Population Genetic Inference with Identity by Descent

Xiaowen Tian

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2020

Reading Committee:

Sharon Browning, Chair

Kelley Harris

Timothy Thornton

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Population Genetic Inference with Identity by Descent

Xiaowen Tian

Chair of the Supervisory Committee:
Sharon Browning
Biostatistics

Identical-by-descent (IBD) segments carry important information on genetic relatedness between individuals and therefore are used widely for statistical inference in population genetics. In this work, we focus on the inference of the genome-wide mutation rate using IBD segments. The two primary methods for estimating the genome-wide mutation rate have been counting de novo mutations in parent-offspring trios and comparing sequence data between closely related species. With parent-offspring trio analysis it is difficult to control for genotype error, and resolution is limited because each trio provides information from only two meioses. Inter-species comparison is difficult to calibrate due to uncertainty in the number of meioses separating species, and it can be biased by selection and by changing mutation rates over time. Existing IBD-based methods are limited to highly inbred samples, or lack robustness to genotype error and error in the estimated demographic history. In this work, we propose mathematical models based on coalescent theory for an IBD-based method that uses sharing of IBD segments among sets of three individuals to estimate the mutation rate as well as the gene conversion rate. With the application to whole genome sequence data from TOPMed Project, we obtain an estimate of 1.58×10^{-8} mutations per base pair per meiosis based on European descendants and 1.52×10^{-8} based on African descendants.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

There are many people that have earned my gratitude for their contribution to my time in graduate school. First, I would like to thank my advisor Sharon Browning for her insights, support, and encouragement. I am deeply grateful to Sharon Browning for introducing me to the research area of population genetics. I am extremely grateful for the help and advice from my dissertation committee members - Kelley Harris, Timothy Thornton, Joseph Felsenstein, and Barbara Wakimoto and members of the Browning Lab. I would like to acknowledge Brian Browning and Ying Zhou for their suggestions and comments on part of this dissertation work. I would also like to thank my research assistant advisor, Holly Janes, for her guidance and generous support. I am also grateful to other faculty and staff in the Department of Biostatistics for making a wonderful graduate program. Last and most, I would like to thank my family for their support and love.

DEDICATION

to my family

Chapter 1

INTRODUCTION

1.1 Identity by descent

A haplotype is a sequence of genetic information on a chromosome that was inherited from a single parent. Two individuals share a haplotype identical by descent if they inherited the haplotype from a common ancestor. Recombination events during meiosis result in the exchange of genetic material between paired homologous chromosomes and therefore break down identity by descent (IBD) segments. As a result, more closely related individuals are expected to share a longer segment identical by descent as compared to distantly related individuals. Figure 1.1 demonstrates the inheritance of genetic material across two generations. The concept of IBD is defined relative to the founders in the given pedigree in this example. Each child's chromosome is a mixture of each parent's two chromosome copies. If we assume the mating happened between unrelated individuals, the full-siblings are expected to share 1/2 their genome identical by descent, and the probability that alleles sampled at random from each individual are IBD (i.e. the kinship coefficient) is 1/4. The first-cousins are expected to share 1/8 of the genome identical by descent since the kinship coefficient between their mothers is 1/4 and their fathers are unrelated. As we compare less closely related individuals, we expect them to have shorter and fewer IBD segments. The length of an IBD segment shared by two individuals inherited from a common ancestor m generations in the past is exponentially distributed with mean $1/(2m)$ Morgans assuming recombinations occur as a Poisson process [24].

The measure of IBD sharing with respect to the founders is well defined in given pedigrees that span a small number of generations. However, the fact of IBD does not depend on

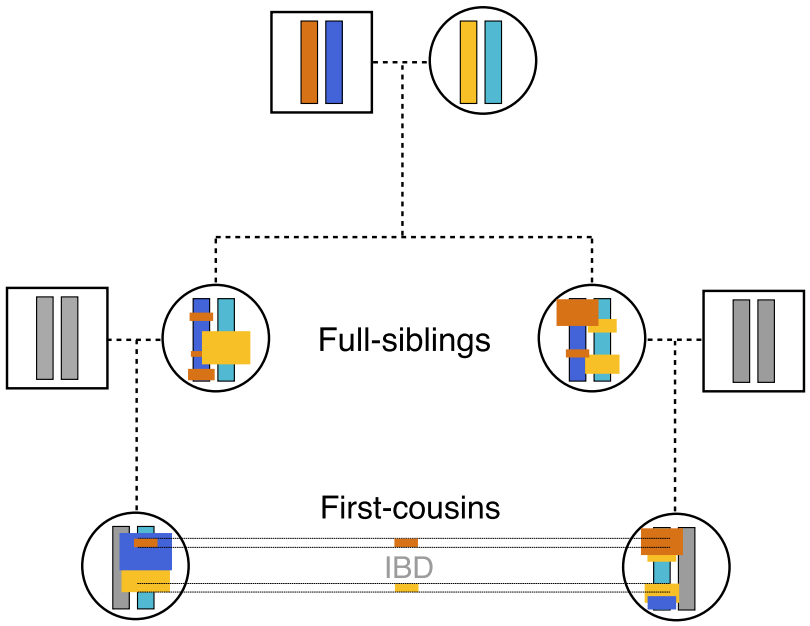


Figure 1.1: Illustration of identity by descent (IBD) segments with respect to a pedigree. The chromosomes are shown in orange, blue, yellow, and green, representing the genetic material inherited from the founders of this family. Each family member is represented by a pair of chromosomes inherited from their two parents. The chromosomes are colored to indicate DNA inherited from the same grandparent. Chromosomes of other members who are not related to the founders are shown in grey. In this example, the first cousins share an orange IBD segment through inheritance from their grandfather. They also share an yellow IBD segment through inheritance from their grandmother.

whether pedigree relationships are known and IBD can be measured relative to the population at some past time point, with the implication that remote genetic relatedness is ignored [52]. In the remainder of this thesis, we define IBD segments as a chromosomal region of a specified length that is transmitted without recombination from a common ancestor who lived at

any time in the past. Despite the name, IBD segments need not be identical. Mutations and gene conversion events in IBD segments may introduce mismatches between haplotypes that are IBD during transmission from a common ancestor to the descendants. Although fewer and shorter IBD segments can be found when analyzing distantly related individuals, if the sample size is large, the total amount of IBD sharing in the sample can become significant as there are $\binom{N}{2}$ possible pairs of individuals to consider for a cohort of N unrelated individuals.

IBD segments can be inferred from single nucleotide polymorphism (SNP) array data or sequence data. SNP array data has lower marker density and most of the markers are common variants, whereas sequence data has higher marker density and most of the markers are rare variants. Most of the statistical approaches for IBD detection fall into two categories: length threshold-based methods, and probabilistic based methods with or without incorporating linkage disequilibrium (LD) [10]. GERMLINE [22] is a computationally efficient length threshold-based approach using sliding windows with a dictionary of haplotypes to discover short exact matches between individuals and then expand these matches to identify long, nearly identical segmental sharing. The program takes phased genotype data as input and allows for haplotype phasing and genotype errors. PLINK [45] is a probabilistic method that does not consider LD. The program takes thinned markers and implements a hidden Markov model (HMM) with the IBD status being the hidden state, the overall relatedness of samples providing information for the transition probabilities, and the IBD status being estimated from the posterior probabilities. PLINK does not allow for genotype errors. Beagle Refined IBD [5] makes use of the advantages of GERMLINE and extends the probabilistic method to incorporate LD to improve accuracy and efficiency. Refined IBD first implements the dictionary approach to identify shared haplotypes of a minimum length and then evaluates the IBD probability of each candidate segment. Refined IBD uses estimated haplotypes that are obtained within the software and it does not allow for genotype errors. IBDseq [6] is a probabilistic based method designed for sequence data. The program uses unphased genotypes and models genotype error, thus it is robust to genotype errors.

Detection of short IBD segments (1-3 cM) has been challenging [38] due to various source of errors. Haplotype phasing errors could break long IBD segments into short segments and lead to underestimation of the length of IBD segments. Genotype errors will also break long IBD segments if the detection method does not allow genotype errors [5]. Moreover, while a large number of comparisons need to be made at each site, the base rate of true IBD segments between pairs of unrelated individuals is low. Therefore, a substantial fraction reported short IBD segments could be false positive errors [3]. Methods based on identity-by-state (IBS) segments may overestimate short IBD segments since the number of IBS segments among unrelated individuals may overwhelm the signal from the real IBD segments [38]. Thus the accuracy of IBD detection methods for short segments is relatively low as compared to the accuracy for longer segments [6]. Setting a large threshold on length of reported IBD segments could reduce the impact from these source of errors, however, long IBD segments can only reflect recent population history rather than ancient history.

1.2 Mutation rate estimation

Mutation adds new genetic variation to populations. This genetic variation is crucial for evolution, and it affects the amount of information available for many common genetic analyses such as genotype imputation, estimation of relatedness, and estimation of ancestral origins. Accurate estimation of the genome-wide mutation rate is important for inferring key demographic parameters such as the timing of population splits [47]. Genome-wide mutation rate estimates are also helpful for understanding the evolution of mutation rate [34]. Despite its importance, measuring mutation rates has been difficult. The direct approach to mutation rate estimation involves sequencing nuclear families and counting de novo mutations in the offspring. However, there are only a small number of de novo mutations per offspring (typically 40-120 genome-wide in humans) [27] and it is difficult to distinguish true mutations from genotype errors and somatic mutations [49]. The choice of filters to remove variants

with higher rates of genotype error, and the assessment of false positive and false negative rates is somewhat arbitrary, which makes it possible for researchers to unintentionally choose filters and methods for assessing error rates that produce an estimate of mutation rate that is close to previously published estimates. Indeed, a recent review found that pedigree-based estimates of mutation rate appear under-dispersed, suggesting a lack of independence across studies [49].

An alternative approach to estimating mutation rates that is less susceptible to genotype error and that uses mutation across large numbers of meiosis is based on the comparison of the human genome with the genomes of closely related species, calibrated by the fossil evidence for the dates of splits between species. The estimates from these inter-species comparisons can be biased by selection, incorrect estimates of average generation length, uncertainty in dating the fossil record, and changes in mutation rates over time. Genome-wide mutation rate estimates from family-based studies are approximately half as high as estimates from inter-species comparisons, suggesting that inter-species estimates are inflated [47].

An alternative approach to mutation rate estimation uses identity by descent (IBD) segments. An IBD segment is a shared portion of a chromosome inherited intact (except for small regions of gene conversion) by two individuals from a common ancestor. The inherited segment will have an identical sequence of alleles in both individuals, except at positions which have mutated since the common ancestor or that were affected by gene conversion. The length of an IBD segment provides information on the number of meioses linking the two haplotypes through their common ancestor, while mismatches in the haplotype sequences provide information regarding the total number of mutations from those meioses. The use of IBD segments to estimate mutation rates has the potential to combine the best features of inter-species and parent-offspring comparisons. Large samples of distantly related individuals can be assayed, leading to assessment of mutations from a large number of meioses. Since

IBD looks back thousands rather than millions of years, there is no danger of confounding the mutation rate in modern humans with that in ancestral human groups and closely-related species.

Recently, Palamara et al. proposed an IBD-based method for estimating mutation rates from accurately phased whole-genome sequence data, such as that obtained from parent-offspring trios [42]. Whereas ordinary trio-based analyses use only meioses within trios, Palamara et al.'s approach uses meioses from IBD between pairs of trio offspring, and thus it draws on many more meiosis than methods that count only de novo mutations. Palamara et al.s method accounts for the effect of genotype error on mutation counts through a regression of the apparent number of mutations in an IBD segment on the estimated time to most common ancestor (TMRCA) of the IBD segment, since the rate of genotype errors is not influenced by the TMRCA, but the number of mutations increases proportionally with the TMRCA. However, genotype error can also affect other key aspects of IBD-based mutation rate estimation in addition to mutation counts, such as the estimation of IBD segment lengths and the estimation of the populations demographic history. The latter two aspects are critical for estimating the TMRCAs of the IBD segments. Palamara et al.s study considered the effect of genotype error on mutation counts, but not its effect on mis-estimation of IBD segment lengths or incorrectly inferred demographic history. In Section 3.3, we find that Palamara et al.s method can give biased estimates of mutation rate, with the amount of bias depending on the level of genotype error and whether the true or inferred demographic history is used.

Another IBD-based approach uses heterozygous genotypes within segments of autozygosity in individuals from populations with high parental relatedness [37, 12]. Advantages of this method over a general IBD-based method is that it is easier to accurately infer long segments of autozgyosity than short segments of IBD, and no estimation of demographic history is needed because one needs only to estimate the degree of parental relatedness of

each individual. A limitation is that it is only applicable to populations for which consanguineous marriages are common. Another approach that utilizes autozygosity rather than between-individual IBD is based on comparing local heterozygosity with estimated TMRCAs along the genome for the two haplotypes in an outbred individual[32]. This latter method incorporates mutations resulting from meioses far back in human history, much further back than IBD-based approaches, and thus requires a very high-resolution recombination map for accurate estimation. Another disadvantage of this method is that it is computationally demanding, and thus it can only be applied to a very small number of individuals, which leads to low precision.

In this thesis, we focus on developing new models and methodologies for mutation rate and gene conversion rate estimation using the IBD segments from population data. We propose a likelihood-based method for estimating genome-wide average mutation rates from sets of three individuals who share a single haplotype identical by descent. We count rare variants shared by two of the three individuals. This avoids the use of singleton variants which have higher genotype error rates [46, 25], and it requires that two genotype errors are need to create any false apparent mutation. The third individual who is IBD with the first two and does not carry the rare alleles provides information on the age of the mutations through the length of IBD sharing between this individual and the other two. We incorporate the distribution of the length of IBD segments, the probability of time to coalescence, the mutation rate, and genotype error into a likelihood function which we maximize to estimate the mutation rate. In the second chapter, we present the likelihood framework designed for accurately phased sequence data without genotype errors or gene conversions, and with ground truth IBD sharing and demographic history. In the third chapter, we address issues with inferring IBD segments, inferring demographic history, and genotype errors. In the fourth chapter, we provide two frameworks for joint estimation of mutation rate and gene conversion rate, one based on regression and the other on likelihood. In the fifth chapter, we generalize our method to unphased datasets with unknown demographic histories and

we present analyses of datasets from the Trans-Omics for Precision Medicine (TOPMed) program [50].

Chapter 2

MUTATION RATE INFERENCE WITH IDENTITY-BY-DESCENT SEGMENTS

2.1 Coalescence probabilities for three haplotypes

Our calculation of the coalescence probabilities are based on the Wright-Fisher model of reproduction [19, 58], which was developed in order to study and quantify genetic variation due to demographic stochasticity. This simple model that describes the population's genealogical relationship among genes, and a number of idealized and simplifying assumptions of this model [26]:

1. Discrete and non-overlapping generations. Reproduction and death are simultaneous and synchronous among all individuals.

2. Haploid individuals. It is common to assume a haploid population of size $2N$ for a diploid population of size N.

3. Constant population size. The number of individuals stays the same over time.

4. Equal fitness. All individuals are equally likely to survive and reproduce.

5. No population structure. Individuals are equally likely to be the offspring from any individuals in the previous generation.

6. No recombination. The entire genetic material is passed from parents to offspring.

Figure 2.1 demonstrates the reproduction process for ten generations under the Wright-Fisher model with ten haplotypes. Although this model is highly idealized, it allows us to

translate the physical model to an elegant mathematical framework while keeping important aspects of biological realism.

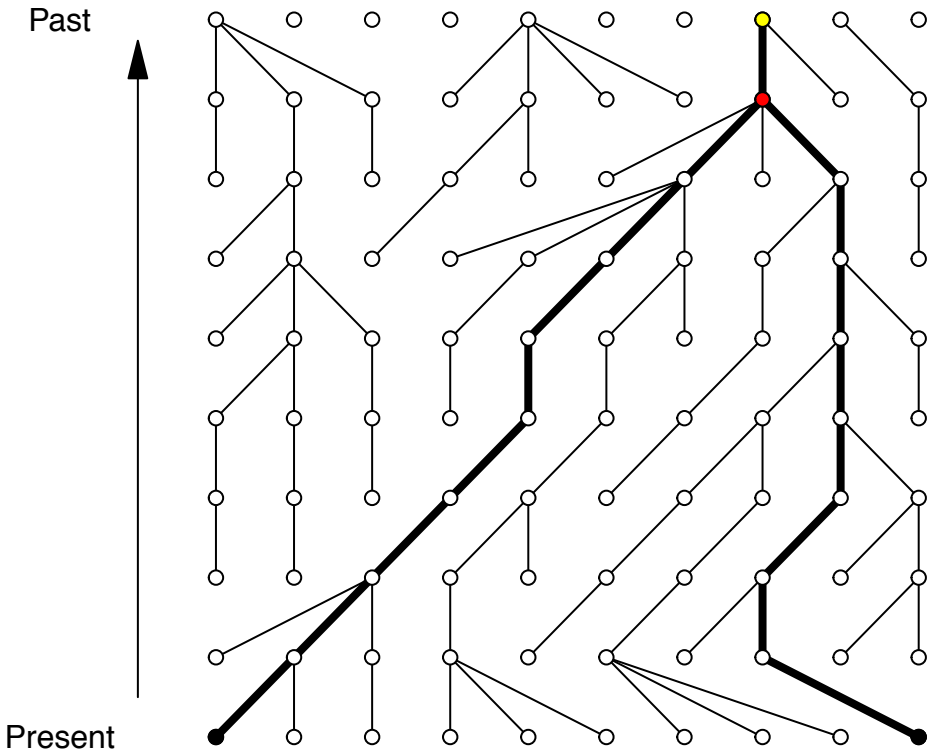


Figure 2.1: The Wright-Fisher model with ten haplotypes. The haploid Wright-Fisher model with ten haplotypes applied for ten generations. The bold line represents the genealogy of two randomly sampled haplotypes (black). The common ancestors of the sampled haplotypes are labeled in red and yellow while the red haplotype represents the most recent common ancestor of the sampled haplotypes.

For a haploid population with size N, the probability that two randomly chosen haplotypes share the same ancestor in the previous generation is $1/N$, since for a given individual i in the previous generation, the probability that individual i is chosen by A and B simulta-

neously as parent is $1/N^2$ and there are N individuals in the previous generation to choose from. Thus, the probability that two present-day haplotypes coalesce at generation g, given that they haven't coalesced more recently, is the probability that they both are assigned the same ancestor out of the N ancestral haplotypes existing in generation g. This probability is 1/N. And the probability of no coalescences in generations 1 to $g-1$ is $(1-1/N)^{g-1}$. Therefore, the probability that two present-day haplotypes coalesce at generation g is $(1-\frac{1}{N})^{g-1}\frac{1}{N}$.

This calculation can be extended to the coalescence for three haplotypes without assuming that the population size is constant over time. In a coalescent tree for three haplotypes (Figure 2.2), there are two coalescence events: the first coalescence, between haplotypes A and B, occurred g_1 generations before present, and the second coalescence, between C and the common ancestor of A and B, occurred g_1+g_2 generations ago. We write $N[g]$ for the haploid effective size g_1 generations in the past. The probability that two present-day haplotypes coalesce at generation g, given that they haven't coalesced more recently, is the probability that they both are assigned the same ancestor out of the $N[g]$ ancestral haplotypes existing in generation g. This probability is $1/N[g]$. Thus the probability that the two haplotype don't coalesce is $1 - 1/N[g]$. Similarly, the probability that no pair of haplotypes among three present day haplotypes coalesce at generation g, given that none of these haplotypes have coalesced more recently, is the product of the probability that the second haplotype is assigned an ancestor that is different from the first haplotypes ancestor, and the probability that the third haplotype is assigned an ancestor that is different from the other two ancestors, which is $(1 - 1/(N[g])(1 - 2/N[g])$. Thus the probability of no coalescence in generations 1 to $g_1 - 1$ is

$$\prod_{g=1}^{g_1-1}(1-\frac{1}{N[g]})(1-\frac{2}{N[g]}).$$

Using similar reasoning, the probability of a coalescence between a given pair of haplotypes (A and B) but no coalescence with the third haplotype (C) at generation g_1, given no coalescence more recently, is $(1 - 1/N[g_1])(1/N[g_1])$. The probability of no coalescence

between C and the common ancestor of A and B between generations (g_1+1) and (g_1+g_2-1) is $\prod_{g=g_1+1}^{g_1+g_2-1}(1-1/N[g])$. The probability of coalescence between C and the common ancestor of A and B at generation (g_1+g_2), given that this coalescence has not occurred more recently, is $1/N[g_1+g_2]$. Thus, the overall probability of this coalescent tree (which we refer to as 'tree3' since it is a tree for three haplotypes) is

$$P(tree3) = \{\prod_{g=1}^{g_1-1}(1-\frac{1}{N[g]})(1-\frac{2}{N[g]})\}\frac{1}{N[g_1]}(1-\frac{1}{N[g_1]})\{\prod_{g=g_1+1}^{g_1+g_2-1}(1-\frac{1}{N[g]})\}\frac{1}{N[g_1+g_2]}.$$

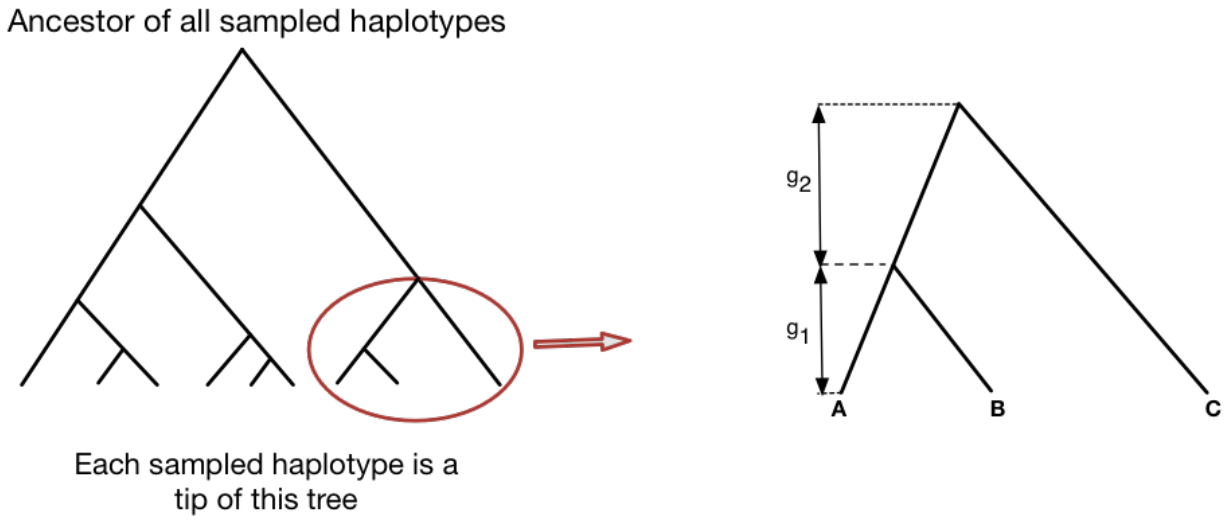


Figure 2.2: An example of the coalescent tree that links three haplotypes. In this example, A, B, and C are the IBD haplotypes that form a set of three-way IBD. Haplotypes A and B coalesce g_1 generations before the present, while C and the common ancestor of A and B coalesce g_1+g_2 generations before the present. The true tree is unknown, and this figure demonstrates one possible tree linking the three haplotypes.

2.2 IBD sharing and time to most recent common ancestor

In the Wright Fisher model, the entire genetic material is passed from parents to offspring under the assumption of no recombination, which makers the sampled haplotypes related by a genealogical tree. However, this assumption needs to be relaxed when we analyze chromosome-wide data which is subject to genetic recombination. During meiosis segments of IBD are broken up by recombination. Therefore, the expected length of an IBD segment depends on the number of generations since the most recent common ancestor at the locus of the segment.

Let g be the number of generations to the most recent common ancestor of two haplotypes at a given randomly-chosen genomic position. A genetic map is used to convert physical base pair positions to genetic positions in Morgans (one Morgan equals one hundred centiMorgans). By definition, the recombination rate is 1 per Morgan per meiosis at any point in the genome. We also assume that recombinations occur as a Poisson process [24]. If we look on one side of the given position, the distribution for length of the IBD segment on that side is $D_g = min(L_1, ..., L_{2g})$ where L_i is exponentially distributed with rate 1 per Morgan since any recombination occurring in either of the lineages would end the IBD segment and the minimum length will determine the length of haplotype that remains intact. The L_i are independent since recombinations in different generations are independent. Therefore, D_g is exponentially distributed with rate $2g$ per Morgan. If we look both upstream and downstream from the chosen site, the distribution of the length of an IBD segment is the sum of two independent exponential distribution each with rate $2g$ per Morgan; that is, the length has a gamma distribution with shape 2 and rate $2g$ per Morgan. We next extend this result to three-way IBD sharing.

When three haplotypes are jointly identical by descent at a given point in the genome, the length of IBD sharing around that position can vary. We consider not only the three-way

region over which all three haplotypes are identical by descent, but also the larger region over which any two of the three haplotypes are identical by descent, because the pairwise IBD segment lengths provide information about the coalescent tree (the ordering of the coalescence events and the coalescence times) in the three-way IBD region. For example, looking to the left of the given position, haplotype C may cease to be IBD with haplotypes A and B at some position, and then at some more distant position A and B also cease to be IBD (Figure 2.3). In Figure 2.3, x_1, x_2, x_3, x_4 are the positions of changes in IBD status measured in Morgans. In this example, the IBD segment shared by haplotypes A and B starts at x_1 and ends at x_3; the IBD segment shared by A and C starts at x_2 and ends at x_4; and x_2 to x_3 is the region shared jointly by A, B, and C. Then if the coalescent tree corresponding to the segment of three-way IBD sharing is that shown in Figure 2.2, the length of the IBD segment shared jointly by A, B, and C has a gamma distribution with shape 2 and rate $3g_1 + 2g_2$ per Morgan, because the total number of meioses in the coalescent tree is $3g_1 + 2g_2$, and a recombination on any one of those meioses will end the joint IBD segment. When the first recombination occurs to end the three-way IBD at the right end (at position x_3), the probability that B is lost rather than A or C is $g_1/(3g_1 + 2g_2)$, and in this case the length of the pairwise IBD segment shared by A and C to the right of x_3 is exponentially distributed with rate $2g_1 + 2g_2$ per Morgan. Similarly, when the first recombination occurs to end the three-way IBD at the left end (at position x_2), the probability that C is lost rather than A or B is $(g_1 + 2g_2)/(3g_1 + 2g_2)$, and in this case the length of the IBD segment shared by A and B to the left of x_2 is exponentially distributed with rate $2g_1$ per Morgan. In this way, we can calculate the probability of the IBD lengths (which we refer to as 'IBD3' since they summarize the three-way IBD sharing for three haplotypes) for any possible coalescent tree.

Let AB-C represent the coalescent tree in Figure 2.2, the other two possible coalescent trees are AC-B and BC-A (Figure 2.4). In AC-B, the coalescence between haplotype A and haplotype C occurred g_1 generations ago and the coalescence between B and the common ancestor of A and C occurred $g_1 + g_2$ generations ago. In BC-A, the coalescence between B

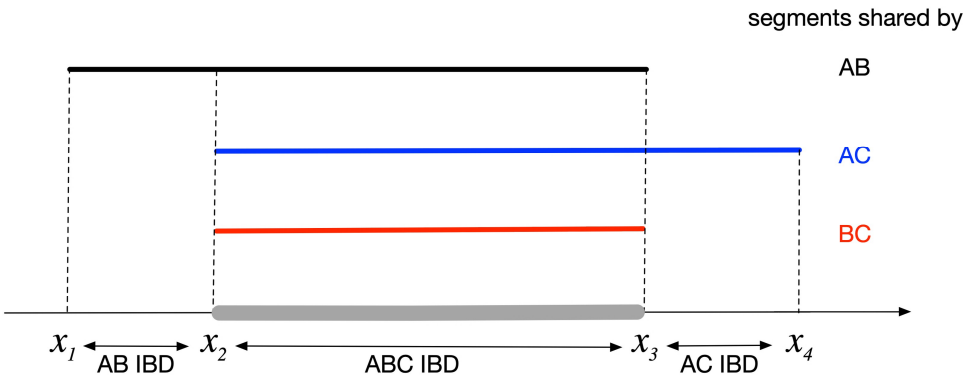
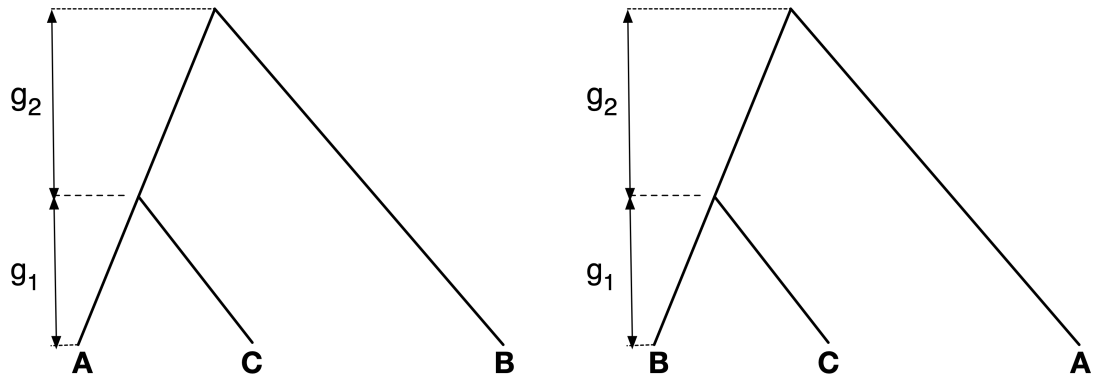


Figure 2.3: An example of three-way IBD. This figure illustrates one possible IBD sharing configuration among three haplotypes denoted A, B, and C. The IBD segment shared by A and B starts at x_1, ends at x_3, and is colored black. The IBD segment shared by A and C starts at x_2, ends at x_4, and is colored blue. The IBD segment shared by B and C starts at x_2, ends at x_3, and is colored red. The gray region from x_2 to x_3 is the IBD region shared jointly by A, B, and C.

and C occurred g_1 generations ago and the coalescence between A and the common ancestor of B and C occurred $g_1 + g_2$ generations ago. Let AB:AC represent the three-way IBD sharing pattern shown in Figure 2.3, where the combination of two haplotypes before the colon symbol shared an IBD segment starts from x_1 and ends at x_2 and the combination of two haplotypes after the colon symbol shared an IBD segment starts from x_3 and ends at x_4. Then there are nine possible three-way IBD sharing patterns in total (Figure 2.3, Figure 2.5).

Let positions x_1, x_2, x_3, x_4 be the changes in IBD status measured in Morgans as shown in Figure 2.3 and Figure 2.5 and g_1, g_2 be the coalescent time measured in generations. Let $f_1(x; \lambda) = \lambda e^{-\lambda x}$ denote an exponential distribution and $f_2(x) = (3g_1 + 2g_2)^2 x e^{-(3g_1+2g_2)x}$ denote the gamma distributions with shape parameter 2 and rate parameter $3g_1 + 2g_2$. Then,



(a) Coalescent tree: AC-B **(b)** Coalescent tree: BC-A

Figure 2.4: Other types of Coalescent trees. In (a), haplotypes A and C coalesce g_1 generations before the present, while B and the common ancestor of A and C coalesce $g_1 + g_2$ generations before the present. In (b), haplotypes B and C coalesce g_1 generations before the present, while A and the common ancestor of B and C coalesce $g_1 + g_2$ generations before the present.

for each of the three-way IBD sharing pattern, we can derive $P(IBD3|tree3)$:

$$
\begin{aligned}
&P(AB:AC \mid AB-C) = P(AB:BC \mid AB-C) \\
&= P(AC:AB \mid AC-B) = P(AC:BC \mid AC-B) \\
&= P(BC:AB \mid BC-A) = P(BC:AC \mid BC-A) \\
&= \frac{(g_1+2g_2)g_1}{(3g_1+2g_2)^2} f_2(x_3-x_2) f_1(x_2-x_1;2g_1) f_1(x_4-x_3;2g_1+2g_2)
\end{aligned}
$$

$$
\begin{aligned}
&P(AC:AB \mid AB-C) = P(BC:AB \mid AB-C) \\
&= P(AB:AC \mid AC-B) = P(BC:AC \mid AC-B) \\
&= P(AB:BC \mid BC-A) = P(AC:BC \mid BC-A)
\end{aligned}
$$

$$= \frac{(g_1+2g_2)g_1}{(3g_1+2g_2)^2} f_2(x_3-x_2) f_1(x_2-x_1; 2g_1+2g_2) f_1(x_4-x_3; 2g_1)$$

$$P(AC:BC \mid AB-C) = P(BC:AC \mid AB-C)$$
$$= P(AB:BC \mid AC-B) = P(BC:AB \mid AC-B)$$
$$= P(AB:AC \mid BC-A) = P(AC:AB \mid BC-A)$$
$$= P(AC:AC \mid AB-C) = P(BC:BC \mid AB-C)$$
$$= P(AB:AB \mid AC-B) = P(BC:BC \mid AC-B)$$
$$= P(AB:AB \mid BC-A) = P(AC:AC \mid BC-A)$$
$$= \frac{g_1^2}{(3g_1+2g_2)^2} f_2(x_3-x_2) f_1(x_2-x_1; 2g_1+2g_2) f_1(x_4-x_3; 2g_1+2g_2)$$

$$P(AB:AB \mid AB-C) = P(AC:AC \mid AC-B) = P(BC:BC \mid BC-A)$$
$$= \frac{(g_1+2g_2)^2}{(3g_1+2g_2)^2} f_2(x_3-x_2) f_1(x_2-x_1; 2g_1) f_1(x_4-x_3; 2g_1)$$

2.3 Probability distribution of mutation counts given the coalescent tree

Given a mutation rate μ per base pair per meiosis, and assuming the infinite sites model [29], the number of mutations accumulated within a genome region of length l base pairs over g meioses has a Poisson distribution with mean $lg\mu$. If the coalescent tree is that shown in Figure 1, in which haplotypes A and B coalesce first, before coalescing with C, with g_2 being the number of meioses from the common ancestor of A and B to the common ancestor of all three haplotypes, then the number of mutations shared by haplotypes A and B but not C across a region of l base pairs within the three-way IBD sharing region is distributed as Poisson($lg_2\mu$). In contrast, considering two of three haplotypes that are not the first coalescing pair, such as haplotypes A and C for this coalescent tree, any apparent mutations shared by these two haplotypes but not the third will be inconsistent with the coalescent tree

(the probability of recurrent mutation is negligible, and is ignored under the infinite sites model). Thus the number of such apparent mutations between any two haplotypes which do not coalesce first is modeled as Poisson with rate zero.

For a given labelling of the three IBD haplotypes, let 'mut3' denote the vector (n_{AB}, n_{AC}, n_{BC}) containing the number of apparent mutations shared by haplotypes A and B but not C (n_{AB}), by haplotypes A and C but not B (n_{AC}), and by haplotypes B and C but not A (n_{BC}) across the region in which all three haplotypes are IBD. An apparent mutation is an allele that is shared by two of the three haplotypes, and has frequency less than the maximum allele frequency threshold. The maximum allele frequency threshold is chosen to be large enough so that all true mutations will be included in the counts, and is never set to a value above 0.5. Thus if two of the three haplotypes share the major allele this will not contribute to the apparent mutation count.

Let $P_\mu(mut3|tree3, IBD3)$ denote the probability of the vector of apparent mutations given the coalescent tree and the IBD endpoints if the mutation rate is μ. Note that after conditioning on tree3 , the distribution of the number of mutations depends on the IBD endpoints only through the base pair length l of the three-way IBD region on which the apparent mutations are counted. Let AB-C, AC-B and BC-A represent the coalescent trees shown in Figure 2.2 and Figure 2.4, then

$$P_\mu(mut3|tree3 = (AB-C, g_1, g_2), IBD3) = \frac{e^{-(lg_2\mu)}(lg_2\mu)^{n_{AB}}}{n_{AB}!} I(n_{AC}=0)I(n_{BC}=0)$$
$$P_\mu(mut3|tree3 = (AC-B, g_1, g_2), IBD3) = \frac{e^{-(lg_2\mu)}(lg_2\mu)^{n_{AC}}}{n_{AC}!} I(n_{AB}=0)I(n_{BC}=0)$$
$$P_\mu(mut3|tree3 = (BC-A, g_1, g_2), IBD3) = \frac{e^{-(lg_2\mu)}(lg_2\mu)^{n_{BC}}}{n_{BC}!} I(n_{AB}=0)I(n_{AC}=0)$$

2.4 Mutation rate estimation with 3-way IBD

The sections above present the components needed to obtain the overall mutation-rate likelihood. Here we combine these components to give the overall likelihood for one set of three-way IBD for three haplotypes around a given position in the genome. The data provide multiple such sets of three-way IBD, and we multiply the likelihoods for each such set. Such sets of three IBD haplotypes are not fully independent, because IBD often occurs in clusters of more than three haplotypes, and we analyze each subset of three haplotypes from such a cluster. Thus the overall likelihood obtained by multiplication is a composite likelihood.

For each set of three IBD haplotypes that we observe in the data, with IBD lengths recorded in IBD3 and apparent mutation counts recorded in mut3, the likelihood of the mutation rate given the data can be obtained using the law of total probability as

$$\begin{aligned} L(\mu) &= P_\mu(IBD3, mut3) \\ &= \sum_{tree3} P_\mu(IBD3, mut3, tree3) \\ &= \sum_{tree3} P_\mu(mut3|tree3, IBD3)P(IBD3|tree3)P(tree3) \end{aligned}$$

The sum over possible coalescent trees, tree3, includes an infinite number of possible trees, however only those with low to moderate coalescent times are consistent with the long IBD segments that we use. In practice we restrict the sum to positive integer coalescent times $g_1 \leq 300$ and $g_2 \leq 300$, as these limits proved to be sufficient in our simulation studies and data analyses, and we sum over the three possible orderings of the coalescent events.

With a large number of such sets of three IBD haplotypes, we can estimate the mutation

rate with precision. We numerically maximize the composite likelihood by performing a grid search. To reduce the computing time required for performing a grid search, we use adaptive grids. We first obtain estimates for the mutation rate from a coarse search grid. We then refine the estimates by applying a finer search grid to a targeted area based on the confidence interval of the initial estimates. For example, we start a search grid for mutation rate on interval $[8.0 \times 10^{-9}, 1.8 \times 10^{-8}]$ with a step size of 1.0×10^{-9}. If we obtain a confidence interval of $[1.2 \times 10^{-8}, 1.4 \times 10^{-8}]$ with the coarse search grid, we further narrow the search grid to interval $[1.1 \times 10^{-8}, 1.5 \times 10^{-8}]$ and reduce the step size from 1.0×10^{-9} to 5.0×10^{-10}. We simultaneously adjust search grids for other parameters (as described in Section 4.2) in a similar way. By repeating the above steps, we can reduce the step size to an ideal resolution.

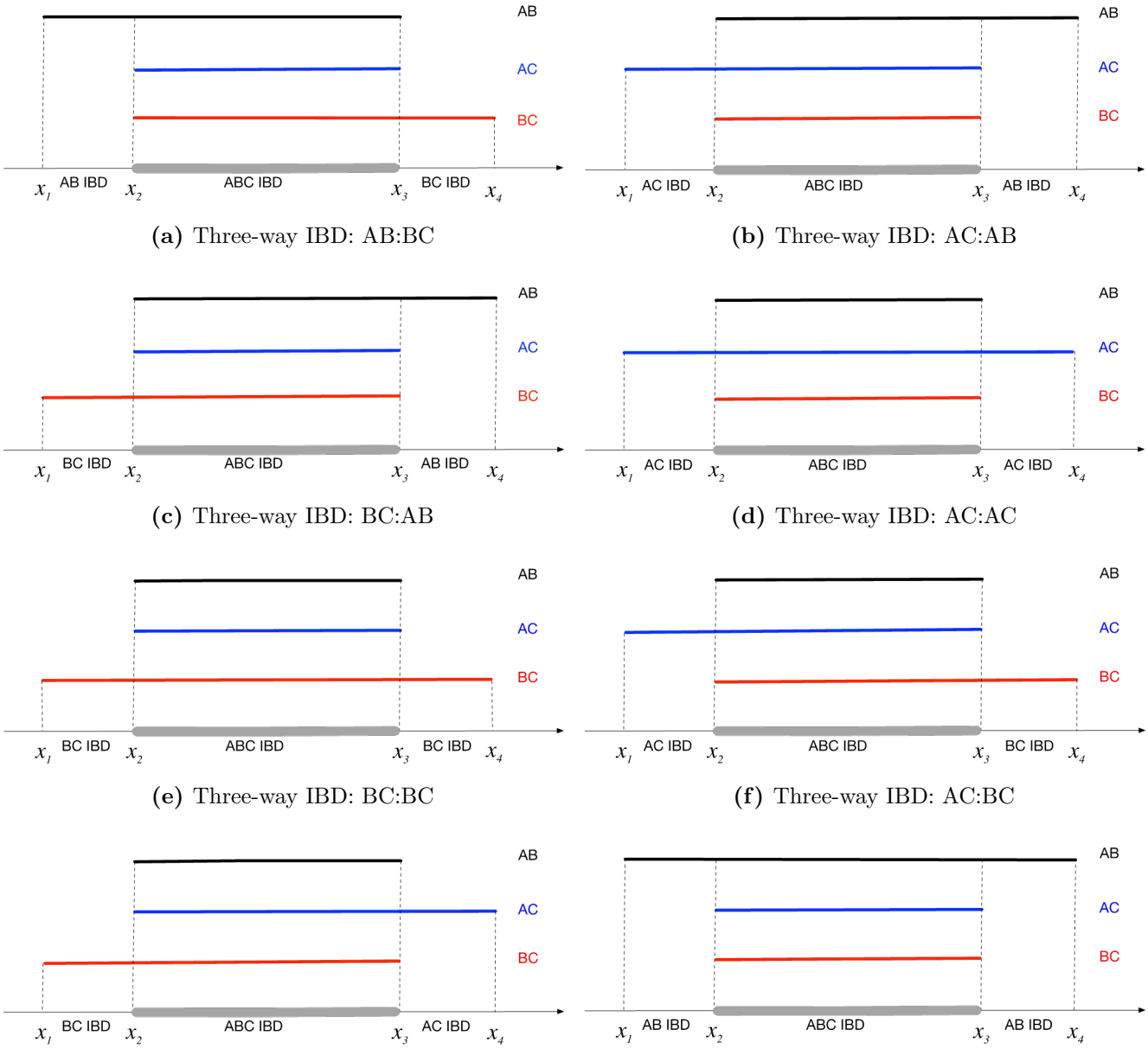
We use bootstrap resampling to assess the precision of the estimated mutation rate. We resample chromosomes (from the 22 autosomal chromosomes) with replacement and obtain a maximum likelihood estimate from the sampled chromosomes in each bootstrap sample. The 95% confidence interval is determined from the 2.5th and 97.5th percentiles of 10,000 bootstrap estimates.

2.4.1 Simulation with true IBD segments

We evaluated the performance of the proposed method on error free simulated data with true IBD segments, known haplotype phase, known demographic history. We used ARGON [41], a discrete-time Wright-Fisher process simulator, to simulate 2000 diploid individuals sampled from a homogeneous population with constant effective population size of 10,000 diploid individuals (the homogeneous model'). The simulated genome size was 30 chromosomes of 100Mb each, with a mutation rate of 1.30×10^{-8} per base pair per meiosis, and a constant recombination rate of 1.0×10^{-8} per base pair per meiosis. In order to reduce the impact of IBD detection errors for future real data analysis, we impose a 3cM minimum length threshold on the pairwise IBD and use this threshold to analyze true IBD segments generated by

ARGON. We then find overlapping IBD segments shared by sets of three individuals. In the three-way IBD regions (e.g. region ABC from x_2 to x_3 in Figure 2.3) we should have IBD between all three pairs of the three individuals (e.g. AB, AC, and BC in Figure 2.3) and the endpoints are necessarily consistent between the three pairwise IBD segments (e.g. AC and BC have the same reported left endpoint x_2 in Figure 2.3). If one of the three IBD segments was not reported (for example if we found AB and AC but not BC), we do not include the three-way IBD segment in the analysis.

Under the homogeneous model with true IBD segments, true effective population size, and true phase, we obtained a mutation rate estimates of 1.30×10^{-8} per base pair per meiosis with a 95% confidence interval of $[1.29 \times 10^{-8}, 1.31 \times 10^{-8}]$ (the simulated mutation rate is 1.30×10^{-8}). The likelihood framework we proposed can provide accurate estimates under the ideal scenario. However, elements of this ideal scenario including the availability of the true IBD segments, known demographic history, perfect phasing quality, absence of genotype errors, and absence of gene conversion are not applicable to real data, so we modify our method to overcome these challenges in the following chapters.



(a) Three-way IBD: AB:BC

(b) Three-way IBD: AC:AB

(c) Three-way IBD: BC:AB

(d) Three-way IBD: AC:AC

(e) Three-way IBD: BC:BC

(f) Three-way IBD: AC:BC

(g) Three-way IBD: BC:AC

(h) Three-way IBD: AB:AB

Figure 2.5: Patterns of three-way IBD. The three-way IBD pattern is coded as W:Z which represents the IBD segment shared by pair W starts at x_1, ends at x_3; the IBD segment shared by pair Z starts at x_2, ends at x_4. The gray region from x_2 to x_3 is the IBD region shared jointly by three haplotypes.

Chapter 3

INFERENCE OF IBD SEGMENTS, AND ACCOUNTING FOR GENOTYPE CALLING ERRORS AND UNKNOWN DEMOGRAPHIC HISTORY

3.1 Apparent mutation counts

The identification of candidate germline mutations requires high-precision sequence data. While genotype calling accuracy continues to improve with developments in sequencing technology, genotype calling errors, particularly at sites of low minor allele frequency, are inevitable due to imperfect sequencing technologies and limitations of current genotype calling algorithms ([40, 44, 59]). For the pedigree-based mutation rate estimation approach, it is difficult to quantify sources of uncertainty due to the choice of quality control filters to reduce the impact of genotype error. Overly stringent filtering will depress the mutation rate estimates [49].

Unlike pedigree-based mutation rate estimation, our method uses cross-family IBD to identify mutations arising over a much larger number of meiosis. With the three-way IBD setting in our approach, we count rare variants shared by two of the three individuals. This avoids the use of singleton variants which have higher genotype error rates [46, 25], and it requires that two genotype errors are need to create any false apparent mutation. Let p be the frequency at which the derived allele is mis-called as ancestral allele, and q be the frequency at which the ancestral allele is mis-called as derived. We assume that mis-calling errors are independent of IBD status. In an IBD trio, the probability of calling a derived allele in each of two IBD haplotypes includes the following components:

- Type 0: variants existing in the data set, but with zero true derived alleles on the

IBD haplotypes. Two mis-calls of ancestral as derived allele in different haplotypes are required to obtain two haplotypes in the trio carrying apparent derived. Probability of double mis-call = $\binom{3}{2}q^2(1-q)$.

- Type 1: one true derived allele on one of three the IBD haplotypes. One mis-call of ancestral as derived allele on one of the other individuals genotypes is needed to obtain two individuals in the trio carrying the apparent derived. Probability of the mis-call = $\binom{2}{1}q(1-q)(1-p)$.

- Type 2: two true derived alleles on two of the three IBD haplotypes.

- Type 3: three true derived alleles on the IBD haplotypes. One mis-call of a derived allele as ancestral is needed to create a false apparent mutation. Probability of the mis-call = $\binom{3}{1}p(1-p)^2$.

We investigated simulated data to determine the rates of such types in IBD trios. In one chromosome of 2000 individuals with 100Mb of simulated data (homogeneous model in Section 2.4.1, i.e. constant population size), we found around 760,000 IBD trios. We considered variants with up to 150 copies in the sample. Each trio has 12685 type 0, 4.2 type 1, 1.3 type 2, and 78 type 3 positions on average. If $p = q = 0.05\%$ there will be an average of 0.0095 type 0 positions, 0.0042 of type 1, and 0.1175 type 3 with two observed (mis-called) derived alleles. Thus under this scenario, of the observed positions in an IBD trio with two called-derived alleles in two of the individuals, over 90% of them represent type 2.

We further reduce the impact of genotype errors through statistical modeling. As described in Section 2.3, given the coalescent tree in Figure 2.2, the number of mutations shared by haplotypes A and B but not C across a region of l base pairs within the three-way IBD sharing region is distributed as Poisson($lg_2\mu$). When the dataset contains genotype errors, some apparent mutations in this region may actually be the result of genotype error. We assume that the rate ϵ of errors of this type (i.e. a miscalled allele in a specific two of three

haplotypes) is constant and does not depend on the coalescence times. Thus across a region of l base pairs, the number of errors of this type is Poisson with rate $l\epsilon$. In consequence, the number of apparent mutations shared by A and B but not C (real mutations and errors) across the region is Poisson with rate $l(g_2\mu + \epsilon)$. In contrast, considering two of three haplotypes that are not the first coalescing pair, such as haplotypes A and C for this coalescent tree, any apparent mutations shared by these two haplotypes but not the third will be genotype errors rather than real mutations because they are inconsistent with the coalescent tree (the probability of recurrent mutation is negligible, and is ignored under the infinite sites model). Thus the number of such apparent mutations between any two haplotypes which do not coalesce first is modeled as Poisson with rate $l\epsilon$ and the probability distribution of mutation counts given the coalescent tree becomes the following and we numerically maximize the composite likelihood with respect to both μ and ϵ (Figure 3.1).

$$P_{\mu,\epsilon}(mut3|tree3 = (AB - C, g_1, g_2), IBD3) = \frac{e^{-(lg_2\mu)}(lg_2\mu)^{n_{AB}}}{n_{AB}!}\frac{e^{-l\epsilon}(l\epsilon)^{n_{AC}}}{n_{AC}!}\frac{e^{-l\epsilon}(l\epsilon)^{n_{BC}}}{n_{BC}!}$$
$$P_{\mu,\epsilon}(mut3|tree3 = (AC - B, g_1, g_2), IBD3) = \frac{e^{-(lg_2\mu)}(lg_2\mu)^{n_{AC}}}{n_{AC}!}\frac{e^{-l\epsilon}(l\epsilon)^{n_{AB}}}{n_{AB}!}\frac{e^{-l\epsilon}(l\epsilon)^{n_{BC}}}{n_{BC}!}$$
$$P_{\mu,\epsilon}(mut3|tree3 = (BC - A, g_1, g_2), IBD3) = \frac{e^{-(lg_2\mu)}(lg_2\mu)^{n_{BC}}}{n_{BC}!}\frac{e^{-l\epsilon}(l\epsilon)^{n_{AB}}}{n_{AB}!}\frac{e^{-l\epsilon}(l\epsilon)^{n_{AC}}}{n_{AC}!}$$

We also investigated the impact on our method of false negative errors, in which copies of the minor allele are mis-called as the major allele. Variants with lower allele frequency are more susceptible to false negative error and these rare variants are usually carried by long IBD segments since they generally arose recently. Hence three-way IBD that involves very long pairwise IBD segments has the most potential to be affected by false negative error. To control the downward bias caused by false negative error, we thus restricted the analysis to IBD segments with length below some threshold. Reducing the maximum length threshold reduces the number of IBD segments that can be used, thus increasing the variance of the estimation. The maximum length threshold is determined from simulation studies so that

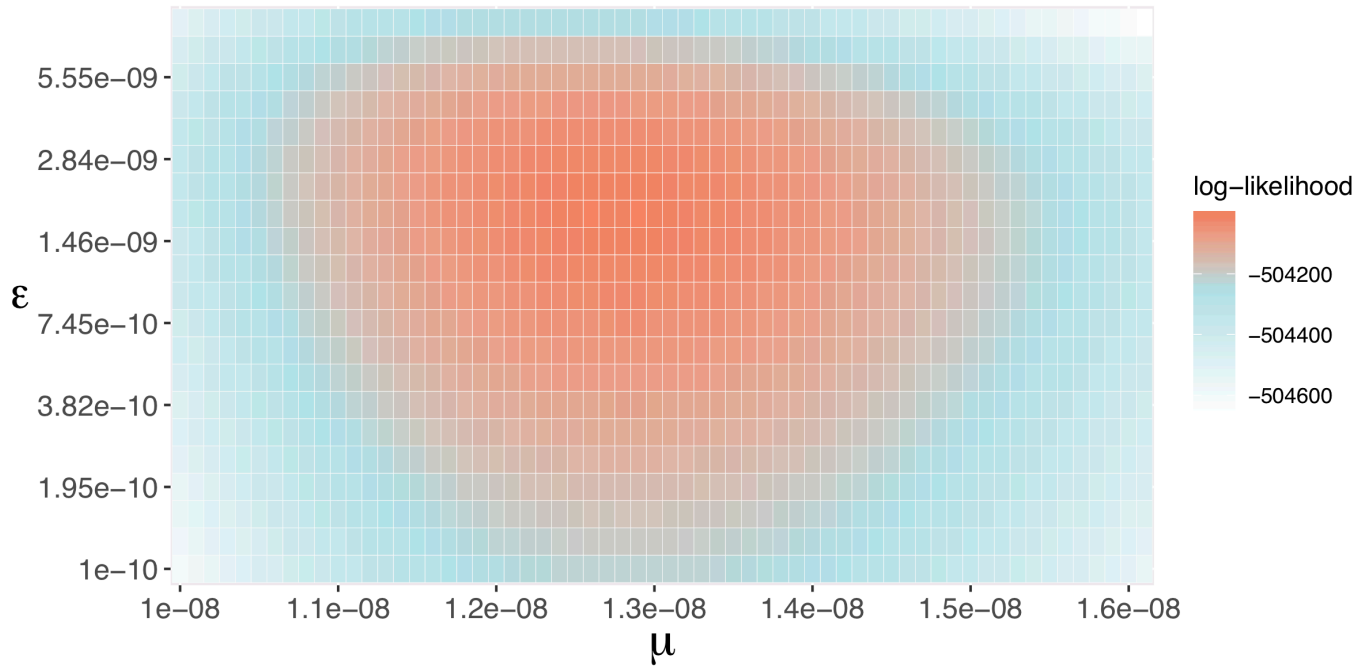


Figure 3.1: An example of the likelihood contour for the mutation rate. Data were simulated under the super-exponential model with errors simulated using the unbiased genotype error scheme (described in Methods). The simulated mutation rate is $1.30 \times 10^{-}8$ per base pair per meiosis, and the error rate is 0.02%. The sample size is 2000 individuals. The likelihood is a function of two parameters: the mutation rate, and an error parameter which is used to control for false apparent mutations cause by genotyping errors. The error parameter is less than the genotype error rate because many genotype errors are removed by the requirement that two of the three IBD haplotypes carry the allele. We use an adaptive search grid to find the values of the parameters that maximize the likelihood.

it provides good control of potential downward bias due to false negative errors while not overly increasing the variance of estimation.

3.2 IBD detection in sequence data

We used Refined IBD [5] in BEAGLE version 4.1 to detect pairwise IBD segments from phased genotypes using only diallelic SNPs with minor allele frequency 10% or higher, with a minimum LOD score threshold of 3 and a minimum length threshold of 1 cM. The minor allele frequency threshold reduces computation time compared to using more variants, and ensures that recent mutations that could contribute to the mutation rate estimation are not used in the IBD detection. We used a minimum length threshold of 1 cM because most IBD segments with a LOD score of 3 or higher have length greater than 1 cM, and because using a smaller threshold would increase computation time. Refined IBD uses a haplotype-based method to detect IBD segments. Consequently, genotype errors and haplotype phase errors can result in gaps in the estimated IBD segments and underestimation of the length of IBD segments. We filled gaps between two detected IBD segments for the same pair of haplotypes when the gap between the IBD segments had a length less than 0.5 cM and the gap contained at most two positions at which the genotypes for the two individuals were inconsistent with IBD (Figure 3.2). The choice of parameters for gap-filling is based on sensitivity analysis of simulated data. This gap-filling step has been shown to make IBD length estimation robust to genotype errors [7]. After the gap-filling step, we impose a 3 cM minimum length threshold on the pairwise IBD segments. We then find overlapping IBD segments shared by sets of three individuals. In the three-way IBD regions (e.g. region ABC from x_2 to x_3 in Figure 2.3) we should have detected IBD between all three pairs of the three individuals (e.g. AB, AC and BC in Figure 2.3). If one of the three IBD segments was not detected (for example if we found AB and AC but not BC), we do not include the three-way IBD segment in the analysis. Because the detected IBD is based on haplotype identity-by-state, the endpoints are necessarily consistent between the three pairwise IBD segments (e.g. AC and BC have

the same reported left endpoint x_2 in Figure 2.3).

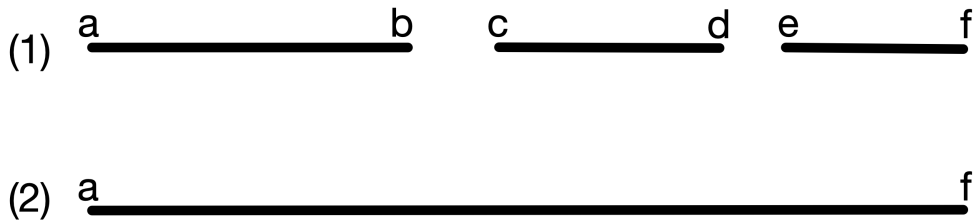


Figure 3.2: An example of the gap-filling procedure. The three detected Refined IBD segments for one pair of haplotypes are shown in (1) as 'ab', 'cd', and 'ef'. If the gap 'bc', between the 'ab' and 'cd' segments has a maximum length of 0.5 cM and the maximum number of genotypes in 'bc' that are inconsistent with IBD is 2, this gap will be filled, as shown in (2). Similar rules are applied to the gap de between the 'cd' and 'ef' segments.

When counting possible mutations, we trim 0.5cM from each end of the region in which we detect three-way IBD sharing. The reason for this trimming is that the observed identity by state often extends somewhat beyond the true IBD region [42, 5]. The number of apparent mutations is the number of rare variants shared by two of the three individuals in this trimmed region. We then adjust the base pair length l, of shared region in the calculation of $P_{\mu,\epsilon}(mut3|tree3, IBD3)$ accordingly.

3.3 Demographic inference

While site frequency spectrum (SFS) based methods have been used extensively for demography inference [39, 23, 21, 28, 33], SFS is not ideal for inferring recent demographic history. SFS utilizes rare variants to infer recent population history, thus a large sample size is required. And the ascertainment of rare variants is subject to false positive and false negative genotype errors which are difficult to account for [28]. Moreover, recent studies showed dis-

tinct demographic models can result in strongly similar SFS [51].

Approximate methods for Ancestral recombination graph (ARG) inference can be used to estimate effective population size, such as methods based on sequentially Markovian coalescent (SMC) [36, 31, 48]. This class of methods utilize whole genome sequences from one [31] or more diploid individuals [48] to infer local time to most recent ancestor (TMRCA), which are informative of past demographic history. However, this type of method can only be applied to a small number of individuals [48] due to computation complexity and thus has limited power to recover recent demographic changes.

Methods like Doris [43] and IBDNe [11] exploit the relationship between the length distribution of IBD segments, TMRCA, and effective population size to infer demographic history. IBD based methods utilize long IBD segments and thus reflect recent demographic history [11]. The parametric approach that Doris takes require a set of pre-specificied demographic parameters [43], which makes it difficult to capture complex demographic histories. IBDNe is based on a non-parametric approach and thus has the flexibility to infer complex population histories [11]. Therefore, we choose to use IBDNe for demographic inference.

The input to the IBDNe program was the set of pairwise gap-filled IBD segments with a minimum length of 2 cM. For the dataset including closely related individuals, we first identified pairs of closely related individuals as those with total length of detected IBD segments exceeding 5% of the genome length, and we removed IBD segments corresponding to such pairs from the IBDNe analysis.

3.4 Validation with simulated data

Other than the model with constant population size (described in Section 2.4.1 as homogeneous model), we simulated two other models, the 'super-exponential' model, and the

'admixture' model. We added genotype errors at different rates to investigate the robustness of our method to genotype errors. We estimated mutation rate with inferred IBD segments and inferred demographic history. The demographic history of the simulated models are shown in Figure 3.3 - Figure 3.5.

In the 'admixture' scenario, we simulated 400 admixed diploid individuals with non-constant effective population size and population structure. The population size was 15,000 until 1000 generations ago, at which time a population split occurred, resulting in two subpopulations with sizes of 10,000 and 5,000. The two subpopulations merged together (admixed) 20 generations ago and then grew at a rate of 1.44% per generation to a current size of 20,000. By simulating a set of unadmixed individuals, we obtained a Weir and Cockerham [55] weighted fixation index (Fst) estimate of 0.07. This Fst is greater than Fst estimates for pairs of European populations, and smaller than Fst estimates for inter-continental pairs [18].

In the 'super-exponential' scenario, we simulated genome-wide data for 10,000 diploid individuals with a demographic model that matches the heterozygosity, magnitude of linkage disequilibrium, and rate of IBD observed in the UK10K sequence data [6]. The demographic model has an initial population size of 24,000 in the distant past, with an out-of-Africa reduction to 3,000 occurring 5,000 generations ago. The population begins to grow 1.4% per generation 300 generations ago. The growth rate increases to 6% and 25% at 60 and 10 generations ago, respectively. The final effective population size is around 21 million.

We created two versions of the simulated data, each with a different type of genotype error. The first type of genotype error includes both false positive errors (major allele mis-called as minor) and false negative errors (minor allele mis-called as major). For diallelic SNPs with minor allele frequency p, we give each allele a probability $\min(\delta, 2p)$ of being changed to the other allelic form (major to minor or vice versa), with the error rate δ taking values of 0.01%, 0.05%, and 0.1% for the homogeneous and admixture models. As the com-

putation time is high for the super-exponential model due to the large sample size, we only added an error rate of 0.02% for this model. We refer to this first error scheme as 'unbiased' error because the rate of error doesn't depend on whether the true allele is the major or minor allele. For rare variants, most of the added errors under this scheme are false positive errors, because there are relatively few minor alleles that could be changed to the major allele. We also wanted to further investigate the effect of false negative error, which may be more prevalent for rare variants. Thus we created a second set of data in which we added only false negative error. For variants with m minor allele copies present in the dataset, we give each copy probability 0.5^m of being changed to the major allele. We refer to this second error scheme as 'false-negative' error. For both types of genotype error, the errors are added prior to IBD detection; therefore, the presence of genotype errors can affect IBD detection, subsequent inference of demographic history with IBDNe, and mutation ascertainment.

Under the super-exponential scenario with a genotype error rate of 0.02%, we obtained a mutation rate estimate of 1.29×10^{-8} per base pair per meiosis with a 95% confidence interval of $[1.281 \times 10^{-8}, 1.301 \times 10^{-8}]$ (the simulated mutation rate is 1.30×10^{-8}). We also obtained accurate estimates of mutation rate under the homogeneous and admixture simulation scenarios (Figure 3.6). Accuracy is maintained even with the highest rate of genotype error considered (0.1%), however at high rates of genotype error fewer segment of IBD are detected and hence confidence intervals are wider.

We also applied IBDMUT, the pairwise IBD method of Palamara et al. [42], to the dataset simulated under the super-exponential scenario. In the IBDMUT analysis, we used IBD segments estimated by GERMLINE [22] with one or two allowed mismatching sites and we used the same estimated demographic history as we used for our method, obtained using IBDNe. Since IBDMUT requires more than 250 gigabytes of memory to process the full simulated dataset, we analyzed a subset of 2000 individuals. We observe biased estimates of mutation rate whether allowing for one or two mismatches, whether using the true or

estimated demographic history, and whether or not the data include genotype errors (Figure 3.7). Up to 8% relative bias is observed.

To investigate the impact on our method of false negative errors, in which copies of the minor allele are mis-called as the major allele, we added genotype errors according to our false-negative genotype error scheme to the homogeneous simulated dataset. In this setting, we found reducing the maximum length threshold to 6 cM provides good control of potential downward bias due to false negative errors while not overly increasing the variance of estimation (Figure 3.8).

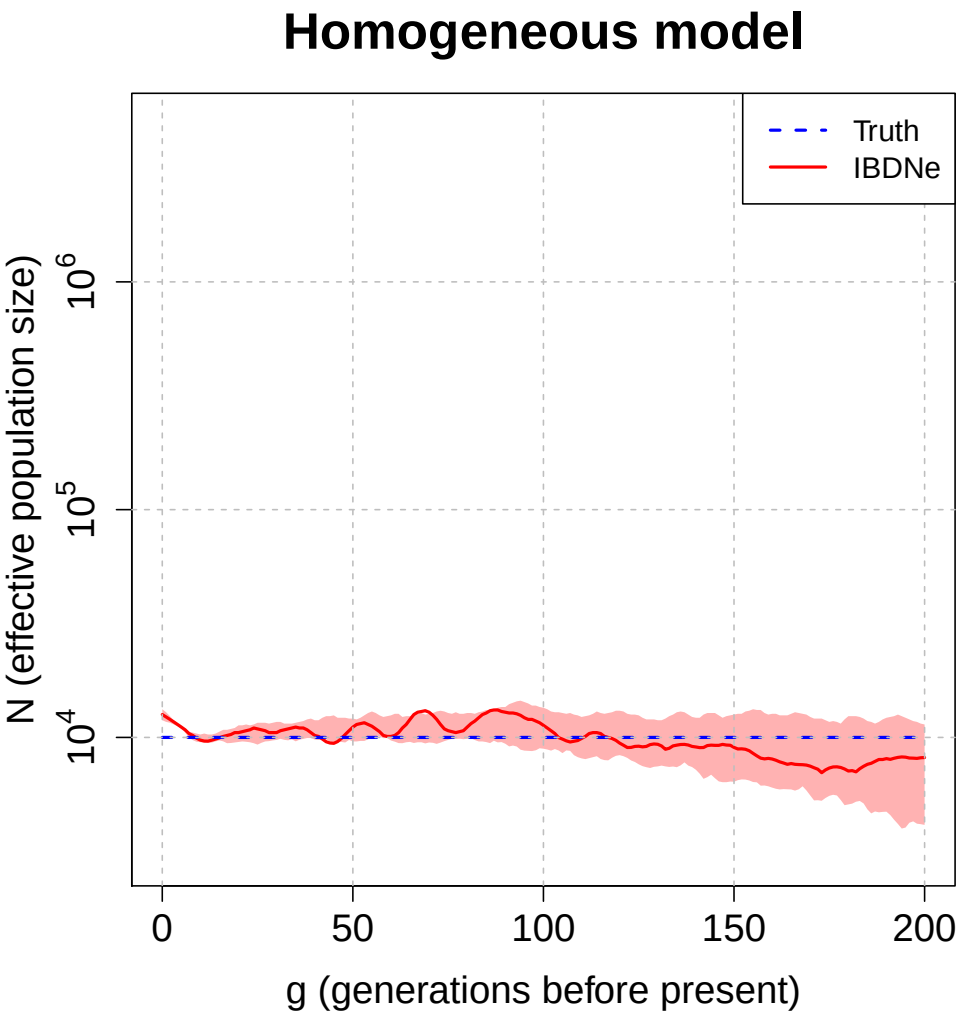


Figure 3.3: Demographic history of 'homogeneous' model. The simulated and estimated recent effective population size for the 'homogeneous' model. Generations before present are shown on the x-axis, and effective population size is shown on the y-axis. The blue dashed line gives the truth, the red line gives the estimated size while the red region gives the 95% bootstrap confidence intervals. The estimates of effective population size comes from applying IBDNe on IBD segments reported from Refined IBD running on dataset with 0.01% genotype error.

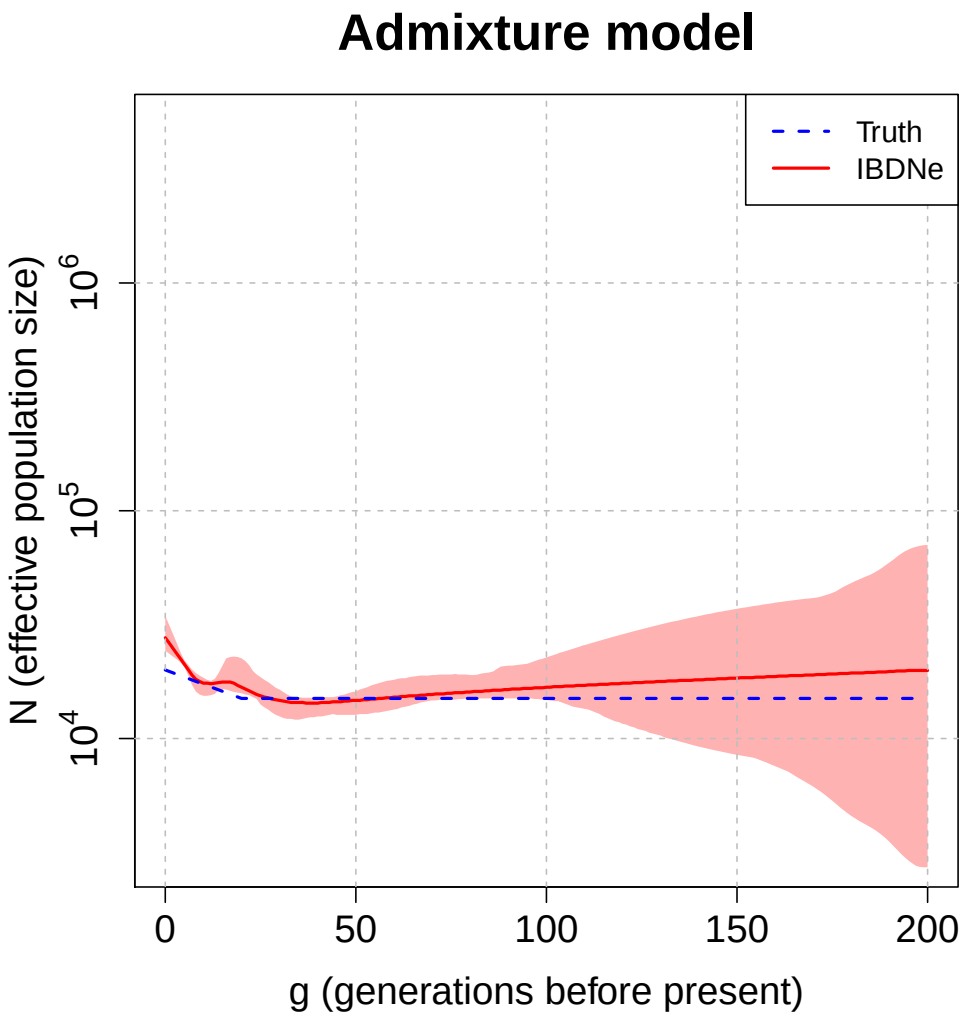


Figure 3.4: Demographic history of 'admixture' model. The simulated and estimated recent effective population size for the 'admixture' model. Generations before present are shown on the x-axis, and effective population size is shown on the y-axis. If there are two populations in the past, the y-axis is showing the total effective population size. The blue dashed line gives the truth, the red line gives the estimated size while the red region gives the 95% bootstrap confidence intervals. The estimates of effective population size comes from applying IBDNe on IBD segments reported from Refined IBD running on dataset with 0.01% genotype error.

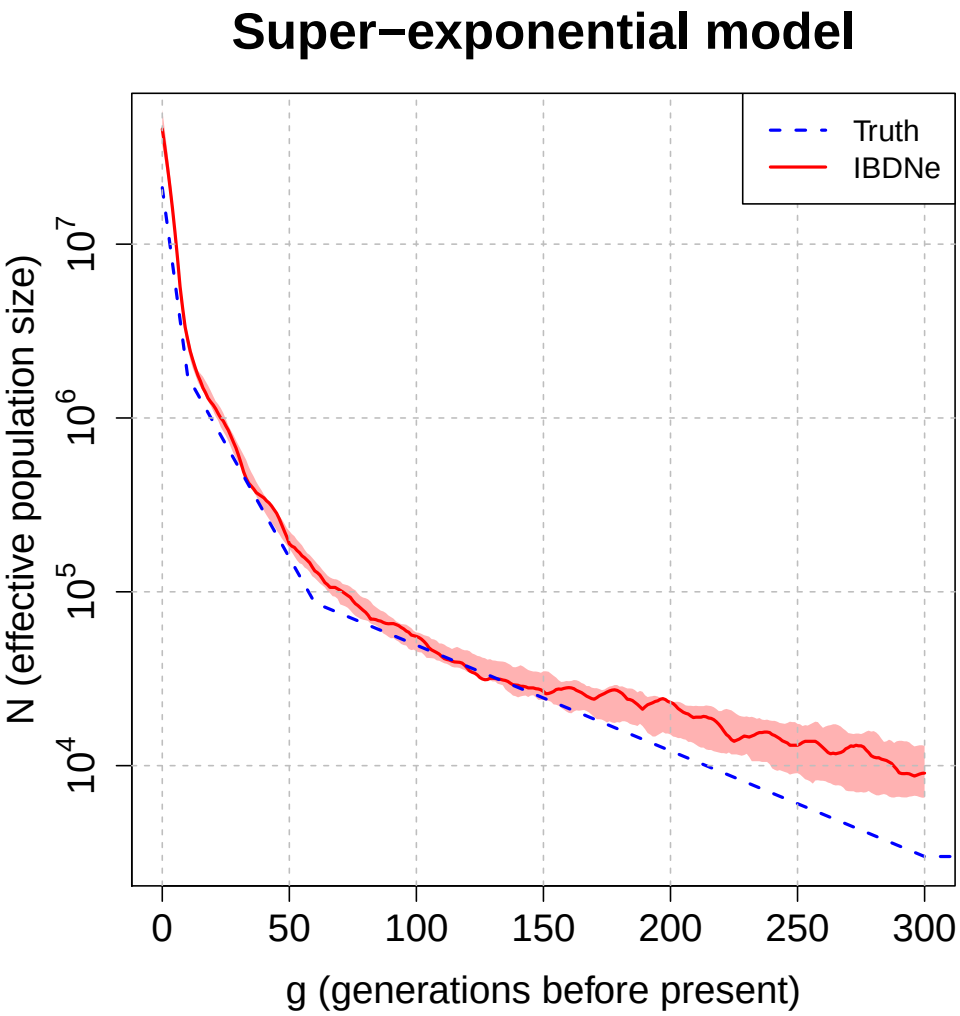


Figure 3.5: Demographic history of 'super-exponential' model. The simulated and estimated recent effective population size for the 'super-exponential' model. Generations before present are shown on the x-axis, and effective population size is shown on the y-axis. The blue dashed line gives the truth, the red line gives the estimated size while the red region gives the 95% bootstrap confidence intervals. The estimates of effective population size comes from applying IBDNe on IBD segments reported from Refined IBD running on dataset with 0.02% genotype error.

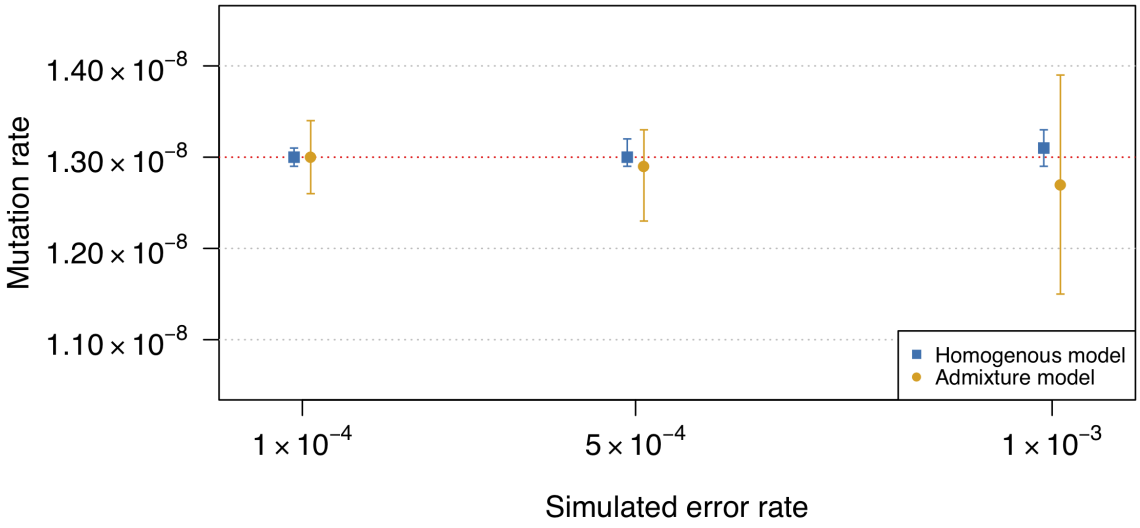


Figure 3.6: Estimated mutation rates under different rates of genotype error. The simulated mutation rate is 1.30×10^{-8} per base pair per meiosis and is indicated with the red dotted line. Point estimates (shapes) and 95% confidence intervals (bars) are shown. The maximum allele frequency threshold for the variants used for the mutation rate estimation is 3.75%. Two different simulation models, the homogeneous model (blue squares) and the admixed model (yellow circle) are assessed at error rates of 1×10^{-4}, 5×10^{-4}, and 1×10^{-3}. Errors are simulated using the 'unbiased' genotype error scheme.

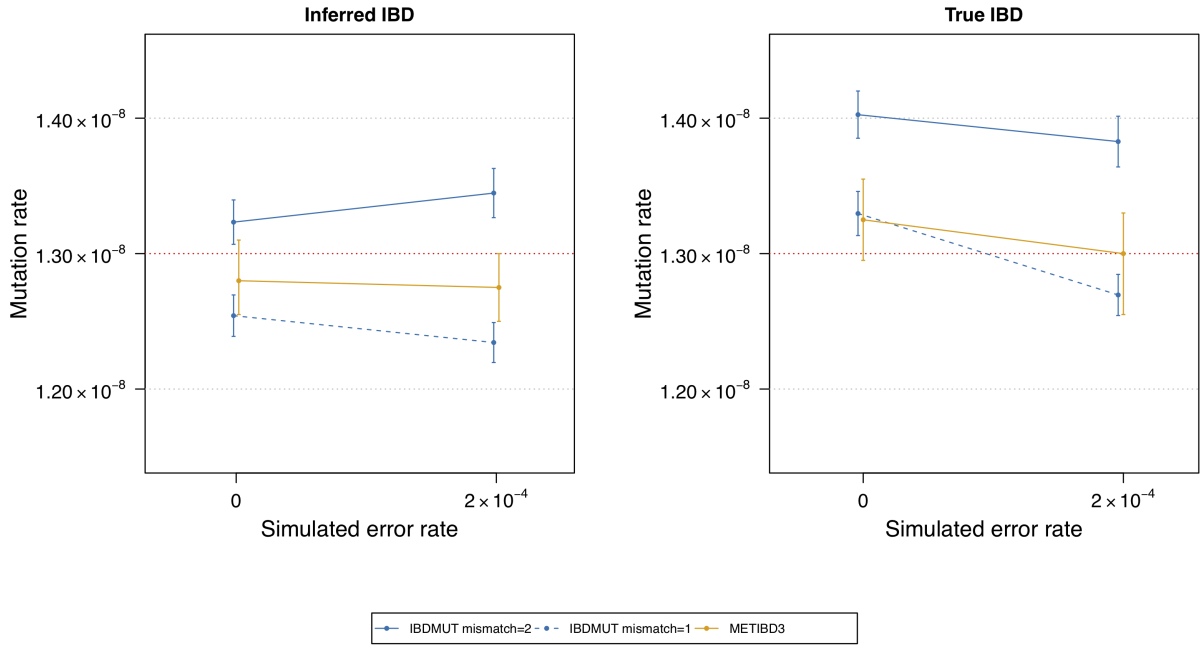


Figure 3.7: Estimated mutation rates from IBDMUT. Data were simulated under the 'super-exponential' model with errors simulated using the 'unbiased' genotype error scheme. The simulated mutation rate is 1.30×10^{-8} per base pair per meiosis and is indicated with the red dotted line. The sample size is 2000 individuals. We used GERMLINE for IBD segment detection with different values for the number of allowed mismatch sites. Genotype errors were added before IBD detection and thus influence the accuracy of IBD inference. We show results when using the inferred effective population size from IBDNe based on the Refined IBD segments (left panel), and also show results when using the true effective population size from the simulation model (right panel). Point estimates (dots and triangles) and 95% confidence intervals (bars) are shown. Results from our method (METIBD3) on the same data are shown for comparison.

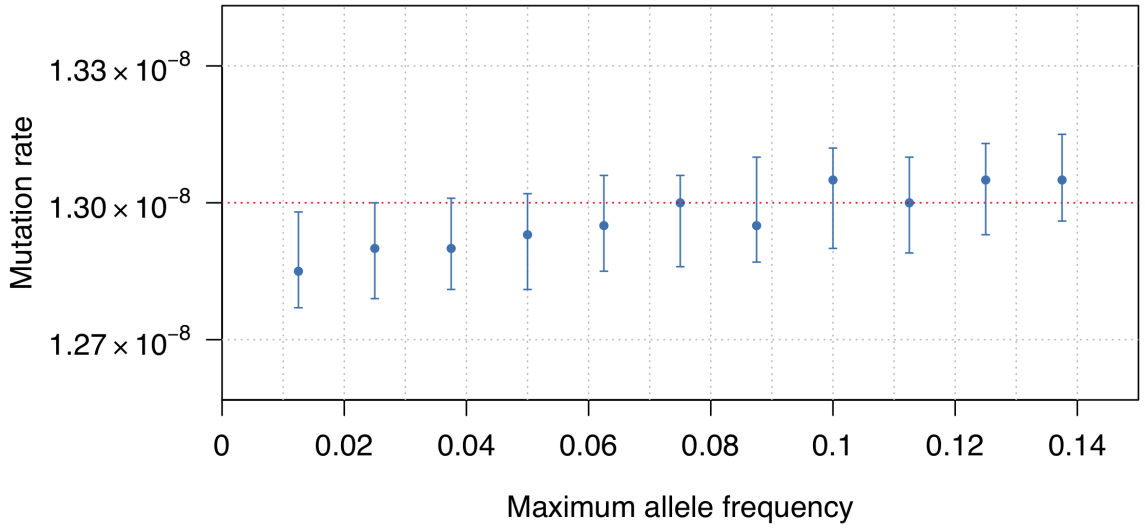


Figure 3.8: Estimated mutation rates under false negative genotype errors. Data were simulated under the 'homogeneous' model with errors simulated using the 'false-negative' genotype error scheme. The simulated mutation rate is 1.30×10^{-8} per base pair per meiosis and is indicated with the red dotted line. We set different thresholds on the maximum length of IBD segments (x-axis) to control the impact of false negative errors. The maximum allele frequency threshold for the variants used for the mutation rate estimation is 3.75%. Point estimates (dots) and 95% confidence intervals (bars) are shown.

Chapter 4

ACCOUNTING FOR NON-CROSSOVER GENE CONVERSION

4.1 Correction for gene conversion through regression

Gene conversion copies variants from one haplotype onto another. Gene conversion can create the appearance of mutation events in the three-way IBD, because a gene conversion event occurring between the common ancestor of the two most-recently coalescing haplotypes and their common ancestor with the third haplotype may introduce a variant that is shared by the two most-recently coalescing haplotypes but not by the third haplotype. The rate of gene-conversion variants in a set of three IBD haplotypes is proportional to the number of generations, g_2, between the more recent coalescence time and the less recent coalescence time (Figure 2.2), as is the number of mutation variants. In fact, gene conversions on other branches of the coalescent tree for the three IBD haplotypes can also cause apparent mutations, but as we discuss below in section 4.2, ignoring the gene conversions on the other branches does not bias the mutation rate estimate.

A major difference between gene-conversion variants and mutation variants is that mutation variants tend to be rare, while gene-conversion variants tend to be common. The probability that an allele is changed via gene conversion is proportional to the heterozygosity of the variant in the population, at the time of the gene conversion, because the recipient allele is only changed if the donor allele (the ancestral individual's other allele) is different from it. Common variants have higher heterozygosity and hence are more likely to be changed via gene conversion.

We thus use only the less common variants when counting mutations, applying a maxi-

mum allele frequency filter. However, gene conversion affects low frequency variants to some extent (albeit less than the effect on high frequency variants), and one cannot set the allele frequency threshold too low or one risks removing some actual mutations. Hence we apply a modified version of the regression adjustment developed by Palamara et al [42].

Consider an allele frequency threshold, f. Only alleles with frequency below this threshold will be included in the apparent mutation counts. If f is sufficiently large, then all mutations occurring on the branch from the most recent common ancestor of all three IBD haplotypes to the most recent common ancestor of the two most recently related IBD haplotypes will have frequency less than f and will be included in the mutation count. If the length of this branch is g_2 generations, then the expected number of such mutations contributing to the overall mutation count across a region of length l basepairs is $lg_2\mu$ as previously described. During the g_2 meioses occurring on this branch, the expected number of basepairs involved in a gene conversion event is $lg_2\theta$, where θ is the rate of gene conversion initiation (per bp per generation) multiplied by the mean gene conversion tract length (in bp). Let h_f denote the rate of heterozygosity (per bp), excluding variants with minor allele frequency greater than f. The probability that a given base is changed and that the change is to an allele with frequency less than f, conditional on the base being involved in a gene conversion event, is the probability that the position was heterozygous in the individual in which the gene conversion took place, and that the heterozygous genotype had the minor allele on the donor haplotype, which is $h_f/2$. Thus the expected contribution to the mutation rate count from gene conversion is $lg_2\theta h_f/2$, the total expected mutation rate count is $lg_2(\mu+\theta h_f/2)+l\epsilon$, and the nominal mutation rate estimated by our method is $\mu_f^* = \mu+\theta h_f/2$. Thus if we apply the method with different values of f (and corresponding different values of h_f), we can regress the estimated nominal mutation rate estimates $\hat{\mu_f}$ against estimated frequency-bounded heterozygosity $\hat{h_f}$ to obtain an estimate of the mutation rate μ in the intercept of the regression.

In the above, we implicitly assumed that the frequency-bounded heterozygosities, h_f, are

fixed across time and geography. Although this assumption may appear doubtful, in fact autosomal heterozygosities are almost identical across populations in Europe [1], for example, even including Finland which has experienced a recent population bottleneck. Thus, for relatively homogeneous (single continental-level ancestry) populations, the assumption of stable heterozyosities gives a reasonable approximation.

For a given allele frequency threshold f, the estimated heterozygosity is

$$\hat{h_f} = \sum_i 2p_i(1-p_i)1_{\{p_i<f\}}/L$$

where the sum is over all variants (indexed by i) in a genome of total length L basepairs. The minor allele frequency of each such variant is p_i and $1_{\{p_i<f\}}$ is the indicator function which takes value 1 if the minor allele frequency is less than f and 0 otherwise. Here we use the expected heterozygosity based on allele frequency and assuming Hardy-Weinberg equilibrium, whereas we could instead use observed heterozygosity; we find that the two approaches give almost equal estimates in the simulated data and real data analysis).

To perform the regression, we use only allele frequency thresholds that are large enough to minimize the possibility of excluding some true mutations. We thus use frequency thresholds between 0.1 and 0.5 in the regression (following Palamara et al. [42]). We choose to use an upper bound of 0.5 to maximize the amount of data in the regression, and we use a lower bound of 0.1 which is high enough to ensure that no true mutations are excluded, yet not so high as to exclude much data from the regression. We also analyzed the data using lower bounds of 0.05 and 0.2 to ensure that the results were robust to this choice. For each frequency threshold, f, we estimate the mutation rate $\hat{\mu^*_f}$, considering as potential mutations only variants with frequency below the threshold. We estimate heterozygosity $\hat{h_f}$ across all variants in the data with frequency below the threshold. We then perform the regression, with the y-axis (mutation rate) intercept providing a gene-conversion-adjusted estimate of mutation rate.

We used MaCS [15], a Markovian coalescent simulator, to simulate data with gene conversion under a European-American demographic history. The European-American model uses the demographic history that we inferred using IBDNe from samples in Framingham Heart Study (Figure 4.1) for generations 1-300, with a population size of 24,000 prior to 5,000 generations ago and a reduction to 6,930 occurring 5,000 generations ago. We simulated 2000 diploid individuals under this scenario. The simulation included gene conversion, with a gene-conversion initiation rate of 1.0×10^{-8} per base pair per meiosis and mean conversion tract length of 100bp [20]. The simulated genome size was 30 chromosomes of 100 Mb each, with a mutation rate of 1.30×10^{-8} per base pair per meiosis, and a constant recombination rate of 1.0×10^{-8} per base pair per meiosis.

We find that the estimated mutation rate continues to increase with increasing maximum frequency of the alleles included in the analysis (Figure 5), as expected under gene conversion. In contrast, under the same simulation scenario but without gene conversion events, we observe that the estimates of mutation rate remain the same for maximum allele frequencies above 2.5%. To correct for the impact of gene conversion events, we performed a regression on heterozygosity (Figure 4.2), and obtained a mutation rate estimate of 1.34×10^{-8} per base pair per meiosis with a 95% confidence interval of $[1.30 \times 10^{-8}, 1.38 \times 10^{-8}]$ (the simulated mutation rate is 1.30×10^{-8}). When we adjusted the lower bound on the range of allele frequency thresholds used in the regression(from 0.1) we obtained similar results: 1.33×10^{-8} $[1.30 \times 10^{-8}, 1.37 \times 10^{-8}]$ for a lower bound of 0.05 and 1.35×10^{-8} $[1.30 \times 10^{-8}, 1.39 \times 10^{-8}]$ for a lower bound of 0.2.

4.2 Incorporating a correction for gene conversion into the likelihood framework

While the regression approach adjusts the effect of gene-conversion through post-processing the maximum likelihood estimates of allele frequency bounded mutation rates, gene conversion events could be modeled parametrically through the likelihood framework so that the gene conversion rate can be estimated jointly with the mutation rate.

As described in Section 4.1, the regression adjustment considers apparent mutations due to gene conversion events occurring between the common ancestor of the two most-recently coalescing haplotypes and their common ancestor with the third haplotype. Thus the total expected apparent mutation count is $lg_2(\mu + \theta h_f/2) + l\epsilon$ when considering alleles with frequency below threshold f. If f is sufficiently large, then all mutations occurring on the branch from the most recent common ancestor of all three IBD haplotypes to the most recent common ancestor of the two most recently related IBD haplotypes will be included in the mutation count. For a sufficiently large f, both mutation events and gene conversion events contribute to the apparent mutation counts, therefore, the number of apparent mutations given the consistent coalescent tree can be modeled with a Poisson distribution with mean $lg_2(\mu + \theta h_f/2) + l\epsilon$. Since mutation variants tend to be rare, when we count apparent mutations using common alleles with frequency above threshold f, such apparent mutation counts are expected to be contributions from gene conversion events only. Let h'_f denote the rate of heterozygosity (per bp) for variants with minor allele frequency above f, the number of apparent mutations can be modeled with a Poisson distribution with mean $lg_2\theta h'_f/2 + l\epsilon$. In consequence, the number of apparent mutations given a consistent coalescent tree is divided into two bins defined by the allele frequency threshold f and modeled through two Possion distributions with different rate. The distribution assumption on the number of apparent mutations given inconsistent coalescent trees stays the same.

Given the coalescent tree in Figure 2.2, we replace the vector 'mut3' (n_{AB},n_{AC}, n_{BC}) by ($n_{AB_1}, n_{AB_2}, n_{AC_1}, n_{AC_2}, n_{BC_1}, n_{BC_2}$) where n_{AB_1} and n_{AB_2} represent number of apparent mutations share by A and B but not C in two disjoint bins defined by allele frequency below and above threshold f (with the same indexing scheme holding for n_{AC} and n_{BC}). Write $w_{11} = lg_2(\mu + \theta h_f/2) + l\epsilon$ for the expected number of apparent mutations with frequency less than f that are shared by haplotypes A and B, $w_{12} = lg_2\theta h'_f/2 + l\epsilon$ for the expected number of apparent mutations with frequency greater than f that are shared by haplotypes A and B, $w_{01} = l\epsilon$ for the expected number of apparent error variants with frequency less than f that are shared by haplotypes A and C (or by haplotypes B and C), and $w_{02} = l\epsilon$ for the expected number of apparent error variants with frequency greater than f that are shared by haplotypes A and C (or by haplotypes B and C). Let h_w represent the probability mass function of a Poisson distribution with mean w. Then

$$
\begin{aligned}
&P_{\mu,\theta,\epsilon}(mut3|tree3 = (AB - C, g_1, g_2), IBD3) \\
&= h_{w_{11}}(n_{AB_1}) h_{w_{12}}(n_{AB_2}) h_{w_{01}}(n_{AC_1}) h_{w_{02}}(n_{AC_2}) h_{w_{01}}(n_{BC_1}) h_{w_{02}}(n_{BC_2})
\end{aligned}
$$

.

The likelihood construction above is analogous to the regression adjustment which only considers gene conversion events occurring between the common ancestor of the two most-recently coalescing haplotypes and their common ancestor with the third haplotype. However, gene conversion events occurring on other branches will also influence the apparent mutation count and the error count.

All three haplotypes will carry the derived allele if the mutation event occurred before the coalescence of the three haplotypes, and the derived allele is carried by their common ancestor. In this case, a gene conversion event occurring between the common ancestor of the three haplotypes and the haplotype that is not in the first coalescing pair may introduce a variant that is shared by the two most-recently coalescing haplotypes but not by the third

haplotype. The rate of such events is proportional to the length of this branch ($g_1 + g_2$ generations). Overall, the expected number of such gene conversion events contributing to the apparent mutation count when using an allele frequency threshold of f is $l(g_1 + g_2)\theta h_f/2$. In this scenario, $h_f/2$ represents the probability that the individual undergoing gene conversion had a heterozygous genotype, with the major allele on the donor haplotype.

Similarly, a recent gene conversion event occurring between the most-recently coalescing haplotypes and their common ancestor may introduce a variant shared by two of the three haplotypes. In this case the two haplotypes sharing the variant include only one of the two most-recently coalescing pair, along with the third haplotype. Such events will contribute to the error count since the sharing of alleles is not consistent with the coalescent tree. The rate of such events is proportional to the length of the branch from one of the most-recently coalescing haplotypes to the common ancestor of the two most-recently coalescing haplotypes (g_1 generations). Overall, the expected number of such gene conversion events contributing to the particular error count (for a given pair of haplotypes that includes only one haplotype from the most-recently coalescing pair) when using an allele frequency threshold f is $lg_1\theta h_f/2$.

In consequence, the rates w_{11}, w_{12}, w_{01}, and w_{02} need to be modified to calculate $P_{\mu,\theta,\epsilon}(mut3|tree3 = (AB - C, g_1, g_2), IBD3)$ for the example of coalescent tree in Figure 2.2. Let $w^*_{11}, w^*_{12}, w^*_{01}$, and w^*_{02} denote the modified rates of the Poisson distributions that are used to model the counts of apparent mutations and errors. Due to gene conversion events occurring between the common ancestor of all three haplotypes and haplotype C, $w^*_{11} = w_{11} + l(g_1 + g_2)\theta h_f/2$ and $w^*_{12} = w_{12} + l(g_1 + g_2)\theta h'_f/2$. Due to gene conversion events occurring between the common ancestor of A and B and haplotype A (or haplotype B), $w^*_{01} = w_{01} + lg_1\theta h_f/2$ and $w^*_{02} = w_{01} + lg_1\theta h'_f/2$. If we re-arrange the parameters, we can see that the terms involving μ are unchanged, which is why ignoring the extra gene conversion events in the regression approach did not bias the estimation of μ. However, the terms

involving θ are changed by a factor of two. Indeed, in our analysis of simulated data, the regression-based estimates of θ were roughly double the true values. In addition, the terms involving ϵ are increased when considering the extra gene conversion events. Thus, modeling the extra gene conversion events could improve the precision in mutation rate estimation and correct the bias in estimation of gene conversion rate.

$$
\begin{array}{c|c}
w_{11} = lg_2(\mu + \theta h_f/2) + l\epsilon & w^*_{11} = lg_2(\mu + \theta h_f) + l(g_1 h_f/2 + \epsilon) \\
w_{12} = lg_2\theta h'_f/2 + l\epsilon & w^*_{12} = lg_2\theta h_f + l(g_1 h'_f/2 + \epsilon) \\
w_{01} = l\epsilon & w^*_{01} = l(g_1 h_f/2 + \epsilon) \\
w_{02} = l\epsilon & w^*_{02} = l(g_1 h'_f/2 + \epsilon)
\end{array}
$$

We compared the performance of the regression adjustment and the likelihood approach on a simulated dataset. We simulated a dataset with a gene-conversion initiation rate of 2.0×10^8 per base pair per meiosis and mean conversion tract length of 300bp, which is close to previously reported estimates of gene conversion rate using human data [56]. Other parameters used for the simulation were kept same as described in Section 4.1. We used hap-IBD [60] to detect pairwise IBD segments from genotypes with true phase using diallelic SNPs with minor allele frequency higher than 0.1. For IBD detection with hap-IBD, we set the minimum seed length to 0.5 cM, the minimum extension length to 0.1cM, and the maximum number of base pairs between seed segments and extensions to be 5000.

For the regression adjustment, we used frequency thresholds between 0.1 and 0.5 as described in Section 4.1 and we obtained a mutation rate estimate of 1.32×10^{-8} per base pair per meiosis (Figure 4.3) with a 95% confidence interval of $[1.28 \times 10^{-8}, 1.35 \times 10^{-8}]$ (the simulated mutation rate is 1.30×10^{-8}). The estimated gene conversion rate is as twice as high as the simulated gene conversion rate (the estimated gene conversion rate is 1.4×10^{-5} per meiosis with 95% confidence interval of $[1.36 \times 10^{-5}, 1.48 \times 10^{-5}]$ and the simulated gene

conversion rate is 6.0×10^{-6}). The bias in the estimates of gene conversion rate exists because the regression adjustment only adjusts for apparent mutations due to gene conversion events occurring between the common ancestor of the two most-recently coalescing haplotypes and their common ancestor with the third haplotype, while apparent mutations due to gene conversion events occurring on other branches are absorbed into the gene conversion term and into the error count. Thus, the estimates for the error term are increasing with the allele frequency thresholds (Figure 4.4).

For the likelihood calculation, we used a frequency threshold of 0.1, i.e. variants with minor allele frequency less than or equal to 0.1 contribute to the estimation of both mutation rate and gene conversion rate, and variants with minor allele frequency greater than 0.1 only contribute to the estimation of gene conversion rate. We obtained a mutation rate estimate of 1.31×10^{-8} per base pair per meiosis with a 95% confidence interval of $[1.28 \times 10^{-8}, 1.33 \times 10^{-8}]$. Compared to the regression adjustment, the likelihood adjustment has the potential to provide more precise estimates of mutation rate. The confidence interval we obtained from the simulated dataset through likelihood adjustment is narrower than that from the regression adjustment, however, replications of simulation studies need to be done to confirm whether the likelihood adjustment is more precise than regression adjustment on average. In our simulation, we found a downward bias in the estimates of gene conversion rate (the estimated gene conversion rate is 5.0×10^{-6} with 95% confidence interval of $[4.8 \times 10^{-6}, 5.0 \times 10^{-6}]$, while the simulated gene conversion rate is 6.0×10^{-6}). The IBD detection uses markers with minor allele frequency greater than 0.1 and may fail to report some IBD segments in which a gene conversion involving a high frequency variant occurred, which would lead to bias in the estimates of the gene conversion rate. At present, estimation of the mutation rate is of primary interest and the gene conversion rate is treated as a nuisance parameter. Future improvements in IBD detection algorithms may provide more accurate IBD calling and have the potential to correct the bias in gene conversion rate estimation.

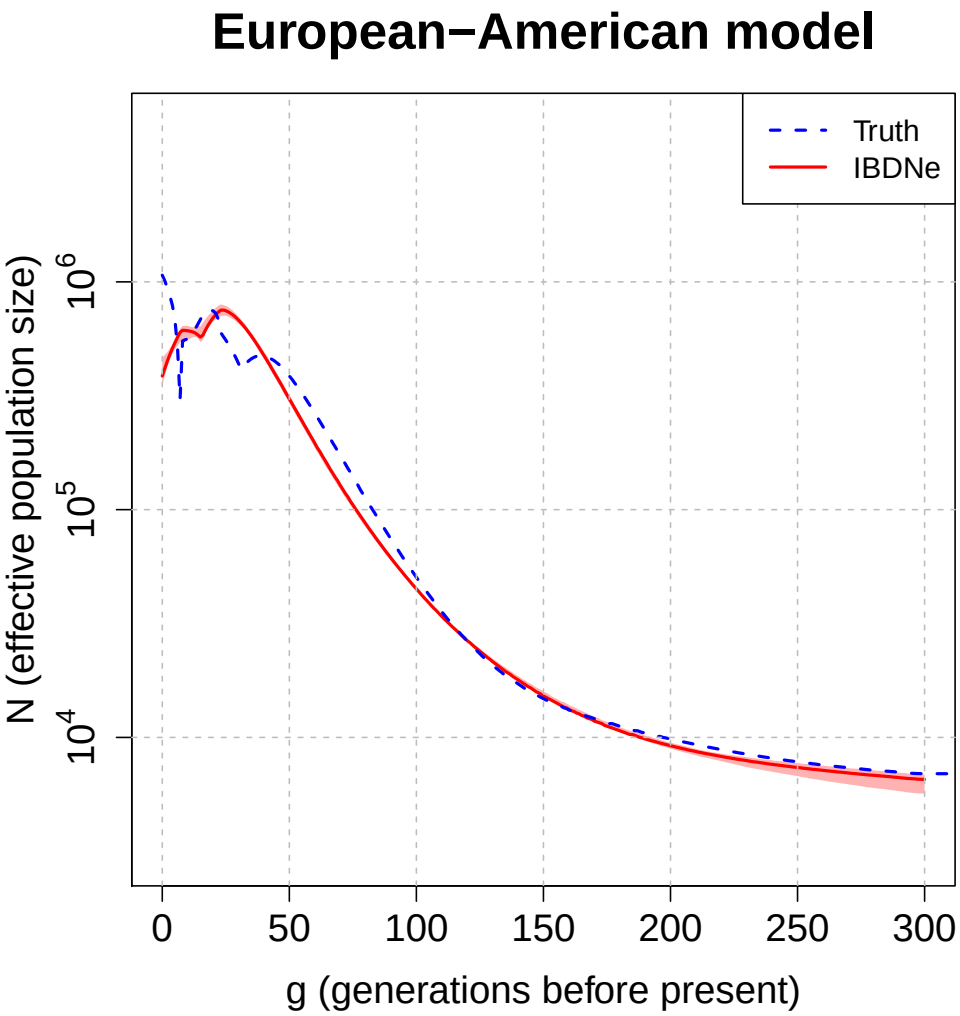


Figure 4.1: Demographic history of 'European-American' model. The simulated and estimated recent effective population size for the 'European-American' model. Generations before present are shown on the x-axis, and effective population size is shown on the y-axis. The blue dashed line gives the truth, the red line gives the estimated size while the red region gives the 95% bootstrap confidence intervals. The estimates of effective population size comes from applying IBDNe on IBD segments reported from Refined IBD running on the dataset simulated with gene conversion.

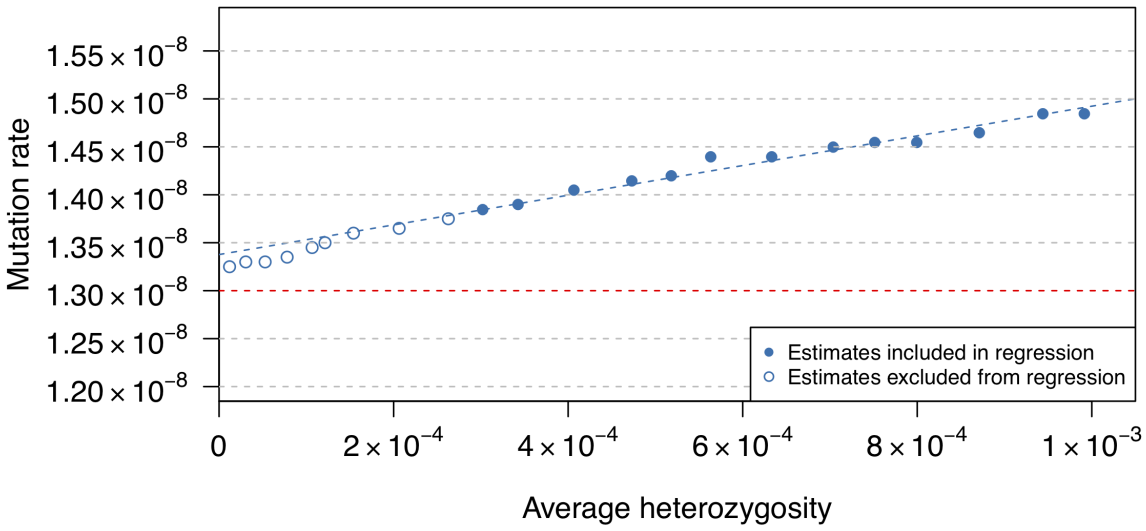


Figure 4.2: Estimated mutation rates in simulated data with gene conversion, as a function of the average heterozygosity of included variants. Data were simulated under the 'European-American' model with gene conversion. The blue dashed line is the fitted regression line; the y-axis intercept of this line gives an overall estimate of mutation rate that is adjusted for the effects of gene conversion. Points corresponding to maximum allele frequency thresholds of 0.1-0.5 are included in the regression (filled points on the plot), as lower thresholds may exclude some true mutations (open points on the plot). For each maximum allele frequency threshold, the average heterozygosity was calculated (x-axis), and the mutation rate estimate was obtained (y-axis). The simulated mutation rate is 1.30×10^{-8} and is shown with the horizontal red dashed line. The simulated gene conversion rate is 1.0×10^{-6}.

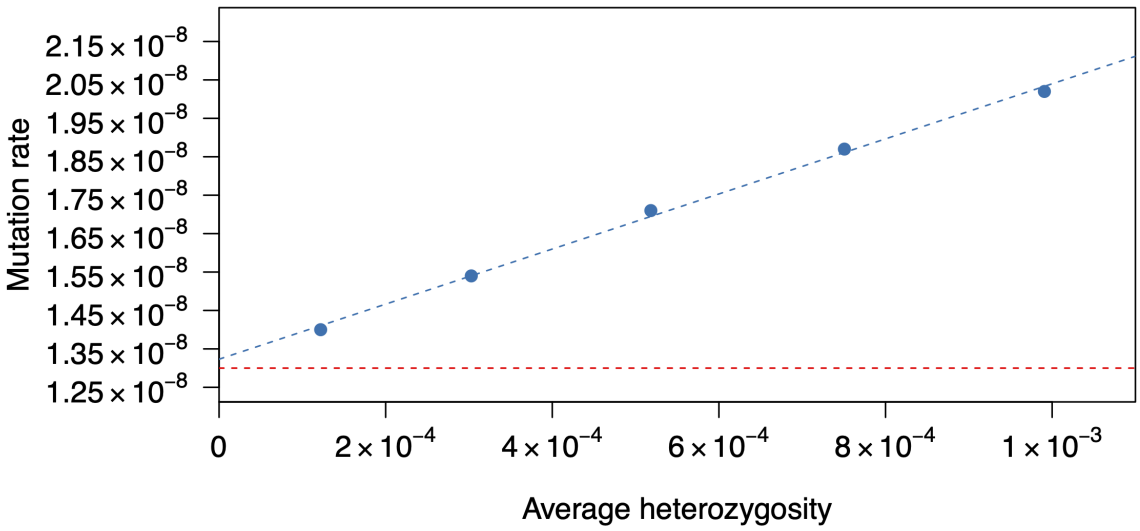


Figure 4.3: Estimated mutation rates in simulated data with gene conversion, as a function of the average heterozygosity of included variants. Data were simulated under the 'European-American' model with gene conversion. The blue dashed line is the fitted regression line; the y-axis intercept of this line gives an overall estimate of mutation rate that is adjusted for the effects of gene conversion. Points corresponding to maximum allele frequency thresholds of 0.1-0.5 are included in the regression. For each maximum allele frequency threshold, the average heterozygosity was calculated (x-axis), and the mutation rate estimate was obtained (y-axis). The simulated mutation rate is 1.30×10^{-8} and is shown with the horizontal red dashed line. The simulated gene conversion rate is 6.0×10^{-6}.

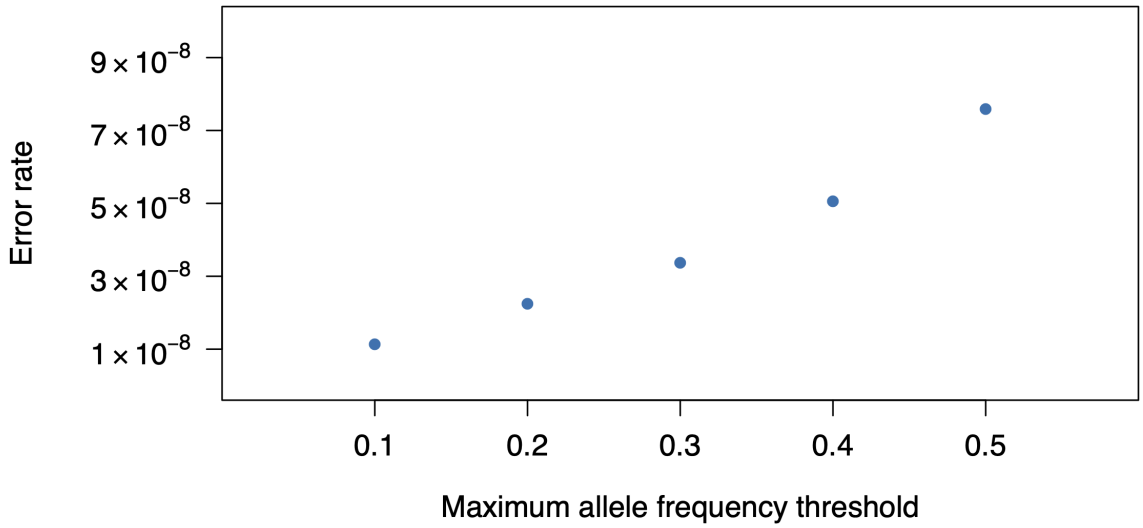


Figure 4.4: Estimated error rates in simulated data with gene conversion, as a function of maximum allele frequency thresholds for regression. Data were simulated under the 'European-American' model with gene conversion. Points corresponding to maximum allele frequency thresholds of 0.1-0.5 are included in the regression. For each maximum allele frequency threshold, the estimate of error rate was obtained (y-axis).

Chapter 5

ANALYSIS OF TOPMED DATA

5.1 Analysis of Framingham Heart Study with regression adjustment

We analyzed the Framingham Heart Study data from the NHLBI TOPMed Project which consists of high-coverage sequence data on 4166 samples with European ancestry (dbGap accession phs000974.v2.p2). We restricted all our analyses to diallelic SNPs passing quality control filters, and we used the Rutgers genetic map [35]. We identified 697 mother-father-offspring trios from a given pedigree and performed trio-based phasing using BEAGLE version 4.0 [4]. Trio-based phasing has high accuracy for both common and rare variants because Mendelian inheritance constraints determine the haplotype phase in the parents and offspring at most positions.

There were 1362 founder individuals from the 697 trios. By using trio parents rather than offspring for the analysis, we double the sample size, and the phasing is well-determined except at those small number of points of crossing-over in the meioses from trio parents to trio offspring. We ran principal component analysis (PCA) on these founders. In order to account for relatedness in the PCA, we removed 194 individuals who were closely related to others in the sample (total length of detected IBD segments exceeding 40cM on chromosome 22) when computing the principal components, and then determined PC scores of the excluded samples by projecting them onto the principal component axes. We found two distinct clusters on the second principal component (PC2$>$0.05 and PC2$\leq$0.05; Figure 5.1), and we inferred the demographic history of each cluster separately using IBDNe. We found that the cluster with PC2$>$0.05 experienced a recent severe population bottleneck around 30 generations ago (Figure 5.2), which is consistent the demographic history of the Ashkenazi

Jewish population [14]. Although our method is robust to population structure (Figure 3.6), we conservatively removed the 55 samples with PC2>0.05 from the mutation rate analysis. We also present results without the exclusion.

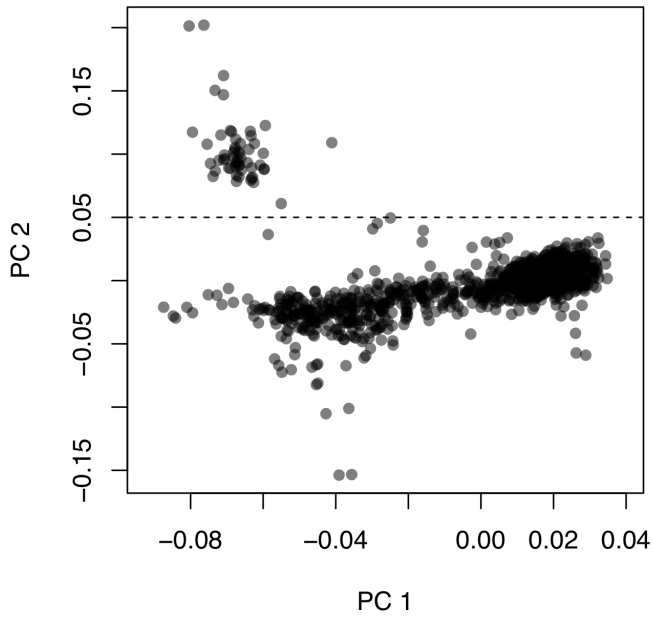


Figure 5.1: The first two principal components for genetic data from 1362 founders in Framingham Heart Study. The dashed line indicates PC2=0.05, which separates the two clusters of individuals.

Similar to the simulated data in Section 3.4, we used Refined IBD [5] in BEAGLE version 4.1 to detect pairwise IBD segments from the phased genotypes using only diallelic SNPs with minor allele frequency 10% or higher. After filling short gaps between segments, we removed segments with length less than 3 cM since the accuracy of IBD detection method tends to be higher for longer IBD segments [6]. We also removed segments with length or

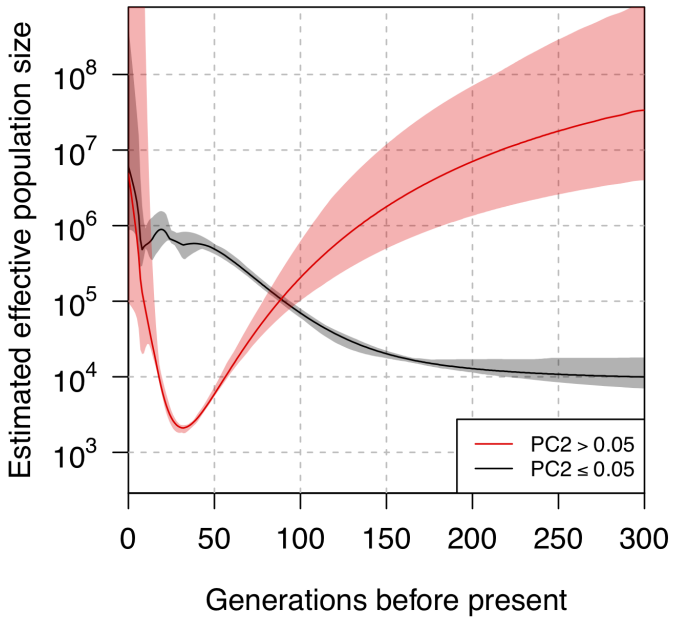


Figure 5.2: Estimated effective population size of two clusters of individuals from the Framingham Heart Study. Generations before present are shown on the x-axis, and estimated effective population size is shown on the y-axis. The solid line in black represents the estimated size from samples with PC2 less than or equal to 0.05, and the estimates are similar to the estimates based on all individuals from the Framingham Heart Study (Figure 5.3) . The solid line in red represents the estimated size from samples with PC2 greater than 0.05. The shaded region gives the 95% bootstrap confidence intervals.

larger than 6 cM as our simulation study showed that removal of segments larger than 6 cM reduces the potential impact of false negative errors.

We used the whole dataset of 4166 samples for the effective population size estimation on the Framingham Heart Study data, which is required to calculate the likelihood of the

mutation rate. We identified 13211 pairs of closely related individuals as those with total length of detected IBD segments exceeding 5% of the genome length, and we removed IBD segments corresponding to such pairs from the IBDNe analysis. The estimated recent effective population size for the Framingham Heart Study data is shown in Figure 5.3.

We excluded regions with extremely high levels of IBD from the analyses. We find that these regions are mainly in areas of the genome with low marker density, such as around centromeres. In such regions, the IBD detection method tends to overestimate the IBD segment lengths, because identity by state extending beyond the end of the IBD segment may cover a large genetic distance. To perform this exclusion, we calculated the level of three-way IBD along the genome in windows of 500 base pairs, and removed any three-way IBD that overlapped windows for which the three-way IBD level exceeded the 99th percentile. Regions that were outside the limits of the genetic map were also removed from the analysis. In addition, some regions of the genome did not contribute to the estimation because they contained no three-way IBD. After removing all these regions, the remaining data covered 2.01 gigabases across the autosomes. Many of the excluded regions are near the centromeres and telomeres due to poor data quality or low variant density in those areas (Figure 5.4).

We ran analyses with a range of maximum allele frequencies (Figure 5.5). The estimates increase as the maximum allele frequency increases, as expected with gene conversion. We thus applied regression on heterozygosity to correct for gene conversion. Our corrected estimate is 1.29×10^{-8} with 95% confidence interval $[1.02 \times 10^{-8}, 1.56 \times 10^{-8}]$ (Figure 5.6). For comparison, when we don't exclude the individuals who were outliers (PC2>0.05) on the principal components analysis, the estimate is 1.21×10^{-8} with 95% CI $[1.00 \times 10^{-8}, 1.45 \times 10^{-8}]$. When we apply the exclusions but change the lower bound on the maximum allele frequency, the estimates are 1.24×10^{-8} $[1.02 \times 10^{-8}, 1.46 \times 10^{-8}]$ for a lower bound of 0.05, and 1.36×10^{-8} $[0.98 \times 10^{-8}, 1.73 \times 10^{-8}]$ for a lower bound of 0.2.

We also applied IBDMUT [42] to these data. We used the same estimated demographic history as for our method (Figure 5.3), and we used IBD segments estimated by GERMLINE [22]. Regions with extremely high levels of pairwise IBD sharing were excluded from the analysis. We removed any segments that overlapped regions in which the pairwise IBD level exceeded the 99th percentile. We further removed segments with length less than 3 cM or larger than 6 cM from the analysis. The mutation rate estimate was 1.31×10^{-8} with 95% confidence interval $[1.20 \times 10^{-8}, 1.42 \times 10^{-8}]$. In this case, IBDMUT's estimate is consistent with our estimate, although our simulations show that this will not always be the case. IBDMUT has a narrower confidence interval because IBDMUT makes use of more meioses. Our method only uses the meioses between the coalescence of the first two haplotypes and their coalescence with the third haplotype in each set of three IBD haplotypes, while IBDMUT uses all the meioses for each pair of IBD haplotypes, including more recent meioses.

5.2 Accounting for haplotype phasing uncertainty

With the pedigree information, we are able analyze the trio parents from the Framingham Heart Study in which rare variants can be phased accurately. Without extending the mutation counting process to account for phasing uncertainty, the application of our method is limited since only small numbers of parent-offspring trios with whole genome sequence data are available. In this section, we describe a modified mutation counting procedure that accounts for the phasing uncertainty and is applicable to analysis of larger data sets of unrelated individuals to increase the precision of the estimates.

Statistical phasing algorithms on unrelated individuals are designed to find the most likely haplotype configuration given genotypes and rely on haplotype frequency information. A variant must be seen several times within its haplotype context to obtain high-confidence phase information since accurate variant frequency and linkage disequilibrium estimates are needed to obtain accurate phase through statistical phasing [9]. In contrast, in

parent-offspring trio based phasing, phase only needs to be determined at a small number of positions where all three individuals are heterozygotes due to Mendelian constraints. Incorporating these constraints into models such as Hidden Markov Models [4] can significantly improve the phasing accuracy, thus trio based phasing has lower error rate than population based phasing algorithms. In the absence of parent-offspring data, it is difficult to phase rare variants computationally. Therefore, it is hard to tell whether the IBD haplotype of interest carries the rare variant, and the phasing uncertainty for rare variants makes it difficult for mutation ascertainment and could bias the mutation count.

For the common variants that can be phased accurately, we obtain apparent mutations from the three haplotypes of interest as shown in Figure 5.7. For the rare variants that are difficult to phase, we use the information on the unphased genotypes of the three individuals instead of the three IBD haplotypes of interest to identify apparent mutations. Let A1, B1, C1 denote the three IBD haplotypes of interest, and let A2, B2, C2 denote the other haplotype of individuals A, B, C, respectively. These haplotypes are estimated based on common variants which can be phased accurately. In the example shown in Figure 5.8, we aim to find candidates for apparent mutations shared between haplotypes A1 and B1 but not C1. Since we focus on rare variants, the possibility that the apparent mutation is also carried by C2 is negligible, thus we do not consider positions where individual C carries the rare variant. If individual A and B both carry two copies of the variant, it is certain that the variant will be shared between the IBD haplotypes (A1 and B1). If individual A carries two copies of the variant and individual B only carries one copy, the probability that the variant is carried by B1 instead of B2 is high given the prior that A1 carries the rare variant and it is IBD with B1; we thus include such a candidate as an apparent mutation. Similar reasoning holds when individual B carries two copies of the variant and individual A only carries one copy of the variant. For a scenario where individuals A and B each carry one copy of the variant, further investigation needs to be done to confirm whether the variant is shared between the two IBD haplotypes of interest.

When the two individuals each carry one copy of the rare variant, we use other individuals in the dataset to confirm whether the variant is shared between the IBD haplotypes of interest. In the example shown in Figure 5.9, the first two dashed rectangles represent candidates for apparent mutations subject to further investigation. For each such configuration, we use phasing of the common variants to identify an extra individual D and one haplotype of this individual (D1) that is IBD with A2 (or B2) but not IBD with A1 (or B1) at the position of interest. In addition, the other haplotype of this individual (D2) should not be IBD with A1 (or B1). If individual D carries the rare variant of interest, we assume that the variant is not shared between the IBD haplotypes of interest (A1 and B1) and exclude such a candidate from the analysis. If there are multiple D individuals with haplotypes that are IBD with A2 (or B2) but not with A1 (or B1), we exclude such a candidate if any of the D individuals carries the rare variant. To avoid the uncertainties at the ends of inferred IBD segments, we trimmed 0.5 cM from each end of the IBD segments before identifying the haplotype D1. If we can't find any D individuals, we keep the candidate in the analysis, which may create a false positive error in mutation counts. However, if the candidate is a false positive error and not shared by A1 and B1, the number of such false positive errors does not depend on g_2 (Figure 2.2) and thus will be accounted by the error term (as described Section 4.2). In Figure 5.9, the candidate in the red dashed rectangle is excluded from the analysis since there is another individual, D, with a haplotype that is IBD with A2 and this individual D also carries the rare variant. The second candidate (second dashed rectangle) is subject to investigation and is included as an apparent mutation because there is no evidence that the variant is carried by A2 or B2. With the adjustment on the mutation counting procedure and the use of the error term that models false positives in mutation counts, we are able to extend our method to account for phasing uncertainty of low frequency variants.

To evaluate the performance of the modified mutation counting scheme in the presence of genotype error, we added 0.01% 'unbiased' genotype error (described in Section 3.3) to the simulated data (described in Section 4.2). Figure 5.10 demonstrates our data processing

pipeline. We first phase markers with minor allele frequency greater than or equal to 1% using Beagle version 5.1 [8]. Markers with minor allele frequency greater than 10% are used for IBD detection with hap-IBD [60]. On the inferred IBD haplotypes, we perform our original counting procedure on the statistically phased markers (markers with frequency greater than 1%). Markers with minor allele frequency greater than 10% will contribute to gene conversion estimation, and markers with minor allele frequency between 1% and 10% will contribute to both mutation and gene conversion estimation. Finally, we perform our modified counting procedure on the unphased low frequency markers and they contribute to both mutation and gene conversion estimation. When applying the modified counting procedure, we trim 0.5cM from each end of the region on the extra individuals (i.e. D) in case the observed identity by state extends beyond the true IBD region [42, 5] in IBD calling. In the simulated data with mutation rate 1.3×10^{-8}, we obtained a mutation rate estimate of 1.29×10^{-8} per base pair per meiosis with a 95% confidence interval of $[1.27 \times 10^{-8}, 1.32 \times 10^{-8}]$. In contrast, when ignoring the phasing uncertainty, we obtained a mutation rate estimate of 1.21×10^{-8} per base pair per meiosis with a 95% confidence interval of $[1.18 \times 10^{-8}, 1.24 \times 10^{-8}]$.

5.3 Analysis of TOPMed data

5.3.1 Framingham Heart Study

We analyzed the full Framingham Heart Study data (Section 5.1) with the modified approach for mutation ascertainment. We restricted all our analyses to diallelic SNPs passing quality control filters, and we performed statistical phasing using BEAGLE version 5.1 [8] on SNPs with minor allele frequency 1% or higher. We used hap-IBD [60] to detect pairwise IBD segments from the phased haplotypes using SNPs with minor allele frequency greater than 10%. For IBD detection with hap-IBD, we set the minimum seed length to 0.5 cM, the minimum extension length to 0.1cM, the maximum number of base pairs between seed segments and extensions to be 5000, the minimum length threshold for reporting IBD segments to 1cM.

After removing duplicated samples, one individual from each pair of monozygotic twins, and offspring with at least one parent in the dataset (identified from the supplied pedigree), there were 2506 individuals left in the dataset. We then performed mutation rate analysis on these 2506 individuals.

When constructing three-way IBD sharing, we restricted the length of IBD segments to 3 cM to 6 cM in order to reduce the impact of genotype calling error and IBD detection error (Section 5.1). Since IBD detection methods tend to overestimate the IBD segment lengths in regions with low marker density (Section 5.1) we therefore excluded regions with extremely high levels of IBD from the analyses. We first calculated the number of three-way IBD along the genome in windows of 500 base pairs (i.e. three-way IBD coverage), and removed any three-way IBD that overlapped windows for which the three-way IBD coverage exceeded the 98th percentile of the genome-wide three-way IBD coverage (Appendix). Regions that were outside the limits of the genetic map were also removed from the analysis. In addition, some regions of the genome did not contribute to the estimation because they contained no three-way IBD. After removing all these regions, the remaining data covered 2.67 gigabases across the autosomes.

We applied the modified mutation ascertainment (Section 5.2) for variants with minor allele frequency below 1%. Variants with minor allele frequency 1% or higher were statistically phased and the original method for mutation ascertainment was used. Apparent mutation counts from variants with minor allele frequency 10% or lower contribute to the estimation of mutation rate and gene conversion rate, while apparent mutation counts from variants with minor allele frequency above 10% only contribute to the estimation of gene conversion rate.

Our estimated mutation rate is 1.58×10^{-8} per base pair per generation with 95% confi-

dence interval $[1.44 \times 10^{-8}, 1.74 \times 10^{-8}]$ and the estimated gene conversion rate is 6.0×10^{-6} per base pair per generation with 95% confidence interval $[5.2 \times 10^{-6}, 6.8 \times 10^{-6}]$. We also performed a subgroup analysis based on PCA scores. We performed principal component analysis on 4166 samples. We first removed 2498 samples who were closely related to other samples in the dataset (total length of detected IBD segments exceeding 25% of the length of whole genome) when computing the principal components, and then determined PC scores of the excluded samples by projecting them onto the principal component axes. The clustering of the PC scores (Figure 5.11) is similar to what we found in Figure 5.1. After removing 105 samples in the cluster that experienced a recent severe population bottleneck, we have 2401 samples left. From the 2401 samples, we obtained a mutation rate estimate of 1.62×10^{-8} per base pair per generation with 95% confidence interval $[1.46 \times 10^{-8}, 1.80 \times 10^{-8}]$ and the estimate for gene conversion rate remains unchanged.

5.3.2 Barbados Asthma Genetic Study and Jackson Heart Study

We also analyzed data from the Barbados Asthma Genetic Study (BAGS) and Jackson Heart Study (JHS) from the TOPMed Project, in which most of the individuals are African descendent. The BAGS data consists of 962 samples and the JHS data consists of 2777 samples. We used the HapMap [18] genetic map instead of Rutgers genetic map [35] since the Rutgers genetic map is based on data from European populations. After removing duplicated samples, one individual from each pair of monozygotic twins, and offspring with two parents in the dataset (identified from the supplied pedigree), there were 738 and 2467 distantly related individuals left in the BAGS and JHS dataset, respectively. Other data processing pipelines are in concordance with the pipeline for analyzing data from the Framingham Heart Study.

We used the default settings of IBDNe[11] to infer the demographic history with IBD segments of length 2 cM and above. We found a recent reduction in effective population size approximately 10 generations ago (Figure 5.12, Figure5.13), which likely reflects the bottleneck effect of migration from Africa.

We removed some regions from the analysis based on the level of three-way IBD along the genome as described in Section 5.3.1. After removing all these regions, the remaining BAGS and JHS data covered 1.8 and 2.7 gigabases across the autosomes. The coverage in BAGS dataset is low due to limited sample size, thus a smaller number of regions were covered with three-way IBD. Regions included in the analysis can be found in Appendix.

Similar to the analysis of Framingham Heart Study data, we performed principal component analysis to capture subgroups with significant different population structure. For Barbados data (Figure 5.14), we first removed 561 (out of 962) samples who were closely related to other samples in the dataset (total length of detected IBD segments exceeding 25% of the length of whole genome) when computing the principal components, and then determined PC scores of the excluded samples by projecting them onto the principal component axes. We identified 33 individuals with PC scores being outliers in the principal components analysis (PC1<-0.15 or PC2<-0.15). These samples may have different ancestry compositions or migration history comparing to other samples in the dataset. Since the number of samples with PC scores being outliers is small, we didn't exclude them from the analysis.

For Jackson Heart Study, we first removed 754 (out of 2777) samples who were closely related to other samples in the dataset (total length of detected IBD segments exceeding 25% of the length of whole genome) when computing the principal components; we didn't identify any outliers in the principal component analysis.

Our estimated mutation rate from the Barbados dataset is 1.52×10^{-8} per base pair per generation with 95% confidence interval $[1.30 \times 10^{-8}, 1.86 \times 10^{-8}]$ and the estimated gene conversion rate is 4.0×10^{-6} per base pair per generation with 95% confidence interval $[3.0 \times 10^{-6}, 5.6 \times 10^{-6}]$. From the Jackson Heart Study, our estimated mutation rate is 1.48×10^{-8} per base pair per generation with 95% confidence interval $[1.40 \times 10^{-8}, 1.60 \times 10^{-8}]$

and the estimated gene conversion rate is 3.2×10^{-6} per base pair per generation with 95% confidence interval $[2.8 \times 10^{-6}, 4.4 \times 10^{-6}]$. The confidence interval for the estimates from Barbados Asthma Genetic Study is wide due to the small sample size. We therefore combined Barbados and Jackson Heart Study data by multiplying the likelihood together for each combination of values in the search grid for mutation rate, gene conversion rate, and the nuisance error parameter. The estimated mutation rate from the combined dataset is 1.52×10^{-8} per base pair per generation with 95% confidence interval $[1.42 \times 10^{-8}, 1.60 \times 10^{-8}]$ and the estimated gene conversion rate is 4.0×10^{-6} per base pair per generation with 95% confidence interval $[2.8 \times 10^{-6}, 4.4 \times 10^{-6}]$ (Figure 5.15,5.16).

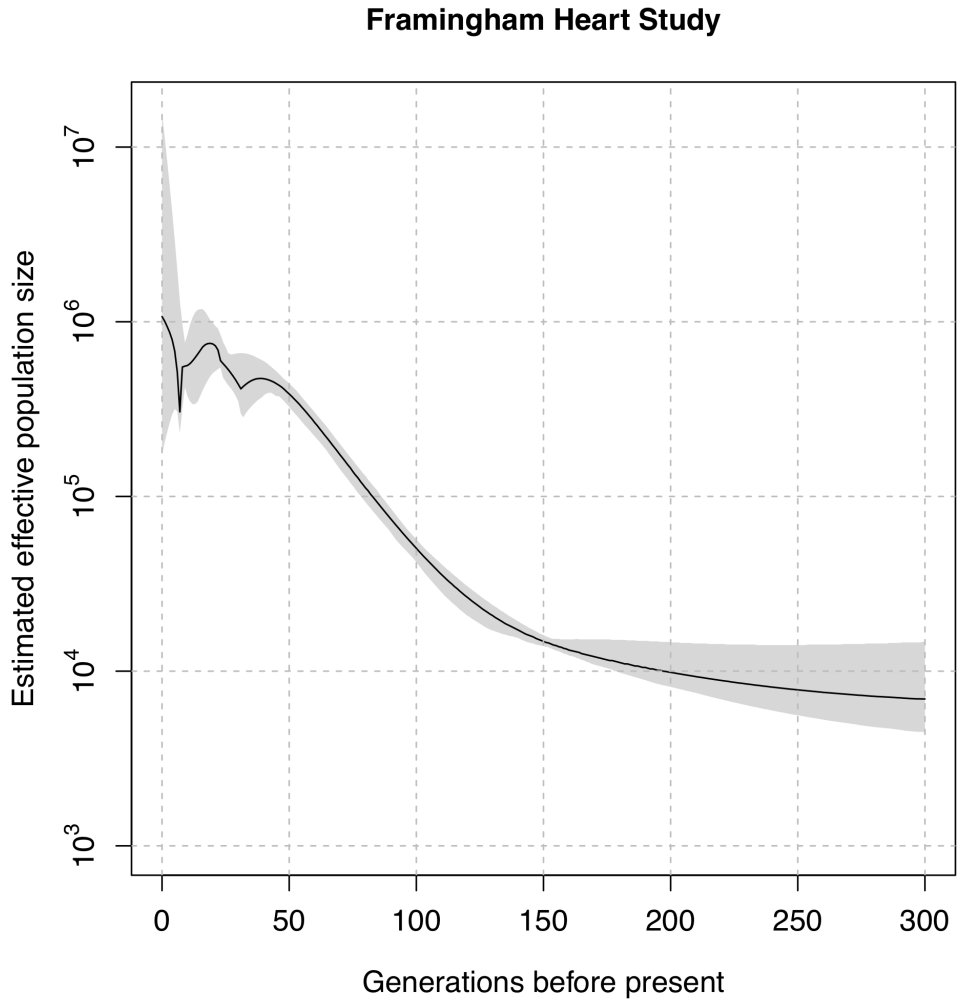


Figure 5.3: Estimated recent effective population size for the Framingham sample. Generations before present are shown on the x-axis, and estimated effective population size is shown on the y-axis. The black line gives the estimated size while the gray region gives the 95% bootstrap confidence intervals. The reduction in estimated effective size approximately 10 generations ago likely reflects the bottleneck effect of European migration to the US. A similar estimated bottleneck was seen in analysis of European-ancestry individuals sampled from Memphis, Tennessee. [7].

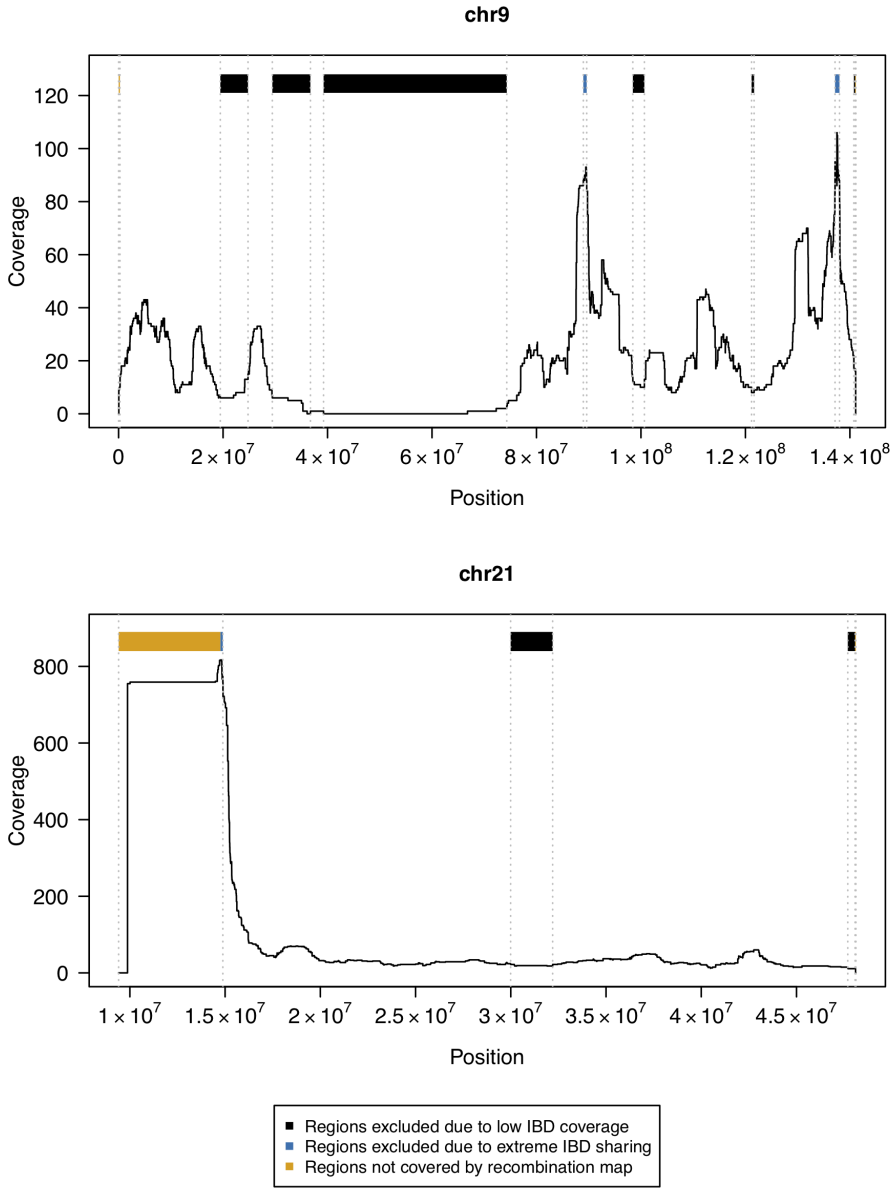


Figure 5.4: Examples of 3-way IBD coverage along the genome in trio-phased Framingham samples. Levels of three-way IBD are shown in windows of 500 base pairs for two representative chromosomes. Black bars represent regions with zero 3-way IBD coverage after removing IBD segments of length greater than 6 cM. Blue bars represent regions excluded from the analysis due to extremely high levels of apparent IBD sharing. Orange bars represent regions not covered by the Rutgers recombination map.

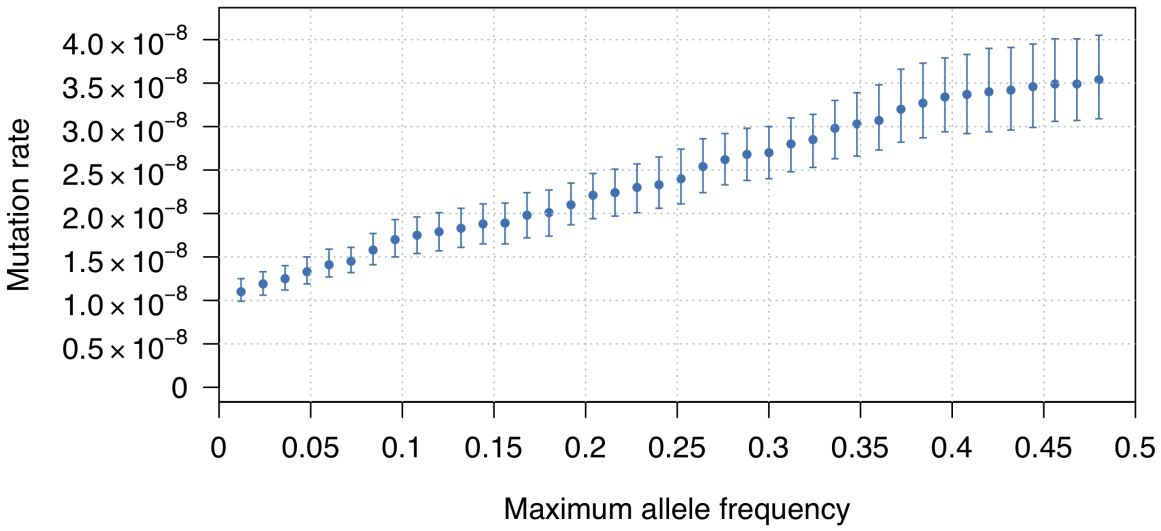


Figure 5.5: Estimated mutation rate from the trio-phased Framingham Heart Study data as a function of maximum allowed allele frequency. Points estimates (dots) and 95% confidence intervals (bars) corresponding to maximum allele frequency thresholds of 0.1-0.5 are included in the regression (filled points on the plot), as lower thresholds may exclude some true mutations (open points on the plot). For each maximum allele frequency threshold, the mutation rate estimate was obtained (y-axis).

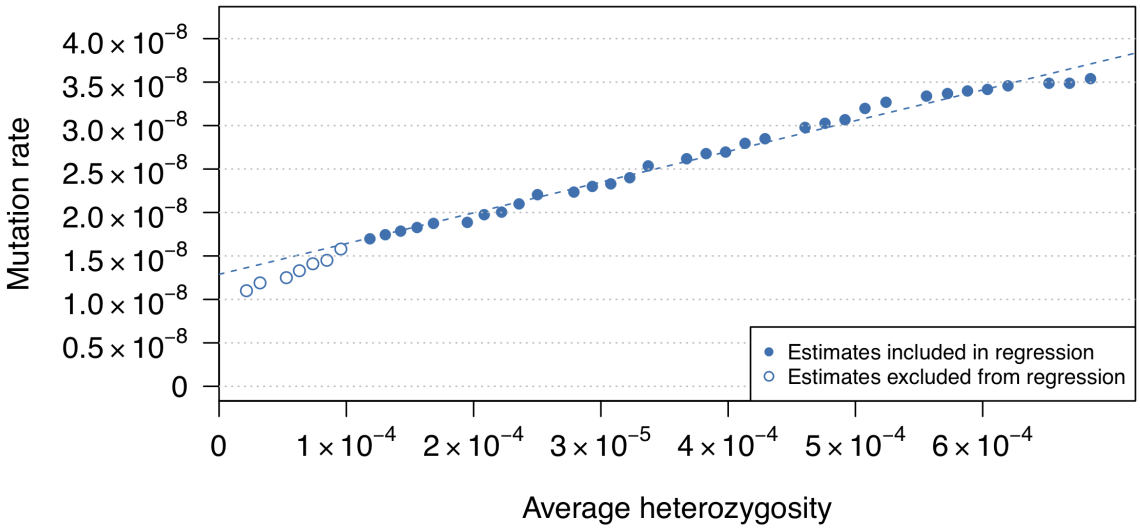


Figure 5.6: Estimated mutation rates from the trio-phased Framingham Heart Study data as a function of the average heterozygosity of included variants. The dashed line represents the fitted regression line; the y-axis intercept of this line gives an overall estimate of mutation rate that is adjusted for the effects of gene conversion. Points corresponding to maximum allele frequency thresholds of 0.1-0.5 are included in the regression (filled points on the plot), as lower thresholds may exclude some true mutations (open points on the plot). For each maximum allele frequency threshold, the average heterozygosity was calculated (x-axis), and the mutation rate estimate was obtained (y-axis).

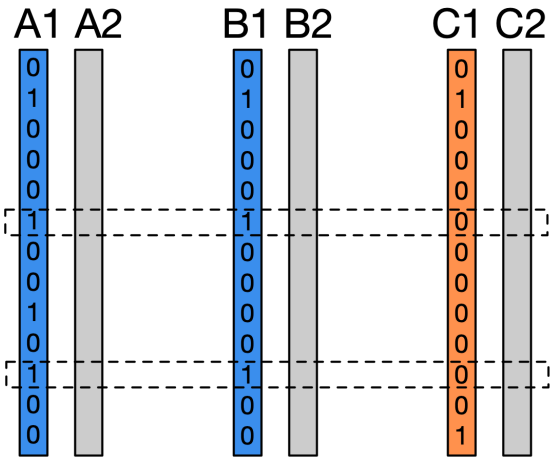


Figure 5.7: Counting process of apparent mutations on accurately phased data. Three IBD haplotypes, A1, B1, and C1, are labeled in blue and orange. In this example, mutations shared by A1 and B1 but C1 are of interest. A2, B2, C2 represent the other haplotypes of individuals A, B, and C respectively. Minor alleles are represented in ones and major allele are represented in zeros. For accurately phased variants, we only use the inferred phase on the IBD haplotypes of interest (A1, B1, and C1) to count apparent mutations. In this example, positions in dashed rectangles demonstrate the apparent mutations share by A1 and B1 but not C1.

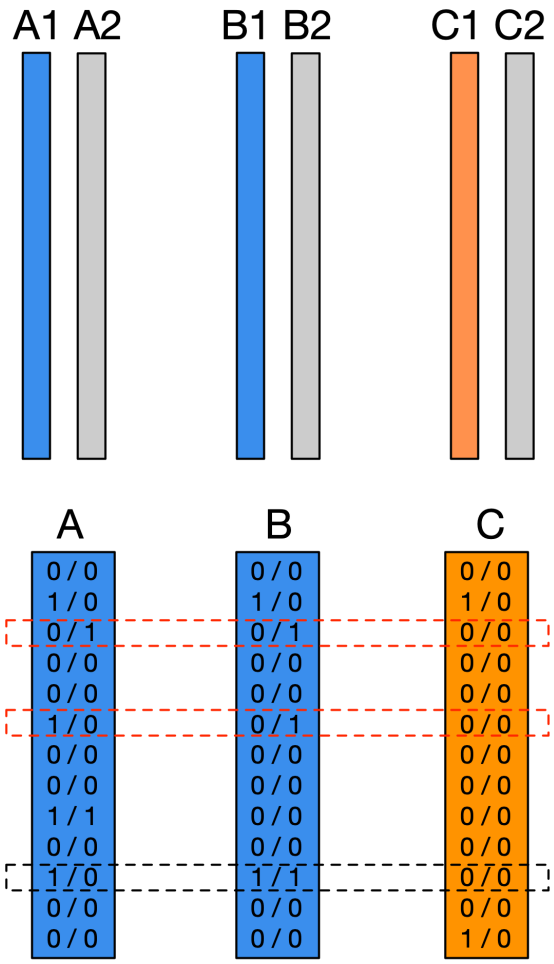


Figure 5.8: Counting process of apparent mutations on unphased data. The figure on the top shows haplotypes. A1, B1, and C1, are labeled in blue and orange. A2, B2, C2 represent the other haplotypes of individuals A, B, and C respectively. In this example, mutations shared by A1 and B1 but C1 are of interest. The IBD haplotypes are estimated from phased common variants. The figure on the bottom shows the unphased genotypes of rare variants for individuals A, B, C. Minor alleles are represented in ones and major allele are represented in zeros. Positions in black dashed rectangles demonstrate the apparent mutations we automatically include in the analysis. Positions in red dashed rectangles demonstrate the candidates for apparent mutations that require further investigation.

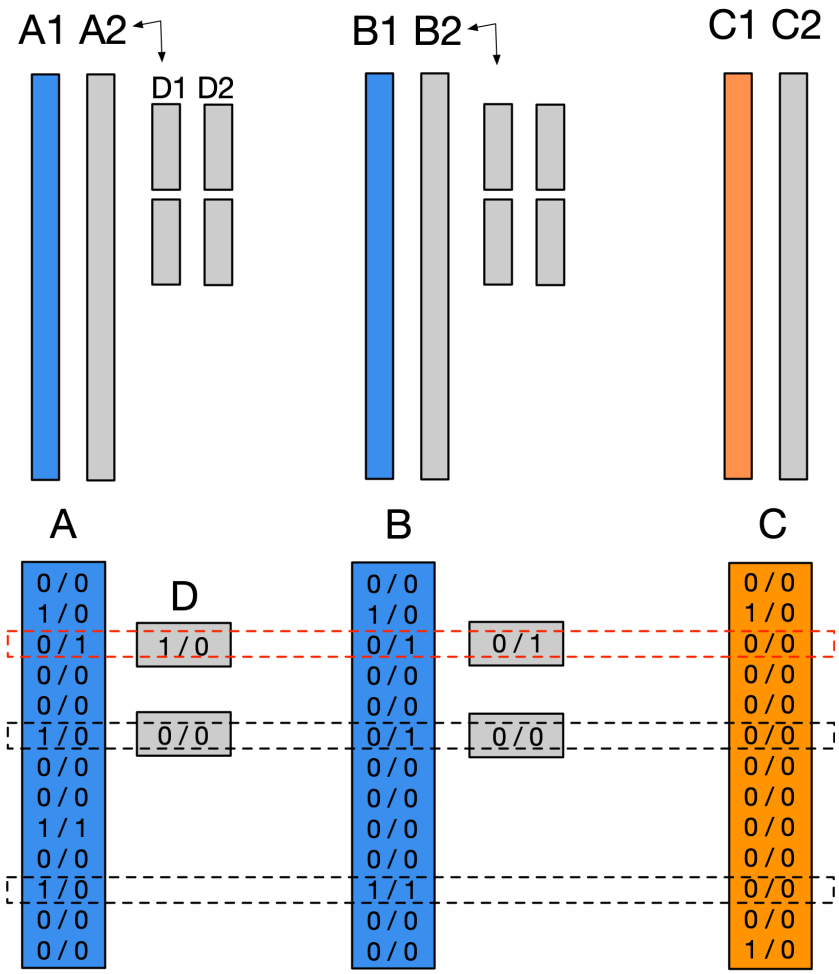


Figure 5.9: Modified counting process of apparent mutations on unphased data. The figure on the top shows haplotypes. A1, B1, and C1, are labeled in blue and orange. A2, B2, C2 represent the other haplotypes of individuals A, B, and C respectively. In this example, mutations shared by A1 and B1 but C1 are of interest. The IBD haplotypes are estimated from phased common variants. The figure on the bottom shows the unphased genotypes of rare variants for individuals A, B, C. Minor alleles are represented in ones and major allele are represented in zeros. D represents an individual with a haplotype D1 that is IBD with A2 or B2 at positions of interest and it is pictured as a short gray segment. Positions in black dashed rectangles demonstrate the apparent mutations we included in the analysis. Positions in red dashed rectangles demonstrate the candidates for apparent mutations that are excluded from analysis due to the carrier status of individual D.

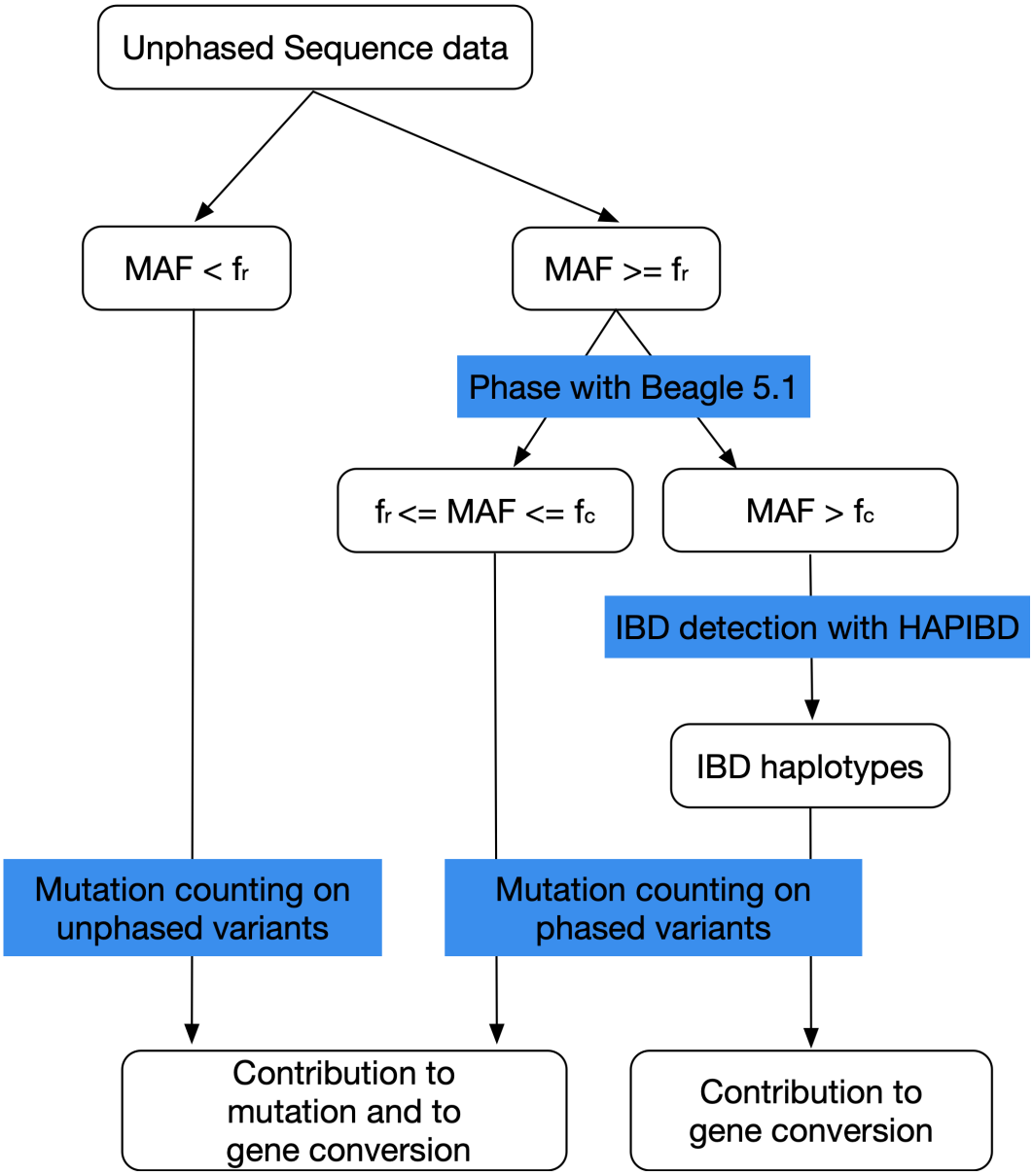


Figure 5.10: Data processing pipeline for unphased sequence data. Variants with minor allele frequency below f_r are kept unphased, other variants are phased with Beagle 5.1. Variants with minor allele frequency above f_c ($f_c \geq f_r$) are used for IBD detection. We use our modified mutation counting procedure on the unphased variants and our original counting procedure on variants that were phased statistically. For simulated and real data analysis, we use $f_r = 0.01$ and $f_c = 0.1$.

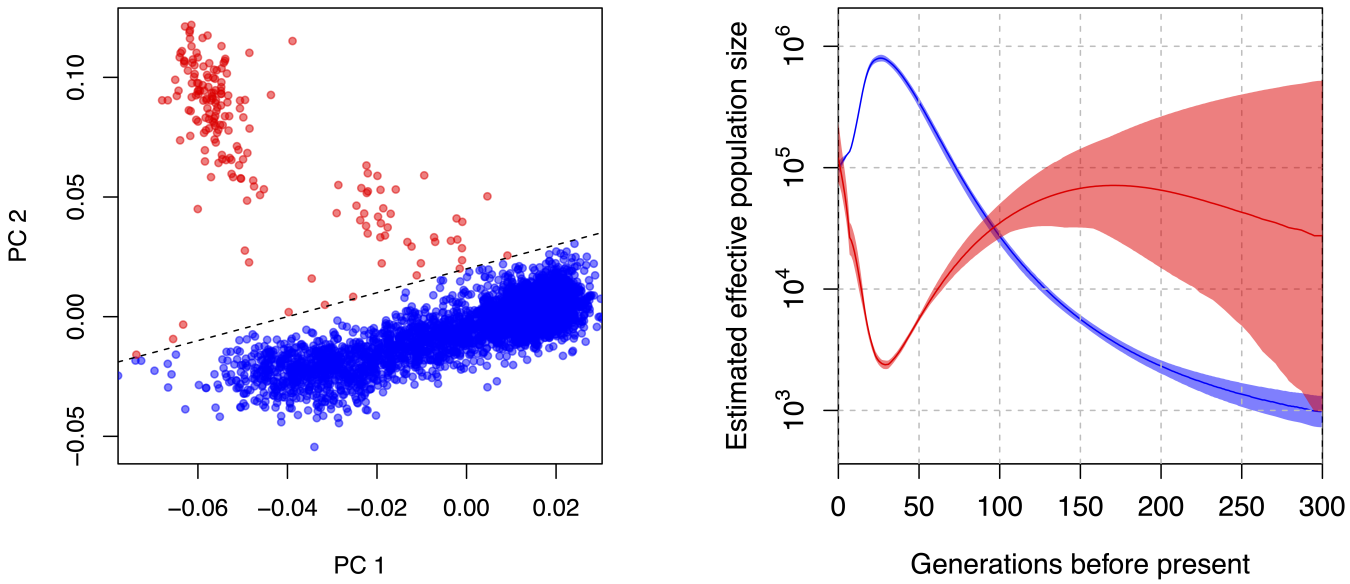


Figure 5.11: The first two principal components for genetic data and estimated effective population size from 4166 samples in Framingham Heart Study. The first two principal components for genetic data from 4166 samples in Framingham Heart Study is shown in the figure on the left. Two clusters of samples are formed based on values of first two principal components and are colored in red and blue. The estimated effective population size of two clusters is shown in the figure on the right. Generations before present are shown on the x-axis, and estimated effective population size is shown on the y-axis. The solid line in red represents the estimated size from samples in the red cluster on the left. The solid line in blue represents the estimated size from samples in the blue cluster on the left. The shaded region gives the 95% bootstrap confidence intervals.

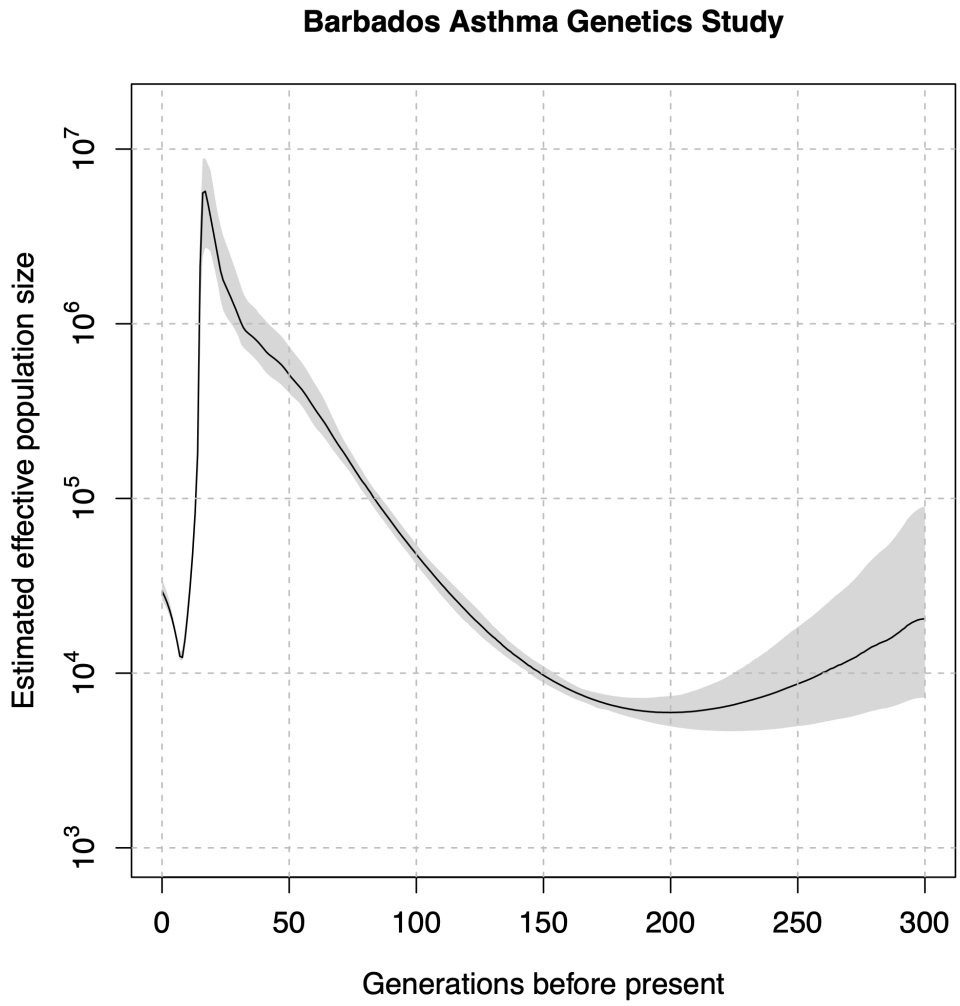


Figure 5.12: Estimated recent effective population size for samples from Barbados Asthma Genetic Study. Generations before present are shown on the x-axis, and estimated effective population size is shown on the y-axis. The black line gives the estimated size while the gray region gives the 95% bootstrap confidence intervals. The reduction in estimated effective size approximately 10 generations ago likely reflects the bottleneck effect of African migration to Barbados.

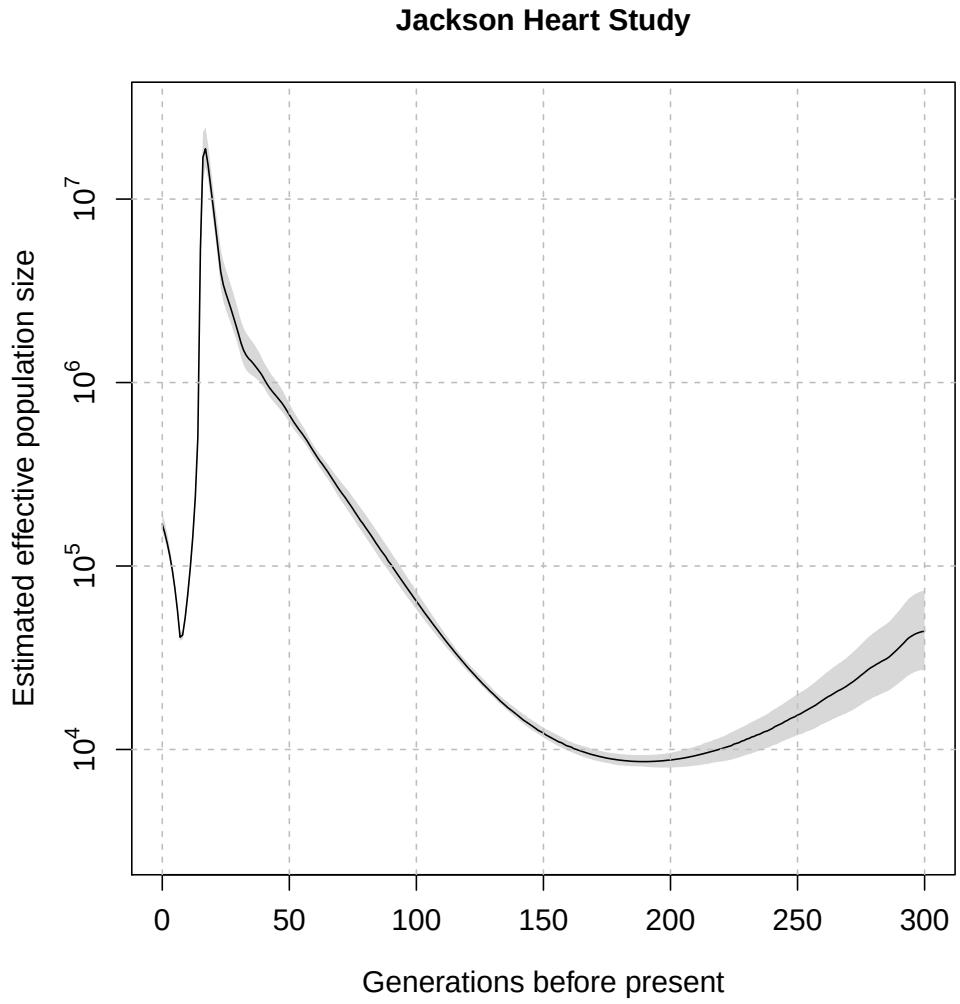


Figure 5.13: Estimated recent effective population size for samples from Jackson Heart Study. Generations before present are shown on the x-axis, and estimated effective population size is shown on the y-axis. The black line gives the estimated size while the gray region gives the 95% bootstrap confidence intervals. The reduction in estimated effective size approximately 10 generations ago likely reflects the bottleneck effect of African migration to the United States.

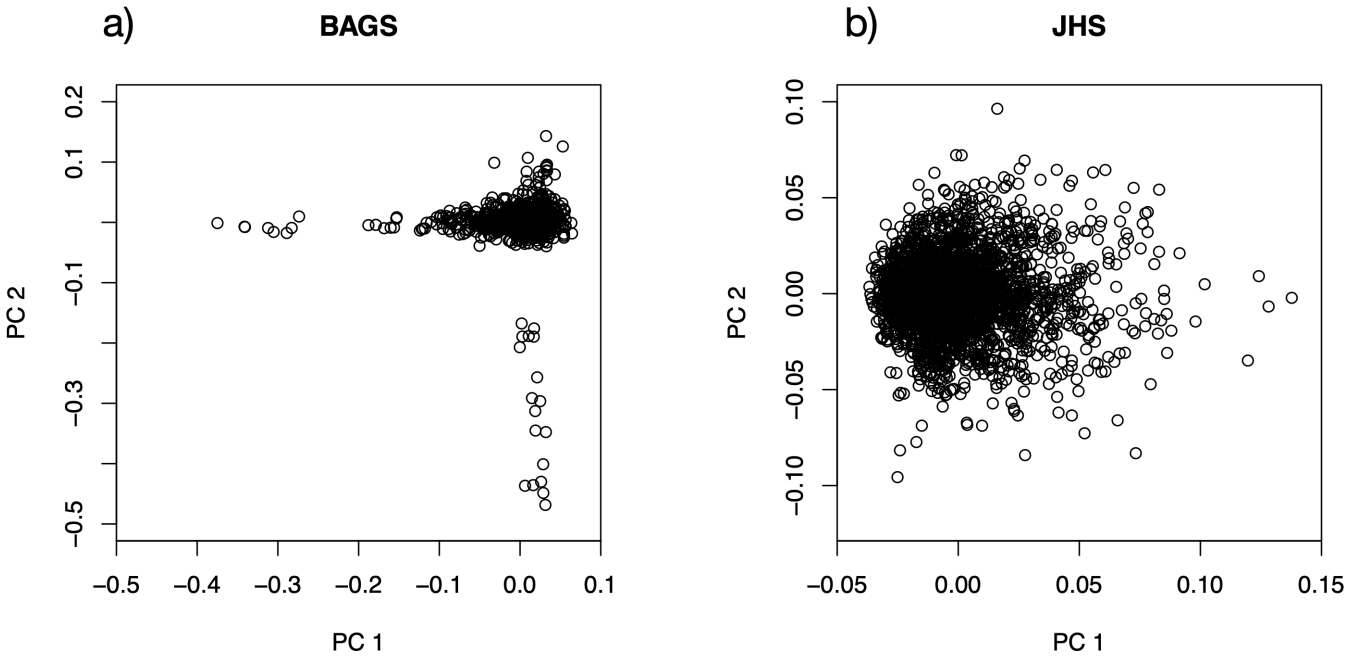


Figure 5.14: The first two principal components for genetic data in the Barbados Asthma Genetic Study and the Jackson Heart Study. a) The first two principal components for genetic data from 962 samples in the Barbados Asthma Genetic Study is shown. b) The first two principal components for genetic data from 2777 samples in the Jackson Heart Study are shown.

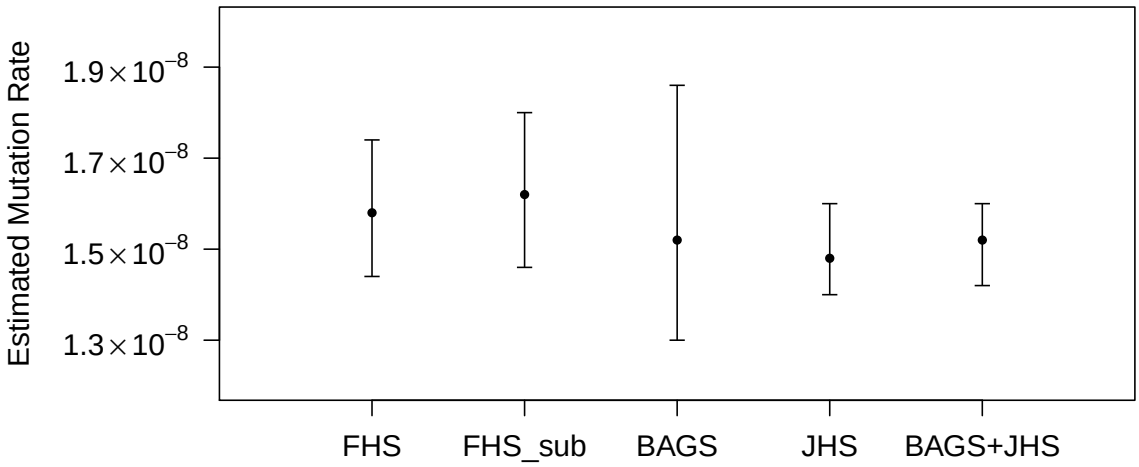


Figure 5.15: Estimated mutation rate from TOPMed studies. Point estimates (dots) and 95% confidence intervals (bars) corresponding to individuals from the Framingham Heart Study, a subset of the Framingham Heart Study based on excluding individuals with PC scores being outliers, the Barbados Asthma Genetic Study, the Jackson Heart Study, and the combined Barbados Asthma Genetic Study and Jackson Heart Study.

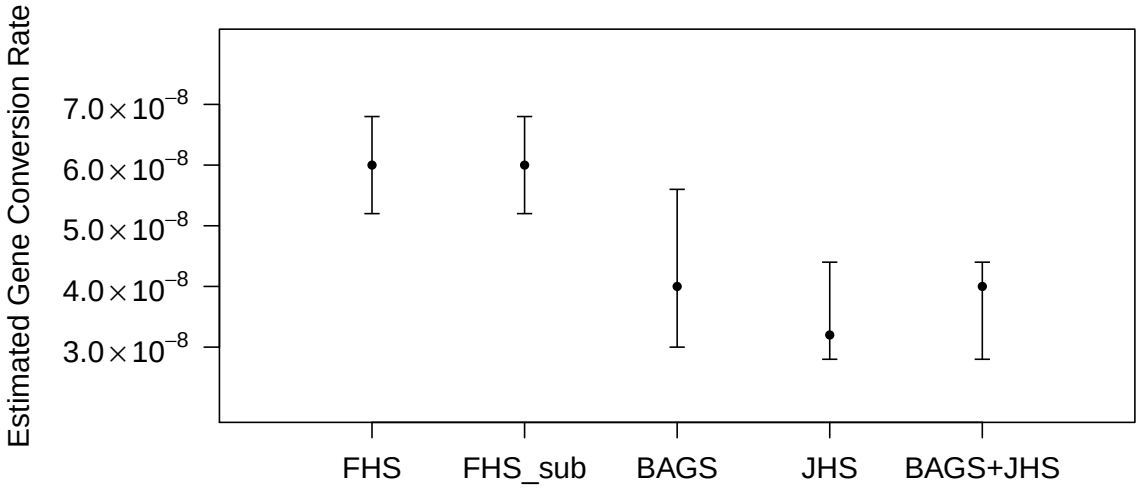


Figure 5.16: Estimated gene conversion rate from TOPMed studies. Point estimates (dots) and 95% confidence intervals (bars) corresponding to individuals from the Framingham Heart Study, a subset individuals of the Framingham Heart Study based on excluding individuals with PC scores being outliers, the Barbados Asthma Genetic Study, the Jackson Heart Study, and the combined Barbados Asthma Genetic Study and Jackson Heart Study.

Chapter 6

CONCLUSIONS AND FUTURE WORK

We analyzed whole genome sequence data from the Framingham Heart Study, the Jackson Heart Study, and the Barbados Asthma Genetic Study in the TOPMed project. Our analysis includes only single nucleotide substitutions; indels and other structural variants are excluded from the analysis. The estimated mutation rate and gene conversion rate for three studies are summarized in Figure 5.15, 5.16.

Our approach uses IBD segments detected with the Refined IBD or hap-IBD algorithm applied to phased data from parent-offspring trios or unphased data. Unlike pedigree-based mutation rate estimation, our method uses cross-family IBD to identify mutations arising over a much larger number of meiosis. The difficulty of distinguishing genotype error from true mutations is handled through statistical modelling, and is also reduced by requiring that the variants be seen in at least two of three identical-by-descent individuals. While the a priori rate of error for a very low frequency variant may be relatively high, evidence for a long IBD segment shared by the two individuals carrying the variant greatly increases the chance that the allele calls for that variant are accurate in those individuals.

When estimating mutation rate, the most serious type of genotype error is false positive error in which a major allele is called as the minor allele. These errors significantly increase the apparent mutation rate if not appropriately accounted for. Our method has strong control of these false positive errors through our error rate modelling. In addition, false negative errors, in which a minor allele is called as the major allele, can also have an effect, somewhat reducing the apparent mutation rate. Genotype error rates are highest for rare and singleton

variants because variant callers use databases of common SNPs to calibrate their results [53]. With our method, singleton variants are ignored. Furthermore, the lowest frequency non-singleton variants, such as variants with only two or three copies in a large data set, are very recent, and have a relatively greater impact on mutation counts for the longest IBD segments which represent very recent common ancestry. Thus we are able to significantly reduce the impact of false negative errors by excluding the very longest IBD segments.

Another potential source of error is gene conversion, which inserts existing alleles from one haplotype onto another haplotype. Such alleles can then differ between IBD haplotypes, which mimics the signal of mutation. However, most of the alleles inserted due to gene conversion are common alleles with high heterozygosity, while recent mutations are low in frequency. With this property, we model the gene conversion events together with mutations under the likelihood framework and separate apparent mutation counts due to gene conversion events from apparent mutation counts due to true mutations.

Our estimates are higher than previous pedigree-based estimates of mutation rate that range from 0.97×10^{-8} to 1.36×10^{-8} per base pair per meiosis [2, 12, 13, 16, 17, 27, 30, 46, 49, 57]. A major and difficult to quantify source of uncertainty in pedigree-based estimates is the choice of quality control filters to reduce the impact of genotype error. Overly stringent filtering will depress the mutation rate estimate [49]. In contrast, our method accounts for genotype error by modelling it, allowing for much reduced dependence on filtering. Pedigree-based methods are also affected by somatic mutation, unless three generations of individuals are genotyped [49]. Our approach is robust to somatic mutations, because such mutations will be carried by only one of the three IBD individuals that we consider, and thus will not be counted.

Our mutation rate estimate from the Framingham Heart Study is consistent with estimates obtained from two other IBD-based methods: An estimate of $1.66 \pm 0.04 \times 10^{-8}$

obtained by Palamara et al. with their pairwise-IBD-based method applied to data from the Genome of the Netherlands study [42], and an estimate of $1.61 \pm 0.26 \times 10^{-8}$ (confidence interval here given as estimate $\pm 2\times$ standard error, rather than estimate $\pm$ standard error given in the original publication) based on segments of ancient autozygosity in eight European and Asian individuals [32]. Our estimate of gene conversion rate from the Framingham Heart Study is consistent with estimates obtained from Palamara et al., which was $5.99 \pm 0.69 \times 10^{-6}$, however, due to the downward bias we observed from simulation studies, the true gene conversion rate might be higher than our estimate.

A major limitation for the pairwise-IBD-based method [42] is the dependency on accurately phased rare variants, which restricts the applicability of the method to data that include families. We extend the mutation counting procedure to account for the phasing uncertainty so that the our approach is applicable to analysis for larger data sets of unrelated individuals. In addition to the Framingham Heart Study, we obtained mutation and gene conversion rate estimates using the Barbados Asthma Genetic Study and the Jackson Heart Study, in which most of the individuals are of African descent.

We didn't observe significant differences in mutation rate estimates comparing the Framingham Heart Study, the Barbados Asthma Genetic Study and the Jackson Heart Study (Figure 5.15). Our estimate of gene conversion rate from the Framingham Heart Study data is much higher than that from the Barbados or Jackson Heart Study data (Figure 5.16). One possible explanation for this difference could be the degree of admixture: we observed that the estimates of gene conversion rate tend to be decreasing as the degree of admixture is increasing. Due to the observed bias in estimates of gene conversion rate in simulation studies, we are not able to draw strong conclusions based on our results. However, this interesting observation merits further investigation. Future improvements in IBD detection algorithms may provide more accurate IBD calling and have the potential to correct the bias in gene conversion rate estimation.

Our method is based on mutations that occurred since common ancestors living at most several hundred generations ago, which is more recent than continental-level population split times. Thus, there is potential to distinguish differences in mutation rates between populations, which may be due to differing environmental exposures or average parental ages [27, 37]. Future studies could also use the framework of three-way IBD sharing to obtain mutation rate estimates for separate mutation classes and facilitate the study of context-specific mutations. With larger sample sizes in future studies, there is also the potential to obtain mutation rate estimates for particular genomic regions or other subsets of the genome, in contrast to the genome-wide estimation that we performed here.

BIBLIOGRAPHY

[1] L. Arbiza, S. Gottipati, A. Siepel, and A. Keinan. Contrasting X-linked and autosomal diversity across 14 human populations. *Am J Hum Genet*, 94(6):827–44, 2014.

[2] P. Awadalla, J. Gauthier, R. A. Myers, F. Casals, F. F. Hamdan, A. R. Griffing, M. Cote, E. Henrion, D. Spiegelman, J. Tarabeux, A. Piton, Y. Yang, A. Boyko, C. Bustamante, L. Xiong, J. L. Rapoport, A. M. Addington, J. L. DeLisi, M. O. Krebs, R. Joober, B. Millet, E. Fombonne, L. Mottron, M. Zilversmit, J. Keebler, H. Daoud, C. Marineau, M. H. Roy-Gagnon, M. P. Dube, A. Eyre-Walker, P. Drapeau, E. A. Stone, R. G. Lafreniere, and G. A. Rouleau. Direct measure of the de novo mutation rate in autism and schizophrenia cohorts. *Am J Hum Genet*, 87(3):316–24, 2010.

[3] Douglas W Bjelland, Uday Lingala, Piyush S Patel, Matt Jones, and Matthew C Keller. A fast and accurate method for detection of IBD shared haplotypes in genome-wide SNP data. *European Journal of Human Genetics*, 25(5):617–624, 2017.

[4] B. L. Browning and S. R. Browning. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet*, 84(2):210–23, 2009.

[5] B. L. Browning and S. R. Browning. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics*, 194(2):459–71, 2013.

[6] Brian L. Browning and Sharon R. Browning. Detecting identity by descent and estimating genotype error rates in sequence data. *The American Journal of Human Genetics*, 93(5):840–851, 2019/10/13 2013.

[7] S. R. Browning, B. L. Browning, M. L. Daviglus, R. A. Durazo-Arvizu, N. Schneiderman, R. C. Kaplan, and C. C. Laurie. Ancestry-specific recent effective population size in the Americas. *PLoS Genet*, 14(5):e1007385, 2018.

[8] Sharon R Browning and Brian L Browning. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American journal of human genetics*, 81(5):1084–1097, 11 2007.

[9] Sharon R Browning and Brian L Browning. Haplotype phasing: existing methods and new developments. *Nature reviews. Genetics*, 12(10):703–714, 09 2011.

[10] Sharon R. Browning and Brian L. Browning. Identity by descent between distant relatives: detection and applications. *Annual Review of Genetics*, 46(1):617–633, 2012. PMID: 22994355.

[11] Sharon R Browning and Brian L Browning. Accurate non-parametric estimation of recent effective population size from segments of identity by descent. *American journal of human genetics*, 97(3):404–418, 09 2015.

[12] C. D. Campbell, J. X. Chong, M. Malig, A. Ko, B. L. Dumont, L. Han, L. Vives, B. J. O'Roak, P. H. Sudmant, J. Shendure, M. Abney, C. Ober, and E. E. Eichler. Estimating the human mutation rate using autozygosity in a founder population. *Nat Genet*, 44(11):1277–81, 2012.

[13] C. D. Campbell and E. E. Eichler. Properties and rates of germline mutations in humans. *Trends Genet*, 29(10):575–84, 2013.

[14] S. Carmi, K. Y. Hui, E. Kochav, X. Liu, J. Xue, F. Grady, S. Guha, K. Upadhyay, D. Ben-Avraham, S. Mukherjee, B. M. Bowen, T. Thomas, J. Vijai, M. Cruts, G. Froyen, D. Lambrechts, S. Plaisance, C. Van Broeckhoven, P. Van Damme, H. Van Marck, N. Barzilai, A. Darvasi, K. Offit, S. Bressman, L. J. Ozelius, I. Peter, J. H. Cho, H. Ostrer, G. Atzmon, L. N. Clark, T. Lencz, and I. Pe'er. Sequencing an Ashkenazi reference panel supports population-targeted personal genomics and illuminates Jewish and European origins. *Nat Commun*, 5:4835, 2014.

[15] G. K. Chen, P. Marjoram, and J. D. Wall. Fast and flexible simulation of DNA sequence data. *Genome Res*, 19(1):136–42, 2009.

[16] D. F. Conrad, J. E. Keebler, M. A. DePristo, S. J. Lindsay, Y. Zhang, F. Casals, Y. Idaghdour, C. L. Hartl, C. Torroja, K. V. Garimella, M. Zilversmit, R. Cartwright, G. A. Rouleau, M. Daly, E. A. Stone, M. E. Hurles, Awadalla P., and 1000 Genomes Project. Variation in genome-wide mutation rates within and between human families. *Nat Genet*, 43(7):712–4, 2011.

[17] 1000 Genomes Project Consortium, G. R. Abecasis, D. Altshuler, A. Auton, L. D. Brooks, R. M. Durbin, R. A. Gibbs, M. E. Hurles, and G. A. McVean. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–73, 2010.

[18] International HapMap Consortium, K. A. Frazer, D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve, R. A. Gibbs, J. W. Belmont, A. Boudreau, P. Hardenbol, S. M. Leal, S. Pasternak, D. A. Wheeler, T. D. Willis, F. Yu, H. Yang, C. Zeng, Y. Gao, H. Hu, W. Hu, C. Li, W. Lin, S. Liu, H. Pan, X. Tang, J. Wang, W. Wang, J. Yu,

B. Zhang, Q. Zhang, H. Zhao, H. Zhao, J. Zhou, S. B. Gabriel, R. Barry, B. Blumenstiel, A. Camargo, M. Defelice, M. Faggart, M. Goyette, S. Gupta, J. Moore, H. Nguyen, R. C. Onofrio, M. Parkin, J. Roy, E. Stahl, E. Winchester, L. Ziaugra, D. Altshuler, Y. Shen, Z. Yao, W. Huang, X. Chu, Y. He, L. Jin, Y. Liu, Y. Shen, W. Sun, H. Wang, Y. Wang, Y. Wang, X. Xiong, L. Xu, M. M. Waye, S. K. Tsui, H. Xue, J. T. Wong, L. M. Galver, J. B. Fan, K. Gunderson, S. S. Murray, A. R. Oliphant, M. S. Chee, A. Montpetit, F. Chagnon, V. Ferretti, M. Leboeuf, J. F. Olivier, M. S. Phillips, S. Roumy, C. Sallee, A. Verner, T. J. Hudson, P. Y. Kwok, D. Cai, D. C. Koboldt, R. D. Miller, L. Pawlikowska, P. Taillon-Miller, M. Xiao, L. C. Tsui, W. Mak, Y. Q. Song, P. K. Tam, Y. Nakamura, T. Kawaguchi, T. Kitamoto, T. Morizono, A. Nagashima, et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164):851–61, 2007.

[19] Ronald Aylmer Fisher. *The genetical theory of natural selection*. The Clarendon press, Oxford,, 1930.

[20] J. Gay, S. Myers, and G. McVean. Estimating meiotic gene conversion rates from population genetic data. *Genetics*, 177(2):881–94, 2007.

[21] Simon Gravel, Brenna M Henn, Ryan N Gutenkunst, Amit R Indap, Gabor T Marth, Andrew G Clark, Fuli Yu, Richard A Gibbs, 1000 Genomes Project, and Carlos D Bustamante. Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences of the United States of America*, 108(29):11983–11988, 07 2011.

[22] A. Gusev, J. K. Lowe, M. Stoffel, M. J. Daly, D. Altshuler, J. L. Breslow, J. M. Friedman, and I. Pe'er. Whole population, genome-wide mapping of hidden relatedness. *Genome Res*, 19(2):318–26, 2009.

[23] Ryan N Gutenkunst, Ryan D Hernandez, Scott H Williamson, and Carlos D Bustamante. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS genetics*, 5(10):e1000695–e1000695, 10 2009.

[24] J. B. S. Haldane. The combination of linkage values and the calculation of distance between the loci of linked factors. *Journal of Genetics*, 8:299–309, 1919.

[25] Z. He, X. Li, S. Ling, Y. X. Fu, E. Hungate, S. Shi, and C. I. Wu. Estimating DNA polymorphism from next generation sequencing data with high error rate by dual sequencing applications. *BMC Genomics*, 14:535, 2013.

[26] Jotun Hein, Mikkel H. Schierup, and Carsten Wiuf. *Gene genealogies, variation and evolution : a primer in coalescent theory*. Oxford University Press, Oxford ; New York, 2005.

[27] H. Jonsson, P. Sulem, B. Kehr, S. Kristmundsdottir, F. Zink, E. Hjartarson, M. T. Hardarson, K. E. Hjorleifsson, H. P. Eggertsson, S. A. Gudjonsson, L. D. Ward, G. A. Arnadottir, E. A. Helgason, H. Helgason, A. Gylfason, A. Jonasdottir, A. Jonasdottir, T. Rafnar, M. Frigge, S. N. Stacey, O. Th Magnusson, U. Thorsteinsdottir, G. Masson, A. Kong, B. V. Halldorsson, A. Helgason, D. F. Gudbjartsson, and K. Stefansson. Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature*, 549(7673):519–522, 2017.

[28] Alon Keinan and Andrew G Clark. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science (New York, N.Y.)*, 336(6082):740–743, 05 2012.

[29] M. Kimura. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, 61(4):893–903, 1969.

[30] A. Kong, M. L. Frigge, G. Masson, S. Besenbacher, P. Sulem, G. Magnusson, S. A. Gudjonsson, A. Sigurdsson, A. Jonasdottir, A. Jonasdottir, W. S. Wong, G. Sigurdsson, G. B. Walters, S. Steinberg, H. Helgason, G. Thorleifsson, D. F. Gudbjartsson, A. Helgason, O. T. Magnusson, U. Thorsteinsdottir, and K. Stefansson. Rate of de novo mutations and the importance of father's age to disease risk. *Nature*, 488(7412):471–5, 2012.

[31] Heng Li and Richard Durbin. Inference of human population history from individual whole-genome sequences. 475(7357):493–496, 07 2011.

[32] M. Lipson, P. R. Loh, S. Sankararaman, N. Patterson, B. Berger, and D. Reich. Calibrating the human mutation rate via ancestral recombination density in diploid genomes. *PLoS Genet*, 11(11):e1005550, 2015.

[33] Xiaoming Liu and Yun-Xin Fu. Exploring population size changes using SNP frequency spectra. *Nature genetics*, 47(5):555–559, 05 2015.

[34] M. Lynch. Evolution of the mutation rate. *Trends Genet*, 26(8):345–52, 2010.

[35] Tara C Matise, Fang Chen, Wenwei Chen, Francisco M De La Vega, Mark Hansen, Chunsheng He, Fiona C L Hyland, Giulia C Kennedy, Xiangyang Kong, Sarah S Murray, Janet S Ziegle, William C L Stewart, and Steven Buyske. A second-generation combined linkage physical map of the human genome. *Genome research*, 17(12):1783–1786, 12 2007.

[36] Gilean A T McVean and Niall J Cardin. Approximating the coalescent with recombination. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 360(1459):1387–1393, 07 2005.

[37] V. M. Narasimhan, R. Rahbari, A. Scally, A. Wuster, D. Mason, Y. Xue, J. Wright, R. C. Trembath, E. R. Maher, D. A. van Heel, A. Auton, M. E. Hurles, C. Tyler-Smith, and R. Durbin. Estimating the human mutation rate from autozygous segments reveals population differences in human mutational processes. *Nat Commun*, 8(1):303, 2017.

[38] Ardalan Naseri, Xiaoming Liu, Kecong Tang, Shaojie Zhang, and Degui Zhi. Rapid: ultra-fast, powerful, and accurate detection of segments identical by descent (IBD) in biobank-scale cohorts. *Genome Biology*, 20(1):143, 2019.

[39] R Nielsen. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics*, 154(2):931–942, 02 2000.

[40] Rasmus Nielsen, Joshua S. Paul, Anders Albrechtsen, and Yun S. Song. Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, 12(6):443–451, 2011.

[41] P. F. Palamara. ARGON: fast, whole-genome simulation of the discrete time Wright-Fisher process. *Bioinformatics*, 32(19):3032–4, 2016.

[42] P. F. Palamara, L. C. Francioli, P. R. Wilton, G. Genovese, A. Gusev, H. K. Finucane, S. Sankararaman, Consortium Genome of the Netherlands, S. R. Sunyaev, P. I. de Bakker, J. Wakeley, I. Pe'er, and A. L. Price. Leveraging distant relatedness to quantify human mutation and gene-conversion rates. *Am J Hum Genet*, 97(6):775–89, 2015.

[43] Pier Francesco Palamara, Todd Lencz, Ariel Darvasi, and Itsik Pe'er. Length distributions of identity by descent reveal fine-scale demographic history. *American journal of human genetics*, 91(5):809–822, 11 2012.

[44] François Pompanon, Aurélie Bonin, Eva Bellemain, and Pierre Taberlet. Genotyping errors: causes, consequences and solutions. *Nature Reviews Genetics*, 6(11):847–859, 2005.

[45] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A R Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I W de Bakker, Mark J Daly, and Pak C Sham. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics*, 81(3):559–575, 09 2007.

[46] Jared C Roach, Gustavo Glusman, Arian F A Smit, Chad D Huff, Robert Hubley, Paul T Shannon, Lee Rowen, Krishna P Pant, Nathan Goodman, Michael Bamshad, Jay Shendure, Radoje Drmanac, Lynn B Jorde, Leroy Hood, and David J Galas. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science (New York, N.Y.)*, 328(5978):636–639, 04 2010.

[47] A. Scally and R. Durbin. Revising the human mutation rate: implications for understanding human evolution. *Nat Rev Genet*, 13(10):745–53, 2012.

[48] Stephan Schiffels and Richard Durbin. Inferring human population size and separation history from multiple genome sequences. *Nature genetics*, (8):919–925, 08 2014.

[49] L. Segurel, M. J. Wyman, and M. Przeworski. Determinants of mutation rate variation in the human germline. *Annu Rev Genomics Hum Genet*, 15:47–70, 2014.

[50] Daniel Taliun, Daniel N. Harris, Michael D. Kessler, Jedidiah Carlson, Zachary A. Szpiech, Raul Torres, Sarah A. Gagliano Taliun, André Corvelo, Stephanie M. Gogarten, Hyun Min Kang, Achilleas N. Pitsillides, Jonathon LeFaive, Seung-been Lee, Xiaowen Tian, Brian L. Browning, Sayantan Das, Anne-Katrin Emde, Wayne E. Clarke, Douglas P. Loesch, Amol C. Shetty, Thomas W. Blackwell, Quenna Wong, François Aguet, Christine Albert, Alvaro Alonso, Kristin G. Ardlie, Stella Aslibekyan, Paul L. Auer, John Barnard, R. Graham Barr, Lewis C. Becker, Rebecca L. Beer, Emelia J. Benjamin, Lawrence F. Bielak, John Blangero, Michael Boehnke, Donald W. Bowden, Jennifer A. Brody, Esteban G. Burchard, Brian E. Cade, James F. Casella, Brandon Chalazan, Yii-Der Ida Chen, Michael H. Cho, Seung Hoan Choi, Mina K. Chung, Clary B. Clish, Adolfo Correa, Joanne E. Curran, Brian Custer, Dawood Darbar, Michelle Daya, Mariza de Andrade, Dawn L. DeMeo, Susan K. Dutcher, Patrick T. Ellinor, Leslie S. Emery, Diane Fatkin, Lukas Forer, Myriam Fornage, Nora Franceschini, Christian Fuchsberger, Stephanie M. Fullerton, Soren Germer, Mark T. Gladwin, Daniel J. Gottlieb, Xiuqing Guo, Michael E. Hall, Jiang He, Nancy L. Heard-Costa, Susan R. Heckbert, Marguerite R. Irvin, Jill M. Johnsen, Andrew D. Johnson, Sharon L.R. Kardia, Tanika Kelly, Shannon Kelly, Eimear E. Kenny, Douglas P. Kiel, Robert Klemmer, Barbara A. Konkle, Charles Kooperberg, Anna Köttgen, Leslie A. Lange, Jessica Lasky-Su, Daniel Levy, Xihong Lin, Keng-Han Lin, Chunyu Liu, Ruth J.F. Loos, Lori Garman, Robert Gerszten, Steven A. Lubitz, Kathryn L. Lunetta, Angel C.Y. Mak, Ani Manichaikul, Alisa K. Manning, Rasika A. Mathias, David D. McManus, Stephen T. McGarvey, James B. Meigs, Deborah A. Meyers, Julie L. Mikulla, Mollie A. Minear, Braxton Mitchell, Sanghamitra Mohanty, May E. Montasser, Courtney Montgomery, Alanna C. Morrison, Joanne M. Murabito, Andrea Natale, Pradeep Natarajan, Sarah C. Nelson, Kari E. North, Jeffrey R. O'Connell, Nicholette D. Palmer, Nathan Pankratz, Gina M. Peloso, Patricia A. Peyser, Wendy S. Post, Bruce M. Psaty, D.C. Rao, Susan Redline, Alexander P. Reiner, Dan Roden, Jerome I. Rotter, Ingo Ruczinski, Chloé Sarnowski, Sebastian Schoenherr, Jeong-Sun Seo, Sudha Seshadri, Vivien A. Sheehan, M. Benjamin Shoemaker, Albert V. Smith, Nicholas L. Smith, Jennifer A. Smith, Nona Sotoodehnia, Adrienne M. Stilp, Weihong Tang, Kent D. Taylor, Marilyn Telen, Timothy A. Thornton, Russell P. Tracy, David J. Van Den Berg, Ramachandran S. Vasan, Karine A. Viaud-Martinez, Scott Vrieze, Daniel E Weeks, Bruce S. Weir, Scott T. Weiss, Lu-Chen Weng, Cristen J. Willer, Yingze Zhang, Xutong Zhao, Donna K. Arnett, Allison E.

Ashley-Koch, Kathleen C. Barnes, Eric Boerwinkle, Stacey Gabriel, Richard Gibbs, Kenneth M. Rice, Stephen S. Rich, Edwin Silverman, Pankaj Qasba, Weiniu Gan, , George J. Papanicolaou, Deborah A. Nickerson, Sharon R. Browning, Michael C. Zody, Sebastian Zöllner, James G. Wilson, L Adrienne Cupples, Cathy C. Laurie, Cashell E. Jaquish, Ryan D. Hernandez, Timothy D. O'Connor, and Gonçalo R. Abecasis. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed program. *bioRxiv*, 2019.

[51] Jonathan Terhorst and Yun S. Song. Fundamental limits on the accuracy of demographic inference based on the sample frequency spectrum. *Proceedings of the National Academy of Sciences*, 112(25):7677–7682, 2015.

[52] Elizabeth A Thompson. Identity by descent: variation in meiosis, across genomes, and in populations. *Genetics*, 194(2):301–326, 06 2013.

[53] J. D. Wall, L. F. Tang, B. Zerbe, M. N. Kvale, P. Y. Kwok, C. Schaefer, and N. Risch. Estimating genotype error rates from high-coverage next-generation sequence data. *Genome Res*, 24(11):1734–9, 2014.

[54] Klaudia Walter, Josine L. Min, Jie Huang, Lucy Crooks, Yasin Memari, Shane McCarthy, John R. B. Perry, ChangJiang Xu, Marta Futema, Daniel Lawson, Valentina Iotchkova, Stephan Schiffels, Audrey E. Hendricks, Petr Danecek, Rui Li, James Floyd, Louise V. Wain, Inês Barroso, Steve E. Humphries, Matthew E. Hurles, Eleftheria Zeggini, Jeffrey C. Barrett, Vincent Plagnol, J. Brent Richards, Celia M. T. Greenwood, Nicholas J. Timpson, Richard Durbin, Nicole Soranzo, Senduran Bala, Peter Clapham, Guy Coates, Tony Cox, Allan Daly, Yuanping Du, Sarah Edkins, Peter Ellis, Paul Flicek, Xiaosen Guo, Xueqin Guo, Liren Huang, David K. Jackson, Chris Joyce, Thomas Keane, Anja Kolb-Kokocinski, Cordelia Langford, Yingrui Li, Jieqin Liang, Hong Lin, Ryan Liu, John Maslen, Dawn Muddyman, Michael A. Quail, Jim Stalker, Jianping Sun, Jing Tian, Guangbiao Wang, Jun Wang, Yu Wang, Kim Wong, Pingbo Zhang, Ewan Birney, Chris Boustred, Lu Chen, Gail Clement, Massimiliano Cocca, George Davey Smith, Ian N. M. Day, Aaron Day-Williams, Thomas Down, Ian Dunham, David M. Evans, Tom R. Gaunt, Matthias Geihs, Celia M. T. Greenwood, Deborah Hart, Audrey E. Hendricks, Bryan Howie, Tim Hubbard, Pirro Hysi, Yalda Jamshidi, Konrad J. Karczewski, John P. Kemp, Genevieve Lachance, Monkol Lek, Margarida Lopes, Daniel G. MacArthur, Jonathan Marchini, Massimo Mangino, Iain Mathieson, Sarah Metrustry, Josine L. Min, Alireza Moayyeri, Kate Northstone, Kalliope Panoutsopoulou, Lavinia Paternoster, John R. B. Perry, Lydia Quaye, J. Brent Richards, Susan Ring, Graham R. S. Ritchie, Hashem A. Shihab, So-Youn Shin, Kerrin S. Small, María Soler Artigas, Lorraine Southam, Timothy D. Spector, Beate St Pourcain, Gabriela Surdulescu, Ioanna Tachmazidou, Nicholas J. Timpson, Martin D. Tobin, Ana M. Valdes, Peter M. Visscher, Louise V. Wain, Kirsten Ward, Scott G. Wilson, Jian Yang, Feng Zhang, Hou-Feng Zheng, Richard Anney, Muhammad Ayub, Jeffrey C. Barrett, Douglas

Blackwood, Patrick F. Bolton, Gerome Breen, David A. Collier, Nick Craddock, Sarah Curran, David Curtis, Louise Gallagher, Daniel Geschwind, Hugh Gurling, Peter Holmans, Irene Lee, Jouko Lönnqvist, Peter McGuffin, Andrew M. McIntosh, Andrew G. McKechanie, Andrew McQuillin, James Morris, Michael C. O'Donovan, Michael J. Owen, Aarno Palotie, Jeremy R. Parr, Tiina Paunio, Olli Pietilainen, Karola Rehnström, Sally I. Sharp, David Skuse, David St Clair, Jaana Suvisaari, James T. R. Walters, Hywel J. Williams, Elena Bochukova, Rebecca Bounds, Anna Dominiczak, I. Sadaf Farooqi, Audrey E. Hendricks, Julia Keogh, Gaëlle Marenne, Andrew Morris, Stephen O'Rahilly, David J. Porteous, Blair H. Smith, Eleanor Wheeler, Saeed Al Turki, Carl A. Anderson, Dinu Antony, Phil Beales, Jamie Bentham, Shoumo Bhattacharya, Mattia Calissano, Keren Carss, Krishna Chatterjee, Sebahattin Cirak, Catherine Cosgrove, David R. Fitzpatrick, A. Reghan Foley, Christopher S. Franklin, Detelina Grozeva, Steve E. Humphries, Matthew E. Hurles, Hannah M. Mitchison, Francesco Muntoni, Alexandros Onoufriadis, Victoria Parker, Felicity Payne, F. Lucy Raymond, Nicola Roberts, David B. Savage, Peter Scambler, Miriam Schmidts, Nadia Schoenmakers, Robert K. Semple, Eva Serra, Olivera Spasic-Boskovic, Elizabeth Stevens, Margriet van Kogelenberg, Parthiban Vijayarangakannan, Kathleen A. Williamson, Crispian Wilson, Tamieka Whyte, Antonio Ciampi, Celia M. T. Greenwood, Audrey E. Hendricks, Karim Oualkacha, Martin Bobrow, Patrick F. Bolton, David R. Fitzpatrick, Heather Griffin, Matthew E. Hurles, Jane Kaye, Karen Kennedy, Alastair Kent, F. Lucy Raymond, Robert K. Semple, Carol Smee, Timothy D. Spector, Nicholas J. Timpson, Ruth Charlton, Rosemary Ekong, Steve E. Humphries, Farrah Khawaja, Luis R. Lopes, Nicola Migone, Stewart J. Payne, Rebecca C. Pollitt, Sue Povey, Cheryl K. Ridout, Rachel L. Robinson, Richard H. Scott, Adam Shaw, Petros Syrris, Rohan Taylor, Anthony M. Vandersteen, Jeffrey C. Barrett, I. Sadaf Farooqi, David R. Fitzpatrick, Matthew E. Hurles, The UK10K Consortium, Writing group, Production group, Cohorts group, Neurodevelopmental disorders group, Obesity group, Rare disease group, Statistics group, Ethics group, Incidental findings group, and Management committee. The UK10K project identifies rare variants in health and disease. *Nature*, 526(7571):82–90, 2015.

[55] B. S. Weir and C. C. Cockerham. Estimating F-Statistics for the analysis of population structure. *Evolution*, 38(6):1358–1370, 1984.

[56] A. L. Williams, G. Genovese, T. Dyer, N. Altemose, K. Truax, G. Jun, N. Patterson, S. R. Myers, J. E. Curran, R. Duggirala, J. Blangero, D. Reich, M. Przeworski, and T. D. Genes Consortium. Non-crossover gene conversions show strong GC bias and unexpected clustering in humans. *Elife*, 4, 2015.

[57] W. S. Wong, B. D. Solomon, D. L. Bodian, P. Kothiyal, G. Eley, K. C. Huddleston, R. Baker, D. C. Thach, R. K. Iyer, J. G. Vockley, and J. E. Niederhuber. New observations on maternal age effect on germline de novo mutations. *Nat Commun*, 7:10486, 2016.

[58] S. Wright. Evolution in mendelian populations. *Genetics*, 16(2):97–159, 1931.

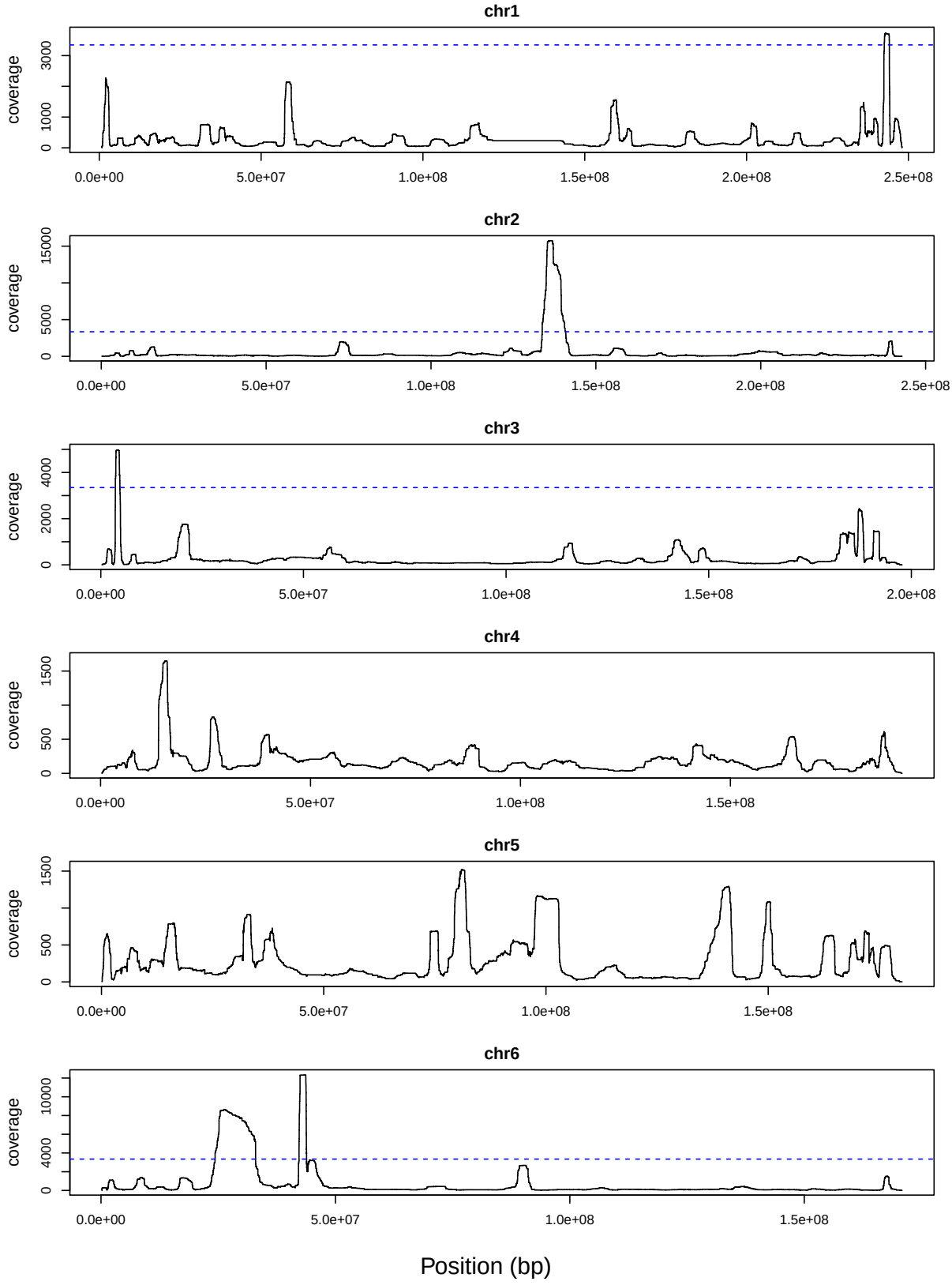
[59] Qi Yan, Rui Chen, James S. Sutcliffe, Edwin H. Cook, Daniel E. Weeks, Bingshan Li, and Wei Chen. The impact of genotype calling errors on family-based studies. *Scientific Reports*, 6(1):28323, 2016.

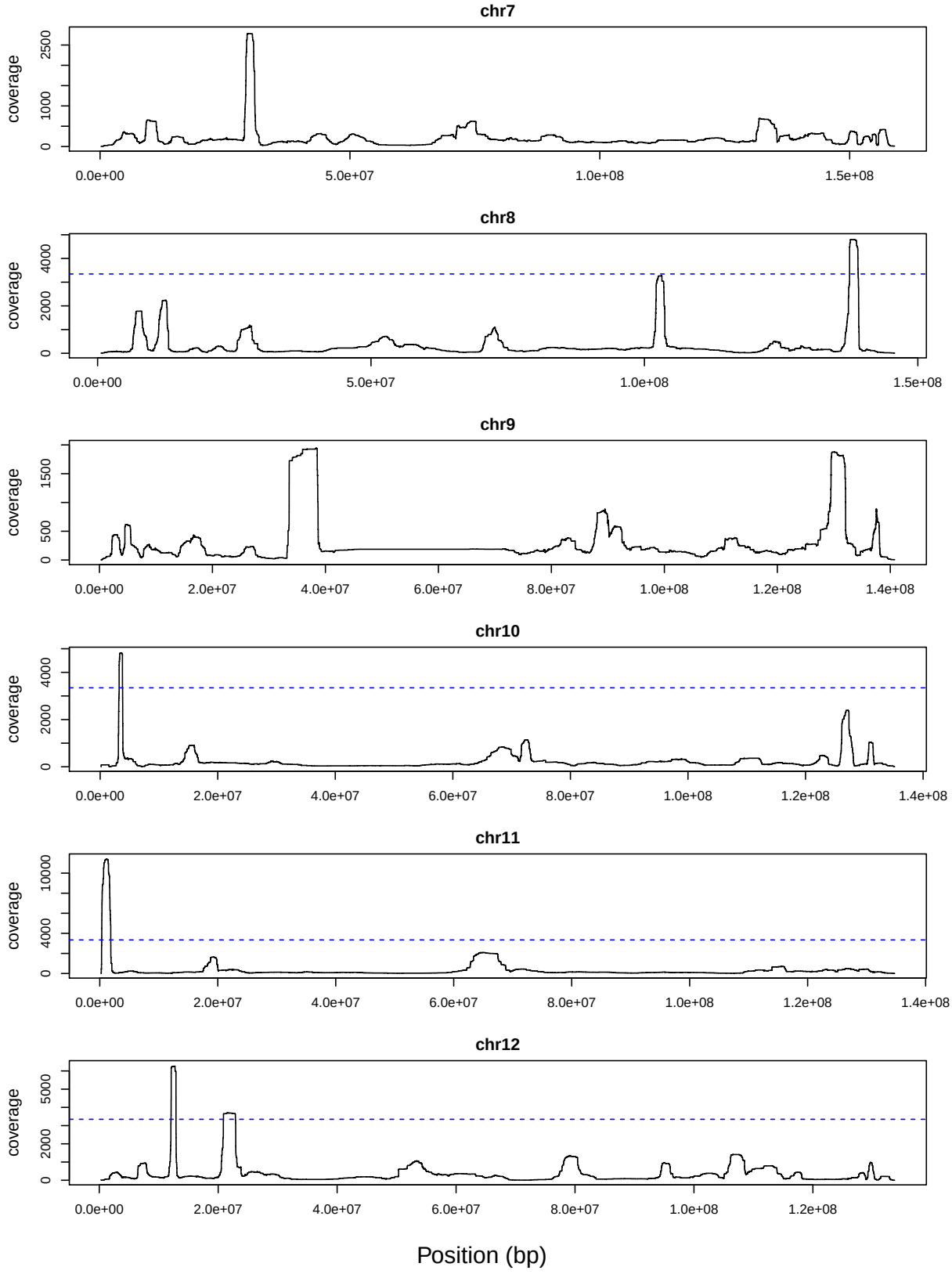
[60] Ying Zhou, Sharon Browning, and Brian L Browning. A fast and simple method for detecting identity by descent segments in large-scale data. *bioRxiv*, 2019.

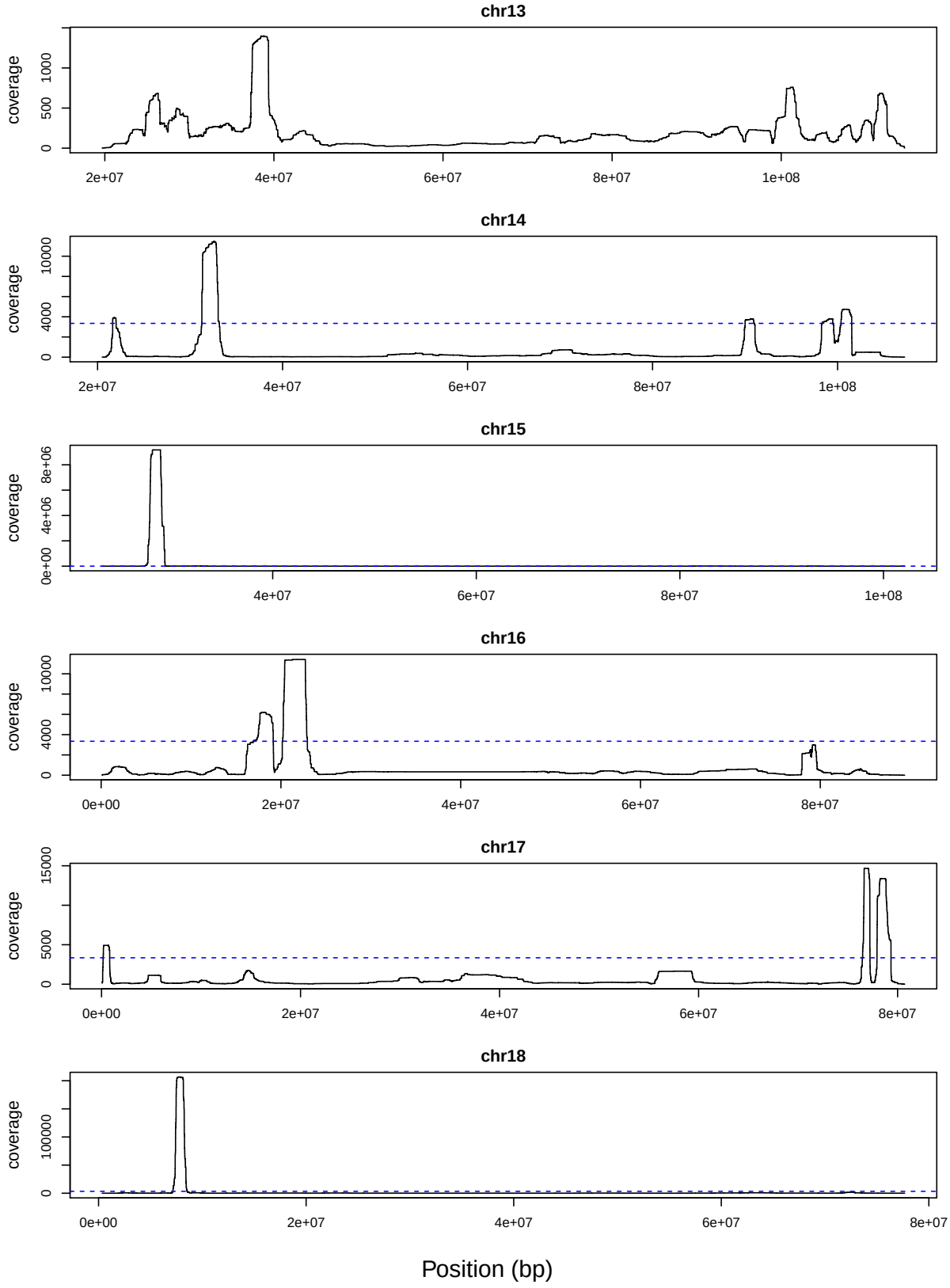
Appendix A

ANALYSIS OF TOPMED DATA

The appendix includes three figures and three tables on three-way IBD coverage along the genome and regions excluded from the analysis for mutation rate in Framingham Heart Study, Barbados Asthma Genetic Study and Jackson Heart Study.


chr1
coverage
0
1000
3000
0.0e+00
5.0e+07
1.0e+08
1.5e+08
2.0e+08
2.5e+08
chr2
coverage
0
5000
15000
0.0e+00
5.0e+07
1.0e+08
1.5e+08
2.0e+08
2.5e+08
chr3
coverage
0
2000
4000
0.0e+00
5.0e+07
1.0e+08
1.5e+08
2.0e+08
chr4
coverage
0
500
1500
0.0e+00
5.0e+07
1.0e+08
1.5e+08
chr5
coverage
0
500
1500
0.0e+00
5.0e+07
1.0e+08
1.5e+08
chr6
coverage
0
4000
10000
0.0e+00
5.0e+07
1.0e+08
1.5e+08
Position (bp)


chr7
coverage
0
1000
2500
0.0e+00
5.0e+07
1.0e+08
1.5e+08
chr8
coverage
0
2000
4000
0.0e+00
5.0e+07
1.0e+08
1.5e+08
chr9
coverage
0
500
1500
0.0e+00
2.0e+07
4.0e+07
6.0e+07
8.0e+07
1.0e+08
1.2e+08
1.4e+08
chr10
coverage
0
2000
4000
0.0e+00
2.0e+07
4.0e+07
6.0e+07
8.0e+07
1.0e+08
1.2e+08
1.4e+08
chr11
coverage
0
4000
10000
0.0e+00
2.0e+07
4.0e+07
6.0e+07
8.0e+07
1.0e+08
1.2e+08
1.4e+08
chr12
coverage
0
2000
5000
0.0e+00
2.0e+07
4.0e+07
6.0e+07
8.0e+07
1.0e+08
1.2e+08
Position (bp)


chr13
coverage
0
500
1000
2e+07
4e+07
6e+07
8e+07
1e+08
chr14
coverage
0
4000
10000
2e+07
4e+07
6e+07
8e+07
1e+08
chr15
coverage
0e+00
4e+06
8e+06
4e+07
6e+07
8e+07
1e+08
chr16
coverage
0
4000
10000
0e+00
2e+07
4e+07
6e+07
8e+07
chr17
coverage
0
5000
15000
0e+00
2e+07
4e+07
6e+07
8e+07
chr18
coverage
0
100000
0e+00
2e+07
4e+07
6e+07
8e+07
Position (bp)

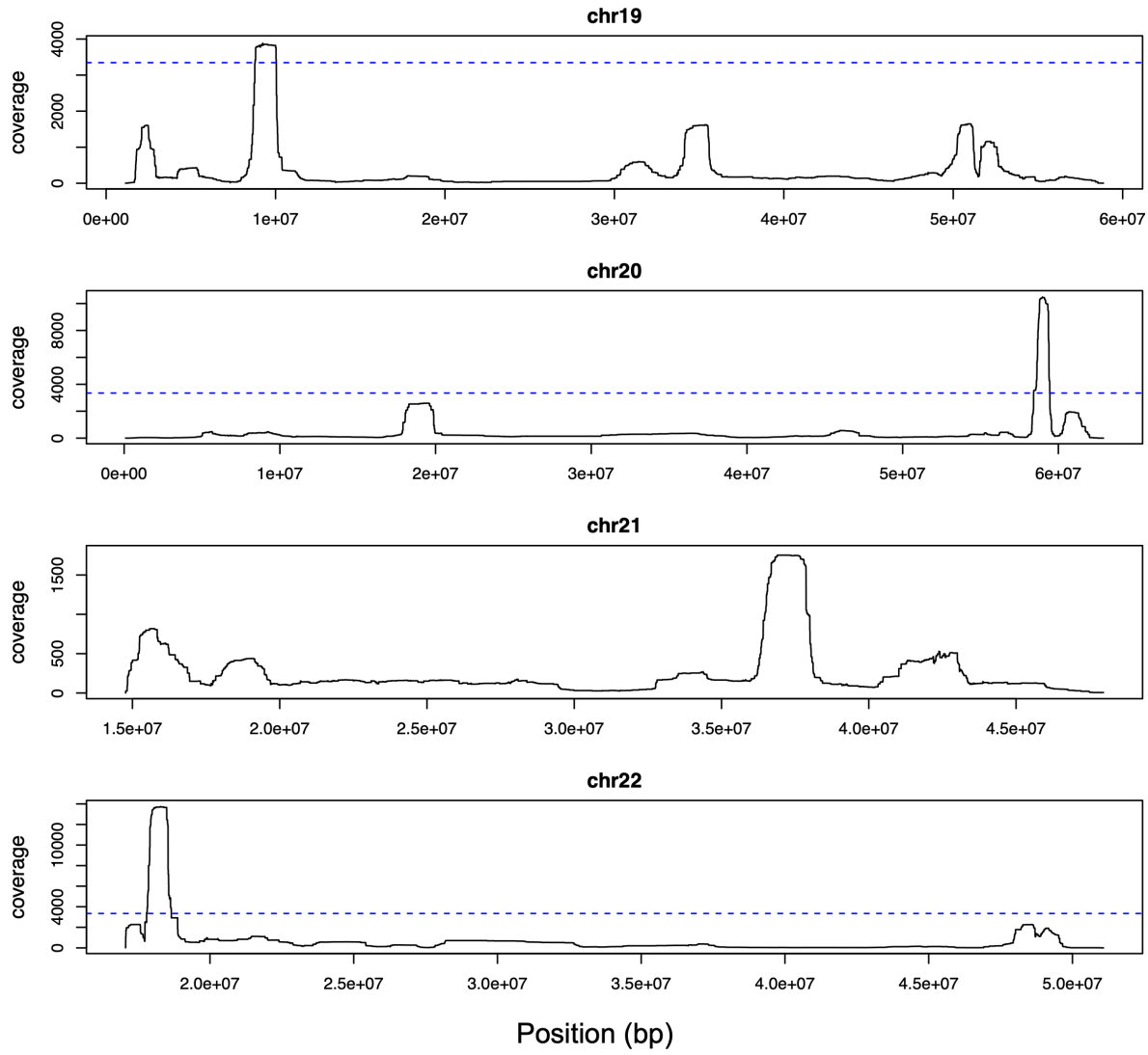
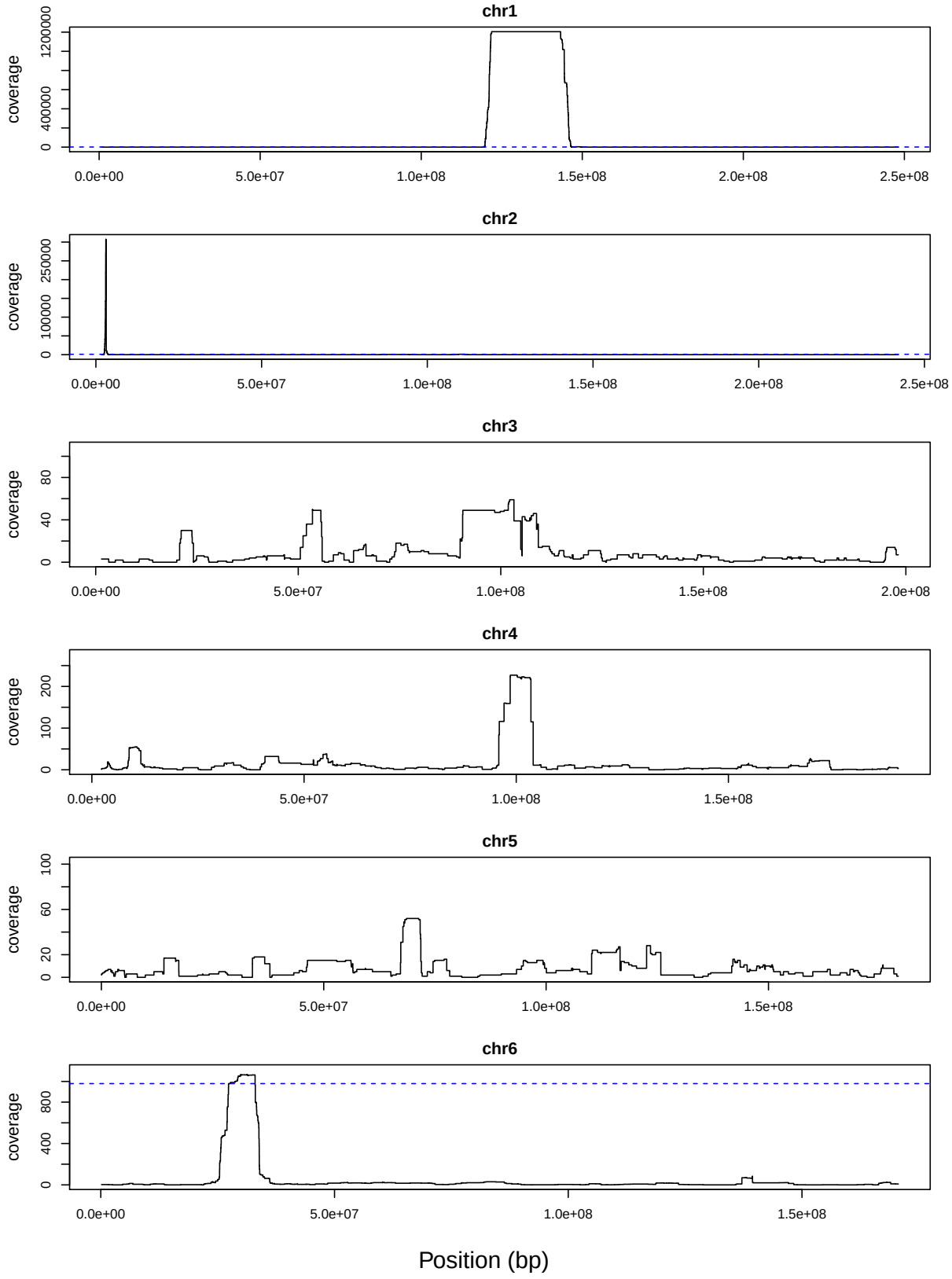
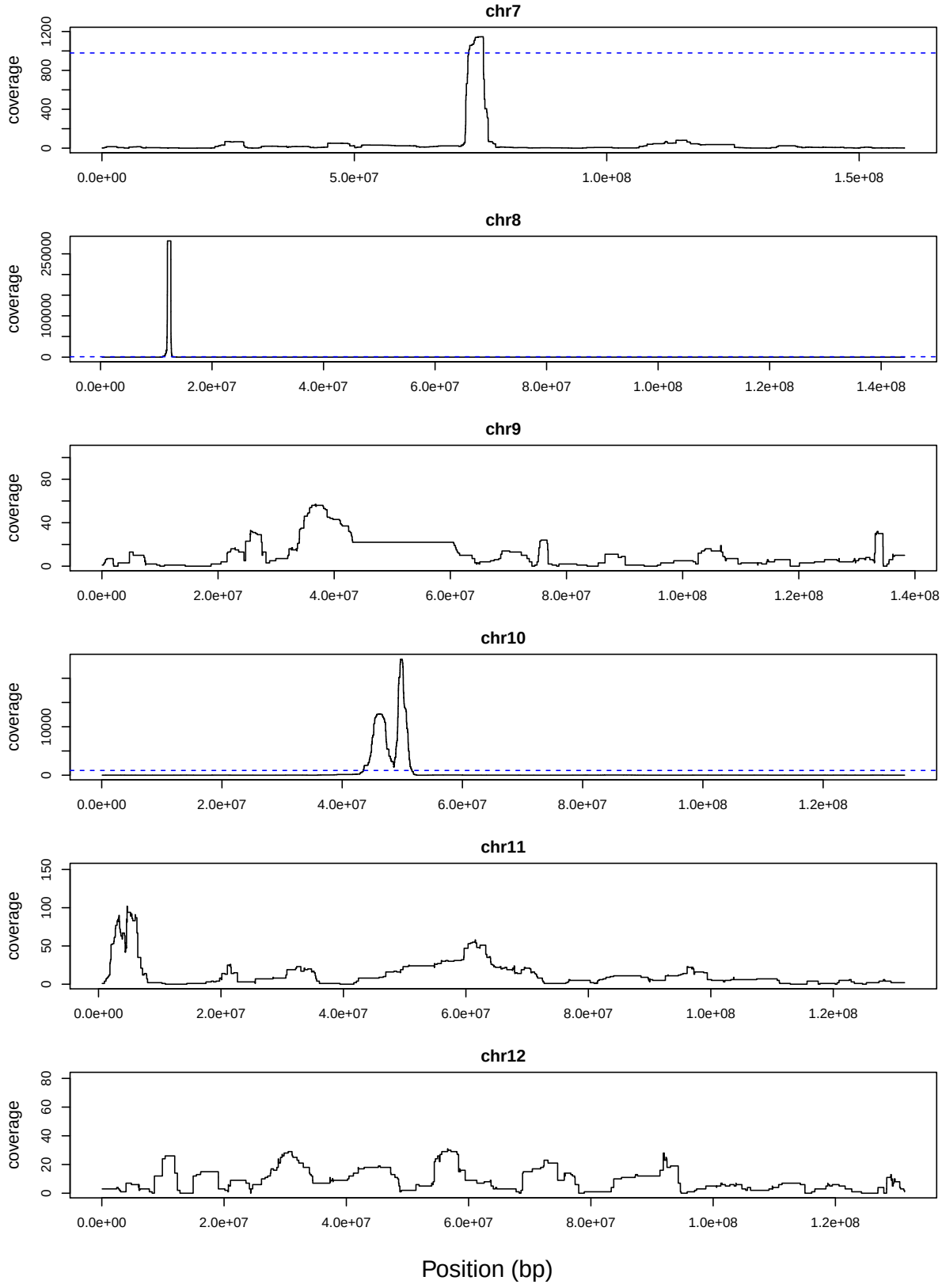


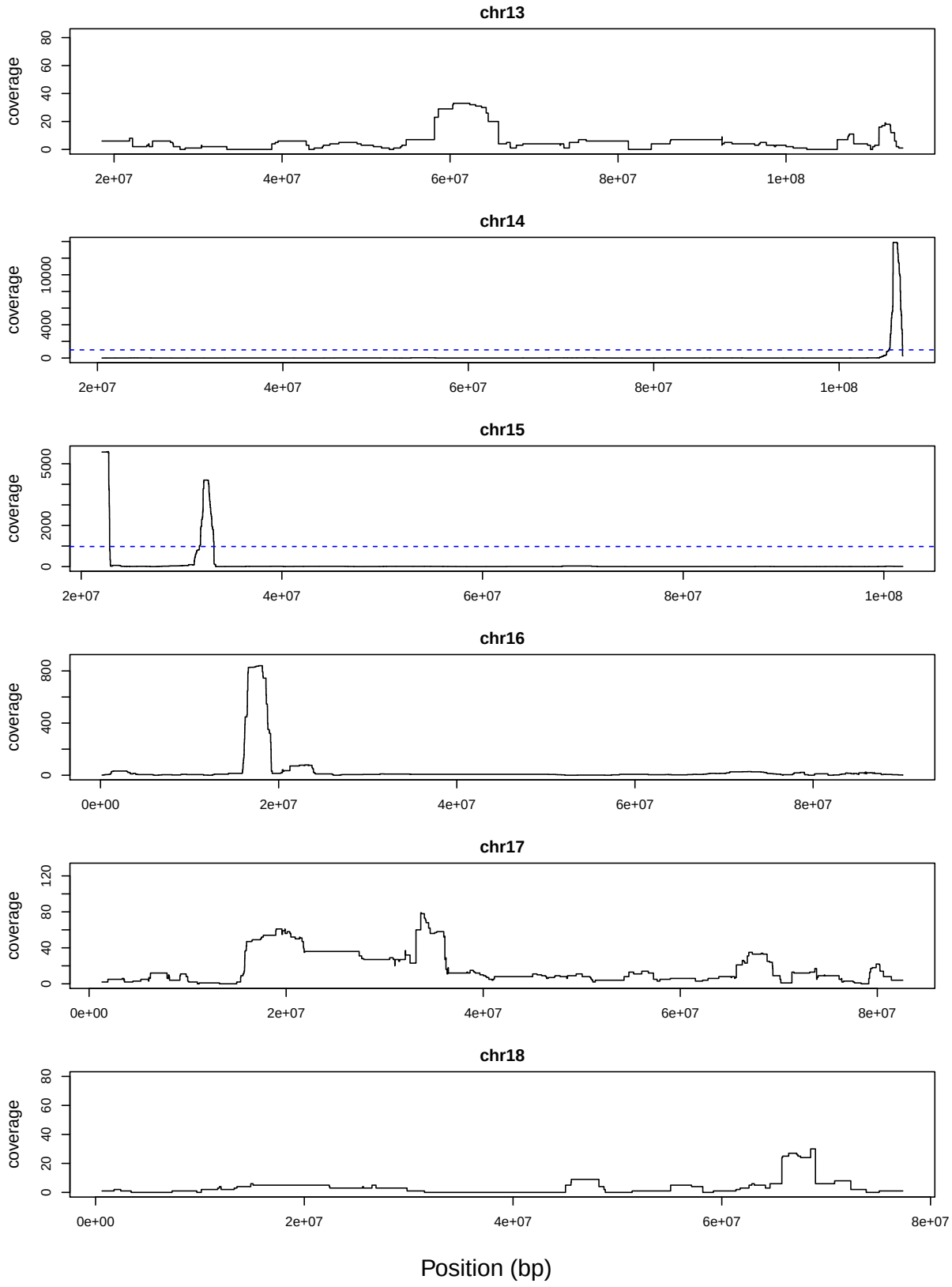
Figure A.1: Three-way IBD coverage along the genome in full Framingham Heart Study data. The y-axis indicates number of three-way IBD along the genome in windows of 500 base pairs. The blue dashed line represent the 98th percentile of genome wide three-way IBD coverage. Regions with three-way IBD coverage exceeded the 98th percentile are removed.

chromosome	from bp	end bp	chromosome	from bp	end bp
1	1000216	242284972	14	20981235	21559960
1	244358023	247403694	14	22212347	30064066
2	743486	132997809	14	33361280	89745757
2	142115598	242090274	14	91324569	98084801
3	672050	3434466	14	101683984	106782735
3	5038131	6967605	15	23428243	27168437
3	6994424	197128094	15	33739358	80728860
4	1251862	190372522	15	86809642	92394309
5	385813	179317412	15	93842962	101858715
6	367535	23911919	16	259887	15541238
6	33543984	41798683	16	23661557	88724445
6	44100816	170036836	17	967716	71918201
7	933249	158480466	17	72024999	76281939
8	964636	136967322	18	596583	7019995
8	139424231	145114729	18	8746074	55045003
9	524467	140342606	18	55501032	77298110
10	346997	3038041	19	1439442	7386613
10	3959996	134562609	19	7508595	8639706
11	2219731	134037830	19	10747254	58195905
12	540246	11539767	20	337219	58147509
12	13090146	20527186	20	59616397	62245936
12	23577720	132896219	21	14960190	47280672
13	20704721	22466545	22	17163278	17688249
13	22477191	99098728	22	19148181	47293063
13	99189475	114150139	22	47305945	50493349

Table A.1: Regions included in mutation rate analysis in full Framingham Heart Study data (build GRCh37)


chr1
coverage
0
400000
1200000
0.0e+00
5.0e+07
1.0e+08
1.5e+08
2.0e+08
2.5e+08
chr2
coverage
0
100000
250000
0.0e+00
5.0e+07
1.0e+08
1.5e+08
2.0e+08
2.5e+08
chr3
coverage
0
40
80
0.0e+00
5.0e+07
1.0e+08
1.5e+08
2.0e+08
chr4
coverage
0
100
200
0.0e+00
5.0e+07
1.0e+08
1.5e+08
chr5
coverage
0
20
60
100
0.0e+00
5.0e+07
1.0e+08
1.5e+08
chr6
coverage
0
400
800
0.0e+00
5.0e+07
1.0e+08
1.5e+08
Position (bp)


chr7
coverage
0
400
800
1200
0.0e+00
5.0e+07
1.0e+08
1.5e+08
chr8
coverage
0
10000
250000
0.0e+00
2.0e+07
4.0e+07
6.0e+07
8.0e+07
1.0e+08
1.2e+08
1.4e+08
chr9
coverage
0
40
80
0.0e+00
2.0e+07
4.0e+07
6.0e+07
8.0e+07
1.0e+08
1.2e+08
1.4e+08
chr10
coverage
0
10000
0.0e+00
2.0e+07
4.0e+07
6.0e+07
8.0e+07
1.0e+08
1.2e+08
chr11
coverage
0
50
100
150
0.0e+00
2.0e+07
4.0e+07
6.0e+07
8.0e+07
1.0e+08
1.2e+08
chr12
coverage
0
20
40
60
80
0.0e+00
2.0e+07
4.0e+07
6.0e+07
8.0e+07
1.0e+08
1.2e+08
Position (bp)


chr13
coverage
0
20
40
60
80
2e+07
4e+07
6e+07
8e+07
1e+08
chr14
coverage
0
4000
10000
2e+07
4e+07
6e+07
8e+07
1e+08
chr15
coverage
0
2000
5000
2e+07
4e+07
6e+07
8e+07
1e+08
chr16
coverage
0
400
800
0e+00
2e+07
4e+07
6e+07
8e+07
chr17
coverage
0
40
80
120
0e+00
2e+07
4e+07
6e+07
8e+07
chr18
coverage
0
20
40
60
80
0e+00
2e+07
4e+07
6e+07
8e+07
Position (bp)

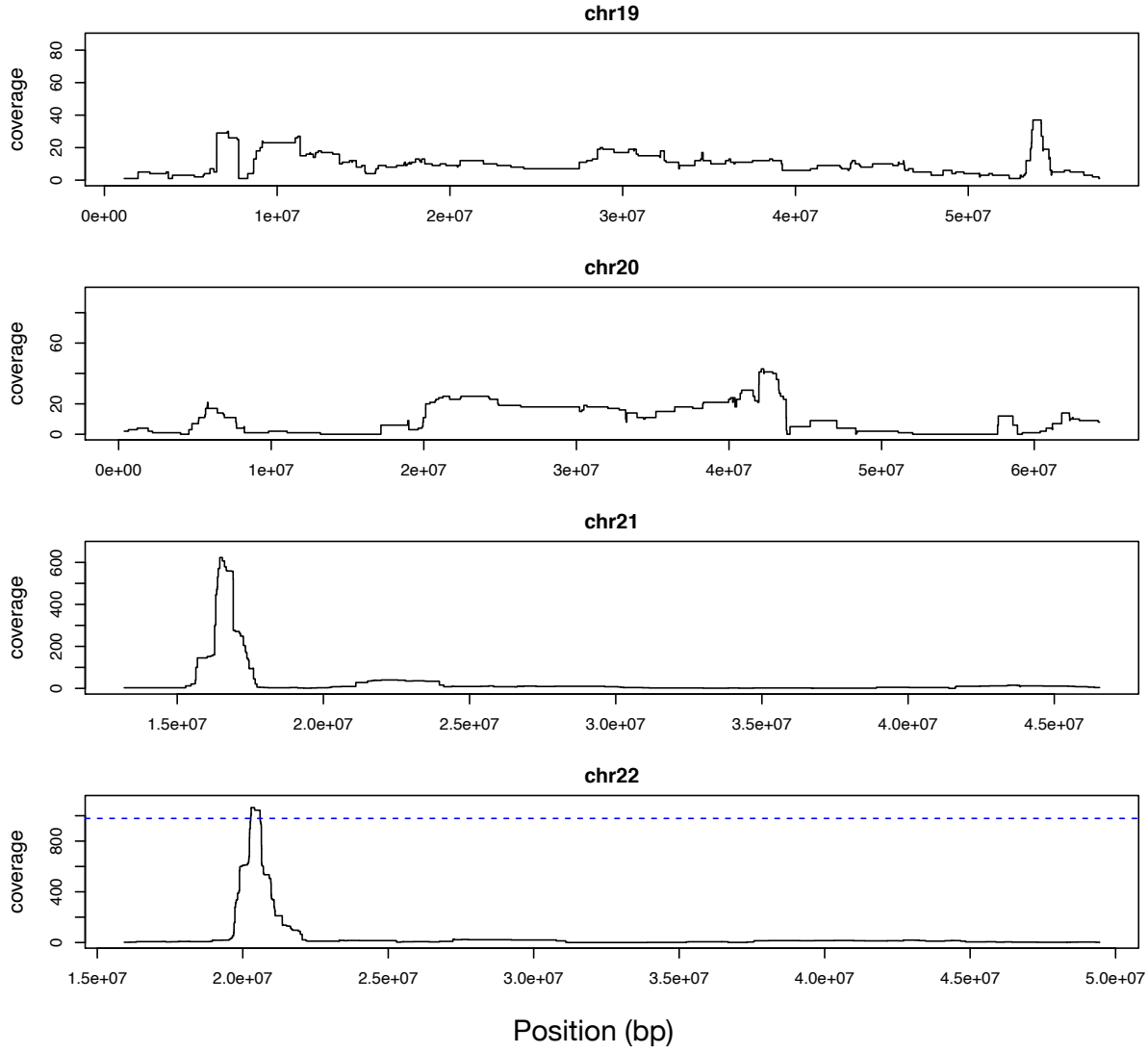


Figure A.2: Three-way IBD coverage along the genome in Barbados Asthma Genetic Study. The y-axis indicates number of three-way IBD along the genome in windows of 500 base pairs. The blue dashed line represent the 98th percentile of genome wide three-way IBD coverage. Regions with three-way IBD coverage exceeded the 98th percentile are removed.

chromosome	from bp	end bp
1	980766	3848839
1	4307390	5297282
1	5601778	7109452
1	7337413	14823397
1	15337049	17226430
1	18738958	19814532
1	21575268	27854846
1	29569715	34850000
1	36554475	42040896
1	44996301	52654655
1	53072982	56842251
1	61560028	64359336
1	65920480	76798548
1	78597309	82500965
1	84268726	87338424
1	89374084	98262283
1	99336894	101833528
1	101924878	115374884
1	154001459	160379503
1	160499860	167018313
1	167518558	168437253
1	171044712	182456106
1	184013256	196027023
1	201462800	203153748
1	203188269	207970984

chromosome	from bp	end bp
1	208007256	209662678
1	210570712	215782719
1	217077567	229015998
1	234520592	235465280
1	236523538	239080290
1	239590931	241657320
1	245497194	247182295
2	4354994	6274951
2	6513820	8064546
2	10966845	11778023
2	13413670	18470687
2	28920933	30788690
2	33436610	38179787
2	39641954	41872526
2	43086239	46069917
2	47343929	65992037
2	72951735	75383338
2	75403125	77159538
2	77464119	82716512
2	85731615	98381450
2	104485126	105590306
2	114226520	120318761
2	120795049	122639599
2	126671075	129934905
2	132013250	138747132

chromosome	from bp	end bp
2	140679666	144484218
2	148815277	157919513
2	158180635	163614802
2	165767975	169344116
2	173027697	182897295
2	185380884	196760605
2	203976189	206857973
2	207115516	211347678
2	211513103	217965599
2	220584891	224068092
2	225123676	233183774
2	234477979	235043935
2	239492182	241834279
3	2741939	3092389
3	5066729	5868470
3	10863601	13175685
3	21349224	23635903
3	24674823	26975377
3	30354767	31809291
3	37688039	46616527
3	50180885	55700712
3	58637736	61645214
3	64267233	66254657
3	66604926	68543295
3	69298792	71669742

chromosome	from bp	end bp
3	73269976	81115122
3	81351214	89975063
3	96060114	107971504
3	108012593	114415253
3	115520688	117179231
3	120606430	124919623
3	127817786	131537851
3	131913028	135565310
3	136785335	140559854
3	141146898	143699653
3	144692479	147777087
3	148742891	155162024
3	159435931	163533275
3	165716720	170025656
3	171322103	173699558
3	174689153	176976138
3	181717778	183004168
3	183470940	184752428
3	186735608	187576188
3	189232708	190114169
3	194878479	197577203
4	3381905	4751812
4	6910543	8577750
4	8627879	15941924
4	15967856	17240336

chromosome	from bp	end bp
4	21836156	25067476
4	29126160	36297588
4	40311314	47480269
4	53794481	74393662
4	75322683	81723396
4	83084423	87595542
4	88037879	91244108
4	91719569	95293628
4	95444927	106037934
4	106901869	112950268
4	113078808	121229278
4	121980979	130426334
4	135975570	138274835
4	138864871	140127773
4	140453732	152603890
4	152661359	157830110
4	160502560	164550453
4	165575089	169041633
4	169637036	173911780
4	177554803	179627537
4	182851076	183782143
4	184476704	185449613
4	186048446	187290661
4	188007628	189711310
5	782372	3678969

chromosome	from bp	end bp
5	4139218	4983508
5	6339593	7808721
5	10795344	12965596
5	13635597	17182658
5	17726542	20085944
5	23492450	30111211
5	34565000	36491114
5	38690273	42861340
5	52132885	56164442
5	56348120	58860630
5	58877676	63238805
5	63256089	73289446
5	74874591	77836621
5	77983434	79710061
5	87167270	98265606
5	100637034	111156746
5	113116675	115716750
5	115740730	118724004
5	118842758	121628699
5	124171838	125198837
5	126342590	129383246
5	135631964	151108352
5	151258461	153673545
5	153704710	157197976
5	159520228	163315750

chromosome	from bp	end bp
5	164073873	165655871
5	166459512	167839454
5	168775065	171620033
5	174167557	174917095
5	175573983	178581572
6	464263	1634803
6	5486899	8786957
6	10121498	12882330
6	18591616	20415968
6	22345416	25437423
6	33888881	36320529
6	37353178	41678496
6	42139310	52335407
6	52335792	56338754
6	63317904	70347942
6	72739790	75826415
6	76615768	88635779
6	90097520	91133854
6	94472195	97326389
6	98853201	102579545
6	104420312	107098734
6	109835323	119885870
6	119900679	125329412
6	136241646	139325007
6	143543289	146651453

chromosome	from bp	end bp
6	146652103	150316302
6	154991384	156270483
6	163078179	164552751
6	166434021	170201271
7	276241	3158477
7	4108407	4889073
7	5545827	8682385
7	8977834	11507160
7	13636643	14255707
7	19065106	20590118
7	21862380	28926813
7	30498399	35135071
7	35576963	36999831
7	37325014	42803978
7	43521321	50352493
7	51082275	56856495
7	64757785	71190068
7	77200456	91204599
7	96119359	101400208
7	101775741	105890995
7	106015507	106987903
7	107820563	125505836
7	128128551	131021972
7	133216400	137870641
7	138063299	149730180

chromosome	from bp	end bp
7	150117360	152548816
7	153301941	154426471
7	154480665	155382201
7	155881185	158196666
8	1003821	2825876
8	3930657	4540540
8	5710411	6877710
8	14045345	17048632
8	18085468	19121929
8	21148227	22690278
8	23126860	25236092
8	29143966	32262348
8	37135057	42119178
8	49591653	54600561
8	56890593	66203669
8	67937087	69528548
8	71377594	74007275
8	75156886	88586993
8	88588732	94063861
8	95191517	97303207
8	98175068	101475583
8	101656734	105498453
8	115643056	118184874
8	118788706	125698826
8	126161282	130505923

chromosome	from bp	end bp
8	131716025	133219382
8	133910541	138110077
8	138243248	139033993
8	139441708	142081232
8	142090827	143472774
9	857631	1792661
9	3829325	4866922
9	4945551	7241273
9	7791716	9517160
9	13433615	14254708
9	19169683	27351579
9	28519572	32737617
9	33101694	38642943
9	65439440	72275978
9	74929147	76930776
9	79161437	81844015
9	86556435	89682601
9	90180287	92175406
9	97526897	106913549
9	107224734	107995897
9	109179057	112129308
9	113305763	114151508
9	116104469	117362492
9	122197643	129531774
9	129543479	134256649

chromosome	from bp	end bp
9	135433556	137553043
10	1121860	3035117
10	3826763	4670914
10	5318785	6367631
10	9390896	12438785
10	12598940	14302465
10	19048647	20987000
10	21028025	24531923
10	29671847	33913184
10	55370790	62068010
10	62655053	71348219
10	77600362	91572161
10	96796968	104519297
10	106022978	112215159
10	113568838	114566573
10	117662717	119459080
10	121331084	124182491
10	126326533	127539605
10	128523430	129254707
10	131248183	133066372
11	1849997	7599536
11	7616849	10350732
11	15445157	18668851
11	19236261	25088304
11	26024106	35480841

chromosome	from bp	end bp
11	36261046	38009035
11	43988997	45899587
11	57416325	68229472
11	68396503	71400375
11	73251318	75843353
11	77125670	80418464
11	82560884	88886121
11	94033320	101477914
11	105623239	111437296
11	116173619	117231648
11	117949596	119870590
11	121393923	122747227
11	123157509	124124066
11	126164383	127446127
11	128032650	128894445
11	129583039	130731190
12	554522	2311603
12	3065991	3946768
12	4160407	5866185
12	6321415	8020707
12	10327202	12578396
12	15668362	20422599
12	21164784	23770314
12	25044072	32926694
12	39988505	47876659

chromosome	from bp	end bp
12	48575125	53711211
12	54290141	62850641
12	64056278	67426150
12	69131401	77022645
12	80037603	83141903
12	87413438	91910559
12	92522677	94634752
12	97324182	105832052
12	105970636	108117743
12	110884264	114748121
12	116599660	120607068
12	121051129	123448631
12	127161782	127659744
12	128526348	129108974
12	129572639	131047681
13	20508045	21980114
13	22313839	23514843
13	24516452	26934054
13	28844673	30051110
13	31120538	32292408
13	39508686	42370949
13	44904236	48154647
13	48159489	51287394
13	57467408	66471657
13	70020364	72257566

chromosome	from bp	end bp
13	73777320	75262869
13	75862857	78247336
13	87710112	92394587
13	92634338	96869321
13	97661318	99884540
13	100726369	101833077
13	106675806	107555369
13	107679810	109125112
13	110798062	113145940
14	20629261	26690748
14	27248662	30076077
14	32018095	33067352
14	33636875	34221270
14	34346599	38525724
14	42506436	45400050
14	46389692	50950968
14	51867041	56126962
14	62153534	65204343
14	67458740	76547328
14	81026571	84322043
14	86201934	89116820
14	89463267	91193330
14	94625609	95283925
14	95798968	96708229
14	98673677	99519872

chromosome	from bp	end bp
14	101253868	104631584
15	23027932	23799147
15	24923575	26370073
15	26889847	27355487
15	28632697	30206161
15	33640942	36394451
15	38181286	49716262
15	51222134	57486813
15	60722641	61619260
15	62641305	66240302
15	68415013	71124744
15	80927839	83946039
15	84861571	87489387
15	88201211	88842276
15	92449686	93211583
15	94034631	95120402
15	95905325	96548274
15	96937613	97552171
15	98563221	101546311
16	837148	3733907
16	3996706	5703056
16	7038195	7782100
16	7856245	8746002
16	9380962	11372623
16	12697704	14857594

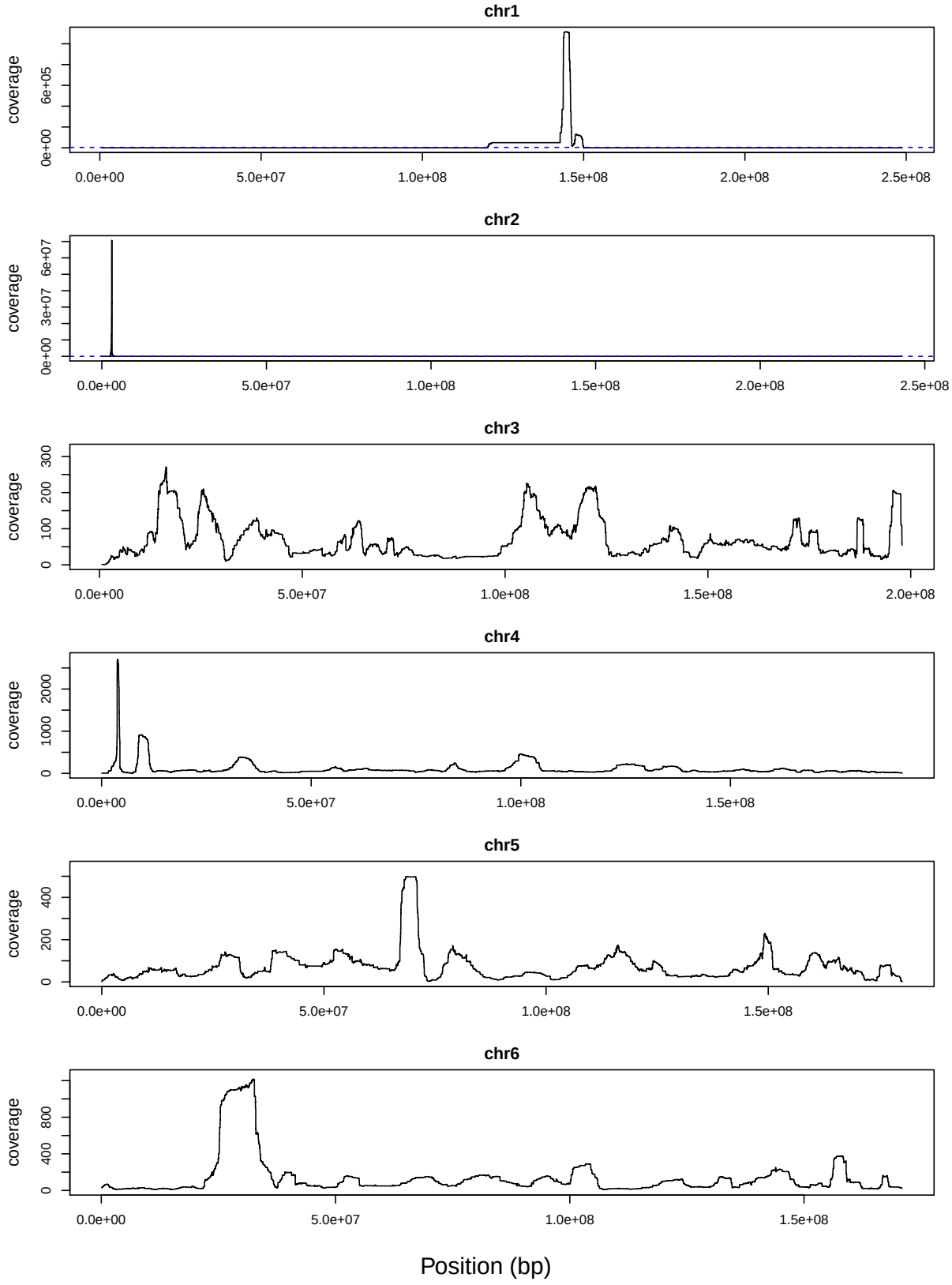
chromosome	from bp	end bp
16	15643055	20300634
16	20407500	23720178
16	24084582	25777817
16	27113122	47950340
16	48700469	51415739
16	54367289	55476839
16	58454215	62741889
16	64495430	74985277
16	77339827	79097091
16	80386226	81603070
16	82662726	85632769
16	85956736	89258160
17	1788926	3254093
17	4502938	5928624
17	6163697	7818016
17	8699094	10816231
17	11597937	13025154
17	16416586	35755646
17	36661033	45383260
17	47231644	50564088
17	50756448	53585886
17	54896270	57074100
17	57511208	59413823
17	61235921	65571694
17	66520002	70260952

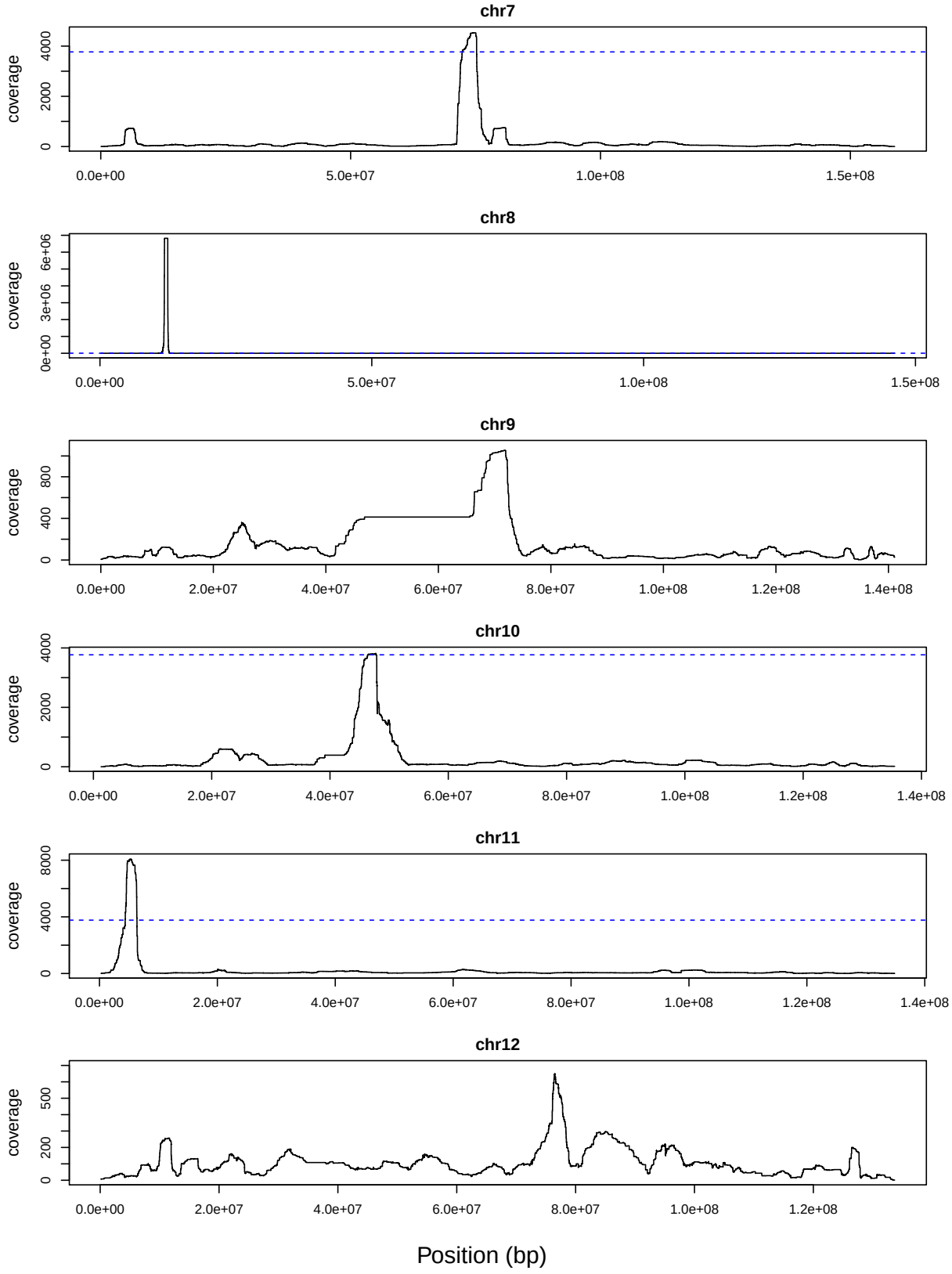
chromosome	from bp	end bp
17	72501940	73255152
17	73358239	77848946
17	79263054	80045803
17	80858441	82074407
18	907032	2874364
18	8232554	8786178
18	10337421	11312876
18	12354835	23147705
18	24044200	25364359
18	26515135	29730097
18	46360267	48670422
18	52264860	55812313
18	56579263	58066015
18	60048948	61921805
18	63013512	68611377
18	69904324	71665706
18	71670423	73315914
18	75854106	76765266
19	1345521	3217001
19	4070853	5852730
19	6322328	6828610
19	6833288	8279728
19	8662963	43454467
19	43454591	47823343
19	47993893	48911322

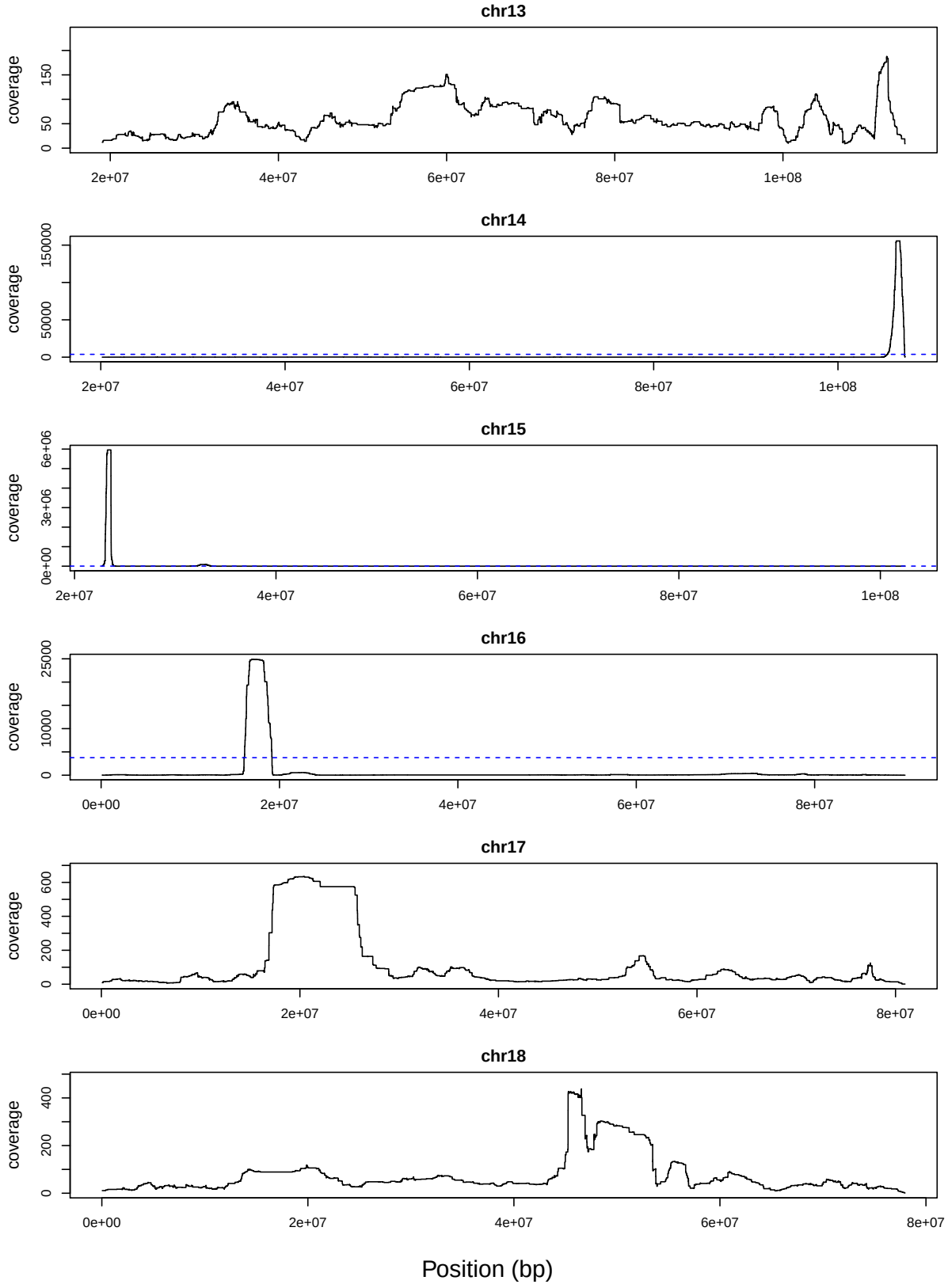
chromosome	from bp	end bp
19	48914742	50215626
19	50602744	51372732
19	51462416	52839033
19	53626867	56549892
20	911854	1894938
20	1936774	3172432
20	4895700	7248329
20	8575959	10474100
20	11074635	12640904
20	17557179	18918496
20	19041963	20971447
20	21110213	43559834
20	44505861	47112828
20	48738672	51714194
20	57717404	58478170
20	59734236	60788241
20	60882800	63950322
21	13828609	15590899
21	16041871	17454882
21	17760167	18957240
21	19683215	29839757
21	30381018	33854439
21	34907710	36223862
21	37881912	39656421
21	39973349	40812559

chromosome	from bp	end bp
21	41963435	45426336
22	16785174	19645297
22	22005208	22938494
22	23471191	30562093
22	34743839	36524995
22	36630895	44529368
22	44594911	45261648
22	46116682	48149375
22	48589014	49265922

Table A.2: Regions included in from mutation rate analysis in Barbados Asthma Genetic Study (build GRCh38)


chr1
coverage
0e+00
6e+05
0.0e+00
5.0e+07
1.0e+08
1.5e+08
2.0e+08
2.5e+08
chr2
coverage
0e+00
3e+07
6e+07
0.0e+00
5.0e+07
1.0e+08
1.5e+08
2.0e+08
2.5e+08
chr3
coverage
0
100
200
300
0.0e+00
5.0e+07
1.0e+08
1.5e+08
2.0e+08
chr4
coverage
0
1000
2000
0.0e+00
5.0e+07
1.0e+08
1.5e+08
chr5
coverage
0
200
400
0.0e+00
5.0e+07
1.0e+08
1.5e+08
chr6
coverage
0
400
800
0.0e+00
5.0e+07
1.0e+08
1.5e+08
Position (bp)


chr7
coverage
0
2000
4000
0.0e+00
5.0e+07
1.0e+08
1.5e+08
chr8
coverage
0e+00
3e+06
6e+06
0.0e+00
5.0e+07
1.0e+08
1.5e+08
chr9
coverage
0
400
800
0.0e+00
2.0e+07
4.0e+07
6.0e+07
8.0e+07
1.0e+08
1.2e+08
1.4e+08
chr10
coverage
0
2000
4000
0.0e+00
2.0e+07
4.0e+07
6.0e+07
8.0e+07
1.0e+08
1.2e+08
1.4e+08
chr11
coverage
0
4000
8000
0.0e+00
2.0e+07
4.0e+07
6.0e+07
8.0e+07
1.0e+08
1.2e+08
1.4e+08
chr12
coverage
0
200
500
0.0e+00
2.0e+07
4.0e+07
6.0e+07
8.0e+07
1.0e+08
1.2e+08
Position (bp)


chr13
coverage
0
50
150
2e+07
4e+07
6e+07
8e+07
1e+08
chr14
coverage
0
50000
150000
2e+07
4e+07
6e+07
8e+07
1e+08
chr15
coverage
0e+00
3e+06
6e+06
2e+07
4e+07
6e+07
8e+07
1e+08
chr16
coverage
0
10000
25000
0e+00
2e+07
4e+07
6e+07
8e+07
chr17
coverage
0
200
400
600
0e+00
2e+07
4e+07
6e+07
8e+07
chr18
coverage
0
200
400
0e+00
2e+07
4e+07
6e+07
8e+07
Position (bp)

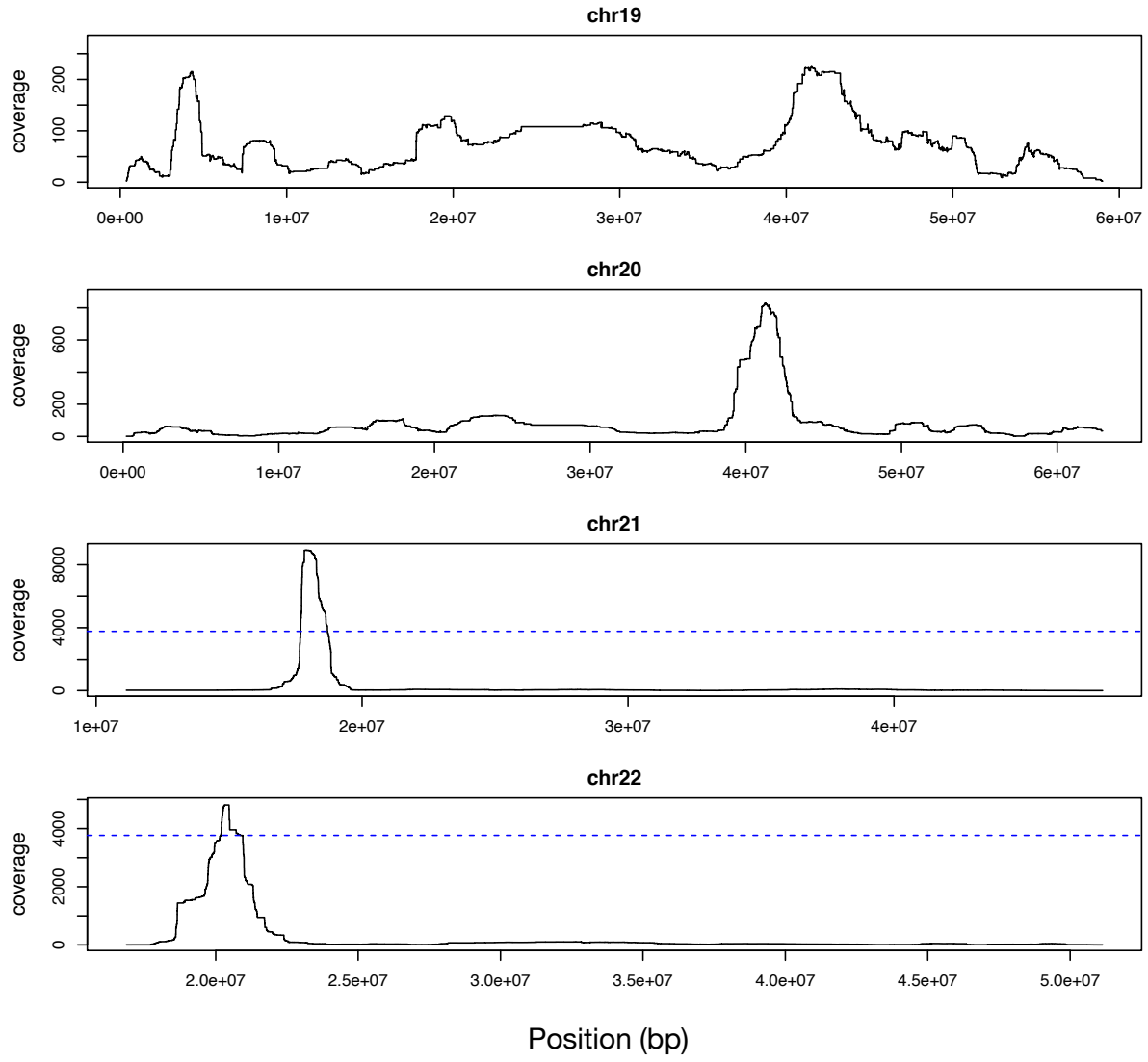


Figure A.3: Three-way IBD coverage along the genome in Jackson Heart Study. The y-axis indicates number of three-way IBD along the genome in windows of 500 base pairs. The blue dashed line represent the 98th percentile of genome wide three-way IBD coverage. Regions with three-way IBD coverage exceeded the 98th percentile are removed.

chromosome	from bp	end bp
1	871953	119503823
1	151846526	247983451
2	4306813	8097173
2	8653332	87734662
2	98978279	107782091
2	114065347	242793379
3	1092332	2053010
3	2158097	30567270
3	30913825	50108269
3	52032080	197305342
4	747288	7203561
4	7295484	190634032
5	721373	73556992
5	74223390	179565541
6	294621	170516802
7	482031	71239379
7	76376085	158171047
8	516419	2954843
8	2971087	10253922
8	13395872	144995423
9	445168	38642940
9	38678786	140445536
10	2099324	44957169
10	48241138	117830342
10	117953796	131729046
10	131840070	135054207

chromosome	from bp	end bp
11	633743	3681828
11	6564852	134659682
12	304900	128019941
12	128047582	132895713
13	19582308	113998593
14	20549740	104631765
15	25061735	30166305
15	33878310	54331540
15	54336200	97078724
15	97516190	102102761
16	590589	13891656
16	19984372	88860869
17	641897	80047905
18	350695	77360212
19	562331	58539252
20	348082	5734002
20	5866210	7080390
20	7487258	56944213
20	57378136	62598183
21	15194729	17311337
21	19089712	33285270
21	33537448	45136668
21	45710051	47013230
22	17361790	19687252
22	20991256	50611362

Table A.3: Regions included in from mutation rate analysis in Jackson Heart Study (build GRCh37)

Appendix B

SOFTWARE RESOURCES

Beagle 4.0: https://faculty.washington.edu/browning/beagle/b4_0.html

Beagle 4.1: https://faculty.washington.edu/browning/beagle/b4_1.html

Beagle 5.1: https://faculty.washington.edu/browning/beagle/beagle.html

Refined IBD: http://faculty.washington.edu/browning/refined-ibd.html

hap-IBD: https://github.com/browning-lab/hap-ibd

IBDNe: http://faculty.washington.edu/browning/ibdne.html

METIBD3: https://github.com/tianxiaowen/mutation_phased