multi-individual IBD detection.
Results are based on simulated phenotypes under the null hypothesis of no genetic association using SNP-array data from random subsets of 1000 from the 124,376 UK Biobank White British individuals considered in the analysis.

	L	2cM			3cM		
	T	**0.25 cM**	**0.5 cM**	**1 cM**	**0.25 cM**	**0.5 cM**	**1 cM**
UK Biobank SNP-array data	Proposed adjustment	0.04	0.035	0.052	0.065	0.048	0.038
	Bonferroni adjustment	0.028	0.023	0.033	0.024	0.026	0.018

On SNP array data from all 124,376 UK Biobank White British individuals considered in this analysis, the estimated multiple-testing p-value threshold for a 5% genome-wide type I error rate was 2.3×10^{-6} for IBD mapping tests using a haplotype length threshold of 2 cM and a trimming threshold of 0.5 cM in multi-individual IBD clustering. The estimated decay parameter $\hat{\alpha}$ of the corresponding modified OU process was 1.62, with a 95% bootstrap confidence interval of 1.35 and 1.92. The 95% bootstrap confidence interval of the observed correlations overlapped with the trajectory of theoretical correlation of the modified OU process with $\hat{\alpha} = 1.62$ (Fig 5.1).

Figure 5.1. Empirical correlation between test statistics from genome-wide IBD mapping tests on simulated null phenotypes using the SNP array data of 124,376 White British individuals in the UK Biobank.
The empirical correlation and their 95% bootstrap confidence intervals are compared to the theoretical correlation when $\hat{\alpha} = 1.62$.



For the genome-wide scan of systolic blood pressure from more than 124k White British individuals in the UK Biobank (Fig 5.1), the IBD mapping test at 17.759 Mb on chromosome 19 (GRCh37) achieved the genome-wide multiple-testing threshold of 2.3×10^{-6} with p-value=2.21×10^{-7}. In addition, a near-significant signal is observed at 3.790 Mb on chromosome 22 (p-value=2.51×10^{-6}). For comparison, we applied the FiMAP test to this dataset at 1 cM intervals across the genome (Fig 5.2). No significant signals were observed relative to the Bonferroni p-value threshold of $0.05/3363 = 1.5 \times 10^{-5}$.

Figure 5.2. Genome-wide IBD mapping on systolic blood pressure data from 124k White British individuals in the UK Biobank.
Negative log10 transformed p-values are plotted against tested positions on each chromosome along the genome. The result that achieved the genome-wide multiple-testing threshold is highlighted in red.

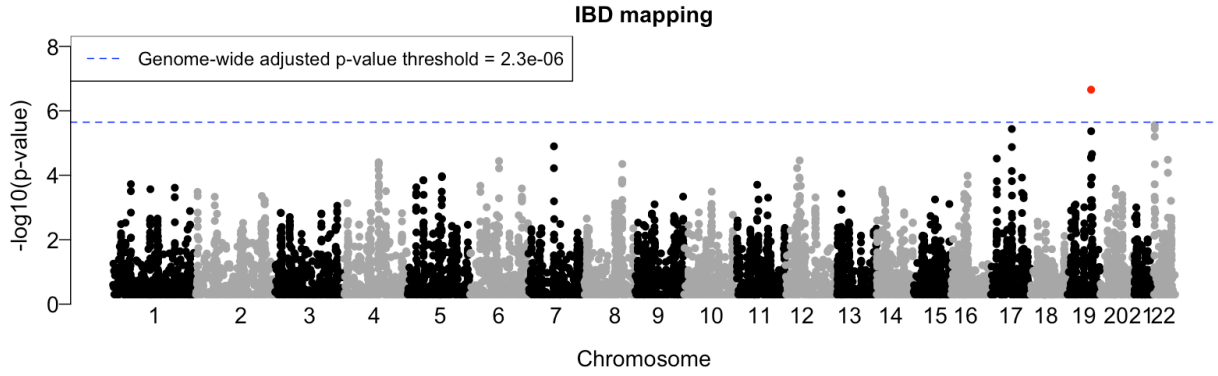
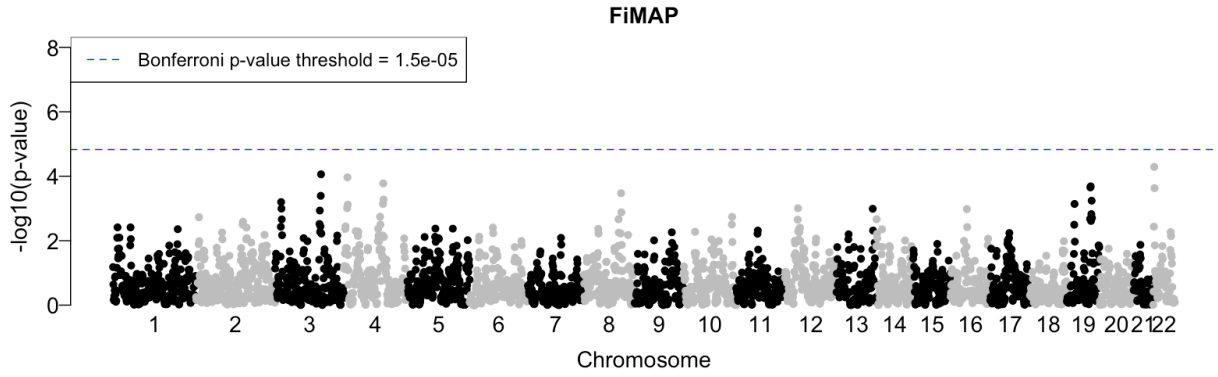


Figure 5.3. Results of applying the FiMAP test at 1 cM intervals across the genome to systolic blood pressure data from 124k White British individuals in the UK Biobank.
Negative log10 transformed p-values are plotted against tested positions on each chromosome along the genome.



5.4 Discussion

Applications of our IBD mapping approach to the UK Biobank systolic blood pressure data revealed a locus on chromosome 19 with IBD mapping p-value that surpassed the genome-wide threshold. This location is close to the *MYO9B* gene and the *USHBP1* gene on chromosome 19, which has been found to harbor variants associated with blood pressure traits among European-ancestry individuals.[148-153] This result illustrates the potential of our method for uncovering

biologically relevant signals in large-scale datasets. In comparison, the FiMAP test did not find any significant signals in the same dataset. However, our method required hundreds of hours to run, whereas FiMAP completed in just a couple of hours, indicating a trade-off between sensitivity and computational cost. While the FiMAP test is well-suited for efficient genome-wide scans in large-scale biobank cohorts, our method may be more advantageous for smaller datasets, where the higher power it offers is more attainable given manageable computational cost.

Nevertheless, the fact that our method requires a significant amount of computational time on biobank-scale cohorts limits its scalability to sample size compared to studies using simpler statistical tests, such as single-variant GWAS or SKAT-type tests. These tests generally rely on more straightforward calculations, such as comparing summary statistics, fitting linear models, or computing score-type statistics under the null hypothesis, which are computationally efficient even for large sample sizes. In contrast, our approach is based on a likelihood ratio framework, where optimizing model likelihoods is computationally intensive and becomes increasingly demanding as the sample size grows, thus restricting the scalability of our method.

Furthermore, the resolution of the significant signals identified by our IBD mapping approach was relatively low, because IBD status changes slowly along the chromosome. This suggests that future work could focus on methods such as haplotype testing to refine the IBD signals and pinpoint the causal variants under complex traits.[154,155] Another limitation of our current approach is its focus on quantitative traits with continuous values; future research extending this method to binary or categorical traits would be useful. Likewise, extending this method to test

gene-environment interaction effects by incorporating interaction terms into the model framework could further enhance its applicability for complex trait analyses.

Chapter 6

CONCLUSIONS AND FUTURE DIRECTIONS

This chapter contains material published in:
Cai R, Browning BL, Browning SR. Identity-by-descent-based estimation of the X chromosome effective population size with application to sex-specific demographic history. *G3: Genes, Genomes, Genetics*. 2023 Oct;13(10):jkad165.
Cai R, Browning S. Identity-by-descent mapping using multi-individual IBD with genome-wide multiple testing adjustment. *bioRxiv.* 2025:2025-01.

The analysis of identity-by-descent (IBD) segments provides a powerful framework for investigating human demographic history and the genetic basis of complex traits. This dissertation advances the application of IBD-based approaches in two key areas: estimating demographic parameters from IBD sharing and leveraging IBD mapping for genetic association studies. First, we introduced a method to infer the X chromosome effective population size and demonstrated how integrating autosomal and X chromosome effective population sizes enables the study of sex-specific demographic history. Second, we developed an IBD-based association testing framework for complex traits and proposed a novel multiple testing correction approach tailored for genome-wide IBD mapping. Together, these contributions enhance the statistical and computational toolkit for IBD-based inference, expanding its applications in population and statistical genetics.

Our approach for IBD-based estimation of X chromosome effective population size offers a new perspective on sex-specific demographic inference, leveraging the distinct coalescent properties of the X chromosome and autosomes. By applying our method to populations represented by

cohorts from the UK Biobank and TOPMed HyperGEN study, we demonstrated that IBD-based estimates can capture historical demographic trends, including differences in male and female effective population sizes over time. The ability to infer time-varying sex-specific N_e distinguishes our approach from traditional methods that assume a constant sex ratio or require predefined demographic epochs.

However, our study also reveals several limitations and highlights potential areas for improvement. First of all, our simulation study assumes a relatively simplified demographic model and does not explicitly account for how the effects of other evolutionary forces, like negative selection, or sex-biased demographic events, such as matrilocality, patrilocality, and sex-biased migration, may influence the inference of autosome or X chromosome effective population size. Since human populations typically experience complex demographic histories, incorporating more realistic models in simulation studies could provide valuable insights that help interpret the patterns of historical effective population sizes observed in real populations.

In addition, several factors may affect the results of sex-specific N_e estimates, including sample size, the marker density of genetic data, the resolution of IBD detection, and phasing accuracy. Differences in marker density affect the identification of short IBD segments, potentially leading to discrepancies in inferred effective population sizes across datasets. Additionally, phase accuracy may play a crucial role in IBD detection on the X chromosome, particularly in female haplotypes where phasing errors can distort IBD-based demographic estimates. While our method captures broad trends in sex-specific population size over hundreds of generations, our

simulations indicate that it is less reliable for testing hypotheses about short-term changes in sex-specific N_e, especially in populations that have undergone recent bottlenecks.

Future research could focus on hybrid approaches that integrate IBD segment analysis with complementary genomic signals, such as the site frequency spectrum or haplotype-based coalescent models, to provide a more comprehensive understanding of sex-specific demographic history, particularly for inferring fine-scale demographic changes in populations with complex histories. Furthermore, as multi-ancestry genetic studies continue to expand, adapting our method to address diverse populations with intricate admixture histories will be crucial. This will necessitate the development of models that explicitly account for complex demographic dynamics that give rise to ancestry-specific patterns of genetic diversity and IBD distribution in structured populations. Overall, our research demonstrates the potential of IBD-based approaches for demographic inference while underscoring the need for continued methodological improvements. As genomic datasets grow in size and resolution, further advancements in IBD analysis will continue to refine our understanding of human evolutionary history.

We also introduced an innovative IBD mapping approach that leverages multi-individual IBD sharing among distantly related individuals in large, outbred populations, constructing local relatedness matrices to assess the contribution of genetic similarities to phenotypic variation. To address computational challenges, we utilized fast matrix inversion techniques and efficient optimization algorithms, enabling the application of genome-wide IBD mapping tests in biobank cohorts. Additionally, we developed an analytical p-value adjustment for genome-wide multiple testing, which models correlations between test statistics using stochastic process theory. This

approach improves computational efficiency compared to traditional resampling methods, such as permutation tests, while enhancing power and ensuring reliable control of type I error rates.

Our simulation studies demonstrated the robustness of the IBD mapping test, which successfully controlled type I error rates and showed strong power to detect causal variants, especially rare or untyped variants, compared to traditional single-variant GWAS approaches. The method maintained consistent performance across both SNP array and sequence data, highlighting its versatility. However, it also revealed that while the IBD mapping test outperforms existing methods in certain contexts, it does not surpass the SKAT test when applying to high-density sequence data or the single-variant GWAS for detecting common causal variants. This emphasizes that the IBD mapping test serves as a complementary tool rather than a replacement for traditional methods. The analysis of systolic blood pressure data from White British individuals in the UK Biobank illustrates the potential of our method for uncovering biologically relevant signals, showcasing its utility in large-scale genomic studies.

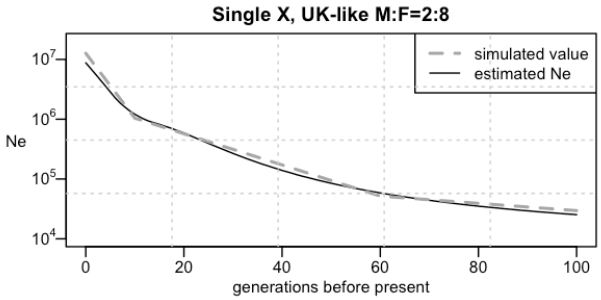
Despite the methodological advancements, our IBD mapping approach has several limitations. First, its performance is inherently dependent on the accuracy of multi-individual IBD detection, which is influenced by parameters such as haplotype length and trimming thresholds. Suboptimal parameter choices can lead to reduced power and increased type I error rates, highlighting the need for further refinement in parameter selection strategies. Second, the computational cost of our likelihood-based framework remains a challenge, particularly for large-scale biobank studies, where alternative methods employing statistical tests with simpler calculations may be preferred. Therefore, additional strategies are needed to further optimize computation efficiency and

enhance applicability for larger datasets. Moreover, the limited resolution of significant signals identified through IBD mapping presents challenges in pinpointing causal variants with high precision. Future work could integrate haplotype-based fine-mapping methods to improve localization of causal variants. Finally, our approach is currently designed for quantitative traits, and extending it to binary traits, gene-environment interactions, and other complex trait architectures would broaden its applicability.

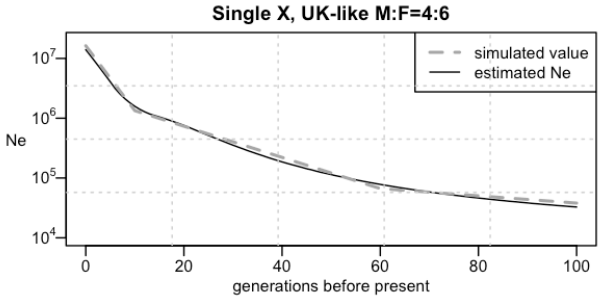
In summary, the IBD mapping approach developed in this study offers a versatile and powerful tool for detecting associations between genomic regions and complex traits, particularly those involving rare or untyped variants. Its robust performance across both SNP array and sequence data highlights its broad applicability to diverse datasets. Future improvements in computational efficiency, integration with higher-resolution methods, and extensions to models incorporating more complex genetic features could further enhance the utility of identity-by-descent information in advancing our understanding of the genetic architecture underlying complex traits across diverse populations.

SUPPLEMENTARY MATERIALS

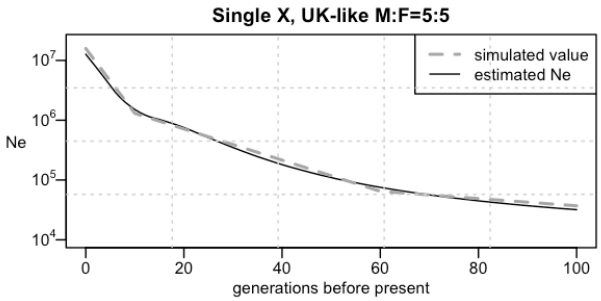
Figure S1. Effective population size estimated from an undivided X chromosome versus splitting the X chromosome into six regions in the UK-like simulation studies.
N_e estimated on a single undivided X chromosome is shown in the left column. N_e estimated by treating six separate regions of the X chromosome as six chromosomes to enable bootstrapping is shown in the right column. In each plot, the Y-axes are on a log scale. From top to bottom, each row displays result from a UK-like simulation with 80%, 60%, 50% (equal sex ratio), 40%, and 20% females, respectively.


Single X, UK-like M:F=2:8
simulated value
estimated Ne
Ne
generations before present

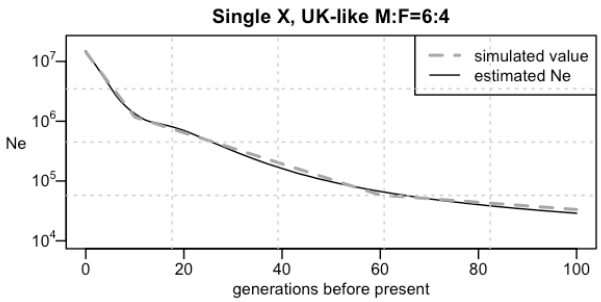

X break to 6 pieces, UK-like M:F=2:8
simulated value
estimated Ne
95% CI
Ne
generations before present


Single X, UK-like M:F=4:6
simulated value
estimated Ne
Ne
generations before present

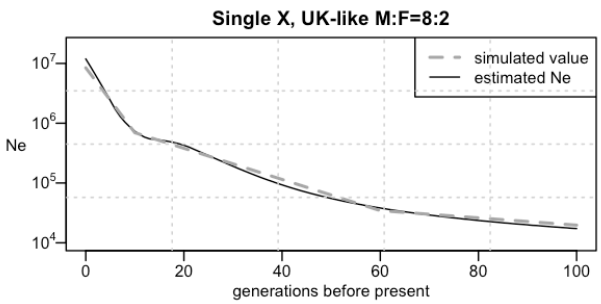

X break to 6 pieces, UK-like M:F=4:6
simulated value
estimated Ne
95% CI
Ne
generations before present


Single X, UK-like M:F=5:5
simulated value
estimated Ne
Ne
generations before present


X break to 6 pieces, UK-like M:F=5:5
simulated value
estimated Ne
95% CI
Ne
generations before present


Single X, UK-like M:F=6:4
simulated value
estimated Ne
Ne
generations before present


X break to 6 pieces, UK-like M:F=6:4
simulated value
estimated Ne
95% CI
Ne
generations before present


Single X, UK-like M:F=8:2
simulated value
estimated Ne
Ne
generations before present

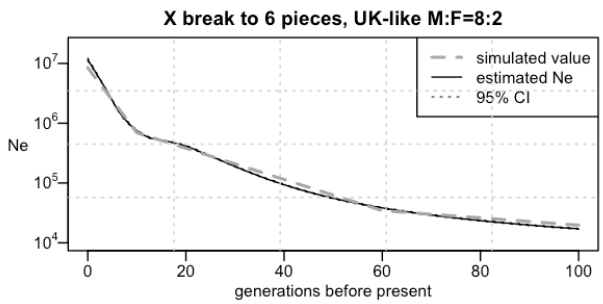
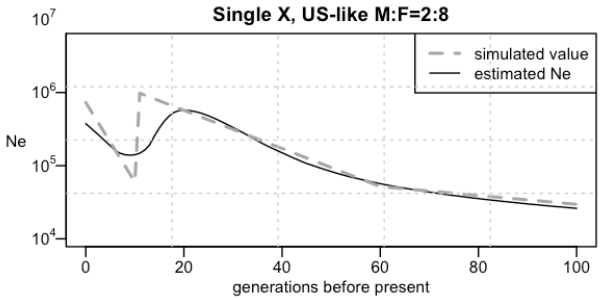
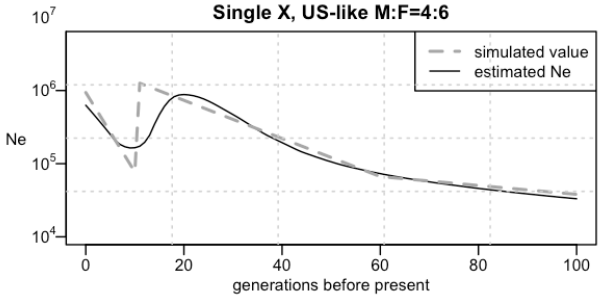
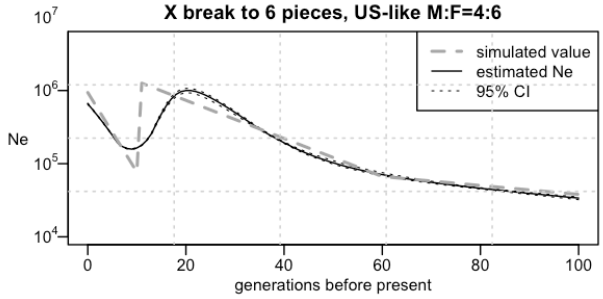

X break to 6 pieces, UK-like M:F=8:2
simulated value
estimated Ne
95% CI
Ne
generations before present

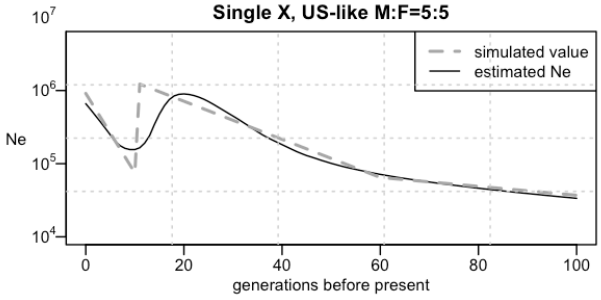
Figure S2. Effective population size estimated from an undivided X chromosome versus splitting the X chromosome into six regions in the US-like simulation studies.
N_e estimated on a single undivided X chromosome is shown in the left column. N_e estimated by treating six separate regions of the X chromosome as six chromosomes to enable bootstrapping is shown in the right column. In each plot, the Y-axes are on a log scale. From top to bottom, each row displays result from a UK-like simulation with 80%, 60%, 50% (equal sex ratio), 40%, and 20% females, respectively.

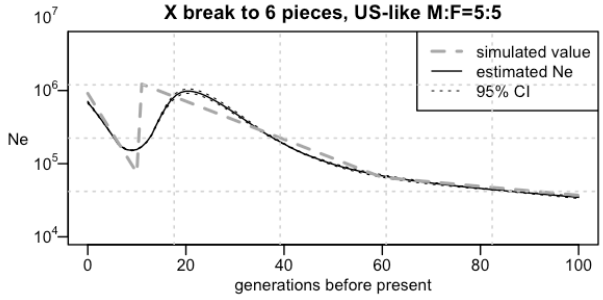

Single X, US-like M:F=2:8
simulated value
estimated Ne
Ne
generations before present

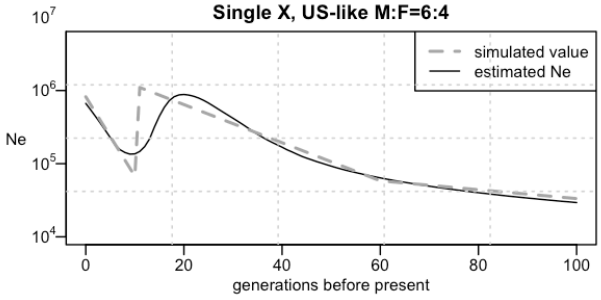

X break to 6 pieces, US-like M:F=2:8
simulated value
estimated Ne
95% CI
Ne
generations before present

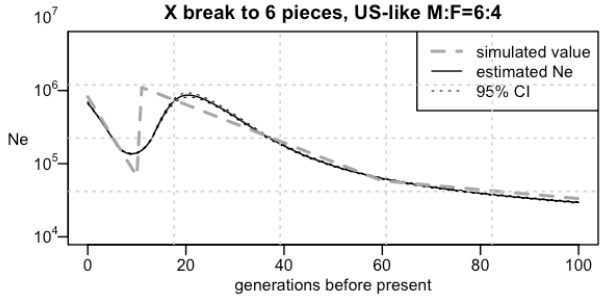

Single X, US-like M:F=4:6
simulated value
estimated Ne
Ne
generations before present

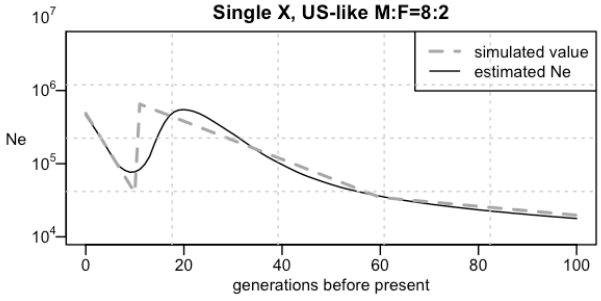

X break to 6 pieces, US-like M:F=4:6
simulated value
estimated Ne
95% CI
Ne
generations before present


Single X, US-like M:F=5:5
simulated value
estimated Ne
Ne
generations before present


X break to 6 pieces, US-like M:F=5:5
simulated value
estimated Ne
95% CI
Ne
generations before present


Single X, US-like M:F=6:4
simulated value
estimated Ne
Ne
generations before present


X break to 6 pieces, US-like M:F=6:4
simulated value
estimated Ne
95% CI
Ne
generations before present


Single X, US-like M:F=8:2
simulated value
estimated Ne
Ne
generations before present

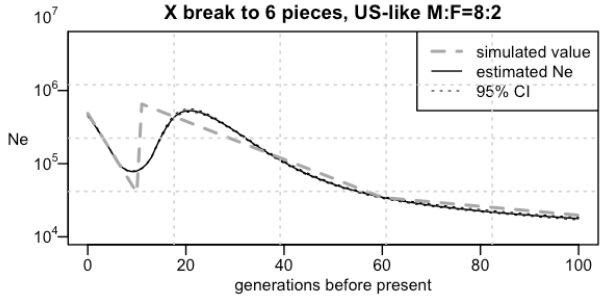
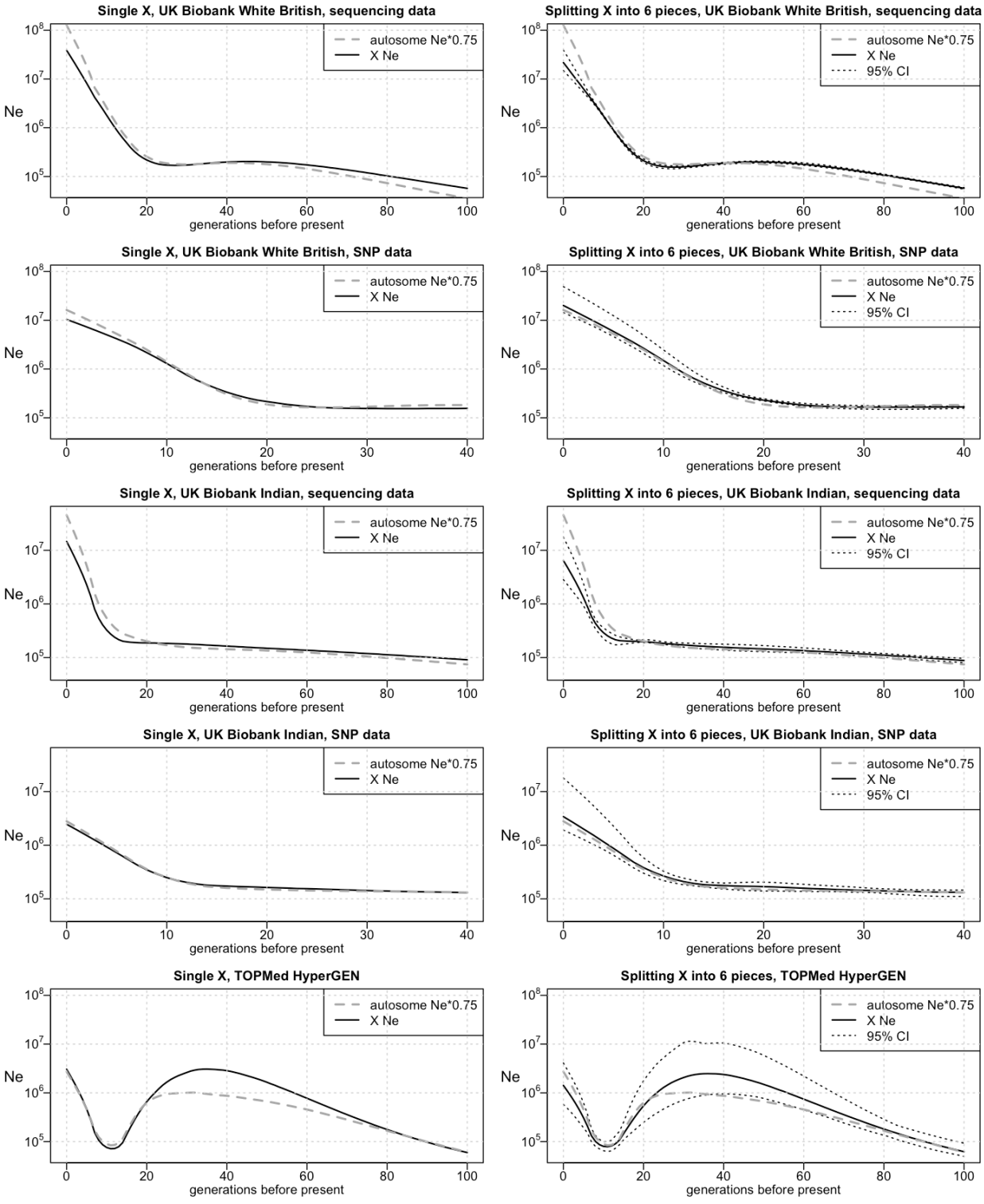

X break to 6 pieces, US-like M:F=8:2
simulated value
estimated Ne
95% CI
Ne
generations before present

Figure S3. Effective population size estimated from an undivided X chromosome versus splitting the X chromosome into six regions in analyses of human populations.
N_e estimated on a single undivided X chromosome is shown in the left column. N_e estimated by treating six separate regions of the X chromosome as six chromosomes to enable bootstrapping is shown in the right column. In each plot, the Y-axes are on a log scale. The first and second rows show results for the UK Biobank White British group either using sequence data or using SNP array data, followed by results for the UK Biobank Indian group using either sequence data or SNP array data in the third and the fourth rows. The last row shows results for the Black non-Hispanic group in the HyperGEN study.


Single X, UK Biobank White British, sequencing data
Splitting X into 6 pieces, UK Biobank White British, sequencing data
Single X, UK Biobank White British, SNP data
Splitting X into 6 pieces, UK Biobank White British, SNP data
Single X, UK Biobank Indian, sequencing data
Splitting X into 6 pieces, UK Biobank Indian, sequencing data
Single X, UK Biobank Indian, SNP data
Splitting X into 6 pieces, UK Biobank Indian, SNP data
Single X, TOPMed HyperGEN
Splitting X into 6 pieces, TOPMed HyperGEN
autosome Ne*0.75
X Ne
95% CI
Ne
generations before present

APPENDICES

Appendix A. Data and code availability

UK Biobank data were downloaded from the European Genome-Phenome Archive (https://ega-archive.org/datasets/EGAD00010001497), under Application Number 19934. Trans-Omics in Precision Medicine (TOPMed) freeze 8 data for study accession phs001293.v2.p1 (HyperGEN) were downloaded from dbGaP (https://www.ncbi.nlm.nih.gov/gap/).

The IBDNe software is available from https://faculty.washington.edu/browning/ibdne.html. The hap-ibd software is available from https://github.com/browning-lab/hap-ibd. The IBDkin software is available from https://github.com/YingZhou001/IBDkin. The ibd-cluster software is available from https://github.com/browning-lab/ibd-cluster. The FiMAP R package is available from https://github.com/hanchenphd/FiMAP. Our code and scripts for running FiMAP analysis on the UK Biobank data adopted examples from https://github.com/hanchenphd/FiMAP-code/. Our IBD mapping test is implemented in the Python package IBDmap, available from https://github.com/RC0515/IBDmap.

Appendix B. Funding acknowledgement

Research in this dissertation was supported by the National Human Genome Research Institute of the National Institutes of Health under award number HG007501.

We gratefully acknowledge the studies and participants who provided biological samples and data for UK Biobank and for TOPMed. Molecular data for the TOPMed program was supported by the National Heart, Lung and Blood Institute (NHLBI). Core support including centralized genomic read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I). Core support including phenotype harmonization, data management, sample-identity QC, and general program coordination were provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN268201800001I). The Hypertension Genetic Epidemiology Network (HyperGEN) Study is part of the NHLBI Family Blood Pressure Program; collection of the data represented here was supported by grants U01 HL054472, U01 HL054473, U01 HL054495, and U01 HL054509; genome sequencing was funded by R01HL055673.

BIBLIOGRAPHY

1. Wright S. Coefficients of inbreeding and relationship. *The American Naturalist*. 1922;56(645):330-338.
2. Cotterman CW. *A calculus for statistico-genetics*. The Ohio State University; 1940.
3. Malécot G. The mathematics of heredity. 1948;
4. Jacquard A. *The genetic structure of populations*. vol 5. Springer Science & Business Media; 2012.
5. Thompson EA. Two-locus and three-locus gene identity by descent in pedigrees. *Mathematical Medicine and Biology: A Journal of the IMA*. 1988;5(4):261-279.
6. Lynch M, Ritland K. Estimation of pairwise relatedness with molecular markers. *Genetics*. 1999;152(4):1753-1766.
7. Thompson EA. Identity by descent: variation in meiosis, across genomes, and in populations. *Genetics*. 2013;194(2):301-326.
8. Haldane J. The combination of linkage values and the calculation of distance between the loci of linked factors. *Journal of Genetics*. 1919;8(4):299-309.
9. Stephens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction from population data. *The American Journal of Human Genetics*. 2001;68(4):978-989.
10. Li N, Stephens M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*. 2003;165(4):2213-2233.
11. Stephens M, Donnelly P. A comparison of bayesian methods for haplotype reconstruction from population genotype data. *The American Journal of Human Genetics*. 2003;73(5):1162-1169.
12. Stephens M, Scheet P. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *The American Journal of Human Genetics*. 2005;76(3):449-462.
13. Scheet P, Stephens M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *The American Journal of Human Genetics*. 2006;78(4):629-644.
14. Browning SR, Browning BL. Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics*. 2011;12(10):703-714.
15. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics*. 2007;81(5):1084-1097.
16. Kong A, Masson G, Frigge ML, et al. Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature genetics*. 2008;40(9):1068-1075.
17. Browning SR, Browning BL. Identity by descent between distant relatives: detection and applications. *Annual review of genetics*. 2012;46:617-633.
18. Leutenegger A-L, Prum B, Génin E, et al. Estimation of the inbreeding coefficient through use of genomic data. *The American Journal of Human Genetics*. 2003;73(3):516-523.
19. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics*. 2007;81(3):559-575.
20. Browning SR. Estimation of pairwise identity by descent from dense genetic marker data in a population sample of haplotypes. *Genetics*. 2008;178(4):2123-2132.

21. Albrechtsen A, Sand Korneliussen T, Moltke I, van Overseem Hansen T, Nielsen FC, Nielsen R. Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*. 2009;33(3):266-274.
22. Browning SR, Browning BL. High-resolution detection of identity by descent in unrelated individuals. *The American Journal of Human Genetics*. 2010;86(4):526-539.
23. Browning BL, Browning SR. Detecting identity by descent and estimating genotype error rates in sequence data. *The American Journal of Human Genetics*. 2013;93(5):840-851.
24. Gusev A, Lowe JK, Stoffel M, et al. Whole population, genome-wide mapping of hidden relatedness. *Genome research*. 2009;19(2):318-326.
25. Browning BL, Browning SR. A fast, powerful method for detecting identity by descent. *The American Journal of Human Genetics*. 2011;88(2):173-182.
26. Browning BL, Browning SR. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics*. 2013;194(2):459-471.
27. Bjelland DW, Lingala U, Patel PS, Jones M, Keller MC. A fast and accurate method for detection of IBD shared haplotypes in genome-wide SNP data. *European Journal of Human Genetics*. 2017;25(5):617-624.
28. Nait Saada J, Kalantzis G, Shyr D, et al. Identity-by-descent detection across 487,409 British samples reveals fine scale population structure and ultra-rare variant associations. *Nature communications*. 2020;11(1):6130.
29. Shemirani R, Belbin GM, Avery CL, Kenny EE, Gignoux CR, Ambite JL. Rapid detection of identity-by-descent tracts for mega-scale datasets. *Nature communications*. 2021;12(1):3546.
30. Durbin R. Efficient haplotype matching and storage using the positional Burrows–Wheeler transform (PBWT). *Bioinformatics*. 2014;30(9):1266-1272.
31. Naseri A, Liu X, Tang K, Zhang S, Zhi D. RaPID: ultra-fast, powerful, and accurate detection of segments identical by descent (IBD) in biobank-scale cohorts. *Genome biology*. 2019;20:1-15.
32. Zhou Y, Browning SR, Browning BL. A fast and simple method for detecting identity-by-descent segments in large-scale data. *The American Journal of Human Genetics*. 2020;106(4):426-437.
33. Freyman WA, McManus KF, Shringarpure SS, et al. Fast and robust identity-by-descent inference with the templated positional burrows–wheeler transform. *Molecular Biology and Evolution*. 2021;38(5):2131-2151.
34. Moltke I, Albrechtsen A, vO Hansen T, Nielsen FC, Nielsen R. A method for detecting IBD regions simultaneously in multiple individuals—with applications to disease genetics. *Genome research*. 2011;21(7):1168-1180.
35. Gusev A, Kenny EE, Lowe JK, et al. DASH: a method for identical-by-descent haplotype mapping uncovers association with recent variation. *The American Journal of Human Genetics*. 2011;88(6):706-717.
36. He D. IBD-Groupon: an efficient method for detecting group-wise identity-by-descent regions simultaneously in multiple individuals based on pairwise IBD relationships. *Bioinformatics*. 2013;29(13):i162-i170.
37. Qian Y, Browning BL, Browning SR. Efficient clustering of identity-by-descent between multiple individuals. *Bioinformatics*. 2014;30(7):915-922.

38. Browning SR, Browning BL. Biobank-scale inference of multi-individual identity by descent and gene conversion. *The American Journal of Human Genetics*. 2024;111(4):691-700.
39. Ramstetter MD, Dyer TD, Lehman DM, et al. Benchmarking relatedness inference methods with genome-wide data from thousands of relatives. *Genetics*. 2017;207(1):75-82.
40. Zhou Y, Browning SR, Browning BL. IBDkin: fast estimation of kinship coefficients from identity by descent segments. *Bioinformatics*. 2020;36(16):4519-4520.
41. Palamara PF, Lencz T, Darvasi A, Pe'er I. Length distributions of identity by descent reveal fine-scale demographic history. *The American Journal of Human Genetics*. 2012;91(5):809-822.
42. Browning SR, Browning BL. Accurate non-parametric estimation of recent effective population size from segments of identity by descent. *The American Journal of Human Genetics*. 2015;97(3):404-418.
43. Browning SR, Browning BL, Daviglus ML, et al. Ancestry-specific recent effective population size in the Americas. *PLoS genetics*. 2018;14(5):e1007385.
44. Palamara PF, Francioli LC, Wilton PR, et al. Leveraging distant relatedness to quantify human mutation and gene-conversion rates. *The American Journal of Human Genetics*. 2015;97(6):775-789.
45. Tian X, Browning BL, Browning SR. Estimating the genome-wide mutation rate with three-way identity by descent. *The American Journal of Human Genetics*. 2019;105(5):883-893.
46. Tian X, Cai R, Browning SR. Estimating the genome-wide mutation rate from thousands of unrelated individuals. *The American Journal of Human Genetics*. 2022;109(12):2178-2184.
47. Zhou Y, Browning BL, Browning SR. Population-specific recombination maps from segments of identity by descent. *The American Journal of Human Genetics*. 2020;107(1):137-148.
48. Albrechtsen A, Moltke I, Nielsen R. Natural selection and the distribution of identity-by-descent in the human genome. *Genetics*. 2010;186(1):295-308.
49. Han L, Abney M. Using identity by descent estimation with dense genotype data to detect positive selection. *European Journal of Human Genetics*. 2013;21(2):205-211.
50. Palamara PF, Terhorst J, Song YS, Price AL. High-throughput inference of pairwise coalescence times identifies signals of selection and enriched disease heritability. *Nature Genetics*. 2018;50(9):1311-1317.
51. Browning SR, Browning BL. Probabilistic estimation of identity by descent segment endpoints and detection of recent selection. *The American Journal of Human Genetics*. 2020;107(5):895-910.
52. Browning SR, Thompson EA. Detecting rare variant associations by identity-by-descent mapping in case-control studies. *Genetics*. 2012;190(4):1521-1531.
53. Chen H, Naseri A, Zhi D. FiMAP: A fast identity-by-descent mapping test for Biobank-scale cohorts. *Plos Genetics*. 2023;19(12):e1011057.
54. Wright S. Evolution in Mendelian populations. *Genetics*. 1931;16(2):97.
55. Ardlie KG, Kruglyak L, Seielstad M. Patterns of linkage disequilibrium in the human genome. *Nature Reviews Genetics*. 2002;3(4):299-309.
56. Pluzhnikov A, Di Rienzo A, Hudson RR. Inferences about human demography based on multilocus analyses of noncoding sequences. *Genetics*. 2002;161(3):1209-1218.
57. Chapman NH, Thompson EA. The effect of population history on the lengths of ancestral chromosome segments. *Genetics*. 2002;162(1):449-458.

58. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS genetics*. 2009;5(10):e1000695.
59. Li H, Durbin R. Inference of human population history from individual whole-genome sequences. *Nature*. 2011;475(7357):493-496.
60. Schiffels S, Durbin R. Inferring human population size and separation history from multiple genome sequences. *Nature genetics*. 2014;46(8):919-925.
61. Hayes BJ, Visscher PM, McPartlan HC, Goddard ME. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome research*. 2003;13(4):635-643.
62. Marth GT, Czabarka E, Murvai J, Sherry ST. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics*. 2004;166(1):351-372.
63. Gravel S, Henn BM, Gutenkunst RN, et al. Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences*. 2011;108(29):11983-11988.
64. Gattepaille LM, Jakobsson M, Blum MG. Inferring population size changes with sequence and SNP data: lessons from human bottlenecks. *Heredity*. 2013;110(5):409-419.
65. Goldgar DE. Multipoint analysis of human quantitative genetic variation. *American journal of human genetics*. 1990;47(6):957.
66. Schork NJ. Extended multipoint identity-by-descent analysis of human quantitative traits: efficiency, power, and modeling considerations. *American journal of human genetics*. 1993;53(6):1306.
67. Amos CI. Robust variance-components approach for assessing genetic linkage in pedigrees. *American journal of human genetics*. 1994;54(3):535.
68. Xu S, Atchley WR. A random model approach to interval mapping of quantitative trait loci. *Genetics*. 1995;141(3):1189-1197.
69. Blangero J, Almasy L. Multipoint oligogenic linkage analysis of quantitative traits. *Genetic epidemiology*. 1997;14(6):959-964.
70. Almasy L, Blangero J. Multipoint quantitative-trait linkage analysis in general pedigrees. *The American Journal of Human Genetics*. 1998;62(5):1198-1211.
71. Blangero J, Williams JT, Almasy L. Quantitative trait locus mapping using human pedigrees. *Human Biology*. 2000:35-62.
72. Day‐Williams AG, Blangero J, Dyer TD, Lange K, Sobel EM. Linkage analysis without defined pedigrees. *Genetic epidemiology*. 2011;35(5):360-370.
73. Glazner C, Thompson E. Pedigree-free descent-based gene mapping from population samples. *Human heredity*. 2015;80(1):21-35.
74. Charlesworth B. Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*. 2009;10(3):195-205.
75. Henden L, Wakeham D, Bahlo M. XIBD: software for inferring pairwise identity by descent on the X chromosome. *Bioinformatics*. 2016;32(15):2389-2391.
76. Buffalo V, Mount SM, Coop G. A genealogical look at shared ancestry on the X chromosome. *Genetics*. 2016;204(1):57-75.
77. Browning SR, Browning BL. Accurate non-parametric estimation of recent effective population size from segments of identity by descent. *Am J Hum Genet*. Sep 3 2015;97(3):404-418. doi:10.1016/j.ajhg.2015.07.012

78. Browning SR, Browning BL, Daviglus ML, et al. Ancestry-specific recent effective population size in the Americas. *PLoS Genet*. 2018;14:e1007385.
79. Ségurel L, Martínez-Cruz B, Quintana-Murci L, et al. Sex-specific genetic structure and social organization in Central Asia: insights from a multi-locus study. *PLoS Genet*. Sep 26 2008;4(9):e1000200. doi:10.1371/journal.pgen.1000200
80. Bryc K, Auton A, Nelson MR, et al. Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proceedings of the National Academy of Sciences*. 2010;107(2):786-791.
81. Heyer E, Chaix R, Pavard S, Austerlitz F. Sex-specific demographic behaviours that shape human genomic variation. *Mol Ecol*. Feb 2012;21(3):597-612. doi:10.1111/j.1365-294X.2011.05406.x
82. Goldberg A, Rosenberg NA. Beyond 2/3 and 1/3: the complex signatures of sex-biased admixture on the X chromosome. *Genetics*. 2015;201(1):263-279.
83. Shringarpure SS, Bustamante CD, Lange K, Alexander DH. Efficient analysis of large datasets and sex bias with ADMIXTURE. *BMC bioinformatics*. 2016;17(1):1-6.
84. Hammer MF, Mendez FL, Cox MP, Woerner AE, Wall JD. Sex-biased evolutionary forces shape genomic patterns of human diversity. *PLoS genetics*. 2008;4(9):e1000202.
85. Bustamante CD, Ramachandran S. Evaluating signatures of sex-specific processes in the human genome. *Nature genetics*. 2009;41(1):8-10.
86. Clemente F, Gautier M, Vitalis R. Inferring sex-specific demographic history from SNP data. *PLoS genetics*. 2018;14(1):e1007191.
87. Musharoff S, Shringarpure S, Bustamante CD, Ramachandran S. The inference of sex-biased human demography from whole-genome data. *PLoS genetics*. 2019;15(9):e1008293.
88. Wang J, Santiago E, Caballero A. Prediction and estimation of effective population size. *Heredity*. 2016;117(4):193-206.
89. Emery LS, Felsenstein J, Akey JM. Estimators of the human effective sex ratio detect sex biases on different timescales. *The American Journal of Human Genetics*. 2010;87(6):848-856.
90. Rosenberg NA. Admixture models and the breeding systems of HS Jennings: A GENETICS Connection. *Genetics*. 2016;202(1):9-13.
91. File Formats Task Team. The Variant Call Format Specication: VCFv4.3 and BCFv2.2. 2020. http://samtools.github.io/hts-specs/VCFv4.3.pdf
92. International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature*. 2007;449(7164):851.
93. Halldorsson BV, Palsson G, Stefansson OA, et al. Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science*. 2019;363(6425):eaau1043.
94. Wright S. *Evolution and the genetics of populations: Vol. 2. The theory of gene frequencies*. 1969.
95. Hartl DL, Clark AG. *Principles of population genetics*. vol 116. Sinauer associates Sunderland; 1997.
96. Haller BC, Messer PW. SLiM 3: forward genetic simulations beyond the Wright-Fisher model. *Molecular biology and evolution*. 2019;36(3):632-637.
97. Kelleher J, Etheridge AM, McVean G. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS computational biology*. 2016;12(5):e1004842.
98. Haller BC, Galloway J, Kelleher J, Messer PW, Ralph PL. Tree-sequence recording in SLiM opens new horizons for forward-time simulation of whole genomes. *Molecular ecology resources*. 2019;19(2):552-566.

99. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*. 2018;562(7726):203-209.
100. Taliun D, Harris DN, Kessler MD, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*. 2021;590(7845):290-299.
101. Fry A, Littlejohns TJ, Sudlow C, et al. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *American journal of epidemiology*. 2017;186(9):1026-1034.
102. Halldorsson BV, Eggertsson HP, Moore KH, et al. The sequences of 150,119 genomes in the UK Biobank. *Nature*. 2022;607(7920):732-740.
103. Browning BL, Tian X, Zhou Y, Browning SR. Fast two-stage phasing of large-scale sequence data. *The American Journal of Human Genetics*. 2021;108(10):1880-1890.
104. Browning BL, Browning SR. Statistical phasing of 150,119 sequenced genomes in the UK Biobank. *The American Journal of Human Genetics*. 2023;110(1):161-165.
105. Bhérer C, Campbell CL, Auton A. Refined genetic maps reveal sexual dimorphism in human meiotic recombination at multiple scales. *Nature communications*. 2017;8(1):1-9.
106. Schaffner SF. The X chromosome in population genetics. *Nature Reviews Genetics*. 2004;5(1):43-51.
107. Ramachandran S, Rosenberg NA, Zhivotovsky LA, Feldman MW. Robustness of the inference of human population structure: a comparison of X-chromosomal and autosomal microsatellites. *Human genomics*. 2004;1(2):1-11.
108. Pinto N, Gusmão L, Amorim A. X-chromosome markers in kinship testing: a generalisation of the IBD approach identifying situations where their contribution is crucial. *Forensic Science International: Genetics*. 2011;5(1):27-32.
109. Keinan A, Mullikin JC, Patterson N, Reich D. Accelerated genetic drift on chromosome X during the human dispersal out of Africa. *Nature genetics*. 2009;41(1):66-70.
110. Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. *Nature reviews genetics*. 2005;6(2):95-108.
111. McCarthy MI, Hirschhorn JN. Genome-wide association studies: potential next steps on a genetic journey. *Human molecular genetics*. 2008;17(R2):R156-R165.
112. Visscher PM, Wray NR, Zhang Q, et al. 10 years of GWAS discovery: biology, function, and translation. *The American Journal of Human Genetics*. 2017;101(1):5-22.
113. McCarthy MI, Abecasis GR, Cardon LR, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature reviews genetics*. 2008;9(5):356-369.
114. Eichler EE, Flint J, Gibson G, et al. Missing heritability and strategies for finding the underlying causes of complex disease. *Nature reviews genetics*. 2010;11(6):446-450.
115. Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*. 2019;20(8):467-484.
116. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS genetics*. 2009;5(2):e1000384.
117. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*. 2011;89(1):82-93.
118. Lee S, Emond MJ, Bamshad MJ, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *The American Journal of Human Genetics*. 2012;91(2):224-237.

119. Liu Y, Chen S, Li Z, Morrison AC, Boerwinkle E, Lin X. ACAT: a fast and powerful p value combination method for rare-variant analysis in sequencing studies. *The American Journal of Human Genetics*. 2019;104(3):410-421.
120. Vacic V, Ozelius LJ, Clark LN, et al. Genome-wide mapping of IBD segments in an Ashkenazi PD cohort identifies associated haplotypes. *Human molecular genetics*. 2014;23(17):4693-4702.
121. Westerlind H, Imrell K, Ramanujam R, et al. Identity-by-descent mapping in a Scandinavian multiple sclerosis cohort. *European Journal of Human Genetics*. 2015;23(5):688-692.
122. Belbin GM, Odgis J, Sorokin EP, et al. Genetic identification of a common collagen disease in puerto ricans via identity-by-descent mapping in a health system. *Elife*. 2017;6:e25060.
123. Hsueh W-C, Nair AK, Kobes S, et al. Identity-by-descent mapping identifies major locus for serum triglycerides in Amerindians largely explained by an APOC3 founder mutation. *Circulation: Cardiovascular Genetics*. 2017;10(6):e001809.
124. Henden L, Twine NA, Szul P, et al. Identity by descent analysis identifies founder events and links SOD1 familial and sporadic ALS cases. *NPJ genomic medicine*. 2020;5(1):32.
125. Chen H, Naseri A, Zhi D. FiMAP: A fast identity-by-descent mapping test for Biobank-scale cohorts. *medRxiv*. 2021:2021.06. 30.21259773.
126. Almasy L, Blangero J. Variance component methods for analysis of complex phenotypes. *Cold Spring Harbor Protocols*. 2010;2010(5):pdb. top77.
127. Page GP, Amos CI, Boerwinkle E. The quantitative LOD score: test statistic and sample size for exclusion and linkage of quantitative traits in human sibships. *The American Journal of Human Genetics*. 1998;62(4):962-968.
128. Chernoff H. On the distribution of the likelihood ratio. *The Annals of Mathematical Statistics*. 1954:573-578.
129. Miller JJ. Asymptotic properties of maximum likelihood estimates in the mixed model of the analysis of variance. *The Annals of Statistics*. 1977:746-762.
130. Self SG, Liang K-Y. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*. 1987;82(398):605-610.
131. Broyden CG. The convergence of a class of double-rank minimization algorithms: 2. The new algorithm. *IMA journal of applied mathematics*. 1970;6(3):222-231.
132. Fletcher R. A new approach to variable metric algorithms. *The computer journal*. 1970;13(3):317-322.
133. Goldfarb D. A family of variable metric updates derived by variational means, v. 24. *Mathematics of Computation*. 1970:21-55.
134. Shanno DF. Conditioning of quasi-Newton methods for function minimization. *Mathematics of computation*. 1970;24(111):647-656.
135. Byrd RH, Lu P, Nocedal J, Zhu C. A limited memory algorithm for bound constrained optimization. *SIAM Journal on scientific computing*. 1995;16(5):1190-1208.
136. Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods*. 2020;17(3):261-272.
137. Shor T, Kalka I, Geiger D, Erlich Y, Weissbrod O. Estimating variance components in population scale family trees. *PLoS Genetics*. 2019;15(5):e1008124.
138. Davis TA. scikit-sparse: Sparse matrix tools extending scipy.sparse. 2021.

139. Kelleher J, Lohse K. Coalescent simulation with msprime. *Statistical Population Genomics*. 2020;986:191-230.
140. Baumdicker F, Bisschop G, Goldstein D, et al. Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*. 2022;220(3):iyab229.
141. Lander ES, Botstein D. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*. 1989;121(1):185-199.
142. Siegmund D, Yakir B. *The statistics of gene mapping*. vol 1. Springer; 2007.
143. Grinde KE, Brown LA, Reiner AP, Thornton TA, Browning SR. Genome-wide significance thresholds for admixture mapping studies. *The American Journal of Human Genetics*. 2019;104(3):454-465.
144. Wolfram Research I. Mathematica. Version 14.1 ed. Champaign, Illinois: Wolfram Research, Inc.; 2024.
145. Browning BL. PRESTO: rapid calculation of order statistic distributions and multiple-testing adjusted P-values via permutation for one and two-stage genetic association studies. *BMC bioinformatics*. 2008;9:1-5.
146. Marees AT, De Kluiver H, Stringer S, et al. A tutorial on conducting genome‐wide association studies: Quality control and statistical analysis. *International journal of methods in psychiatric research*. 2018;27(2):e1608.
147. Warren HR, Evangelou E, Cabrera CP, et al. Genome-wide association analysis identifies novel blood pressure loci and offers biological insights into cardiovascular risk. *Nature genetics*. 2017;49(3):403-415.
148. Hoffmann TJ, Ehret GB, Nandakumar P, et al. Genome-wide association analyses using electronic health records identify new loci influencing blood pressure variation. *Nature genetics*. 2017;49(1):54-64.
149. Giri A, Hellwege JN, Keaton JM, et al. Trans-ethnic association study of blood pressure determinants in over 750,000 individuals. *Nature genetics*. 2019;51(1):51-62.
150. Sakaue S, Kanai M, Tanigawa Y, et al. A cross-population atlas of genetic associations for 220 human phenotypes. *Nature genetics*. 2021;53(10):1415-1424.
151. Plotnikov D, Huang Y, Khawaja AP, et al. High blood pressure and intraocular pressure: a Mendelian randomization study. *Investigative Ophthalmology & Visual Science*. 2022;63(6):29-29.
152. Zhu X, Zhu L, Wang H, Cooper RS, Chakravarti A. Genome‐wide pleiotropy analysis identifies novel blood pressure variants and improves its polygenic risk scores. *Genetic epidemiology*. 2022;46(2):105-121.
153. Koskeridis F, Fancy N, Tan PF, et al. Multi-trait association analysis reveals shared genetic loci between Alzheimer's disease and cardiovascular traits. *Nature Communications*. 2024;15(1):9827.
154. Browning SR. Multilocus association mapping using variable-length Markov chains. *The American Journal of Human Genetics*. 2006;78(6):903-913.
155. Browning BL, Browning SR. Efficient multilocus association testing for whole genome association studies using localized haplotype clustering. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*. 2007;31(5):365-375.