Statistical Methods to Estimate Evolutionary and Technical Parameters Using Whole Genome Sequence Data

Nobuaki Masaki

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
Trevor Bedford
Brian L. Browning
Guanghao Qi

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Statistical Methods to Estimate Evolutionary and Technical Parameters Using Whole Genome Sequence Data

Nobuaki Masaki

Chair of the Supervisory Committee:
Sharon R. Browning
Department of Biostatistics

Whole genome sequence data are widely used in humans and other species to reveal evolutionary patterns and recent demographic history. In this dissertation, we introduce three new statistical methods that can be used to estimate technical parameters such as genotype error rates, as well as parameters related to genome evolution and recent demographic history, using whole genome sequence data from humans and SARS-CoV-2. In our first method, we propose a model that calculates the likelihood of observed parent-offspring trio genotypes, adjusting for both genotype errors and uncalled deletions. We fit our model to SNVs in 77 White British trios identified in the UK Biobank whole genome sequence data, obtaining estimates for the genotype error and uncalled deletion rates in this dataset. In our second method, we formulate a model to estimate the mean length of gene conversion tracts. Our model uses a separate per-site allele conversion rate for each observed tract. We fit this model to gene conversion tracts detected from the UK Biobank whole autosome sequence data and infer the mean length of gene conversion tracts in humans. Finally, in our third method, we propose a hidden Markov model that accounts for mutations and genotype errors to detect recombinant SARS-CoV-2 sequences.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I would like to thank my research advisor Dr. Sharon Browning for teaching me how to conduct research and for guiding me throughout my Ph.D. Research was a much slower process at first, but I feel as though I have now gained a level of proficiency in the scientific process, thanks to her mentorship. Sharon trusted me to work autonomously on my projects, which helped shape me into an independent researcher.

I am grateful for Dr. Brian Browning for proposing and supervising the project that forms Chapter 2 of this dissertation. This project resulted in my first publication as a lead author, and Brian's patient guidance taught me how to write and revise a scientific manuscript.

I want to thank Dr. Trevor Bedford for introducing me to virus genomics and genomic epidemiology, for encouraging me to apply my training in statistical genetics to his research field, and for cultivating a healthy and vibrant research environment at his lab. Working at the Bedford Lab has made me confident in my decision to pursue research as a career path.

My Ph.D. journey, which started in the midst of the pandemic in 2020, was incredibly challenging. I want to thank my Ph.D. cohort for helping me through my classes, especially in the first and second years. I also want to thank Dr. Antonio Olivas, who tutored me during the challenging second year theory classes. I'm grateful to Dr. Tracey Marsh, who was my RA supervisor during my first year of Ph.D. Tracey introduced me to early detection of cancer research and helped me stay productive during this difficult time. I am thankful to my friends Sungtaek Son, James Buenfil, Derry Cheng, Andy Su, Carlos Avendano, Tyler Smith, Ko Fukushima, Tyler Chang, Pierce McDonnell, and Shengzhi Wang for their support. Finally, I'm grateful to all of my past and present lab colleagues–especially Seth Temple and Ruoyi Cai for their countless advice on research and academics–and to John Huddleston,

Cecile Tran Kiem, Kim Andrews, Jennifer Chang, Victor Lin, and Philippa Steinberg for making research engaging and motivating me to come to the lab every day.

DEDICATION

To my parents

Chapter 1

INTRODUCTION

Whole genome sequence data for humans and pathogens such as SARS-CoV-2 are now widely available. Furthermore, there are a variety of statistical tools already developed to infer evolutionary processes from these datasets. When inferring aspects of evolution in humans and SARS-CoV-2, it is important to consider fundamental differences in their genome lengths, generation and replication times, mutation rates, and recombination mechanisms.

The human genome is 3.2 billion bp long, far exceeding the 30 kb SARS-CoV-2 genome. Furthermore, generation time in humans is much longer (around 20-30 years). A study on the replication of SARS-CoV-2 in cell culture found that the growth curve of SARS-CoV-2 plateaued at around 14 hours post-infection [46], so the replication cycle of SARS-CoV-2 is likely less than 14 hours. Studies have also reported progeny viruses being released into the cell culture at 6 hours post-infection [15].

Furthermore, the rate of mutations per site per generation is much smaller in humans. Using TOPMed whole genome sequence data, Tian et al. estimated the genome-wide mutation rate for humans to be 1.24×10^{-8} per generation for SNVs [63]. In contrast, SARS-CoV-2 mutation rate estimates range from 1×10^{-6} to 2×10^{-6} mutations per nucleotide per replication cycle [40].

Finally, recombination processes differ in these two organisms. In humans, genetic recombination primarily occurs during meiosis and are initiated by double-strand breaks [25, 36]. In SARS-CoV-2, recombination happens when a single cell is co-infected with two viruses and RdRp-mediated template switching occurs during RNA synthesis [67].

These contrasting evolutionary processes result in different methods being applied to genomic data in humans and SARS-CoV-2. Because SARS-CoV-2 samples capture many

replication cycles over short periods of time, phylogenetic methods, including real-time tree building techniques, are often employed [24]. On the other hand, large-scale human whole genome sequence data usually spans a few generations at most, so methods using present-day sequences to infer evolutionary parameters or recent demographic events are frequently used. These methods include coalescent models that consider the lengths of identity-by-descent segments across pairs of individuals to estimate the historical effective population size [48].

Coalescent models are also used to infer aspects of SARS-CoV-2 evolution, but these methods must be applicable to sequences sampled at differing points in time. Because SARS-CoV-2 has a short replication cycle, we cannot assume that all sequences are from the same generation. The Bayesian Skyline model can accommodate sampling times by assigning a prior distribution to sampling times and effective population sizes between sampling times [17]. Bayesian methods are useful in this context because it allows the model to account for error associated with phylogenetic reconstruction when quantifying the uncertainty in other parameters such as the effective population size [17].

In this dissertation, we develop three statistical methods designed for whole genome sequence data in humans and SARS-CoV-2. In Chapter 2, we describe a method we can use to estimate genotype error and uncalled deletions from parent-offspring trio genotypes, derived from the UK Biobank whole genome sequence data. In Chapter 3, we describe a method for estimating the mean length of gene conversion tracts in humans, using past gene conversions detected from the UK Biobank whole autosome sequence data using identity-by-descent segments. Finally, in Chapter 4, we look at a method designed to detect recombinant SARS-CoV-2 genomes in a surveillance dataset of SARS-CoV-2 genomes. We describe the context for each project below.

1.1 Genotype errors and uncalled deletions

Genotype errors from high-throughput genotyping technologies can affect the conclusions reached by downstream statistical analysis. For example, genotype errors can reduce the

statistical power to detect loci associated with the trait of interest in both pedigree-based linkage studies and case-control genetic association studies [34, 22, 31]. Researchers have attempted to mitigate the effects of genotype errors by masking genotypes that are estimated to have a high probability of error prior to statistical analysis [27, 13]. Several methods have also integrated the possibility of genotype errors into linkage analysis to alleviate their effects on the conclusions reached [19, 16].

Given the effects that genotype errors can have on downstream statistical results, researchers may be interested in quantifying the frequency of genotype errors in a particular study. One common approach used to estimate genotype error rates relies on individuals with multiple genotyped samples [29]. However, discordant genotypes will not be observed at a site if all samples from an individual share the same genotype error. Thus, only a lower bound for the genotype error rate can be estimated with this approach.

Some genotype errors can be detected by finding Mendelian-inconsistent genotypes in parent-offspring trios. However, only a small percentage of genotype errors at biallelic markers are detectable from Mendelian inconsistencies [28], so probabilistic models have been used in conjunction with the observed number of Mendelian inconsistencies to estimate genotype error rates [32, 26], One limitation of these approaches is that an overall error rate is estimated, rather than error rates that are conditional on the true genotype.

More recent approaches have modeled error rates that depend on both the true genotype and the type of miscall (e.g. the rate at which a major homozygous genotype is miscalled as a heterozygous genotype). Wang et al. use apparent phase violations in three-generation pedigrees to estimate these rates [52], Browning and Browning use a likelihood-based method to estimate the genotype error rates conditional on the true genotype from parent-offspring trio genotypes [6].

1.2 Gene conversions

During meiosis, homologous chromosomes undergo genetic recombination resulting in the transfer of genetic material. Double strand breaks that occur during recombination are

resolved in two distinct ways. Crossovers result in a long tract of DNA (typically spanning millions of base pairs) being exchanged between homologous chromosomes. On the other hand, non-crossover gene conversions typically result in a non-reciprocal transfer of alleles within a short tract [4]. These gene conversion events are thought to most commonly occur via the synthesis-dependent strand annealing mechanism, where a double stranded break is repaired by the invasion of a protruding 3' end into the donor chromatid. Gene conversion events may also occur via the resolution of two Holliday junctions [44].

Gene conversions can be detected in humans by analyzing sequence data from pedigrees or sperm samples and identifying positions in which the allele of one homologous chromosome has been replaced by the other [4, 3, 35, 9]. The distance between these positions, where alleles are thought to have been converted by a gene conversion event, can be used to estimate the length of the gene conversion tract. Using SNP array and whole genome sequence data from 34 three-generation pedigrees, Williams et al. determined that tract lengths are in the order of 100-1,000 bp based on detected allele conversions [4]. Using three-generation pedigrees helps to distinguish between allele conversions and genotype errors. It can be difficult to distinguish between allele conversions and genotype errors when using two-generation pedigrees or sperm samples.

Williams et al. further identified apparent clusters of gene conversion tracts spanning 20-30 kb, which may have resulted from discontinuous gene conversion events occurring in close proximity during the same meiosis [4]. This phenomenon has previously been referred to as complex gene conversions. Complex gene conversions as long as 100 kb were also found by Halldorsson et al. [9]. Complex gene conversions could arise from mechanisms such as GC-biased repair across long stretches of DNA [4]. In this study, we will focus on individual gene conversion tracts where the length spanning the furthest allele converted markers does not exceed 1.5 kb.

Efforts have been made to model the length distribution of gene conversion tracts using detected gene conversion tracts in humans and other species [30, 45]. Recently, Palsson et al. detected 12,948 paternal and 15,712 maternal gene conversions transmitted to 5,420 trios

in 2,132 Icelandic families [21]. Using their model, they estimated the mean length of gene conversion tracts to be 123 bp (95% CI: [94, 135]) and 102 bp (95% CI: [71, 125]) for paternal and maternal transmissions respectively [45, 21].

Palsson et al. also found that the frequency of observed gene conversions was much higher in crossover recombination hotspots (22.4-fold and 13.7-fold for paternal and maternal transmissions respectively) [21]. While the relative frequencies of gene conversions in hotspots and non-hotspot regions have been characterized, differences in the length distribution of gene conversion tracts between these regions have not been studied in great detail.

1.3 Recombination in SARS-CoV-2

Recombination is thought to occur in coronaviruses via a copy-choice mechanism in which the viral RNA-dependent RNA polymerase switches template strands during negative strand synthesis [14]. Co-infections involving multiple SARS-CoV-2 strains can result in recombinant viral genomes.

This has been verified by genomic surveillance. Trémeaux et al. ran real-time whole genome sequencing of SARS-CoV-2 sequences in 6,829 samples collected from 6,411 individuals in 2022, and found minor recombinant haplotypes in four out of 17 individuals co-infected by SARS-CoV-2 sequences belonging to two lineages. The recombinant viruses detected in the four individuals were all mosaics between Omicron lineages (BA.2-BA.4, BA.2.12.1–BA.5, and BA.1.1–BA.2) [64].

One of the most significant recombinant lineages that emerged during the pandemic is XBB, thought to be derived from a recombination event between two Omicron lineages (BJ.1 and BM.1.1.1) using phylogenetic analysis [60]. Substitutions in the XBB Spike protein obtained from both BJ.1 and BM.1.1.1 were found to jointly confer resistance against immunity induced by infection by previous lineages and vaccination. Furthermore, the effective reproduction number (R_e) of XBB was estimated to be 1.23 and 1.20 times higher than its parental lineages BJ.1 and BM.1.1.1, respectively, using epidemic data from late 2022 [60]. XBB.1.5 reached a peak prevalence of 55% globally in epidemiological week 12 of 2023 [1]. In the

United States, XBB.1.5 reached a peak prevalence of 84.1% by April 1, 2023 [61]. Because recombination can combine mutations in different SARS-CoV-2 lineages that jointly confer a growth advantage, systematic surveillance and robust statistical detection of recombinant lineages are crucial.

Chapter 2

SIMULTANEOUS ESTIMATION OF GENOTYPE ERROR AND UNCALLED DELETION RATES

This chapter is adapted, with minor modifications, from Masaki et al. (2024) [43].

2.1 Introduction

No method to our knowledge has adjusted for the effect of uncalled deletions when estimating genotype error rates. An inherited, uncalled deletion creates two genotype errors in a parent-offspring trio. A genotype in a parent with an uncalled deletion will likely be miscalled as homozygous for the non-deleted allele, and the same type of error will be seen in the offspring if the deletion is inherited.

In our study, we model uncalled deletions using an uncalled deletion rate, defined as one half times the probability of an uncalled deletion being present in a randomly chosen genotype in our sample. We develop a model that takes both uncalled deletions and genotype errors into account when calculating the probability of observing a parent-offspring trio genotype, and we use maximum likelihood to simultaneously estimate the uncalled deletion rate and genotype error rates based on observed trio genotypes. Genotype error rates can depend on a marker's MAF, so we count observed trio genotypes within predefined MAF windows and fit a separate model for each MAF window.

We ran two simulation studies to assess model performance. We show that our method results in less biased estimators of genotype error rates if uncalled deletions are present, compared to a model that does not account for uncalled deletions. We also show that our method can accurately estimate the overall genotype error rate and the proportion of genotype errors attributable to uncalled deletions for markers with MAF > 0.001.

Finally, we fit our model to SNV variants in UK Biobank whole genome sequence data for 77 White British trios to obtain estimates and bootstrap confidence intervals for the uncalled deletion and genotype error rates in this dataset. The overall genotype error rate and uncalled deletion rate for SNVs with MAF > 0.001 were estimated to be 3.2×10^{-4} (90% CI: $[2.8 \times 10^{-4}, 6.2 \times 10^{-4}]$) and 1.2×10^{-4} (90% CI: $[1.0 \times 10^{-4}, 2.7 \times 10^{-4}]$) respectively. We further estimate that 77% (90% CI: [73%, 88%]) of genotype errors at these markers are attributable to uncalled deletions. Using the Akaike information criterion (AIC) [23], we show that our model fits this dataset better than a model that does not take uncalled deletions into account.

2.2 UK Biobank sequence data

The UK Biobank release of 200,031 sequenced genomes [11] includes 77 parent–offspring trios whose members have been classified as White British by the UK Biobank [12]. We restricted the markers to SNVs having "PASS" in the VCF Filter field, fewer than 5% missing genotypes, and AAScore > 0.95, and we excluded any SNVs overlapping with a called deletion or structural variant in any of the 200,031 sequenced genomes [11, 7]. After marker filtering, there were 441,608,608 autosomal SNVs.

2.3 Parental and trio genotypes

If a marker has multiple observed alternate alleles, we combine them into a single alternate allele. We label the major and minor alleles as A and B respectively. Because we restrict our analysis to markers that do not overlap a called deletion, an observed allele is never a deletion. Major allele homozygous, heterozygous, and minor allele homozygous genotypes are denoted AA, AB, and BB respectively. In equations, the AA, AB, and BB genotypes are denoted using the minor allele dose (0, 1, and 2 respectively).

At each marker, we define a trio genotype to be the three genotypes of a parent-offspring trio. We consider the genotypes of the parents to be interchangeable, meaning that two trio genotypes will be considered (and labeled) the same if the father and mother's genotypes

are swapped. Parental genotypes are listed in alphabetical order. For example, AA-AB is a parental genotype in which the two parents have the AA and AB genotypes. Trio genotypes are written as a triplet of genotypes: the two parents' genotypes are listed first in alphabetical order followed by the offspring's genotype. As an example, AA-AB-AB represents the trio genotype for which the two parents have the AA and AB genotypes, and the offspring the AB genotype. Table 2.1 shows the observed trio genotype counts from the 77 UK Biobank White British trios (UKB-WB) [12].

Trio genotype	Integer repr.	UKB–WB	Inconsistent
AA-AA-AA	0,0,0	33,754,589,673	
AA-AA-AB	0,0,1	37,896	✗
AA-AA-BB	0,0,2	195	✗
AA-AB-AA	0,1,0	74,717,572	
AA-AB-AB	0,1,1	74,518,651	
AA-AB-BB	0,1,2	34,644	✗
AA-BB-AA	0,2,0	37,305	✗
AA-BB-AB	0,2,1	23,288,271	
AA-BB-BB	0,2,2	15,943	✗
AB-AB-AA	1,1,0	11,597,232	
AB-AB-AB	1,1,1	23,436,236	
AB-AB-BB	1,1,2	11,648,545	
AB-BB-AA	1,2,0	12,778	✗
AB-BB-AB	1,2,1	12,085,809	
AB-BB-BB	1,2,2	12,100,952	
BB-BB-AA	2,2,0	74	✗
BB-BB-AB	2,2,1	1,985	✗
BB-BB-BB	2,2,2	3,927,557	
Missing	N/A	1,785,549	

Table 2.1: Autosomal trio genotype counts for 77 UK Biobank White British trios (UKB-WB). Trio genotype counts were from 441,608,608 SNV variants that do not overlap a called deletion. Major allele homozygous, heterozygous, and minor allele homozygous genotypes are denoted AA, AB, and BB respectively. In equations, these genotypes are represented using by their minor allele dose (0, 1, or 2). The last column indicates whether the trio genotype is Mendelian inconsistent. A trio genotype is considered missing if any member of the trio has a missing genotype.

2.4 Minor allele frequency intervals

We fit our model to the observed trio genotypes from markers whose MAFs are within a specified MAF interval, and we fit a separate model for each interval. This allows model parameters (the genotype error and uncalled deletion rates) to depend on the MAF interval. The 101 MAF intervals are $(0, 0.001]$, $(0.001, 0.005]$, $(0.005, 0.01]$, $(0.01, 0.015]$, ..., $(0.495, 0.5]$. The first two intervals have a length of 0.001 and 0.004, and the remaining intervals have a length of 0.005.

2.5 Overview of the method

Our method assumes that the observed trio genotypes within a MAF interval $I \subseteq [0, 0.5)$ are generated under a probabilistic model that incorporates uncalled deletions and miscalled genotypes.

In our model, we start with parental genotypes within MAF interval I that have no deleted alleles, which we refer to as the pre-deletion parental genotypes. Deletions are added to a random subset of these genotypes, the frequency of which is determined by an uncalled deletion rate. The genotype of the offspring in each trio is then sampled assuming transmission equilibrium, conditional on the parental genotype.

We assume the trio genotypes in our final sample are drawn proportionally to their true frequencies and that genotype errors occur independently on each genotype during genotyping. Each type of genotype error, defined by the true and observed genotypes, occurs at a fixed rate.

Based on the model described above, we calculate the likelihood of seeing our sample of observed trio genotypes for MAF interval I as a function of the uncalled deletion rate and genotype error rates. We use maximum likelihood estimation to obtain an estimate of these parameters. In the following sections, we describe each component of our model in detail.

2.6 Estimating pre-deletion parental genotype frequencies

Pre-deletion parental genotypes are parental genotypes within MAF interval I that precede the inclusion of uncalled deletions. There are six possible pre-deletion parental genotypes (AA-AA, AA-AB, AA-BB, AB-AB, AB-BB, BB-BB).

We assume that the frequencies of these parental genotypes (e.g., the frequency of the AA-AB parental genotype) are the same for all markers whose MAFs are within I. Because the MAF intervals are narrow, this assumption should be reasonably accurate if the markers are in Hardy-Weinberg equilibrium. We denote the pre-deletion parental genotype frequency for parental genotypes i, j in MAF interval I as $\Pi^{i,j}_{\text{pre},I}$. For example, $\Pi^{0,1}_{\text{pre},I}$ denotes the pre-deletion parental genotype frequency of AA-AB in MAF interval I. Because these frequencies are unknown, we estimate $\Pi^{i,j}_{\text{pre},I}$ as the observed proportion of parental genotypes i, j in MAF interval I after excluding Mendelian-inconsistent trios. These estimates are denoted $\hat{\Pi}^{i,j}_{\text{pre},I}$.

2.7 Genotypes with uncalled deletions

Deletions are not present in the observed trio genotypes because we excluded SNVs that overlap with a called deletion. However, we consider the possibility that deleted alleles are present at some markers but are miscalled as the major or minor allele. We label genotypes with a deleted allele as AD or BD, depending on whether the remaining allele is the major or minor allele respectively. We expect deleted alleles to have low frequency at a SNV marker if the called genotypes contain no deleted alleles. Consequently, we do not account for genotypes in which both alleles are deleted because this should rarely occur. In equations, the AD and BD genotypes are denoted using the integers 3 and 4 respectively (recall that the AA, AB, and BB genotypes are denoted using 0, 1, and 2 respectively).

2.8 Estimating true trio genotype frequencies

In our model, the proportion of genotypes with a deleted allele in MAF interval I depends on the uncalled deletion rate Γ_I. Each genotype has probability $2\Gamma_I$ of containing a deleted

allele, meaning one of the pre-deletion genotype's alleles (chosen with equal probability) is deleted. This process models the inclusion of uncalled deletions in our sample of trio genotypes.

We denote the parental genotype frequencies after adjusting for uncalled deletions in each MAF interval as $\Pi_I^{i,j}$. Using Γ_I together with our previous estimates of the pre-deletion parental genotype frequencies, $\hat{\Pi}_{\text{pre},I}^{i,j}$, we obtain adjusted estimates of the parental genotype frequencies in each MAF interval, denoted $\hat{\Pi}_I^{i,j}$ (see Appendix A). We then use the estimated parental genotype frequencies, $\hat{\Pi}_I^{i,j}$, and transmission equilibrium to obtain an estimate of the true trio genotype frequencies $\hat{\Pi}_I^{i,j,k}$, where k is the offspring genotype. We will sometimes use the notation $\hat{\Pi}_I^{i,j,k}(\Gamma_I)$ to emphasize the dependency of $\hat{\Pi}_I^{i,j,k}$ on Γ_I. Note that the parental genotypes i, j in $\hat{\Pi}_I^{i,j}$ include deleted alleles (i.e., AD and BD). For example, $\hat{\Pi}_I^{1,3}$ denotes the estimated frequency of the AB-AD parental genotype.

2.9 Modeling genotype errors

In our model, an observed trio genotype is obtained by sampling a trio genotype proportionally to the true trio genotype frequencies $\Pi_I^{i,j,k}$, and then introducing genotype errors according to an error model. Let $\Theta_I^{x,y}$ denote the probability that the true genotype $x \in \{0,1,2,3,4\}$ is called as the genotype $y \in \{0,1,2\}$ at any marker whose MAF is in interval I. Genotypes cannot be called as AD or BD because the observed data are SNVs that do not overlap called deletions. We refer to all $\Theta_I^{x,y}$ for which $x \neq y$ as genotype error rates. The true trio genotype frequencies $\Pi_I^{i,j,k}$ and the genotype error rate parameters $\Theta_I^{x,y}$ are used to derive the probability of observing a specific trio genotype in MAF interval I. We replace each true trio genotype frequency $\Pi_I^{i,j,k}$ with its estimate $\hat{\Pi}_I^{i,j,k}$, because we do not observe $\Pi_I^{i,j,k}$ directly.

The rates at which genotypes are called correctly ($\Theta_I^{0,0}$, $\Theta_I^{1,1}$, and $\Theta_I^{2,2}$), or as homozygous for the non-deleted allele when the true genotype carries a deleted allele ($\Theta_I^{3,0}$ and $\Theta_I^{4,2}$), can be expressed in terms of the other ten genotype error rates:

$$\Theta_I^{0,0} = 1 - \Theta_I^{0,1} - \Theta_I^{0,2},$$
$$\Theta_I^{1,1} = 1 - \Theta_I^{1,0} - \Theta_I^{1,2},$$
$$\Theta_I^{2,2} = 1 - \Theta_I^{2,0} - \Theta_I^{2,1},$$
$$\Theta_I^{3,0} = 1 - \Theta_I^{3,1} - \Theta_I^{3,2},$$
$$\Theta_I^{4,2} = 1 - \Theta_I^{4,0} - \Theta_I^{4,1}.$$

There are 12 error rates in total ($\Theta_I^{3,0}$, $\Theta_I^{4,2}$, and the ten error rates on the right-hand side of the preceding equations).

In our model, genotype errors occur independently within trio genotypes, meaning that the probability that two genotype errors occur in a single trio genotype is simply the product of the two genotype error rates. Genotype errors also occur independently across different trio genotypes.

To make the optimization problem slightly easier, we assume that $\Theta_I^{0,1} = \Theta_I^{3,1}$, $\Theta_I^{0,2} = \Theta_I^{3,2}$, $\Theta_I^{2,0} = \Theta_I^{4,0}$, and $\Theta_I^{2,1} = \Theta_I^{4,1}$. This means that a genotype with a deletion (AD or BD) will be observed as AA, AB, or BB with the same probabilities as the genotype that is homozygous for the non-deleted allele. This assumption is reasonable because all sequence reads overlapping a marker will carry the same allele when a genotype is homozygous for that allele and when that allele is the non-deleted allele in a genotype carrying a deletion.

So far, we have described the deletion and genotype error processes in our model (see Figure 2.1 for a graphical summary of these processes). Next, we describe how to calculate the likelihood for observed trio genotypes using this model.

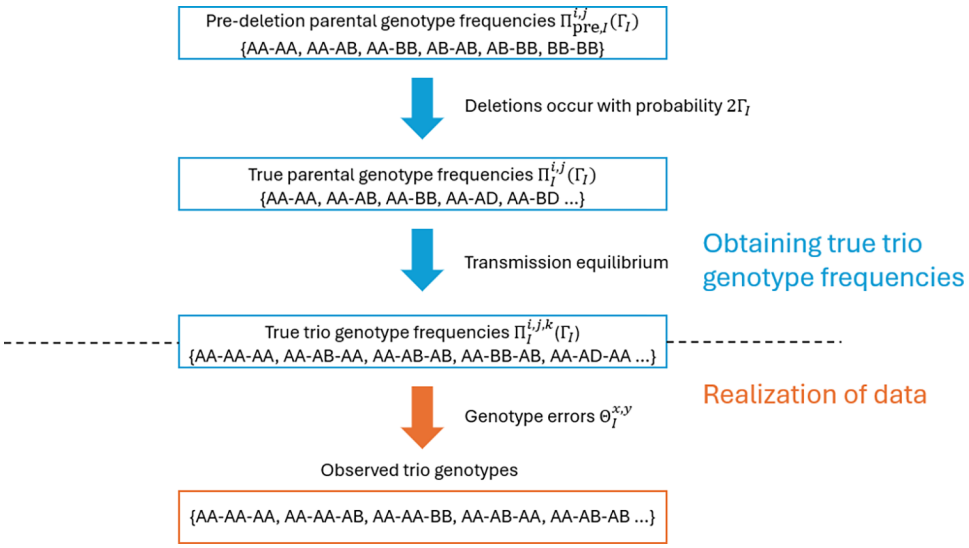

Figure 2.1: Overview of proposed model. We allow deletions to occur on each pre-deletion genotype with a probability of $2\Gamma_I$. The frequencies of the resulting parental genotypes are referred to as the true parental genotype frequencies. True trio genotype frequencies are calculated assuming transmission equilibrium. The observed trio genotypes are drawn proportionally to these frequencies, and genotype errors are introduced during the sampling process.

2.10 Observed trio genotype probabilities

We can obtain the probability of an observed trio genotype G_m at any marker m that falls within MAF interval I by summing over all true trio genotypes. For each true trio genotype, we take its frequency as estimated by our model and take the product of this frequency with the genotype error rates corresponding to genotype errors that would lead us to observe G_m (see Equation 1 of Figure 2.2).

$$P(G_m = a, b, c) = \sum_{i=0}^{4}\sum_{j=i}^{4}\sum_{k=0}^{4} \hat{\Pi}_I^{i,j,k}(\Gamma_I)(\Theta_I^{i,a}\Theta_I^{j,b} + 1_{a\neq b}\Theta_I^{i,b}\Theta_I^{j,a})\Theta_I^{k,c} \quad (1)$$

$$\ell(\boldsymbol{\Theta}_I, \Gamma_I) = \sum_{a=0}^{2}\sum_{b=a}^{2}\sum_{c=0}^{2} G_I^{a,b,c} \log\left(\sum_{i=0}^{4}\sum_{j=i}^{4}\sum_{k=0}^{4} \hat{\Pi}_I^{i,j,k}(\Gamma_I)(\Theta_I^{i,a}\Theta_I^{j,b} + 1_{a\neq b}\Theta_I^{i,b}\Theta_I^{j,a})\Theta_I^{k,c}\right) \quad (2)$$

Figure 2.2: Log-likelihood of observed trio genotypes. Equation 1 shows how the probability of a single observed trio genotype is calculated. Equation 2 shows the joint log-likelihood of all the observed trio genotypes in the sample.

Next, we let $G_I^{i,j,k}$ denote the observed count of trio genotype i, j, k in I. For example, $G_I^{0,1,1}$ represents the observed count of trio genotype AA-AB-AB in MAF interval I. The log-likelihood of observing all trio genotypes within MAF interval I is obtained by summing the log-likelihood of observing each trio genotype (see Equation 2 of Figure 2.2).

2.11 Maximum likelihood estimation

We estimate the genotype error parameters $\Theta_I^{x,y}$ and the uncalled deletion rate Γ_I by maximizing the log-likelihood $\ell(\Theta_I, \Gamma_I)$ (Equation 2 of Figure 2.2) in the log scale of the parameters using the simulated annealing algorithm (SANN) implemented in the `optim()` function in R with the default settings [62]. The SANN method produced larger maximized likelihoods than other optimization methods implemented in the `optim()` function when we fitted our model to the UK Biobank sequence data. Maximum likelihood estimates were exponentiated to their original scales after maximization.

2.12 Estimating the overall genotype error rate and overall uncalled deletion rate

The overall genotype error rate for a set of MAF intervals is the probability that a randomly chosen SNV genotype in the MAF intervals is miscalled. Similarly, the overall uncalled

deletion rate for a set of MAF intervals is one half times the probability that a randomly chosen genotype in the MAF intervals contains a deleted allele.

The frequency of genotype $k \in \{0, 1, 2, 3, 4\}$ in MAF interval I can be obtained by summing up the true trio genotype frequencies for which the offspring has genotype k. Replacing the true trio genotype frequencies by their estimates, we obtain

$$\hat{\Pi}_I^k = \sum_{i=0}^{4} \sum_{j=i}^{4} \hat{\Pi}_I^{i,j,k}.$$

We define the genotype error rate in MAF interval I, Δ_I, as the probability that a randomly chosen genotype in MAF interval I is miscalled. We estimate Δ_I using

$$\hat{\Delta}_I = \sum_{k=0}^{2} \hat{\Pi}_I^k \left(1 - \hat{\Theta}_I^{k,k}\right) + \sum_{k=3}^{4} \hat{\Pi}_I^k.$$

We estimate the overall genotype error rate for a set of MAF intervals as the average of the estimated genotype error rate in each MAF interval, $\hat{\Delta}_I$, weighted by the number of trio genotypes in each MAF interval.

The first MAF interval of $(0, 0.001]$ contains most of the trio genotypes (96.9%) in the UK Biobank sequence data, so including this MAF interval in the weighted average would make the overall genotype error rate for the set of all MAF intervals highly dependent on the genotype error rate in the first MAF interval. Because of this, we report the estimated genotype error rate for $(0, 0.001]$ separately from the estimated overall genotype error rate for the MAF intervals in $(0.001, 0.5]$.

To estimate the overall uncalled deletion rate for a set of MAF intervals, we average the estimated uncalled deletion rates in each MAF interval, $\hat{\Gamma}_I$, weighting by the number of observed trio genotypes in each MAF interval. We report the estimated overall uncalled deletion rate for the 100 MAF intervals in $(0.001, 0.5]$ and the estimated uncalled deletion rate for $(0, 0.001]$ separately.

Finally, the proportion of genotype errors attributable to deletions for the MAF intervals in $(0.001, 0.5]$ can be estimated by dividing two times the estimated overall uncalled deletion rate for the MAF intervals in $(0.001, 0.5]$ by the estimated overall genotype error rate for the MAF intervals in $(0.001, 0.5]$.

2.13 Confidence intervals

We calculate 95% bootstrap confidence intervals for the uncalled deletion rate and genotype error rates in each MAF interval. To obtain each bootstrap sample, we sample with replacement at the trio level and aggregate the observed trio genotype counts across the sampled trios for each MAF interval. The number of trios in each bootstrap sample is the same as the original number of trios in the dataset. We refit our model on the observed trio genotype counts within each MAF interval in the bootstrap sample and store the resulting estimates of the genotype error rates and uncalled deletion rate.

We repeat this process 100 times to estimate the standard error of each maximum likelihood estimator (corresponding to a genotype error rate or the uncalled deletion rate in the log scale) using the sample standard deviation of the corresponding parameter estimates stored from the 100 iterations. For each parameter, we calculate its 95% bootstrap confidence interval centered around its estimate, assuming a normal distribution for the maximum likelihood estimators in the log scale, and exponentiate the bounds to obtain an interval in the original scale. This prevents the lower bound of our confidence interval from being negative. In the simulation study, this confidence interval had empirical coverage rates closer to 95% compared to a bootstrap confidence interval derived by assuming a normal distribution for the maximum likelihood estimators in the original scale. This assumption may not hold in the original scale because the parameters are rates bounded between 0 and 1.

90% percentile bootstrap confidence intervals for the overall genotype error rate, overall uncalled deletion rate, and proportion of genotype errors attributable to deletions for SNVs with MAF > 0.001 were derived by recalculating these rates for each bootstrap iteration and obtaining the 0.05 and 0.95 quantiles of the resulting bootstrap distributions. We used the

percentile method to derive bootstrap confidence intervals for these averaged rates because it was not appropriate to assume that our estimates for these rates followed a normal distribution. The nominal rate of 90% was chosen because of the relatively small number (100) of bootstrap iterations.

2.14 Simulation study 1

We perform a simulation study to calculate the bias of maximum likelihood estimators for the uncalled deletion rate and genotype error rates when observed trio genotypes within a single MAF interval are generated though the deletion and genotype error processes described earlier. We also calculate coverage probabilities of bootstrap confidence intervals for the uncalled deletion rate and genotype error rates. Finally, we compare our estimates of genotype error rates with those obtained from the Browning and Browning method, which does not account for deletions [6].

In the simulation, we generate observed trio genotype counts for four MAF intervals, $(0.01, 0.015]$, $(0.05, 0.055]$, $(0.25, 0.255]$, and $(0.49, 0.495]$, by applying the deletion and genotype error processes with known parameters. These processes are identical to what was described earlier in this section (see Figure 2.1 for a graphical summary of these processes). To start, we generate pre-deletion genotypes for 100 parental pairs at 3×10^8 markers in each MAF interval, resulting in a total of 3×10^{10} pre-deletion parental genotypes per MAF interval.

Minor allele frequencies for the 3×10^8 markers in a MAF interval are drawn from a uniform distribution on the MAF interval. These markers are assumed to be in Hardy-Weinberg equilibrium, and genotypes for parents are drawn based on expected genotype frequencies under random mating.

We use an uncalled deletion rate of 2×10^{-4} per allele to add deleted alleles on genotypes. Thus, a deletion occurs on a genotype with a probability of 4×10^{-4}, in which case a randomly chosen allele is deleted.

We then sample an offspring genotype for each parental pair assuming transmission equi-

librium. For example, if a parental genotype is AB-AD, the genotypes AA, AB, AD, and BD occur with equal probability in the offspring at the marker. If by chance, the offspring's genotype in a true trio genotype is DD, we remove the corresponding trio genotype. This removal maintains the relative frequencies of the true trio genotypes. A DD genotype in the offspring occurs very infrequently because the uncalled deletion rate is low in this simulation study.

We then simulate genotype errors on the true trio genotypes using the genotype error rates in Table 2.2. These genotype error rates were estimated in an analysis of the UK Biobank data that used a preliminary version of our model. We store the resulting observed trio genotypes.

Model parameter	**Value**
$\Theta_I^{0,1}$	2×10^{-4}
$\Theta_I^{0,2}$	5×10^{-7}
$\Theta_I^{1,0}$	9×10^{-5}
$\Theta_I^{1,2}$	2×10^{-4}
$\Theta_I^{2,0}$	2×10^{-6}
$\Theta_I^{2,1}$	9×10^{-4}
Γ_I	2×10^{-4}

Table 2.2: Model parameters used to generate observed trio genotypes in the simulation.

Finally, we fit our likelihood model and the model described in Browning and Browning [6] to estimate genotype error rates (and the uncalled deletion rate when using our model) using the simulated observed trio genotypes (see Section 2.11).

95% bootstrap confidence intervals for genotype error rates (and the uncalled deletion rate when using our model) were generated using the method described in Section 2.13, except we aggregate the observed trio genotype counts across the sampled trios for a single

MAF interval, as opposed to for multiple MAF intervals, and refit our model only on this MAF interval. We repeated the above simulation 200 times to calculate the empirical bias of estimators and the coverage probabilities of confidence intervals.

2.15 Simulation study 2

We perform a second simulation to evaluate our model's estimation of the overall genotype error rate, the overall uncalled deletion rate, and the proportion of genotype errors attributable to uncalled deletions for the MAF intervals in $(0.001, 0.5]$. We first generate observed trio genotypes for the 100 MAF intervals $\big((0.001, 0.005], (0.005, 0.01], (0.01, 0.015], \ldots, (0.495, 0.5]\big)$.

In each MAF interval, we first generate pre-deletion genotypes for 100 parental pairs. We set the number of simulated markers in each MAF interval so that the observed trio genotype count is similar to that of the corresponding interval in the UK Biobank sequence data. Minor allele frequencies for markers in each MAF interval are drawn from a uniform distribution on the corresponding MAF interval. Genotypes for parents are drawn based on expected genotype frequencies under random mating. Then, we used the same deletion and genotype error processes employed in the previous simulation to generate observed trio genotypes for each of the 100 MAF intervals. The uncalled deletion rate and genotype error rates used are fixed across MAF intervals and are identical to the previous simulation (Table 2.2).

We obtain the true genotype error rate for each MAF interval, Δ_I, using the true trio genotype counts (before genotype errors are added) and the true genotype error rates. We then average Δ_I across the 100 MAF intervals, weighting by the trio genotype count in each MAF interval, to calculate the overall genotype error rate for the MAF intervals in $(0.001, 0.5]$.

We obtain the true proportion of genotype errors attributable to deletions for the MAF intervals in $(0.001, 0.5]$ by dividing two times the overall uncalled deletion rate by the overall genotype error rate for the MAF intervals in $(0.001, 0.5]$.

To estimate these quantities, we first fit our likelihood model to the observed trio genotype

counts to estimate the uncalled deletion rate and genotype error rates in each MAF interval. Using the process described in Section 2.6, we estimate the overall genotype error rate, the overall uncalled deletion rate, and the proportion of genotype errors attributable to deletions for the MAF intervals in $(0.001, 0.5]$. We obtain 90% bootstrap confidence intervals for these quantities using the method described in Section 2.13.

Finally, we compare our estimates of the genotype error rate and uncalled deletion rate in each MAF interval, as well as the overall genotype error rate, the overall uncalled deletion rate, and the proportion of genotype errors attributable to deletions for the MAF intervals in $(0.001, 0.5]$ to their true values.

2.16 Results

2.16.1 Simulation study 1

We show our results from the first simulation study in Table 2.3. In Figs S1 and S2, we plot estimates for all parameters obtained from the 200 iterations in each of the MAF intervals used to generate the data. The results vary considerably across the different MAF intervals.

	True value	Bias		SE		Coverage	
		No deletions	Deletions	No deletions	Deletions	No deletions	Deletions
MAF interval $(0.01, 0.015]$							
$\Theta_I^{0,1}$	2×10^{-4}	-9.19×10^{-5}	$\mathbf{-8.91 \times 10^{-5}}$	$\mathbf{2.00 \times 10^{-5}}$	2.03×10^{-5}	**0.135**	0.125
$\Theta_I^{0,2}$	5×10^{-7}	1.15×10^{-8}	$\mathbf{9.66 \times 10^{-9}}$	$\mathbf{2.24 \times 10^{-8}}$	2.37×10^{-8}	**0.940**	0.935
$\Theta_I^{1,0}$	9×10^{-5}	3.57×10^{-3}	$\mathbf{3.45 \times 10^{-3}}$	$\mathbf{8.38 \times 10^{-4}}$	8.59×10^{-4}	**0.135**	0.105
$\Theta_I^{1,2}$	2×10^{-4}	1.51×10^{-4}	$\mathbf{-4.93 \times 10^{-5}}$	$\mathbf{1.21 \times 10^{-5}}$	1.07×10^{-4}	0.005	**0.925**
$\Theta_I^{2,0}$	2×10^{-6}	3.59×10^{-3}	$\mathbf{3.53 \times 10^{-3}}$	5.18×10^{-3}	$\mathbf{5.14 \times 10^{-3}}$	0.335	**0.375**
$\Theta_I^{2,1}$	9×10^{-4}	3.34×10^{-3}	$\mathbf{3.08 \times 10^{-3}}$	6.00×10^{-3}	$\mathbf{5.95 \times 10^{-3}}$	**0.960**	0.965
Γ_I	2×10^{-4}	NA	$\mathbf{6.98 \times 10^{-7}}$	NA	$\mathbf{1.11 \times 10^{-4}}$	NA	**0.915**
MAF interval $(0.05, 0.055]$							
$\Theta_I^{0,1}$	2×10^{-4}	-7.91×10^{-5}	$\mathbf{-7.07 \times 10^{-5}}$	2.90×10^{-5}	$\mathbf{2.40 \times 10^{-5}}$	**0.805**	0.570
$\Theta_I^{0,2}$	5×10^{-7}	$\mathbf{9.65 \times 10^{-9}}$	1.44×10^{-8}	3.28×10^{-8}	$\mathbf{3.16 \times 10^{-8}}$	**0.935**	0.905
$\Theta_I^{1,0}$	9×10^{-5}	6.84×10^{-4}	$\mathbf{6.13 \times 10^{-4}}$	2.70×10^{-4}	$\mathbf{2.26 \times 10^{-4}}$	0.695	**0.710**
$\Theta_I^{1,2}$	2×10^{-4}	1.70×10^{-4}	$\mathbf{-1.79 \times 10^{-5}}$	$\mathbf{1.49 \times 10^{-5}}$	1.04×10^{-4}	0.010	**0.930**

Continued on next page

Table 2.3 (continued)

	True value	Bias		SE		Coverage	
		No deletions	Deletions	No deletions	Deletions	No deletions	Deletions
$\Theta_I^{2,0}$	2×10^{-6}	4.98×10^{-4}	$\mathbf{4.49 \times 10^{-4}}$	3.88×10^{-4}	$\mathbf{3.71 \times 10^{-4}}$	0.135	**0.140**
$\Theta_I^{2,1}$	9×10^{-4}	$\mathbf{1.00 \times 10^{-3}}$	1.30×10^{-3}	$\mathbf{1.69 \times 10^{-3}}$	1.83×10^{-3}	**0.935**	0.925
Γ_I	2×10^{-4}	NA	$\mathbf{-1.72 \times 10^{-5}}$	NA	$\mathbf{1.03 \times 10^{-4}}$	NA	**0.885**
MAF interval $(0.25, 0.255]$							
$\Theta_I^{0,1}$	2×10^{-4}	-1.60×10^{-4}	$\mathbf{-1.97 \times 10^{-5}}$	$\mathbf{3.77 \times 10^{-5}}$	4.95×10^{-5}	0.415	**0.975**
$\Theta_I^{0,2}$	5×10^{-7}	$\mathbf{-3.18 \times 10^{-8}}$	3.37×10^{-8}	$\mathbf{5.02 \times 10^{-8}}$	7.11×10^{-8}	0.925	**0.960**
$\Theta_I^{1,0}$	9×10^{-5}	2.37×10^{-4}	$\mathbf{2.70 \times 10^{-5}}$	$\mathbf{5.65 \times 10^{-5}}$	7.36×10^{-5}	0.605	**0.960**
$\Theta_I^{1,2}$	2×10^{-4}	1.83×10^{-4}	$\mathbf{-1.64 \times 10^{-6}}$	$\mathbf{1.85 \times 10^{-5}}$	8.74×10^{-5}	0.005	**0.915**
$\Theta_I^{2,0}$	2×10^{-6}	$\mathbf{4.43 \times 10^{-6}}$	5.97×10^{-6}	$\mathbf{2.14 \times 10^{-6}}$	3.32×10^{-6}	0.120	**0.140**
$\Theta_I^{2,1}$	9×10^{-4}	-8.39×10^{-4}	$\mathbf{7.08 \times 10^{-6}}$	$\mathbf{3.31 \times 10^{-5}}$	4.90×10^{-4}	0.005	**0.895**
Γ_I	2×10^{-4}	NA	$\mathbf{-6.34 \times 10^{-6}}$	NA	$\mathbf{7.10 \times 10^{-5}}$	NA	**0.960**
MAF interval $(0.49, 0.495]$							
$\Theta_I^{0,1}$	2×10^{-4}	-1.37×10^{-4}	$\mathbf{-5.74 \times 10^{-6}}$	$\mathbf{6.25 \times 10^{-5}}$	1.13×10^{-4}	0.760	**0.890**
$\Theta_I^{0,2}$	5×10^{-7}	$\mathbf{1.82 \times 10^{-7}}$	3.68×10^{-7}	$\mathbf{1.68 \times 10^{-7}}$	3.21×10^{-7}	**0.790**	0.675
$\Theta_I^{1,0}$	9×10^{-5}	7.25×10^{-5}	$\mathbf{3.07 \times 10^{-6}}$	$\mathbf{3.26 \times 10^{-5}}$	5.84×10^{-5}	0.570	**0.925**
$\Theta_I^{1,2}$	2×10^{-4}	3.21×10^{-4}	$\mathbf{1.42 \times 10^{-4}}$	$\mathbf{3.15 \times 10^{-5}}$	1.13×10^{-4}	0.060	**0.945**
$\Theta_I^{2,0}$	2×10^{-6}	-1.84×10^{-7}	$\mathbf{8.69 \times 10^{-9}}$	$\mathbf{2.68 \times 10^{-7}}$	3.68×10^{-7}	0.925	**0.970**
$\Theta_I^{2,1}$	9×10^{-4}	-6.50×10^{-4}	$\mathbf{-2.94 \times 10^{-4}}$	$\mathbf{6.68 \times 10^{-5}}$	2.30×10^{-4}	0.815	**0.915**
Γ_I	2×10^{-4}	NA	$\mathbf{-7.42 \times 10^{-5}}$	NA	$\mathbf{6.83 \times 10^{-5}}$	NA	**0.870**

Table 2.3: Simulation results. Genotype error rates were estimated for two models: one with and one without uncalled deletions. The bias of each estimator is calculated by taking the sample mean of each parameter estimate across the 200 simulations and subtracting the true value. The standard error of each estimator is estimated using the sample standard deviation of the parameter estimates across the 200 simulations. The coverage was calculated as the proportion of times the 95% bootstrap confidence intervals captured the true value across the 200 simulations. The bolded values indicate the model that performed better in terms of the corresponding metric (bias, standard error, or coverage).

For the two simulations in which the observed trio genotypes were generated from markers contained in the MAF interval of $(0.01, 0.015]$ or $(0.05, 0.055]$, our model and the model in

Browning and Browning [6], which does not account for uncalled deletions, produce similar estimates and bootstrap confidence intervals for all genotype error rates except for $\Theta_I^{1,2}$. Our model produces a substantially less biased estimator of $\Theta_I^{1,2}$. The bootstrap confidence interval generated using our model also captures the true value of $\Theta_I^{1,2}$ at a rate that is much closer to the nominal rate of 0.95. Both models generate accurate estimates for $\Theta_I^{0,2}$, but not for error rates that result in a higher count of observed AA or AB genotypes ($\Theta_I^{1,0}, \Theta_I^{2,0}, \Theta_I^{0,1}, \Theta_I^{2,1}$). This could be due to an imbalance between genotype counts when these MAF intervals are used (expected BB genotype counts are much smaller than expected AA or AB genotype counts) which make increases in BB genotype counts from genotype errors more apparent than increases in AA or AB genotypes.

For the remaining two simulations in which the observed trio genotypes were generated from markers contained in the MAF interval of $(0.25, 0.255]$ or $(0.49, 0.495]$, our model produces reasonable error rate estimates for the most part, despite some estimators being biased. Overall, our model resulted in estimators for genotype error rates that were less biased but more variable than the Browning and Browning model [6].

Our estimator for the uncalled deletion rate Γ_I seemed to be effectively unbiased, except when minor allele frequencies for markers were drawn from the highest MAF interval of $(0.49, 0.495]$. Here, our model tended to underestimate Γ_I.

2.16.2 Simulation study 2

In Figure 2.3, we plot the estimated and true genotype error rates for each of the 100 MAF intervals in $(0.001, 0.5]$. The true overall genotype error rate for the 100 MAF intervals was 6.3×10^{-4}, while the estimated overall genotype error rate was 6.6×10^{-4} (90% CI: $[4.2 \times 10^{-4}$, $7.1 \times 10^{-4}]$).

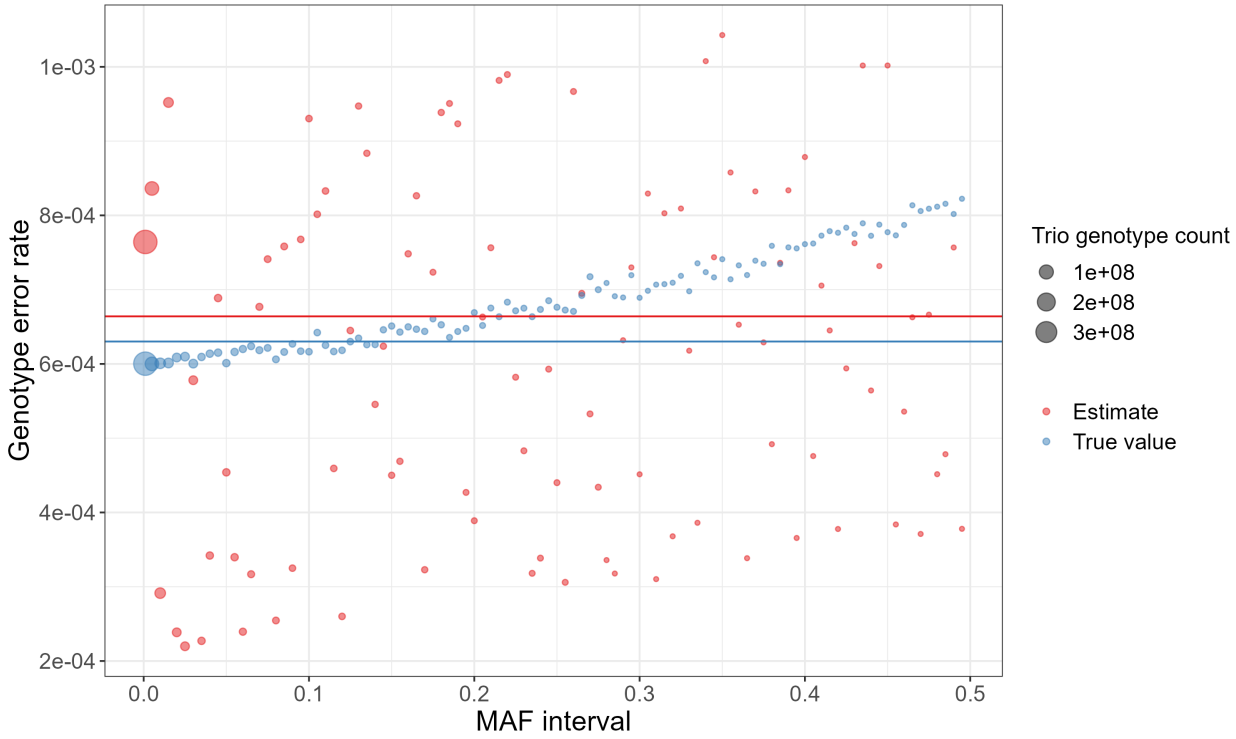

Figure 2.3: True and estimated genotype error rates for 100 MAF intervals in $(0.001, 0.5]$. The blue and red dots respectively represent the true and estimated genotype error rates for each MAF interval. The size of the dots indicates the number of trio genotypes in the corresponding MAF interval. The blue and red horizontal lines respectively indicate the true and estimated overall genotype error rate for the 100 MAF intervals.

In Figure 2.4, we plot the estimated uncalled deletion rate for each of the 100 MAF intervals in $(0.001, 0.5]$. The true overall uncalled deletion rate for the 100 MAF intervals was 2×10^{-4}, while the estimated overall uncalled deletion rate was 2.2×10^{-4} (90% CI: $[9.2 \times 10^{-5}, 2.4 \times 10^{-4}]$). The true proportion of genotype errors attributable to uncalled deletions was 0.63, while the estimated proportion of genotype errors attributable to uncalled deletions was 0.65 (90% CI: [0.44, 0.68]).

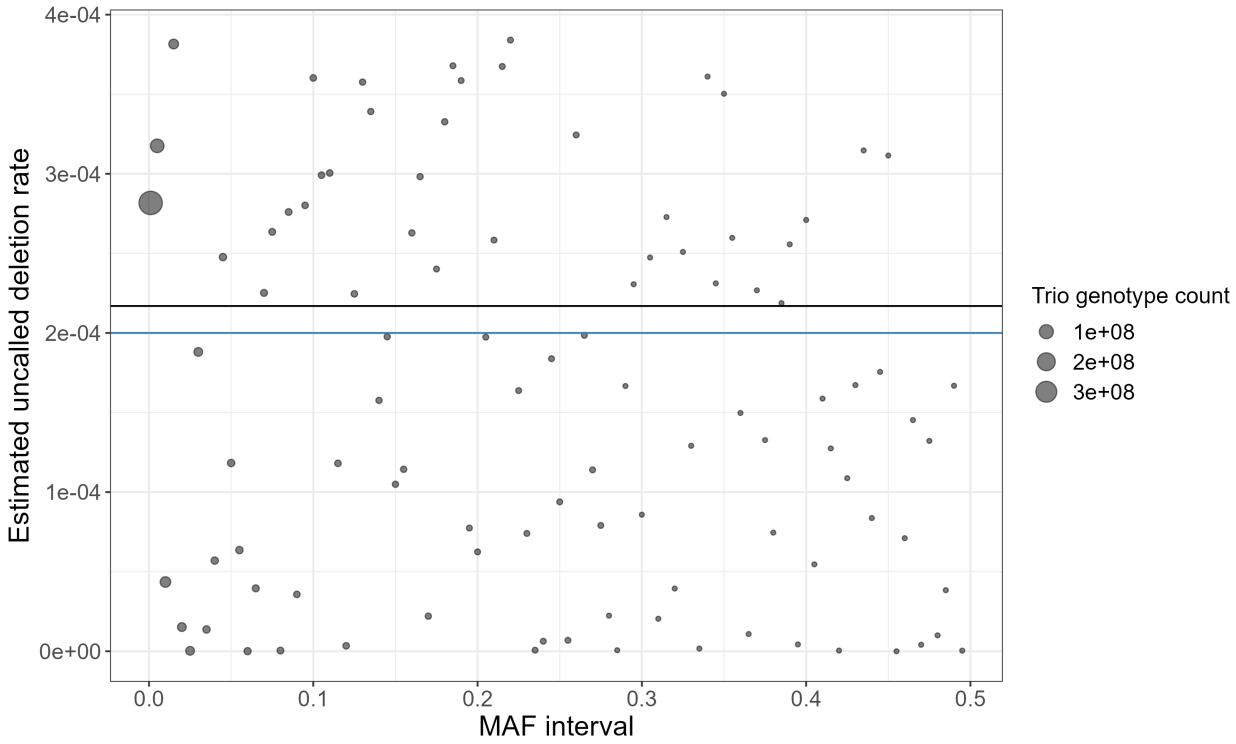

Figure 2.4: Estimated uncalled deletion rates for 100 MAF intervals in $(0.001, 0.5]$. The dots represent the estimated uncalled deletion rate for each MAF interval. The size of the dots indicates the number of trio genotypes in the corresponding MAF interval. The blue and black horizontal lines respectively indicate the true and estimated overall uncalled deletion rate for the 100 MAF intervals.

2.16.3 UK Biobank sequence data

Both our model and the Browning and Browning model were fit on each MAF interval in the UK Biobank sequence data [6]. For each model fit, we obtained the parameter estimates, 95% bootstrap confidence interval for each estimated parameter, and AIC of each model.

In Figure 2.5, we plot the estimates and 95% bootstrap confidence intervals from both models. Although bootstrap confidence intervals for parameters calculated using the two models often overlap within the same MAF interval, the estimates between the two models

have a noticeably different trend across the lower to higher MAF intervals, except for $\Theta_I^{2,0}$ and $\Theta_I^{0,2}$. We see that for the other error rates ($\Theta_I^{0,1}$, $\Theta_I^{1,0}$, $\Theta_I^{1,2}$, and $\Theta_I^{2,1}$), the model with deletions generates higher estimates for error rates that result in a higher count of observed AB genotypes ($\Theta_I^{0,1}$, $\Theta_I^{2,1}$) and lower estimates for error rates that result in a higher count of observed AA or BB genotypes ($\Theta_I^{1,0}$, $\Theta_I^{1,2}$).

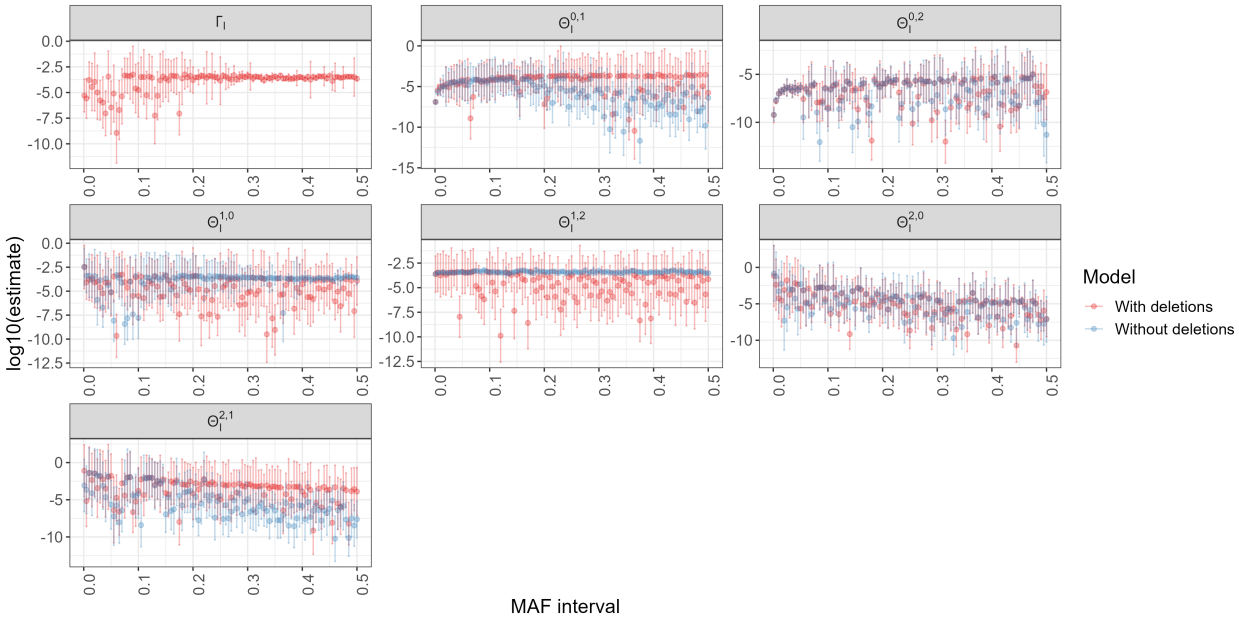

Figure 2.5: Genotype error rate and uncalled deletion rate estimates and confidence intervals for the UK Biobank sequence data. Genotype error rates were estimated for two models: one with and one without uncalled deletions. Each panel represents a parameter that is being estimated. The x-axis of each panel represents the MAF interval, and the y-axis represents the estimate on the $\log_{10}$ scale.

Notably, our model estimates an uncalled deletion rate Γ_I that appears to be largely constant across MAF intervals. This is consistent with our intuition that the rate at which uncalled deletions are present in genotypes should not depend on the MAF of the underlying marker.

We also calculated the AIC of both models fit to each MAF interval. Smaller AIC values indicate better model fit. Compared to the Browning and Browning model [6], our model had a smaller AIC in 81 of the 101 MAF intervals. In Figure 2.6, we plot absolute differences in AIC between the two models for each MAF interval. We see that AIC slightly prefers the Browning and Browning model at lower MAF intervals.

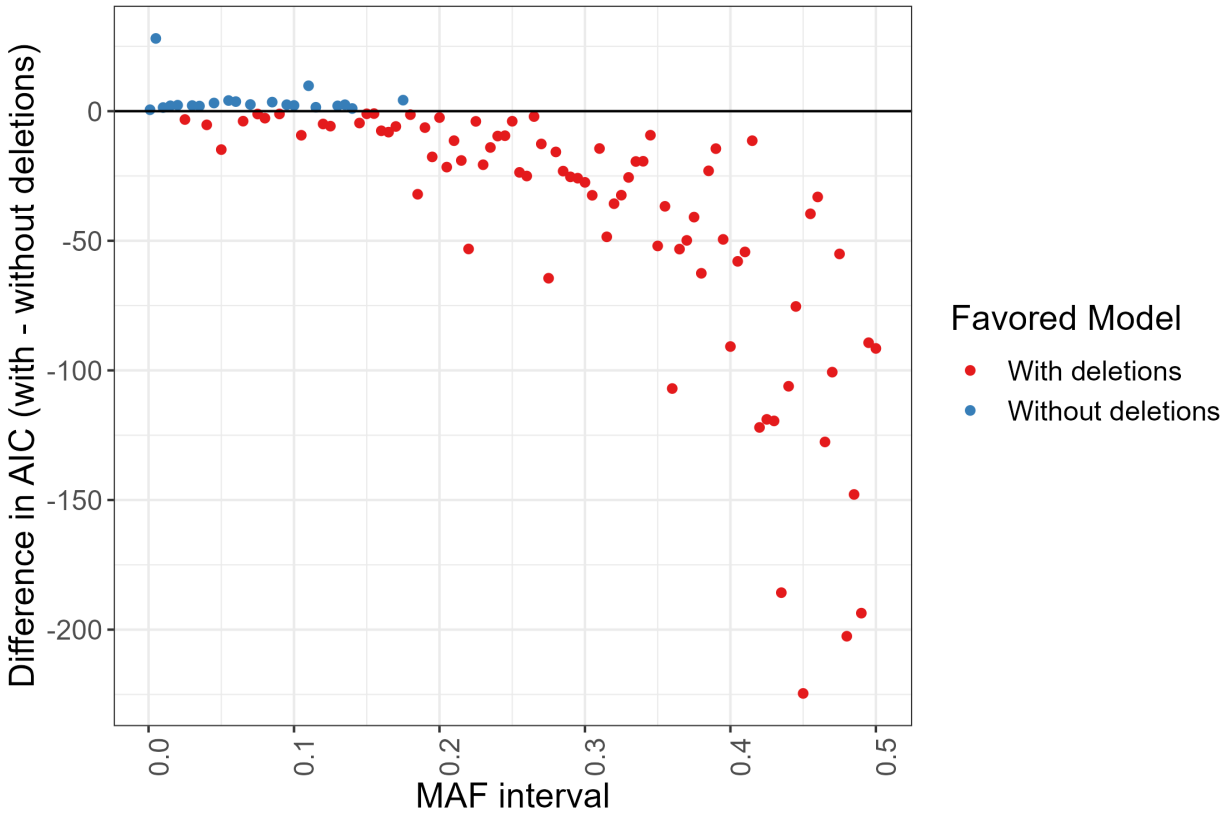

Figure 2.6: Absolute differences in AIC between the model with and without deletions in the UK Biobank sequence data. The difference in AIC between the models was calculated for each MAF interval in the UK Biobank sequence data. Points below the horizontal line at 0 indicate a preference for the model with deletions, while points above the horizontal line indicate a preference for the model without deletions.

Using our model, the overall genotype error rate and uncalled deletion rate for SNVs

with MAF in $(0.001, 0.5]$ were estimated to be 3.2×10^{-4} (90% CI: $[2.8 \times 10^{-4}, 6.2 \times 10^{-4}]$) and 1.2×10^{-4} (90% CI: $[1.0 \times 10^{-4}, 2.7 \times 10^{-4}]$) respectively. We estimated the proportion of genotype errors attributable to uncalled deletions for SNVs with MAF in $(0.001, 0.5]$ to be 0.77 (90% CI: [0.73, 0.88]). For SNVs with MAF in $(0, 0.001]$, the genotype error and uncalled deletion rates were estimated to be 1.0×10^{-5} (90% CI: $[1.8 \times 10^{-6}, 6.8 \times 10^{-4}]$) and 5.1×10^{-6} (90% CI: $[8.0 \times 10^{-7}, 3.4 \times 10^{-4}]$) respectively.

2.16.4 Discussion

Previous studies have attempted to estimate genotype error rates, but no study to our knowledge has controlled for the presence of uncalled deletions in the collected sample, which can lead to biased estimates for these rates. In this study, we present a method for estimating genotype error rates from parent-offspring trio data that allows for the presence of uncalled deleted alleles. Our model allows a proportion of parental genotypes to have uncalled deletions, which can be inherited by the offspring. These uncalled deletions, along with other genotype errors, affect the observed genotypes in our sample. We use a maximum likelihood estimation procedure to estimate the rate of uncalled deletions and genotype errors.

In our simulation studies, our model produced an unbiased estimator for the uncalled deletion rate across a range of marker minor allele frequencies when uncalled deletions are present. Our model also resulted in genotype error rate estimators that were generally less biased than estimators from a model that did not control for uncalled deletions. Finally, our model produced an accurate estimate of the overall genotype error rate, overall uncalled deletion rate, and the proportion of genotype errors attributable to uncalled deletions for a set of MAF intervals.

Using our model, we estimated the overall genotype error rate and uncalled deletion rate at SNVs with MAF in $(0.001, 0.5]$ in 77 sequenced White British parent-offspring trios in the UK Biobank to be 3.2×10^{-4} (90% CI: $[2.8 \times 10^{-4}, 6.2 \times 10^{-4}]$) and 1.2×10^{-4} (90% CI: $[1.0 \times 10^{-4}, 2.7 \times 10^{-4}]$) respectively. We further estimated the proportion of genotype errors attributable to uncalled deletions for these SNVs to be 0.77 (90% CI: [0.73, 0.88]).

This indicates that uncalled deletions are the primary source of genotype error at SNVs with MAF > 0.001 in these data. Based on the AIC, our model was a better fit to the UK Biobank sequence data than a model that does not account for uncalled deletions.

Our model makes some assumptions that may not be met when fitting the model to real data. First, our model assumes that uncalled deletions occur at a constant rate for every genotype in our sample and that genotyped markers are in Hardy–Weinberg equilibrium. The method could potentially be extended to allow for a more flexible process for introducing uncalled deletions into the sample. Second, our model assumes that genotype errors occur independently on each genotype. Third, we assume that genotypes with deleted alleles (AD or BD) are called as heterozygous (AB) and homozygous for the non-deleted allele at the same rates as genotypes that are homozygous for the non-deleted allele. Relaxing this assumption will increase the number of parameters in the current model, which may require developing a more efficient and robust optimization algorithm. Similarly, we could allow for the observation of more than two alleles as an alternative to combining all minor alleles into a composite minor allele, but this would also require increasing the number of model parameters.

Our model is restricted to SNVs that do not overlap called deletions or other structural variants. We expect genotype error rates to be lower for SNVs that do not overlap structural variants compared to SNVs that do overlap structural variants. Extending the model to SNVs overlapping called deletions would allow us to estimate genotype error rates for genotypes with a called deletion. This would give the added benefit of using genotype data from more markers. However, this would also require us to add more parameters to the model, such as the probability of calling each true genotype as a genotype with a deleted allele.

Our model cannot be readily extended to estimate rates of uncalled duplications because it is not clear how to define the true genotype when duplicate copies carry different alleles. If there is an uncalled duplication, we do not know which allele is present at the location to which the sequence reads were mapped. Estimating rates of uncalled duplications will likely require a model framework that can account for this type of uncertainty.

In this study, we propose a novel approach that uses trio genotype data to simultaneously estimate the rate of uncalled deletions and genotype errors. We show that estimated genotype error rates in UK Biobank sequence data depend on whether uncalled deletions are included in the model, and that the error model that includes uncalled deletions appears to better fit the observed sequence data.

Chapter 3

MODELING THE LENGTH DISTRIBUTION OF GENE CONVERSION TRACTS IN HUMANS FROM THE UK BIOBANK SEQUENCE DATA

This chapter is adapted, with minor modifications, from the bioRxiv preprint by Masaki and Browning (2025) [42].

3.1 Introduction

A large number of gene conversion tracts can be detected from biobank-scale sequence data using inferred identity-by-descent (IBD) clusters. A gene conversion event occurring after the most recent common ancestor of an IBD cluster will transfer new alleles onto the haplotype, if the individual undergoing meiosis has at least one heterozygous marker within the gene conversion tract. Allele conversions cause discordant alleles within the IBD cluster in the current population, which can be used to detect past gene conversion events. Because discordant alleles can prevent the detection of the IBD cluster, Browning and Browning devised a method to use non-overlapping regions of each chromosome for detecting IBD clusters and gene conversions that have occurred on each IBD cluster [59]. Applying their method to whole autosome sequence data from 125,361 individuals from the UK Biobank, they found 9,313,066 allele conversions inferred to belong to 5,961,128 gene conversion tracts. To detect an allele conversion, this method requires at least two haplotypes within an IBD cluster to have the same alternate allele. This means that genotype errors will not be falsely identified as allele conversions, unless the same genotype error occurs twice in the same IBD cluster.

In our study, we propose a statistical method to model the length distribution of gene conversion tracts detected from the UK Biobank whole autosome data. In our method, we

account for the difference in the true length of a gene conversion tract and its observed length, which we define as the distance between the furthest allele converted markers inside this tract. The gene conversion tracts that we detect are from past transmissions in the population, for which the parental genotypes are not known. Allele conversions can only occur at heterozygous sites within a gene conversion tract in the transmitting parent, but we do not have access to the transmitting parent's genotype data. This is not a problem in pedigree studies, where the positions of heterozygous sites in both parents are known. To appropriately account for the difference in the true and observed length of each gene conversion tract in our study without access to the transmitting parent's genotype data, we assume that allele conversions occur with the same probability at each position within the same gene conversion tract. We estimate the allele conversion probability for each detected gene conversion tract using the heterozygosity rate of markers near the tract. Additionally, to account for the effects of linkage disequilibrium on the distribution of allele conversions, we found it necessary to exclude observed gene conversion tract lengths of one bp from our dataset, and we account for this exclusion in our analyses (see Appendix C).

We allow the length distribution of gene conversion tracts to follow a geometric random variable, a sum of two geometric random variables, or a mixture of two geometric components. A geometric distribution is appropriate if the gene conversion tract grows one bp at a time, and after each extension, there is a fixed probability that it continues extending to the next bp, independent of previous steps. This distribution has been found to accurately model the length distribution of gene conversion tracts in Drosophila melanogaster [2]. A sum of two geometric random variables is appropriate if the gene conversion tract extends outward in both directions from a central position, with each side following the same extension process as in the geometric case. Here, we assume that the probability of extending by one bp is the same in both directions. A mixture of two geometric components is appropriate if some proportion of gene conversion tracts have a smaller mean length relative to the remaining tracts. This phenomenon has previously been observed in mammals. For example, Wall et al. estimated, applying this distribution to gene conversion tracts from a captive baboon

colony, that more than 99% of all gene conversion tracts were very short (mean 24 bp), but the remaining tracts were much longer (mean 4.3 kb) [30]. Furthermore, Palsson et al. similarly estimated that within shorter gene conversion tracts (< 1 kb) in both sexes, the majority of gene conversion tracts had a smaller mean compared to the remaining tracts [21]. For each tract length distribution, we derive a closed form expression for the distribution of observed tract lengths to efficiently calculate the joint likelihood for nearly one million detected gene conversion tracts during maximum likelihood estimation. After fitting our model for each tract length distribution, we use the Akaike Information Criterion (AIC) to choose the best fitting tract length distribution [23].

We validate our model by fitting it to detected gene conversion tracts from a coalescent simulation, originally described in Browning and Browning (2024), that incorporates evolutionary and technical factors such as mutations, genotype errors, and potential artifacts introduced by the multi-individual IBD detection method used to identify gene conversion tracts [59]. This coalescent simulation was conducted using msprime, which only allows gene conversion tract lengths to be drawn from a geometric distribution [20]. Thus, to test the robustness of our method to different tract length distributions, we run an additional simulation study drawing gene conversion tract lengths from various distributions, including a mixture of two geometric components (see Appendix D).

Finally, we apply our model to estimate the mean length of gene conversion tracts detected from the UK Biobank whole autosome data. In addition to estimating the mean length for all detected tracts, we stratify detected tracts based on whether they overlap with a crossover recombination hotspot, and estimate the mean length separately for both sets of detected tracts.

3.2 UK Biobank whole autosome data

We ran our analysis on whole autosome sequence data from 125,361 individuals from the UK Biobank, who identified themselves as 'white British' in the initial release of 150,119 sequenced genomes. The UK Biobank study was reviewed and approved by the North

West Research Ethics Committee and all subjects gave informed consent [11]. The data were obtained under UK Biobank application number 19934, and the 150,119 genomes were phased using Beagle 5.4 [8, 7].

3.3 Detecting gene conversion tracts

We used gene conversion tracts previously detected in the UK Biobank whole autosome data using IBD clusters [59]. IBD clusters are sets of haplotypes at a locus that have a recent common ancestor. If a recent gene conversion event transfers new alleles onto a haplotype in the IBD cluster, there will be discordant alleles within the IBD cluster, which can then be used to detect this gene conversion event. The detection method splits the genome into short, interleaved regions where IBD clusters are inferred or where gene conversion tracts are detected based on the inferred IBD clusters. These regions were each 9 kb long, for a total of 18 kb per IBD inference and gene conversion detection region pair, and this 18 kb pattern was repeated throughout each chromosome. Furthermore, this 18 kb pattern was offset by 0, 6, and 12 kb, and the analysis repeated for each offset to ensure that allele conversions at all positions could be detected.

Allele conversions were detected at markers where two haplotypes in an IBD cluster shared one allele and two others shared the alternative allele, minimizing the false detection of genotype errors as allele conversions. Furthermore, only markers with MAF ≥ 0.05 were considered to avoid misclassifying mutations as allele conversions.

After allele conversions were detected, they were clustered to form detected gene conversion tracts. Allele conversions were merged into the same gene conversion tract if they were located within 1.5 kb of each other, and their allele bipartitions were concordant. For this condition to be satisfied, each subcluster at the first site (representing a single allele) needs to overlap with a different subcluster at the other site.

Across all the autosomes, 9,313,066 allele conversions were detected [59]. These allele conversions were inferred to belong to 5,961,128 detected gene conversion tracts. Furthermore, 4,943,183 (82.9%) of the detected gene conversion tracts were comprised of a single

allele conversion [59]. 1,017,945 (17.1%) of the detected tracts were comprised of two or more allele conversions. We refer to the length spanning the furthest allele converted markers in a detected gene conversion tract as the observed tract length of the gene conversion tract. If a detected gene conversion tract is comprised of a single allele conversion, the observed tract length is one bp.

We label the observed tract lengths of all detected gene conversion tracts as $\{l_j | j = 1, ..., m\}$. The procedure used to detect gene conversion tracts in each offset assumes that gene conversion tract lengths do not exceed 1.5 kb. To take this into account, we exclude any observed tract lengths exceeding 1.5 kb when estimating the mean gene conversion tract length. This results in the exclusion of 141,361 tracts (2.4% of all detected tracts). We also exclude observed tract lengths of one bp prior to estimation, because our model assigns a higher probability mass at one bp compared to what we observe in the data (see Appendix C). This is likely because we do not account for linkage disequilibrium in our model. Linkage disequilibrium can cause heterozygous genotypes to be observed together at nearby positions more frequently than would be expected if these positions were independent, causing gene conversions to span two heterozygous sites more frequently than would be expected under our model. Although we exclude observed tract lengths of one bp when estimating the mean gene conversion tract length, the proportion of observed tract lengths of one bp is used to understand the effect of linkage disequilibrium on the distribution of observed tract lengths (see Appendix C). We appropriately account for the omission of these tracts in our model by truncating the marginal distribution of observed tract lengths (derived in a later section) at one bp and 1.5 kb. After removing both detected tracts of 1 bp and those exceeding 1.5 kb, we are left with 876,584 detected tracts. Although excluding these tracts reduces the amount of data used in the estimation procedure, results from our simulation study suggest that the resulting estimates are unbiased under the truncated model.

3.4 Definitions and overview of model

We model N, the length of a gene conversion tract, as a geometric random variable, a sum of two independent and identically distributed geometric random variables, or a mixture of two geometric components. We further let L be a random variable representing the observed tract length of a gene conversion tract, which is the length spanning the furthest allele converted markers within the gene conversion tract. The event $L = 0$ represents no allele conversions occurring within the tract, and $L = 1$ represents one allele conversion occurring within the tract. In the following sections, we derive the conditional distribution of L given N and the marginal distribution of L. We further describe the procedure we use to obtain a maximum likelihood estimate of ϕ, $\hat{\phi}$, using the observed tract lengths $\{l_j | j = 1, ..., m\}$ detected from the UK Biobank whole autosome data.

3.5 The distribution of the observed tract length conditional on the gene conversion tract length

The observed tract length of a gene conversion tract, represented by the random variable L, depends on where allele conversions occur on the gene conversion tract. We will first assume that allele conversions happen with probability ψ at every position within some gene conversion tract that is exactly n bp long. Under this scenario, the following conditional distribution has previously been derived [18].

$$P(L = l \mid N = n) = \begin{cases} (1-\psi)^n, & l = 0, \\ n\,\psi\,(1-\psi)^{n-1}, & l = 1, \\ (n-l+1)\,\psi^2\,(1-\psi)^{n-l}, & 2 \leq l \leq n. \end{cases}$$

In the probability above, we conditioned on the gene conversion tract length, represented by the random variable N, being n bp long. Obtaining an observed tract length of zero bp is equivalent to allele conversions not occurring within the gene conversion tract, which

happens with a probability of $(1-\psi)^n$. Next, obtaining an observed tract length of one bp is equivalent to an allele conversion occurring at exactly one position within the gene conversion tract. There are n possible positions in which the allele conversion can occur, and each configuration happens with a probability of $\psi(1-\psi)^{n-1}$. Lastly, to obtain an observed tract length of l bp, where $2 \leq l \leq n$, we need to observe two allele conversions that span exactly l positions, and allele conversions cannot occur at the $n-l$ positions flanking the two allele conversions. There are $n-l+1$ ways to overlay these two allele conversions on the gene conversion tract, and each configuration occurs with a probability of $\psi^2(1-\psi)^{n-l}$.

3.6 Deriving the marginal distribution of the observed tract length

If the gene conversion tract length N is drawn from a geometric distribution with mean ϕ, we have

$$P(N=n) = \left(1-\frac{1}{\phi}\right)^{n-1} \frac{1}{\phi}.$$

Letting $\lambda = 1/\phi$,

$$P(L=l) \;=\; \sum_{n=l}^{\infty} P\big(L=l \mid N=n\big)\, P(N=n) \;=\; \begin{cases} \dfrac{\lambda(1-\psi)}{\lambda+\psi-\lambda\psi}, & l=0, \\ \dfrac{\lambda\psi}{(\lambda+\psi-\lambda\psi)^2}, & l=1, \\ \dfrac{\lambda(1-\lambda)^{l-1}\psi^2}{(\lambda+\psi-\lambda\psi)^2}, & l \geq 2. \end{cases}$$

This is the marginal distribution of the observed tract length L. A closed-form expression for L was not derived previously, but this form is crucial for accelerating likelihood computations, given that we compute the joint likelihood of nearly one million observed tract lengths during maximum likelihood estimation. We further truncate this distribution to appropriately model observed tract lengths detected in the UK Biobank sequence data using the multi-individual IBD method [59]. Recall that we only retain observed tract lengths between

2 and 1,500 bp during estimation, so we account for this by truncating the distribution of L between 2 and 1,500 bp.

$$P(2 \leq L \leq 1500) \;=\; \sum_{l=2}^{1500} \frac{\lambda(1-\lambda)^{l-1}\psi^2}{(\lambda+\psi-\lambda\psi)^2} \;=\; \frac{\psi^2\left[(1-\lambda)-(1-\lambda)^{1500}\right]}{(\lambda+\psi-\lambda\psi)^2}.$$

Then,

$$P(L=l \mid 2 \leq L \leq 1500) \;=\; \frac{P(L=l)}{P(2 \leq L \leq 1500)} \;=\; \frac{\lambda(1-\lambda)^{l-1}}{(1-\lambda)-(1-\lambda)^{1500}}.$$

Notice that conditioning on $2 \leq L \leq 1500$ removed the parameter ψ from our model. As mentioned earlier, $\{l_j \mid j = 1, \ldots, m\}$ represents the observed tract lengths in our dataset. When fitting the model, we use the filtered set of observed tract lengths, $\{l_j \mid j = 1, \ldots, m,\, 2 \leq l_j \leq 1500\}$. Henceforth, we will also index our random variable L using j. L_j represents the random variable corresponding to the observed tract length of detected gene conversion tract j in our dataset. We have

$$P(L_j = l_j \mid 2 \leq L_j \leq 1500, \lambda) \;=\; \frac{\lambda(1-\lambda)^{l_j-1}}{(1-\lambda)-(1-\lambda)^{1500}}.$$

We also consider two alternative distributions for N: a sum of two independent and identically distributed geometric random variables, and a mixture of two geometric components. The derivations of $P(L_j = l_j \mid 2 \leq L_j \leq 1500)$ under both settings are provided in Appendix B.2. Under these settings, $P(L_j = l_j \mid 2 \leq L_j \leq 1500)$ depends on ψ_j, the allele conversion probability for each detected tract, so we estimate ψ_j for each tract j before estimating ϕ. The procedure to estimate ψ_j for each tract j is described in the following section.

3.7 Estimating the allele conversion probability for each detected tract

Recall that ψ_j represents the probability that an allele conversion will occur at each position within detected gene conversion tract j. When N is a sum of two geometric random variables

or a mixture of two geometric components, the likelihood of the observed tract length for detected gene conversion tract j, $P(L_j = l_j \mid 2 \leq L_j \leq 1500)$, depends on ψ_j (see Appendix B.2), so we need to estimate ψ_j for $j = 1, \ldots, m$ to obtain a maximum likelihood estimate for the mean gene conversion tract length ϕ.

Allele conversions occur at positions within each gene conversion tract where the individual is heterozygous. Therefore, the probability that a randomly selected individual from the population is heterozygous at a given marker can be used to estimate the probability that an allele conversion will happen at this marker, once it is included in a gene conversion tract. However, it is difficult to derive a closed-form expression for the marginal distribution of L when we only allow allele conversions to occur at SNV positions, and with differing rates at each SNV position. Thus, we let allele conversions occur with the same probability ψ_j at all positions within detected gene conversion tract j. We use the average heterozygosity rate of positions near detected tract j to estimate ψ_j.

Letting a_j and b_j $(a_j \leq b_j)$ represent the positions on the chromosome corresponding to the furthest allele-converted markers within detected gene conversion tract j, we average the heterozygosity rate across the set of positions $[a_j - 5000,\, b_j + 5000]$ to estimate ψ_j:

$$\hat{\psi}_j = \frac{1}{b_j - a_j + 10001} \sum_{i=a_j-5000}^{b_j+5000} 2p_i(1 - p_i).$$

Here, p_i denotes the minor allele frequency (MAF) of position i on the chromosome in which the gene conversion event occurred. p_i is calculated using the sample of 125,361 White British individuals from the UK Biobank. Variants with MAF less than 5% were excluded when detecting allele conversions, so we cannot observe allele conversions at these positions (see Section 3.3). Therefore, if the MAF is less than 5% at position i, we set $p_i = 0$. The formula $2p(1-p)$ for heterozygosity at a marker assumes that Hardy–Weinberg equilibrium holds, which is a reasonable approximation for common variants in a relatively homogeneous population.

If either $a_j - 5000$ or $b_j + 5000$ exceeds the end of the chromosome, the averaging only

takes place within the bounds of the chromosome (e.g. if $a_j = 100$ and $b_j = 200$, we only average the heterozygosity rate from positions 1 to 5,200).

3.8 Maximum likelihood estimation of the mean gene conversion tract length

Given observed tract lengths $\{l_j \mid j = 1, \ldots, m\}$, we propose the following maximum likelihood estimator for ϕ, the mean gene conversion tract length, when the gene conversion tract length N is drawn from a geometric distribution. Recall that the version of the model in which N is geometric was parameterized by $\lambda = 1/\phi$, but we can simply maximize with respect to ϕ. In other words,

$$\hat{\phi} = \arg\max_{\phi} \sum_{j \in I_2^{1500}} \log P\big(L_j = l_j \mid 2 \leq L_j \leq 1500, \phi\big),$$

where $I_2^{1500} = \{\, j = 1, \ldots, m \mid 2 \leq l_j \leq 1500\}$. When N is a sum of two geometric random variables, we parameterize the distribution of L using $\gamma = 2/\phi$ (see Appendix B.2). Unlike the geometric case, our marginal distribution of L_j truncated between 2 and 1,500 still depends on ψ_j, so for each j, we plug in our estimated $\hat{\psi}_j$ in place of ψ_j. Then, we can again maximize with respect to ϕ:

$$\hat{\phi} = \arg\max_{\phi} \sum_{j \in I_2^{1500}} \log P\big(L_j = l_j \mid 2 \leq L_j \leq 1500, \phi, \psi_j = \hat{\psi}_j\big).$$

When N is a mixture of two geometric components, we have three unknown parameters ϕ_1, ϕ_2, and w_1, which represent the mean of the first component, the mean of the second component, and the mixing proportion of the first component (see Appendix B.2). Again, our marginal distribution of L_j truncated between 2 and 1,500 still depends on ψ_j, so for each j, we plug in our estimated $\hat{\psi}_j$ in place of ψ_j. Then, we can maximize with respect to ϕ_1, ϕ_2, and w_1:

$$\big(\hat{\phi}_1, \hat{\phi}_2, \hat{w}_1\big) = \arg\max_{\phi_1, \phi_2, w_1} \sum_{j \in I_2^{1500}} \log P\big(L_j = l_j \mid 2 \leq L_j \leq 1500, \phi_1, \phi_2, w_1, \psi_j = \hat{\psi}_j\big).$$

To find the arg max when N is geometric or a sum of two geometric random variables, we use the L-BFGS-B algorithm implemented in the `scipy.optimize.minimize` function from the SciPy Python library [47]. When N is a mixture of two geometric components, we define a grid for w_1 ranging from 0.002 to 0.5, using increments of 0.00025 between 0.002 and 0.01, and increments of 0.05 between 0.05 and 0.5. We chose a finer grid at smaller values of w_1 because preliminary analyses of observed tract lengths from the UK Biobank whole autosome data consistently inferred w_1 to be close to zero. Then, for each w_1 value in the grid, we again ran the L-BFGS-B algorithm from four starting values of (ϕ_1, ϕ_2): $(0.0005, 0.0005)$, $(0.0005, 0.1)$, $(0.1, 0.0005)$, and $(0.1, 0.1)$. Multiple starting values were used because the likelihood of (ϕ_1, ϕ_2) (fixing w_1) appeared to have multiple local maxima. The final maximum likelihood estimates were selected as the set of (w_1, ϕ_1, ϕ_2) values achieving the highest joint likelihood across all grid points of w_1 and starting values of L-BFGS-B. A closed-form for the maximum likelihood estimator is available for geometric N if we use a fixed allele conversion probability ψ for all tracts and we condition on $L \geq 1$ (see Appendix B.1).

To choose between the three distributions of N, we propose calculating the Akaike Information Criterion (AIC) under each version of the model [23]. Lower AIC indicates that the distribution of N that is used is a better fit to the data.

3.9 Bootstrap confidence intervals

We calculate 95 % bootstrap confidence intervals for ϕ (w_1, ϕ_1, ϕ_2 in the case where N is a mixture of two geometric components). We denote the number of detected gene conversion tracts with observed tract length between 2 and 1,500 bp as $|I_2^{1500}|$. To obtain each bootstrap sample, we sample with replacement $|I_2^{1500}|$ observed tract lengths from the set $\{\, l_j \mid j = 1, \ldots, m,\ 2 \leq l_j \leq 1500\}$. Each bootstrap sample consists of the set of observed tract lengths $\{l_j\}$ and allele conversion probabilities $\{\psi_j\}$ corresponding to the resampled indices.

We refit our model to 500 bootstrap samples and obtain a new maximum likelihood estimate of ϕ (or w_1, ϕ_1, ϕ_2 in the case where N is a mixture of two geometric components)

for each bootstrap sample. We take the 0.025 and 0.975 quantiles of the resulting bootstrap distributions and use these as the bounds of our 95 % bootstrap confidence intervals.

3.10 Simulation study

We use simulated data described in Browning and Browning (2024) [59]. 20 regions of length 10 Mb were generated for 125,000 individuals using the coalescent simulator msprime v1.2 [20]. The demographic model for the simulation was an exponentially growing population with an initial size of 10,000 and a growth rate of 3% per generation for the past 200 generations. To simulate recombination and mutation, a crossover rate of 1 cM/Mb and a mutation rate of 1.5×10^{-8} per bp per meiosis were used. The mutation rate used is similar to previously inferred mutation rates using IBD segments [68, 49]. Gene conversions were simulated with an initiation rate of 0.02 per Mb and gene conversion lengths were simulated from a geometric distribution with a mean tract length of 300 bp. The processes used to add uncalled deletions and genotype errors are described in Browning and Browning (2024) [59]. Variants with MAF ≤ 0.01 were excluded, the phase information was removed, and Beagle 5.4 was used to statistically phase the genotypes [8]. The multi-individual IBD analysis detected 284,838 allele conversions belonging to 226,007 detected gene conversion tracts across the 20 regions. We fit our model to the detected gene conversion tracts in each of the 20 regions to estimate the mean gene conversion tract length in each region. For the purposes of this simulation study, we refer to the detected gene conversion tracts in each region as a separate replicate dataset. We refer to fitting our model to the detected gene conversion tracts in each of the 20 regions as a separate replicate of this simulation study.

We fit our model under all three distributions for the true tract length (geometric, sum of two geometric random variables, and mixture of two geometric components). Because the true tract lengths in this simulation study are drawn from a geometric distribution, we are interested in whether the version of the model in which the tract length is geometric will be favored using AIC.

msprime only allows gene conversion tract lengths to be drawn from a geometric distribu-

tion [20]. Thus, to test the robustness of our method to different tract length distributions, we run an additional simulation study drawing gene conversion tract lengths from various distributions, including a mixture of two geometric components (see Appendix D).

3.11 UK Biobank analysis

We previously described how we obtain the observed tract lengths of all detected gene conversion tracts from the UK Biobank whole autosome data, denoted $\{l_j \mid j = 1, \ldots, m\}$. We fit our model on this dataset, using all three tract length distributions (geometric, sum of two geometric random variables, and mixture of two geometric components). We further compare model fit under each of these distributions using AIC.

In addition, we run a stratified analysis, stratifying observed tract lengths based on whether they overlapped with a crossover hotspot. To avoid ascertainment bias, where longer tracts are more likely to overlap a crossover hotspot by chance, we defined overlap based on whether the midpoint of the detected gene conversion tract was inside a crossover hotspot. To define crossover hotspots, we use the deCODE genetic map from Halldorsson et al. and follow their definition of crossover hotspots as regions with crossover rates exceeding ten times the genome-wide average [10].

We calculate local crossover rates between nearby markers on each chromosome by dividing the genetic distance between the two markers by their physical distance. Initially, we calculate the local crossover rate between the first marker in the genetic map and the marker closest to it that is distant by at least 2 kb. We next calculate the local crossover rate between this newly identified marker and the marker closest to it that is distant by at least 2 kb. We repeat this process until the last marker on this chromosome is included in a local crossover rate calculation, or until we cannot identify further markers that are at least 2 kb away.

If the local crossover rate between two markers is more than ten times the genome-wide average, we classify the region spanning these markers as a crossover hotspot. We stratify the observed tract lengths $\{l_j \mid j = 1, \ldots, m\}$ based on whether the midpoint of the

corresponding detected gene conversion tract was inside a crossover hotspot. We then fit our model, separately for each set of tracts. We again use all three tract length distributions to fit the model in this stratified analysis, and compare model fit using AIC.

3.12 Results

3.12.1 Simulation study

We fit our model to the observed tract lengths from each replicate of the simulation study. The number of observed tract lengths between 2 bp and 1.5 kb across the 20 replicates ranged from 2,005 to 2,314. Recall that a geometric distribution with mean 300 bp was used to simulate gene conversion tract lengths in this simulation study. We estimate the mean tract length under all three tract length distributions (geometric, sum of two geometric random variables, and mixture of two geometric components).

Estimates and confidence intervals using the geometric setting are shown in Figure 3.1. The average estimate of the mean tract length across the 20 replicates is 289.5 bp under the geometric setting, which is slightly lower than the true mean of 300 bp used to simulate the gene conversion tracts. Under the geometric setting, the true mean of 300 bp is contained in our 95% bootstrap confidence intervals in 14 out of the 20 replicates.

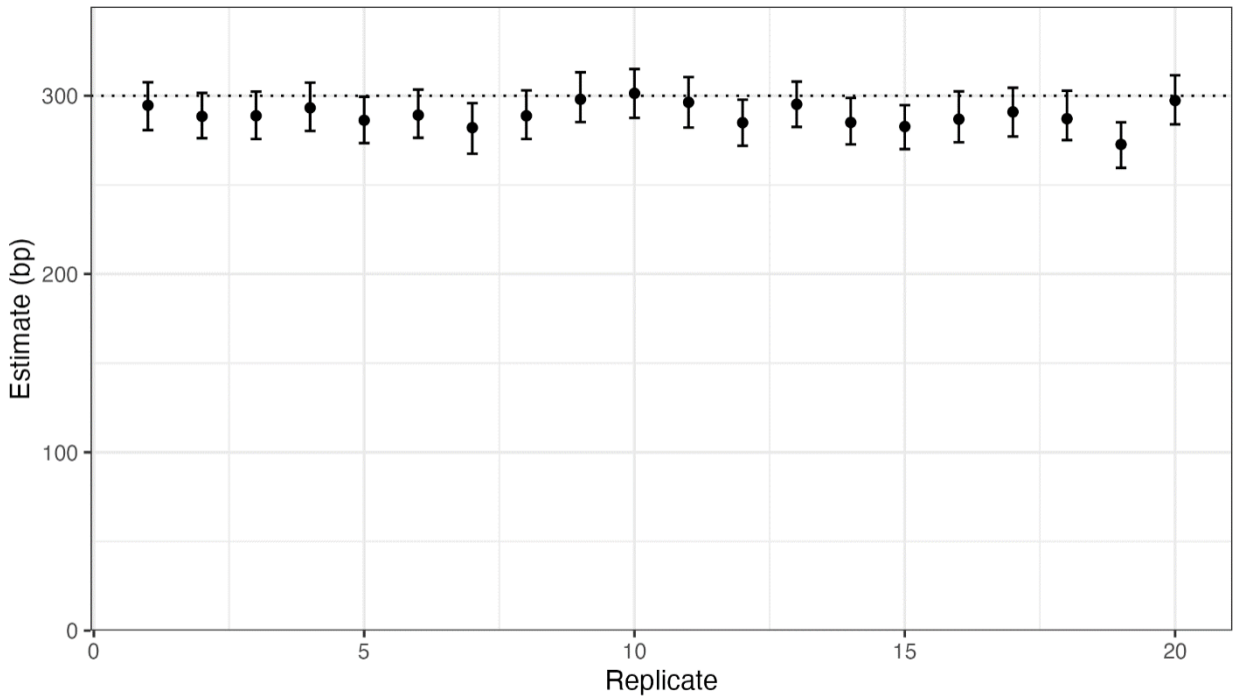

Figure 3.1: The estimated mean gene conversion tract length under the geometric setting across replicate simulations. The dotted horizontal line represents the true mean gene conversion tract length. Gene conversion tract lengths were simulated using a geometric distribution. We plot our estimate and 95% bootstrap confidence interval under the geometric setting for each replicate simulation.

The geometric setting results in the smallest AIC in 16 out of the 20 replicates. For the remaining four replicates, AIC is lowest when gene conversion tract lengths are assumed to be drawn from a mixture of two geometric components. Estimates for these four replicates using the mixture setting are shown in Figure 3.2.

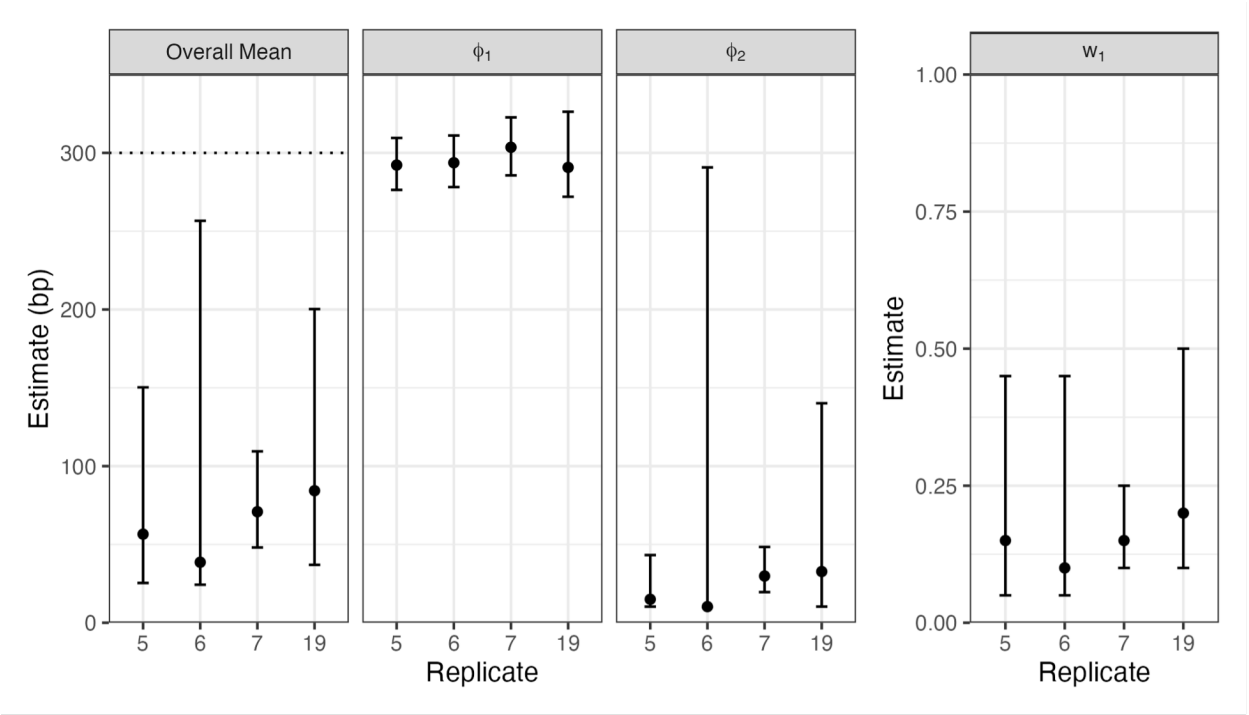

Figure 3.2: Parameter estimates for four replicates using the mixture distribution. The dotted horizontal line represents the true mean gene conversion tract length. We plot the estimated parameter values with 95% bootstrap confidence intervals for each replicate simulation.

For these four replicates, we see that the mixture setting underestimates the overall mean of 300 bp. Notice that the mean of the first component is estimated to be close to 300 bp for these replicates, but the mean of the second component is estimated to be much lower. The mixing proportion of the first component is estimated to be between 0.1 and 0.2 across the four replicates. 95% confidence intervals for parameters tend to be wide, except for the mean of the first component.

Although the mixture setting results in estimates of the overall mean that are much lower compared to the geometric setting for these four replicates, the difference in AIC between these settings is very small for two of the four replicates (1.6 and 0.8). The difference in AIC

for the remaining two replicates is 22.9 and 14.6.

Across all 20 replicates, the difference in AIC between the geometric and mixture settings (positive values preferring the mixture setting) ranges from 22.9 to -4. An AIC difference of -4 indicates that the log-likelihoods of the two settings were equal, and the difference between the AICs is because of the two additional parameters used under the mixture setting. Because the geometric distribution is nested within the mixture of two geometric components, the log-likelihood under the geometric setting cannot exceed that of the mixture setting.

3.12.2 UK Biobank analysis

We applied our estimation method to the observed tract lengths detected from the UK Biobank whole autosome data. The AIC is lowest (indicating best fit) under the setting where the true tract length distribution is assumed to be a mixture of two geometric components (11,860,323). The AIC for the geometric and sum of two geometric settings were 12,201,916 and 12,268,153 respectively. The difference in AIC between the mixture setting and the geometric setting, which had the next lowest AIC, was 341,593, providing strong evidence in favor of the mixture setting.

When assuming that gene conversion tract lengths are a mixture of two geometric components, we estimate the mixing proportion for the first component to be 0.00525 (95% CI: [0.005, 0.00525]). We estimate the mean of the first and second components to be 724.7 bp (95% CI: [720.1, 728.7]) and 16.9 bp (95% CI: [16.4, 17.0]) respectively. We estimate the overall mean to be 20.6 bp (95% CI: [19.9, 20.7]).

For the stratified analysis, we calculated the genome-wide average crossover rate to be 1.23 cM/Mb. We classify any regions exceeding ten times this rate as a crossover hotspot. Of the 876,584 tracts detected from the UK Biobank sequence data, the midpoints of 290,766 (33.2%) were contained within a crossover hotspot. For both tract sets—the set of tracts with midpoint in a crossover hotspot and the remaining tracts—the lowest AIC was obtained under the mixture setting, so we report our results from assuming that gene conversion tract lengths are drawn from the mixture distribution.

For detected tracts with midpoint located within a crossover hotspot, we estimate the mean of the first and second components to be 579.8 bp (95% CI: [574.8, 585.5]) and 20.3 bp (95% CI: [19.7, 21.1]) respectively. We further estimate the mixing proportion for the first component to be 0.0095 (95% CI: [0.00925, 0.01]). We estimate the overall mean to be 25.6 bp (95% CI: [24.9, 26.7]).

For detected tracts with midpoint not located within a crossover hotspot, we estimate the mean of the first and second components to be 813.9 bp (95% CI: [807.7, 819.3]) and 15.5 bp (95% CI: [14.9, 15.6]) respectively. We further estimate the mixing proportion for the first component to be 0.004 (95% CI: [0.00375, 0.004]). We estimate the overall mean to be 18.7 bp (95% CI: [17.9, 18.8]).

3.13 Discussion

Previous studies have tried to measure gene conversion tract lengths in humans by detecting allele conversions from pedigree and sperm-typing data [4, 3, 35, 9]. However, in these studies, it is only possible to detect gene conversion events occurring in a relatively small number of meioses. Efforts to detect gene conversions from pedigree data have been limited by the number of multi-generational pedigrees that have been genotyped. Sperm-typing studies have also been limited by the availability of appropriate data. In sperm-typing studies, distinguishing genotype errors from allele conversions is also difficult.

By applying the multi-individual IBD method to the UK Biobank whole autosome data, we were able to detect gene conversion events across multiple meioses in the ancestral history of this population [59]. Using this method, 5,961,128 gene conversion tracts were detected, which is several orders of magnitude larger than what had been detected in humans in the past. In the largest pedigree study conducted to detect gene conversions, fewer than 30,000 gene conversion events were detected from 5,420 trios [21].

We proposed a likelihood-based estimation method to infer the mean gene conversion tract length. Our method is inspired by a previous approach developed by Betran et al., which was applied to gene conversion tracts detected in 34 *Drosophila subobscura* sequences [18].

However, we made several key improvements. First, we define a separate allele conversion probability for each gene conversion tract, based on the density and heterozygosity rate of markers near each tract. Second, we allow gene conversion tract lengths to follow multiple distributions, including a mixture of two geometric components, which has previously been found to appropriately model gene conversion tract lengths in other mammals [30]. Third, we derive the closed-form expression for the distribution of observed tract lengths for each true tract length distribution, which allows for fast and exact calculation of the joint likelihood during maximum likelihood estimation. Finally, we allow for the selection of the best fitting tract length distribution using AIC.

We ran a coalescent simulation incorporating gene conversion events to validate our estimation method. Since we used msprime for the simulation, gene conversion tract lengths were necessarily drawn from a geometric distribution. Nonetheless, this simulation allowed us to accurately capture potential biases arising from evolutionary and technical factors such as mutations and genotype errors, as well as potential artifacts introduced by the multi-individual IBD detection method used to identify gene conversion tracts [59]. We found that our model accurately estimated the mean gene conversion tract length when the length distribution of gene conversion tracts was correctly specified to be geometric. Our model resulted in biased estimates of the mean gene conversion tract length when the length distribution was incorrectly specified. In most replicates of this simulation study (16 out of 20 replicates), AIC correctly determined the best fitting distribution to be geometric.

To further assess the robustness of our model to the misspecification of the tract length distribution, we conducted a separate simulation study where gene conversion tract lengths were drawn from multiple distributions (see Appendix D). In this study, we found that the model selected by AIC consistently produced relatively unbiased estimates across a range of tract length distributions. Furthermore, when the true tract length distribution was one of the three distributions that we allow for in our model, we found that AIC selects the true distribution in most cases.

Applying our method to observed tract lengths detected from the UK Biobank whole

autosome data, we found that the mixture setting, which had the lowest AIC by a large margin, estimated most tracts have a small mean of 16.9 (95% CI: [16.4, 17.0]), and only a small proportion of tracts have a much larger mean of 724.7 (95% CI: [720.1, 728.7]). The mixing proportion for the geometric distribution with the smaller mean was estimated to be 0.00525 (95% CI: [0.005, 0.00525]). We estimate the overall mean to be 20.6 (95% CI: [19.9, 20.7]).

Our estimate of the mean gene conversion tract length is very sensitive to the assumed tract length distribution. When assuming that gene conversion tract lengths are geometric, our model estimates the mean gene conversion tract length to be 459.0 bp (95% CI: [457.3, 460.5]), which is much higher than our estimate under the mixture setting. However, given the large AIC difference between these two models (341,593), we are confident that the mixture distribution is a much better fit to the data. This result aligns with previous findings in humans. Palsson et al. found that among tracts shorter than 1 kb, the majority had a smaller mean length compared to the longer tracts [21]. The higher estimate we obtained under the geometric distribution is also consistent with our simulation results. In the simulation assessing the robustness of our method, where we draw gene conversion tract lengths from various distributions (see Appendix D), we found that assuming a geometric distribution when the true distribution is a mixture of two geometric components can lead to an inflated estimate of the mean tract length, particularly when one component has a substantially larger mean but contributes relatively few tracts (see Table D.1).

We estimated the overall mean gene conversion tract length to be 20.6 bp (95% CI: [19.9, 20.7]), which is shorter than previous estimates. For instance, Palsson et al. reported mean tract lengths of 123 bp (95% CI: [94, 135]) for paternal and 102 bp (95% CI: [71, 125]) for maternal transmissions [21]. Methodological differences between our approach and the NCOurd model used by Palsson et al. may account for this discrepancy [45]. NCOurd requires specifying a penetrance parameter, defined as the probability that a heterozygous marker within a gene conversion tract is allele converted. In our framework, we set the allele conversion probability within each tract equal to the local mean heterozygosity rate.

This effectively assumes that, for shorter gene conversion tracts (< 1.5 kb), all heterozygous markers are allele converted. This would correspond to using a penetrance of 1 in NCOurd. In contrast, Palsson et al. estimate a fixed penetrance of 0.66 for all detected tracts by using a grid of penetrance values and selecting the one that maximizes the model likelihood. This implies that roughly a third of heterozygous sites within a gene conversion tract do not undergo allele conversion, leading to longer estimated tract lengths. Importantly, penetrance may vary with tract length, making the use of a single penetrance value potentially inappropriate. However, estimating penetrance as a function of the tract length is challenging, especially for short tracts, which often do not overlap with many markers. This limitation has been noted in the original NCOurd publication [45].

There are a few other findings on the length distribution of gene conversion tracts in humans, most notably, in the sperm-typing study by Jeffreys and May, which concluded that the mean length is in the range of 55–290 bp [3]. Jeffreys and May inferred the range of mean gene conversion tract lengths (55–290 bp) by comparing observed gene conversion lengths to simulated tracts under geometrically and normally distributed gene conversion tract lengths. However, our simulation where tract lengths are drawn from a mixture distribution suggests that modeling all tracts using a single distribution, without explicitly accounting for outliers, can lead to an inflated estimate of the mean when a small proportion of tracts are much longer than the rest (see Appendix D).

Wall et al. analyzed gene conversion tracts shorter than 10 kb in a captive baboon colony using a mixture of two geometric distributions [30]. They estimated that 99.8% of tracts had a mean length of 24 bp (95% CI: [18, 31]), while the remaining tracts had a mean of 4.3 kb (95% CI: [2.6, 4.9]). Both the mixing proportion and the mean of the shorter component are similar to our estimates.

We ran an additional analysis in which we stratified detected gene conversion tracts from the UK Biobank whole autosome data by whether their midpoints were located within a crossover hotspot. In both sets of tracts—the set of tracts with midpoints located within a crossover hotspot and the remaining tracts—AIC was smallest when assuming a mixture

distribution for the true tract length distribution. Comparing the estimated parameters for the mixture distribution in each set, detected tracts with midpoints located within a hotspot were estimated to have a larger proportion of longer tracts (0.0095; 95% CI: [0.00925, 0.01]) compared to the remaining detected tracts (0.004; 95% CI: [0.00375, 0.004]). The mean of the longer component of the mixture distribution was estimated to be smaller for hotspot tracts (579.8 bp; 95% CI: [574.8, 585.5]) compared to the remaining tracts (813.9 bp; 95% CI: [807.7, 819.3]). The mean of the shorter component of the mixture distribution was estimated to be larger for hotspot tracts (20.3 bp; 95% CI: [19.7, 21.1]) compared to the remaining tracts (15.5 bp; 95% CI: [14.9, 15.6]). The overall mean was larger for hotspot tracts (25.6 bp; 95% CI: [24.9, 26.7]) compared to the remaining tracts (18.7 bp; 95% CI: [17.9, 18.8]). These differences in the proportion of longer tracts, and in the mean lengths of the shorter and longer components were significant. This is a preliminary finding and we recommend further analysis to confirm this result. Recombination hotspots correlate with other genomic features such as GC rate [58], so the difference may be caused by factors other than the recombination rate itself.

It is important to acknowledge that our method omits observed tract lengths exceeding 1.5 kb, because we cannot accurately detect observed tract lengths corresponding to longer gene conversion tracts. Complex gene conversion events, which result in both allele converted and non-allele converted markers, often span more than 1.5 kb [9]. To appropriately model the lengths of these longer tracts, we would need to apply a detection method that can reliably detect these tracts.

In this study, we did not extend the mixture distribution, which was strongly favored by AIC, to have more than two components. While a mixture model with additional components may better capture the true distribution of gene conversion tract lengths, exploring such models proved computationally challenging due to the complexity of the optimization procedure and the large number of detected gene conversion tracts. Future work may consider more flexible models, such as three-component mixtures, particularly as methods for detecting longer or complex gene conversion events from population-level sequence data

become available.

Chapter 4

DETECTING THE PANGO LINEAGE ANCESTRY OF RECOMBINANT SARS-COV-2 SEQUENCES

4.1 Introduction

A wide range of computational approaches have been developed to detect recombination in viruses. Broadly, similarity methods such as SimPlot visualize how a query sequence's similarity shifts across the genome relative to putative parents [56, 55]. RDP4 examines all triplets within a set of sequences and applies a suite of tests (e.g., GENECONV, MaxChi, Bootscan, 3SEQ) to detect recombination breakpoints and assign parental sequences [41, 57, 50, 37]. However, the number of comparisons is cubic with respect to the sample size, which is infeasible for large-scale datasets.

Phylogeny–based methods such as GARD detect breakpoints by fitting phylogenies to alignment segments and comparing model fit across candidate partitions [33]. The repeated tree-fitting and model-comparison steps are computationally intensive, so GARD is generally applied to downsampled alignments rather than to surveillance-scale datasets comprising millions of genomes.

More recently, SARS-CoV-2-specific tools have been designed to operate on surveillance-scale datasets. Bolotie uses an hidden Markov model (HMM) where the latent states represent SARS-CoV-2 lineages [66]. The Viterbi algorithm is used to predict the sequence of lineages contributing ancestry at each position. RIPPLES identifies candidate recombinant sequences by scanning a global mutation-annotated phylogeny for unusually long branches that may represent recombination events [65]. For each candidate recombinant sequence, RIPPLES partitions the genome into multiple segments and re-places each onto the global phylogeny using maximum parsimony. RecombinHunt compares segment-wise mutation pat-

terns on a query sequence to lineage-specific profiles [5]. It constructs a cumulative likelihood profile across the genome and uses the Akaike Information Criterion to choose between three models with zero, one, or two breakpoints.

Although these SARS-CoV-2–specific tools scale to surveillance-scale datasets, each has method-specific limitations. Bolotie's HMM does not model de novo mutations or genotype errors, which can induce spurious state switches when the query sequence harbors mutations absent from the mutation profile of its true lineage. The HMM's transition probability is also user-specified, making breakpoint detection sensitive to this choice. RIPPLES relies on a single mutation-annotated global phylogeny. Uneven sampling and sequencing artifacts can inflate or deflate the long-branch signal used to identify candidate recombinants. Moreover, the threshold for the long-branch signal is defined by the user, and the initial candidate set of recombinant sequences is sensitive to this chosen cutoff. RecombinHunt relies on several hard evidence gates (e.g., declaring a genome non-recombinant when it differs from the most likely lineage by ≤ 2 mutations), and these thresholds are likewise sensitive to de novo mutations and genotype errors. Finally, both RIPPLES and RecombinHunt permit at most two breakpoints, even though recombinant lineages with more breakpoints have been detected.

In this paper, we develop a method to detect recombinant SARS-CoV-2 sequences within a test set of sequences collected over a short interval (a few days to a week), using a representative set of past sequences. Our method employs an HMM inspired by the Li and Stephens model [38] that accounts for de novo mutations and genotype errors in both recombinant and non-recombinant sequences. For each query sequence, we estimate a pseudo-frequency for observing an allele not present in the lineage providing ancestry at that position, as well as the probability of lineage transitions between consecutive sites. Our method does not rely on a phylogeny, user-defined parameters, nor hard cutoffs, and can accommodate any number of breakpoints.

We evaluate performance in a simulation where we generated synthetic recombinants and controls from SARS-CoV-2 genomes sampled between January and March 2022. We report

sensitivity and specificity for labeling a synthetic sequence as a recombinant, parental-lineage identification accuracy, defined as the proportion of synthetic recombinant sequences for which the inferred parental lineages matches the truth, and breakpoint localization error, measured as the absolute nucleotide distance between the true and inferred breakpoints, along with other metrics.

We applied our method to GenBank SARS-CoV-2 genomes collected from September 2020 to March 2024, partitioning sequences into contiguous, non-overlapping windows. In total, 440,307 unique sequences were evaluated, of which 7,619 were classified as recombinant (1.73%; 95% CI: [1.69%, 1.77%]).

4.2 Obtaining SARS-CoV-2 sequences and clustering Pango lineages

We obtained SARS-CoV-2 sequences and associated metadata from GenBank, processed using the Nextstrain pipeline [24]. After filtering for sequences collected in England between September 2020 and March 2024, we clustered Pango lineages based on their sequence count. Any Pango lineages with fewer than 10,000 sequences were collapsed into their parental lineage using unaliased versions of Pango lineages [53]. This was done iteratively to ensure that all collapsed Pango lineages contained at least 10,000 sequences. Lineages without a defined parent were grouped into a shared "other" category. We collapsed 2304 Pango lineages that existed during this period to 41 collapsed lineages (including other). Unless otherwise specified, all mentions of Pango lineages refer to the collapsed lineages resulting from this procedure.

4.3 Reference and test sets

From the sequences collected in England between September 2020 and March 2024, we generated sliding windows of reference and test set pairs. Each sliding window consisted of 43 days, and these windows were incremented by 7 days at a time to generate 185 sliding windows.

Sequences collected in the first 36 days and last 7 days of each 43-day sliding window

were respectively assigned to the reference and test set for this sliding window. If the first 36 days of a sliding window contained more than 100,000 sequences, we drew a uniform random sample of 100,000 sequences and assigned this to the reference set. Similarly, if the last 7 days of a sliding window contained more than 3,000 sequences, we drew a uniform random sample of 3,000 sequences and assigned this to the test set.

4.4 Calculating the nucleotide frequency matrix for each reference set

For each of the 185 reference sets from each sliding window, we calculated a nucleotide frequency matrix that contains the frequency of each nucleotide (A, T, C, G) at every genome position for each Pango lineage. The nucleotide frequency was calculated by dividing the per-site nucleotide counts by the total sequence count in each Pango lineage. When calculating frequencies, we excluded all non-standard nucleotides (i.e., those other than A, T, C, or G). If no sequences in a Pango lineage carried any of the standard nucleotides at a given position, we assigned equal probabilities (0.25 each) to A, T, C, and G.

4.5 Predicting local Pango lineage ancestries for test sequences

The local Pango lineage ancestry of a SARS-CoV-2 sequence in a given test set refers to the specific Pango lineage contributing ancestry to each genomic position of the sequence.

If a sequence in a given test set results from a recombination event that involves parental sequences from distinct Pango lineages, its local ancestry will consist of segments derived from these distinct lineages, with transitions between segments marking recombination breakpoints. Conversely, for non-recombinant sequences, the local Pango lineage ancestry will be uniform across the genome, corresponding to a single parental lineage. It is important to note that the true Pango lineage ancestry at each position of a test sequence is defined in relation to the Pango lineages present in the corresponding reference set. For example, lineages L_1, L_2, and L_3 may all be present in the corresponding reference set, with lineage L_3 arising from a recombination event between two sequences in lineages L_1 and L_2 respectively. In this case, a sequence from lineage L_3 in the test set will have L_3 as the true local Pango

lineage ancestry at all positions of the genome.

We predict the local Pango ancestry of all sequences in each test set using the nucleotide frequency matrix calculated from the corresponding reference set and an HMM inspired by the Li and Stephens model [38].

This HMM jointly models the latent sequence of local Pango lineage ancestry and the observed nucleotide sequence for each test sequence. It does so by considering three key components: (i) the probability of each Pango lineage contributing ancestry at the first position (initial state probabilities), (ii) the probability of transitioning between different Pango lineages along the genome (transition probabilities), and (iii) the probability of observing each nucleotide at a given position, conditional on the Pango lineage (emission probabilities). Transitions between lineages correspond to recombination events.

4.6 Hidden Markov model to predict local Pango lineage ancestry

In this section, we mathematically define the HMM used to predict the local Pango lineage ancestry of each test sequence.

Let the genome length be denoted by N, and let $t \in \{1, 2, \ldots, N\}$ index genomic positions. We define the latent Pango lineage ancestry at position t as c_t, where $c_t \in \{1, 2, \ldots, M\}$ and M is the number of distinct Pango lineages represented in the corresponding reference set. Each value of c_t corresponds to one of these M lineages.

We further denote the observed nucleotide at position t as O_t. O_t takes values in the set $\{\mathrm{A}, \mathrm{T}, \mathrm{C}, \mathrm{G}\}$.

In the following sections, we define the three key components of this HMM, which are the initial state probabilities, the transition probabilities, and the emission probabilities.

4.6.1 Initial state probabilities

The initial state probabilities represent the probability of each Pango lineage contributing ancestry at the first genomic position. We define the initial state probability of Pango lineage i ($i \in \{1, 2, \ldots, M\}$) as $\pi_i = P(c_1 = i)$, where c_1 denotes the ancestral lineage at the first

position. In our model, we set π_i proportional to the frequency of lineage i in the reference set. Let n_i be the number of sequences assigned to lineage i in the reference set, and let $n_{\text{total}} = \sum_{j=1}^{M} n_j$ be the total number of sequences across all M lineages. Then,

$$\pi_i = \frac{n_i}{n_{\text{total}}}, \quad i \in \{1, 2, \ldots, M\}.$$

4.6.2 Transition probabilities

Transition probabilities represent the probability that we will transition from one Pango lineage ancestry to another between consecutive positions on the test sequence. Here, transitions between distinct Pango lineages correspond to recombination breakpoints. We define the transition probability from Pango lineage i to Pango lineage j as

$$a_{ij} = P(c_{t+1} = j \mid c_t = i), \quad i, j \in \{1, 2, \ldots, M\}, \quad t \in \{1, 2, \ldots, N-1\}.$$

Here, a_{ij} represents the probability that the local Pango lineage ancestry of the test sequence changes from i to j between any consecutive positions on the genome. In our model, we set transition probabilities as

$$a_{ij} = \begin{cases} \sigma, & \text{if } i = j, \\ \frac{1-\sigma}{M-1}, & \text{if } i \neq j. \end{cases}$$

Thus, σ is the probability that we do not encounter a recombination breakpoint between consecutive positions on the genome. We also assume that transitions between any two Pango lineages $i \neq j$ occur with the same probability. Because σ is an unknown parameter, we later describe our method for estimating σ.

4.6.3 Emission probabilities

Emission probabilities of the HMM define the probability of observing each nucleotide (i.e., A, T, C, G) at a particular position on the test sequence, conditional on the local Pango lineage ancestry at that position. We define the emission probability of observing nucleotide k at position t, conditional on the local Pango lineage ancestry being i at position t, as

$$b_{i,t}(k) = P(O_t = k | c_t = i), \quad k \in \{\mathrm{A}, \mathrm{T}, \mathrm{C}, \mathrm{G}\}, \quad i \in \{1, 2, \ldots, M\}, \quad t \in \{1, 2, \ldots, N\}.$$

$b_{i,t}(k)$ depends on the nucleotide frequency matrix calculated from the corresponding reference set. We use $f_{i,t}(k)$ to denote the frequency of nucleotide k at position t in Pango lineage i in the corresponding reference set. To adjust for possible mutations and genotype errors that could occur on the test sequence, we apply a pseudo-frequency ϵ. Specifically, we let

$$b_{i,t}(k) = \frac{f_{i,t}(k) + \epsilon}{1 + 4\epsilon}.$$

The pseudo-frequency ϵ assigns a non-zero probability of observing a nucleotide at position t, when the Pango lineage assigned to be the ancestral state at t contains no sequences that have this nucleotide in the reference set. We want to allow for this non-zero probability in case the test sequence acquires a mutation (or genotype error) at position t that leads to an observed nucleotide that is not contained in the Pango lineage. A small value of ϵ allows occasional mutations or genotype errors without forcing a lineage switch in the HMM. Because ϵ is an unknown parameter, we describe our method for estimating ϵ in the following section.

Occasionally, there will be an ambiguous nucleotide on a test sequence. Because an ambiguous symbol conveys no information about which nucleotide is present, we set the emission to have a probability of one in this case.

Symbol	Description
N	Genome length
M	Number of Pango lineages in the reference set
c_t	Pango lineage ancestry at position t
O_t	Observed nucleotide at position t ($\in \{\mathrm{A}, \mathrm{T}, \mathrm{C}, \mathrm{G}\}$)
σ	Non-transition probability ($P(c_{t+1} = c_t)$)
ϵ	Pseudo-frequency for emissions (accounts for mutations and genotyping errors)
a_{ij}	Transition probability from lineage i to lineage j
$b_{i,t}(k)$	Emission probability of nucleotide k at t given $c_t = i$
π_i	Initial probability that $c_1 = i$

Table 4.1: Summary of symbols used in the hidden Markov model.

4.7 Maximum likelihood estimation of parameters in the hidden Markov model

We have two unknown parameters in our HMM. σ represents the probability that the local Pango lineage ancestry does not change between consecutive positions and ϵ is our psuedo-frequency used to adjust emission probabilities.

To perform maximum likelihood estimation on these two parameters, we want to obtain the marginal probability of the observed nucleotide sequence of a test sequence, conditional on these two parameters. In this section, we describe the procedure we use to obtain this marginal probability.

Using the transition and emission probabilities described in the previous section, it is relatively straightforward to obtain the joint probability of a latent sequence of local Pango lineage ancestry and the observed nucleotide sequence for a test sequence. Let $\{i_1, i_2, \ldots, i_N\}$ be an arbitrary sequence of Pango lineage ancestry and $\{k_1, k_2, \ldots, k_N\}$ be the observed nucleotide sequence for a test sequence. Then,

$$P(c_1 = i_1, c_2 = i_2, ..., c_N = i_N, O_1 = k_1, O_2 = k_2, ..., O_N = k_N \mid \sigma, \epsilon)$$
$$= \pi_{i_1} b_{i_1,1}(k_1) a_{i_1 i_2} b_{i_2,2}(k_2) a_{i_2 i_3} b_{i_3,3}(k_3) \cdots a_{i_{N-1} i_N} b_{i_N,N}(k_N).$$

To obtain the marginal probability of the observed nucleotide sequence conditional on the model, we can simply sum up this joint probability across all possible sequences of local Pango lineage ancestry, as shown below.

$$P(O_1 = k_1, O_2 = k_2, ..., O_N = k_N \mid \sigma, \epsilon)$$
$$= \sum_{i_1, i_2, ..., i_N \in \{1,2,...,\mathrm{M}\}} P(c_1 = i_1, c_2 = i_2, ..., c_N = i_N, O_1 = k_1, O_2 = k_2, ..., O_N = k_N \mid \sigma, \epsilon).$$

This procedure can be done efficiently using the forward algorithm described by Rabiner [51].

We can maximize this marginal probability with respect to our two parameters to obtain our maximum likelihood estimates, as shown below.

$$\hat{\sigma}, \hat{\epsilon} = \arg\max_{\sigma, \epsilon} P(O_1 = k_1, O_2 = k_2, ..., O_N = k_N \mid \sigma, \epsilon).$$

Maximum likelihood estimation of σ and ϵ is done for each test sequence.

Optimization was carried out with the limited-memory BFGS algorithm subject to box constraints, using `scipy.optimize.minimize` (`method = "L-BFGS-B"`) [47].

During numerical optimization, we reparameterize σ to $\tau = (1 - \sigma)(N - 1)$, which represents the number of expected transitions for the test sequence. Furthermore, we optimized ϵ on the log scale and later exponentiated to obtain our estimate in the original scale. The search was initialized at $(\log(\epsilon), \tau) = (\log(0.005), 1)$ and restricted to the intervals $\log(\epsilon) \in [\log(10^{-8}), \log(0.02)]$ and $\tau \in [0, 3]$. The reparameterization of σ to τ was done to avoid possible numerical instabilities that might arise when trying to optimize σ directly,

because we expect σ to be 1 or very close to 1. We similarly optimized ϵ in the log scale because we expect ϵ to be close to 0. We chose the upper bound of three for τ because most discovered recombinant lineages were detected to have three or fewer breakpoints.

4.8 Obtaining the most likely sequence of Pango lineage ancestry

We apply the Viterbi algorithm described by Rabiner [51] to each test sequence to obtain the maximum-likelihood path of latent Pango lineage ancestry states, and hence the locations of detected recombination breakpoints. When applying the Viterbi algorithm, we use our maximum likelihood estimates of the two frequencies, σ and ϵ, described in earlier sections. Specifically, we compute the sequence of Pango lineage ancestry that maximizes the joint probability of the ancestry path and the observed nucleotide sequence, given our maximum likelihood estimates. In other words,

$$\hat{i}_1, \hat{i}_2, \ldots, \hat{i}_N = \underset{i_1, i_2, \ldots, i_N}{\arg\max}\, P(c_1 = i_1, c_2 = i_2, \ldots, c_N = i_N, O_1 = k_1, O_2 = k_2, \ldots, O_N = k_N \mid \hat{\sigma}, \hat{\epsilon}).$$

Here, $\hat{i}_1, \hat{i}_2, \ldots, \hat{i}_N$ denotes the predicted sequence of Pango lineage ancestry. Pango lineage transitions in the predicted sequence of Pango lineage ancestry are interpreted as recombination breakpoints.

Note that because $P(O_1 = k_1, O_2 = k_2, \ldots, O_N = k_N \mid \hat{\sigma}, \hat{\epsilon})$ is constant with respect to the sequence of local Pango lineage ancestry, the above is equivalent to maximizing the posterior probability of the sequence of local Pango lineage ancestry given the observed nucleotide sequence and our maximum likelihood estimates. In other words,

$$\hat{i}_1, \hat{i}_2, \ldots, \hat{i}_N = \underset{i_1, i_2, \ldots, i_N}{\arg\max}\, P(c_1 = i_1, c_2 = i_2, \ldots, c_N = i_N \mid O_1 = k_1, O_2 = k_2, \ldots, O_N = k_N, \hat{\sigma}, \hat{\epsilon}).$$

In Figure 4.1, we show an illustration of the Viterbi algorithm.

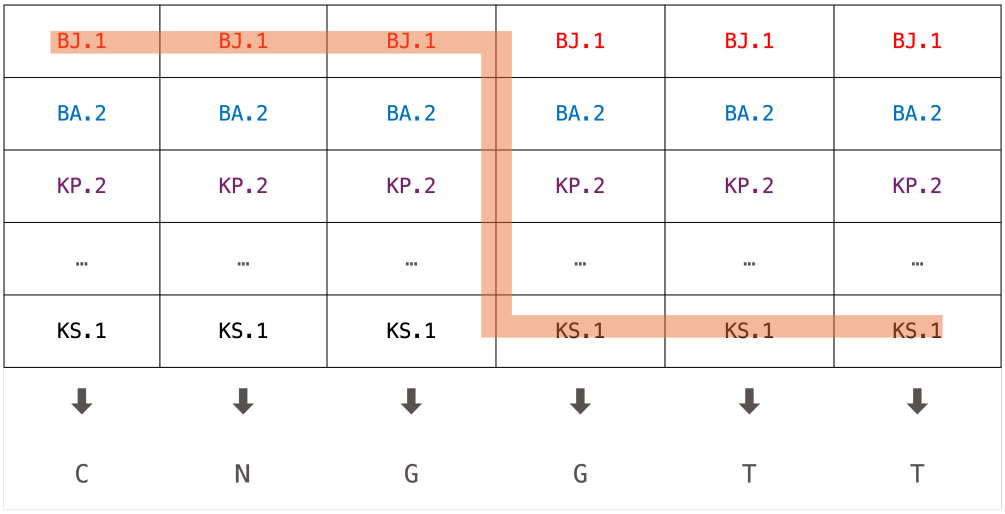

Figure 4.1: Illustration of the Viterbi algorithm. We find a maximum-likelihood path across a grid of Pango lineages, which maximizes the joint probability of the latent path of Pango lineage ancestry (shown using the orange path) and the observed nucleotide sequence shown in the bottom row.

4.9 Simulation study

To assess our method's ability to detect recombination and accurately assign local Pango lineage ancestry, we conducted a simulation study using synthetic SARS-CoV-2 sequences with known local Pango lineage ancestries. These synthetic sequences were generated from real SARS-CoV-2 genomes.

To generate these synthetic sequences, we selected the reference set comprised of 100,000 SARS-CoV-2 sequences collected in England between January 30 and March 6, 2022. Using the sequences in this reference set, we simulated 1,000 recombinant sequences with two parental lineages and 1,000 control sequences with one parental lineage. Of the 1,000 recombinant sequences, 500 of them were generated using a single recombination breakpoint. To generate these sequences, we randomly sampled two parental sequences from different Pango lineages in the reference set and copied nucleotides from one parent up to a breakpoint chosen uniformly on the genome, and from the other parent thereafter. The remaining 500

recombinant sequences were generated using two breakpoints. For these sequences, we again sampled two parental sequences from different Pango lineages. We chose two breakpoints uniformly from all possible breakpoint combinations and inserted a middle segment from one sequence between these breakpoints, replacing the corresponding region in the genome of the other sequence. When a synthetic recombinant sequence was an exact copy of one of its parental sequences, we discarded this sequence and repeated the sequence generation process. To mimic mutations and genotyping errors, we introduced random point mutations into all recombinant sequences at a rate of 0.0002 per site (which corresponds to approximately six mutations per genome on average).

For each of the 2,000 synthetic sequences, we used the method described in Section 4.8 to predict the local Pango lineage ancestry. Emission probabilities for the HMM are based on the nucleotide frequency matrix calculated from the reference set.

To evaluate performance, we conducted several quantitative assessments. First, we estimated the sensitivity and specificity of our method for detecting recombinant sequences. We classified a test sequence as a recombinant if the detected local ancestry contained at least one lineage transition. Controls were treated as true negatives, and synthetic recombinants as true positives. Second, we estimated the mean position-by-position accuracy of the detected local ancestry across synthetic sequences by comparing the detected Pango lineage at each genomic position to the true parental lineage. Third, we assessed whether the set of parental lineages was correctly recovered for each synthetic sequence by comparing the set of true parental lineages with the set of lineages detected at any genomic position. Both the mean position-by-position accuracy and the set-level recovery rate of parental lineages were estimated separately for recombinant and non-recombinant sequences.

For the sensitivity, specificity, and set-level recovery rate of parental Pango lineages, we report 95% confidence intervals derived using the Clopper-Pearson method. For mean position-by-position accuracy, we calculated 95% bootstrap confidence intervals by sampling 500 times with replacement from synthetic sequences (either from the set of recombinants or the set of non-recombinant sequences), calculating the mean position-by-position accuracy

in each bootstrap sample, and taking the 2.5th and 97.5th percentiles of the bootstrapped estimates.

Because the probability of detecting a recombinant is higher when the parental sequences of the recombinant are less similar, we also modeled the sensitivity $S(d)$ as a logistic function of the genome-wide Hamming distance d between the parental sequences:

$$\text{logit}\big(S(d)\big) = \beta_0 + \beta_1 d.$$

Here $\exp(\beta_1)$ represents the multiplicative difference in the odds of detection for two synthetic sequences whose parental Hamming distances are one unit apart. To fit this model, we used the model's predicted label for all of the synthetic recombinants (1 if the model detected the sequence as a recombinant and 0 otherwise).

To quantify breakpoint resolution, we calculated the distance between each inferred breakpoint and its corresponding true genomic position. We restricted analysis to synthetic recombinants whose detected breakpoint count matched the number of true breakpoints. For synthetic recombinants with one breakpoint, we simply calculated the distance between the true and detected breakpoint. For synthetic recombinants with two breakpoints, we paired true and detected breakpoints by ordering both the true and detected breakpoints from 5' to 3' and matching them in order (first with first, second with second). Within each pair, we computed the distance between the true and detected breakpoint, and calculated the mean of these pairwise distances. In cases where the number of detected breakpoints differed from the number of true breakpoints, we did not compute distances. However, we assessed the number of cases where there was a mismatch in the number of breakpoints by tabulating the number of detected and true breakpoints for each sequence.

4.10 Real data analysis

We applied our method to the full set of SARS-CoV-2 sequences collected in England between September 2020 and March 2024. As described in Section 4.3, the sequences were divided

into temporally matched reference and test sets using a sliding 43-day window, with a 36-day reference period followed by a 7-day test period. We detected the local Pango lineage ancestry for sequences in each test period using the method described in Section 4.8. If there were more than 3,000 sequences in a 7-day test period, we randomly sampled 3,000 sequences within the 7-day period and detected the local Pango lineage ancestry for these 3,000 sequences. Otherwise, we detected the local Pango lineage for all sequences within the 7-day period.

Emission probabilities for the HMM were derived from the nucleotide frequency matrix computed from sequences in the corresponding reference set.

After obtaining the predicted local Pango lineage ancestries for all test sequences, we classified any sequences with a recombination breakpoint (meaning that there are at least two inferred parental lineages) to be a recombinant sequence. We later look at the number of detected recombinant sequences in each test set, and their proportion relative to the total number of sequences in each test set. We further investigate all of the detected breakpoint positions across all detected recombinant sequences.

4.11 Results

4.11.1 Simulation study

To evaluate the performance of our method for detecting recombinant SARS-CoV-2 sequences, we conducted a series of assessments based on the predicted local Pango lineage ancestry of synthetic SARS-CoV-2 sequences. The process used to generate these sequences are described in Section 4.9.

We first quantified sensitivity and specificity using the presence of at least one inferred lineage transition (i.e., recombination breakpoint) as the classification criterion. Our method achieved a sensitivity of 0.583 (95% CI: [0.552, 0.614]) and a specificity of 0.995 (95% CI: [0.988, 0.998]) in distinguishing recombinant from non-recombinant sequences.

To assess the accuracy of local ancestry inference, we computed the mean position-by-

position accuracy separately for recombinant and control sequences. On average, the inferred Pango lineage matched the true parental lineage at 78.4% of genomic positions for recombinant sequences (95% CI: [77.0%, 79.9%]). Among control sequences, mean position-by-position accuracy was 99.2% (95% CI: [98.6%, 99.6%]).

We further evaluated whether the set of true parental lineages was correctly detected. The exact set of parental lineages was detected in 35.6% of synthetic recombinant sequences (95% CI: [32.6%, 38.7%]), and there was an overlap between the true and detected lineages in 98.1% of synthetic recombinant sequences (95% CI: [97.0%, 98.9%]). The correct lineage was detected in 99.0% of synthetic control sequences (95% CI: [98.2%, 99.5%]) and there was an overlap between the true and detected lineages in 99.4% of synthetic control sequences (95% CI: [98.7%, 99.8%]).

We next looked at set-level detection stratified by the true parental lineage combination. In Table 4.2, we observe the proportion of times we detected the exact set of true parental lineages for synthetic recombinant sequences, stratified by lineage combinations with at least ten synthetic recombinants.

True lineages	Num. samples	Recovered	Prop.	2.5 % CI	97.5 % CI
(BA. 1.1, BA. 2.9)	17	12	0.706	0.440	0.897
(BA. 1.1, BA. 2.1)	12	8	0.667	0.349	0.901
(BA. 1.1, BA. 1.15)	14	8	0.571	0.289	0.823
(BA. 1.1, BA. 2.3)	16	9	0.563	0.299	0.802
(BA. 1.17.2, BA. 2)	68	38	0.559	0.433	0.679
(BA. 1.15, BA. 2)	17	9	0.529	0.278	0.770
(BA. 1.1, BA. 2)	225	114	0.507	0.439	0.574
(BA. 1.15.1, BA. 2)	22	11	0.500	0.282	0.718
(BA. 1.1.15, BA. 2)	16	8	0.500	0.247	0.753
(BA. 1, BA. 2)	142	66	0.465	0.381	0.550
(BA. 1.1, BA. 2.10)	17	6	0.353	0.142	0.617
(BA. 1.1, BA. 1.16)	13	4	0.308	0.091	0.614
(BA. 1.1, BA. 1.17.2)	75	22	0.293	0.194	0.410
(BA. 1.1, BA. 1.15.1)	11	3	0.273	0.060	0.610
(BA. 1.1, BA. 1.1.15)	10	1	0.100	0.003	0.445
(BA. 1, BA. 1.1)	132	2	0.015	0.002	0.054
(BA. 2, BA. 2.10)	12	0	0.000	0.000	0.265
(BA. 2, BA. 2.1)	11	0	0.000	0.000	0.285
(BA. 2, BA. 2.3)	10	0	0.000	0.000	0.308
(BA. 2, BA. 2.9)	14	0	0.000	0.000	0.232
(BA. 1, BA. 1.17.2)	46	0	0.000	0.000	0.077

Table 4.2: Set-level detection for lineage pairs with more than ten synthetic sequences.

We then evaluated whether the sensitivity of our method to detect synthetic recombinant sequences (regardless of whether we detect the true parental lineages) is associated with the Hamming distance between the two parental sequences of each synthetic recombinant sequence. Using logistic regression, we found a positive association between the parental Hamming distance and the sensitivity ($p = 3.31 \times 10^{-84}$ using two-sided Wald test). We estimate that for two recombinant sequences that differ by one unit in their parental Ham-

ming distances, the odds of detection is 1.11 times higher in the recombinant sequence with the higher parental Hamming distance (95% CI: [1.10, 1.12]). The relationship between the parental Hamming distance and detection probability is shown in Figure 4.2, which displays the fitted logistic curve and the associated 95% pointwise confidence band.

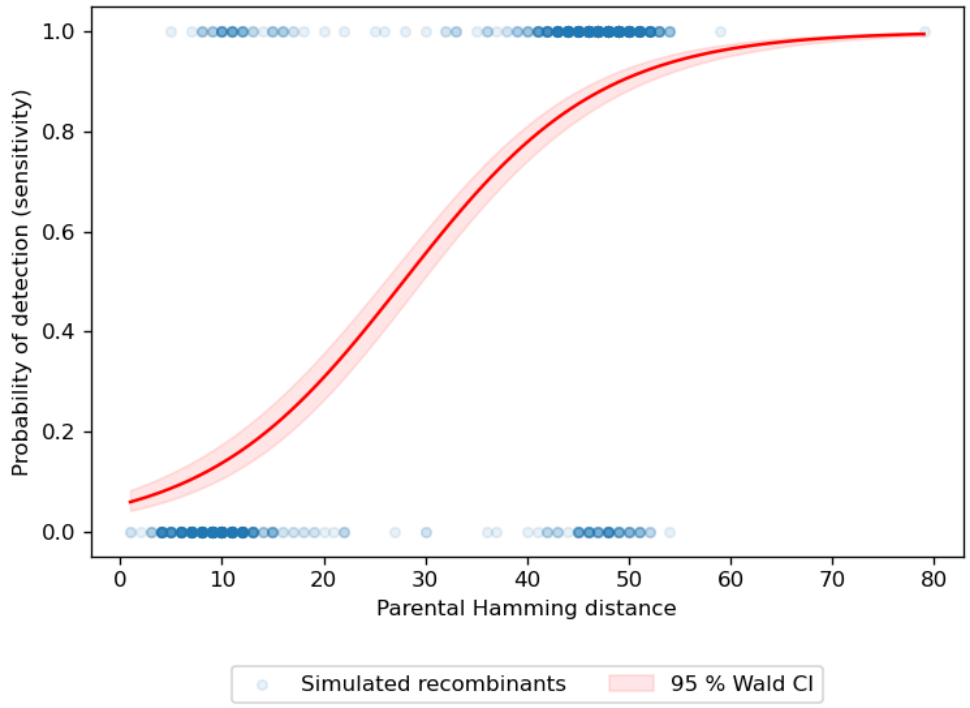

Figure 4.2: Relationship between the Hamming distance and the probability of correctly classifying a sequence as a recombinant. The red curve shows the fitted logistic regression model, and the shaded region indicates the 95% confidence interval.

To assess breakpoint resolution, we evaluated the accuracy of inferred breakpoint positions for recombinant sequences. Among sequences with one breakpoint (that had one inferred breakpoint), the average breakpoint distance was 984 nucleotides (95% CI: [785, 1,218]). For sequences with two breakpoints (that had two inferred breakpoints), the average mean distance was 1,282 nucleotides (95% CI: [1,140, 1,420]). The average mean

distance for synthetic recombinant sequences with two breakpoints was computed in two steps. For every sequence we first took the distance between each true breakpoint and its corresponding detected breakpoint and averaged those two distances within that sequence. We then averaged those per-sequence means across all sequences. It is difficult to assess breakpoint resolution for a recombinant sequence whose inferred breakpoint count does not match its true breakpoint count. A confusion matrix of inferred and true breakpoint counts for recombinant sequences is shown in Table 4.3.

True breakpoints	**Inferred breakpoints**			
	0	1	2	3
1	200	294	4	2
2	217	95	186	2

Table 4.3: Confusion matrix of inferred versus true breakpoint counts.

4.11.2 Real data analysis

We used our method to predict the local Pango lineage ancestry for 440,307 sequences across 185 test windows. Each test window spanned a period of 7 days with no gaps between successive windows. Of the 440,307 sequences across our test sets, 7,619 were detected to be recombinant sequences using our method (1.73%; 95% CI: [1.69%, 1.77%]).

In Figure 4.3, we plot the estimated frequency of recombinants in each test window (the number of detected recombinants divided by the number of sampled sequences in the test window). We see that there is an upwards trend in the estimated frequency of recombinants across time, with the maximum frequency at around 7%. However, there is more uncertainty around our estimate as we move forward in time, because the number of sequences in each test window decreases over time (see Section 4.3). Our method is only able to detect recombinants whose parents belong to different Pango lineages, and the number of Pango lineages is higher

in later test windows. Thus, the proportion of recombinants that are detectable using our method is likely higher in later test windows.

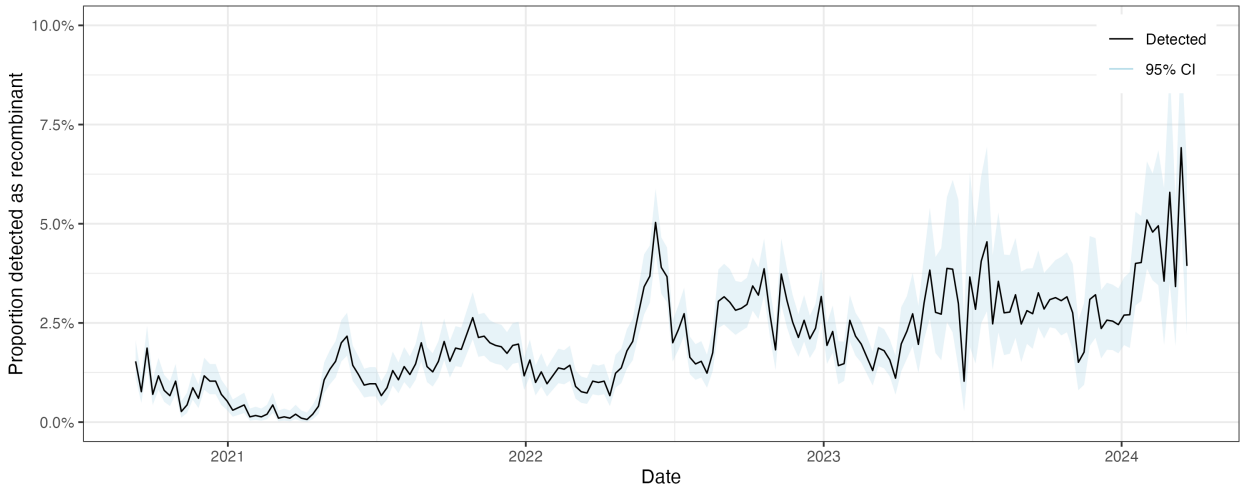

Figure 4.3: The frequency of detected recombinant sequences in each time window.

In Figure 4.4, we plot the number of detected recombinants in each test window, with colors representing different predicted numbers of parental lineages. The predicted number of parental lineages for a detected recombinant sequence is the number of distinct parental lineages in the detected local Pango lineage ancestry. The majority of detected recombinants have two predicted parental lineages. Detected recombinants had five predicted parental lineages at most. Note that the detected recombinant frequency in Figure 4.3 is not proportional to the count in Figure 4.4 because not all test windows have the same number of sampled sequences.

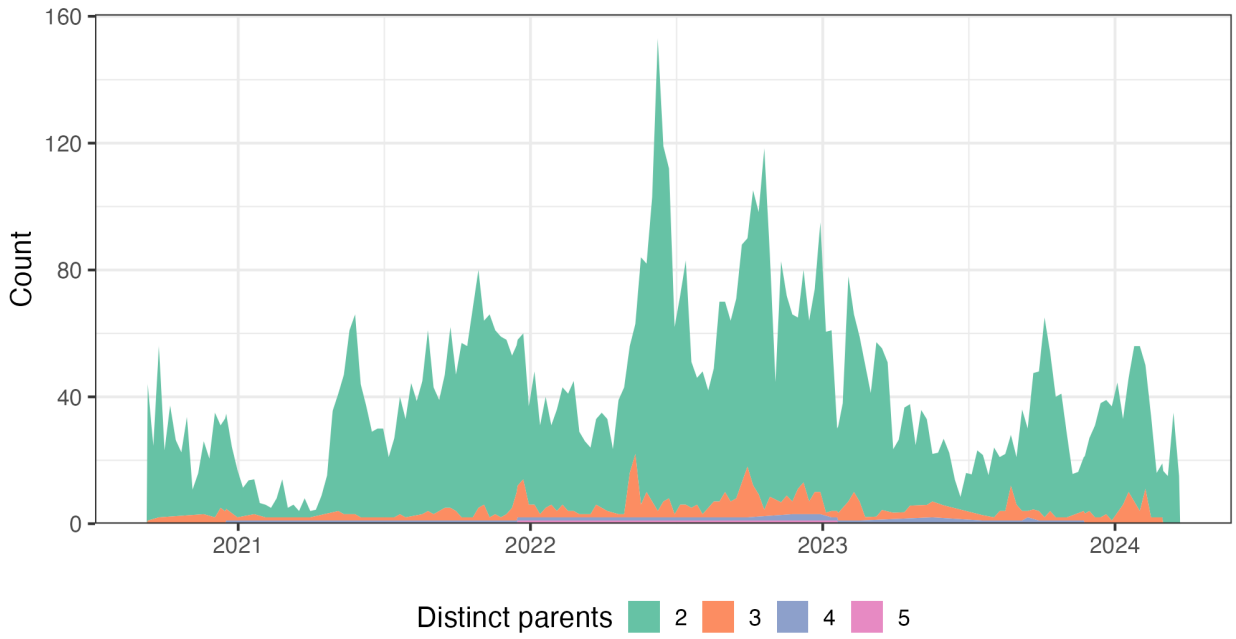

Figure 4.4: The number of detected recombinants in each test window, colored by the predicted number of parental lineages.

Because most detected recombinants only had two predicted parental lineages, we chose to focus on these cases. It is also relatively easy to group these detected recombinants by their predicted parental lineage pairs.

Taking detected recombinants with two predicted parental lineages, we plotted their counts and proportions by parental lineage combination across all test windows in Figure 4.5. In this figure, we color only the top 18 Pango lineages by proportion summed across test windows to reduce visual clutter. All remaining lineages are shown in grey. Recombinants with at least one parent from these grey lineages are also represented in grey, rather than with a combination of two colors.

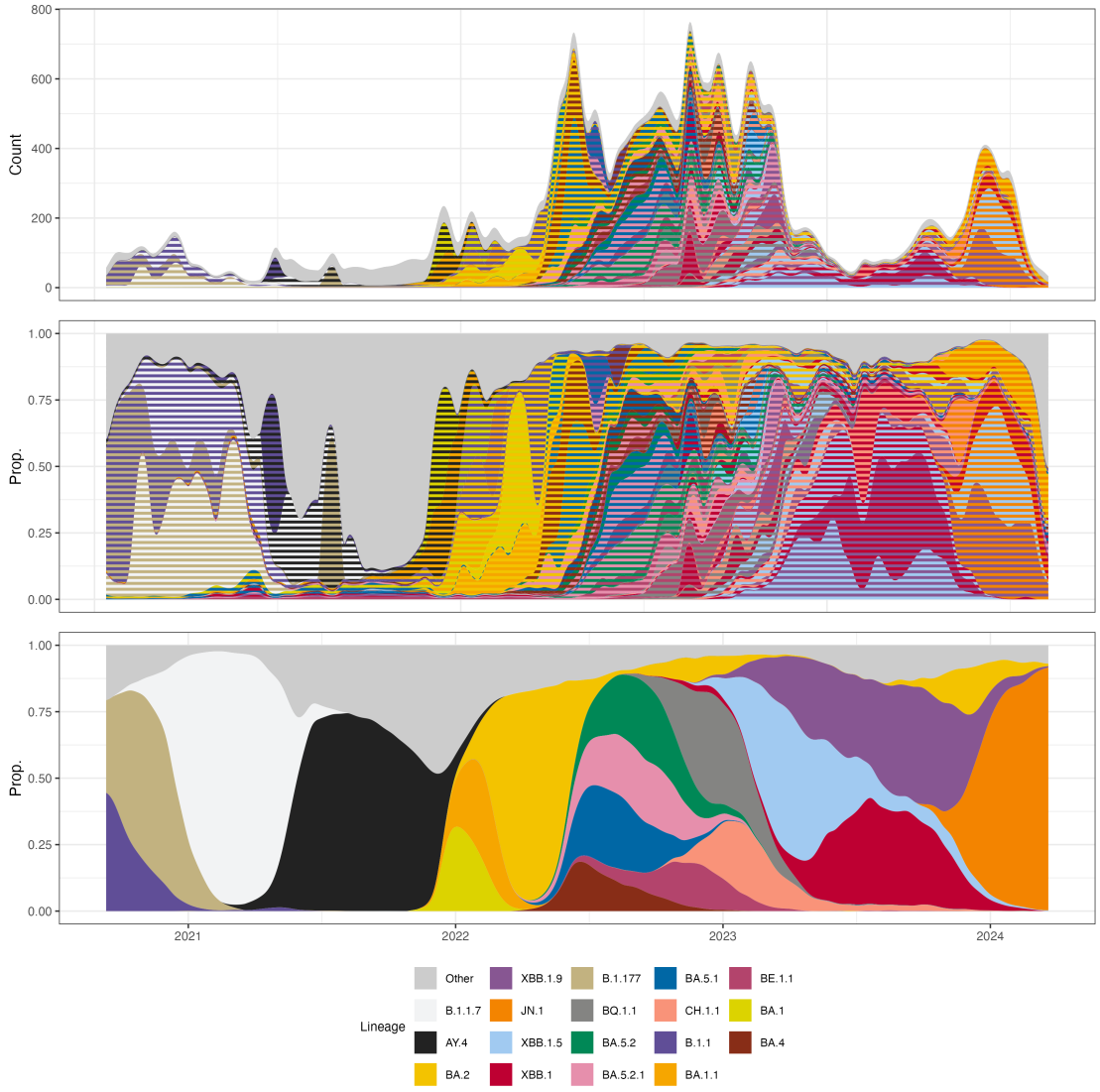

Figure 4.5: Counts and proportions of detected recombinant sequences with two predicted lineage combinations. The bottom panel shows the frequency of each Pango lineage across time. The middle panel shows the proportion of detected recombinants with each predicted parental lineage combination. To indicate lineage combinations, we use striped colors comprised of the two lineages contributing ancestry. Finally, the top panel represents the number of detected recombinants with each predicted lineage combination. To avoid using too many colors, we highlighted the top 18 Pango lineages in terms of their frequency summed across test windows. We smoothed each lineage trajectory with a cubic smoothing spline to dampen small week-to-week variations in the counts and proportions.

Finally, each recombinant sequence detected using our method has at least one detected recombination breakpoint. In Figure 4.6, we plot the genomic position of each detected breakpoint from all detected recombinant sequences. We see enrichment of recombination breakpoints in the Spike protein. We observe an additional enrichment of recombination breakpoints immediately upstream of the nucleocapsid gene.

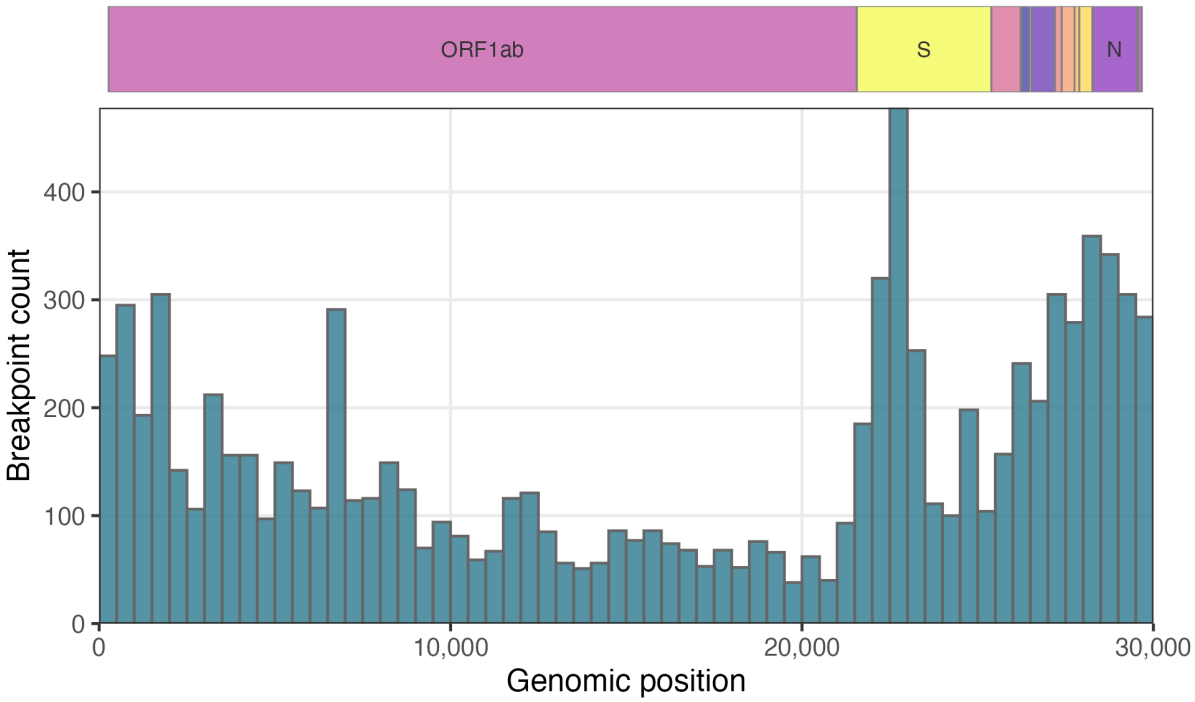

Figure 4.6: Histogram of detected recombination breakpoints.

4.12 Discussion

Recombination detection forms a key part of genomic surveillance by enabling the identification of recombinant lineages, which can grow to be the most prevalent lineage by combining mutations that jointly confer a growth advantage. Here, we develop a method to detect the local Pango lineage ancestry of query SARS-CoV-2 sequences in a test set. Our method does not depend on an existing phylogeny, nor any user-defined parameters. Furthermore, we do

not employ any strict cutoffs in the process of classifying a sequence as a recombinant.

Using synthetic sequences simulated from real SARS-CoV-2 genomes, we show that our method has moderate sensitivity (0.583; 95% CI: [0.552, 0.614]) and high specificity (0.995; 95% CI: [0.988, 0.998]) for classifying sequences as recombinant or non-recombinant. For true recombinants in the simulation, the model often did not recover the exact set of parental lineages—the correct set was recovered for only 35.6% of recombinants (95% CI: [32.6%, 38.7%]). Position-by-position ancestry assignment accuracy was still high for recombinant sequences (78.4%; 95% CI: [77.0%, 79.9%]). This suggests that the HMM frequently infers the parental lineage contributing more ancestry to the recombinant sequence as the sole contributor of ancestry.

There are several likely reasons why the model fails to recover the true parental lineages. One is that the recombinant may differ from one or both parents at only a few base pairs, making it difficult for the model to distinguish whether such differences arise from mutation or recombination. In this case, the model may incorrectly assign only one parental lineage to the recombinant sequence. Consistent with this, we observed that model sensitivity decreases as the similarity between parental lineages increases. Another likely reason is the presence of many closely related lineages circulating in the reference set (samples collected between January and March of 2022), such that the recombining region may plausibly be from multiple lineages. This period is a particularly challenging period to detect recombinants, because there are many closely related Omicron lineages that are circulating. We found that the sensitivity is highly dependent on the reference set that we generate the synthetic recombinant sequences from.

Even though the model frequently underestimates the number of recombination breakpoints for recombinant sequences in the simulation (see Table 4.3), when the correct number of breakpoints were estimated, their positions were generally accurate.

Applying our model to real SARS-CoV-2 sequences from September 2020 to March 2024, we found 7,619 recombinant sequences across 440,307 sequences (1.73%; 95% CI: [1.69%, 1.77%]). However, considering that our model has limited sensitivity, this is not a good

estimate of the overall prevalence of recombinants. An estimator for the prevalence dependent on the sensitivity and specificity of a test has previously been derived by Rogan and Gladen [54]. Assuming that the sensitivity and specificity estimates can be generalized to the real data, we estimate using the Rogan-Gladen estimator that the prevalence is 2.13% (95% CI: [1.30%, 2.81%]). The 95% confidence interval was derived by independently sampling from the binomial distributions for the observed positive fraction (7,619 of 440,307), the sensitivity (583 of 1,000), and the specificity (995 of 1,000), recomputing the Rogan–Gladen estimate at each draw and taking the 2.5th and 97.5th percentiles of the resulting distribution. The prevalence estimate and confidence interval should be interpreted with caution because we do not necessarily expect the sensitivity and specificity estimates to generalize well to the real data analysis. Furthermore, our 95% confidence interval assumes that our sensitivity, specificity, and positive proportion estimates are independent, which may not be reasonable given that synthetic sequences were generated using real sequences which may overlap with the sequences that we are using to obtain the observed positive fraction. In the future, it may be useful to find gold-standard sets of recombinants and non-recombinants in the real data that we can use to evaluate the sensitivity and specificity of the model.

Using RIPPLES, Turakhia et al. found 2.7% of sampled genomes inferred to have detectable recombinant ancestry [65]. This proportion is inside our 95% confidence interval for the prevalence of recombinants, but we should keep in mind that Turakhia et al. only analyze sequences up to May 2021, before the emergence of XBB. Furthermore, the detectable set of recombinants is not the same using our method and RIPPLES, because our method cannot detect inter-lineage recombination. Another notable difference is that our method will not classify a sequence as having recombinant ancestry if the recombinant lineage of the sequence is defined as a lineage in the reference set.

We further looked at the locations of detected breakpoints from the real data and found enrichment of recombination breakpoints in the Spike protein. We observe an additional enrichment of recombination breakpoints immediately upstream of the nucleocapsid gene. This is consistent with previous work on recombination hotspots in SARS-CoV-2 and related

coronaviruses [39]. However, it may be important to account for the fact that breakpoints at the ends of the genome may be less detectable, because the resulting recombinant may not differ by many mutations to one of its parental sequences.

In future work, we will evaluate the detection model on simulations that incorporate time-varying lineage frequencies, co-infection rates, and sampling rates to better mirror surveillance data. Using these simulations, we can better quantify how sensitivity varies with parental similarity and breakpoint location, and refine prevalence estimates by allowing sensitivity and specificity to vary over time. We eventually want to see whether the recombinant sequences we detect in the real data and their parental lineage compositions are consistent with past infection dynamics.

Chapter 5

CONCLUSION AND FUTURE DIRECTIONS

In this dissertation, we looked at three projects involving whole-genome sequencing data from humans and SARS-CoV-2. In the first project, we formulated a statistical model to jointly estimate genotype error and uncalled deletion rates from the UK Biobank whole genome sequence data. We found that uncalled deletions may be responsible for many of the genotype errors at SNVs not overlapping called deletions or other structural variants. Specifically, we estimate that 77% of the genotype errors at these markers are attributable to uncalled deletions (90% CI: [73%, 88%]).

A natural extension of our method would include the estimation of genotype error rates for SNVs that overlap with called deletions or structural variants, because we expect genotype error rates to be higher for these SNVs. However, this would be more challenging as we would need to model more rates (e.g., the rate of miscalling the major allele as a deletion).

In our second project, we developed another statistical model to infer the mean length of gene conversion tracts detected from the UK Biobank whole autosome data. Using a mixture of two geometric components for the tract length distribution, we estimate that the smaller component has a mean of 16.9 bp (95% CI: [16.4, 17.0]), and the larger component has a mean of 724.7 bp (95% CI: [720.1, 728.7]). We further estimate the proportion of tracts in the second component to be 0.00525 (95% CI: [0.005, 0.00525]). After stratifying by crossover-hotspot overlap, we infer that tracts whose midpoints lie within crossover hotspots are, on average, longer than the remaining tracts.

One natural extension of this method would consider additional components of the geometric distribution. Exploring such models is computationally challenging due to the added number of parameters. However, if we are able to reliably detect longer gene conversion

tracts, and these longer tracts do not fit well to the two-component model, we may want to consider more components in our model.

In our third project, we developed a method to detect recombinant SARS-CoV-2 sequences in test sets comprised of sequences collected within a span of a few days to a week. Our method uses a HMM that jointly models the latent sequence of local Pango lineage ancestry and the nucleotide sequence of the query sequence, while accounting for mutations and genotype errors. Applying our model to real SARS-CoV-2 sequences from September 2020 to March 2024, we found 7,619 recombinant sequences across 440,307 sequences (1.73%; 95% CI: [1.69%, 1.77%]).

In this project, we plan to work on another simulation study where we incorporate epidemiological and surveillance processes such as time-varying lineage frequencies, co-infection rates, and sampling rates. Our goal is to evaluate the method under conditions more closely aligned with observed surveillance data. We then want to obtain real-data prevalence estimates corrected for misclassification using sensitivity and specificity evaluated under these more realistic conditions.

In our last two projects, we made use of specific evolutionary processes in humans and SARS-CoV-2 to uncover unobservable aspects of evolution. For example, it is possible to use identity-by-descent segments to detect past gene conversions in humans, because meiotic recombination breaks up shared segments at a predictable rate across many generations of meioses. In the final project on recombination detection in SARS-CoV-2, we adjusted for recent de novo mutations on query sequences, given the rapid mutation rate of the virus.

As we discussed in the introduction, certain statistical tools lend themselves more to genomic data in humans or SARS-CoV-2, but certain frameworks, like the Li and Stephens model used commonly for local ancestry inference [38], are applicable to both human and SARS-CoV-2 genomic data. However, it is still necessary to tailor these frameworks to better fit the specific species under investigation. As a further example, certain viruses like influenza undergo reassortment, where the number of possible recombination breakpoints is limited to segment boundaries, and analyses can be tailored to account for this.

BIBLIOGRAPHY

[1] Xbb.1.5 updated risk assessment. Technical report, World Health Organization, June 2023. URL https://www.who.int/docs/default-source/coronaviruse/20230620xbb.1.5.pdf. Accessed July 24, 2025.

[2] Hilliker AJ, Harauz G, Reaume AG, Gray M, Clark SH, and Chovnick A. Meiotic gene conversion tract length distribution within the rosy locus of drosophila melanogaster. *Genetics*, 137:1019–1026, 1994. doi: 10.1093/genetics/137.4.1019.

[3] Jeffreys AJ and May CA. Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nat Genet*, 36:151–156, 2004. doi: 10.1038/ng1287.

[4] Williams AL, Genovese G, Dyer T, Altemose N, Truax K, Jun G, et al. Non-crossover gene conversions show strong gc bias and unexpected clustering in humans. *eLife*, 4: e04637, 2015. doi: 10.7554/eLife.04637.

[5] Tommaso Alfonsi, Anna Bernasconi, Matteo Chiara, and Stefano Ceri. Data-driven recombination detection in viral genomes. *Nature Communications*, 15(1):3313, April 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-47464-5. URL https://www.nature.com/articles/s41467-024-47464-5. Publisher: Nature Publishing Group.

[6] Browning BL and Browning SR. Genotype error biases trio-based estimates of haplotype phase accuracy. *Am J Hum Genet*, 109(6):1016–1025, 2022. doi: 10.1016/j.ajhg.2022.04.019.

[7] Browning BL and Browning SR. Statistical phasing of 150,119 sequenced genomes in the uk biobank. *Am J Hum Genet*, 110(1):161–165, 2023. doi: 10.1016/j.ajhg.2022.11.008.

[8] Browning BL, Tian X, Zhou Y, and Browning SR. Fast two-stage phasing of large-scale sequence data. *The American Journal of Human Genetics*, 108:1880–1890, 2021. doi: 10.1016/j.ajhg.2021.08.005.

[9] Halldorsson BV, Hardarson MT, Kehr B, Styrkarsdottir U, Gylfason A, Thorleifsson G, et al. The rate of meiotic gene conversion varies by sex and age. *Nat Genet*, 48: 1377–1384, 2016. doi: 10.1038/ng.3669.

[10] Halldorsson BV, Palsson G, Stefansson OA, Jonsson H, Hardarson MT, Eggertsson HP, et al. Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science*, 363:eaau1043, 2019. doi: 10.1126/science.aau1043.

[11] Halldorsson BV, Eggertsson HP, Moore KHS, Hauswedell H, Eiriksson O, Ulfarsson MO, et al. The sequences of 150,119 genomes in the uk biobank. *Nature*, 607(7920): 732–740, 2022. doi: 10.1038/s41586-022-04965-x.

[12] Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018. doi: 10.1038/s41586-018-0579-z.

[13] Laurie CC, Doheny KF, Mirel DB, Pugh EW, Bierut LJ, Bhangale T, et al. Quality control and quality assurance in genotypic data for genome-wide association studies. *Genet Epidemiol*, 34(6):591–602, 2010. doi: 10.1002/gepi.20516.

[14] Brianna Sierra Chrisman, Kelley Paskov, Nate. Stockham, Kevin Tabatabaei, Jae-Yoon Jung, Peter Washington, Maya Varma, Min Woo Sun, Sepideh Maleki, and Dennis P. Wall. Indels in SARS-CoV-2 occur at template-switching hotspots. *BioData Mining*, 14:20, March 2021. ISSN 1756-0381. doi: 10.1186/s13040-021-00251-0. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7980745/.

[15] Mirko Cortese, Ji-Young Lee, Berati Cerikan, Christopher J. Neufeldt, Viola M. J. Oorschot, Sebastian Köhrer, Julian Hennies, Nicole L. Schieber, Paolo Ronchi, Giulia

Mizzon, Inés Romero-Brey, Rachel Santarella-Mellwig, Martin Schorb, Mandy Boermel, Karel Mocaer, Marianne S. Beckwith, Rachel M. Templin, Viktoriia Gross, Constantin Pape, Christian Tischer, Jamie Frankish, Natalie K. Horvat, Vibor Laketa, Megan Stanifer, Steeve Boulant, Alessia Ruggieri, Laurent Chatel-Chaix, Yannick Schwab, and Ralf Bartenschlager. Integrative Imaging Reveals SARS-CoV-2-Induced Reshaping of Subcellular Morphologies. *Cell Host & Microbe*, 28(6):853–866.e5, December 2020. ISSN 1931-3128, 1934-6069. doi: 10.1016/j.chom.2020.11.003. URL https://www.cell.com/cell-host-microbe/abstract/S1931-3128(20)30620-X. Publisher: Elsevier.

[16] Cartwright DA, Troggio M, Velasco R, and Gutin A. Genetic mapping in the presence of genotyping errors. *Genetics*, 176(4):2521–2527, 2007. doi: 10.1534/genetics.106.063982.

[17] A. J. Drummond, A. Rambaut, B. Shapiro, and O. G. Pybus. Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences. *Molecular Biology and Evolution*, 22(5):1185–1192, May 2005. ISSN 0737-4038. doi: 10.1093/molbev/msi103. URL https://doi.org/10.1093/molbev/msi103.

[18] Betran E, Rozas J, Navarro A, and Barbadilla A. The estimation of the number and the length distribution of gene conversion tracts from population dna sequence data. *Genetics*, 146:89–99, 1997.

[19] Sobel E, Papp JC, and Lange K. Detection and integration of genotyping errors in statistical genetics. *Am J Hum Genet*, 70(2):496–508, 2002. doi: 10.1086/338920.

[20] Baumdicker F, Bisschop G, Goldstein D, Gower G, Ragsdale AP, Tsambos G, et al. Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*, 220:iyab229, 2022. doi: 10.1093/genetics/iyab229.

[21] Palsson G, Hardarson MT, Jonsson H, Steinthorsdottir V, Stefansson OA, Eggertsson HP, et al. Complete human recombination maps. *Nature*, 639:700–707, 2025. doi: 10.1038/s41586-024-08450-5.

[22] Abecasis GR, Cherny SS, and Cardon LR. The impact of genotyping error on family-based analysis of quantitative traits. *Eur J Hum Genet*, 9(2):130–134, 2001. doi: 10.1038/sj.ejhg.5200594.

[23] Akaike H. Information theory and an extension of the maximum likelihood principle. In *Selected papers of Hirotugu Akaike*, pages 199–213. Springer, 1998.

[24] James Hadfield, Colin Megill, Sidney M Bell, John Huddleston, Barney Potter, Charlton Callender, Pavel Sagulenko, Trevor Bedford, and Richard A Neher. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*, 34(23):4121–4123, 2018.

[25] Neil Hunter. Meiotic Recombination: The Essence of Heredity. *Cold Spring Harbor Perspectives in Biology*, 7(12):a016618, December 2015. ISSN 1943-0264. doi: 10.1101/cshperspect.a016618. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4665078/.

[26] Saunders IW, Brohede J, and Hannan GN. Estimating genotyping error rates from mendelian errors in snp array genotypes and their impact on inference. *Genomics*, 90 (3):291–296, 2007. doi: 10.1016/j.ygeno.2007.05.011.

[27] Douglas JA, Boehnke M, and Lange K. A multipoint method for detecting genotyping errors and mutations in sibling-pair linkage data. *Am J Hum Genet*, 66(4):1287–1297, 2000. doi: 10.1086/302861.

[28] Douglas JA, Skol AD, and Boehnke M. Probability of detection of genotyping errors and mutations as inheritance inconsistencies in nuclear-family data. *Am J Hum Genet*, 70(2):487–495, 2002. doi: 10.1086/338919.

[29] Wall JD, Tang LF, Zerbe B, Kvale MN, Kwok PY, Schaefer C, et al. Estimating genotype error rates from high-coverage next-generation sequence data. *Genome Res*, 24(11):1734–1739, 2014. doi: 10.1101/gr.168393.113.

[30] Wall JD, Robinson JA, and Cox LA. High-resolution estimates of crossover and non-crossover recombination from a captive baboon colony. *Genome Biology and Evolution*, 14:evac040, 2022. doi: 10.1093/gbe/evac040.

[31] Lebrec JJ, Putter H, Houwing-Duistermaat JJ, and van Houwelingen HC. Influence of genotyping error in linkage mapping for complex traits–an analytic study. *BMC Genet*, 9:57, 2008. doi: 10.1186/1471-2156-9-57.

[32] Hao K, Li C, Rosenow C, and Hung Wong W. Estimation of genotype error rate using samples with pedigree information–an application on the genechip mapping 10k array. *Genomics*, 84(4):623–630, 2004. doi: 10.1016/j.ygeno.2004.05.003.

[33] Sergei L. Kosakovsky Pond, David Posada, Michael B. Gravenor, Christopher H. Woelk, and Simon D. W. Frost. GARD: a genetic algorithm for recombination detection. *Bioinformatics (Oxford, England)*, 22(24):3096–3098, December 2006. ISSN 1367-4811. doi: 10.1093/bioinformatics/btl474.

[34] Hou L, Sun N, Mane S, Sayward F, Rajeevan N, Cheung KH, et al. Impact of genotyping errors on statistical power of association tests in genomic analyses: A case study. *Genet Epidemiol*, 41(2):152–162, 2017. doi: 10.1002/gepi.22027.

[35] Odenthal-Hesse L, Berg IL, Veselis A, Jeffreys AJ, and May CA. Transmission distortion affecting human noncrossover but not crossover recombination: A hidden source of meiotic drive. *PLOS Genetics*, 10:e1004106, 2014. doi: 10.1371/journal.pgen.1004106.

[36] Matthew C. LaFave and Jeff Sekelsky. Mitotic Recombination: Why? When? How? Where? *PLoS Genetics*, 5(3):e1000411, March 2009. ISSN 1553-7390. doi: 10.1371/journal.pgen.1000411. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2648873/.

[37] Ha Minh Lam, Oliver Ratmann, and Maciej F Boni. Improved Algorithmic Complexity for the 3SEQ Recombination Detection Algorithm. *Molecular Biology and Evolution*,

35(1):247–251, January 2018. ISSN 0737-4038. doi: 10.1093/molbev/msx263. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5850291/.

[38] Na Li and Matthew Stephens. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165(4):2213–2233, December 2003. ISSN 0016-6731. doi: 10.1093/genetics/165.4.2213.

[39] Spyros Lytras, Joseph Hughes, Darren Martin, Phillip Swanepoel, Arné de Klerk, Rentia Lourens, Sergei L Kosakovsky Pond, Wei Xia, Xiaowei Jiang, and David L Robertson. Exploring the Natural Origins of SARS-CoV-2 in the Light of Recombination. *Genome Biology and Evolution*, 14(2):evac018, February 2022. ISSN 1759-6653. doi: 10.1093/gbe/evac018. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8882382/.

[40] Peter V. Markov, Mahan Ghafari, Martin Beer, Katrina Lythgoe, Peter Simmonds, Nikolaos I. Stilianakis, and Aris Katzourakis. The evolution of SARS-CoV-2. *Nature Reviews Microbiology*, 21(6):361–379, June 2023. ISSN 1740-1534. doi: 10.1038/s41579-023-00878-2. URL https://www.nature.com/articles/s41579-023-00878-2. Publisher: Nature Publishing Group.

[41] Darren P. Martin, Ben Murrell, Michael Golden, Arjun Khoosal, and Brejnev Muhire. RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evolution*, 1(1):vev003, March 2015. ISSN 2057-1577. doi: 10.1093/ve/vev003. URL https://doi.org/10.1093/ve/vev003.

[42] N Masaki and SR Browning. Modeling the length distribution of gene conversion tracts in humans from the uk biobank sequence data. *bioRxiv*, 2025. doi: 10.1101/2024.12.30.630818. URL https://www.biorxiv.org/content/early/2025/05/28/2024.12.30.630818.

[43] N Masaki, SR Browning, and BL Browning. Simultaneous estimation of genotype error and uncalled deletion rates in whole genome sequence data. *PLOS Genetics*, 20(5):

e1011297, 2024. doi: 10.1371/journal.pgen.1011297. URL https://doi.org/10.1371/journal.pgen.1011297.

[44] McMahill MS, Sham CW, and Bishop DK. Synthesis-dependent strand annealing in meiosis. *PLoS Biol*, 5:e299, 2007. doi: 10.1371/journal.pbio.0050299.

[45] Hardarson MT, Palsson G, and Halldorsson BV. Ncourd: modelling length distributions of nco events and gene conversion tracts. *Bioinformatics*, 39:btad485, 2023. doi: 10.1093/bioinformatics/btad485.

[46] Natacha S. Ogando, Tim J. Dalebout, Jessika C. Zevenhoven-Dobbe, Ronald W.A.L. Limpens, Yvonne van der Meer, Leon Caly, Julian Druce, Jutte J. C. de Vries, Marjolein Kikkert, Montserrat Bárcena, Igor Sidorov, and Eric J. Snijder. SARS-coronavirus-2 replication in Vero E6 cells: replication kinetics, rapid adaptation and cytopathology. *The Journal of General Virology*, 101(9):925–940, September 2020. ISSN 0022-1317. doi: 10.1099/jgv.0.001453. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7654748/.

[47] Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nat Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

[48] Pier Francesco Palamara, Todd Lencz, Ariel Darvasi, and Itsik Pe'er. Length Distributions of Identity by Descent Reveal Fine-Scale Demographic History. *The American Journal of Human Genetics*, 91(5):809–822, November 2012. ISSN 0002-9297, 1537-6605. doi: 10.1016/j.ajhg.2012.08.030. URL https://www.cell.com/ajhg/abstract/S0002-9297(12)00472-7. Publisher: Elsevier.

[49] Palamara PF, Francioli LC, Wilton PR, Genovese G, Gusev A, Finucane HK, et al. Leveraging distant relatedness to quantify human mutation and gene-conversion rates.

The American Journal of Human Genetics, 97:775–789, 2015. doi: 10.1016/j.ajhg.2015.10.006.

[50] D. Posada and K. A. Crandall. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proceedings of the National Academy of Sciences of the United States of America*, 98(24):13757–13762, November 2001. ISSN 0027-8424. doi: 10.1073/pnas.241370698.

[51] L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, February 1989. ISSN 1558-2256. doi: 10.1109/5.18626. URL https://ieeexplore.ieee.org/document/18626. Conference Name: Proceedings of the IEEE.

[52] Wang RJ, Radivojac P, and Hahn MW. Distinct error rates for reference and non-reference genotypes estimated by pedigree analysis. *Genetics*, 217(1):1–10, 2021. doi: 10.1093/genetics/iyaa014.

[53] Cornelius Roemer. pango_aliasor: Utility to alias and dealias pango lineages. https://github.com/corneliusroemer/pango_aliasor, 2023.

[54] W. J. Rogan and B. Gladen. Estimating prevalence from the results of a screening test. *American Journal of Epidemiology*, 107(1):71–76, January 1978. ISSN 0002-9262. doi: 10.1093/oxfordjournals.aje.a112510.

[55] M. O. Salminen, J. K. Carr, D. S. Burke, and F. E. McCutchan. Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning. *AIDS research and human retroviruses*, 11(11):1423–1425, November 1995. ISSN 0889-2229. doi: 10.1089/aid.1995.11.1423.

[56] Stéphane Samson, Étienne Lord, and Vladimir Makarenkov. SimPlot++: a Python application for representing sequence similarity and detecting recombination. *Bioinformat-*

ics, 38(11):3118–3120, May 2022. ISSN 1367-4803. doi: 10.1093/bioinformatics/btac287. URL https://doi.org/10.1093/bioinformatics/btac287.

[57] S Sawyer. Statistical tests for detecting gene conversion. *Molecular Biology and Evolution*, 6(5):526–538, September 1989. ISSN 0737-4038. doi: 10.1093/oxfordjournals.molbev.a040567. URL https://doi.org/10.1093/oxfordjournals.molbev.a040567.

[58] Fullerton SM, Bernardo Carvalho A, and Clark AG. Local rates of recombination are positively correlated with gc content in the human genome. *Molecular Biology and Evolution*, 18:1139–1142, 2001. doi: 10.1093/oxfordjournals.molbev.a003886.

[59] Browning SR and Browning BL. Biobank-scale inference of multi-individual identity by descent and gene conversion. *The American Journal of Human Genetics*, 111:691–700, 2024. doi: 10.1016/j.ajhg.2024.02.015.

[60] Tomokazu Tamura, Jumpei Ito, Keiya Uriu, Jiri Zahradnik, Izumi Kida, Yuki Anraku, Hesham Nasser, Maya Shofa, Yoshitaka Oda, Spyros Lytras, Naganori Nao, Yukari Itakura, Sayaka Deguchi, Rigel Suzuki, Lei Wang, MST Monira Begum, Shunsuke Kita, Hisano Yajima, Jiei Sasaki, Kaori Sasaki-Tabata, Ryo Shimizu, Masumi Tsuda, Yusuke Kosugi, Shigeru Fujita, Lin Pan, Daniel Sauter, Kumiko Yoshimatsu, Saori Suzuki, Hiroyuki Asakura, Mami Nagashima, Kenji Sadamasu, Kazuhisa Yoshimura, Yuki Yamamoto, Tetsuharu Nagamoto, Gideon Schreiber, Katsumi Maenaka, Takao Hashiguchi, Terumasa Ikeda, Takasuke Fukuhara, Akatsuki Saito, Shinya Tanaka, Keita Matsuno, Kazuo Takayama, and Kei Sato. Virological characteristics of the SARS-CoV-2 XBB variant derived from recombination of two Omicron subvariants. *Nature Communications*, 14:2800, May 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-38435-3. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10187524/.

[61] CDC COVID-19 Response Team. Genomic surveillance for sars-cov-2 variants: Circulation of omicron lineages — united states, january 2022–may 2023. *MMWR. Morbidity*

and Mortality Weekly Report, 72(24):649–654, 2023. URL https://www.cdc.gov/mmwr/volumes/72/wr/mm7224a2.htm. Peak XBB.1.5 prevalence 84.1% by April 1, 2023.

[62] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2006.

[63] Xiaowen Tian, Ruoyi Cai, and Sharon R. Browning. Estimating the genome-wide mutation rate from thousands of unrelated individuals. *The American Journal of Human Genetics*, 109(12):2178–2184, December 2022. ISSN 0002-9297. doi: 10.1016/j.ajhg.2022.10.015. URL https://www.sciencedirect.com/science/article/pii/S0002929722004633.

[64] Pauline Trémeaux, Justine Latour, Noémie Ranger, Vénicia Ferrer, Agnès Harter, Romain Carcenac, Pauline Boyer, Sofia Demmou, Florence Nicot, Stéphanie Raymond, and Jacques Izopet. SARS-CoV-2 Co-Infections and Recombinations Identified by Long-Read Single-Molecule Real-Time Sequencing. *Microbiology Spectrum*, 11(4):e00493–23, June 2023. doi: 10.1128/spectrum.00493-23. URL https://journals.asm.org/doi/10.1128/spectrum.00493-23. Publisher: American Society for Microbiology.

[65] Yatish Turakhia, Bryan Thornlow, Angie Hinrichs, Jakob McBroome, Nicolas Ayala, Cheng Ye, Kyle Smith, Nicola De Maio, David Haussler, Robert Lanfear, and Russell Corbett-Detig. Pandemic-scale phylogenomics reveals the SARS-CoV-2 recombination landscape. *Nature*, 609(7929):994–997, September 2022. ISSN 1476-4687. doi: 10.1038/s41586-022-05189-9. URL https://www.nature.com/articles/s41586-022-05189-9. Publisher: Nature Publishing Group.

[66] Ales Varabyou, Christopher Pockrandt, Steven L. Salzberg, and Mihaela Pertea. Rapid detection of inter-clade recombination in SARS-CoV-2 with Bolotie. *Genetics*, 218(3):iyab074, July 2021. ISSN 1943-2631. doi: 10.1093/genetics/iyab074.

[67] Heather L. Wells, Cassandra M. Bonavita, Isamara Navarrete-Macias, Blake Vilchez,

Angela L. Rasmussen, and Simon J. Anthony. The coronavirus recombination pathway. *Cell Host & Microbe*, 31(6):874–889, June 2023. ISSN 1931-3128, 1934-6069. doi: 10.1016/j.chom.2023.05.003. URL https://www.cell.com/cell-host-microbe/abstract/S1931-3128(23)00196-8. Publisher: Elsevier.

[68] Tian X, Cai R, and Browning SR. Estimating the genome-wide mutation rate from thousands of unrelated individuals. *The American Journal of Human Genetics*, 109: 2178–2184, 2022. doi: 10.1016/j.ajhg.2022.10.015.

Appendix A

ESTIMATING PARENTAL GENOTYPE FREQUENCIES

Our estimates for the true parental genotype frequencies $\hat{\Pi}_I^{i,j}$ in MAF interval I are derived from the estimated pre-deletion parental genotype frequencies $\hat{\Pi}_{\text{pre},I}^{i,j}$ and the uncalled deletion rate Γ_I. Recall that $\hat{\Pi}_{\text{pre},I}^{i,j}$ is simply the observed proportion of each parental genotype pair in MAF interval I after we exclude Mendelian-inconsistent trios.

Note that there are only six possible pre-deletion parental genotype pairs (AA–AA, AA–AB, AA–BB, AB–AB, AB–BB, BB–BB), but there are 15 possible true parental genotype pairs once deleted alleles (D) are allowed (AA–AA, AA–AB, AA–BB, AA–AD, AA–BD, AB–AB, AB–BB, AB–AD, AB–BD, BB–BB, BB–AD, BB–BD, AD–AD, AD–BD, BD–BD). We estimate $\hat{\Pi}_I^{i,j}$ by enumerating every way these 15 pairs can arise through the replacement of one or more alleles by deletions in the six pre-deletion pairs. For example, the pair AA–AD can come from AA–AA if any of the four alleles is deleted, or from AA–AB if the B allele is deleted. Hence

$$\hat{\Pi}_I^{0,3} = 4\,\hat{\Pi}_{\text{pre},I}^{0,0}\,\Gamma_I(1-2\Gamma_I) + \hat{\Pi}_{\text{pre},I}^{0,1}\,\Gamma_I(1-2\Gamma_I).$$

All adjusted frequencies are obtained in the same way:

$$\begin{aligned}
\hat{\Pi}_I^{0,0} &= \hat{\Pi}_{\text{pre},I}^{0,0}(1-2\Gamma_I)^2,\\
\hat{\Pi}_I^{0,1} &= \hat{\Pi}_{\text{pre},I}^{0,1}(1-2\Gamma_I)^2,\\
\hat{\Pi}_I^{0,2} &= \hat{\Pi}_{\text{pre},I}^{0,2}(1-2\Gamma_I)^2,\\
\hat{\Pi}_I^{0,3} &= 4\hat{\Pi}_{\text{pre},I}^{0,0}\Gamma_I(1-2\Gamma_I) + \hat{\Pi}_{\text{pre},I}^{0,1}\Gamma_I(1-2\Gamma_I),\\
\hat{\Pi}_I^{0,4} &= 2\hat{\Pi}_{\text{pre},I}^{0,2}\Gamma_I(1-2\Gamma_I) + \hat{\Pi}_{\text{pre},I}^{0,1}\Gamma_I(1-2\Gamma_I),\\
\hat{\Pi}_I^{1,1} &= \hat{\Pi}_{\text{pre},I}^{1,1}(1-2\Gamma_I)^2,\\
\hat{\Pi}_I^{1,2} &= \hat{\Pi}_{\text{pre},I}^{1,2}(1-2\Gamma_I)^2,
\end{aligned}$$

$$\hat{\Pi}_I^{1,3} = 2\hat{\Pi}_{\mathrm{pre},I}^{0,1}\Gamma_I(1-2\Gamma_I) + 2\hat{\Pi}_{\mathrm{pre},I}^{1,1}\Gamma_I(1-2\Gamma_I),$$

$$\hat{\Pi}_I^{1,4} = 2\hat{\Pi}_{\mathrm{pre},I}^{1,2}\Gamma_I(1-2\Gamma_I) + 2\hat{\Pi}_{\mathrm{pre},I}^{1,1}\Gamma_I(1-2\Gamma_I),$$

$$\hat{\Pi}_I^{2,2} = \hat{\Pi}_{\mathrm{pre},I}^{2,2}(1-2\Gamma_I)^2,$$

$$\hat{\Pi}_I^{2,3} = 2\hat{\Pi}_{\mathrm{pre},I}^{0,2}\Gamma_I(1-2\Gamma_I) + \hat{\Pi}_{\mathrm{pre},I}^{1,2}\Gamma_I(1-2\Gamma_I),$$

$$\hat{\Pi}_I^{2,4} = 4\hat{\Pi}_{\mathrm{pre},I}^{2,2}\Gamma_I(1-2\Gamma_I) + \hat{\Pi}_{\mathrm{pre},I}^{1,2}\Gamma_I(1-2\Gamma_I),$$

$$\hat{\Pi}_I^{3,3} = 4\hat{\Pi}_{\mathrm{pre},I}^{0,0}\Gamma_I^2 + 2\hat{\Pi}_{\mathrm{pre},I}^{0,2}\Gamma_I^2 + \hat{\Pi}_{\mathrm{pre},I}^{1,1}\Gamma_I^2,$$

$$\hat{\Pi}_I^{3,4} = 4\hat{\Pi}_{\mathrm{pre},I}^{0,2}\Gamma_I^2 + 2\hat{\Pi}_{\mathrm{pre},I}^{0,1}\Gamma_I^2 + 2\hat{\Pi}_{\mathrm{pre},I}^{1,2}\Gamma_I^2 + 2\hat{\Pi}_{\mathrm{pre},I}^{1,1}\Gamma_I^2,$$

$$\hat{\Pi}_I^{4,4} = 4\hat{\Pi}_{\mathrm{pre},I}^{2,2}\Gamma_I^2 + 2\hat{\Pi}_{\mathrm{pre},I}^{1,2}\Gamma_I^2 + \hat{\Pi}_{\mathrm{pre},I}^{1,1}\Gamma_I^2.$$

Appendix B

DERIVATIONS FOR THE MARGINAL DISTRIBUTION OF THE OBSERVED TRACT LENGTH

B.1 Deriving a maximum likelihood estimator for ϕ under the constant ψ model

In this section, we suppose that ψ is constant across tracts l_j. We have

$$P(L = l) = \sum_{n=\max(1,l)}^{\infty} P(L = l \mid N = n)\, P(N = n).$$

We assume N is geometric with mean ϕ and the distribution of $L \mid N$ in Section 3.6. Letting $\lambda = 1/\phi$,

$$P(L = l) \;=\; \sum_{n=l}^{\infty} P\big(L = l \mid N = n\big)\, P(N = n) \;=\; \begin{cases} \dfrac{\lambda(1-\psi)}{\lambda+\psi-\lambda\psi}, & l = 0, \\ \dfrac{\lambda\psi}{(\lambda+\psi-\lambda\psi)^2}, & l = 1, \\ \dfrac{\lambda(1-\lambda)^{l-1}\psi^2}{(\lambda+\psi-\lambda\psi)^2}, & l \geq 2. \end{cases}$$

We can easily condition on $L \neq 0$ to obtain

$$P(L = 1 \mid L \neq 0) = \frac{\lambda\psi}{(\lambda\psi-\lambda-\psi)^2} \cdot \frac{\lambda+\psi-\lambda\psi}{\psi} = \frac{\lambda}{\lambda+\psi-\lambda\psi},$$
$$P(L = l \mid L \neq 0) = \frac{\lambda(1-\lambda)^{l-1}\psi^2}{(\lambda\psi-\lambda-\psi)^2} \cdot \frac{\lambda+\psi-\lambda\psi}{\psi} = \frac{\lambda(1-\lambda)^{l-1}\psi}{\lambda+\psi-\lambda\psi} \quad (l \geq 2).$$

Furthermore,

$$\begin{aligned}\mathbb{E}[L \mid L \neq 0] &= \frac{\lambda}{\lambda+\psi-\lambda\psi} + \frac{\lambda\psi}{\lambda+\psi-\lambda\psi}\sum_{l=1}^{\infty}(l+1)(1-\lambda)^{l} \\ &= \frac{\lambda}{\lambda+\psi-\lambda\psi} + \frac{(\lambda-1)^2\psi}{\lambda\left(\lambda+\psi-\lambda\psi\right)}.\end{aligned}$$

Let $\mathbf{1}\{\}$ be the indicator function. Maximum likelihood estimation is relatively straightforward when conditioning on $L \neq 0$. Under this scenario,

$$\mathcal{L}(\lambda \mid l) = \prod_{j=1}^{m}\left[\frac{\lambda(1-\lambda)^{l_j-1}\psi^{\mathbf{1}\{l_j\geq 2\}}}{\lambda+\psi-\lambda\psi}\right] = \left(\frac{\lambda}{\lambda+\psi-\lambda\psi}\right)^{m}\prod_{j=1}^{m}(1-\lambda)^{l_j-1}\psi^{\mathbf{1}\{l_j\geq 2\}}.$$

$$\frac{\partial}{\partial\lambda}\log\mathcal{L}(\lambda \mid l) = \frac{m}{\lambda} + \frac{m(\psi-1)}{\lambda+\psi-\lambda\psi} + \frac{m-\sum_{j=1}^{m} l_j}{1-\lambda} = 0,$$

which gives the quadratic

$$-(1-\psi)\Big(\textstyle\sum_j l_j - m\Big)\lambda^2 - \psi\Big(\textstyle\sum_j l_j\Big)\lambda + m\psi = 0,$$

and hence

$$\hat{\lambda}_{\text{MLE}} = \frac{\psi\sum_j L_j - \sqrt{(\psi\sum_j L_j)^2 + 4m\psi(1-\psi)\big(\sum_j L_j - m\big)}}{-2(1-\psi)\big(\sum_j L_j - m\big)}.$$

Conditioning on $L \neq 0$, we can also derive the Fisher information for λ, denoted $\mathcal{I}_{L\neq 0}(\lambda)$:

$$\begin{aligned}f_{L\neq 0}(l;\lambda) &= \frac{\lambda(1-\lambda)^{l-1}\psi^{\mathbf{1}(l\geq 2)}}{\lambda+\psi-\lambda\psi}, \\ \mathcal{I}_{L\neq 0}(\lambda) &= -\mathbb{E}\left[\frac{\partial^2}{\partial\lambda^2}\log f_{L\neq 0}(L;\lambda)\,\middle|\, L\neq 0\right] \\ &= \frac{1}{\lambda^2} + \frac{\lambda}{(1-\lambda)^2(\lambda+\psi-\lambda\psi)} + \frac{\psi}{\lambda(\lambda+\psi-\lambda\psi)} - \frac{1}{(1-\lambda)^2} - \frac{(1-\psi)^2}{(\lambda+\psi-\lambda\psi)^2}.\end{aligned}$$

B.2 Deriving the marginal distribution of the observed tract length under two alternative settings

We first consider the case in which N is distributed as a sum of two independent and identically distributed geometric random variables, each with mean $\phi/2$. We have

$$P(N=n) = (n-1)\Big(1-\tfrac{2}{\phi}\Big)^{n-2}\Big(\tfrac{2}{\phi}\Big)^{2}.$$

Letting $\gamma = 2/\phi$,

$$P(L=l) = \sum_{n=l}^{\infty} P(L=l \mid N=n)\, P(N=n)$$

$$= \begin{cases} \dfrac{\gamma^2(1-\psi)^2}{(\gamma+\psi-\gamma\psi)^2}, & l=0, \\ \dfrac{2\gamma^2\psi(1-\psi)}{(\gamma+\psi-\gamma\psi)^3}, & l=1, \\ \dfrac{\gamma^2(1-\gamma)^{\,l-2}\psi^2\big[(l-3)(\gamma+\psi-\gamma\psi)+2\big]}{(\gamma+\psi-\gamma\psi)^3}, & l\geq 2. \end{cases}$$

Then

$$P(2\leq L\leq M) = \sum_{l=2}^{M} \frac{\gamma^2(1-\gamma)^{\,l-2}\psi^2\big[(l-3)(\gamma+\psi-\gamma\psi)+2\big]}{(\gamma+\psi-\gamma\psi)^3}$$

$$= \frac{(\gamma+\psi-\gamma\psi)\,\psi^2\big[(3-M)\gamma(1-\gamma)^{\,M-1}-(1-\gamma)^{\,M-1}-2\gamma+1\big]+2\gamma\psi^2\big[1-(1-\gamma)^{\,M-1}\big]}{(\gamma+\psi-\gamma\psi)^3}.$$

Hence

$$P(L=l \mid 2\leq L\leq M)$$

$$= \frac{(\gamma+\psi-\gamma\psi)(l-3)\,\gamma^2(1-\gamma)^{\,l-2}+2\gamma^2(1-\gamma)^{\,l-2}}{(\gamma+\psi-\gamma\psi)\big[(3-M)\gamma(1-\gamma)^{\,M-1}-(1-\gamma)^{\,M-1}-2\gamma+1\big]+2\gamma\big[1-(1-\gamma)^{\,M-1}\big]}.$$

Similarly to the case where N is geometric, we index our random variable L using j so that L_j represents the random variable corresponding to the observed tract length for detected

tract j in our dataset. This time, we also index ψ using j so that an allele conversion happens with probability ψ_j at every position within the jth detected tract (the estimation of ψ_j is described in Section 3.7). We have

$$P(L_j = l_j \mid 2 \le L_j \le M)$$
$$= \frac{(\gamma + \psi_j - \gamma\psi_j)(l_j - 3)\,\gamma^2(1-\gamma)^{\,l_j-2} + 2\gamma^2(1-\gamma)^{\,l_j-2}}{(\gamma + \psi_j - \gamma\psi_j)\big[(3-M)\gamma(1-\gamma)^{\,M-1} - (1-\gamma)^{\,M-1} - 2\gamma + 1\big] + 2\gamma\big[1 - (1-\gamma)^{\,M-1}\big]}.$$

We next consider the case where N is distributed as a mixture of two geometric components. We let the two geometric means be ϕ_1 and ϕ_2, and let w_1 represent the mixing proportion of the first component. We have

$$P(N = n) = w_1\big(1 - \tfrac{1}{\phi_1}\big)^{n-1}\tfrac{1}{\phi_1} \;+\; (1 - w_1)\big(1 - \tfrac{1}{\phi_2}\big)^{n-1}\tfrac{1}{\phi_2}.$$

Letting $\lambda_1 = 1/\phi_1$ and $\lambda_2 = 1/\phi_2$,

$$P(L = l) = \sum_{n=l}^{\infty} P(L = l \mid N = n)\, P(N = n)$$
$$= \begin{cases} \dfrac{w_1\lambda_1(1-\psi)}{\lambda_1 + \psi - \lambda_1\psi} + \dfrac{(1-w_1)\lambda_2(1-\psi)}{\lambda_2 + \psi - \lambda_2\psi}, & l = 0, \\ \dfrac{w_1\lambda_1\psi}{(\lambda_1 + \psi - \lambda_1\psi)^2} + \dfrac{(1-w_1)\lambda_2\psi}{(\lambda_2 + \psi - \lambda_2\psi)^2}, & l = 1, \\ \dfrac{w_1\lambda_1(1-\lambda_1)^{\,l-1}\psi^2}{(\lambda_1 + \psi - \lambda_1\psi)^2} + \dfrac{(1-w_1)\lambda_2(1-\lambda_2)^{\,l-1}\psi^2}{(\lambda_2 + \psi - \lambda_2\psi)^2}, & l \ge 2. \end{cases}$$

Then

$$P(2 \le L \le M) = \frac{w_1\psi^2\big[(1-\lambda_1) - (1-\lambda_1)^M\big]}{(\lambda_1 + \psi - \lambda_1\psi)^2} \;+\; \frac{(1-w_1)\psi^2\big[(1-\lambda_2) - (1-\lambda_2)^M\big]}{(\lambda_2 + \psi - \lambda_2\psi)^2}.$$

Hence

$$P(L = l \mid 2 \leq L \leq M) = \frac{\dfrac{w_1\lambda_1(1-\lambda_1)^{l-1}\psi^2}{(\lambda_1+\psi-\lambda_1\psi)^2} + \dfrac{(1-w_1)\lambda_2(1-\lambda_2)^{l-1}\psi^2}{(\lambda_2+\psi-\lambda_2\psi)^2}}{\dfrac{w_1\psi^2\left[(1-\lambda_1)-(1-\lambda_1)^M\right]}{(\lambda_1+\psi-\lambda_1\psi)^2} + \dfrac{(1-w_1)\psi^2\left[(1-\lambda_2)-(1-\lambda_2)^M\right]}{(\lambda_2+\psi-\lambda_2\psi)^2}}.$$

Again using j to index detected tracts,

$$P(L_j = l_j \mid 2 \leq L_j \leq M) = \frac{\dfrac{w_1\lambda_1(1-\lambda_1)^{l_j-1}\psi_j^2}{(\lambda_1+\psi_j-\lambda_1\psi_j)^2} + \dfrac{(1-w_1)\lambda_2(1-\lambda_2)^{l_j-1}\psi_j^2}{(\lambda_2+\psi_j-\lambda_2\psi_j)^2}}{\dfrac{w_1\psi_j^2\left[(1-\lambda_1)-(1-\lambda_1)^M\right]}{(\lambda_1+\psi_j-\lambda_1\psi_j)^2} + \dfrac{(1-w_1)\psi_j^2\left[(1-\lambda_2)-(1-\lambda_2)^M\right]}{(\lambda_2+\psi_j-\lambda_2\psi_j)^2}}.$$

In practice, we plug in $M = 1500$ because we exclude all observed tract lengths longer than 1500 bp detected from the UK Biobank whole autosome data.

Appendix C

THE EFFECT OF LINKAGE DISEQUILIBRIUM ON THE DISTRIBUTION OF OBSERVED TRACT LENGTHS

In this section, we specify gene conversion tract lengths to be geometric. Recall that ϕ is the mean gene conversion tract length. Then, the observed tract length distribution for detected gene conversion tract j, truncated between 1 and 1,500 bp, is

$$P(L_j = l_j \mid 1 \leq L_j \leq 1500, \lambda, \psi_j) = \begin{cases} \dfrac{\lambda\psi_j}{\lambda\psi_j + \psi_j^2\left[1-\lambda-(1-\lambda)^{1500}\right]}, & \text{if } l_j = 1, \\ \dfrac{\lambda(1-\lambda)^{l_j-1}\psi_j^2}{\lambda\psi_j + \psi_j^2\left[1-\lambda-(1-\lambda)^{1500}\right]}, & \text{if } l_j \geq 2, \end{cases}$$

where $\lambda = 1/\phi$ and ψ_j is the allele conversion probability for detected tract j.

In the main text, we described a method for obtaining $\hat{\psi}_j$, our estimate of ψ_j, for all detected tracts j. Plugging in $\psi_j = \hat{\psi}_j$, we can obtain the probability mass at 1 bp, conditioned on $1 \leq L_j \leq 1500$ and λ:

$$P(L_j = 1 \mid 1 \leq L_j \leq 1500, \lambda, \psi_j = \hat{\psi}_j) = \frac{\lambda\hat{\psi}_j}{\lambda\hat{\psi}_j + \hat{\psi}_j^2\left[1-\lambda-(1-\lambda)^{1500}\right]}.$$

We can estimate the proportion of detected tracts with an observed tract length of 1 bp (among detected tracts with an observed tract length less than or equal to 1,500 bp) by taking the mean of $P(L_j = 1 \mid 1 \leq L_j \leq 1500, \lambda, \psi_j = \hat{\psi}_j)$ across all detected tracts j with an observed tract length that is less than or equal to 1,500 bp. Let $\hat{\pi}(L = 1 \mid 1 \leq L \leq 1500, \lambda)$ denote this estimated proportion, computed as:

$$\hat{\pi}(L = 1 \mid 1 \leq L \leq 1500, \lambda) = \frac{1}{|I_1^{1500}|} \sum_{j \in I_1^{1500}} P(L_j = 1 \mid 1 \leq L_j \leq 1500, \lambda, \psi_j = \hat{\psi}_j),$$

where $I_1^{1500} = \{j = 1, \ldots, m \mid 1 \leq l_j \leq 1500\}$ and $|I_1^{1500}|$ represents the number of detected tracts with an observed tract length that is less than or equal to 1,500 bp. In this summation, the only quantity that varies across tracts is $\hat{\psi}_j$, the tract-specific estimate of the allele conversion probability. Notice how $\hat{\pi}(L = 1 \mid 1 \leq L \leq 1500, \lambda)$ depends on $\lambda = 1/\phi$, for which we can plug in an appropriate value (an estimate or the true value if it is known).

Once we obtain the observed tract lengths of detected gene conversion tracts, denoted $\{l_j \mid j = 1, \ldots, m\}$, using the multi-individual IBD method [59], we know the proportion of detected tracts with an observed tract length of 1 bp (among detected tracts with an observed tract length less than or equal to 1,500 bp) in our dataset. If our estimate $\hat{\pi}(L = 1 \mid 1 \leq L \leq 1500, \lambda)$ differs from this proportion, our model may not be fitting well to the data.

Browning and Browning ran a coalescent simulation incorporating gene conversions, where they fixed the mean gene conversion tract length to be 300 bp [59]. Twenty regions of length 10 Mb were generated for 125,000 individuals, and the multi-individual IBD analysis detected 284,838 allele conversions belonging to 226,007 detected gene conversion tracts across the twenty regions. This simulation is described in more detail in the main text and in Browning and Browning (2024) [59].

From this simulation study, the actual proportion of detected tracts with an observed tract length of 1 bp (among detected tracts with an observed tract length less than or equal to 1,500 bp) was 0.807. However, $\hat{\pi}(L = 1 \mid 1 \leq L \leq 1500, \phi = 300) = 0.860$. This indicates that our model is overestimating the proportion of detected tracts with an observed tract length of 1 bp in the coalescent simulation.

We can similarly compare the actual proportion of detected tracts with an observed tract length of 2 bp or longer to the distribution $P(L = l \mid 2 \leq L \leq 1500, \phi)$ derived in the main text. This time, we do not have to worry about varying $\hat{\psi}_j$ for each tract j, because our distribution conditional on $2 \leq L \leq 1500$ no longer depends on the allele conversion probability. Thus, we do not have to average across detected tracts like we did earlier. For example, we can compare the actual proportion of detected tracts with an observed tract

length of 3 bp (among detected tracts with an observed tract length between 2 and 1,500 bp) to $P(L = 3 \mid 2 \leq L \leq 1500, \phi = 300)$. To facilitate this comparison, we define the CDF of L truncated between 2 and 1,500 bp as

$$F_2^{1500}(l \mid \phi) = \sum_{k=2}^{l} P(L = k \mid 2 \leq L \leq 1500, \phi).$$

In Figure C.1, we plot this and the empirical CDF of observed tract lengths between 2 and 1,500 bp detected in the coalescent simulation. We see from Figure C.1 that our truncated distribution of L fits well to the actual proportion of observed tract lengths between 2 and 1,500 bp.

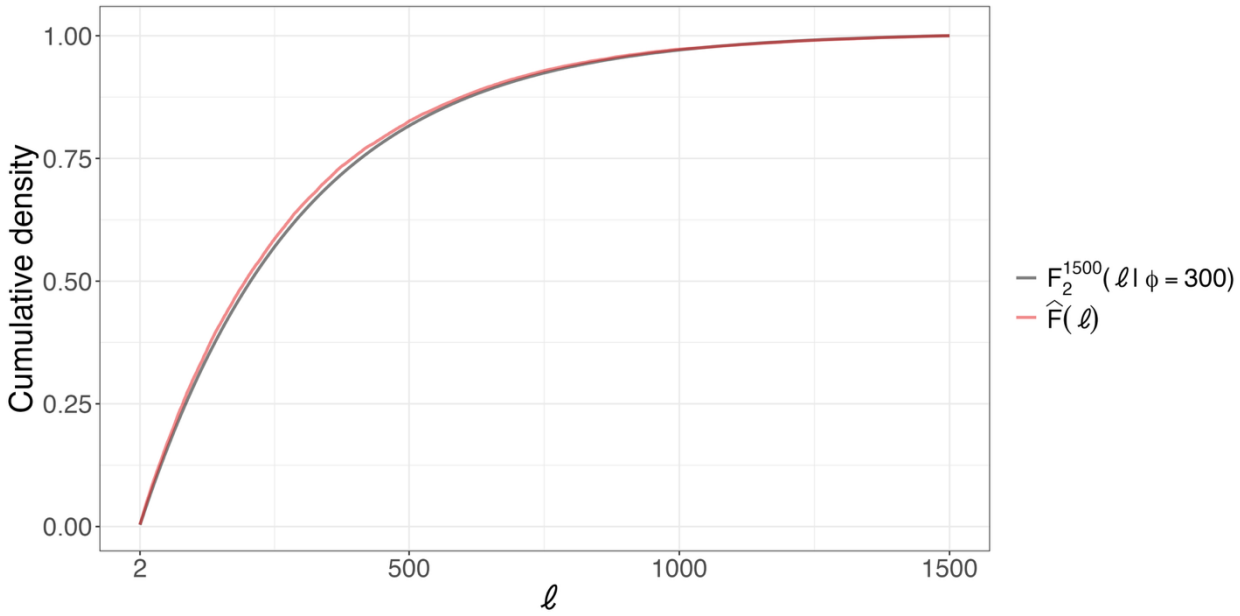

Figure C.1: Comparing the CDF of L and the empirical CDF of observed tract lengths detected in the coalescent simulation. We plot the CDF of L truncated between 2 and 1,500 bp (in grey) and the empirical CDF of observed tract lengths between 2 and 1,500 bp detected in the coalescent simulation (in red).

We want to figure out why our model is not fitting well to the actual proportion of

detected tracts with an observed tract length of 1 bp in the coalescent simulation. We think this is likely because our model does not account for linkage disequilibrium, even though linkage disequilibrium is present in the simulated regions.

Our model assumes that all positions within a gene conversion tract have the same probability of allele conversion. This means that an allele conversion occurring at one position does not make it more or less likely that an allele conversion will occur at another nearby position within the same gene conversion tract. This assumption is used to derive the marginal distribution of L in the main text. However, in this coalescent simulation and in real populations, linkage disequilibrium can cause heterozygosity to be correlated between nearby positions, leading to allele conversions occurring together at nearby positions more frequently than if these positions were independent from one another. This may explain why the actual proportion of detected tracts with an observed tract length of 1 bp in the coalescent simulation is smaller than what the model predicts.

To test whether linkage disequilibrium is causing a smaller proportion of detected tracts to have an observed tract length of 1 bp compared to what the model predicts, we simulate observed tract lengths under a setting where markers are independent. For this simulation, we use the population heterozygosity rates of markers on chromosome 1 from the UK Biobank whole-autosome data. We use the following steps to simulate observed tract lengths:

1. We generate 10^6 gene conversion tracts by uniformly sampling the starting position on chromosome 1 and drawing the length of the gene conversion tract from a geometric distribution with mean 300. The start and end positions of each tract are saved.

2. We let an allele conversion occur at each position i within each gene conversion tract with probability $2p_i(1-p_i)$, where p_i is the minor allele frequency at position i.

3. For each gene conversion tract, we obtain the observed tract length of the gene conversion tract by taking the length spanning the furthest allele-converted positions.

In step 2, we set $p_i = 0$ if the minor allele frequency is less than 5% at position i to prevent

detecting allele conversions at these markers, like in the multi-individual IBD method [59].

From this simulation, the actual proportion of observed tract lengths that were 1 bp was 0.813 (among detected tracts with an observed tract length between 1 and 1,500 bp) whereas $\hat{\pi}(L = 1 \mid 1 \leq L \leq 1500, \phi = 300) = 0.818$. This time, our model only slightly overestimates this proportion. From Figure C.2, we also see that our model closely fits the empirical distribution of observed tract lengths between 2 and 1,500 bp generated from this simulation.

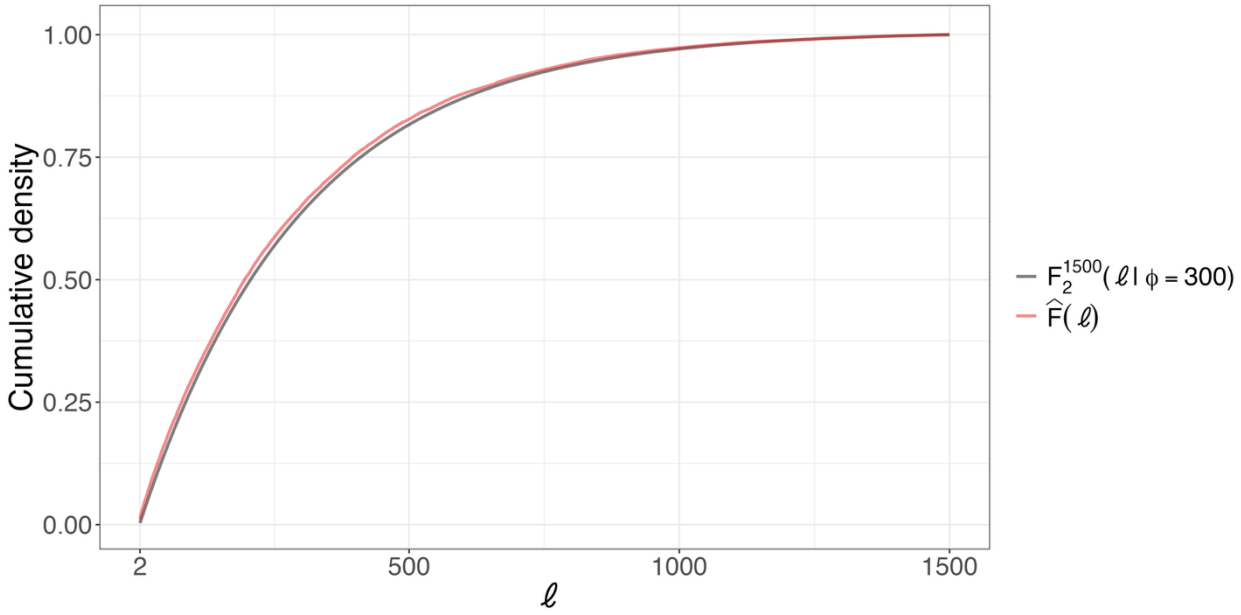

Figure C.2: Comparing the CDF of L and the empirical CDF of observed tract lengths generated in the simulation without linkage disequilibrium. We plot the CDF of L truncated between 2 and 1,500 bp (in grey) and the empirical CDF of observed tract lengths between 2 and 1,500 bp generated in the simulation without linkage disequilibrium (in red).

Compared to the coalescent simulation, our model better predicts the proportion of observed tract lengths that are 1 bp from this simulation, in which observed tract lengths are generated under a setting where markers are independent. Recall that in the coalescent

simulation, the model overestimates the proportion of observed tract lengths that are 1 bp. This indicates that linkage disequilibrium may cause the proportion of observed tract lengths that are 1 bp to be lower than what the model predicts. When estimating the mean length of gene conversion tracts, we can avoid this issue by only considering observed tract lengths between 2 and 1,500 bp and by truncating the marginal distribution of L between 2 and 1,500 bp before model fitting, as we have done in the main paper.

Appendix D

SIMULATION STUDY TO ASSESS THE ROBUSTNESS OF THE MODEL

We run a simulation study to assess how well our model can estimate the mean tract length ϕ when gene conversion tract lengths are from various distributions. We simulate observed tract lengths $\{l_j \mid j = 1, \ldots, m\}$ using five distributions for the length distribution of gene conversion tracts (see Figure D.1):

1. Geometric distribution with mean 100 bp.

2. Sum of two geometric random variables, each with mean 50 bp.

3. Sum of three geometric random variables, each with mean 33.3 bp.

4. Discrete uniform distribution with support from 1 to 199 bp.

5. Mixture of two geometric components with means 700 bp and 68.4 bp, with 5% of tracts drawn from the first component.

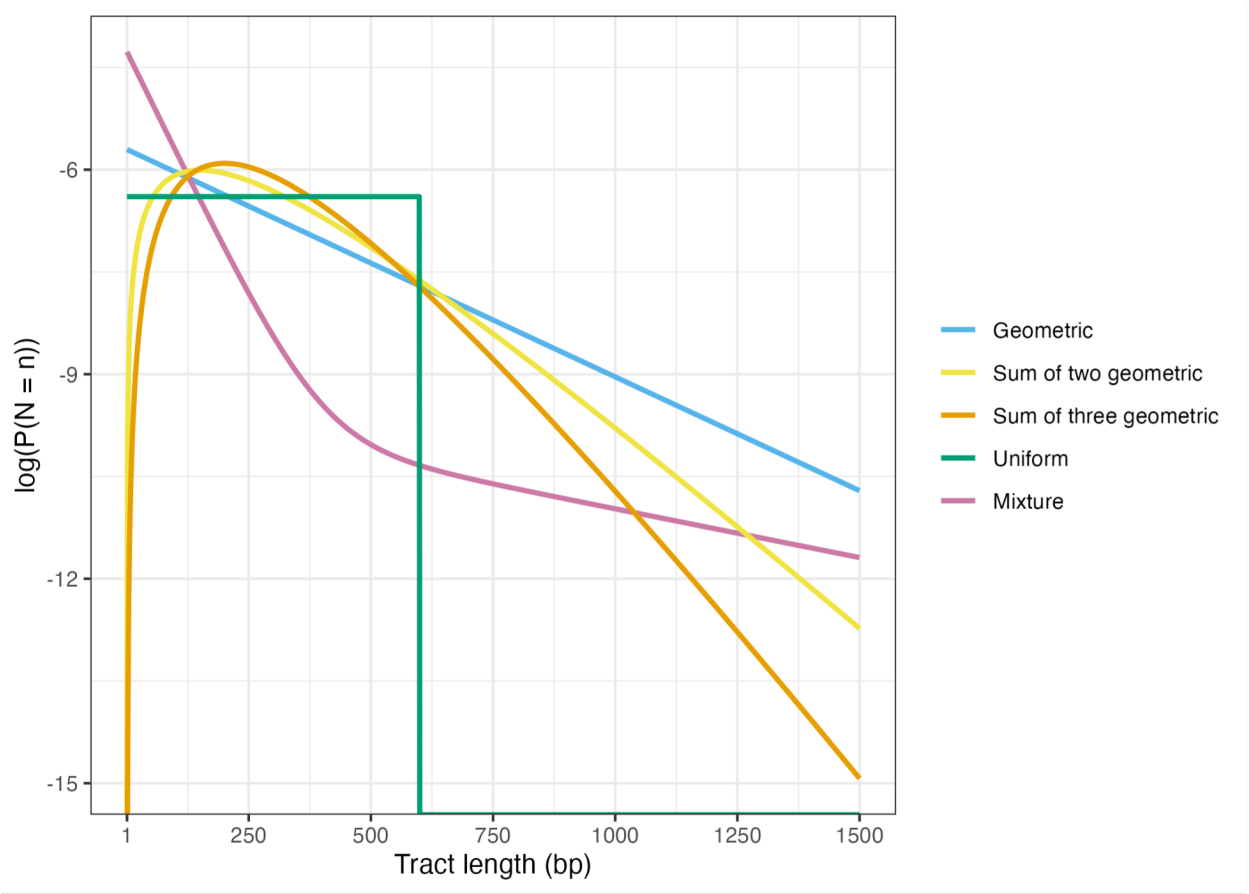

Figure D.1: Probability distribution functions (log scale) of the five distributions used to simulate gene conversion tract lengths. We plot the distribution functions of the geometric distribution, the sum of two geometric random variables, the sum of three geometric random variables, the discrete uniform distribution, and the mixture of two geometric components that we draw the gene conversion tract lengths from the simulation study used to assess the robustness of the model.

All five distributions have an overall mean of 100 bp. Recall that in the previous coalescent simulation we generated 20 regions of length 10 Mb for 125,000 individuals. In this simulation study, we generate observed tract lengths by simulating gene conversion tracts on the first region (out of the 20 regions) from the previous coalescent simulation. To simulate one set

of observed tract lengths, we first sample 100,000 individuals with replacement from the 125,000 individuals. For each resampled individual, we follow these steps:

1. We randomly select a starting position for the gene conversion tract, chosen uniformly across the 10 Mb region.

2. We draw the length of the gene conversion tract from one of the five specified distributions.

3. We determine the observed tract length as the length spanning the furthest heterozygous markers within the simulated gene conversion tract.

This procedure results in 100,000 observed tract lengths, some of which may be 0 bp due to the absence of heterozygous markers within the corresponding gene conversion tracts. For each of the five distributions listed earlier, we repeat this procedure 100 times to obtain 100 sets of observed tract lengths. Then, we fit our model under all distributions of the true tract length (geometric, sum of two geometric random variables, and a mixture of two geometric components) to each set of observed tract lengths. Because the number of observed tract lengths differs for each set, we sample 200 observed tract lengths between 2 and 1,500 bp in each set to make sure that we use the same number of observed tract lengths for estimation.

For each set of observed tract lengths, and for each assumed distribution for the true tract length, we obtain both a point estimate and a 95% bootstrap confidence interval for the mean tract length. Table D.1 reports the empirical bias and empirical standard deviation of our estimate of the mean, as well as the empirical coverage probability of our 95% confidence interval under all model settings across 100 sets of observed tract lengths generated using each of the five distributions. Under the AIC-selected setting, we use the estimate and confidence interval from the assumed tract length distribution with the smallest AIC value in each set of observed tract lengths. Table D.2 reports the number of times each assumed tract length distribution was preferred by AIC across the 100 sets of observed tract lengths generated using each of the five distributions.

Distribution	Chosen Setting	Bias	SD	Coverage
Geom	AIC-selected	13.9	29.2	0.54
	Mixture	-21.1	26.8	0.79
	Geom	-2.5	7.9	0.85
	Geom2	45.9	11.8	0.01
Geom2	AIC-selected	-5.0	13.9	0.79
	Mixture	-34.8	7.7	0.00
	Geom	-33.3	4.5	0.00
	Geom2	-0.5	6.6	0.93
Geom3	AIC-selected	-18.4	8.3	0.12
	Mixture	-45.1	5.8	0.00
	Geom	-43.9	3.3	0.00
	Geom2	-16.4	4.9	0.15
Mixture	AIC-selected	7.3	27.0	0.98
	Mixture	7.3	27.0	0.98
	Geom	265.6	34.6	0.00
	Geom2	434.8	45.6	0.00
Uniform	AIC-selected	-23.6	4.4	0.00
	Mixture	-48.3	3.1	0.00
	Geom	-48.3	3.1	0.00
	Geom2	-23.6	4.4	0.00

Table D.1: Results from simulation study to assess robustness. We assess the performance of our method under each distribution that we use to simulate the true tract lengths (first column) and the chosen setting of the tract length distribution (second column). We report the empirical bias (third column) and standard deviation (fourth column) of our estimate of the mean, as well as the empirical coverage of our 95% confidence interval (fifth column) across 100 replicates of the simulation study. Under the AIC-selected setting, we use the estimate and confidence interval from the distributional setting with the smallest Akaike Information Criterion (AIC) value in each of the 100 replicates.

Distribution	Chosen Setting	Times Selected by AIC
Geom	Mixture	3
	Geom	59
	Geom2	38
Geom2	Mixture	0
	Geom	14
	Geom2	86
Geom3	Mixture	0
	Geom	7
	Geom2	93
Mixture	Mixture	100
	Geom	0
	Geom2	0
Uniform	Mixture	0
	Geom	0
	Geom2	100

Table D.2: Number of replicates each distributional setting was selected by the Akaike Information Criterion (AIC). For each of the five data-generating distributions, we simulated 100 sets of observed tract lengths. We then counted how many times each distribution of N was selected as the best fitting distribution based on AIC.

Appendix E

SUPPLEMENTARY FIGURES

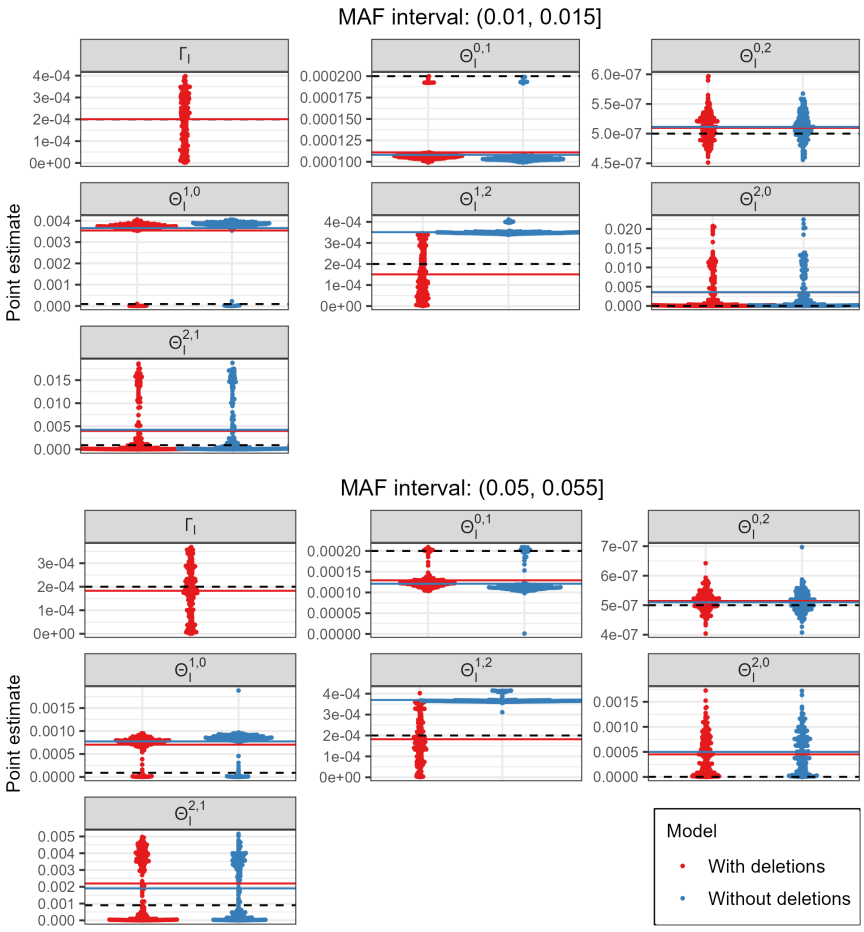

Figure S1: Distribution of parameter estimates from simulated trio genotype counts from markers in (0.01,0.015] and (0.05,0.055]. The dashed black line represents the true parameter value from which the observed trio genotypes are simulated. The red and blue lines represent the sample mean of the estimates from the model with and without deletions respectively.

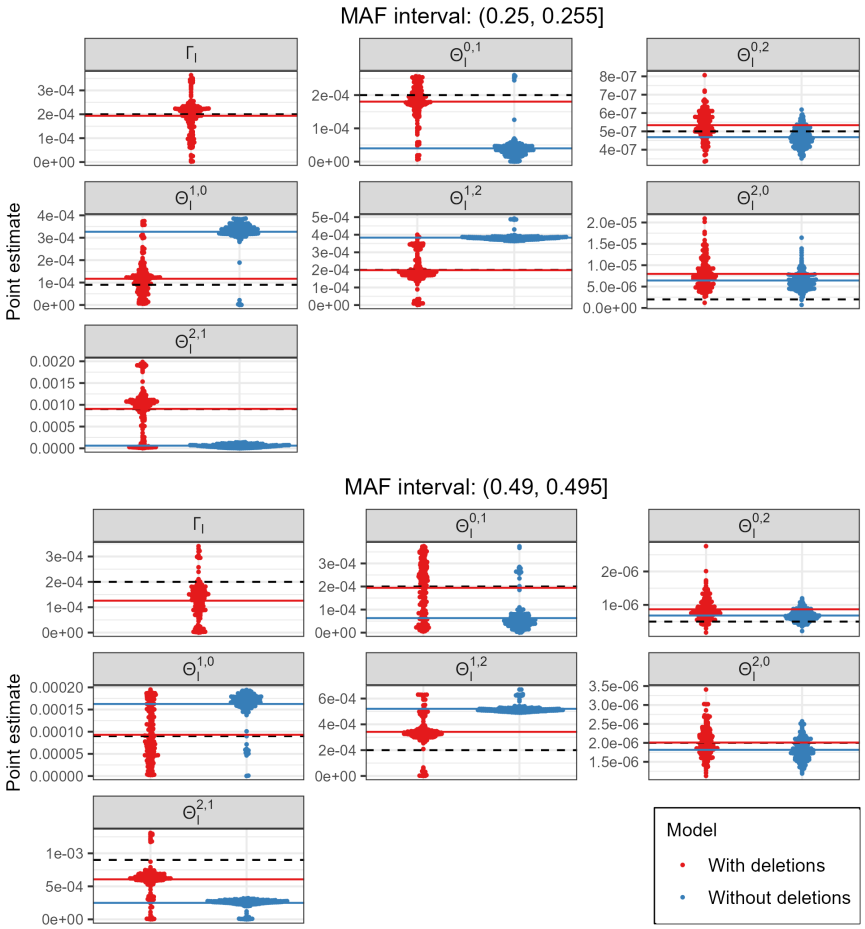

Figure S2: Distribution of parameter estimates from simulated trio genotype counts from markers in (0.25,0.255] and (0.49,0.495]. The dashed black line represents the true parameter value from which the observed trio genotypes are simulated. The red and blue lines represent the sample mean of the estimates from the model with and without deletions respectively.

Appendix F

SUPPLEMENTARY MATERIALS

S1 Funding

The methodological and analytical work performed in this study was supported by the National Human Genome Research Institute under award numbers R01 HG008359 and R01 HG005701. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

This work is supported by NIH NIGMS (R35 GM119774 to T.B.). T.B. is a Howard Hughes Medical Institute Investigator.

S2 Data acknowledgements

This research has been conducted using the UK Biobank Resource under application number 19934.

We gratefully acknowledge the investigators and laboratories that generated, submitted, and shared the sequence data and metadata via GenBank (NCBI), on which this research is based.