# Scalable Methods for the Inference of Identity by Descent

Fiona Grimson

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

Doctor of Philosophy

University of Washington

2017

Reading Committee:

Elizabeth Thompson, Chair

Sharon Browning

Marina Meila-Predoviciu

Program Authorized to Offer Degree:
Statistics

University of Washington

**Abstract**

Scalable Methods for the Inference of Identity by Descent

Fiona Grimson

Chair of the Supervisory Committee:
Professor Elizabeth Thompson
Statistics

Identity by descent (IBD) describes the shared inheritance of DNA and underlies genetic similarity between individuals. Estimated IBD graphs describing the IBD relationships among individuals have many uses in statistical genetics. An important application is to detect, through linkage analysis, the location of genes that cause genetic diseases. Accurate estimation of IBD graphs among large groups of individuals is essential.

IBD is typically estimated either among individuals in small family pedigrees or among distantly related individuals sampled from a population with an unknown pedigree relationship. Both pedigree and population approaches require different modeling assumptions and are applied to different study designs. In this thesis, scalable methods of estimating IBD that combine both pedigree and population estimation methods are developed for family based studies. IBD is estimated between the founders of the family pedigrees to incorporate more information about shared inheritance without additional data collection.

A combined IBD model is developed for sib-pair studies, and is demonstrated on siblings from a 50 generation simulated population. A merging algorithm is also developed to combine pedigree and population IBD estimates for family based studies with larger component pedigrees. The merging method is demonstated on pedigrees from the simulated population and on an Alzheimer's disease family study. Combined IBD models are shown to increase power and resolution for locating genes that cause genetic diseases.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# Chapter 1

# INTRODUCTION

One of the goals of statistical genetics is to determine the location of genes that contribute to the expression of genetically influenced phenoytpes. Analyses rely on the shared coancestry of individuals. DNA is passed from parent to child in large segments resulting in close relatives having more DNA in common and thus being more phenotypically similar. If two individuals share DNA inherited from a common ancestor on a segment of their genome they are considered identical-by-descent (IBD) at that location. The focus of this thesis is the development of scalable methods for estimation of IBD, combining estimation techniques to improve gene detection using human genetic data.

Observed genetic data for pedigree-based genetic studies is described in Section 1.1. Identity-by-descent, and a generative model for identity-by-descent as a function of inheritance vectors is in Section 1.2. A generative model for observed genotypes conditional on IBD is Section 1.3. Techniques for the calculation of probabilities on pedigree graphs that use the conditional dependencies described in these models are discussed in Section 1.4. Methods for IBD inference are discussed in Section 1.5. The primary application of IBD in this thesis is for the detection of the location of disease-causing genetic variants. Linkage analysis methods use IBD inferred from pedigrees and are described in Section 1.6; IBD mapping methods use IBD inferred from distant relatives and are described in Section 1.7; Association tests are described in Section 1.8 and do not use IBD inference. The contributions of this thesis to the literature are described in Section 1.9.

### *1.1 Observed Genetic Data*

Humans are diploid with 23 pairs of chromosomes, 22 homologous pairs of autosomes and the $XY$ sex chromosomes. Each haploid genome consists of over 3 billion base pairs of DNA. Genetic variation among individuals is studied by comparing genetic markers such as single-nucleotide-polymorphisms (SNPs) at specific base pair locations (loci) along the genome. While the majority of the genome is identical in all humans, SNPs are single base pairs of DNA that may take different nucleotide types among individuals. Most common SNPs have two possible nucleotide types that are the alleles of the SNP, denoted $A$ and $B$, and are two of the DNA bases $A, T, C, G$. The alleles occur in a population with frequencies $p_l$ and $q_l = 1 - p_l$ at loci $l = 1, ..., L$. The major allele is the more common type and the minor allele the less common. The cost of obtaining SNP data has rapidly decreased over the years. The International HapMap project [The International HapMap Consortium, 2007] and 1000 Genomes project [The 1000 Genomes Project Consortium, 2015] have mapped the location of over 84 million SNPs from thousands of genome sequences across global populations. The allele frequencies of SNPs vary by population as a result of population genetic processes such as migration, natural selection, non-random mating and random genetic drift.

Genetic information is passed from parent to child through a random process called meiosis. In meiosis the homologous copies of each chromosome duplicate, cross over, recombine, and separate into four haploid gametes. Under the law of independent assortment each meiosis is independent and each pair of homologous chromosomes duplicate, cross over, and recombine independently of other pairs [Merriam, 2001]. A child has two copies of DNA, comprised of DNA from two gametes, one inherited from each parent. The allelic type of the SNPs on one DNA copy of an individual make up a haplotype. The two haplotypes of an individual together make up the genotype. In the genotype, the parental origin of each allele is not known.

In pedigree-based studies, a pedigree structure is observed on a set of individuals $\mathcal{F}$ and describes the family tree of the individuals. In this thesis, individuals will be indexed by

$i = 1, ..., N$. The two ordered DNA copies (paternal then maternal) of individual $i$ are $(i_1, i_2)$. All ordered DNA copies will be indexed by $j = 1, ..., 2N$. The family pedigree structure may have disjoint components that represent several small families and is typically only 1-5 generations deep. The family individuals $\mathcal{F}$ are part of a larger population $\mathcal{P}$ that spans many generations for which the pedigree structure is unknown. Multilocus marker genotypes are observed on a set of individuals $\mathcal{G} \subset \mathcal{F} \subset \mathcal{P}$ at loci $l = 1, ..., L$ with known population allele frequencies. Genotypes are represented by $\mathbf{g}$ where $\mathbf{g}_{il} \in \{AA, AB, BB\}$ is the genotype for individual $i$ at locus $l$. The unobserved haplotypes are latent variables denoted $\mathbf{h}$ where $\mathbf{h}_{jl} \in \{A, B\}$ is the allelic type of DNA copy $j$ at locus $l$. The two haplotypes belonging to the two DNA copies of an individual are $\mathbf{h}_{il} = (\mathbf{h}_{i_1 l}, \mathbf{h}_{i_2 l}) \in \{(A, A), (A, B), (B, A), (B, B)\}$. Genotypes are a function of the haplotypes, $\psi(\mathbf{h}_{il}) = \mathbf{g}_{il}$, where $\psi((A, A)) = AA$, $\psi((A, B)) = \psi((B, A)) = AB$ and $\psi((B, B)) = BB$. We also observe trait data on individuals $\mathcal{Y} \subset \mathcal{F}$ where $\mathcal{Y}$ and $\mathcal{G}$ may overlap or not.

### *1.2 Identity by Descent*

Two or more homologous copies of DNA at a locus are identical by descent (IBD) relative to a past population if they are descended from the same ancestral copy in that population. The IBD state among a set of individuals at a locus is the partition of the constituent DNA copies into classes where DNA copies in the same class are IBD. The IBD states over all loci for DNA copies in $\mathcal{G}$ relative to population founders are denoted $\mathbf{s}^{\mathcal{P}}$ where $\mathbf{s}_l^{\mathcal{P}}$ is the IBD state at locus $l$, and the IBD states relative to family founders are $\mathbf{s}^{\mathcal{F}}$.

The IBD state at locus $l$, $\mathbf{s}_l^{\mathcal{P}}$ (or $\mathbf{s}_l^{\mathcal{F}}$) is a partition over ordered DNA copies $j \in \mathcal{G}$. As marker data are observed as genotypes on individuals rather than haplotypes on chromosomes, we are also interested in genotypically equivalent IBD states. The genotypically equivalent IBD states $\tilde{\mathbf{s}}_l^{\mathcal{P}}$ (or $\tilde{\mathbf{s}}_l^{\mathcal{F}}$) are formed by collapsing over IBD states such that

$$P(\tilde{\mathbf{S}}_l^{\mathcal{P}} = \tilde{\mathbf{s}}_l^{\mathcal{P}}) = P(\mathbf{S}_l^{\mathcal{P}} \in \mathcal{S}) \tag{1.1}$$

where $\mathcal{S}$ is the set of $\mathbf{s}_l^{\mathcal{P}}$ that are equivalent under different orderings of the two DNA copies

of each individual [Thompson, 1974].

The IBD state at a locus $\mathbf{s}_l$ can be represented in many different ways. Labeling methods and an R package to change between them are described in more detail in Appendix A. In this thesis two labeling methods are used. For two individuals, or four DNA copies, there are 15 possible IBD states at each locus. These are referred to by numerical Jacquard labels [Jacquard, 1970], and are listed in Table A.1. For larger sets of DNA copies one of the most useful representations is as a partition of the copies into classes, where co-membership of a class indicates IBD. In this thesis the IBD partition on an arbitrary number of DNA copies is represented by a vector of founder genome labels (FGLs). The FGL representation and its basis in inheritance is described here.

### *1.2.1 FGL Model*

When the pedigree structure is known, the realized descent of DNA at each locus through a pedigree can be represented by inheritance vectors (IVs). IVs are latent variables describing inheritance through meioses that are parental to the DNA copy. The IV is denoted $\mathbf{V} = \mathbf{v}$ where $\mathbf{v}_{jl} \in \{0, 1\}$ is the inheritance indicator for DNA copy $j$ at locus $l$. If the DNA is from is the individual's paternal (maternal) genome, and the DNA at the locus was inherited from the father's (mother's) paternal copy, it takes the value 1; if the DNA was inherited from the father's (mother's) maternal copy it takes the value 0. Under the assumptions of Mendelian segregation, the probabilities of maternal or paternal inheritance are equal, and the IV state is independent over the ordered DNA copies. However, IVs are not independent over loci.

IVs for a DNA copies differ between adjacent loci when there is a recombination between the loci. Under the assumption of no genetic interference, IVs have Markov dependence over loci [Lander and Green, 1987]. The probability of a recombination occurring between loci

$l-1$ and $l$ is independent of the interval between $l$ and $l+1$. That is,

$$P(\mathbf{V}=\mathbf{v}) = \prod_j P(\mathbf{V}_{j.} = \mathbf{v}_{j.}) \tag{1.2}$$

$$= \prod_j \left[ P(\mathbf{V}_{j,1} = \mathbf{v}_{j,1}) \prod_{l=2}^{L} P(\mathbf{V}_{j,l} = \mathbf{v}_{j,l} | \mathbf{v}_{j,l-1}) \right]. \tag{1.3}$$

Equation (1.2) shows independence over the DNA copies $j = 1, ..., 2N$, Equation (1.3) shows Markov dependence over loci. The prior probabilities of inheritance are $P(\mathbf{V}_{j1} = 0) = P(\mathbf{V}_{j1} = 1) = 0.5$.

A recombination occurs between two loci on a chromosome if the DNA at the two loci derives from two different parental chromsomes. This happens if there are an odd number of crossovers between the two loci during meiosis. A recombination is reflected in an IV transition between loci $l-1$ and $l$, separated by $d_l$ centimorgans (cM) is the recombination fraction $\theta_l$ given by Haldane's map function [Haldane, 1919],

$$P(\mathbf{V}_{j,l} = \mathbf{v}_{j,l} \neq \mathbf{v}_{j,l-1}) = 0.5(1 - e^{-0.02 d_l}) = \theta_l. \tag{1.4}$$

For small $d_l$ this is approximated by $\theta_l \approx 0.01 d_l$. One centimorgan (cM) is approximately 1 million base pairs (Mbp) in humans although this varies along the genome and by gender [Kong et al., 2002]. The cM position of markers at given bp positions is the genetic map, the most commonly used genetic map is the Rutgers map [Matise et al., 2007].

The inheritance vectors on a pedigree can be used to determine founder genome labels (FGLs) which describe IBD states. The FGLs are denoted $\mathbf{x}^{\mathcal{F}}$ where $\mathbf{x}^{\mathcal{F}}_{jl}$ is the label of the haploid founder genome of $\mathcal{F}$ that supplied the original copy of DNA inherited by individual $j$ at locus $l$. Given a labelling f the genomes of the pedigree founders, the FGLs are a deterministic function of IVs, $\phi(\mathbf{v}) = \mathbf{x}^{\mathcal{F}}$, and thus the labels refer to specific founder genomes of $\mathcal{F}$. The FGL of a DNA copy at a locus depends only in its inheritance indicator $\mathbf{v}_{il}$ and the FGLs in the parent that produced the gamete. If the DNA is from an individual's paternal gamete, the inheritance indicator indicates whether the DNA was inherited from the father's paternal gamete $P(j)$ or father's maternal gamete $M(j)$. The function $\phi$ is therefore defined

recursively,

$$\mathbf{x}_{jl} = \phi(\mathbf{v}_{.l}) = \begin{cases} \mathbf{x}_{P(j)l} & \text{if } \mathbf{v}_{jl} = 0 \\ \mathbf{x}_{M(j)l} & \text{if } \mathbf{v}_{jl} = 1. \end{cases} \tag{1.5}$$

The inheritance vectors point to the parental FGL, independently of the value of the parental FGLs or the process that generated them. There is not a 1-1 correspondence between the IV state and FGL state in general, as several paths of inheritance can lead to the same FGL. The IV and FGL contain the same information if the pedigree structure describes siblings with founder parents, discussed further in Chapter 3. FGL state transitions are in general not Markov and the length of a shared FGL segments depends on the number of meioses to the common ancestor. For example, half-second-cousins will cease to be IBD (share the same FGL) if there is a recombination in any one of 6 gamete chromosomes. If the cousins are not IBD, the probability of returning to IBD depends on how many recombinations away from IBD they are.

The function $\sigma(\mathbf{x}^{\mathcal{F}}) = \mathbf{s}^{\mathcal{F}}$ maps the FGLs to the IBD state. The mapping $\sigma$ is not 1-1 as many FGLs can give the same IBD state. The FGL refers to a specific founder genome of $\mathcal{F}$ and for IBD it only matters that the DNA came from the same founder genome, not which founder genome it was.

FGLs can also be used to represent IBD relative to population founders, $\mathbf{s}^{\mathcal{P}}$, when the population pedigree structure is unknown. In this case IVs are not used as a latent variable. The FGLs are arbitrary, in the sense that if two DNA copies share an FGL we know they are IBD relative to a founder genome of $\mathcal{P}$, but we do not know which founder genome has that label. The FGLs at each locus can be re-labeled in the canonincal form without any loss of information. If $\mathbf{x}_{.l}^{\mathcal{P}} = (x_1, ..., x_{2N})$ are the FGLs at locus $l$ for all DNA copies, the canonical labeling satisfies conditions 1.6 and 1.7:

$$x_1 = 1 \tag{1.6}$$

$$x_{j+1} \leq \max_{j'=1,...,j} (x_{j'}) + 1, \; \forall j = 1, ..., 2N - 1. \tag{1.7}$$

The function $\mathbf{s}^{\mathcal{P}} = \sigma(\mathbf{x}^{\mathcal{P}})$ that maps the canonical FGLs to IBD states is 1-1 as a shared FGL indicates IBD only.

To summarize, the probability of FGLs $\mathbf{x}^{\mathcal{F}}$ of $\mathcal{F}$ is

$$P(\mathbf{X}^{\mathcal{F}} = \mathbf{x}^{\mathcal{F}}) = \sum_{\mathcal{V}} P(\mathbf{V} \in \mathcal{V}) \tag{1.8}$$

where $\mathcal{V} = \{\mathbf{v} : \phi(\mathbf{v}_j) = \mathbf{x}_j^{\mathcal{F}} \forall j\}$ and the probabilities $P(\mathbf{V} = \mathbf{v})$ that make up $P(\mathbf{V} \in \mathcal{V})$ are given in Equation (1.3). The FGLs state $\mathbf{x}^{\mathcal{F}}$ is a representation of the IBD state $\mathbf{s}^{\mathcal{F}}$.

### *1.2.2 IBD Segments*

Under Haldane's model [Haldane, 1919] we can model recombination points as occurring along the chromosome as a Poisson process with rate 1 per Morgan [Boehnke, 1994]. If IBD between a pair of chromosomes is defined relative to the most recent common ancestor (MRCA) $g$ generations ago, then any recombination in the $2g$ meioses separating the pair will end the IBD segment. If we consider the lengths of all IBD segments between such chromosomes, the lengths are distributed exponentially with rate $2g$ per Morgan, ignoring truncation at the end of a chromosome. After 25 generations the mean IBD segment length is 2cM with sd 1 and after 50 generations it is 1cM with sd 2. We can also consider the distribution of lengths for IBD segments that cover a specific position on the chromosome. The distance from the specified position to the endpoints of the segment are exponentially distributed with rate $2g$, so the total length of the segment is distributed Erlang with rate $2g$ per Morgan and shape 2 Boehnke [1994], Palamara et al. [2012]. The mean length of an IBD segment relative to an ancestor DNA copy 25 generations ago is 4cM with sd $\sqrt{2}/50$ and relative to 50 generations is 2cM with sd $\sqrt{2}/100$.

In this thesis, IBD is defined relative to the founder genomes of the population. The MRCA is typically more recent than the founders of the population. A segment of DNA that is identical relative to the MRCA may be a composite of segments inherited from founders. A segment of IBD in current genomes relative to population founder genomes may be a composite of founder genomes and have changes in FGL while remaining IBD.

### *1.3 Conditional Independence of Genotypes*

As described in Section 1.1, we observe multilocus genotypes $\mathbf{g}$ for a set of individuals $\mathcal{G}$ at certain positions (loci) along a chromosome. The individuals are assumed to be members of a family $\mathcal{G} \subset \mathcal{F}$ with an observed pedigree structure that is part of a larger population $\mathcal{P}$. The probability of observing $\mathbf{g}$ can be conditioned on the IBD state or FGLs of the individuals,

$$P(\mathbf{G} = \mathbf{g}) = \sum_{\mathbf{x} \in \mathcal{X}^{\mathcal{F}}} P(\mathbf{G} = \mathbf{g} | \mathbf{x}^{\mathcal{F}}) P(\mathbf{X}^{\mathcal{F}} = \mathbf{x}^{\mathcal{F}}) \tag{1.9}$$

where $\mathcal{X}^{\mathcal{F}}$ is the set of possible unique founder genome labelings using the founders of $\mathcal{F}$. The generative model in Equation (1.9) for $\mathbf{g}$ has two parts: a model for FGLs, and a model for genotypes given FGLs. The model for FGLs, $P(\mathbf{X}^{\mathcal{F}} = \mathbf{x}^{\mathcal{F}})$, is described in Section 1.2, Equation (1.8). The model for genotypes given FGLs, $P(\mathbf{G} = \mathbf{g} | \mathbf{x}^{\mathcal{F}})$ is described in this section.

The FGLs specify which founder chromosome provided the DNA at each locus for each chromosome. The allelic types that make up the haplotype of the chromosome are therefore a function of the FGLs, $\mathbf{h}_{jl} = \alpha(\mathbf{x}_{jl})$ where $\alpha()$ gives the allelic type of the founder $\mathbf{x}_{jl}$. If DNA copies are IBD at a locus they will have the same FGL and, in the absence of mutation, will share an allelic type. Sharing an allelic type is referred to as being identical by state (IBS).

Linkage disequilibrium (LD) is correlation in allelic type between loci on the same chromosome, giving elevated frequencies of certain haplotypes in a population compared to what would be expected for independent loci. LD results broadly from two causes [Thompson, 2013]. The first is a new variant arising on local haplotypic background creating a strong association between the variant and the alleles of that background. The LD can be broken down by recombination, but for tightly linked loci this can take thousands of generations. The other cause of LD is genetic drift, leading to different haplotype frequencies in different subpopulations. There can be association in the population as a whole, even if there is no LD in the subpopulations. This type of LD is still a reflection of coancestry as individuals in

the same subpopulation are more closely related. LD is not IBD but it is a reflection of IBD and is therefore a confounding variable in IBD estimation from IBS and haplotype sharing.

We do not model LD, so when assigning allelic types to FGLs we assume that each founder has allele $A$ or $B$ at locus $l$ with probabilities $p_l$ and $q_l$ respectively, independently over loci. If $k$ is a founder label, $P(\alpha(k) = A) = p_l$ and $P(\alpha(k) = B) = q_l$. For the non-founder individuals, define $\mathbf{h}_{il} = (\mathbf{h}_{i_1 l}, \mathbf{h}_{i_2 l}) = (\alpha(\mathbf{x}_{i_1 l}), \alpha(\mathbf{x}_{i_2 l}))$ to be the allelic types of the ordered pair of DNA copies $(i_1, i_2)$ for individual $i$. Let $K_l$ be the unique FGLs present at a given locus, then $n_l = \sum_{K_l} \mathbb{1}(\alpha(K_l) = A)$ is the number of FGLs that were assigned an $A$ allele and $m_l = \sum_{K_l} \mathbb{1}(\alpha(K_l) = B)$ is the number of FGLs that were assigned a $B$ allele. Then,

$$P(\mathbf{H} = \mathbf{h}|\mathbf{x}) = \prod_l P(\mathbf{H}_{\cdot l} = \mathbf{h}_{\cdot l}|\mathbf{x}_{\cdot l}) = P(\alpha(\mathbf{x}_{\cdot l})) = \prod_l p_l^{n_l} q_l^{m_l}. \tag{1.10}$$

Genotype data on individuals $i$ is formed by $\psi(\mathbf{h}_{il}) = \psi((\mathbf{h}_{i_1 l}, \mathbf{h}_{i_2 l})) = \mathbf{g}_{il}$, a function that collapses genotypically equivalent allelic types of the two ordered DNA copies of the individual. Genotypically equivalent ordered allelic types are defined by interchanging the two DNA copies of an individual [Thompson1974]. Specifically,

$$P(\mathbf{G}_{il} = AA|\mathbf{x}_{il}) = P(\mathbf{H}_{i_1 l} = A \cap \mathbf{H}_{i_2 l} = A|\mathbf{x}_{il}) = P(\mathbf{H}_{i_1 l} = A|\mathbf{x}_{il}) P(\mathbf{H}_{i_2 l} = A|\mathbf{x}_{il}), \tag{1.11}$$

$$\begin{aligned} P(\mathbf{G}_{il} = AB|\mathbf{x}_{il}) &= P(\mathbf{H}_{i_1 l} = A|\mathbf{x}_{il}) P(\mathbf{H}_{i_2 l} = B|\mathbf{x}_{il}) \\ &\quad + P(\mathbf{H}_{i_1 l} = B|\mathbf{x}_{il}) P(\mathbf{H}_{i_2 l} = A|\mathbf{x}_{il}), \end{aligned} \tag{1.12}$$

$$P(\mathbf{G}_{il} = BB|\mathbf{x}_{il}) = P(\mathbf{H}_{i_1 l} = B|\mathbf{x}_{il}) P(\mathbf{H}_{i_2 l} = B|\mathbf{x}_{il}). \tag{1.13}$$

To summarize,

$$P(\mathbf{G} = \mathbf{g}|\mathbf{x}) = \prod_l \left[ \sum_{\mathcal{H}} P(\mathbf{H}_{\cdot l} \in \mathcal{H}|\mathbf{x}_{\cdot l}) \right]. \tag{1.14}$$

where $\mathcal{H} = \{\mathbf{h}_{\cdot l} : \psi(\mathbf{h}_{il}) = \mathbf{g}_{il},\ \forall i\}$. The probabilities $P(\mathbf{H}_{\cdot l} \in \mathcal{H}|\mathbf{x}_{\cdot l})$ can be calculated with Equation (1.10).

Observing IBS does not guarantee IBD, as the alleles may have descended from different founders that happen to have the same allelic type. Differences in allelic type, in addition to environmental and other factors, cause differences in phenotypes among individuals. Thus,

IBD underlies the genetic and phenotypic similarities between individuals and is the framework that connects evolutionary and coalescent theory with the analysis of genetic marker and trait data observed on individuals [Thompson, 2013].

## *1.4 Calculation of Probabilities on Graphs*

This section describes methods of calculating probabilities on pedigree graphs, which will be used in the calculation of the probability of observed data on a pedigree. As in Section 1.1 let $\mathcal{F}$ be individuals for whom a pedigree structure is observed, with multilocus marker genotype data observed on individuals $\mathcal{G} \subset \mathcal{F}$. In pedigree calculations, the unobserved population pedigree $\mathcal{P}$ that encompasses $\mathcal{F}$ is ignored. Sections 1.2 and 1.3 described the generative model for the genotype data. The probability of the observed genotype data under the generative model is given in Equation (1.9). All methods rely on the fundamental independence assumption in Equation (1.9) that dependence in allelic types derives only from IBD relative to the founders of the pedigree.

### *1.4.1 Exact calculation by peeling*

An early method for the calculation of probabilities of data observed on pedigrees was developed by Elston and Stewart [1971]. The algorithm was computes the probability of trait data on a pedigree, given trait data, pedigree structure, and model parameters. The trait values are $\mathbf{y}$ on observed individuals $\mathcal{Y} \subset \mathcal{F}$. Trait expression is controlled by $\mathbf{t}$, the allelic type of the individuals at a single trait locus. The trait locus is not necessarily one of the marker loci, and the trait locus alleles are not necessarily marker alleles. At the trait locus there are two alleles, $N$ and $D$, so for individual $i$, $\mathbf{t}_i \in \{NN, ND, DD\}$.

The Elston-Stewart model for the marginal probability of the trait data on the pedigree

has three parts:

$$P(\mathbf{Y}=\mathbf{y}) = \sum_{\mathbf{t}} P(\mathbf{Y}=\mathbf{y}|\mathbf{t})P(\mathbf{T}=\mathbf{t}) \tag{1.15}$$

$$= \sum_{\mathbf{t}} \left[\prod_{\mathcal{F}^{\mathcal{F}}} P(\mathbf{T}_i=\mathbf{t}_i)\right]\left[\prod_{\mathcal{F}^{\mathcal{N}}} P(\mathbf{T}_i=\mathbf{t}_i|\mathbf{t}_{i_F}\mathbf{t}_{i_M})\right]\left[\prod_{\mathcal{Y}} P(\mathbf{Y}_i=\mathbf{y}_i|\mathbf{t}_i)\right]. \tag{1.16}$$

The population model is $P(\mathbf{T}_i=\mathbf{t}_i)\forall i \in \mathcal{F}^{\mathcal{F}}$, the probability of trait genotypes in the pedigree founders. These probabilities can be determined in the same manner as the allocation of allelic types to founder genomes in Section 1.3. The transmission or meiosis model is $P(\mathbf{T}_i = \mathbf{t}_i|\mathbf{t}_{i_F}\mathbf{t}_{i_M})\ \forall i \in \mathcal{F}^{\mathcal{N}}$, the probability of the trait genotype in a non-founder given the trait genotypes of the father $i_F$ and mother $i_M$. These probabilities can also be determined by inheritance vectors, as described in Section 1.2. The penetrance model is $P(\mathbf{Y}_i=\mathbf{y}_i|\mathbf{t}_i)\ \forall i \in \mathcal{Y}$, the probability of a trait value given the trait genotype for observed individuals. As in Equation (1.9), any trait genotypes and trait values are conditionally independent over individuals given the inheritance. Parametric models for trait data are still specified in the manner of Elston-Stewart, even if the Elston-Stewart peeling is not used.

The Elston and Stewart [1971] algorithm was extended by [Ott, 1974] to consider two loci and allow the calculations of LOD scores, which are described in Section 1.6. It was also generalized to arbitrary pedigrees [Cannings et al., 1978] and more complex models [Cannings et al., 1980] and is now well known as a method for calculating probabilities on graphs [Lauritzen, 1992]. Elston-Stewart peeling is used to calculate the probability of trait values on a pedigree by factorizing the likelihood as in Equation (1.16) and computing the components. The probability of multilocus genotypes can also be calculated by factorizing in the same manner, so

$$P(\mathbf{G}=\mathbf{g}) = \sum_{\mathbf{t}} P(\mathbf{G}=\mathbf{g}|\mathbf{t})P(\mathbf{T}=\mathbf{t}) \tag{1.17}$$

$$= \sum_{\mathbf{t}} \left[\prod_{\mathcal{F}^{\mathcal{F}}} P(\mathbf{T}_i=\mathbf{t}_i)\right]\left[\prod_{\mathcal{F}^{\mathcal{N}}\setminus\mathcal{G}} P(\mathbf{T}_i=\mathbf{t}_i|\mathbf{t}_{i_F}\mathbf{t}_{i_M})\right]\left[\prod_{\mathcal{G}} P(\mathbf{G}_{i\cdot}=\mathbf{g}_{i\cdot}|\mathbf{t}_{i_F}\mathbf{t}_{i_M})\right], \tag{1.18}$$

where $\mathbf{t}_i$ is now an unobserved multilocus marker genotype.

The complexity of the calculation is linear in the number of unobserved individuals but exponential in the number of loci, as the number of possible genotypes increases exponentially with the number of loci. Elston-Stewart peeling is therefore ideal for large pedigrees that have many unobserved individuals. However, it is limited in the number of loci that can be jointly computed over before the number of multilocus genotypes to sum over becomes prohibitively large [Thompson, 2000, Abecasis and Wigginton, 2005]. Elston-Stewart peeling is used in, for example, LIPED [Ott, 1974], LINKAGE/FASTLINK [Lathrop, 1985] and VITESSE [O'Connell and Weeks, 1995].

An alternative peeling algorithm is Lander-Green [Lander and Green, 1987] which factorizes the likelihood using inheritance vectors $\mathbf{v}$ as latent states. The marginal probability of multilocus marker data on the pedigree is

$$P(\mathbf{G} = \mathbf{g}) = \sum_{\mathbf{v}} P(\mathbf{G} = \mathbf{g}|\mathbf{v})P(\mathbf{V} = \mathbf{v}) \tag{1.19}$$

$$= \sum_{\mathbf{v}} \left[\prod_{l} P(\mathbf{G}_{\cdot l} = \mathbf{g}_{\cdot l}|\mathbf{v}_{\cdot l})\right] \left[P(\mathbf{V}_{\cdot 1} = \mathbf{v}_{\cdot 1}) \prod_{l>1} P(\mathbf{V}_{\cdot l} = \mathbf{v}_{\cdot l}|\mathbf{v}_{\cdot l-1})\right], \tag{1.20}$$

where $\mathbf{v}_{\cdot l}$ are the inheritance indicators for all ordered DNA copies $j \in \mathcal{F}$ and $\mathbf{g}_{\cdot l}$ are the genotypes for all individuals $i \in \mathcal{G}$. The first component is the model for a genotype given inheritance vectors, which is discussed in Section 1.3. The second component is a Markov model for the inheritance vectors, also discussed in Section 1.3. Again, as in Equation (1.9), the genotypes are conditionally independent given the inheritance.

The complexity of Lander-Green peeling is linear in the number of loci but exponential in the number of non-founders in the pedigree. The number of possible inheritance vector configurations to sum over increases exponentially with the number of non-founders. Compared to Elston-Stewart peeling, Lander-Green peeling allows more loci to be analyzed jointly. However, the pedigree size is limited to a small number of meioses.

The Lander-Green peeling algorithm is used in for example GENEHUNTER [Kruglyak and Lander, 1995, Kruglyak et al., 1996], ALLEGRO [Halperin et al., 2009], MERLIN [Abecasis

et al., 2002] and MORGAN [Tong and Thompson, 2008].

### *1.4.2 Approximate Calculation by MCMC*

In pedigrees where exact computation is infeasible due to number of meioses, number of loci or both, a common strategy is to split the pedigree into smaller easier to analyze sub-pedigrees. This has been shown to result in significant power loss in linkage analysis [Dyer et al., 2001]. An alternative that allows the use of the full pedigree are Markov-Chain Monte-Carlo (MCMC) methods. Examples are LOKI [Heath, 1997], SIMWALK2 [Sobel and Lange, 1996] and MORGAN 3.3 [MORGAN, 2016]. Such methods provide approximations to the likelihood and have the advantage that restrictions on pedigree size, number of markers and inbreeding are removed. The complexity is linear in the number of markers and number of individuals. The disadvantage is that many iterations may be required for the sample distribution to converge to the true posterior and the calculations can still be computationally intensive.

To realize inheritance vectors and thus FGLs jointly over multiple loci conditional on genotype data, the MORGAN program `gl_lods` uses an LMM-sampler method Tong and Thompson [2008]. The realizations, denoted $\mathbf{x}^{(r)}$ for $r = 1, ..., R$ are from the equilibrium distribution $P(\mathbf{X}^{\mathcal{F}} = \mathbf{x}^{\mathcal{F}}|\mathbf{g})$ where $\mathbf{x}^{\mathcal{F}}$ is for all individuals in $\mathcal{F}$ and $\mathbf{g}$ is over genotyped individuals $\mathcal{G} \in \mathcal{F}$.

The LMM sampler is used to update the IBD state inheritance vectors which are defined at all loci and all meioses. As complex computations are performed on the MCMC samples, realizations are only saved at intervals to reduce autocorrelation in sequential samples [Geyer, 1992]. The update is done either by locus (L-sampler) or by multiple meiosis (MM-sampler). Further details on the L-sampler can be found in Thompson [2000] and the MM-sampler in Tong and Thompson [2008]. If the update is by the L-sampler, for each locus the inheritance vectors are updated based on the genotype at the locus and the current realization of inheritance vectors at neighboring loci. For instance a paternal copy may be inherited instead of the maternal. The Markov property is assumed for inheritance vectors along the chromosome. If the update is made by the MM-sampler, inheritance vectors at

all loci are updated for multiple meioses, a subset of meioses small enough for Lander-Green calculations, to reflect altered recombination points along the chromosomes.

### *1.5 IBD Inference*

Correlation in observed genotypes, or IBS, has long been used to measure relatedness.IBS can be used to estimate the local IBD state, or the presence or absence of IBD at a given locus. Evidence for local IBD is in segments of IBS along the chromosome. Chromosomes share a haplotype that due to its population frequency is not expected to be shared by a set of chromosomes of that size selected randomly from the population. The smaller the by-chance probability the stronger the evidence of IBD. Haplotypes with a larger number of markers have lower frequency so when shared provide clearer evidence of more recent IBD. For shorter IBD segments evidence of IBD can be hard to distinguish from linkage disequilibrium (LD).

The number of possible IBD states among a set of individuals increases exponentially with the number of individuals. For two individuals there are 15 possible IBD states at each locus, for five there are more than $10^5$, and for ten more than $10^{13}$. If the pedigree structure is observed, it can be used as an informative prior on the IBD state among the individuals. For example, if there are no loops in the pedigree there is zero probability of IBD between the two DNA copies of an individual. Over multiple loci, the expected length of IBD segments is informed by the pedigree. Closer relatives are separated by fewer meioses and are expected to share longer segments IBD, as discussed in Section 1.2. The pedigree structure also defines the set of founder individuals that the IBD is relative to, and any IBD shared between founders relative to an ancestral population is ignored. IBD inference when the pedigree structure is known is described in Section 1.5.1.

IBD inference is also possible when the pedigree structure is not observed. There are no pedigree constraints that limit the IBD state space. For inference, assumptions must be made about the unobserved population pedigree structure. IBD inference when the pedigree structure is unknown is described in Section 1.5.2.

### *1.5.1 When Pedigree is Known*

We are interested in inference about the IBD state among all individuals in $\mathcal{F}$ relative to the family founders $\mathcal{F}^{\mathcal{F}}$, $P(\mathbf{x}^{\mathcal{F}}|\mathbf{g})$. As described in Section 1.1, in pedigree-based IBD inference, we observe the family pedigree structure and multilocus marker data on a subset of the family individuals $\mathcal{G} \subset \mathcal{F}$.

In a pedigree, inheritance vectors, $\mathbf{v}$, have a Markov dependence over loci. This is a key dependence in the Lander-Green method in Section 1.4.1, and was also discussed in Section 1.2. Recall from Section 1.2 that FGLs are functions of inheritance vectors, $\phi(\mathbf{v}) = \mathbf{x}$, but do not in general have Markov dependence across loci. One exception is when the only family structure is sibships. For siblings, the FGL states are Markov and a hidden Markov model can be used for FGL inference. IBD inference in siblings is discussed in Chapter 3.

For a general pedigree, MCMC realization of FGLs from the pedigree conditional on observed marker genotypes is used to approximate $P(\mathbf{x}^{\mathcal{F}}|\mathbf{g})$. The MORGAN `gl_auto` program, described in Section 1.4.2, [Tong and Thompson, 2008, MORGAN, 2016] samples MCMC realizations using an LMM sampler.

The resolution of IBD segment detection is limited in pedigree-based methods. IBD is relative to pedigree founders so there are very few meioses and IBD segments are long [Boehnke, 1994]. Additionally, linkage marker panels typically contain no more than 2 SNPs per cM, which is sparse compared to available SNP panels. The sparsity of the linkage panel allows markers to be selected that are in linkage equilibrium, and the IBD segments are long enough that they will still cover many markers. When MCMC simulation is used, sparse markers allow better mixing of the Gibbs samplers. On dense panels the probability of recombination and state change between markers becomes very small.

In a pedigree setting the definition of IBD relative to the pedigree founders can also considered a limitation. There is no allowance for IBD between the pedigree founders who may themselves be members of the same population. For example, in the ERF pedigrees described in Section 2.5, pedigree founders are at least members of the same isolated popula-

tion and are likely to be close relatives. The extent of a pedigree is somewhat arbitrary and may have had branches and ancestors pruned to simplify pedigree calculations or remove uncertain relationships.

### *1.5.2 When Pedigree is Unknown*

Suppose the pedigree structure that relates individuals $\mathcal{G}$ with observed marker genotypes is not observed. We are interested in the IBD state among $\mathcal{G}$ relative to the founders of the population $\mathcal{P}$. The population founders $\mathcal{P}^{\mathcal{F}}$ are typically more ancient than the founders in a pedigree-based analysis and the relationships between $\mathcal{G}$ are unknown as they are not close relatives. There are a large number of meioses separating individuals in $\mathcal{G}$ from their most recent common ancestor, thus IBD segments are short and few. Any similarity in allelic types that is a result of IBD may be hard to distinguish from local LD.

Markers for the estimation of short segments of IBD should be both dense and have a low level of LD. These are conflicting goals as the closer the spacing of markers the more LD between them in general. Definitions of dense vary by application, for instance Browning and Browning [2010] use SNPs spaced from 150-450 per cM but account for LD in their model. Brown et al. [2012], Zheng et al. [2014] do not model LD and use a dense panel of approximately 50 SNPs per cM. In this thesis panels of 40-50 SNPs per cM are considered dense whereas a linkage panel of <2 SNPs per cM is considered sparse. Panels are selected such that LD between the markers in the panel is minimized. A simple method to reduce LD is to thin markers, reducing the number of markers and increasing the distance between them. From a dense panel of $N$ markers per cM that are roughly evenly spaced, a sparse panel of $n < N$ markers per cM can be formed by selecting every $N/n$th marker. A more sophisticated and commonly used method is PLINK LD-pruning [Purcell et al., 2007b, Purcell, 2015] where markers are selected specifically to minimize LD. Purcell et al. [2007b] suggest a very high threshold allowing $R^2$ between markers of up to 0.5. In this thesis PBAP [Nato et al., 2015] was used for LD-pruning with a more stringent $R^2$ threshold, see Chapter 2 for details.

For estimation of IBD in a population setting, assumptions must be made about population IBD sharing. Browning and Browning [2012] review methods for the detection of IBD in remote relatives. Methods broadly fall into two categories: rule-based and model-based. In rule-based methods, searches are made for shared haplotypes in large population samples. Methods include GERMLINE [Gusev et al., 2009], Kong's method [Kong et al., 2008] and BEAGLE `fastIBD`, [Browning and Browning, 2011]. In model-based approaches, the probability of a randomly selected pair of chromosomes being IBD at a locus and the expected length of IBD segments must be supplied. Direct estimation of the ancestral recombination graph (ARG) that describes the true IBD relationship is computationally difficult [Kuhner and Smith, 2007]. Furthermore, the ARG and hence IBD in the population sample is not Markov along the chromosome, but can be approximated closely with a Markov model [McVean and Cardin, 2005]. The majority of model-based methods use a HMM in which the latent states are the unobserved IBD states at marker loci along the chromosome and the emissions are the genotypes. The Markov assumption is made for FGLs over the chromosome, that is,

$$P(\mathbf{x}_{i\cdot}, \mathbf{g}_{i\cdot}) = P(\mathbf{g}_{i1}|\mathbf{x}_{i1})P(\mathbf{x}_{i1}) \prod_{l=2}^{L} P(\mathbf{x}_{il}|\mathbf{x}_{il-1})P(\mathbf{g}_{il}|\mathbf{x}_{il}). \tag{1.21}$$

Emission probabilities, $P(\mathbf{g}_{il}|\mathbf{x}_{il})$ can be calculated with a genotype model as in Section 1.3. A transition model between IBD states along the chromosome must be assumed. Without a pedigree there are no explicit constraints on possible transitions between hidden IBD states between adjacent loci but the transition model expresses the fact that adjacent states are likely to be similar. Leutenegger et al. [2003] introduced a two-state HMM to model IBD/non-IBD for a pair of homologous chromosomes in an individual and the same model is used in Browning [2008] for pairs of phased haplotypes from a population. The first pedigree-free model for IBD between a pair of diploid individuals was PLINK [Purcell et al., 2007a].

Models typically assume linkage equilibrium between the markers. LD-pruning or thinning markers reduces LD but also decreases the number of markers available thus increasing

the minimum detectable segment length. Improvements that take LD into account have been developed [Albrechtsen et al., 2009, Browning, 2008, Browning and Browning, 2010] and require additional information from large samples of haplotypes from well-studied populations in order to fit an LD model.

In most populations the probability of IBD between the two DNA copies of a single individual is at least as great as between DNA copies in different individuals [Thompson, 2013]. Models with more hidden states that reflect all possible IBD states between four DNA copies have been developed [Thompson, 2008] and have been extended to any number of chromosomes [Thompson, 2009, Brown et al., 2012]. In Brown et al. [2012], the transitions between hidden IBD states along the chromosome are modeled with a modified Chinese restaurant process (CRP). The transition probabilities are given in Table B1 of Brown et al. [2012]. A new "potential" chromosome is added to an IBD group at rate $\beta j$ where $j$ is the number of chromosomes already in the group and $\beta$ is the pairwise probability of IBD, or forms a new group not IBD to any existing group at rate $(1 - \beta)$. Instantaneously, with each new addition a random one of the $n + 1$ chromosomes is removed and the new chromosome assumes the identity of the exiting chromosome. This defines a rate matrix $Q$ that is multiplied by a rate-of-change parameter $\alpha$ to control the length of IBD segments. For a randomly selected pair of chromosomes, the pointwise probability of IBD relative to population founders is $\beta$, and the expected length of such a segment is $(2\alpha)^{-1}$ cM. The stochastic process has Ewens sampling formula Ewens [1972] as its stationary distribution. The set of emissions from the HMM along the chromosome are genotypes or haplotypes. Allowance is made for genotyping error, however, LD is not modeled.

A limitation of the HMM framework for IBD detection is the number of individuals whose IBD state can be estimated jointly. The state space grows rapidly making forward-backward hidden-state probability calculations intractable. The direct estimation of larger joint graphs has been attempted. Moltke et al. [2011] use a MCMC method with simplified latent state and transition model. Zheng et al. [2014] use a reversible jump MCMC which is limited to haplotypic data. A particle filter approach was used by Glazner [2014]. All the MCMC

methods are intensive computationally so are limited in number of individuals and number of SNPs. For example, Zheng et al. [2014] use up to 40 haplotypes and 860 markers over 10 Mbp and Moltke et al. [2011] use up to 10 individuals and 501 markers over 8 Mbp.

Methods that construct a joint graph from pairwise IBD estimates include Gusev et al. [2011], He [2013] and Qian et al. [2014]. Glazner and Thompson [2015] introduced a sequential HMM method for joint IBD estimation from population samples. Monte-Carlo realizations of the joint state are produced, each one built up by sequentially adding individuals. As individuals are added their IBD states are modeled by an HMM conditional on marker data and individuals already in the graph. The method is implemented in `ibd_stitch`. `ibd_stitch` was demonstrated on sets of up to 40 chromosomes on regions of up to 40Mbp, however, good estimates of joint IBD were only achieved for 10-15 individuals. In larger sets, as individuals are added there are too many restrictions placed on IBD states available to the new individual. Correctly estimated IBD sharing with individuals to be added can be contradictory to intermediate states leading to incorrect estimates and long computation times. LD is not modeled. The `ibd_stitch` model assumes the chromosomes are sampled from a population and IBD between the chromosomes is relative to a distant population founder ancestor. The `ibd_stitch` model used in this thesis does not account for any close pedigree relationships between the chromosomes which can provide strong priors on IBD state probabilities.

The inputs to the `ibd_stitch` program are the kinship change rate $\alpha$ and population kinship $\beta$ of the CRP [Brown et al., 2012], the null fraction, and the genotyping error rate. The null fraction parameter is used to mix the stationary distribution of IBD states with the discrete jump chain to ensure that there are no zero-probability transitions between very close markers. The null fraction is the proportion of the stationary distribution that is used.

Gene detection using IBD-based mapping is discussed in Section 1.7. There are other uses for IBD from population samples in population genetics including demographic history and ancestral population size estimation [Browning and Browning, 2015, Palamara and Pe'er, 2013] and migration rates [Palamara and Pe'er, 2013]. Genome-wide IBD estimates

are also used to adjust for relatedness in genome-wide association studies (GWAS) [Euahsunthornwattana et al., 2014].

## *1.6 Linkage Analysis*

The application of IBD in this thesis is the detection of the location of genes contributing to expression of a trait. The trait models used are described in Section 1.6.1. In a pedigree-based study linkage analysis for trait locus detection, described in Section 1.6.2. The probability of the trait data is conditioned on IBD relative to pedigree founders. In Section 1.7 IBD-based mapping methods are described. IBD-based mapping methods condition on IBD relative to population founders in distantly related individuals. In Section 1.8 association tests are described. Association tests do not condition on IBD.

### *1.6.1 Trait Model*

There are broadly two models for the genetic basis of complex human diseases [Iyengar and Elston, 2007]. Diseases are described as rare or common depending upon their prevalence in the population. A disease may also be caused by a rare or common variant. Under the common variant hypothesis there are a few common allelic variants that account for genetic variation in disease susceptibility. Under the rare variant hypothesis there is allelic heterogeneity - there are many different DNA sequence variations in a gene that can result in disease susceptibility. Both disease mechanisms have been found in human diseases. A commonly cited common variant is the APOE allele implicated in Alzheimer's disease [Corder and et al, 1993]. Cystic Fibrosis is caused by mutations in the CFTR gene. It has both a common mutation that accounts for 70% of cases, but thousands of rare mutations that also increase risk [Sosnay et al., 2013].

In a pedigree, trait values $\mathbf{y}$ are observed on individuals $\mathcal{Y} \subset \mathcal{G}$. In this thesis we assume that there is one locus controlling trait expression. The trait model is specified in the manner of Elston-Stewart in Equation (1.16). The alleles at this locus are a neutral allele $N$ and a disease allele $D$. The trait genotype is $\mathbf{t}$ where $\mathbf{t}_i \in \{NN, ND, DD\}$. If the trait

is qualitative, such as case/control status, the trait value for individual $i$ is $\mathbf{y}_i \in \{0,1\}$ and the penetrance probabilities are $P(\mathbf{Y}_i = \mathbf{y}_i|\mathbf{t}_i)$. If the trait is quantitative we assume $\mathbf{Y}_i \sim N(\mu_i, \sigma_i)$ where $\mu_i$ and $\sigma_i$ are functions of $\mathbf{t}_i$.

In this thesis under the rare variant hypothesis, the trait allele status is tied to the FGL at the trait locus. Each FGL that is designated $D$ represents a unique, rare, trait variant that developed on a different haplotypic background. Under the common variant hypothesis, the trait allele status is tied to the allelic type of the FGL at the trait locus. For instance, the allele $A$ is assigned to $N$ and the allele $B$ is assigned to $D$. The simulation of trait values, and penetrance under either hypothesis is described in Chapter 2, Section 2.2.1 for the sibling data set and 2.3.1 on larger pedigrees.

### *1.6.2 Parametric Linkage Analysis*

Assume that trait values $\mathbf{y}$ have been observed for individuals $\mathcal{Y} \subset \mathcal{F}$ and multilocus marker genotypes have been observed for individuals $\mathcal{G} \subset \mathcal{F}$. A pedigree structure is observed for $\mathcal{F}$. A linkage analysis is performed to identify marker loci that are linked to the unknown locus that controls trait expression.

In a parametric linkage analysis, a trait model is assumed in the manner of Elston-Stewart in Section 1.4, Equation (1.16). A likelihood ratio test is performed at hypothesised trait locations, which in this thesis are the marker loci. The null hypothesis, $H_0$ is that the trait is unlinked to all marker loci. The alternative, $H_1$, is that the trait is at the hypothesised location. The test statistic is the map-specific multipoint LOD score at each locus $l$,

$$LOD = \log_{10} \frac{P_{H_1}(\mathbf{Y} = \mathbf{y}, \mathbf{G} = \mathbf{g})}{P_{H_0}(\mathbf{Y} = \mathbf{y}, \mathbf{G} = \mathbf{g})}. \tag{1.22}$$

The calculation of the LOD is discussed further in Section 1.6.3. Using data from multiple markers in the LOD score for a single locus leads to gains in power and fewer false positive signals [Wijsman and Amos, 1997] as does using larger pedigrees [Wijsman and Amos, 1997, Wright et al., 1999].

Typically, 3 is used as the threshold for evidence of linkage between the trait and marker.

A LOD of 3 indicates that the odds that the loci are linked are 1000 times greater than the odds they are not. The LOD is calculated along the genome at multiple (non-independent) loci and the threshold does not take into account multiple testing. A LOD of zero or less indicates no evidence against the null hypothesis of no linkage.

Robustness to mis-specification of the trait model is a concern with parametric methods as the true trait model is almost always unknown. To avoid specification of the trait model, non-parametric models have been developed. These models rely on the detection of excess IBD sharing at the trait locus. Non-parametric methods were first suggested by Penrose [1953], comparing IBD distributions for affected siblings. Risch [1990] generalized this to evaluate evidence for linkage. Examples of non-parametric linkage tests are affected sib pair tests, such as Kong and Cox [1997], described in Section 1.6.4.

### *1.6.3 Calculation of LOD scores*

When available marker data were sparse along the chromosome, the LOD score was calculated exactly by Elston-Stewart peeling. With modern dense marker data it is typically calculated by Lander-Green peeling or by MCMC [Thompson, 2000]. Peeling algorithms are described in Section 1.4. The MCMC calculation of the LOD implemented in MORGAN [MORGAN, 2016] is described here. The calculation of the LOD score is separated into two distinct steps. The first step is obtaining MCMC realizations of FGLs conditional on marker data, implemented in `gl_auto` and described in Section 1.4.2. The second step is calculating probabilities of trait data at each locus for the FGL realizations, implemented in `gl_lods`.

In the absence of linkage, the trait and marker data are independent, so the LOD score can be written

$$LOD = \log_{10} \frac{P_{H_1}(\mathbf{Y} = \mathbf{y}, \mathbf{G} = \mathbf{g})}{P_{H_0}(\mathbf{Y} = \mathbf{y}, \mathbf{G} = \mathbf{g})} \tag{1.23}$$

$$= \log_{10} \frac{P_{H_1}(\mathbf{Y} = \mathbf{y}, \mathbf{G} = \mathbf{g})}{P_{H_0}(\mathbf{Y} = \mathbf{y}) P_{H_0}(\mathbf{G} = \mathbf{g})} \tag{1.24}$$

$$= \log_{10} \frac{P_{H_1}(\mathbf{Y} = \mathbf{y} | \mathbf{G} = \mathbf{g})}{P_{H_0}(\mathbf{Y} = \mathbf{y})}. \tag{1.25}$$

Assiming that the trait value and genotype are independent given the IBD state, the numerator in (1.25) can be expressed as

$$P_{H_1}(\mathbf{Y}=\mathbf{y}|\mathbf{G}=\mathbf{g}) = \sum_{\mathbf{x}} P_{H_1}(\mathbf{Y}=\mathbf{y}|\mathbf{x})P(\mathbf{X}=\mathbf{x}|\mathbf{g}) = E\left[P_{H_1}(\mathbf{Y}=\mathbf{y}|\mathbf{x})\right] \tag{1.26}$$

where $E[]$ is the expectation under the model for marker data. The calculation of (1.26) is approximated with MCMC realizations of the FGLs $\mathbf{x}^{(r)}$, $r = 1, ..., R$,

$$E\left[P_{H_1}(\mathbf{Y}=\mathbf{y}|\mathbf{x})\right] \approx \frac{1}{R}\sum_{r}\left[P_{H_1}(\mathbf{Y}=\mathbf{y}|\mathbf{x}^{(r)})\right] \tag{1.27}$$

The denominator of the LOD score in Equation (1.25), $P_{H_0}(\mathbf{Y}=\mathbf{y})$, must also be supplied to `gl_lods`. This probability calculation is for a single locus over a large pedigree, so in most cases it can be calculated exactly with Elston-Stewart peeling. In analyses in this thesis the value was computed by running the MORGAN program `lm_linkage` without any MCMC computations.

### *1.6.4 Sib-Pair Tests*

Sib-pair tests are linkage tests that compare the distribution of IBD states in pairs of siblings, conditioning on their phenotype status. The most common sib-pair test is the affected sib-pair (ASP) test where both sibs are affected with a dichotomous trait. No trait model is assumed. The distribution of IBD states between pairs of affected siblings is compared to the null distribution ( 1/4 share 0, 1/2 share 1 and 1/4 share 2 copies IBD).

Some early work with the ASP method considered IBD sharing in the HLA region, implicating HLA in diabetes [Cudworth and Woodrow, 1975]. ASP methods have also been used with other complex diseases like Alzheimer's [Pericak-Vance et al., 1991], atopy [Moffat et al., 1992], Schizophrenia [Gill et al., 1996], colorectal cancer [Tomlinson et al., 2007] and Type-2 diabete [Ghosh et al., 2000]. As parental data are not required, sib-pair tests can be particularly useful for late-onset diseases. The ASP test has also been extended in a variety of ways. Extensions to the joint analysis of quantitative traits [Haseman and Elston, 1972], multiple loci [Fulker and Cardon, 1994, Kruglyak and Lander, 1995], other

types of relative [Weeks and Lange, 1988, Cordell et al., 2000], and environmental risk factor covariates [Khoury et al., 1991].

There is no uniformly most powerful test for the hypothesis of no linkage. Dudoit and Speed [1999, 2000] propose a sib-pair test that is locally most powerful for alternatives close to the null. The Dudoit test can be used on either qualitative or quantitative traits and uses sib pairs of all phenotypes. The test is based on the likelihood of IBD conditional on the phenotypes of all pairs, thus avoiding unrealistic sampling assumptions and allowing the use of pairs from different ascertainment mechanisms in a single analysis. The test statistic is a score statistic in the recombination fraction between a marker locus and the unknown trait locus,

$$D = 16 \sum_{i=1,...,N} (\pi_{2i} - \pi_{0i})(N_{2i} - N_{0i}), \tag{1.28}$$

where $i = 1, ..., N$ are all sib pairs, $\pi_{ij}$ is the probability that pair $i$ shares $j$ DNA copies IBD at the trait locus and $N_{ij}$ is the indicator that pair $i$ shares $j$ DNA copies IBD at the marker locus. $\pi_{ij}$ is calculated under the trait model and $N_{ij}$ is based on estimated IBD. A general genetic model can be used, with multiple genes unlinked to each other, arbitrary penetrances, and no population genetic assumptions such as random mating or HWE. In the case of qualitative trait data, the statistic is a linear combination of the mean IBD statistic of Blackwelder and Elston [1985] and Knapp et al. [1994] for affected, unaffected, and discordant sib pairs with weights depending on the genetic model.

A major disadvantage of sib-pair tests is that sib-pair data are rarely collected for large numbers of pairs. It is useful, however, to illustrate the advantages of descent-based tests in general and the potential of combining population and pedigree information in Chapter 3. Allowing for population-level relatedness in the parents of sib-pairs should give additional power especially in the case of a recessive trait where in affected sib-pairs both are carriers.

## *1.7 IBD Mapping*

IBD mapping methods are used for detecting trait loci in distantly related individuals, using IBD relative to population founders. An overview of methods for detection of genes using IBD-based mapping between individuals not known to be related is given in Browning and Thompson [2012]. Tests generally use either "clustering" statistics that model haplotype clustering or "pairwise IBD" methods where pairwise IBD estimates are used directly. Glazner and Thompson [2015] uses the a direct estimate of the joint IBD graph.

Clustering methods rely on the detection of shared haplotypes, clustering shared haplotypes into IBD classes at a locus. All haplotypes in the same class are IBD with each other. Each individual, having two haplotypes, is a member of two classes. Clusters are tested for association with case-control status [Gusev et al., 2011]. Browning and Thompson [2012] note that if ancestry is recent there will be few individuals in each cluster so power is low. Conversely, if ancestry is more distant haplotypes are shorter and are shared with less certainty. Another limitation is that sharing of haplotypes is determined between each pair. The methods must then resolve joint IBD clusterings from the pairwise estimates. When IBD is not estimated with 100% certainty there may be situations where, for instance, if A and B are estimated to be IBD, and B and C are estimated to be IBD, but A and B are not.

Pairwise methods compare the rate of IBD in case/case to case/control or control/control pairs, to detect excess IBD sharing expected the trait locus [Purcell et al., 2007a]. Unlike affected sib-pairs or other relative pairs the individuals do not have a known relationship to help detect IBD. The lack of knowledge of the relationship also means that control individuals are necessary to determine the "background" level of IBD in controls [Browning and Thompson, 2012].

In samples of distantly related individuals, pairwise IBD mapping methods have had success where association tests have failed. Browning and Thompson [2012] demonstrate that IBD mapping has higher power compared to association testing when there are multiple rare variants within a gene that contribute to disease susceptibility. Some successes of IBD

mapping in distantly related samples are Gusev et al. [2011, 2012] who used IBD mapping to discover genome-wide significant regions in isolated populations and outbred populations where association tests failed; Francks et al. [2010] used IBD mapping to identify potential susceptibility locus for schizophrenia and bipolar disorder with genotype data in case-control samples; Ionita-Laza et al. [2013] found a genome-wide significant linkage signal in a dataset of multiple sclerosis patients and Letouze et al. [2012] searched for founder mutations in cancer samples.

Glazner and Thompson [2015] uses the joint IBD graph between two or more individuals. The test statistic is similar to the linkage test statistic in Equation 1.22. The numerator is calculated using realizations of the joint IBD graph from `ibd_stitch` are used in place of realizations of pedigree IBD graphs. The `ibd_stitch` method for the joint IBD realizations was described in Section 1.5.2. The LOD score denominator cannot be calculated without the pedigree structure. Glazner and Thompson [2015] instead uses a permutation test, permuting trait values over individuals at each locus.

## *1.8 Association Tests*

Association studies are an alternative to IBD-based methods for mapping the location of disease genes, such as linkage analysis in Section 1.6, or IBD mapping in Section 1.7. Association studies test whether alleles are associated with the trait without conditioning on IBD state. There are three possible causes for allelic association with a trait. The ideal situation is that the allele is a causal variant that increases disease susceptibility. There will also be association with the allele if there is LD between the allele and a causal variant as the allele and the causal variant will be correlated. Finally, if there is population stratification there may be a spurious correlation between the allele and the trait if the allele and trait occur at different frequencies in different sub-populations. There are population-based and family-based association tests, each utilize different strategies to minimize the effect of population stratification.

Population-based association tests use "unrelated" cases and controls and cross-classify by

genotype. The "unrelated" individuals are in fact distantly related individuals whose shared pedigree relationship is unknown. A chi-squared test or logistic regression is commonly used. The most common population-based association design is the genome-wide association study (GWAS) that tests association on panels of millions of SNPs across the genome. The p-value for significance is adjusted for multiple testing, conventionally $5 \times 10^{-8}$ [Bush and Moore, 2012]. GWAS relies on the common disease common variant hypothesis for the genetic basis of traits, see Section 1.1. GWAS can account for population structure by partitioning the sample into subpopulations [Falush et al., 2003], principle components analysis of sample structure [Price et al., 2006] and accounting for estimates of pairwise relatedness between the individuals Kang et al. [2010].

Family based association tests use cases and controls with a known family relationship, and use expected allele transmission rates. Family-based tests provide some protection against population stratification and admixture. The first family-based association test was the transmission disequilibrium test (TDT) [Spielman et al., 1993]. The original TDT uses genotype data from trios - affected offspring and parents. The null hypothesis is no linkage and no association between the marker locus and the trait locus, the alternative is that the marker is linked and associated with the trait locus. Both linkage and association must be present in order to reject the null. In population-based tests only association is detected. If there is linkage but no association the marker and trait are transmitted together but with different markers in different families so there is no overall association with a particular allele. If there is association but no linkage (e.g. admixture or population structure) there is no tendency for marker and trait to be transmitted together. The TDT does not require the specification of a disease model or assumptions about the distribution of the disease in a population. Laird and Lange [2006] and Ott et al. [2011] contrast family and population based association tests.

One generalization of the TDT is the sibling transmission disequilibrium test (STDT) Spielman et al. [1993]. The test is applicable to a dichotomous trait and applies to discordant sib pairs. The test compares the genotype counts of affected sibs and counts expected under

the null hypothesis of transmission equilibrium, conditional on the genotypes of the entire sibship. The family-based association test (FBAT) [Laird et al., 2000] is a generalization of the TDT that allows for more general traits including quantitative and dichotomous phenotypes, covariates, more general family relationships and multiple markers [Ott et al., 2011]. The theoretical basis for the FBAT is given in Rabinowitz and Laird [2000]. Population and family-based designs have also been combined, to leverage both types of relationship. For example Ott et al. [2011] use family designs to screen SNPs for use in association tests. Laird and Lange [2008] explore the separation of population and family information. Population information is subject to bias by population substructure, and is used for screening or model development.

Association tests, compared to descent based tests such as linkage analysis, can have more power for disease genes with weak effect [Ott et al., 2011]. However, the inability to detect variants to explain much of the heritability in the most common disorders with association tests has led to renewed interest in descent-based methods. Descent-based methods are particularly useful for rare risk variants - when variants occur infrequently, conclusive evidence of the disease link will require observation of co-inheritance within families.

## *1.9 Thesis Contributions*

IBD is an important tool for understanding the shared ancestry of a group of individuals and for detecting loci that contribute to trait expression. Methods of IBD detection are specialized for either small numbers of distantly related individuals or family pedigrees. The major goal of this thesis is to develop methods for IBD detection that combine population and pedigree information. Methods for IBD estimation should also be scalable to allow for joint IBD states between larger groups. Methods developed in this thesis use pedigree-IBD augmented with population-IBD estimates of IBD between population founders. This allows the estimation of IBD that is the result of a larger ancestral population pedigree without using the ancestral pedigree structure - which is likely unknown or inaccurate.

Glazner and Thompson [2013] showed that it was possible to improve linkage analysis on a pedigree by incorporating IBD estimated between pedigree founders. The method merged pedigree-based IBD estimates with pairwise IBD states estimated between individuals not known to be related. Joint IBD states were formed at each locus by adding pairwise states to the pedigree IBD graphs in order of descending probability, provided they did not conflict with existing IBD. The method was shown to be capable of producing accurate LOD scores in a large pedigree using only information about relationships in small subpedigrees. However, uncertainty in the pairwise inferences was not accounted for and there was a lack of smoothing across marker loci.

Real and simulated data used in this thesis are described in Chapter 2. To test methods a large scale data simulation of a population and haplotypes was performed, described in Section 2.1. Methods are also applied to the GAW and Alzheimer's disease data sets, described in Sections 2.4 and 2.5.

Methods developed for IBD estimation and trait locus detection in sib pairs are described in Chapter 3. IBD detection takes both the sibling relationship and population-level relationships between the parents into account. Association tests are compared to IBD-based tests for trait locus detection on the simulated data set. A merging method for combining IBD estimated for large pedigrees with population-IBD is described in Chapter 4. The merging method is applied to simulated data in Chapter 5 and to real data in Chapter 6. Conclusions and future work are discussed in Chapter 7.

# Chapter 2

# DATA

This chapter describes the data sets used in this thesis. A simulated population was created for this thesis to test IBD estimation and detection of trait loci when the truth is known. A description of the simulation is given in Section 2.1. From the simulated population a data set of sib pairs and a data set of small pedigrees were created. These are described in Sections 2.2 and 2.3. The GAW data set is described in Section 2.4. This data has real pedigrees and genotypes but simulated trait values. A real Alzheimer's disease data set is described in Section 2.5.

## *2.1 Simulation of Population Data*

In order to create simulated pedigree data sets of the type described in Section 1.1, a simulated population was created. The simulation follows the generative model given in Sections 1.2 and 1.3. The creation of founder haplotypes by assigning allelic types to FGLs is described in Section 2.1.1. The simulation of a multi-generational population pedigree structure and inheritance vectors is described in Section 2.1.2. The formation of haplotypes given the inheritance vectors and founder allelic types is described in Section 2.1.3.

### *2.1.1 Creation of Founder Haplotypes*

A model for the assignment of allelic types to founder genome loci was described in Chapter 1 (Equation (1.10)) that assumed no LD between the loci. The IBD detection methods in this thesis use the no-LD model. To test the robustness of IBD estimation methods to this assumption, founder haplotypes with realistic LD patterns were simulated.

LD structure and allele frequencies in the simulated founder haplotypes were modeled on

a set of haplotypes from the 1000 Genomes Project [The 1000 Genomes Project Consortium, 2015]. The 1000 Genomes haplotypes used were a mixture of Asian and European population samples. The Asian populations, with population codes, were Chinese Dai in Xishuangbanna (CDX), Han Chinese in Beijing (CHB), Southern Han Chinese (CHS), Japanese in Tokyo (JPT) and Vietnamese from Ho Chi Minh City (KHV). The European populations were Europeans with northern and western ancestry (CEU), Finnish in Finland (FIN), English and Scottish in Great Britain (GBR), Iberian in Spain (IBS) and Toscani in Italy (TSI). The mixture of Asian and European haplotypes should result in more LD than an Asian-only or European-only sample due to haplotypic similarity within each race.

Phased marker data on the haplotypes was obtained from the version 2 (February 2012) of the 20110521 release [1000 Genomes, 2012b]. Allele frequency and map position information was obtained from 1000 Genomes vcf files [1000 Genomes, 2012a]. The segment of genome used came from the shorter arm of chromosome 1.

A dense set of SNP markers were selected from the available 1000 Genomes markers. This step was performed by Prof. Mary Kuhner. The selection was made with the following criteria on the minor allele : the frequency of the reference allele should be between 0.05 and 0.95 in both populations; the frequency in at least one population should be between 0.2 and 0.8; the difference in allele frequency between populations should be greater than 0.05; and the inter-SNP distance should be greater than 25,000 bp. These criteria result in 4303 SNP markers spaced over 120 Mbp, with an average intermarker distance of 27890 bp.

A sparser subset of markers was selected for linkage analysis. The sparse set has a total of 361 markers spaced an average intermarker distance of 26,208 bp. They were selected to be relatively evenly spaced, and to give the most heterozygotes in a set of individuals in the 47-50th generation. These individuals were the members of the set of 50 pedigrees described in Section 2.3.

A BEAGLE model [Browning, 2006] was fitted to the 1000 Genomes haplotypes on the dense markers using BEAGLE version 3.3.2. The BEAGLE model is a graphical model for haplotype clusters in the form of a directed acyclic graph. At each marker there are nodes

that correspond to haplotype clusters and between adjacent markers edges join or split the clusters into new groupings. Following a path through the DAG results in a haplotype, with rare haplotypes having a low probability path and common haplotypes having a high probability path. BEAGLE uses a scale factor to determine whether nodes should be merged to reduce the number of haplotype clusters at a given node. A higher value means more merging and fewer clusters. More independence between markers gives fewer haplotype clusters, so fewer clusters can indicate lower LD. In this simulation, a value of 1 was used for the scale factor to fit a model with relatively high LD.

Founder haplotypes with the LD structure of the 1000 Genomes data as modeled by the BEAGLE DAG were simulated with the MORGAN `ibd_create` sub-program `beaglesim` Brown et al. [2012]. A total of 40,000 haplotypes were simulated for the 20,000 population founders.

### *2.1.2 Simulation of Descent through Pedigree*

The model for inheritance vectors given in Section 1.2 describes the descent of DNA on a fixed pedigree structure. For the simulated population, both the population structure and the inheritance vectors were simulated with the MORGAN `ibd_create` sub-program `simpop_fgl`, using MORGAN version 3.3.2. The inputs to `simpop_fgl` are the length of the chromosome in Mbp, the recombination rate of % per Mbp, population size, number of generations and remating probability. The remating probability is described below. Mutations are not simulated. The output describes the base-pair start and end points of each FGL along the two ordered chromosomes of individuals in the selected generations. In terms of the model in Section 1.3, it simulates the IVs, $\mathbf{v}$, and from the output the FGLs $\mathbf{x}^{\mathcal{P}} = \phi(\mathbf{v})$ can be determined for any point on the chromosome. The creation of haplotypes on these chromosomes is described in Section 2.1.3.

In the simulation, the length of the chromosome was 120 Mbp with a recombination rate of 0.9% per Mbp. The recombination rate of 0.9% per Mbp gives a genetic distance of 108 cM over 120 Mbp. The chromosome 1 sex-averaged centromere-included recombination rate is 0.96% per Mbp, the sex-averaged centromere-excluded rate is 1.08% per Mbp. Note that the

rate used in the simulation was the result of a bug in the MORGAN `ibd_create` sub-program `simpop_fgl`, which used a recombination rate of 1.08% per 1.2 Mbp resulting in the final rate of 0.9% per Mbp. This bug was later corrected in MORGAN version 3.3.1.

The population size was 20,000 individuals over 50 generations. The population contains 10,000 males and 10,000 females per generation. A new generation is formed by selecting a random male and a random female to mate and form recombined chromosomes for a male and female child, 10,000 times. Assuming a 25 year generation time, this is a period of 1250 years. A remating probability of 0.33 was used.

*Remating Probability*

The inbreeding effective population size is the size of a Wright-Fisher population that has the same rate of change of inbreeding coefficient as the population under consideration [Crow and Kimura, 1970]. A Wright-Fisher population has a constant size, generations that do not overlap, and random mating. The formula for effective population size is

$$N_e = \frac{4N - 2}{2 + var(K)}$$

where $N$ is the actual population size and $K$ is a random variable denoting the number of gametes contributed by each individual [Crow and Kimura, 1970]. Anything that increases $var(K)$ will reduce $N_e$, and cause more rapid genetic drift, that is, more rapid loss of genetic variation in the population. To maintain $N_e$ approximately equal to $N$ over successive generations $var(K)$ must be 2.

In the simulation, a constant effective population size of approximately 20,000 was maintained over the 50 generations in genetic isolation by setting the probability of selecting an individual for a second or subsequent marriage and thus controlling $var(K)$. In the simulated population the actual population size is 20,000 and each pairing passes two gametes per parent to the next generation. If the selection of each parent is a simple random sample (SRS) with replacement the expected number of matings per individual is 1 with variance $\approx 1$, so the variance in the number of gametes is 4 and the effective population size is 13,333. To

maintain the effective population size of 20,000, the variance in the number of gametes, controlled by the number of multiple marriages, needs to be reduced. There needs to be fewer individuals with no marriages, and fewer individuals with multiple marriages. To reduce the possibility of a second or subsequent marriage a rejection sampling scheme was used. An individual is selected from the population by SRS and the probability of acceptance $p^m$ where $m$ is the number of previous marriages and $p$ is a probability between 0 and 1. If the individual is rejected another is sampled until one is accepted. Accepting all sampled individuals is equivalent to $p = 1$ so $p < 1$ will reduce the probability of a second marriage and thus reduce the variance in number of marriages.

In the rejection sampling scheme, $var(K)$ depends on the number of marriages $M$ of an individual,

$$var(K) = var(2M) = 4\left[\sum_{m=0}^{10000} mP(M=m) - \sum_{m=0}^{10000} m^2 P(M=m)\right]. \tag{2.1}$$

The maximum $m$ is 10,000 if the same individual is selected for every marriage. Exact computation of $P(M = m)$ is intractible. Empirical results for effective population size under different sampling probabilities $p$ are presented in Table 2.1, simulating 100,000 draws with this sampling scheme for $N = 20,000$. To obtain an effective population size of 20,000, we need $var(K) = 2$. A value of $p = 0.33$ maintains the actual and effective population sizes close to 20,000. In Figure 2.1 simulated populations with $p = 1$ and $p = 0.33$ are compared. As expected, the distribution of the number of marriages per individual has lower variance when $p = 0.33$.

| $p$ | $var(k)$ | $N_e$ |
|---|---|---|
| 1 | 1 | 13,333 |
| 0.5 | 2.36 | 18,349 |
| 0.33 | 2.11 | 19,455 |

Table 2.1: Empirical effective population size for different $p$

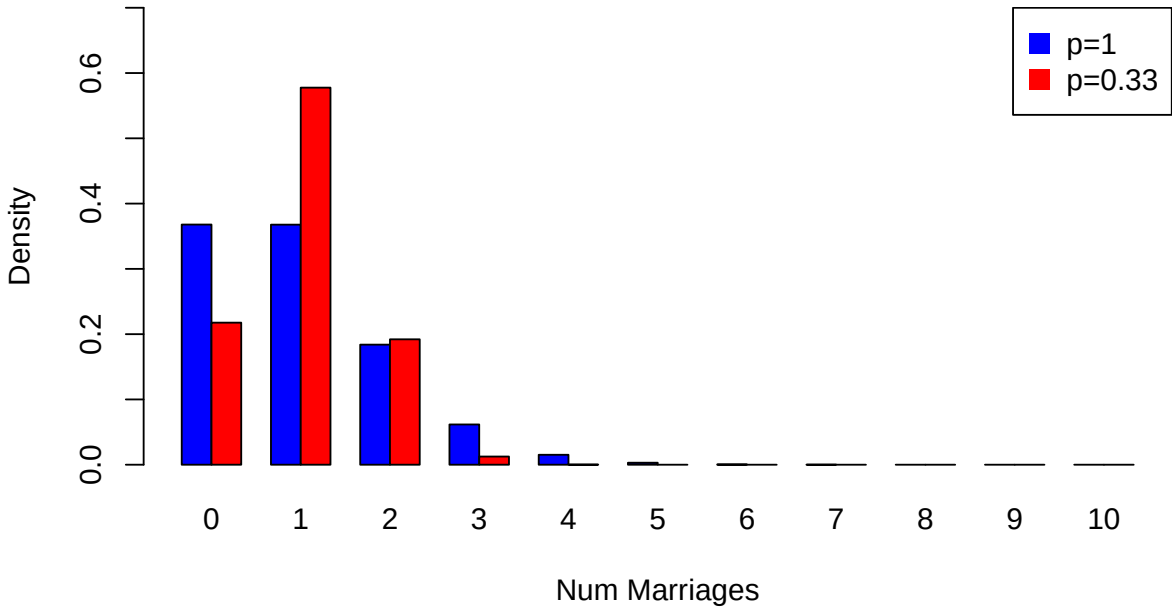



Figure 2.1: Number of marriages per individual with $p = 1$ and $p = 0.33$

A person has two parents and at most four grandparents and eight great-grandparents. In the simulated population the number of unique great grandparents of a person was determined. For the 20,000 individuals in the 4th generation, the number of great-grandparents they each had is given in Table 2.2. A small number of individuals had relatively high inbreeding with for instance only 4 great-grandparents. Overall there were fewer individuals with less than 8 grandparents when $p = 0.33$.

Sampling 10,000 random pairs from the 10,000 males from the 4th generation, the number of pairs that share any parents, grandparents or great grandparents is given in Table 2.3. In the particular realizations shown for $p = 1$ there was one sibling marriage and for $p = 0.33$ there were two sibling marriages. There is a reduced chance of sharing recent ancestors when $p = 0.33$. Of those that shared great-grandparents, there is also a reduction in the number shared, as seen in Table 2.4.

| | 4 | 6 | 7 | 8 |
|---|---|---|---|---|
| $p = 1$ | 2 | 14 | 40 | 19944 |
| $p = 0.33$ | 2 | 2 | 18 | 19978 |

Table 2.2: Number of great-grandparents for 20,000 individuals in generation 4, out of possible eight.

| | Parents | G.Parents | G.G.Parents |
|---|---|---|---|
| $p = 1$ | 1 | 10 | 53 |
| $p = 0.33$ | 2 | 9 | 61 |

Table 2.3: Number of pairs that share parents, grandparents or great-grandparents in a sample of 10,000 pairs of males in the 4th generation.

| | 1 | 2 | 4 |
|---|---|---|---|
| $p = 1$ | 39 | 11 | 3 |
| $p = 0.33$ | 25 | 12 | 3 |

Table 2.4: Number of great-grandparents shared, given sharing, in a sample of 10,000 pairs of males in the 4th generation

| | G.Parents Generation | Parents Generation | Same Generation |
|---|---|---|---|
| $p = 1$ | 23.74 | 51.10 | 104.82 |
| $p = 0.33$ | 16.67 | 33.38 | 66.89 |

Table 2.5: Average number of descendants in each generation, of the great-grandparents of 100 males in the 4th generation

In a sample of 100 of the 10,000 males from the 4th generation, the great-grandparents were identified for each individual and the total number of descendants were counted. In Table 2.5 the average number of descendants in each generation is shown, with a decrease in the number of descendants when $p = 0.33$.

### *2.1.3 Creation of Haplotypes for Population Members*

Haplotypes were created for individuals in the simulated population for whom marker data is required. In Section 1.3 this is the step $\mathbf{h}_{ijl} = \alpha(\mathbf{x}_{jl})$ where the founder allelic type corresponding to an FGL is determined at marker loci. Haplotype creation was done with MORGAN (version 3.2) IBD create programs. The sub-program `flg2haplo` matches the FGL segments along the chromosome of the individuals simulated by `simpop` with the founder haplotypes simulated in `beaglesim`.

For the `flg2haplo` program the inputs were the `simpop_fgl` output file, the `beaglesim` output file, marker positions in cM and rate of cM per Mbp. In version 3.2 of MORGAN the FGL segments in the `simpop_fgl` output file were given with breakpoint positions in base pairs (bp) and the marker positions in cM. Marker positions were converted from cM to bp within `flg2haplo` using the supplied rate. To form the output haplotypes the FGL for each marker on each chromosome is matched with the corresponding founder allele. The output is a file containing the haplotypes of the simulated population members.

Observed LD along the simulated haplotypes, and the change in the LD pattern over generations is described in Figures 2.2 and 2.3. LD is measured by $R^2$, the squared correlation

between two indicator variables for the presence of particular alleles at two loci [Hill and Robertson, 1968]. The figures show a 12cM segment with a strong LD block. In Figure 2.2 the color gradient displays $R^2$ values from 0 to 1, and in Figure 2.3 the color gradient displays $R^2$ values from 0 to 0.001. Comparing the two generations in Figure 2.2 and Figure 2.3 shows that short-range LD, within the LD blocks along the diagonal at distances of less than 2 cM, is maintained over the 50 generations despite recombination. There is, however, buildup of long-range LD between SNPs in different LD blocks. For instance in generation 50 in Figure 2.3 there is new LD between SNPs approximately 6 cM apart. The buildup in LD is due to genetic drift, the disappearance of particular haplotypes that are not passed on to the next generation due to randomness in mating and meiosis [Koch et al., 2013].

## *2.2 Data Set: Simulated Sib-pairs*

From the population simulation, a large data set of sib-pairs was created for comparing IBD estimation methods and sib-pair tests. The sib-pair data set is used in Chapter 3. At each generation the population has 20,000 individuals which are 10,000 males and 10,000 females, made up of 10,000 male-female sibling pairs as described in Section 2.1. Sib-pairs from generations 1-4, 22-25, and 47-50 are used as data sets for IBD estimation in Section 3.1. The sib-pairs data sets are used to compare association and IBD-based testing in Section 3.2. Sib-pair data sets were prepared for up to 500 generations to demonstrate the buildup of association over generations.

### *2.2.1 Binary Trait Simulation*

For the sibling data replications of both a multiple rare variant (MRV) and single common variant (SCV) trait were simulated, see Section 1.1. For the MRV trait, at the selected trait locus a sample of the FGLs in the population at that locus will be labeled causal and the rest non-causal. The probability of expression of the simulated trait will be a function of the number of causal copies each individual has. By using FGLs to determine causal alleles we simulate causal variants arising on different haplotypic backgrounds giving

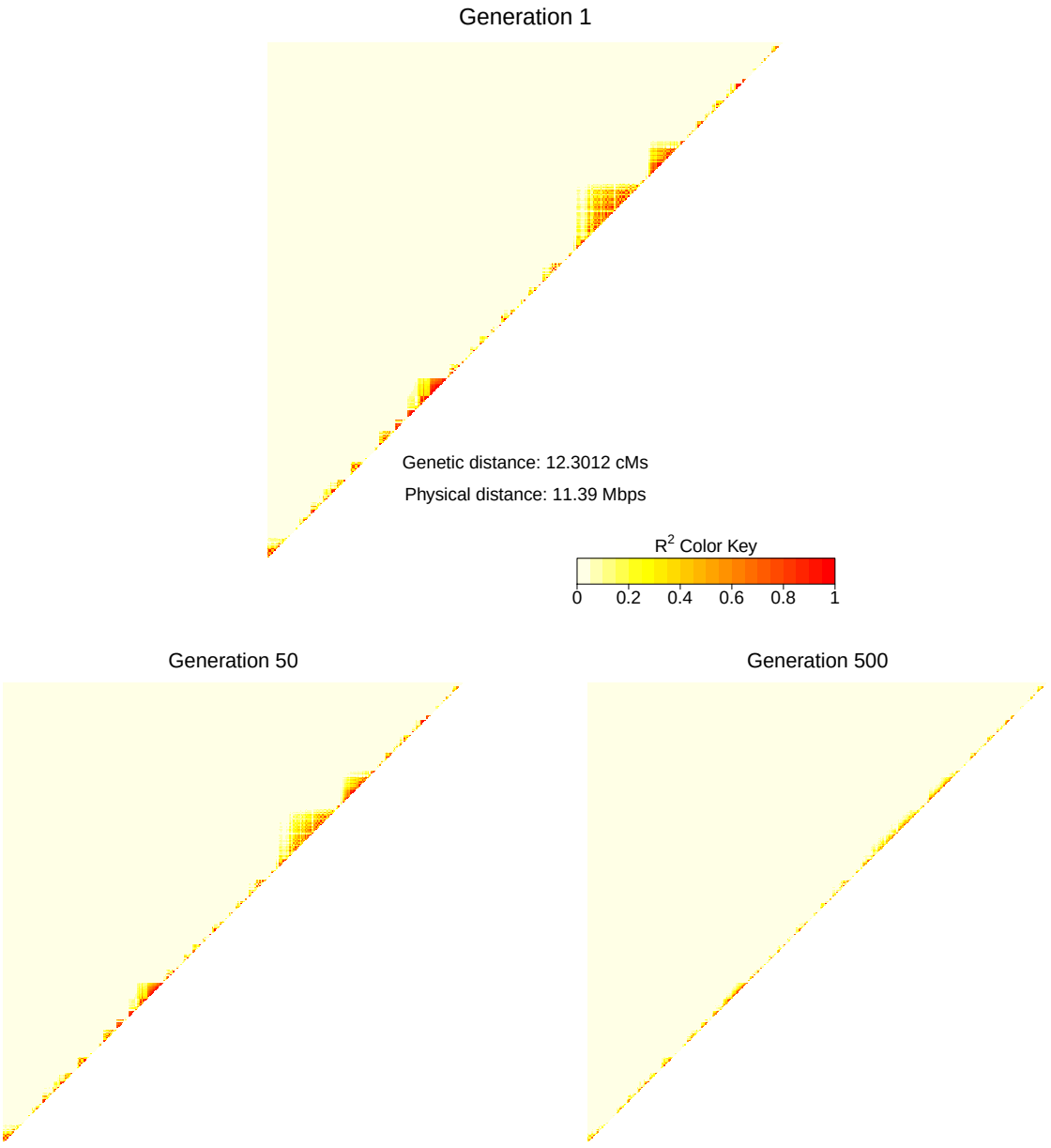



Figure 2.2: LD on a segment of simulated chromosome, measured by pairwise $R^2$ at generation 1, 50, and 500. Scale 0 to 1.

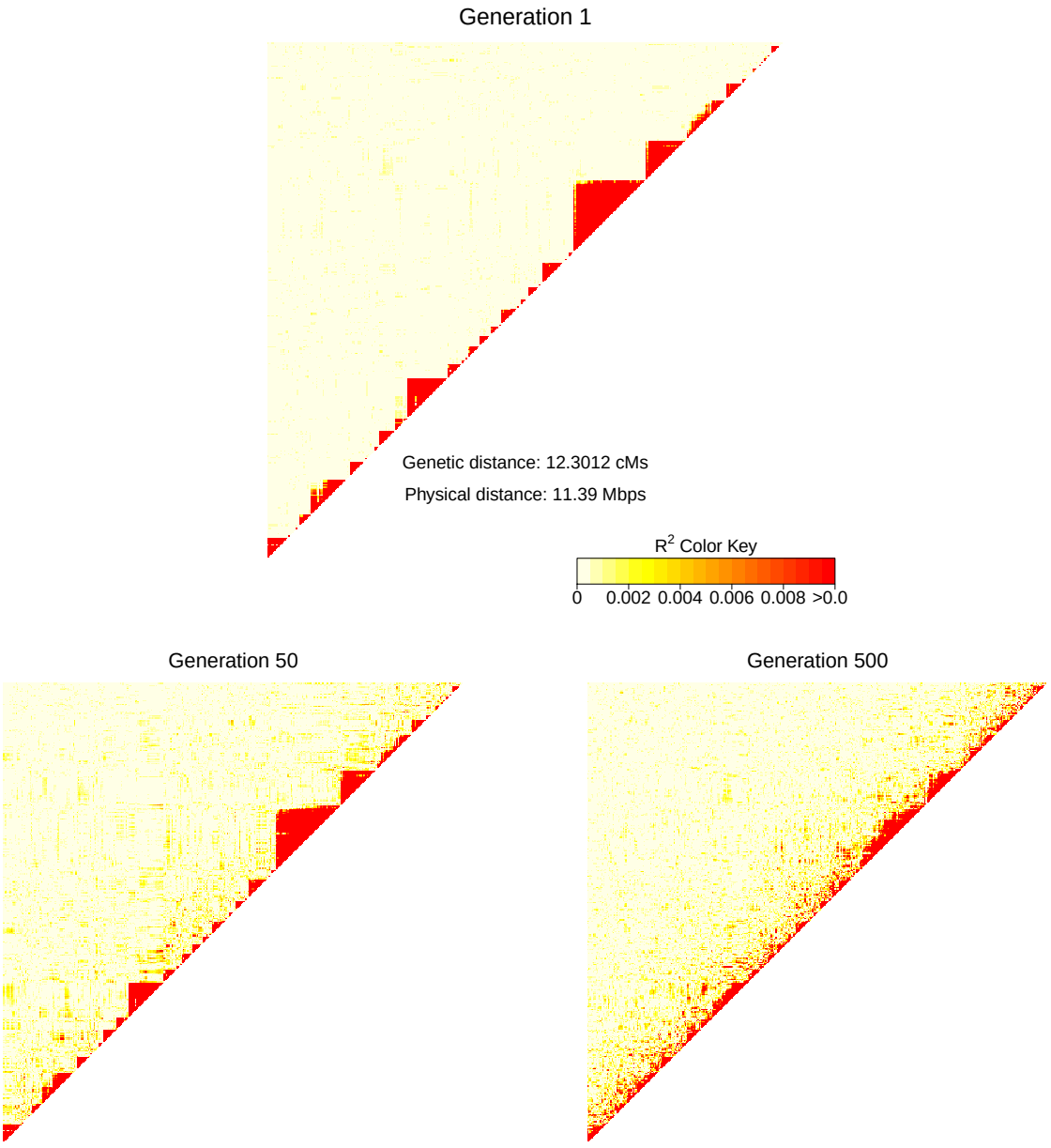



Figure 2.3: LD on a segment of simulated chromosome, measured by pairwise $R^2$ at generation 1, 50, and 500. Scale 0 to 1.

allelic heterogeneity. Biologically, variants would occur at different locations within the same functional gene. For the SCV trait, at the selected trait locus the minor allele was selected as the causal variant. Trait expression is again a function of the number of causal copies the individual has. In the SCV case the causal variants all arise on a single haplotypic background at the functional gene. Biologically, this could happen if a mutation ocurred many generations ago and has been inherited by many individuals in the current population. The SNP marker is in or close to the functional gene.

Given the number of copies of the causal allele, the following trait models were used to determine affected status. There is a single trait locus at a position 31cM along the chromosome, with a disease and neutral allele $D$ and $N$ with $P(D) = 0.18$. A binary trait is simulated with penetrances $P(Y = 1|DD) = 0.9$, $P(Y = 1|ND) = 0.3$ and $P(Y = 1|NN) = 0.1$. The overall disease prevalence is $P(Y = 1) = 0.2$. The MRV and SCV traits differ in the determination of trait alleles. For the MRV trait a sample of FGLs is taken from the FGLs present in generation 1 of the population. If a haplotype has one of the sampled FGLs at the trait locus, it has allele $D$. The allelic types of markers are not taken into account, so there will be no association between marker alleles and trait alleles at the first generation of the population. For the SCV trait, the SNP allele at the trait locus is used. In the marker data set there will be a marker allele that is perfectly correlated with the unobserved trait allele. This represents an ideal situation where there happens to be a marker at the trait locus.

### *2.3 Data Set: Simulated Pedigrees*

A set of four-generation sub-pedigrees were extracted from generations 47-50 of the simulated population. These pedigrees are for use in pedigree-based analyses, and as they come from the larger population the pedigree members also have population-level relationships.

A large set of potential sub-pedigrees were identified by tracing all the descendants from each individuals in generation 47. This gave 20,000 potential pedigrees, with mean size of 22 individuals. Of these, pedigrees that did not have individuals in all four generations

were eliminated, leaving 14,937 potential pedigrees. Identical pedigrees, such as cases where the male and female of a sib-pair had identical descendants, were also eliminated leaving 11,783 potential pedigrees. Next, a smaller set of pedigrees was sampled such that none of the pedigrees had individuals in common at the 50th generation. The sampling was done by starting with an empty set and, in random order, adding a pedigree if there was no intersection in individuals at generation 50 with pedigrees already in the set and rejecting the pedigree otherwise. The sampling resulted in a set of 1902 pedigrees. These pedigrees contained 38123 unique individuals over all 1902 pedigrees. The average size of pedigrees was 20.58, with 6.95 members in the final generation. The distributions of the size of the reduced list of pedigrees is given in Figure 2.4. There were 14 looped pedigrees and two pedigrees with a full sibling marriage.

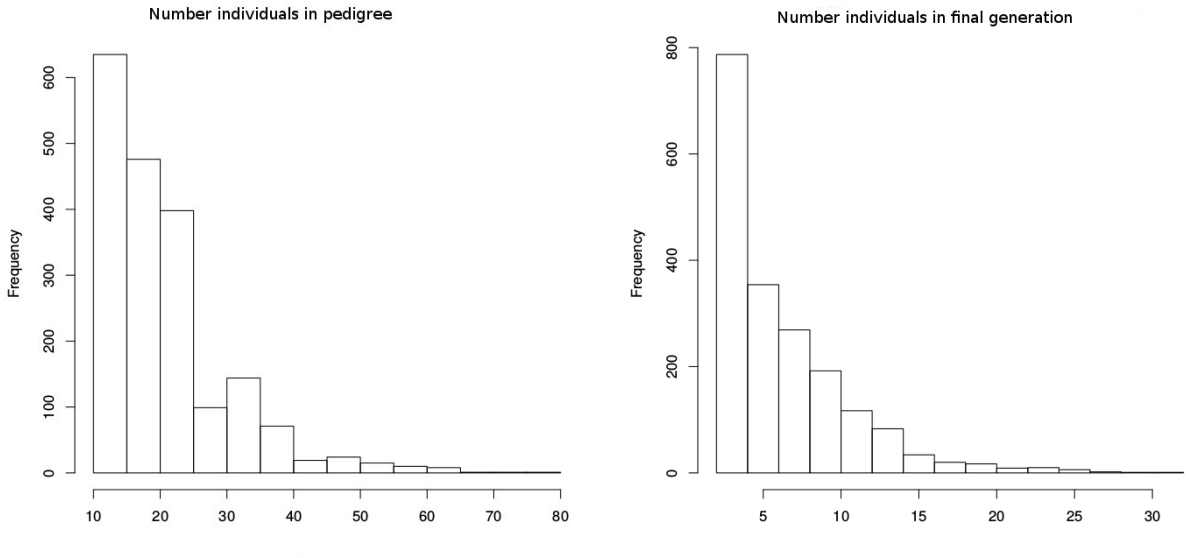



Figure 2.4: Total number of individuals and number of individuals in final generation of pedigrees

From the list of 1902 pedigrees a set of 2 and a set of 4 pedigrees were selected to test the merging procedure described in Chapter 4, the results of which are given in Chapter 5. A

simple random sample of 50 pedigrees were inspected to select these sets. The requirements were that the pedigrees should be of similar size so that one does not dominate the LOD score, to be of sufficient size with enough meioses to detect a linkage signal and have IBD between the pedigree founders.

The first set of pedigrees was "Merge2": a set of 2 pedigrees where one founder in each pedigree make up a pair of siblings. The close relationship between the founder siblings should allow a large increase in linkage signal if the sibling IBD is detected and used effectively. The two pedigrees in Merge2 are plotted in Figure 2.5. The kinships between all the individuals in the Merge2 pedigrees calculated by nominal pedigree relationship are compared to the observed IBD sharing on the chromosome in Figure 2.6. Figure 2.6 shows that the IBD sharing observed in the simulated data is greater than the pedigree relationships indicate, with IBD sharing between the nominally separate pedigrees. The excess IBD observed that is not explained by the pedigree is what we hope to identify and incorporate in the merging procedure. The second set was "Merge4", a set of 4 pedigrees that did not have any pedigree relationships between the founders within the 4 generations of the population. To select the four pedigrees, the FGLs present at the sparse markers were inspected. The four pedigrees had regions along the chromosome where there were FGLs present in more than one family, indicating IBD between pedigree founders. The pedigrees selected were all a similar size and moderately large, around 25 individuals per pedigree.

Note that in the extracted pedigrees there are multiple marriages and individuals with a large number of offspring. The population simulation described in Section 2.1.2 was designed to limit this. The observed patterns are an artifact of sampling, as individuals with large numbers of offspring are more likely to have descendants in the final generation.

### *2.3.1 Quantitative Trait Simulation*

Rare and common variant trait models are described in Section 1.6.1 and were applied to the simulated sib-pair data to simulate a binary trait in Section 2.2.1. For the simulated pedigree data, a quantitative trait with both rare and common causal variants was simulated.

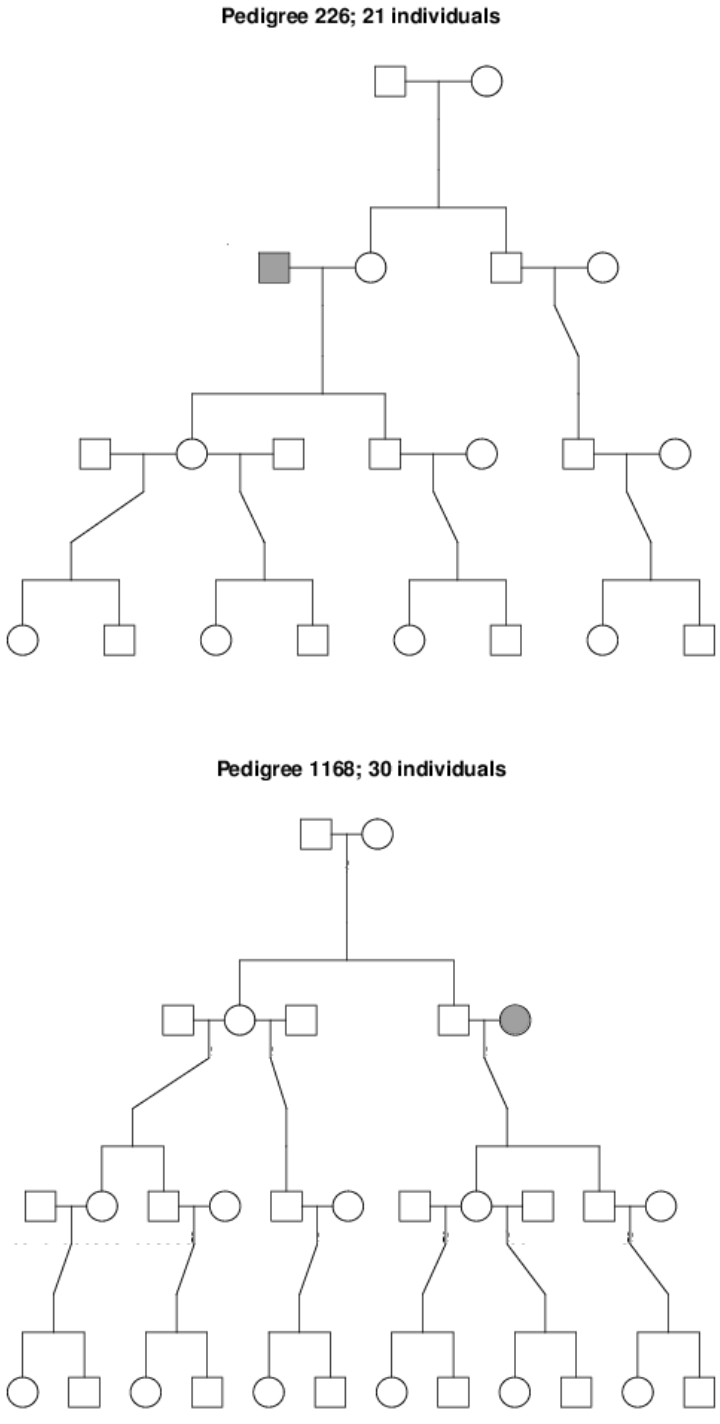



Figure 2.5: The two pedigrees in the "Merge2" set. Individuals 964612 and 964611 are siblings, filled in grey.

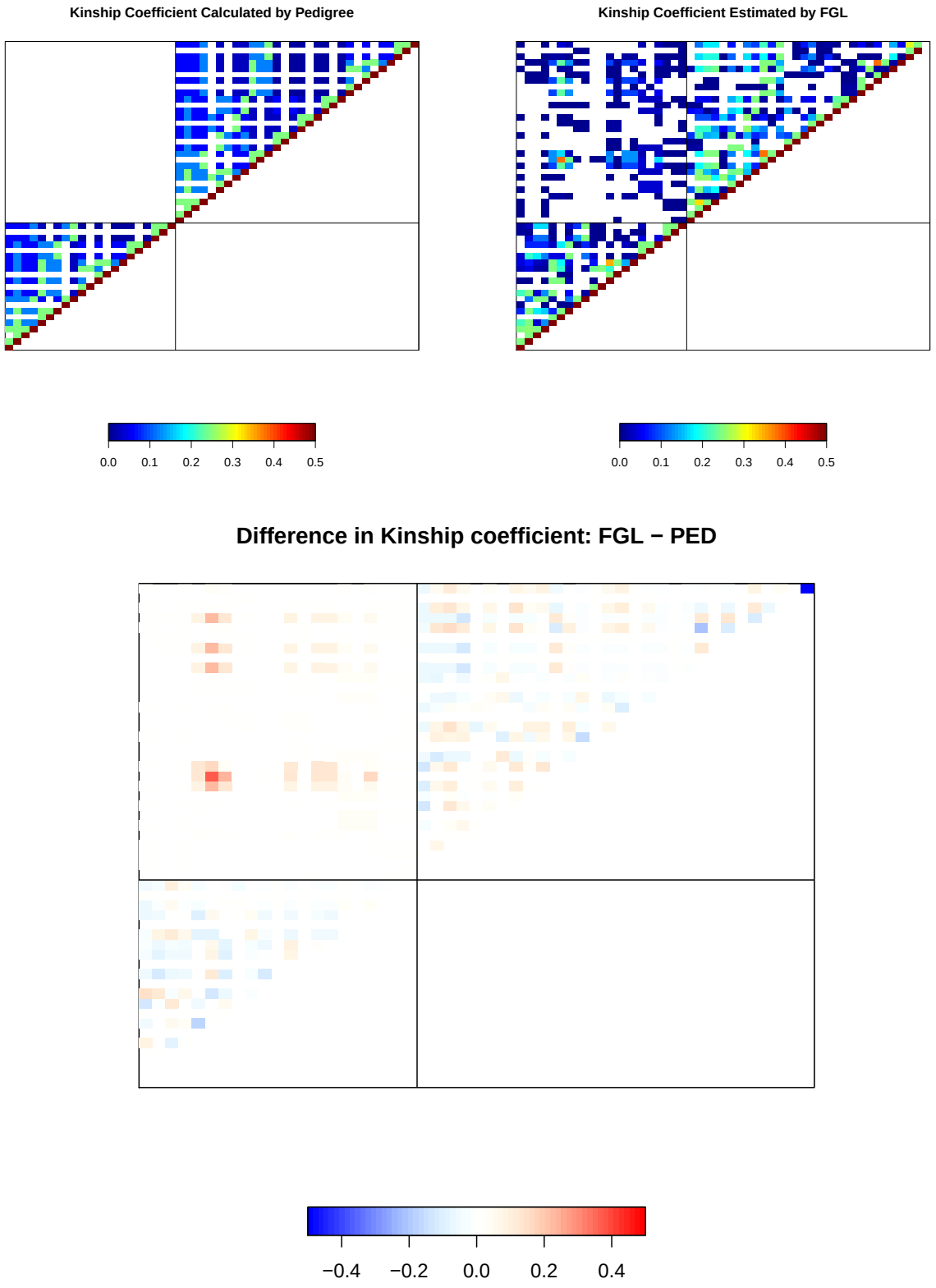



Figure 2.6: Comparison of kinship coefficients from pedigree relationship (PED) and observed IBD sharing (FGL)

At the trait locus, founders that would have a disease-causing allele $D$ were selected by sampling FGLs from all unique FGLs in the pedigrees at the trait locus. A total 30% of chromosomes had a causal allele, $P(D) = 0.3$. An exact test for Hardy-Weinberg equilibrium [Levene, 1949, Haldane, 1954] showed no evidence that the distribution of trait genotypes was different to Hardy-Weinberg equilibrium.

Trait values were normally distributed with means that depend on trait genotypes. The means were $E(\mathbf{Y}_i|\mathbf{t}_i = DD) = 2.46, E(\mathbf{Y}_i|\mathbf{t}_i = ND) = 0.82, E(\mathbf{Y}_i|\mathbf{t}_i = NN) = 0.27$ and the variance is 1. If the 0.9 percentile is used to discretize the trait into affected $\mathbf{y}'_i = 1$ and unaffected $\mathbf{y}'_i = 0$, then $P(\mathbf{Y}_i = 1|\mathbf{t}_i = DD) = 0.69, P(\mathbf{Y}_i = 1|\mathbf{t}_i = ND) = 0.13$ and $P(\mathbf{Y}_i = 1|\mathbf{t}_i = NN) = 0.05$. Thus, if discretized, the model corresponds to a relative risk of 3 per copy of the causal allele. Given affected status, the discrimination among genotypes is $P(\mathbf{t}_i = DD|\mathbf{y}_i = 1) = 0.20, P(\mathbf{t}_i = ND|\mathbf{y}_i = 1) = 0.56$ and $P(\mathbf{t}_i = NN|\mathbf{y}_i = 1) = 0.24$.

### *2.4 Data Set: GAW19 Pedigrees*

Another data set used in this thesis was a set of large pedigrees from the Genetic Analysis Workshop 19 (GAW19). A merging analysis for this dataset was published in Saad et al. [2016], and is described and updated in Chapter 5, Section 5.1.

For GAW19, the dataset provided was from the previous workshop, GAW18, and is described in Almasy et al. [2014]. In GAW18, genetic marker data came from the Type 2 Diabetes Genetic Exploration by Next-generation sequencing in Ethnic Samples (T2D-GENES) consortium [T2D], an international effort to identify genes influencing susceptibility to type 2 diabetes. T2D-GENES Project 2 is a study based on complex pedigrees, designed to identify low frequency or rare variants influencing susceptibility to type 2 diabetes using information from whole genome sequencing. The T2D-GENES Project 2 pedigrees contain 1043 individuals in 20 pedigrees from the San Antonio Family Studies (SAFS). SAFS families were obtained by sampling Mexican Americans aged 40 to 60 at random from low-income Mexican American census tracts in San Antonio, Texas, and recruiting their spouses and relatives.

The GAW18 data set contains real pedigree structures from the 20 T2D-GENES SAFS pedigrees. Genetic marker data were almost 500,000 SNP genotypes across the genome. All 959 individuals were genotyped. SNP panels, marker selection and quality control were described in Almasy et al. [2014]. The GWAS marker loci were very dense, on chromosome 3 there were 48,892 SNPs over 225cM, an average intermarker distance of 0.0046cM. The sparse set of loci selected by Blue et al. [2014] for linkage analysis was a subset of 351 SNPs with average intermarker distance 0.62cM. Marker positions were taken from the Rutgers sex-averaged interpolated positions of dbSNP Build 134 Matise et al. [2007]

Simulated trait data was available on all individuals. The trait data were 200 replicates of a simulated diastolic blood pressure phenotype. The trait simulation model was which based on real blood pressure distributions, frequency of hypertension, medication use, and tobacco smoking taken from the SAFS data. There are 55 causal variants across the genome, the largest of which is MAP4. MAP4 is at 69cM on chromosome 3 and explains 2.3% of the variance in the trait. Trait data on each individual was adjusted for age, sex and use of anti-hypertensive medications.

Figure 2.7 is the pedigree structures for the seven GAW pedigrees. Colors of individuals indicate the mean trait value over the trait replicates. On the vertical scale, the expected trait values for zero, one, or two copies of the causal SNP allele under the trait model are indicated by AA, Aa and aa respectively. Individuals with labels highlighted in yellow form the set selected in the merging analysis by Saad et al. [2016], details of which can be found in Chapter 6. There are 3 individuals highlighted in pedigree 5, one in pedigree 6, one in pedigree 7, one in pedigree 8, 6 in pedigree 10, 9 in pedigree 21, and one in pedigree 25 giving a total of 22 individuals.

## *2.5 Data Set: Alzheimer's Disease Pedigrees*

The real data set used to demonstrate the merging algorithm in Chapter 6 is composed of two pedigrees from an isolated European population with individuals affected by Alzheimer's disease. The data came from the Alzheimer's Disease Sequencing Project (ADSP) [ADSP,

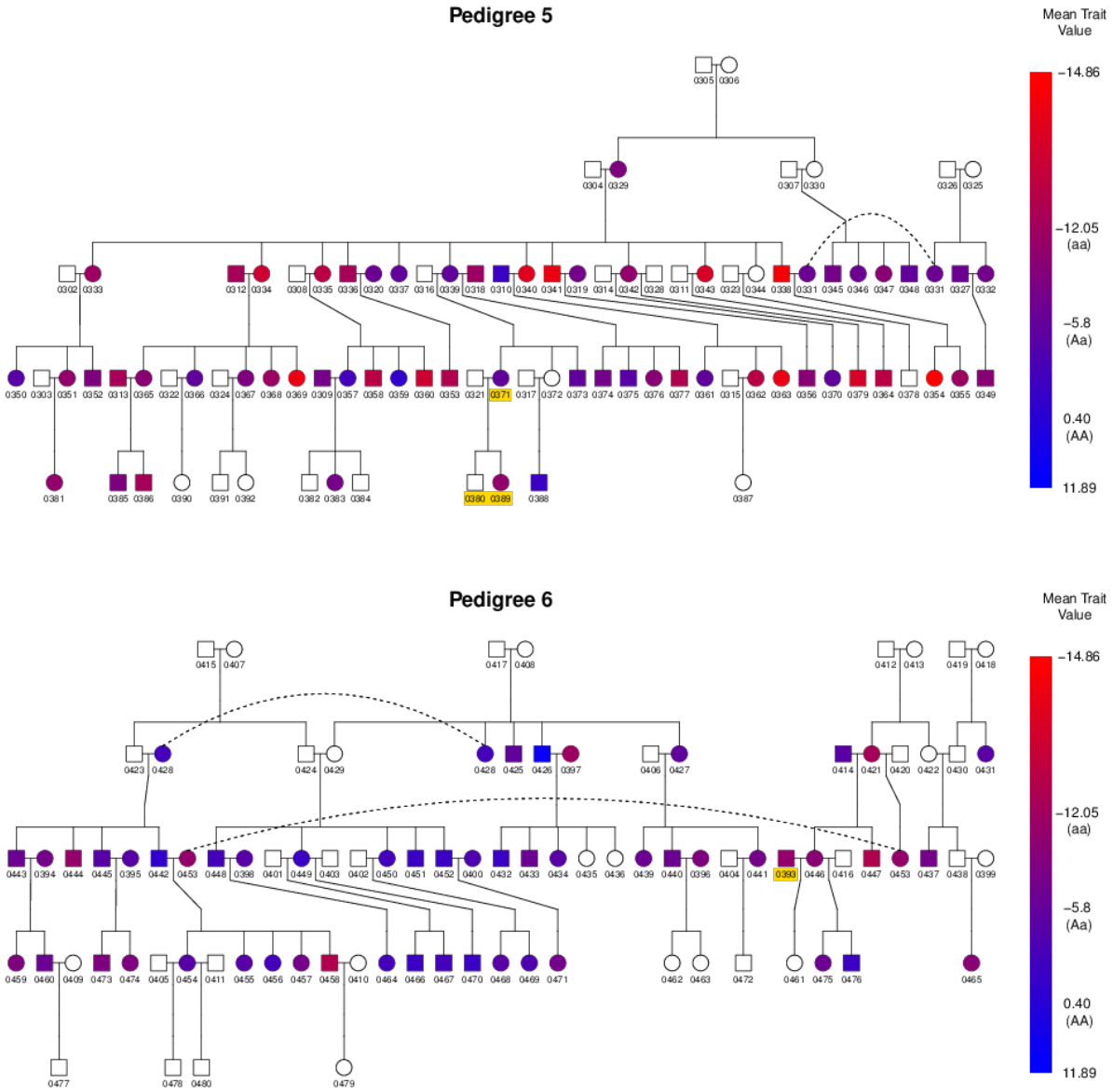



Figure 2.7: GAW pedigrees; Colors represent observed trait values, yellow highlighted individuals were selected for merging.

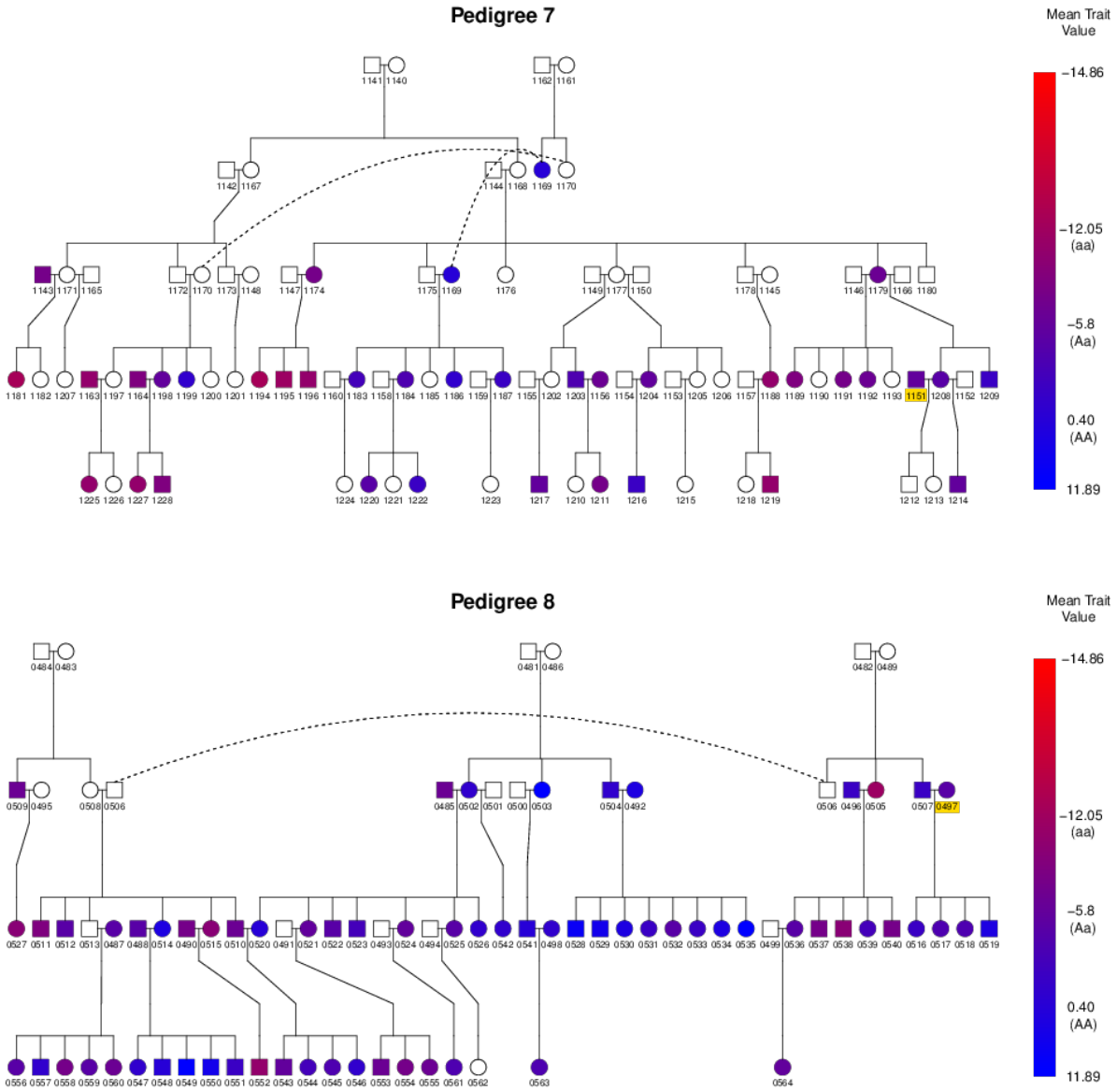



Figure 2.7: GAW pedigrees; Colors represent observed trait values, yellow highlighted individuals were selected for merging.

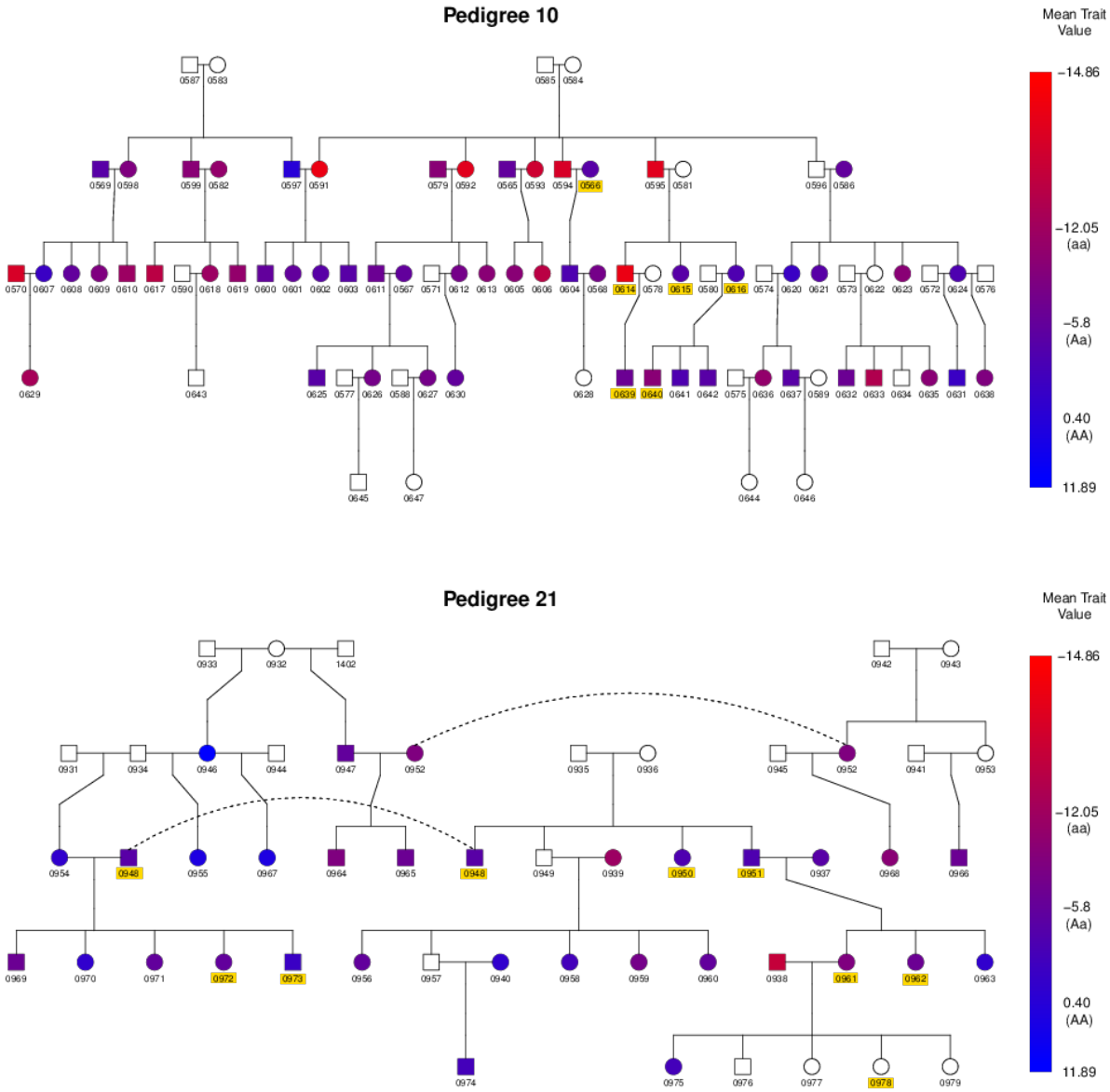



Figure 2.7: GAW pedigrees; Colors represent observed trait values, yellow highlighted individuals were selected for merging.

2016] and Erasmus Rucphen Family study (ERF) [ERF, 2016].

The ERF study was of an isolated population in the south-west Netherlands. The pedigree relationships were determined by tracing the genealogical records of 103 probable Alzheimer's patients. A genealogy containing 4645 members over 18 generations was first presented in Liu et al. [2007]. The genealogy was traced through church records of 22 families that had at least five children baptized in the community church between 1850 and 1990. All living descendants of these couples and spouses were invited to take part in the study. Data collection on 2065 individuals was performed from June 2002 to February 2005. Genotyping protocols for SNP data are described in Liu et al. [2007].

The data set used in this thesis is a small part of the ERF study, consisting of two pedigrees named ERF201 and ERF203. The two pedigree structures are disjoint - pedigree plots are given for ERF201 in Figure 2.8 and ERF203 in Figure 2.9. The full genealogy that relates the founders of each pedigree is shown in 2.10 sourced from Liu et al. [2007], but was not used in the analysis. Figures were created using pedfiddler 0.6.1 [PED, 2010]. Individuals are classified as either affected with the Alzheimer's disease trait or of unknown status. Unknown status is appropriate for Alzheimer's disease as it is a late onset onset disease and may not yet be presenting at the time of death or observation. Trait status is indicated by blue (affected) or yellow (unknown) in the pedigree plots.

Multilocus marker genotype data over all 22 autosomes are available on the individuals in each pedigree who are underlined in Figures 2.8 and 2.9. There are 2 individuals in ERF201 and 22 individuals in ERF203 with SNP data. There are a total of 5 individuals in ERF201 and 30 individuals in ERF203 that are affected with the trait. The individuals with SNP data are all affected. Details of the SNP panels chosen are given in the data analysis section 6.2. The data has already been through standard quality control procedures and used in other analyses. Liu et al. [2007] presented a genome-wide linkage analysis on all ERF pedigrees, identifying regions on chromosomes 1, 3, 10 and 11 as significant for linkage. A linkage analysis has also been performed by the Wijsman group that is described in Chapter 6. Population allele frequencies were obtained from the 1000 Genomes data set,

using the European samples.

The pedigree-based kinship between the genotyped individuals in ERF201, 201_46 and 201_44, is 0.0078125. Kinships for individuals in ERF203 are in Table 2.6. The closest relationships are between two sibling pairs with a kinship of 0.25 and two avuncular pairs with kinships 0.12.

| | 167 | 5 | 175 | 176 | 186 | 169 | 188 | 168 | 173 | 166 | 181 | 174 | 184 | 189 | 180 | 172 | 177 | 183 | 171 | 182 | 170 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 203_178 | 0.016 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 203_167 | | 0.031 | 0.008 | 0.008 | 0.008 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 203_5 | | | 0.031 | 0.031 | 0.031 | 0.016 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.004 | 0.000 |
| 203_175 | | | | 0.250 | 0.062 | 0.016 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 203_176 | | | | | 0.062 | 0.016 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 203_186 | | | | | | 0.016 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.016 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 203_169 | | | | | | | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 203_188 | | | | | | | | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.008 | 0.016 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 203_168 | | | | | | | | | 0.016 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.016 | 0.000 | 0.000 | 0.000 | 0.000 |
| 203_173 | | | | | | | | | | 0.016 | 0.016 | 0.031 | 0.000 | 0.000 | 0.000 | 0.000 | 0.016 | 0.000 | 0.000 | 0.000 | 0.000 |
| 203_166 | | | | | | | | | | | 0.250 | 0.031 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 203_181 | | | | | | | | | | | | 0.031 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 203_174 | | | | | | | | | | | | | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 203_184 | | | | | | | | | | | | | | 0.062 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.004 | 0.000 |
| 203_189 | | | | | | | | | | | | | | | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.004 | 0.000 |
| 203_180 | | | | | | | | | | | | | | | | 0.125 | 0.016 | 0.000 | 0.031 | 0.016 | 0.000 |
| 203_172 | | | | | | | | | | | | | | | | | 0.000 | 0.000 | 0.062 | 0.031 | 0.000 |
| 203_177 | | | | | | | | | | | | | | | | | | 0.016 | 0.000 | 0.000 | 0.000 |
| 203_183 | | | | | | | | | | | | | | | | | | | 0.000 | 0.008 | 0.000 |
| 203_171 | | | | | | | | | | | | | | | | | | | | 0.125 | 0.000 |
| 203_182 | | | | | | | | | | | | | | | | | | | | | 0.008 |
| 203_170 | | | | | | | | | | | | | | | | | | | | | |

Table 2.6: Kinships calculated from pedigree relationship between genotyped individuals in ERF201 pedigree

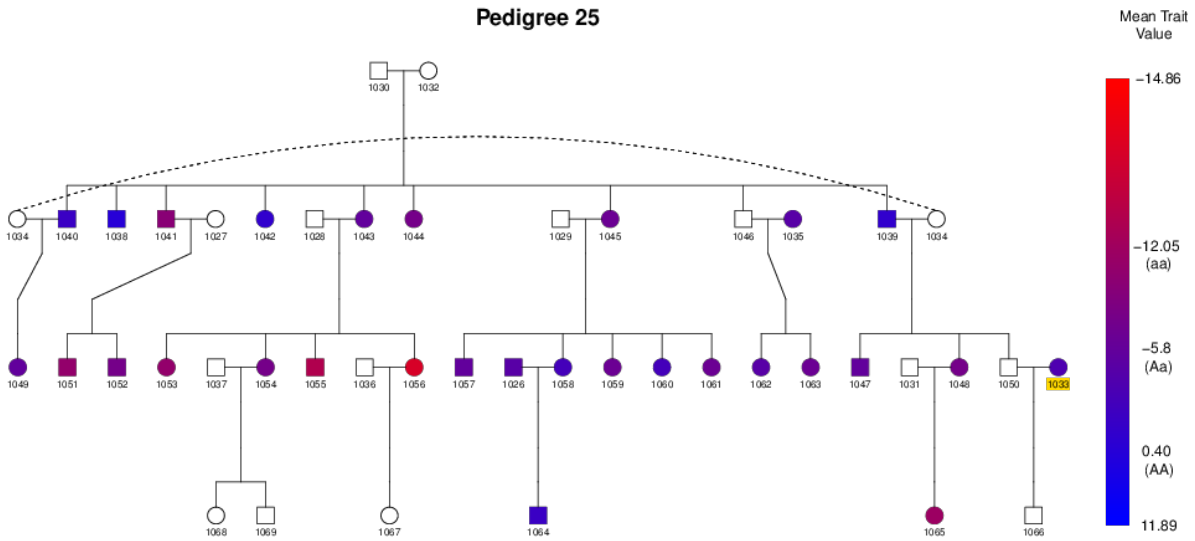



Figure 2.7: GAW pedigrees; Colors represent observed trait values, yellow highlighted individuals were selected for merging.

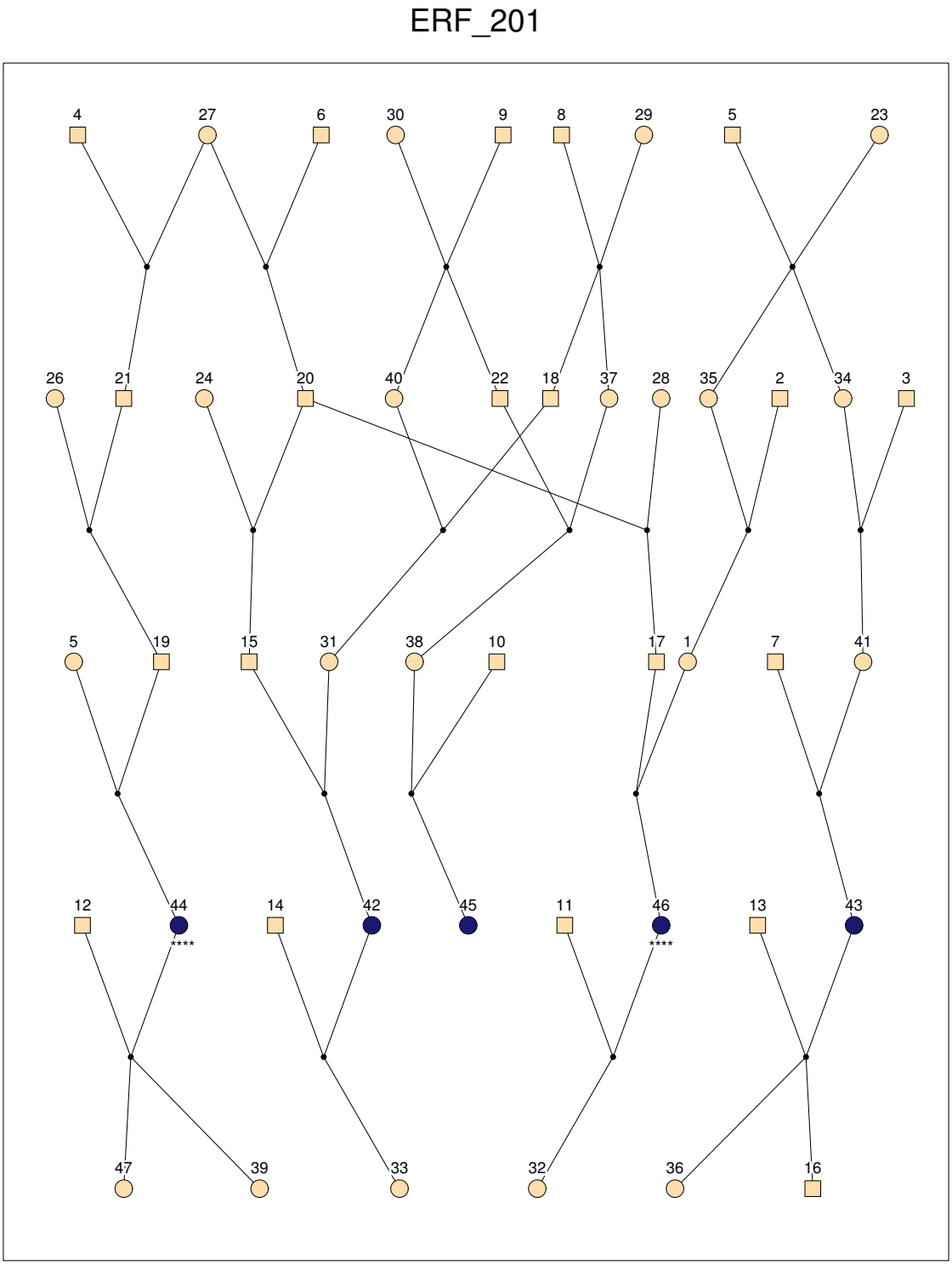



Figure 2.8: The ERF_201 Pedigree. Affected individuals are blue; genotyped individuals are underlined.

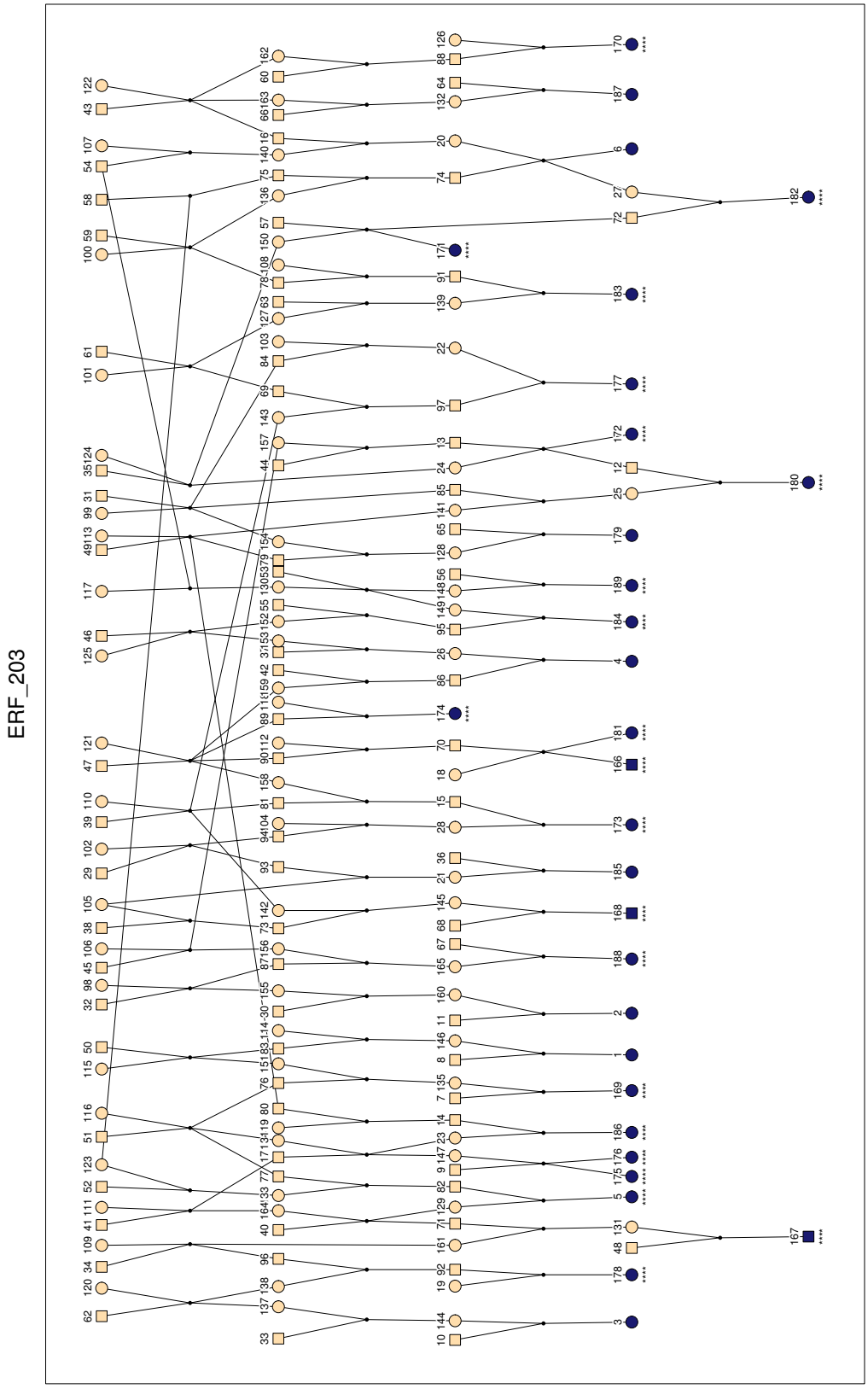



Figure 2.9: The ERF_203 Pedigree. Affected individuals are blue; genotyped individuals are underlined.

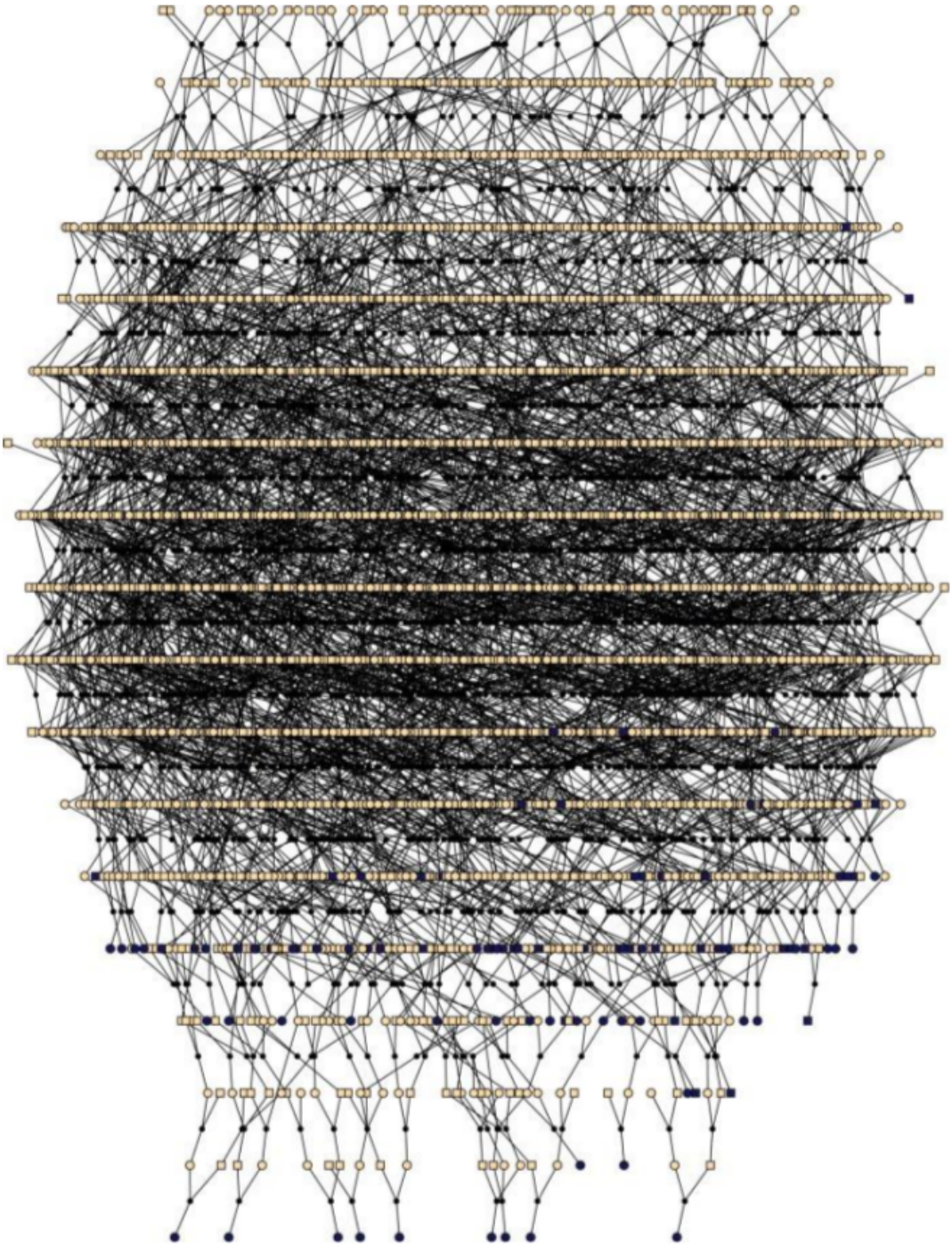


Figure 2.10: The full ERF pedigree from which ERF201 and ERF203 were drawn.

# Chapter 3

# IBD ESTIMATION AND TRAIT LOCUS DETECTION IN SIBLINGS

In this chapter genetic data on siblings is used for IBD estimation and detection of trait loci. Section 3.1 deals with IBD estimation. A Markov model for IBD in siblings with related parents is developed in Section 3.1.2. The method is applied to IBD inference using both an HMM and a composite likelihood model, described in Section 3.1.3, and results for simulated data are given in Section 3.1.4. In Section 3.2 tests for trait locus detection are compared on the simulated sibling data. Both family-based and population-based association tests are compared to IBD-based methods.

## *3.1 IBD estimation*

IBD estimation in a pair of siblings is similar to IBD estimation on a pedigree. Multilocus marker data $\mathbf{g}$ is observed on two individuals $\mathcal{G}$. The pedigree structure observed is on the set $\mathcal{F}$ that comprises the siblings and their two parents. Estimation of IBD relative to the founders of $\mathcal{F}$, the parents, is described in Section 3.1.1. Estimation of IBD relative to the founders of a larger population $\mathcal{P}$ that encompasses $\mathcal{F}$ is described in Section 3.1.2.

### *3.1.1 IBD in Sibs Relative to Parents*

Estimating IBD in siblings relative to their parents is a special case of IBD estimation in a pedigree, described in Section 1.5.1 based on the IBD model described in Section 1.2.1. For sibs the mapping between inheritance vectors and FGLs, $\phi(\mathbf{v}) = \mathbf{x}$, is one-to-one. Over the 4 chromosomes in the sib pair there are $2^4 = 16$ possible inheritance vectors at each locus. As each inheritance indicator points to one of two possible founder genomes, each IV gives a

specific one of the 16 possible FGL vectors. The IVs are modeled as having Markov dependence over loci, as in Equation (1.3). In this case, the FGLs also have Markov dependence, as the 1-1 mapping guarantees that there are no pooled states with different transition probabilities. Also, in the four chromosomes, recombinations along each chromosome are Markov and a recombination in any chromosome changes the FGL state. Therefore,

$$P(\mathbf{X} = \mathbf{x}) = \prod_j P(\mathbf{X}_{j,1} = \mathbf{x}_{j,1}) \prod_{l>1} P(\mathbf{X}_{j,l} = \mathbf{x}_{j,l} | \mathbf{x}_{j,l-1}), \tag{3.1}$$

where $j = 1, ..., 4$ are the ordered DNA copies of the siblings and $l = 1, ..., L$ are loci. The transitions between FGLs are modeled as in Equation (1.4),

$$P(\mathbf{X}_{j,l} \neq \mathbf{x}_{j,l-1}) = \theta_l = 0.5(1 - e^{-0.02 d_l}), \tag{3.2}$$

where $d_l$ cM is the distance between markers $l-1$ and $l$. The transition between FGL vectors is therefore

$$P(\mathbf{X}_{.,l} = \mathbf{x}_{.,l} | \mathbf{x}_{j,l-1}) = \theta_l^{r_l} (1 - \theta_l)^{n_l} \tag{3.3}$$

where $r_l$ is the number of chromosomes that change state and $n_l$ is the number that do not change state between loci $l-1$ and $l$. The possible FGL states, if converted to canonical labeling as in Equations (1.6) and (1.7), collapse to four IBD states. Refer to Appendix A for IBD labeling descriptions. The four states are $\mathbf{S}^{\mathcal{F}} \in \{1212, 1213, 1232, 1234\}$. The collapsed states retain the Markov property as the states that are combined each have the same transition probabilities, and a recombination in any of the chromosomes will still change the $\mathbf{S}^{\mathcal{F}}$ state. The states collapse further to three genotypically equivalent IBD states which are $\tilde{\mathbf{S}}^{\mathcal{F}} \in \{$sharing 0, 1 or 2 copies IBD$\}$. For dense markers, a first order approximation to the IBD transition matrix for genotypically equivalent IBD states is

$$P = \begin{array}{c} \\ \tilde{S}_{l-1} = 0 \\ \tilde{S}_{l-1} = 1 \\ \tilde{S}_{l-1} = 2 \end{array} \begin{array}{c} \begin{array}{ccc} \tilde{S}_l = 0 & \tilde{S}_l = 1 & \tilde{S}_l = 2 \end{array} \\ \begin{bmatrix} \exp(-0.04 d_l) & 1 - \exp(-0.04 d_l) & 0 \\ 0.5(1 - \exp(-0.04 d_l)) & \exp(-0.04 d_l) & 0.5(1 - \exp(-0.04 d_l)) \\ 0 & 1 - \exp(-0.04 d_l) & \exp(-0.04 d_l) \end{bmatrix} \end{array}, \tag{3.4}$$

where $S$ indicates the number of copies shared. The transitions in IBD states $\tilde{\mathbf{S}}^{\mathcal{F}}$ are Markov because the two states 1213 and 1232 that combine to form $S = 1$ have the same transition probabilities. A Markov chain with transition matrix (3.4) has stationary distribution $(0.25, 0.5, 0.25)$ which are the overall IBD sharing probabilities for siblings.

### *3.1.2 IBD in Sibs Relative to Population*

The IBD state in the siblings relative to the founders of $\mathcal{P}$ can be conditioned on the IBD state of the parents as in Equation (1.5). The FGLs on the sibs relative to the founders of $\mathcal{P}$, denoted $\mathbf{x}^{\mathcal{P}}_{sibs}$, is a function of the inheritance vectors from the parents to the sibs ($\mathbf{v}_{sibs}$) and the FGL in the parents relative to the population ($\mathbf{x}^{\mathcal{P}}_{pars}$). That is, $\phi(\mathbf{v}_{sibs}, \mathbf{x}^{\mathcal{P}}_{pars}) = \mathbf{x}^{\mathcal{P}}_{sibs}$, the inheritance indicators for each sib DNA copy point to the FGL of the population founder that supplied the DNA in the parent. For a particular sib DNA copy $j$ at locus $l$, if the DNA copy is the paternal copy of the individual, it inherited DNA from either the father's paternal copy $P(j)$ or father's maternal copy $M(j)$, so $\mathbf{x}^{\mathcal{P}}_{jl} = \mathbf{x}^{\mathcal{P}}_{P(j)l}$ if $\mathbf{v}_{jl} = 0$ and $\mathbf{x}^{\mathcal{P}}_{jl} = \mathbf{x}^{\mathcal{P}}_{M(j)l}$ if $\mathbf{v}_{jl} = 1$.

The model for IVs is Markov along the chromosome, as described in Section 1.2.1, Equation (1.3), and specifically for sib pairs in Section 3.1.1. The model for FGLs in the parents relative to the population founders can also be modeled as Markov along the chromosome, as seen in IBD inference when the pedigree structure is unknown, in Section 1.5.2. The Brown et al. [2012] CRP Markov model is used. There are 16 possible states for $\mathbf{v}_{sibs}$ and 15 canonically labeled FGL states for $\mathbf{x}^{\mathcal{P}}_{pars}$. In total there are $16 \times 15 = 240$ combinations. The transition matrix between these combinations can be formed by a Kronecker product of the IV and CRP transition matrices because the state of the parents is independent of the state of the siblings given the parents, as per Equation (1.5). The 240 combinations map to 15 canonically labeled FGL states under $\phi(\mathbf{v}_{sibs}, \mathbf{x}^{\mathcal{P}}_{pars}) = \mathbf{x}^{\mathcal{P}}_{sibs}$. The canonically labeled FGL states are equivalent to the 15 IBD states on 4 chromosomes.

The equilibrium distributions for the 15 IBD states among 4 chromosomes are plotted in Figure 3.1. The CRP is the Brown et al. [2012] model for IBD transitions in individuals with

no known pedigree structure, and represents $\mathbf{x}^{\mathcal{P}}_{pars}$. The distribution for sibs with unrelated parents represents $\mathbf{x}^{\mathcal{F}}_{sibs}$. The distribution for siblings with related parents is formed by combining the two processes as described above. The probability of being in state 15 (no IBD) is reduced in the combined process.

### *3.1.3 IBD Inference*

Inference on IBD states between a pair of siblings can be performed with a Hidden Markov Model (HMM). The hidden states are FGLs, $\mathbf{x}^{\mathcal{F}}$ if the IBD is relative to the parents, or $\mathbf{x}^{\mathcal{P}}$ if the IBD is relative to the population. The transitions between hidden states are a Markov chain over the marker loci, with transition matrices described in Section 3.1.1 and 3.1.2 respectively. The emissions are the genotypes at each loci. Emissions probabilities are modeled by $P(\mathbf{g}|\mathbf{x})$, described in Section 1.3.

An alternative to the HMM is a composite likelihood model (CM). Composite likelihoods [Lindsay, 1988] are a general class of pseudo-likelihoods based on marginal or conditional probabilities. A composite likelihood is formed by multiplying a collection of component likelihoods, where each component is a conditional or marginal density. Composite likelihoods have been used in many clustered or spatially dependent data applications [Varin et al., 2011]. They have also been used in many statistical genetics applications [Larribe and Fearnhead, 2011]. Composite likelihoods have been used to estimate parameters that vary along the chromosome, including the recombination rate [McVean et al., 2004] and of LD [Fearnhead and Donnelly, 2001]. We consider the composite likelihood of the FGL state at a target locus, a single marker at which IBD is to be estimated. The composite likelihood is constructed under the modeling assumption that the FGL state at each marker locus is independent, conditional on the target locus. This independence is not the case in practice, as FGL states at adjacent loci are dependent. The likelihood of the FGL state at the target locus, $\mathbf{x}_0$ is

$$L(\mathbf{x}_0) = \sum_{\mathbf{x}_{\cdot 1}} \cdots \sum_{\mathbf{x}_{\cdot L}} \left[ P(\mathbf{g}_{\cdot 0}|\mathbf{x}_{\cdot 0}) \prod_l P(\mathbf{g}_{\cdot l}|\mathbf{x}_{\cdot l}) P(\mathbf{x}_{\cdot l}|\mathbf{x}_{\cdot 0}) \right] \tag{3.5}$$

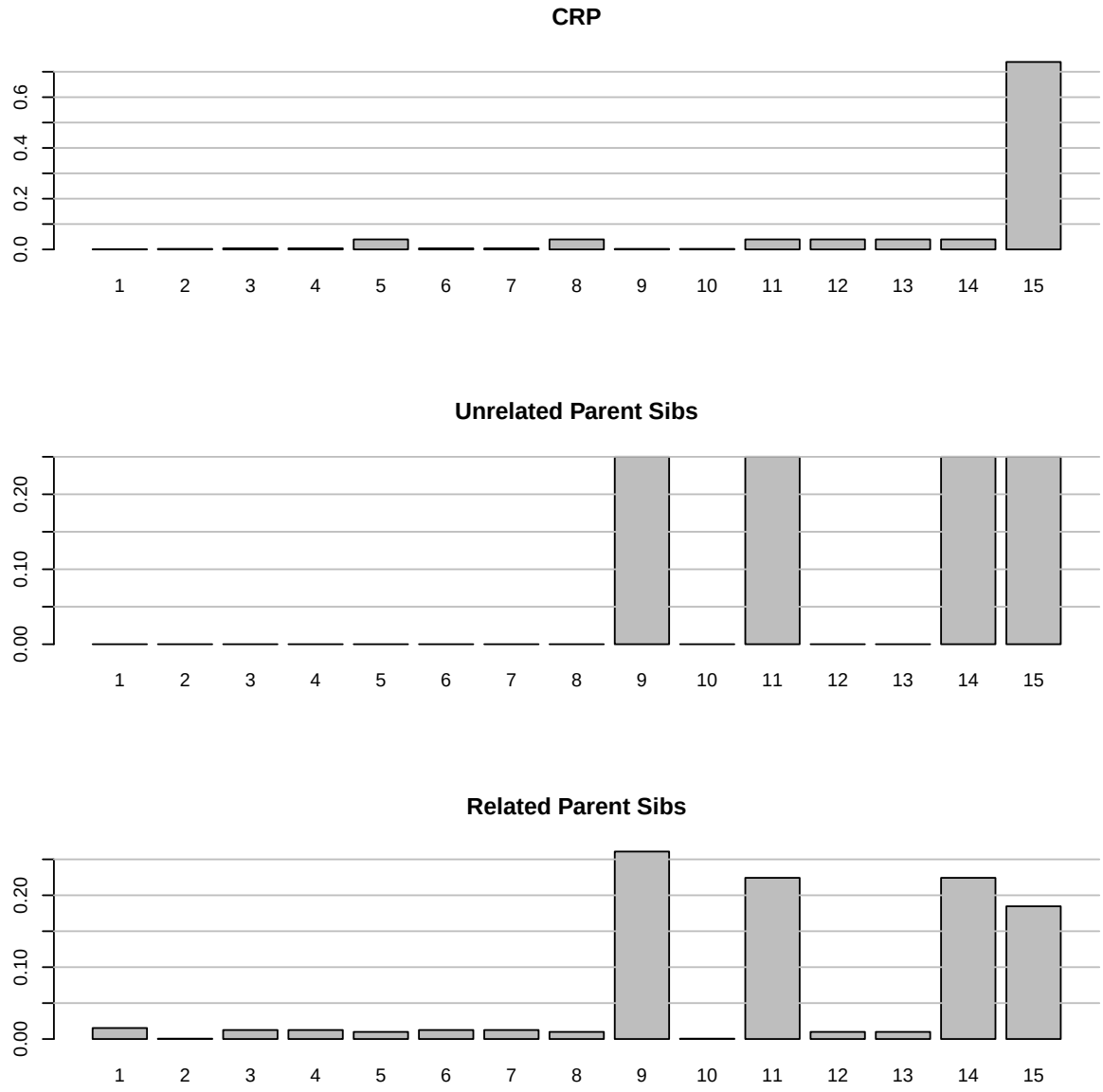



Figure 3.1: Equilibrium distributions for $\alpha = 0.2$, $\beta = 0.05$.

where $l = 1, ..., L$ are loci in a window around the target locus, and $l = 0$ is the target locus.

The properties of composite likelihood methods and the estimators they define are well known, see [Chandler and Bate, 2007] for example. Maximum likelihood estimates from

conditional likelihoods are unbiased if the marginals are correctly specified. However, as components are multiplied regardless of independence, the composite likelihood has the behavior of a mis-specified model and there is a loss of efficiency compared to the correct specification of the full likelihood. Composite methods have some advantages over methods that seek to explicitly model the dependence in the data. One motivation for their use is often computational - by using marginal distributions we avoid computing or modeling the complex joint distribution. Another motivation is a notion of robustness to misspecification of the full joint distribution. In this case, robustness to LD that is not accounted for in the joint distribution.

Computationally, the calculation of the composite likelihood can be more efficient than the HMM as the probabilities $P(\mathbf{x}_{.l}|\mathbf{x}_{.0})$ depend only on the distance to the target locus. Computations of state probabilities for composite likelihood are of the order $O(sn)$ where $s$ is the number or hidden states and $n$ is the number of loci. For the HMM the complexity is $O(s^2n)$. In the HMM the posterior probabilities at all loci are calculated in the forward-backward algorithm. In the CM, however, the posterior probability is only calculated at one target locus and a new calculation must be performed for a new target. The HMM and the Composite likelihoods are compared in Section 3.1.4 on the simulated sib pair data set.

In terms of robustness, for the CM to give an unbiased estimate of the IBD state, the marginals need to be correctly specified. The marginal transition and emission probabilities used are the same for the CM and HMM methods. A first order approximation to the true transition probabilities is used for both methods, as in Equation (3.4). The approximation is appropriate for dense markers where it is likely that there has only been zero or one crossovers. If too large a window is used for the CM method, the distance between the target marker and markers on the edge of the window can become large enough that this is not a safe assumption to make. The transition probabilities do not account for a variable recombination rate or LD. In the CM method, however, we assume no LD between the target marker and markers within 2cM of the target locus. In the HMM method we assume no LD between adjacent markers. The effect of LD on estimates from the CM and HMM methods

is also explored in Section 3.1.4.

### *3.1.4 IBD Inference on Simulated Sib Pairs*

*Data*

The data used for the simulated sib-pairs is described in Section 2.2. Comparisons will be made between a simple random sample of 2000 sib-pairs at generation 1 and 50 of the simulated pedigree. The siblings at generation 1 have parents that are population founders, so the IBD transition matrix for siblings with unrelated parents in Section 3.1.1 is the correct model. At generation 50 the parents of the siblings are related and share IBD relative to the founders of the population. The IBD transition matrix for siblings with related parents in Section 3.1.2 is the appropriate model. In this simulation we estimate IBD states from haplotype data $\mathbf{h}$ on ordered DNA copies, that has not been collapsed into genotypes on individuals. All 15 IBD states will be identifiable with the haplotype data. Estimates of IBD state are made at 361 sparse markers along the chromosome, but based on the denser 4303 markers along the chromosome. The HMM models are run on all 4303 markers. The CM models are run on a window of 55 dense markers (spanning approximately 2cM) on either side of each target marker. When calculating transition probabilities for population-IBD transitions in the CRP, the parameters are $\alpha = 0.001, \beta = 0.05$ which allows for more IBD than is expected in generation 49. The parameters were described in Section 3.1.2 and specify a model where the pointwise probability of IBD between two randomly selected chromosomes relative to the population founders is 0.05 and the expected length of such a segment is 500 cM. We found that, as in Brown et al. [2012], having a small $\alpha$ gave better performance. A small $\alpha$ gives a smaller prior probability of an IBD state change and thus stronger evidence from marker data is required for a state change. The detected IBD segments are longer and there are more markers available to provide evidence of a change in the segment.

*Computation Time*

Table 3.1 contains computation times for the HMM and CM methods on the data set, at 361 target markers and 181 target markers. The HMM method is run on the entire chromosome with posterior probabilities calculated from the forward-backward probabilities at specified loci. Computation time is dependent on the length of the chromosome and not the number of target loci. The CM method is run on a window around each target so computation time is dependent on the number of targets. The CM method is therefore more efficient for a lower number of targets spaced over a larger area.

| | user | system | elapsed |
|---|---|---|---|
| HMM 361 | 218.140 | 0.107 | 218.238 |
| CM 361 | 576.622 | 1.597 | 747.016 |
| HMM 181 | 218.647 | 0.176 | 218.786 |
| CM 181 | 284.041 | 0.220 | 284.229 |

Table 3.1: Computation times for HMM and CM methods at 361 and 181 targets

*Error Rates in CM and HMM*

The HMM and CM methods are compared in Table 3.2 by overall error rate. The overall error rate is defined as the proportion of all loci where the estimated IBD state (out of 15 possible states) is not the same as the true IBD state (out of 15 possible states). The comparison is made for HMM and CM methods with either the related or unrelated parents transition matrix. The CM method has lower errors than the HMM method for both generations and both transition matrices. The reason for the lower error rate is explored further below. The CM and HMM make different types of errors in IBD estimation, with the CM tending to underestimate and the HMM overestimate the amount of IBD. We will see that this is due primarily to the effect of LD and the use of marker-to-marker transitions in the HMM.

We will also explore the effect of using either the related or unrelated transition matrix on the error rate. The unrelated transition matrix allows a larger state space than the related transition matrix. The larger state space can result in a higher overall error rate, due to false positive estimates of the additional states.

| | Gen 1 | Gen 50 |
|---|---|---|
| HMM unrel | 0.036 | 0.034 |
| HMM rel | 0.056 | 0.048 |
| CM unrel | 0.015 | 0.019 |
| CM rel | 0.012 | 0.013 |

Table 3.2: Fraction of states estimated incorrectly, over all sib-pairs and all loci

The direction of errors in the HMM and CM methods are different. Table 3.3 classifies the direction of the errors for the methods in generation 1. The HMM tends to over-estimate IBD due to LD between adjacent loci. The CM tends to under-estimate IBD due to only considering dependence between each marker locus and the target.

| | Too Little | Too Much |
|---|---|---|
| HMM Unrel | 0.000 | 0.036 |
| HMM Rel | 0.000 | 0.042 |
| CM Unrel | 0.014 | 0.002 |
| CM Rel | 0.008 | 0.004 |

Table 3.3: Direction of errors made by different methods in generation 1. Too little(much) if estimated state has fewer(more) copies shared IBD between the sibs than true IBD state.

Table 3.2 shows that in generation 1 the HMM with transitions for unrelated parents has fewer errors than with related parents. This is due to the rarity of the related-parent states

- it is much more likely to get a false positive than a true positive. The generation 50 error rates are also higher than generation 1 for each HMM method. This is due to the tendency of the HMM method to overestimate IBD and the reduction of non-IBD states in generation 50, explored below.

For the CM method, when related parent transitions were used there was a lower error rate than using unrelated parent transitions in both generations in Table 3.2. The tendency of the CM method is to estimate too little IBD, as seen in Table 3.3. Modeling related parents reduces the likelihood of a non-IBD state, as seen in Figure 3.1. The effect on the errors is demonstrated in Table 3.4: there is a reduction in the number of incorrect estimates of state 15 (no-IBD). The prior has a minimal effect on the accuracy of the method. Table 3.5 shows the errors for generation 1 and 50 with both related and unrelated transition matrices and different priors.

The HMM and CM methods also differ in how well they estimate the states that are only possible for related parents. In Table 3.6 the errors in generation 50 are classified by the true state - the true states that are only possible with related parents (labels 1-8,10,12,13) and those possible with unrelated parents (labels 9,11,14,15). The methods that use the sibs with unrelated parents only allow estimation of the states 9,11,14 and 15 - the other states are only possible if the parental relatedness is allowed. If the related-parents transitions are used the HMM method estimates these states with a much lower error than the CM method. The related-parents-only states make up around 1% of the states in generation 50 and require evidence from adjacent loci for accurate estimation.

*Effect of LD*

The errors in the HMM method are much more affected by LD than in the CM method. In Table 3.7 the fraction of errors by each method are compared in segments of chromosome with high and low LD. The LD level was calculated by the mean pairwise $R^2$ in the 55 dense locus window around the target. The high LD region was 20 target SNPs from 53cM to 61cM with mean pairwise $R^2$ of 0.05. The low LD region was 20 target SNPs from 61cM to

| | | Truth | | | |
|---|---|---|---|---|---|
| | | 9 | 11 | 14 | 15 |
| Est | 9 | 175174 | 277 | 251 | 0 |
| Unrel | 11 | 2521 | 181041 | 16 | 284 |
| | 14 | 2409 | 6 | 175149 | 293 |
| | 15 | 27 | 2449 | 2511 | 179592 |
| Est | 1 | 4 | 3 | 1 | 5 |
| Rel | 3 | 0 | 9 | 0 | 4 |
| | 4 | 0 | 0 | 14 | 16 |
| | 5 | 0 | 0 | 0 | 14 |
| | 6 | 1 | 7 | 0 | 11 |
| | 7 | 0 | 0 | 6 | 1 |
| | 8 | 0 | 0 | 0 | 5 |
| | 9 | 176742 | 554 | 601 | 6 |
| | 11 | 1733 | 182071 | 19 | 936 |
| | 12 | 0 | 0 | 0 | 3 |
| | 13 | 0 | 0 | 0 | 4 |
| | 14 | 1645 | 10 | 176092 | 970 |
| | 15 | 6 | 1119 | 1194 | 178194 |

Table 3.4: CM estimated and true states assuming unrelated or related parents in generation 1

68cM with mean pairwise $R^2$ of 0.01. The HMM methods have up to 19 times the errors in the high LD vs low LD regions. The CM methods had up to 3 times the errors in the high LD vs low LD regions.

The correlation between error and LD is given in Table 3.8. The correlation between error and LD is similar for both HMM and CM methods. Both methods are affected by

| | Gen 1 | | | Gen 50 | | |
|---|---|---|---|---|---|---|
| Transitions | Unrel | Rel | Unrel | Unrel | Rel | Unrel |
| Prior | Unrel | Rel | Rel | Unrel | Rel | Rel |
| 1 | | | | 1.000 | 0.711 | 1.000 |
| 2 | | | | 1.000 | 1.000 | 1.000 |
| 3 | | | | 1.000 | 0.410 | 1.000 |
| 4 | | | | 1.000 | 0.523 | 1.000 |
| 5 | | | | 1.000 | 0.310 | 1.000 |
| 6 | | | | 1.000 | 0.500 | 1.000 |
| 7 | | | | 1.000 | 0.412 | 1.000 |
| 8 | | | | 1.000 | 0.437 | 1.000 |
| 9 | 0.028 | 0.019 | 0.027 | 0.030 | 0.020 | 0.029 |
| 10 | | | | | | |
| 11 | 0.015 | 0.009 | 0.015 | 0.015 | 0.009 | 0.015 |
| 12 | | | | 1.000 | 0.387 | 1.000 |
| 13 | | | | 1.000 | 0.273 | 1.000 |
| 14 | 0.016 | 0.010 | 0.016 | 0.016 | 0.010 | 0.016 |
| 15 | 0.003 | 0.011 | 0.003 | 0.002 | 0.009 | 0.002 |
| All | 0.015 | 0.012 | 0.015 | 0.019 | 0.013 | 0.019 |

Table 3.5: Errors for CM method by true states with either related or unrelated parent transition matrices and prior probabilities

LD. The HMM and CM differ, however, in the magnitude of the errors. The magnitude of the errors along the chromosome are plotted in Figure 3.2. The HMM methods have larger magnitude errors than CM that appear in areas with high LD. The HMM method models transitions between pairs of adjacent loci whereas CM uses loci in the window paired with the target locus. LD tends to be short-range, giving stronger associations between adjacent

| | Sibs w/ rel. parents | All Sibs |
|---|---|---|
| HMM unrel | 1.00 | 0.03 |
| HMM rel | 0.08 | 0.05 |
| CM unrel | 1.00 | 0.02 |
| CM rel | 0.42 | 0.01 |

Table 3.6: Errors at generation 50 for states that are only possible for sibs with related parents, and states possible for all sibs.

| | Gen 1 | | Gen 50 | |
|---|---|---|---|---|
| | LOW | HIGH | LOW | HIGH |
| HMM unrel | 0.01 | 0.10 | 0.01 | 0.10 |
| HMM rel | 0.01 | 0.19 | 0.01 | 0.17 |
| CM unrel | 0.01 | 0.03 | 0.02 | 0.03 |
| CM rel | 0.01 | 0.03 | 0.01 | 0.03 |

Table 3.7: Error rates in areas of high and low LD

loci than pairs of loci in the window that are not adjacent. The LD associations are evidence for IBD in the models.

*IBD Estimation Recommendations*

The HMM and CM methods each have advantages and disadvantages. The CM method has the potential to reduce computational burden for IBD estimation compared to the HMM over a long segment of genome if the number of target IBD states is low. The CM method is less likely than the HMM to detect IBD as it does not use marker-to-marker transitions along the chromosome. The marker-to-marker transitions provide evidence for low-probability IBD

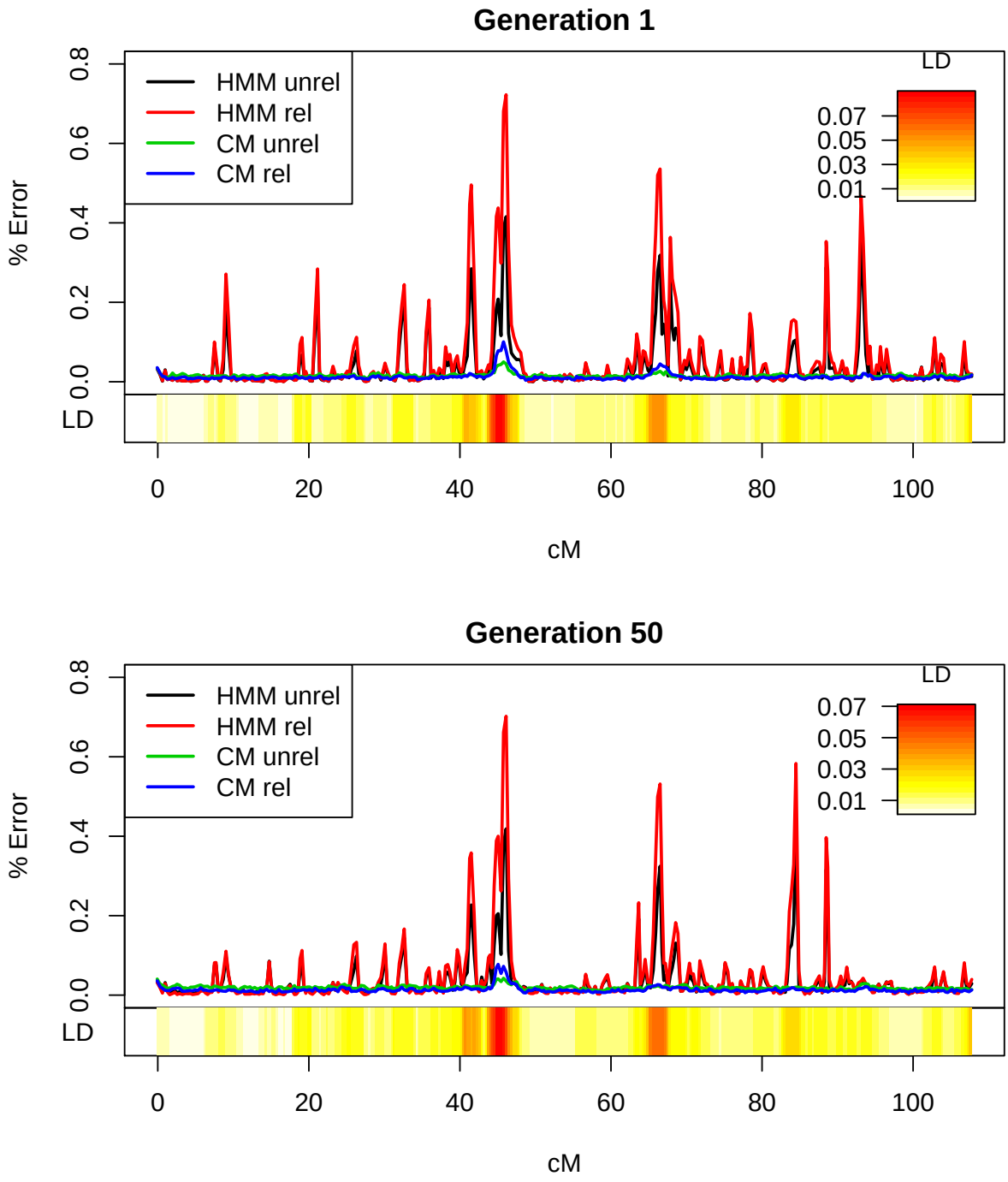



Figure 3.2: Estimation methods error per locus against mean LD

| | Gen 1 | Gen 50 |
|---|---|---|
| HMM unrel | 0.54 | 0.58 |
| HMM rel | 0.66 | 0.68 |
| CM unrel | 0.64 | 0.49 |
| CM rel | 0.79 | 0.75 |

Table 3.8: Correlation (r) between error rate and LD at locus

states that may be short, such as states with IBD sharing in the parents. The CM method, therefore, has fewer true-positive IBD states. However, the CM method also detects fewer false-positive IBD states. The marker-to-marker transition probabilities do not account for LD, so estimate the low-probability states too frequently, mistaking LD for IBD. The HMM method is much more sensitive to LD than the CM method.

### *3.2 Association and IBD Mapping*

In this section, trait locus detection using association and IBD mapping is compared on the simulated sib pair data set. The data set is described in Section 2.2, with trait simulation in Section 2.2.1. Association Tests are described in Section 1.8. In this section, a population based association test is compared to a family based association test. The population based association test is a chi-square test for association between case/control status and genotype. The family based test is a sibling transmission disequilibrium test (STDT). IBD mapping is described in Section 1.7. In this section, an affected sib pair test is compared to the Dudoit test on all sib pairs.

Samples of 1000 sib pairs from the total 10,000 pairs of siblings in generations 1-4, 22-25, 47-50 are used to compare testing methods in Section 3.2.1. To demonstrate the effect of association buildup over more generations, the population is extended to 500 generations in Section 3.2.2. The simulated traits are replicates of either a multiple rare variant or single common variant trait with the same frequency of trait disease allele and trait penetrance.

1000 replicates of each trait are used.

### *3.2.1 First 50 generations*

On generations 1-4, 22-25 and 47-50 a family-based association test, a population-based association test, an affected sib pair test and the Dudoit test for all sib pairs are compared for single common variant and multiple rare variant traits. Results are shown in figure 3.3. In these figures, power is computed as the proportion of realizations in which the p-value of the test is less than 0.05. The upper panels show the power at a marker locus very close to the trait locus over generations of the population, the lower panels show the power at a given generation at locations along the chromosome.

For the association tests, both the population-based GWAS test and the family-based STDT test had high power for the common variant. There was very low power in both association tests for the rare variant, as there is very low correlation between the marker alleles and trait alleles. There is a slight increase in power over the 50 generations, which is explored further in Section 3.2.2 below. The power for the rare variant trait at loci along the chromosome is given in the lower panel. The power is slightly increased at markers near the trait locus for the GWAS test at generation 50. There is no benefit to using the family-based association test, as both association and linkage need to be present and there is no association between the trait and marker alleles [Ott et al., 2011].

For the IBD-based tests, the Dudoit test that used all 1000 sib pairs had high power at a marker close to the trait locus over all generations. This is expected as the marker and trait locus are linked, the IBD shared at the marker locus is likely to be the same at the trait locus. There was no change in power over generations, or between the common and rare variant. As shared descent is modeled, the trait type does not affect the power. This demonstrates the utility of IBD-based tests for traits that are caused by multiple rare variants, or that have allelic heterogeneity. The ASP test only used affected sib pairs and thus had a much smaller sample size, around 200 compared to 1000 pairs in the Dudoit test. The power for the ASP test was therefore lower and more variable over the generations. In

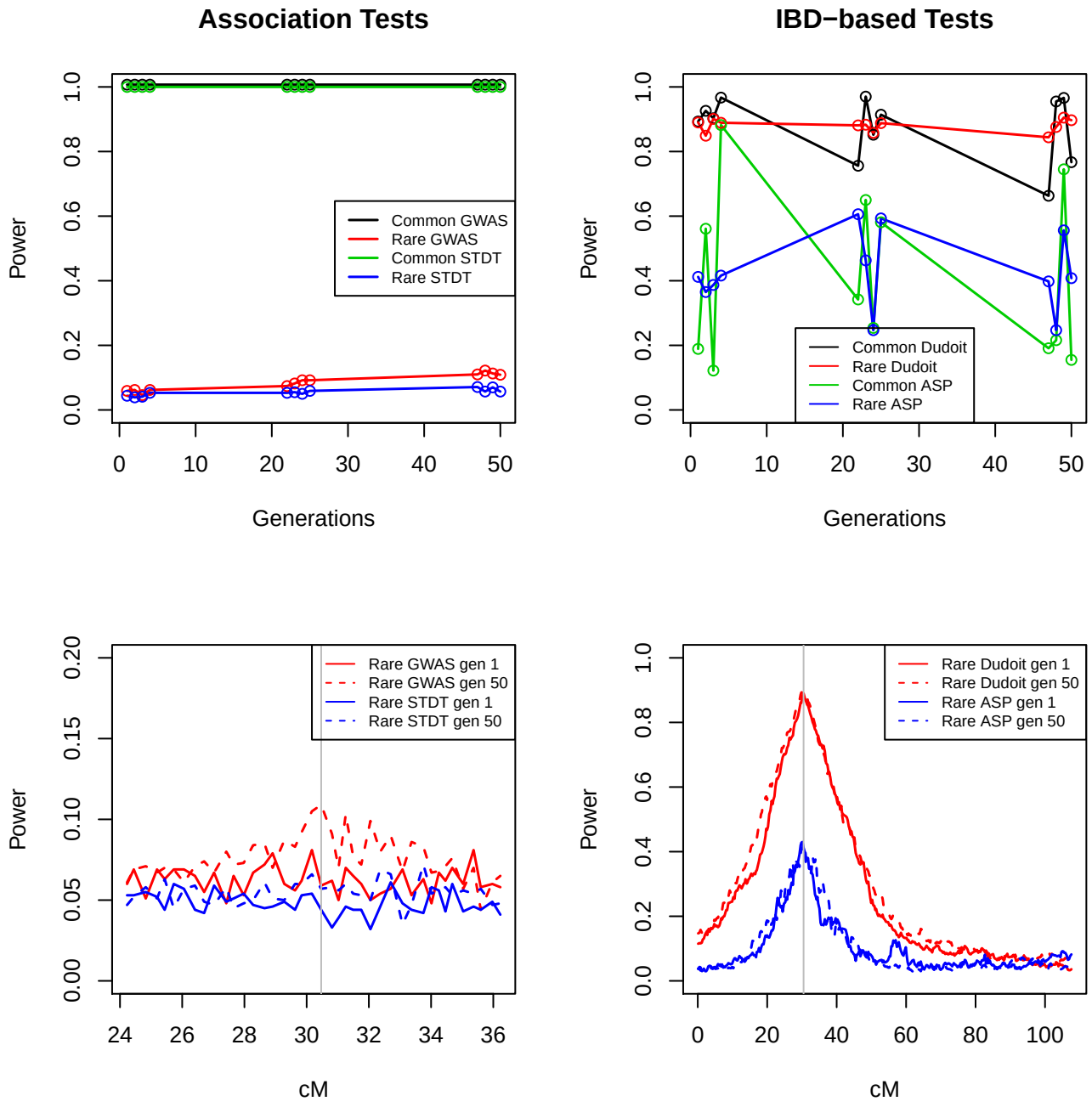



Figure 3.3: Power at 5% level for association and IBD based tests over first 50 generations (1-4, 22-25 and 47-50) and along chromosome.

the lower panel, the power for the rare variant trait over the chromosome is highest near the trait locus, and has a fairly wide peak due to long sections of shared IBD.

The IBD based analysis in Figure 3.3 used the true IBD state, relative to the population founders. In Figure 3.4 the effect of using estimated IBD is shown, for the rare variant trait at the 1st and 50th generations. IBD was estimated by either the HMM or CM method as in Section 3.1.4. At generation 1 the unrelated parents transition matrix was used, and at generation 50 the related parents transition matrix was used. The power for the test using true IBD and IBD estimated with the CM method was similar. The power for the test with IBD estimated by HMM was inflated at locations along the genome where high LD caused excess IBD to be detected by the HMM method.

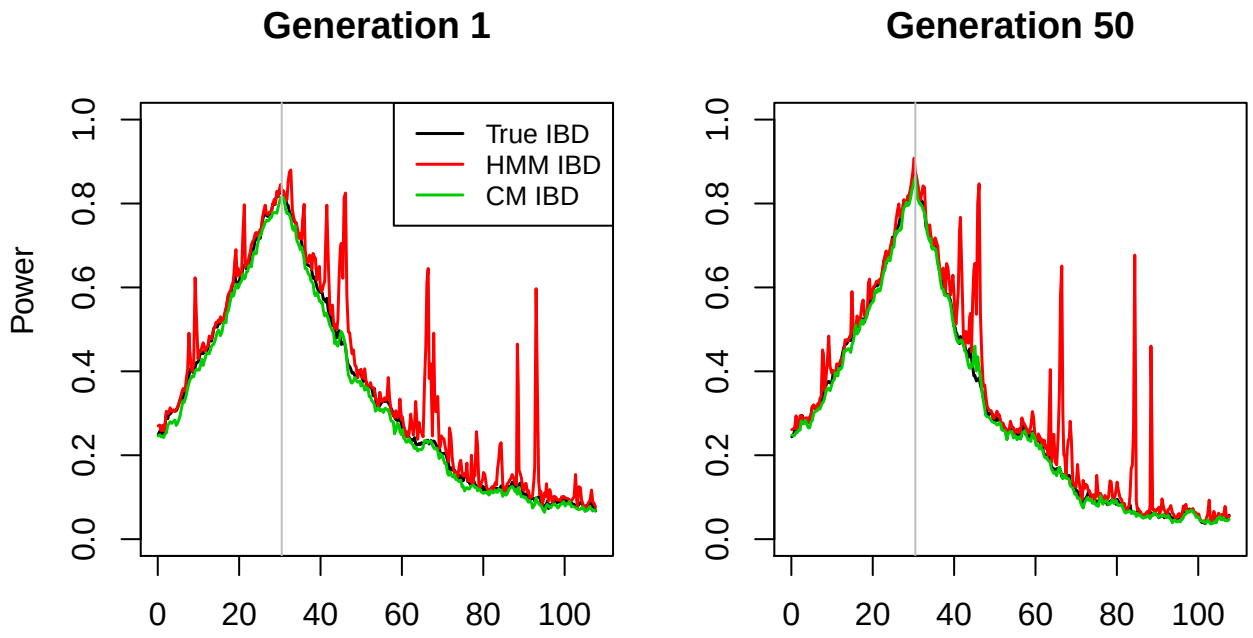



Figure 3.4: Power of IBD-based Dudoit test for a rare variant trait at markers along the chromosome, with true and estimated IBD.

### *3.2.2 After 500 generations*

Over the generations of the population, genetic drift results in the loss of some founder haplotypes and increasing frequency of others, relative to the population founders. There is no selection, so any of the founder haplotypes is equally likely to be lost from the population. Over the generations of the population, the number of unique FGLs in the population decreases, as seen in the first panel of Figure 3.5. For one realization of the trait, the percentage of FGLs in the population in a given generation that are disease alleles remains relatively constant over the generations, shown in the center panel of Figure 3.5. This results in an increasing correlation between the disease alleles and the marker alleles. The correlation coefficient ($R^2$) between disease alleles and marker alleles on the population haplotypes is shown in the third panel of Figure 3.5. The loss of haplotypes can also be seen in the LD heatmaps in Figure 2.2 and 2.3. Short-range LD blocks are broken by recombination, and short-range LD is built up between loci that previously had no association. The buildup in short-range LD between is another interpretation of the cause of the increasing correlation between trait and marker alleles. Long-range LD is built up as haplotypes are lost.

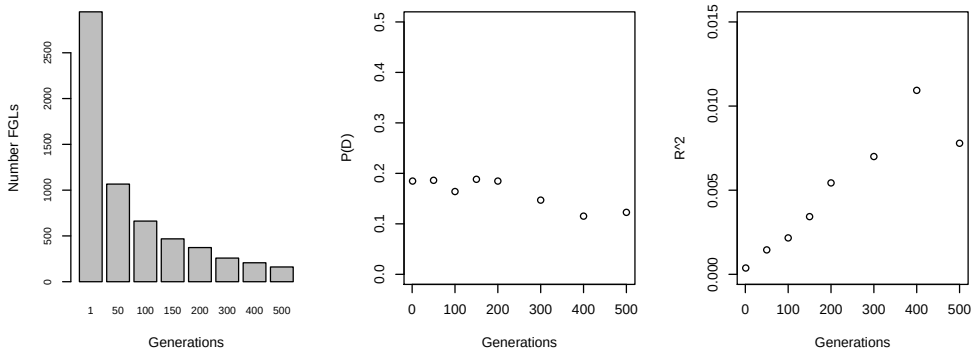



Figure 3.5: Number of unique FGLs, percentage of FGLs that are disease alleles $P(D)$, and correlation ($R^2$) between disease and marker alleles over generations of the population for one realization of the multiple rare variant trait

The effect on the power of the association and IBD-based test is shown in Figure 3.6. The tests are performed are the GWAS test and Dudoit test, at a marker locus very close to the trait locus. The null distribution of IBD sharing was adjusted for the Dudoit test at the 500th generation to reflect increased rates of sharing 1 or 2 copies IBD, and reduced rates of sharing 0 copies IBD. The association test has high power for the common variant over all generations, as there is a high correlation between the allelic type of the marker and the allelic type of the trait. In the rare variant, it is only as correlation between the trait alleles and marker alleles increases that there is any power to detect the trait. The IBD-based test, on the other hand, has high power for both the rare and common variants over all generations. The IBD test is not affected by the trait type or increasing association.

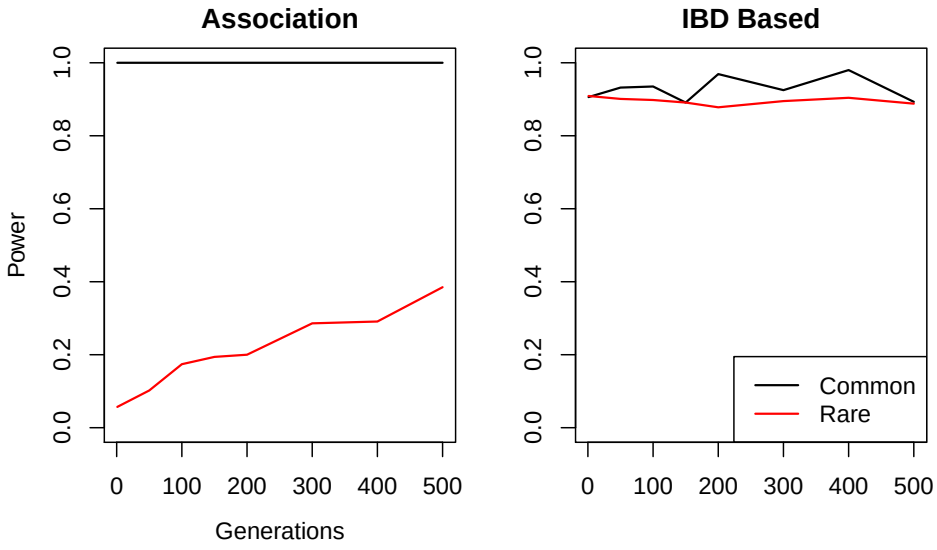



Figure 3.6: Power to detect trait at 5% level of significance in association and IBD-based tests over generations of the population, for the single common variant and multiple rare variant traits. Tests are performed at a marker locus very close to the trait locus.

In the case of a common variant, the power of both association and IBD-based tests are high. Figure 3.7 shows the power to detect the single common variant trait at marker loci

along the simulated chromosome for the association and IBD-based tests. The association test has a much higher precision than the IBD-based tests. The IBD-based tests have a larger area on the chromosome around the trait locus where the trait is detected due to long segments of IBD shared between siblings. There is some increase in power for the IBD test at locations further from the trait locus due to buildup in IBD. These simulations demonstrate that if the causal mechanism is a common variant, and dense marker data is available, association tests have high power to detect the trait locus with a high degree of specificity.

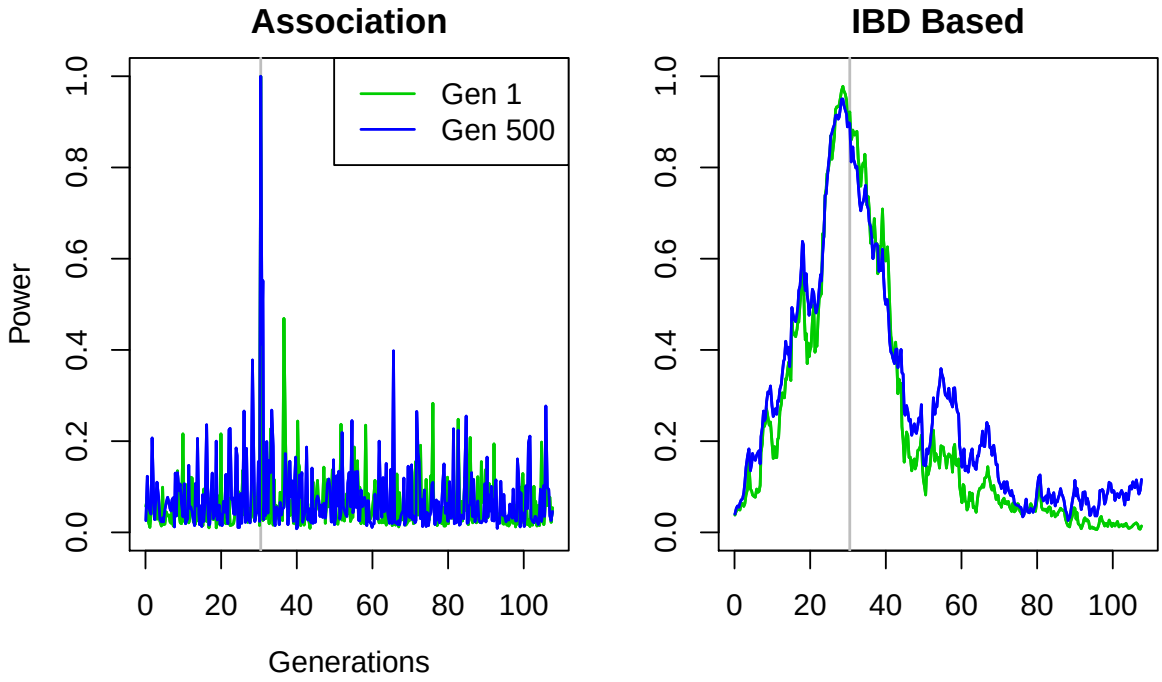



Figure 3.7: Power to detect trait at 5% level of significance, in association and IBD-based tests for a single common variant trait. Tests are performed at marker loci along the chromosome, the trait locus is indicated by the grey line.

For the association test on the rare variant, the increase in LD increases the power at locations close to the trait locus. Figure 3.8 shows the increase in power near the trait locus

over the 500 generations. The increase after 500 generations is much larger than after only 50 generations in Figure 3.3.

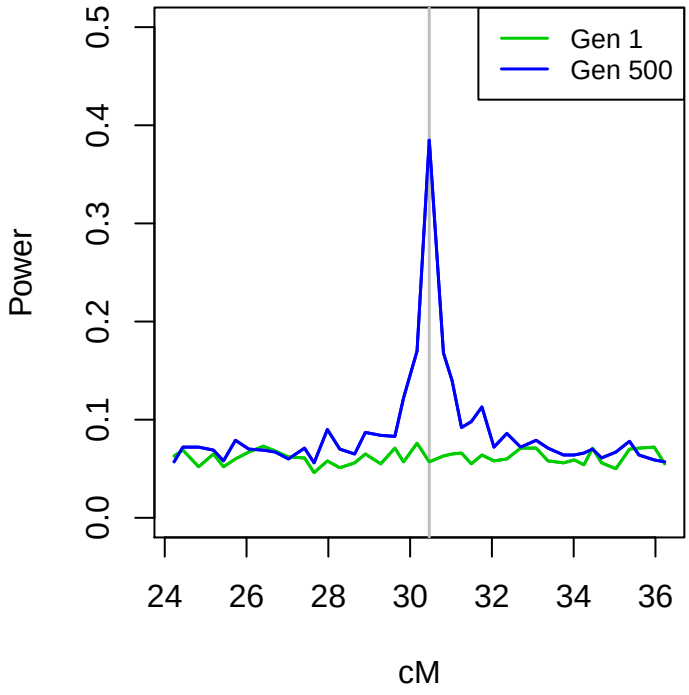



Figure 3.8: Power to detect trait at 5% level, in association test on a multiple rare variant trait. Tests are performed at marker loci along the chromosome, the trait locus is indicated by the grey line.

### *3.2.3 Summary*

Association tests are powerful and have high resolution for traits caused by a single common variant. When the trait is caused by multiple rare variants, however, association tests have very low power to detect the trait unless there is strong local LD. IBD-based tests are powerful for traits caused by either rare or common variants. They have less precision than association tests due to the length of IBD segments, and rely on accurate IBD estimation.

The simulation results demonstrate that accurate IBD estimation and the use of IBD-based tests is important for trait locus detection when there are multiple rare variants, variants that are private to a family, and allelic heterogeneity.

# Chapter 4

# ALGORITHM FOR MERGING POPULATION AND PEDIGREE IBD

This chapter describes a method for merging IBD estimated on individuals in family pedigrees relative to pedigree founders with IBD estimated on the pedigree founders relative to the population founders. The result is estimates of IBD among individuals in a pedigree, relative to population founders. The merged IBD can be used for linkage analysis. In Chapter 5 the method is applied to simulated data sets and in Chapter 6 the method is applied to a real data set. The data for these applications are described in Chapter 2.

## *4.1 Merging method overview*

As in Section 1.1, assume there are a set of genotyped individuals $\mathcal{G}$ who are part of known family pedigrees $\mathcal{F}$. The families are part of a larger population $\mathcal{P}$. We wish to estimate the IBD state among individuals in $\mathcal{F}$ relative to the founders of $\mathcal{P}$, denoted $\mathbf{s}_i^{\mathcal{P}}$ for $i \in \mathcal{F}$. Under the model in Section 1.2, Equation (1.5), IBD among individuals in the family pedigrees relative to the pedigree founders is independent of IBD among the pedigree founders relative to the population founders.

Ideally, we would estimate IBD among all founders of the pedigrees relative to the population founders. Both population- and pedigree-IBD estimates would be independent realizations from the posterior distribution of IBD given marker data and pedigree structure. The population- and pedigree-IBD distributions would be combined by merging population-IBD and pedigree-IBD realizations. In the merging step, chromosomes from paternal and maternal gametes in the population- and pedigree-IBD estimates would be correctly matched for each individual.

In reality, there are several limitations that make this impossible. In population-IBD estimation, pedigree-IBD estimation, and in merging. First, in population-IBD estimation, not all founders in the pedigree are genotyped. High quality IBD estimates cannot be obtained on all founders as the data are not available. We limit population-IBD estimation to genotyped individuals only to ensure the quality of estimates. Practical limitations in the method for population-IBD estimation mean that the joint IBD state can only be estimated well for sets of less than 10 individuals, which further restricts population-IBD estimation. Second, in pedigree-IBD estimation, realizations of IBD partitions are not independent samples from the posterior. They are MCMC samples so are dependent and require good mixing of the Markov Chain to accurately represent the posterior. Finally, the parental origin of chromosomes cannot be observed and can only be inferred in some cases if pedigree information is available. This means that the correct matching of maternal and paternal copies of DNA in population- and pedigree- IBD realizations cannot be guaranteed.

In the merging algorithm presented in this chapter, the aim is to incorporate population-IBD estimates into the pedigree-IBD estimates. The steps of the algorithm are displayed as a flowchart in Figure 4.1. The first step is to estimate pedigree-IBD on all individuals in all pedigrees by obtaining MCMC realizations of IBD partitions. In the second step a subset of individuals for population-IBD estimation is selected, independently of the pedigree-IBD estimation, and realizations of population-IBD are obtained. As a further measure to ensure the quality of population-IBD estimates, the realizations are summarized into a single consensus partition representing IBD estimated above a given threshold. The quality of the estimates is demonstrated on simulated data in Section 5.2. The motivation for only adding high quality IBD estimates is described in Section 4.6, where it is shown that adding IBD increases the LOD score in most cases. Adding false-positive IBD will increasing the LOD score incorrectly and result in a false linkage signal. The thresholding places a further check on population-IBD estimates, only allowing IBD to be added if it estimated with a high frequency in the population-IBD realizations. There is a loss of information, however, as after thresholding no probablities are used. In the final merging step, the consensus partition

is merged with each of the pedigree-IBD realizations to give a set of merged pedigree- and population-IBD realizations.

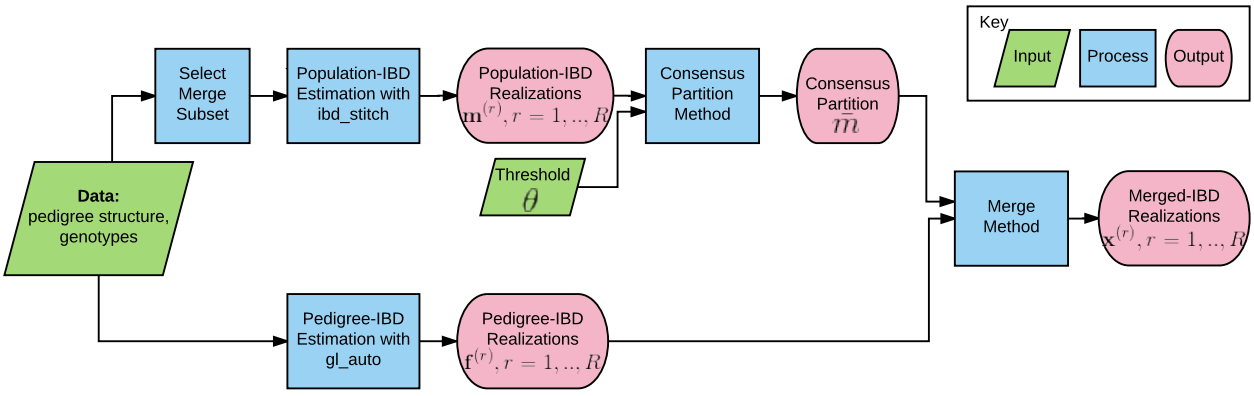



Figure 4.1: Flowchart of steps for obtaining merged IBD realizations

Section 1.5.1 described methods of estimating pedigree-IBD on individuals in $\mathcal{F}$ relative to the founders of the family pedigrees $\mathcal{F}$. The pedigree-IBD is denoted $\mathbf{s}^{\mathcal{F}}$ and represented by FGL partitions $\mathbf{x}^{\mathcal{F}}$ where $\mathbf{x}^{\mathcal{F}}_{jl}$ is the label of the founder of $\mathcal{F}$ that provided the DNA to chromosome $j \in \mathcal{F}$ at locus $l \in \lambda$. As IBD is estimated, the FGL partitions are modeled as a random variable $\mathbf{X}^{\mathcal{F}}$ with a distribution that depends on the pedigree structure and is conditioned on observed genotype data on $\mathcal{G}$. The MORGAN program `gl_auto` produces MCMC realizations from $\mathbf{X}^{\mathcal{F}}$. The output is recoded to give FGL matrices $\mathbf{f}^{(r)}, r = 1, ..., R$ where $\mathbf{f}^{(r)}$ is a $2N \times L$ matrix containing FGL labels for the $2N$ pedigree chromosomes at each of $L$ loci. For the merging algorithm we will assume that we have realizations of pedigree-IBD $\mathbf{f}^{(r)}$ on individuals in the pedigrees $\mathcal{F}$ over a sparse panel of markers $\lambda$ indexed by $l = 1, ..., L$.

Section 1.5.2 described methods of estimating population-IBD on individuals in $\mathcal{G}$ relative to the founders of the larger population $\mathcal{P}$. The population-IBD is denoted $\mathbf{s}^{\mathcal{P}}$ and represented by FGLs $\mathbf{x}^{\mathcal{P}}$ where $\mathbf{x}^{\mathcal{P}}_{jl}$ is the FGL of chromosome $j \in \mathcal{G}$ at locus $l \in \Lambda$. For a given subset of individuals $\mathcal{M} \subset \mathcal{G}$, the FGLs are modeled as a random variable $\mathbf{X}^{\mathcal{P}}$ with a distribution

that depends on a population-IBD model and is conditioned on observed genotype data on $\mathcal{M}$. The `ibd_stitch` program produces independent realizations from $\mathbf{X}^{\mathcal{P}}$. The output is recoded to give FGL matrices $\mathbf{m}^{(r)}, r = 1, ..., R$ where $\mathbf{m}^{(r)}$ is a $2n \times L$ matrix containing FGL labels for the $2n$ chromosomes in the merge subset at each of the $L$ loci. The population-IBD estimation is performed using a denser set of markers than is used for pedigree-IBD estimation, however, only the output at the same $L$ loci used in pedigree-IBD estimation is used. The selection of an optimal merge set $\mathcal{M} \subset \mathcal{G}$ is discussed in Section 4.2. The realizations of population-IBD on $\mathcal{M}$ will be summarized by a consensus partition, which is described in Section 4.3.

The merging step, where the consensus partition will be merged into each realization of pedigree-IBD, is described in Section 4.4. Finally, section 4.5 describes the calculation of a LOD score using the merged realizations and the effect of merging on the LOD score is explored in Section 4.6.

### *4.2 Selection of Merge Set*

The merge set is the subset of individuals in the pedigree who are selected for estimation of joint population-IBD. As discussed in Section 4.1, the ideal merge set is all the pedigree founders; however, this is not possible in real data. In human pedigrees, the youngest generation is the most likely to be genotyped. The pedigree founders may be deceased, or it may be otherwise impossible to obtain genotype data on them. Additionally, even if it were possible to obtain genotypes on all founders, limitations in estimation methods mean that it is difficult to obtain joint IBD states over large numbers of individuals. The limitations of `ibd_stitch` are discussed in Section 1.5.2.

In practice we select as large a set as possible of genotyped individuals who are "unrelated" according to the pedigree structure. Unrelated individuals are selected for population-IBD estimation because the population-IBD model does not account for pedigree relationships. IBD relative to pedigree founders is better estimated by `gl_auto`, where pedigree relationships are used to provide an informative prior. When selecting between a large num-

ber of potential individuals, preference should be given to a set with a high level of genotypic similarity over the genome. Selecting distantly related individuals with genotypic similarity increases the chance that they share large amounts or long segments of IBD that may be important in the analysis. The target for IBD estimation is all population-IBD between all individuals in the family, and the target LOD score is conditional on the population-IBD. By selecting individuals who share the most IBD, we will be closer to capturing all population-IBD.

The selection of the merge set $\mathcal{M}$ has not been automated, and was performed as a manual step in this thesis. First, a list of genotyped individuals who are unrelated according to the pedigree structure was obtained by inspecting pedigree kinships, defined in Section 1.5. The matrix of the kinship coefficients between all pairs of genotyped individuals has a block structure. The block structure is formed by zeros between groups of individuals who have no pedigree ancestors in common. Furthermore, within a pedigree, subgroups of individuals that do not share a common ancestor in the pedigree will also have zero pairwise kinships. One individual is selected from each block to give a list of unrelated individuals. Within a block the oldest individual should be selected. For instance, if a block is made up of a parent-child pair the parent is selected. In order to reduce the length of the list to a size appropriate for `ibd_stitch`, the individuals with highest IBS sharing were selected. The realized genotypic similarity (based on observed IBS) was calculated between all pairs. The pairs were ranked from highest to lowest, and individuals were selected starting with the highest ranked pairs till the desired size of merge set was selected.

### *4.3 Consensus Partition*

Consensus clustering is a method of combining multiple partitions to find the most representative partition, without accessing the features or processes that determined the clustering. Consider the realizations of population-IBD from `ibd_stitch` at a given locus $l$, $\mathbf{m}_l^{(r)}, r = 1, ..., R$ to be an ensemble of $R$ partitions of objects ($2N$ DNA copies for $N$ individuals) into classes (up to $2N$ unique FGLs). Taking the consensus over `ibd_stitch`

realizations allows us to determine IBD probabilities in the posterior given the uncertain ordering of DNA copies in the realizations. We then impose a threshold on those probabilities for IBD detection as quality control for the population-IBD. Merging the consensus partition as opposed to the individual realizations improves performance. It is shown in Chapter 5 how sensitive the linkage signal is to false positive IBD, so we want to be as conservative as possible in adding IBD states. The improvement in using the consensus partition over individual realizations is demonstrated in the GAW analysis in Section 5.1.

A membership matrix is a representation of a partition. The IBD partition for realization $r$ at locus $l$ over the ordered DNA copies is represented as a single ordered FGL vector $\mathbf{m}_l^{(r)}$ in the `ibd_stitch` output. It can also be represented as a membership matrix, $\mathbf{M}^{(r)}$, where

$$\mathbf{M}_{jk}^{(r)} = \mathbb{1}(\text{Copy } j \text{ in class } k) \tag{4.1}$$

The ordering of the two DNA copies of an individual, however, varies between realizations of `ibd_stitch` as there is no pedigree structure to identify which DNA copy is paternal or maternal in origin. A new membership matrix was defined to allow for the varying ordering of the DNA copies between realizations. The importance of not assuming parental origin is known is explored in Section 4.3.1. Define a new type of membership matrix $\mathbf{M}^{(r)} = \mu(\mathbf{m}_l^{(r)})$ where $\mathbf{M}^{(r)}$ has two rows for each individual $i$, denoted $(i_1, i_2)$ for a total of $2N$ rows and a column for each unique FGL class $k = 1, ..., 2N$. The elements are

$$\mathbf{M}_{i_1,k}^{(r)} = \mathbb{1}(\text{ either copy of indiv } i \text{ in class } k) \tag{4.2}$$

$$\mathbf{M}_{i_2,k}^{(r)} = \mathbb{1}(\text{ both copy of indiv } i \text{ in class } k), \tag{4.3}$$

so $i_1$ is the row for either DNA copy, and $i_2$ is the row for both DNA copies. Note that $\mathbf{M}_{i_1,k}^{(r)} \geq \mathbf{M}_{i_2,k}^{(r)}$ because if both DNA copies are in a class then either DNA copy must be in the class. Also, the sum over both rows for an individual must be two, $\sum_j \sum_k \mathbf{M}_{ijk}^{(r)} = 2$. Either both non-zero indicators for an individual are in the "either" row for different classes and the row sums are 2 and 0; or one is in the "either" row and the other in the "both" row and the row sums are 1 and 1. The maximum number of FGL classes is equal to the number

of DNA copies as we need to allow for no IBD, where each DNA copy has a unique class label. The class labels are arbitrary and will differ between realizations.

In general, consensus clustering is an optimization task, to find a clustering $\bar{\mathbf{M}}$ that minimizes a loss function of the form

$$L(\bar{\mathbf{M}}) := \sum_{r=1}^{R} w_r d^2(\mathbf{M}^{(r)}, \bar{\mathbf{M}}), \tag{4.4}$$

where $w_r$ are weights and $d^2()$ is a dissimilarity measure. Consensus clustering methods vary in the definition of $\bar{\mathbf{M}}$, by the definition of $d^2$ and $w$, the search space for $\bar{\mathbf{M}}$, and by the optimization of the search. These components are defined below.

The dissimilarity function $d^2$ in Equation (4.4) will be the Euclidean dissimilarity of Dimitriadou et al. [2002], based on the L2 norm. As the class labels are arbitrary and differ between realizations, dissimilarity functions are designed to be invariant to the relabeling of classes. The relabeling of classes corresponds to permutations, $\Pi$, of the columns of $\mathbf{M}^{(r)}$. The dissimilarity function is,

$$d^2(\mathbf{M}^{(r)}, \bar{\mathbf{M}})^2 := \sum_{i=1}^{N} \sum_{j=1}^{2} \sum_{k=1}^{K} |\Pi_r(\mathbf{M}^{(r)}_{i_j k}) - \bar{\mathbf{M}}_{i_j k}|^2, \tag{4.5}$$

where $\Pi_r$ is the optimal permutation for realization $r$ that minimizes $d^2$. The weighting in Equation (4.4) will be set to $w_r = 1$ for all $r$, to give equal weight to each realization from `ibd_stitch`. We assume that each realization is an independent sample from the posterior and thus each realization should have equal weight.

The search space for $\bar{\mathbf{M}}$ is over all soft partitions. In hard consensus partitions [Strehl and Ghosh, 2002] the output is a single consolidated graph where each object, in our case $i_j$, is allocated to one of the classes. Soft consensus partitions are a generalization [Punera and Ghosh, 2006] that give a posterior probability of membership of each object to each class. That is,

$$\bar{\mathbf{M}}_{i_1 k} = P(\text{ either copy of indiv } i \text{ in class } k | \mathbf{M}^{(1)}, ..., \mathbf{M}^{(R)}) \tag{4.6}$$

$$\bar{\mathbf{M}}_{i_2 k} = P(\text{ both copies of indiv } i \text{ in class } k | \mathbf{M}^{(1)}, ..., \mathbf{M}^{(R)}). \tag{4.7}$$

Note that while in $\mathbf{M}^{(r)}$ each element is an indicator that takes the value 0 or 1, in $\bar{\mathbf{M}}$ each element is a probability in the interval [0,1]. The rows in $\bar{\mathbf{M}}$ are not constrained to a particular sum. For a group of distantly related individuals there may be strong evidence for IBD between a certain pair of DNA copies and weak or no evidence for other IBD at the locus. The posterior probabilities in the soft partition allow us to determine the strength of evidence for estimated IBD which is not possible in a hard partition.

Obtaining $\bar{\mathbf{M}}$ is a numerical optimization problem that iterates between two steps: finding the optimal $\bar{\mathbf{M}}$ for the current permutations and finding the optimal permutations given the current $\bar{\mathbf{M}}$. The algorithm initializes with a randomly chosen set of permutations. For given permutations, $\bar{\mathbf{M}}$ can be found analytically. For instance, if $\Pi_r$ were the identity matrix for all $r$, then the estimator for element $\bar{\mathbf{M}}_{i_j k}$ would be $\frac{1}{R}\sum_{r=1}^{R} \mathbb{1}(i_j \text{ in class } k \text{ in realization } r)$. That is, the proportion of realizations in which each event was observed. Finding $\bar{\mathbf{M}}$ given the permuations never increases the value of the loss function. The other step, finding the optimal permutations given $\bar{\mathbf{M}}$, is computationally difficult. There are up to $K!$ possible permutations of the $K$ columns in each realization $\mathbf{M}^{(r)}$, and the optimal permutation must be found for each realization. Finding the optimal permutation is an instance of the multi-dimensional assignment problem (MAP) which has been shown to be NP-hard [Dimitriadou et al., 2002]. The solution was approximated using numerical optimizations from the R package CLUE [Hornik, 2005, 2015]. CLUE, in the `cl_consensus` function using the `''SE''` method. This method comes from Dimitriadou et al. [2002]. For each realization, a confusion matrix $\mathbf{C}$ is created that cross-tabulates the class assignments in $\mathbf{M}^{(r)}$ and the current $\bar{\mathbf{M}}$. The matrix $\mathbf{C}$ has therefore $K$ rows for each class in $\mathbf{M}^{(r)}$ and $K$ columns for each class in $\bar{\mathbf{M}}$. The elements are defined $\mathbf{C}_{k_1,k_2} = \sum_i \sum_j \mathbf{M}^{(r)}_{i_j k_1} \bar{\mathbf{M}}_{i_j k_2}$, that is, the sum of probabilities in $\bar{\mathbf{M}}$ for class $k_2$ over the rows in which $\mathbf{M}^{(r)}$ has a 1 in class $k_1$. The element will have a value between 0 (for instance if in $\mathbf{M}^{(r)}$ all indicators on all rows are 0 for class $k_1$) up to $K$ (if in $\mathbf{M}^{(r)}$ all indicators on all rows are 1 for class $k_1$, and in $\bar{\mathbf{M}}$ there is probability 1 for class $k_2$ on all rows). Label matching is done by a greedy algorithm on the confusion matrix, which starts by matching the two class labels with the highest term in the confusion

matrix. The row and column containing the highest element are removed from the matrix, and the highest term in the reduced matrix is found. The label matching determines the new permutation $\Pi_r$. As each new $\Pi_r$ is found with a greedy algorithm, it is not guaranteed to decrease the contribution of realization $r$ to the loss function. The optimization procedure can end at a local minimum.

For this thesis, the CLUE code was modified to implement a custom function for constructing the newly defined IBD membership matrices $\mathbf{M}^{(r)}$ from the FGL partitions $\mathbf{m}_l^{(r)}$. The $\bar{\mathbf{M}}$ with the lowest total dissimilarity is selected over 5 runs of the numerical optimization from different initial conditions.

The need for multiple runs of the consensus method is demonstrated on the Alzheimer's data, which is described in Section 2.5. The consensus procedure was run on `ibd_stitch` realizations on loci in a segment of chromosome 21 for a set of 7 individuals (set3). The segment contains 109 marker loci. 500 runs of the consensus function were performed at each marker rather than the normal 5 runs. At 56 of the 109 markers, IBD was indicated between individuals 201_44 and 203_180 in 100% of the runs of the consensus function. At the remaining 53 loci, the IBD was present in only 94% of the runs of the consensus function. In the 6% of runs that did not find the IBD, the numerical optimization stopped at a local minimum. For example, the two DNA copies of individual 203_180 are in classes 1 and 2, and individual 201_44 has one DNA copy in class 3 and one DNA copy equally likely to be in either class 1 or 2. If the best of two runs is used the IBD is found 100% of the time at all loci. Five runs is appropriate for these data, but more than five runs may be necessary in other data sets.

The purpose of summarizing the population-IBD into a soft consensus was to evaluate the probability of population-IBD estimated by `ibd_stitch`. Only IBD with a high probability will be incorporated into the pedigree-IBD to form merged IBD estimates. The matrix $\bar{\mathbf{M}}$ contains probabilities that DNA copies are members of arbitrarily labeled classes. These probabilities are not IBD probabilities, but they are related to IBD. For instance, if two DNA copies each have a high probability of membership to a certain class, they are likely both in

the class and thus are IBD. There may be classes where there is no DNA copy with a high probability of membership, this indicates that the class is empty and fewer than $K$ classes are required to partition the DNA copies into IBD classes. To determine whether there is sufficient evidence of IBD for merging, the probabilities in $\bar{\mathbf{M}}$ are thresholded by $\theta > 0$ to give a hard partition $\bar{\mathbf{m}}$. The partition $\bar{\mathbf{m}}$ is a deterministic function of the probabilities $\bar{\mathbf{M}}$ and the threshold, that is, $\bar{\mathbf{m}} = \mu(\bar{\mathbf{M}}, \theta)$. The function $\mu$ is described below, it assigns DNA copies to an IBD class if the estimated membership probability exceeds the threshold. The partition $\bar{\mathbf{m}}$ is then merged into every realization of pedigree-IBD, $\mathbf{f}^{(r)}$, as described in Section 4.4.

Pseudocode for the thresholding function $\bar{\mathbf{m}} = \mu(\bar{\mathbf{M}}, \theta)$ is given in Algorithm 1. For each individual we need to assign labels to the individual's two DNA copies in $\bar{\mathbf{m}}$. If there is a high probability that the individual has both copies in the same class, both copies should get that class label in $\bar{\mathbf{m}}$. We determine this by testing if $\bar{\mathbf{M}}_{i_2k} = P(\text{both copies of } i \text{ in class } k) > \theta$ for some $k$. If this is not the case, then the copies must be in different classes. We look for classes where $\bar{\mathbf{M}}_{i_1k} = P(\text{either copy of } i \text{ in class } k) > \theta$. If there are two such classes, we give one of the class labels to each DNA copy in $\bar{\mathbf{m}}$. If there is only one such class, one of the DNA copies gets the class label but the other copy must get a unique class label; there is not strong enough evidence to determine a definitive class membership for the other DNA copy. The class labels in the probability matrix $\bar{\mathbf{M}}$ are $1, ..., K$ so we use unique class labels from $K + 1$ onwards to ensure that no other DNA copy will have the label. Finally, if there are no classes where $P(\text{either copy of } i \text{ in class } k) > \theta$, then both DNA copies get unique class labels and are not IBD with any other DNA copy in $\bar{\mathbf{m}}$.

### *4.3.1 Consensus assuming known parental origin*

The importance of defining a membership matrix that does not assume parental origin is illustrated on a pair of individuals in the Merge2 pedigrees, described in Section 2.3. The pair of individuals are siblings according to the unobserved population pedigree, but are unrelated in the observed family pedigree structure. This investigation was performed as

**INPUT** : Probability matrix $\bar{\mathbf{M}}$ with probabilities for classes 1...$K$; threshold $\theta$

**OUTPUT** : Hard partition vector $\bar{\mathbf{m}}$

Initialize new class labels $c = K + 1$ and output vector $\bar{\mathbf{m}}$
**for** *each individual* **do**
  **if** $P($*both copies in class* $k) > \theta$ *for some* $k$ **then**
    set both labels for individual in $\bar{\mathbf{m}}$ to $k$
  **else**
    Identify classes $k$ where $P$(either copy in class $k$) $> \theta$
    **if** *two classes* $k_1$ *and* $k_2$ **then**
      set labels for individual in $\bar{\mathbf{m}}$ to $k_1$ and $k_2$
    **else if** *one class* $k$ **then**
      set labels for individual in $\bar{\mathbf{m}}$ to $k$ and $c$, increase $c = c + 1$
    **else**
      set labels for individual in $\bar{\mathbf{m}}$ to $c$ and $c + 1$, increase $c = c + 2$
    **end**
  **end**
**end**

**Algorithm 1:** Thresholding algorithm for function $\bar{\mathbf{m}} = \mu(\bar{\mathbf{M}}, \theta)$

earlier versions of the merging algorithm did not account for uncertain parental origin.

These siblings happen to share approximately half of the simulated chromosome with one copy IBD and half with two copies IBD. Table 4.1 shows how well this IBD is estimated. Quality of IBD estimates are indicated by the % of loci at which the correct IBD state was estimated and by the relative root mean squared error (RRMSE) of number of IBD class descriptive statistics. The descriptive statistics are the size of the largest IBD class at a locus, the number of single DNA copies at a locus, the average class size at a locus, and the

number of classes at a locus. The RRMSE, defined for a given descriptive statistic $c$, is

$$RRMSE(\hat{\mathbf{c}}, \mathbf{c}) = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{c}_i - c_i)^2}}{\frac{1}{n}\sum_{i=1}^{n} c_i} \tag{4.8}$$

where $\hat{\mathbf{c}} = \hat{c}_1, ..., \hat{c}_n$ are the descriptive statistics calculated on the estimated IBD state over $n$ loci and $\mathbf{c} = c_1, ..., c_n$ the descriptive statistic calculated on the true IBD state at each locus. The RRMSE measure is appropriate as the true IBD classes do not vary greatly over loci. Good estimates were obtained when the parental origin is correctly assumed to be unknown (T), with the membership matrix as defined in Equations (4.2) and (4.3). Good IBD estimates are not obtained when parental origin is falsely assumed to be known (F), with the membership matrix as defined in Equation (4.1). Within the unknown origin consensuses, the lower threshold 51% performed better than the 80% and 99% threshold.

| | Correct State % | Largest Class | Single Copies | Avg Class Size | Number Classes |
|---|---|---|---|---|---|
| F_51 | 32.41 | 0.41 | 1.48 | 0.40 | 0.56 |
| F_80 | 0.00 | 0.50 | 1.72 | 0.45 | 0.64 |
| F_99 | 0.00 | 0.50 | 1.72 | 0.45 | 0.64 |
| T_51 | 99.45 | 0.00 | 0.49 | 0.03 | 0.03 |
| T_80 | 98.89 | 0.00 | 0.48 | 0.04 | 0.04 |
| T_99 | 95.29 | 0.08 | 0.53 | 0.06 | 0.09 |

Table 4.1: Consensus IBD partitions from ibd_stitch compared to simulation truth for a pair of siblings (4 DNA copies). Comparisons are % correct for correct state, otherwise RRMSE (Eq 4.8). Methods are assuming known (K) or unknown (U) parental origin at different thresholds.

## *4.4 Merging Algorithm*

The purpose of the merging stage of the algorithm is to combine the pedigree-IBD realizations with the consensus partition from the population-IBD realizations to form merged population- and pedigree-IBD realizations. The merged realizations approximate the distribution of IBD states among the pedigree individuals relative to the population founders. A realization of pedigree IBD will have two or more IBD classes merged into one class when the consensus partition $\bar{\mathbf{m}}$ indicates that they should be the same class.

The merging of the consensus partition $\bar{\mathbf{m}}$ for locus $l$ into each realization of pedigree-IBD $\mathbf{f}_l^{(1)}, ..., \mathbf{f}_l^{(R)}$ to form the merged IBD realizations $\mathbf{x}_l^{(1)}, ..., \mathbf{x}_l^{(R)}$ will be done for each realization $r = 1, ..., R$ and at each locus $l = 1, ..., L$. The inputs are the population-IBD partition for the locus $\bar{\mathbf{m}}$ and a single pedigree-IBD realization $\mathbf{f}_l^{(r)}$. The output is vector $\mathbf{x}_l^{(r)}$ of the same dimension as $\mathbf{f}_l^{(r)}$, which will be used as merged-IBD realization. The implementation of the algorithm used in analyses in this thesis is described in Section 4.4.1. Limitations of the implementation are discussed in Section 4.4.2.

The main challenge in the merging step is that the parental origins of the two DNA copies belonging to an individual may be aligned differently in $\bar{\mathbf{m}}$ and $\mathbf{f}_l^{(r)}$. The parental origins of DNA copies in population-IBD estimation are not known, and in the pedigree-IBD realizations the individuals in the merge set are founders so the parental origins are also unknown. The copies are therefore matched randomly in each run of the merge algorithm. If in $\bar{\mathbf{m}}$ a DNA copy of an individual is IBD with a DNA copy in another individual, $\mathbf{f}_l^{(r)}$ is updated by adding IBD among any of the four copy combinations with equal probability. The selection of copies to form the new IBD group can result in either a large IBD class in $\mathbf{x}_l^{(r)}$ if that label is shared with many other pedigree individuals or a small class if not. The random selection ensures that the effect of the choice is averaged over the realizations.

### *4.4.1 Implementation of merge algorithm*

The implementation of the algorithm used for analyses in this thesis is given in pseudocode in Algorithm 2 and described here. The inputs are the population-IBD consensus vector $\bar{\mathbf{m}}$ and the pedigree-IBD realization $\mathbf{f}_l^{(r)}$. The output will be the merged-IBD realization $\mathbf{x}_l^{(r)}$.

The output is initialized to the pedigree-IBD, and will be updated to reflect the additional population-IBD. We also initialize an empty vector $x_{used}$. The highest level iteration is over the population-IBD classes that are to be merged into the pedigree-IBD. Population-IBD classes are identified by inspecting $\bar{\mathbf{m}}$ and identifying labels that are shared by more than one DNA copy. For a given IBD class, the individuals with at least one copy in the class, according to the population-IBD, are identified. The current labels for both their DNA copies are identified in $\mathbf{x}_l^{(r)}$. We need to find both labels as we do not know which copy will be used for the new IBD group. The label that will be used for the new IBD group is $x_{new}$. It is selected from the set of current labels. Next, the individuals with at least one DNA copy in the population-IBD class are iterated over, updating the pedigree-IBD labels. An individual who is autozygous in the population-IBD have both current labels in $\mathbf{x}_l^{(r)}$ updated to $x_{new}$. When a label is updated, it is replaced everywhere it appears in $\mathbf{x}_l^{(r)}$. Individuals who are not autozygous in the population-IBD should have one of their DNA copy labels updated to $x_{new}$. When choosing which label of an individual to replace, we need to ensure that the label has not been used before to merge a previous IBD group. This is the purpose of the $x_{used}$ vector. If the population-IBD specifies that an individual has each DNA copy in a different IBD class, then one copy should be used to merge the first class and the other copy for the second class. The copy used in the first class is chosen at random, which fixes the copy to be used in the second class. Once the labels are updated for all the individuals, the label is added to $x_{used}$.

There are three sources of uncertainty in the creation of the merged IBD state: uncertainty in the pedigree IBD estimates, uncertainty in the population IBD estimates and in the merging procedure itself. The uncertainty in pedigree-IBD is reflected in the merged

**INPUT** : Population-IBD consensus partition $\bar{\mathbf{m}}$ for locus $l$; Pedigree-IBD partition realization for locus $l$, $\mathbf{f}_l^{(r)}$

**OUTPUT** : Merged-IBD realization for locus $l$, $\mathbf{x}_l^{(r)}$

Initialize output vector to pedigree-IBD FGLs $\mathbf{x}_l^{(r)} = \mathbf{f}_l^{(r)}$
Initialize empty vector $x_{used}$.
**for** *Each class in* $\bar{\mathbf{m}}$ *with more than one DNA copy* **do**
    Identify individuals with at least one copy in $\bar{\mathbf{m}}$ class and identify current labels in $\mathbf{x}_l^{(r)}$ of both copies for these individuals
    Choose new label $x_{new}$ from current labels, excluding any label in $x_{used}$
    **for** *Each individual in* $\bar{\mathbf{m}}$ *class* **do**
        **if** *Individual is autozygous in* $\bar{\mathbf{m}}$ **then**
            Update all instances of both the individual's labels in $\mathbf{x}_l^{(r)}$ to $x_{new}$
        **else if** *Individual has one current label in* $x_{used}$ **then**
            Update all instances of the individual's other label in $\mathbf{x}_l^{(r)}$ to $x_{new}$
        **else if** *Individual has no current labels in* $x_{used}$ **then**
            Select one of the individual's labels in $\mathbf{x}_l^{(r)}$ at random with equal probability
            Update all instances of the label in $\mathbf{x}_l^{(r)}$ to $x_{new}$
        Add label $x_{new}$ to $x_{used}$.
    **end**
**end**

**Algorithm 2:** Merging algorithm for combining $\bar{\mathbf{m}}$ and $\mathbf{f}_l^{(r)}$.

IBD states as each realization of pedigree-IBD is used. The uncertainty in population-IBD is reduced by summarizing the realizations into a consensus partition where the probability must exceed the threshold $\theta$. The distribution of population-IBD estimates is not reflected in the merged states. Uncertainty in the merge procedure itself comes from random decisions made in the merging algorithm about which DNA copies to make IBD due to unknown

parental origins.

Random matching of DNA copies in the population-IBD and pedigree-IBD realizations is an unbiased approach when there are unknown parental origins. When a merge is performed, the matching of the DNA copies used is an independent sample from the uniform distribution over all possible matchings. The probability of a merged IBD state $\mathbf{x}$ can be expressed as an expected value over all possible matches. The probability of a merged IBD state can be approximated by using the independently realized matches $match^{(r)}$ for realizations $r = 1, .., R$. That is,

$$P(\mathbf{x}) = \sum_{matches} P(\mathbf{x}|match)P(match) = E_{matches}[P(\mathbf{x}|match)] \approx \frac{1}{R}\sum_{r=1}^{R} P(\mathbf{x}|match^{(r)}).$$

In the merging procedure, pedigree-IBD is never removed. Consider two cases - when the merge set is composed only of individuals with zero kinship, and when there are individuals with non-zero kinship. If the merge set has only individuals with zero kinship, there is no overlap in the IBD that is detected by `ibd_stitch` and `gl_auto`. Any population-IBD detected by `ibd_stitch` cannot be estimated by `gl_auto` as there are no common ancestors in the pedigree structure. Thus, population-IBD should be added and no pedigree-IBD removed. If the merge set contains individuals that have non-zero kinship, `ibd_stitch` may detect IBD between the individuals that is due to the pedigree relationship. In this case, the `gl_auto` estimates for this pedigree-IBD are preferred. The `gl_auto` estimates are made conditional on the pedigree relationship and the `ibd_stitch` estimates are not. Therefore, pedigree-IBD that is estimated by `gl_auto` should not be removed, even if it is not detected by `ibd_stitch`. Note that more care must also be taken when adding population-IBD in this case, discussed further in 4.4.2.

### *4.4.2 Limitations of Implemented Algorithm*

The output of the implemented algorithm is equivalent to the output of Tarjan's disjoint sets algorithm Gabow and Tarjan [1985], also known as the Union-Find algorithm. The steps described in Algorithm 2 achieve the same output as the Tarjan algorithm applied to a

version of $\bar{\mathbf{m}}$ with the DNA copies of individuals given in a random order within individuals. In the Tarjan algorithm, the partitions would be merged by iterating over pairs of DNA copies. First a "find" step is performed to determine if in the given $\bar{\mathbf{m}}$ the two copies are in the same IBD class, and if they are, performing a "union" step to ensure that the two classes are merged in the output. The Tarjan algorithm is a more efficient method of obtaining the output than the implemented algorithm. The Tarjan algorithm was not used in the analyses presented in this thesis as during development other options for merging were explored, such as checking for existing IBD in the pedigree-IBD realization. While these options were not implemented in the final version, the algorithm as implemented was written to allow these modifications. Two options considered to improve the merging algorithm were the use of marker data and checking for existing IBD in the pedigree-IBD realizations.

Marker data could potentially be used to assist in matching the parental origins of DNA copies in the population- and pedigree-IBD realizations. The challenge in doing this is that marker data at multiple loci and between multiple individuals would need to be used in order to match the DNA copies. This could be investigated further in future research, using ideas from statistical haplotype phasing.

Merging could also be improved by checking for existing IBD relationships in the pedigree-IBD realizations before merging in population-IBD. If the merge set has individuals with non-zero kinship, the population-IBD estimates will contain IBD groups that reflect the parent-child relationship that are already present in the pedigree-IBD realization. When merging a population-IBD group, we would want to check whether the IBD relationship is already represented in the pedigree-IBD. In the implemented algorithm, both pedigree-IBD DNA copy labels from all individuals with at least one copy in the population-IBD group are listed. At this stage, immediately after line 4 in Algorithm 2, a check for existing IBD could be performed. Checking for existing IBD is an important future development for the algorithm, as in practice it may be difficult to select a merging set with no individuals related by the pedigree. It would, however, complicate the merging. There may be more than one existing IBD class that satisfies the population-IBD, or more than one class that partially

satisfies the population-IBD. The order in which classes are merged also becomes important. An existing IBD class may partially satisfy multiple population-IBD classes, some better than others.

### *4.5 Calculation of LOD score from Merged Graphs*

The LOD score equation was introduced in Section 1.6.2, Equation (1.22) and the method of calculating the LOD for pedigree IBD realizations was described in Section 1.6.3. The calculation of the LOD score numerator can be performed in the same manner with the merged realizations $\mathbf{x}^{(r)}$ in place of the pedigree realizations $\mathbf{f}^{(r)}$. The denominator of the LOD score, the base log likelihood, is the probability of observing the trait data caused by an unlinked locus, conditional on the pedigree structure. Methods described in Section 1.6.3 cannot be directly applied as the relatedness modeled is now with respect to the founders of $\mathcal{P}$ and the pedigree structure of $\mathcal{P}$ is unknown. Allowing for additional paths of inheritance in the population pedigree should increase the base log likelihood.

In the pedigree-free LOD score calculations in Glazner and Thompson [2015] a permutation method is used to approximate the base log likelihood. The permutation of trait values over individuals does not take into account the non-exchangeability of the individuals, some of whom are more closely related than others for instance. In the merging analyses, however, we have access to the pedigree structure of $\mathcal{F}$ which describes the closest relationships between the individuals. The simplest strategy is to use the base log likelihood calculated on the pedigree structure for $\mathcal{F}$. The use of the unadjusted base log likelihood was investigated using the simulated data, described below. It was found that if the closest pedigree relationships are correctly specified in the pedigree, the adjustment to the base log likelihood was negligible.

The simulated population is described in Section 2.3. The MORGAN simulation program `simpop_fgl` that was used for the pedigree simulation was modified to return only pedigree information (ID, father ID, mother ID) for the full population of 20,000 individuals per generation over 50 generations. The pedigree structure was stored as a DAG using the R `igraph`

[Csardi and Nepusz, 2006] package, and the most recent common ancestor (MRCA) for all individuals in the "Merge2" subpedigrees were found using the `DAGancestor` function. All of the descendants of the MRCA, excluding those that had no descendants in the Merge2 pedigrees, were used to form the population pedigree. There were a total 277,091 individuals, or 294,020 including dummy founders. Note that this is not the full specification of all possible relationships between the individuals, but it does include the majority of the full population pedigree in the lower generations. Likelihood calculations were performed using MORGAN `mc_null_lods` with trait data on the Merge2 individuals and the rest of the individuals unobserved. Burn in was 10% of MC realizations. Table 4.2 contains a comparison of the summed contributions from the pedigree components (first row) with the base log likelihood calculated with the use of increasingly large subsets of the full population pedigree.

The LOD score contributions were -15.49 and -19.66 from each pedigree in Merge2 giving a total -35.16 compared to -34.36 using the population pedigree graph. The increasing size of the population pedigree shows the LOD score stabilizing quickly with little change to the LOD score after adding just two individuals. The two individuals are the parents of the two siblings whose relationship was not specified in the Merge2 pedigree structure. The use of the summed base log likelihood contributions from the unmarked pedigree components as a base log likelihood for the merged graph is a good approximation provided that close pedigree relationships are correctly specified.

### *4.6 Effect of adding IBD on LOD scores*

This section contains a discussion of the effect that adding IBD through merging has on the LOD score. When IBD is added through merging, the numerator of the LOD score given in Equation (1.22) is affected. The numerator is the probability of the trait data given the IBD states, $P_{H_1}(\mathbf{Y} = \mathbf{y}|\mathbf{x})$, and is changed when $\mathbf{x}$ is changed. In this section, a simulated data scenario is used to demonstrate the effects of adding IBD to $\mathbf{x}$. The effect on the LOD score will typically, but not always, be an increase in the value of the LOD score if IBD is added between phenotypically similar individuals. As the merging procedure only adds IBD,

| Num. Indivs | Num. Switches | Time (hh:mm:ss) | Num. MC realizations | LOD | SD |
|---|---|---|---|---|---|
| 51 | | | | -35.15818 | |
| 53 | 66 | 0:00:01 | 10,000 | -34.38005 | 0.2886 |
| 222 | 136 | 0:00:01 | 10,000 | -34.35338 | 0.3077 |
| 427 | 240 | 0:00:01 | 10,000 | -34.36214 | 0.3074 |
| 833 | 452 | 0:00:01 | 10,000 | -34.35797 | 0.3100 |
| 833 | 452 | 0:00:02 | 100,000 | -34.36157 | 0.3083 |
| 833 | 452 | 0:02:25 | 1,000,000 | -34.35984 | 0.3093 |
| 3133 | 1766 | 0:05:33 | 1,000,000 | -34.36028 | 0.3106 |
| 18731 | 15805 | 0:31:15 | 1,000,000 | -34.36142 | 0.3128 |
| 42685 | 52804 | 1:34:54 | 1,000,000 | -34.36222 | 0.3139 |

Table 4.2: Comparison of base log likelihood calculated with only individuals in Merge2 pedigrees (first row) to increasingly large subsets of the population pedigree.

the LOD score will almost always be increased after merging. This is the motivation behind being conservative with the IBD that is added to avoid false positive IBD that will inflate the LOD score.

Assume that there are sufficient data to determine the FGL partition at a locus. We will consider the two small pedigrees given in Figure 4.2, box A with affected individuals colored orange. Potential alternative IBD graphs for these pedigrees are given, labeled B through J, with numbers indicating arbitrary FGLs and colored boxes indicating IBD added between founders in graph A. The change in LOD score is compared for each graph relative to graph A without any adjustment in the base log likelihood, trait values or mode of inheritance. These changes are given in Table 4.3. Dominant and recessive trait models for a causal allele $B$ with frequency $P(B) = 0.1$ are compared. In the dominant model the penetrances are $P(\mathbf{Y} = 1|AA) = 0.001$ and $P(\mathbf{Y} = 1|AB) = P(\mathbf{Y} = 1|BB) = 0.99$. In the recessive mode

of inheritance the penetrances are $P(\mathbf{Y} = 1|AA) = P(\mathbf{Y} = 1|AB) = 0.001, P(\mathbf{Y} = 1|BB) = 0.99$. Observed and unobserved refer to whether the trait status of the unaffected individuals is observed. If the trait status is not observed it is coded as unknown.

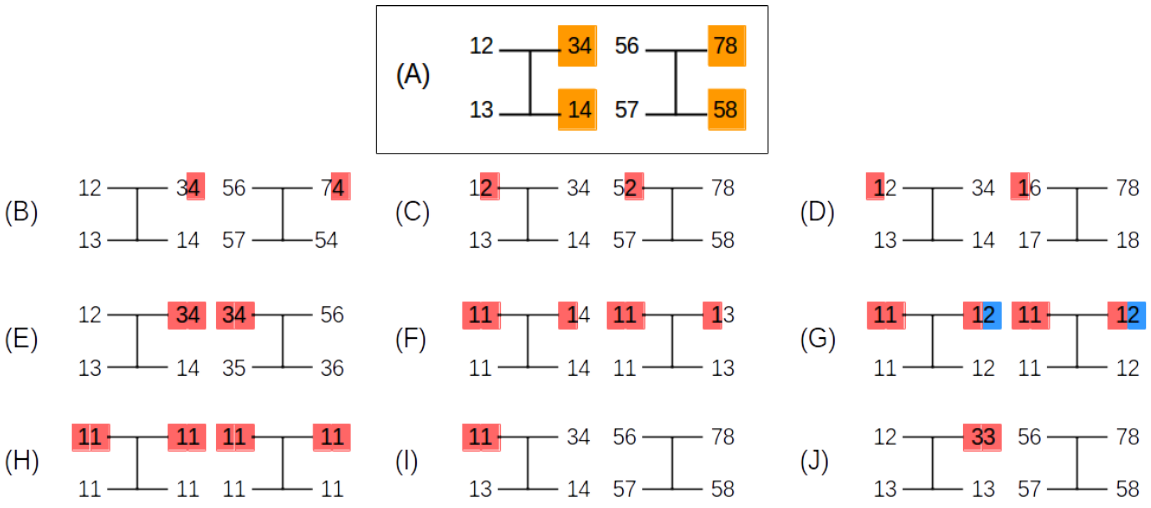



Figure 4.2: Example pedigree with alternative FGL patterns. Graph A is the baseline IBD sharing pattern, with shaded individuals affected with the trait. Graphs B-J contain additional IBD sharing relative to A.

Adding IBD between affected individuals, as in B, results in an increase in the LOD score. When adding IBD between unaffected individuals, as in C and D, the larger the IBD group that is formed the larger the increase in the LOD score. In C the new IBD group is two chromosomes and in D six chromosomes. The effect sizes are larger if the unaffected individuals are observed to have unaffected trait status rather than unknown trait status. The same pattern is present in the recessive model. When IBD is added between an affected and an individual with unobserved trait status as in E, there is an increase in likelihood. If the person with unobserved trait status is in fact unaffected, however, we see a decrease. In general, the addition of the IBD increases the likelihood if there are fewer FGLs between

| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| Dominant Unobserved | 0.00 | 0.92 | 0.00 | 0.03 | 0.03 | 0.96 | 1.20 | 0.94 | 0.00 | -0.04 |
| Dominant Observed | 0.00 | 1.00 | 0.04 | 0.05 | -1.91 | 0.27 | 1.23 | 1.29 | 0.05 | -1.95 |
| Recessive Unobserved | 0.00 | 1.00 | 0.00 | 0.99 | 0.99 | 2.98 | 3.98 | 5.00 | 0.00 | 1.00 |
| Recessive Observed | 0.00 | 0.97 | 0.02 | 0.62 | -0.69 | 4.13 | 5.13 | 8.15 | -0.43 | 0.82 |

Table 4.3: LOD score changes under different IBD graphs, dominant and recessive modes of inheritance

potentially phenotypically similar individuals but not if the individuals are phenotypically different. If the individuals are phenotypically different this forces the individual to have unaffected trait status and at least one copy of the causal allele, which has low probability. The effect of size of the introduced IBD groups is also illustrated in F, G and H. In F there are three remaining FGLs, in G there are two and in H one. Although IBD is introduced between phenotypically different individuals the size of these groups overcomes the negative effects. In F we have the same sharing that caused a decrease in E, but the overall effect is an increase in likelihood. Observing unaffected individuals increases the magnitude of the effect. In I we see the effects of adding IBD between the two copies of an individual who is unaffected. When this individual is observed there is an increase in likelihood in the dominant model and a decrease in the recessive model. The dominant model allows a single copy of the causal allele $B$ to be shared between phenotypically different individuals but the recessive does not. When summing over all possible genotypes the recessive model is forced into the homozygous $BB$ case. This does not happen in D for instance, where there is sharing between phenotypically dissimilar individuals but without the forcing of a homozygote. Similarly, in J we see the effect of adding IBD between the two copies of an affected individual. This increases the likelihood in the recessive case as the affected individual is guaranteed two identical copies. However, in the dominant case the likelihood decreases.

Consider the calculation of $P(\mathbf{Y}|\mathbf{x})$. Take the example IBD graphs A and J, with just two observed and affected individuals in the left hand pedigree. The calculation can be expressed as

$$P(\mathbf{Y}|\mathbf{x}) = \sum_{\alpha(\mathbf{x})} P(\mathbf{Y}|\alpha(\mathbf{x}))P(\alpha(\mathbf{x})) \tag{4.9}$$

where $\alpha$ are possible allelic assignments to each FGL. This becomes

$$A : P(\mathbf{Y}|\mathbf{x} = A) = \sum_{3\in\{A,B\}} \left[ \sum_{4\in\{A,B\}} \left[ P(\mathbf{Y}|34)P(3)P(4) \sum_{1\in\{A,B\}} P(\mathbf{Y}|14)P(1) \right] \right] \tag{4.10}$$

$$J : P(\mathbf{Y}|\mathbf{x} = J) = \sum_{3\in\{A,B\}} \left[ P(\mathbf{Y}|33)P(3) \sum_{1\in\{A,B\}} P(\mathbf{Y}|13)P(1) \right] \tag{4.11}$$

In graph A compared to B there is multiplication by an additional allele frequency $P(4)$ which decreases $P(\mathbf{Y}|\mathbf{x})$ . There is also the addition of probabilities of allelic assignments that are different between FGL 3 and 4 which increases $P(\mathbf{Y}|\mathbf{x})$ . It is thus possible to decrease the LOD score even when adding IBD between two affected individuals. Although there are fewer allelic copies to multiply there can be a greater probability contribution from the extra allelic configurations. In this case the probability of a heterozygote in the affected parent is larger.

In summary adding IBD between individuals who are phenotypically similar will typically increase the LOD score numerator. The amount of the increase depends on the size of the resulting IBD group with a larger group generally giving a larger increase in LOD score. As the IBD added propagates to other individuals the phenotypes of all the individuals with DNA copies in the group affects the change in likelihood. The trait model has an important impact on the size and direction of the effect especially when unlikely IBD-phenotype combinations are introduced. Counterintuitive changes in the likelihood can occur when autozygous IBD is introduced. Although IBD was added between individuals with identical phenotypes in the simulated scenario, it eliminated the possibility of a heterozygote which had the effect of an overall reduction in likelihood.

# Chapter 5

# MERGING ALGORITHM WITH SIMULATED DATA

This chapter contains the results of applying the merging algorithm described in Chapter 4 to two data sets. The analysis on the GAW data set is in Section 5.1. The analysis on the simulated pedigrees data set is in Section 5.2.

## *5.1 Analysis of GAW Data*

The GAW data set is described in Chapter 2, Section 2.4. Blue et al. [2014] performed a pedigree-only linkage analysis on this data for GAW18. An early version of the merging method was used for GAW19, published in Saad et al. [2016], and described in in Section 5.1.1. The Chapter 4 version of the merging method was used to update these results, described in Section 5.1.2. While marker genotype data was available on all odd numbered chromosomes, focus was restricted to chromosome 3 which contains the MAP4 gene at position 69cM, the variant that made the largest contribution to the expression of the simulated trait.

Blue et al. [2014] performed a linkage analysis on seven disjoint pedigrees (numbers 5, 6, 7, 8, 10, 21, 25) with total 529 individuals that showed evidence of cryptic relatedness. Cryptic relatedness is recent common ancestry between individuals that is not reflected in the given pedigree structure. Cryptic relatedness was detected by Thornton et al. [2014], by estimating kinship coefficients on all pairs of individuals from all available marker data. Blue et al. [2014] selected a panel of 351 markers over the 224cM of chromosome 3 at average intermarker distance 0.64cM. The linkage analysis was performed using MORGAN `gl_auto` realizations of pedigree FGLs, $\mathbf{f}^{(r)}, r = 1, ..., 1000$. Three trait models were tested, the best of which was Model 1, developed based on epidemiology in Bonaa and Thelle [1991]. The trait

allele frequency in this model is $P(D) = 0.0318$ and genotype means are 0.3958, -5.8277 and -12.0512 for 0, 1 and 2 copies respectively of the causal allele at the trait locus, with within genotype variance of 78.8972. The parameters were defined by the SNP with the biggest contribution to the simulated trait variance, in the MAP4 gene. The SNP only explains 0.0229% of the trait variance, so model 1 tests whether we can detect a locus with a small effect size if it is modeled perfectly.

### *5.1.1 GAW 19 Linkage Analysis*

This section describes the merging analysis for GAW 19 published in Saad et al. [2016]. A merging set $\mathcal{M}$ was selected from pairs of individuals who were not related in the pedigree structure but showed high levels of IBD sharing genome-wide. The 17 pairs contain 21 unique individuals from all the 7 pedigree components. Many of the 17 pairs of individuals share IBD from 50 to 75 cM along chromosome 3 and all pairs gave a strong signal of IBD 69 cM along the chromosome at the simulated trait locus. The pairs were previously identified in GAW18.

A denser marker panel $\Lambda$ was selected that included all linkage panel marker loci $\lambda$ from Blue et al. [2014]. $\Lambda$ contains a total of 48,892 marker loci with an average intermarker distance of 0.005cM. Genotypes were observed at all loci for all individuals and minor allele frequencies were all greater than 0.05 in the genotyped individuals.

Realizations of pedigree FGLs $\mathbf{X}^{\mathcal{F}}$ on all individuals in $\mathcal{F}$ at sparse loci $\lambda$, denoted $\mathbf{f}^{(r)}, r = 1, ..., 1000$ were reused from Blue et al. [2014]. Realizations of population FGLs $\mathbf{X}^{\mathcal{P}}$ on individuals in $\mathcal{M}$ at dense loci $\Lambda$, denoted $\mathbf{m}^{(r)}, r = 1, ..., 1000$ were made using `ibd_stitch`. Parameters used (see Section 1.5.2) were a population kinship $\beta = 0.05$,change rate $\alpha = 0.05$, null fraction 0.05, and genotyping error rate 0.01.

It was found that a linkage analysis using `gl_lods` on a merged graph over all 529 individuals on all seven pedigrees was not computationally feasible. The analysis was split into two groups: those individuals in pedigrees 5, 6, 21 and 25, and those individuals in pedigrees 10, 8 and 7. The two components have an approximately equal number of individuals and

none of the cryptically related pairs are split between the two groups. The population FGL realizations were estimated over all individuals in $\mathcal{M}$. The realizations were later partitioned into these two groups and merging was performed separately on each group.

Merging for this analysis was done using an earlier version of the algorithm described in Chapter 4. The matching algorithm used was Algorithm 2 of Glazner [2014]. In this version each realization of `ibd_stitch`, $\mathbf{m}^{(r)}$, was paired with a realization of `gl_auto`, $\mathbf{f}^{(r)}$, and merging was performed on the two realizations. There was no calculation of a consensus graph. In each merge, for each individual that appeared in both the $\mathbf{m}^{(r)}$ and $\mathbf{f}^{(r)}$ graphs, an arbitrary matching was made between its two DNA copies in $\mathbf{m}^{(r)}$ and its two DNA copies in $\mathbf{f}^{(r)}$. There was no accommodation for unknown paternal origin; it is expected that the correct DNA copies will match half the time. After matching, the DNA copies in $\mathbf{m}^{(r)}$ that were indicated to be IBD were identified and the corresponding entries in $\mathbf{f}^{(r)}$ were altered to have the same FGL.

Figure 5.1 has the LOD scores from this merging analysis. The vertical line indicates the trait location. The dashed LOD score is the average over the 200 traits of the summed unmarked pedigree component LOD scores. The grey lines are the merged graph LOD scores for each of the 200 simulated traits and the black line indicates the mean merged LOD over the 200 traits.

The unmerged analysis identified a region of interest from around 50-75cM along the chromosome with LOD scores around 1.5. The merged analysis showed a clear increase in the LOD score in this region with LOD scores of around 4. The merging also resulted in inflation of the LOD score along the whole chromosome with loci as far as 100cM away from the trait locus increasing in LOD score from around 0 to around 1. There is also a high variance from locus to locus in the merged LOD score even when averaged over traits. The large differences between adjacent loci are due to very different realized IBD states at the loci both in the same realization and between realizations. This is a result of both very short segments of IBD identified by `ibd_stitch` and flaws in the merging algorithm that were later updated.

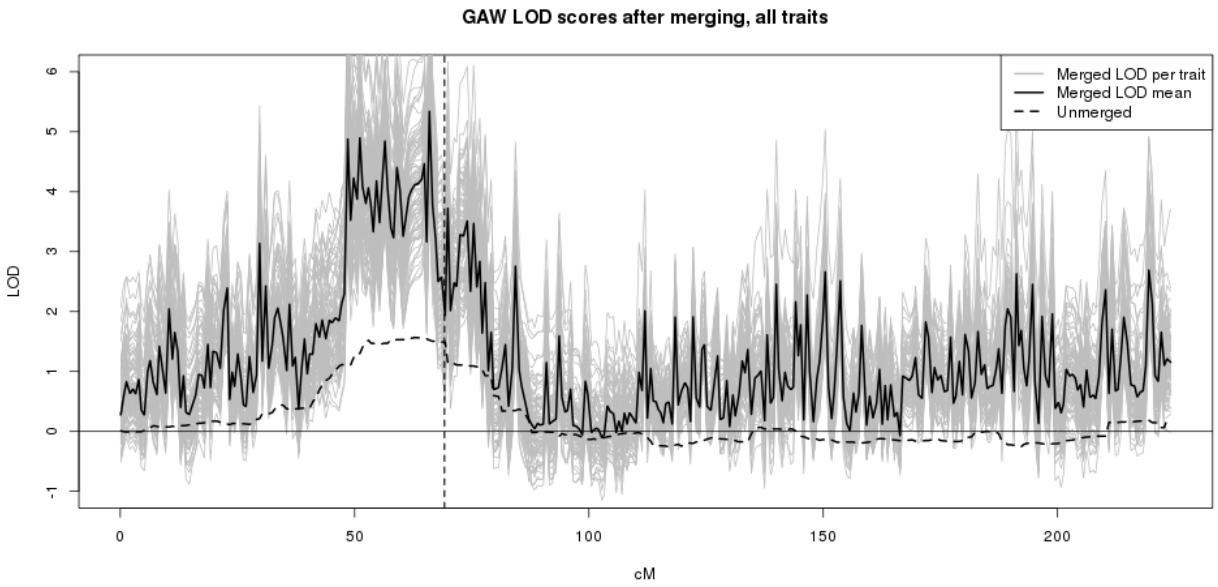



Figure 5.1: GAW pedigrees; LOD scores for 200 simulated traits.

A consensus procedure was later conducted on the population-IBD estimates from `ibd_stitch` to evaluate the quality of the IBD estimates. Figure 5.2 shows the segments in the consensus partition of the `ibd_stitch` outputs at the 80% threshold for the component containing individuals in pedigrees 5, 6, 21 and 25. Only this component is plotted as it made a much larger contribution to the LOD score than the other component containing individuals in 10, 8 and 7. The blue IBD segments are IBD added between the two chromosome copies of one individual, the red segments are IBD between chromosomes in different pedigrees and the black segments are IBD between chromosomes from the same pedigree but different individuals.

While we expect short segments of IBD between population members, the large number of very short population-IBD segments and lack of long segments even between known parents, children and siblings indicates that `ibd_stitch` estimates were of poor quality. For example, the bottom line of the figure shows IBD between individuals 05_0371 and 05_0380. Although

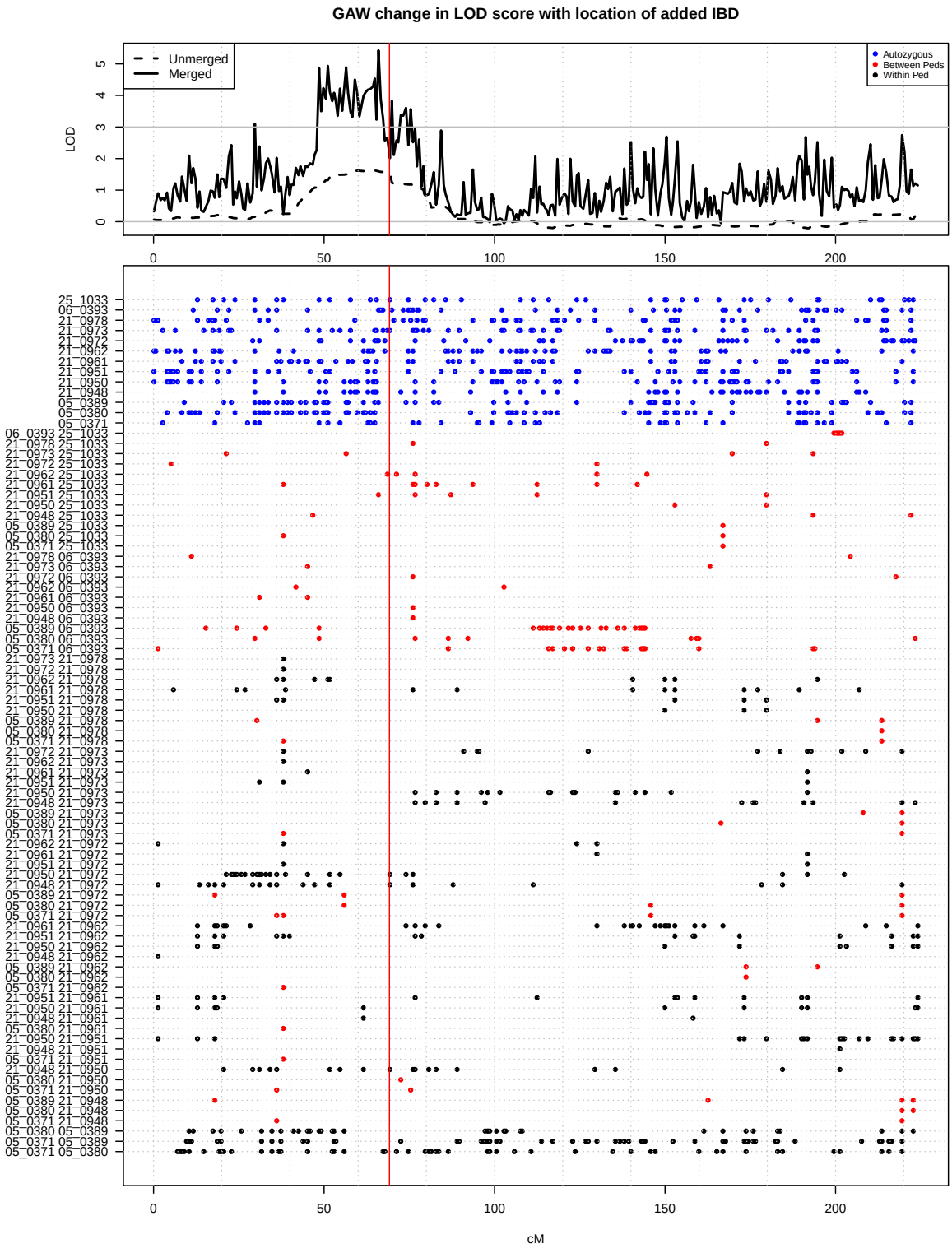



Figure 5.2: Change in LOD score for GAW dataset after merging, compared to IBD consensus partition at 80% threshold for group 1.

these individuals are parent and child and we would expect to see long segments of IBD, the `ibd_stitch` estimates switch in and out of IBD with very short segments. The large number of individuals in the analysis (21) made the estimation of the joint graph difficult. Furthermore, close pedigree relationships between certain individuals are not modeled in the `ibd_stitch` model Glazner and Thompson [2015]. The `ibd_stitch` parameters, rather than the pedigree, give a prior on IBD. The parameters used in this analysis had a high $\alpha$ and low $\beta$ for IBD between a parent-child pair.

The earlier version of the algorithm also merges all estimated IBD without checking for existing IBD, taking into account uncertain phasing, or performing any quality control on the `ibd_stitch` estimates. In the case of close relatives a large amount of IBD is being estimated both by pedigree and population methods. The merging step did not include a check for existing IBD, so the same IBD may be added twice. For example, for two individuals with DNA copies (A,B) and (C,D), there may be pedigree-IBD already present between A and C. If `ibd_stitch` indicates IBD between these individuals as due to the pedigree relationship, it may be correctly placed between A-C but is equally likely to be placed between A-D or B-C which result in a 3 DNA copy IBD group, or B-D which results in two IBD pairs.

Figure 5.2 shows the large amount of autozygosity estimated by `ibd_stitch`. All 13 individuals have autozygosity along the length of the chromosome. Autozygosity is detected when there are long stretches of homozygosity along the two chromosome copies belonging to the individual. The estimated autozygosity is not a result of excess homozygosity in these individuals. Excess homozygosity compared to what is expected under Hardy-Weinberg equilibrium, can be a result of the population being a mixture of sub-populations with different allele frequencies. This is possible in a Mexican American population. However, a Fisher's exact test for decreased heterozygosity [Fisher, 1922, Levene, 1949] found only 2811 out of 48892 loci, or 5.7%, significant at the 5% level without accounting for multiple testing. Fisher's exact test is a directional test for the distribution of the cells of a contingency table. The p-value, which can be computed exactly, is the probability of observing a sample configuration that is even less likely than the one being evaluated, conditional on the observed

allele counts.

The estimated autozygosity is likely due to a lack of haplotypic variation in the population and strong local LD. Local LD can be a result of mixture and admixture in the Mexican American population [Falush et al., 2003]. A mixture of sub-populations with differing allele frequencies at loci along the genome leads to correlations among markers along the genome. Admixture also causes LD due to unbroken segments of chromosome inherited from one or other ancestral population.

### *5.1.2 Updates to GAW analysis*

The analysis in Section 5.1.1 can be improved with better selection of individuals for $\mathcal{M}$, better population-IBD detection, and the use of the updated merging method. The updated analysis is described in this section.

A new subset of individuals for merging, $\mathcal{M}_2$, was selected. Criteria for an optimal merge set are described in Section 4.2. The original merge set was $\mathcal{M}_1$ had 21 individuals and $\mathcal{M}_2$ has only 5 individuals to improve the quality of `ibd_stitch` estimates. The individuals in $\mathcal{M}_2$ were selected so that all pairwise kinships were zero. No pairs were related so there was no overlap between pedigree-IBD and population-IBD estimation. The five individuals were selected from the same list of pairs identified to be highly cryptically related. They are 5_371, 6_393, 7_1151, 10_566 and 10_616. As pedigrees 8, 21 and 25 had none of the merging subset and no pedigree-only linkage signal, these were not used in the analysis.

To reduce LD between the dense markers from the original GWAS analysis a new dense panel was selected. A panel with less LD should result in fewer false-positive IBD signals. All of the 48,892 SNPs on chromosome 3 were used in the population-IBD estimation panel in the GAW 19. For the updated analysis every 5th marker was selected giving 9779 markers. This increased the average intermarker distance from 0.0046cm, or 217 per cM, to 0.23cm, or 43 per cM. This panel is still denser than the linkage panel which had only 351 markers and an intermarker distance of 0.62cM.

Population-IBD was again estimated using `ibd_stitch`. The parameters were $\alpha = \beta =$

0.03 with genotype error 0.01. Reductions of $\alpha$ to 0.01 and 0.0001 to increase estimated segment length and reduce the number of switches in and out of IBD states were tested, but had only minor effects on the `ibd_stitch` consensus. The population-IBD was merged into the pedigree-IBD using the algorithm in Chapter 4.

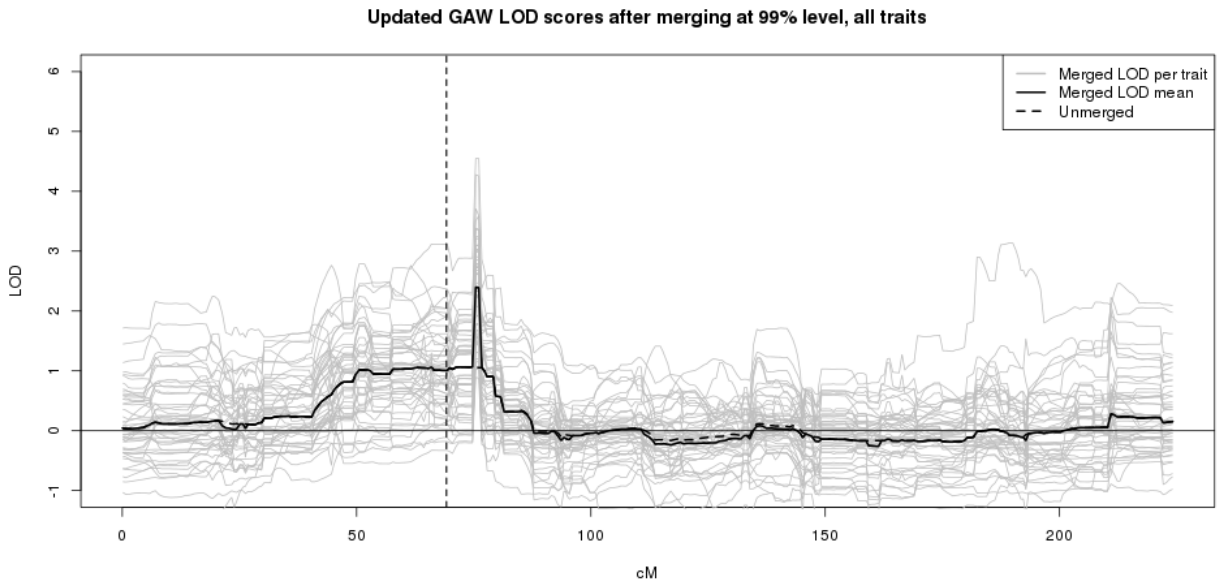



Figure 5.3: GAW pedigrees; Updated LOD scores for simulated traits.

Linkage analysis results are given in Figure 5.3. The figure has a grey line for the LOD score for each of the 200 simulated trait realizations with a black line for the average over trait realizations. Compared to the original analysis in Figure 5.1, Figure 5.3 has a smooth LOD score, with a single major increase in LOD score around 75cM along the chromosome which increases the LOD from approximately 1 to 2. The sharp increase is not exactly at the simulated trait locus at 69cM, but is in the region from 40-80cM with an LOD score signal. There is no longer inflation of the LOD score at loci more distant from the trait locus.

Figure 5.4 shows the average LOD over trait realizations, and the population-IBD segments estimated by `ibd_stitch`. Compared to Figure 5.2, the IBD estimate quality has

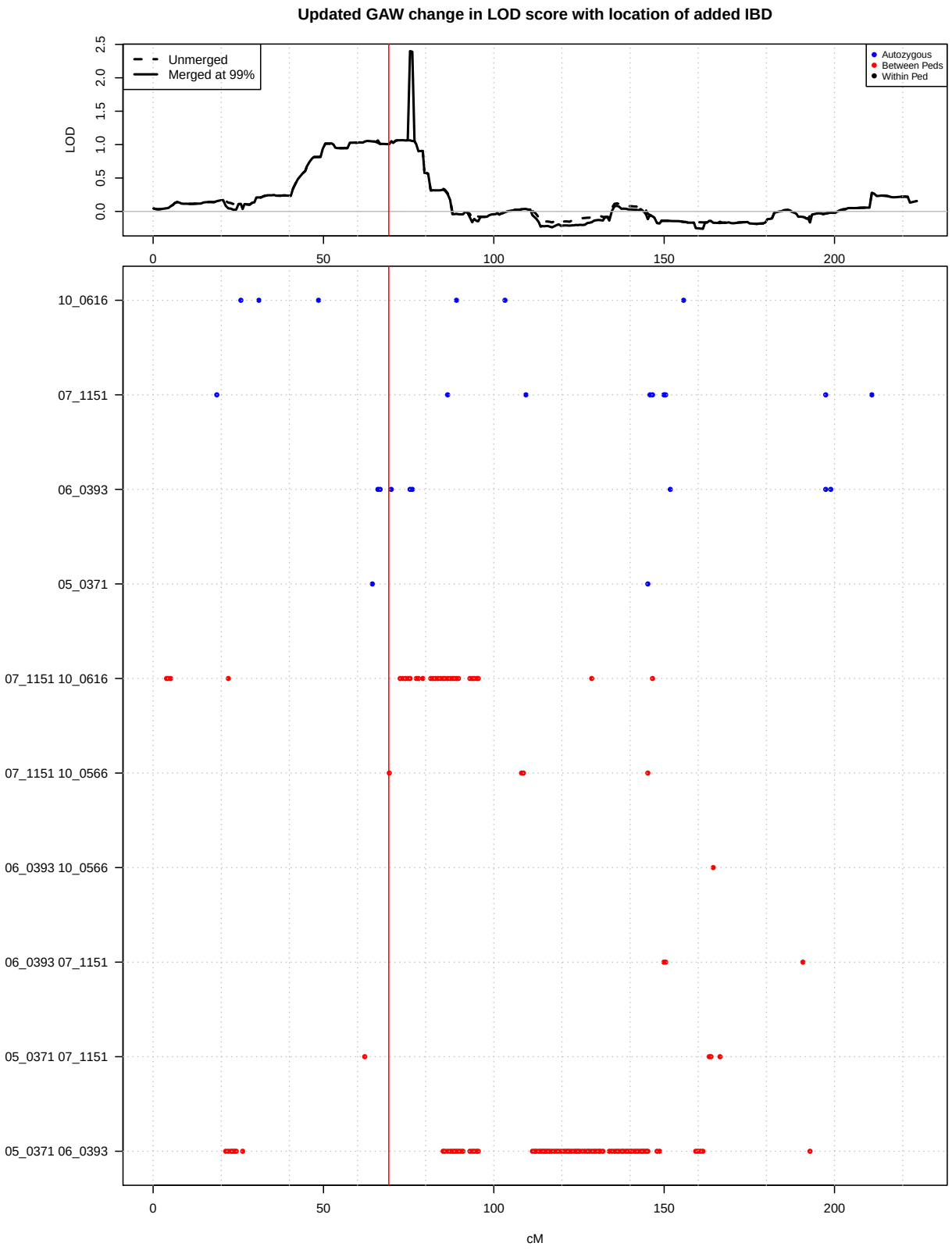



Figure 5.4: Change in LOD score for GAW dataset with updated analysis, merged at 99% threshold.

improved. Segments in the updated analysis have much less autozygosity and the segments that are present are longer and less broken. This can be seen, for example, in the segment between 07_1151 and 10_0616, and 05_0371 with 06_0393. The main increase in LOD score is due to a segment of autozygosity in 06_0393 at around 75cM which corresponds to a 1.7cM segment of unbroken homozygosity in this individual. This IBD segment requires further investigation to determine if it can be discounted as the result of a genotyping artifact. Stretches of what appears to be homozygous loci can in fact be hemizygous. The individual may have only one copy of the chromosome in this region of the genome due to deletion, or only one of their two copies was able to be typed. The longer IBD segments between individuals in different pedigrees (red segments in Figure 5.4) made a minimal difference to the average LOD score - the merged and unmerged LODs are very close.

### *5.1.3 Summary*

Overall the analysis of this dataset performed for GAW 19 [Saad et al., 2016], described in Section 5.1.1, indicated the potential of the merging approach to introduce informative IBD information and enhance the linkage signal. The LOD score was increased from an average of 1.5 to 4 around the trait locus. There was, however, inflation of the LOD score over the whole chromosome, including loci at the other end of the chromosome. The merged LOD score was also a very jagged curve. Improvements needed to be made to ensure the quality of IBD estimates and to better take into account uncertain phasing and existing pedigree-IBD in the merging algorithm. An analysis with these updates was performed, described in Section 5.1.2, that showed the updated algorithm and steps to ensure IBD quality resulted in a much smoother LOD signal and no inflation at unlinked loci. There was, however, no increase in LOD at the causal locus. The spike in LOD caused by autozygosity in one individual indicates a possible data quality issues. The autozygosity can be genuine due to inbreeding, or an artifact of the sample such as hemizygosity.

## *5.2 Analysis of Simulated Pedigree Data*

Chapter 2 Section 2.1 describes the simulation of the population, extraction of pedigrees and simulation of trait. In this section the simulated pedigree sets Merge2 and Merge4, described in Section 2.3, are used for merging. Section 5.2.1 describes the creation of the merged graphs. Section 5.2.2 compares the consensus graphs to the simulation truth for different thresholds, and Section 5.2.3 compares the merged graphs to the simulation truth for different thresholds and using true and estimated population IBD for merging. Use of merged graphs in LOD score analysis with a simulated trait is described in Section 5.2.4. The section is summarized in 5.2.5.

### *5.2.1 Creation of Merged IBD Graphs*

To form merging subsets, individuals were selected from the founders of the simulated pedigrees. The pedigrees in Merge2 have a total of 19 founders and the pedigrees in Merge4 have a total of 57 founders. Subsets of increasing size were compared. For Merge2 the smallest set was a pair of founder individuals, one from each pedigree, whose true relationship was siblings. In Merge4 a subset of 9 founder individuals were selected for high levels of IBD sharing.

The simulation truth IBD states were obtained from the simulated population. Estimated IBD on the pedigrees is compared to the true pedigree-IBD relative to the pedigree founders. Estimated population-IBD among individuals in the merging set is compared to the true population-IBD relative to the population founders. After merging, the merged pedigree- and population-IBD estimates are also compared to the simulation truth.

Pedigree-IBD was estimated with the MORGAN program `gl_auto`. A total of 50,000 MCMC iterations with a burn in of 500 iterations were performed on the pedigrees, saving every 50th iteration as a realization of pedigree-IBD. Population-IBD was obtained from 1000 realizations of `ibd_stitch` on the merging subset using parameters $\alpha = 0.05, \beta = 0.05$, null fraction 0.05, no genotyping error and assuming unphased data.

Consensus partitions were made at threshold levels from 51% to 99% with 5 runs of the consensus algorithm, as discussed in Section 4.3. In Section 5.2.2 the effectiveness of `ibd_stitch` is compared for different numbers of DNA copies and different thresholds. Merging was done using the method in Chapter 4. Descriptions of the merged graphs are given in Section 5.2.3. LOD score results are given in Section 5.2.4.

### *5.2.2 Consensus at different thresholds*

The performance of the consensus method assuming unknown parental origin was tested on different size sets of individuals at different thresholds. The results for these are given in Table 5.1. The sets are the 2 individuals from Merge2 and all 19 founders of Merge2 and from Merge4 the set of 9 founders and the set of all 57 founders. The perfomance is measured in RRMSE, defined in Equation 4.8.

`ibd_stitch` performs very well for a set of two individuals at all thresholds. For 9 individuals `ibd_stitch` performs reasonably well, with best results at the 80% threshold - the RRMSE for the largest class is 0.15. At 80% there is a correlation of around .6 between true and estimated values for the size of largest class, the number of single DNA copies, the average class size and the number of classes. For 19 individuals the performance of `ibd_stitch` is not as good, but the best results are again around 80%. The RRMSE for the largest class is now 0.26. At 80% there is a correlation of .36 in the largest class size and around .45 on the other measures. For 56 individuals the performance of `ibd_stitch` is very poor. Again the optimal threshold is around 80% but the RRMSE is now 0.35 for the largest class. There is a correlation of 0.34 for the largest class, but negative or very low correlations on the other measures. Error in average class size is very low across the board due to the class sizes being either 1 or 2 DNA copies in almost all cases.

These results indicate that `ibd_stitch` estimates are reliable for smaller groups of less than 10 individuals but unreliable for larger sets. This behavior is expected for `ibd_stitch`, as discussed in Section 1.5.2.

Figure 5.5 compares the performance of the estimation method on the siblings in the

| | | Largest Class | Single Copies | Avg Class Size | Number Classes |
|---|---|---|---|---|---|
| 2 indivs | 51 | 0.00 | 0.49 | 0.03 | 0.03 |
| | 60 | 0.00 | 0.49 | 0.04 | 0.04 |
| | 70 | 0.00 | 0.48 | 0.04 | 0.04 |
| | 80 | 0.00 | 0.48 | 0.04 | 0.04 |
| | 90 | 0.04 | 0.49 | 0.05 | 0.06 |
| | 99 | 0.08 | 0.53 | 0.06 | 0.09 |
| 9 indivs | 51 | 0.40 | 0.12 | 0.09 | 0.07 |
| | 60 | 0.33 | 0.10 | 0.07 | 0.05 |
| | 70 | 0.23 | 0.07 | 0.04 | 0.04 |
| | 80 | 0.15 | 0.07 | 0.03 | 0.03 |
| | 90 | 0.22 | 0.09 | 0.04 | 0.04 |
| | 99 | 0.41 | 0.12 | 0.06 | 0.06 |
| 19 indivs | 51 | 0.45 | 0.12 | 0.09 | 0.07 |
| | 60 | 0.38 | 0.11 | 0.07 | 0.06 |
| | 70 | 0.33 | 0.09 | 0.06 | 0.05 |
| | 80 | 0.26 | 0.08 | 0.05 | 0.04 |
| | 90 | 0.36 | 0.11 | 0.05 | 0.05 |
| | 99 | 0.41 | 0.13 | 0.06 | 0.06 |
| 57 indivs | 51 | 0.35 | 0.08 | 0.04 | 0.04 |
| | 60 | 0.23 | 0.07 | 0.03 | 0.03 |
| | 70 | 0.24 | 0.08 | 0.04 | 0.04 |
| | 80 | 0.35 | 0.10 | 0.05 | 0.05 |
| | 90 | 0.43 | 0.10 | 0.05 | 0.05 |
| | 99 | 0.52 | 0.11 | 0.05 | 0.05 |

Table 5.1: Consensus IBD partitions from `ibd_stitch` compared to simulation truth for sets of individuals of varying size. Statistics are RRMSE, truth is simulation truth pedigree IBD and population IBD of merge set. Methods are assuming unknown (U) parental origin at different thresholds.

2 individual set from Merge2 who also appear in the 19 individual set. The same marker data and IBD model for the siblings was used in both cases but when the estimation was performed on the larger set the estimation was less accurate. The confidence in the IBD is lower - there is only IBD present at lower thresholds and when present fewer copies are estimated to be shared. This is due to the sequential building of the IBD state, where IBD states among previously added individuals restrict the possibilities for subsequent individuals and can result in the correct state being unavailable - the IBD between the pair is restricted by the other 17 individuals.

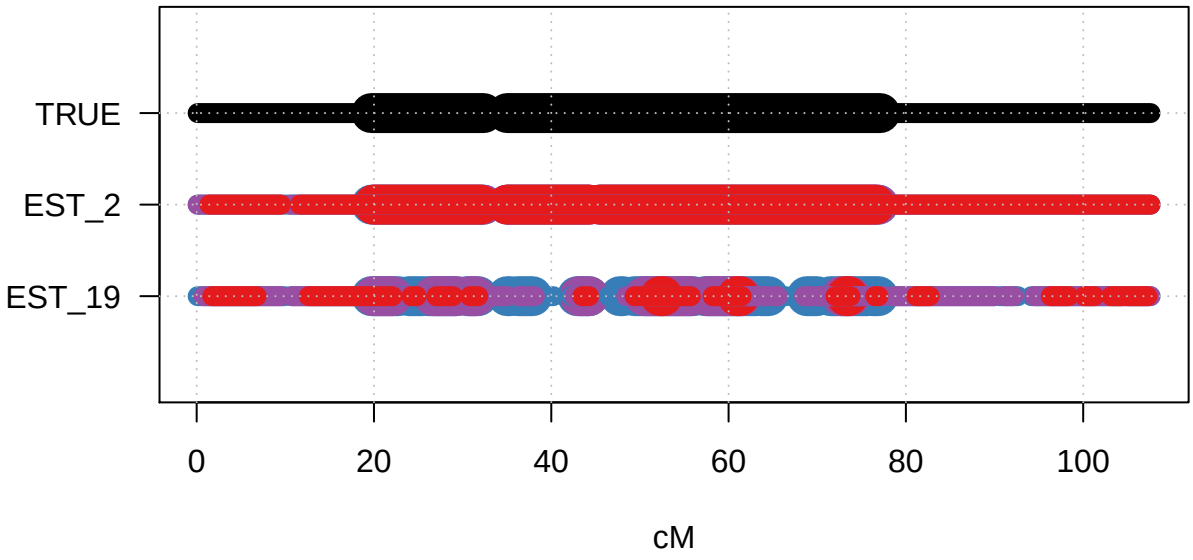



Figure 5.5: IBD segments for pair of siblings. True IBD (black) compared to estimated IBD from either the set of 2 or 19 individuals. Colors indicate threshold of 80% (blue), 90% (purple) and 99% (red), and the pair share either one (thin line) or two (thick line) copies IBD.

The full depiction of IBD states, true and estimated, for the set of 19 individuals is in Figure 5.6. Where there are black true segments but no colored estimated segments IBD

has been missed, a false negative. Likewise, where there are colored segments but no black segment there is a spurious false positive IBD segment. Spurious IBD segments tend to be short as they are caused by local LD. Figure 5.7 shows the tendency for spurious IBD segments to be estimated where there is high local LD.

### *5.2.3 Merging using true and estimated population IBD*

Measures of the closeness of the merged graphs to the true IBD is given in Table 5.2. For two individuals the reduction in RRMSE from pedigree-only to merged-true is small. As the estimated IBD is close to the true IBD, as seen in 5.1, there is little difference between using the true and estimated IBD. The difference between merged-true and true increases as set size increases. The difference between merged-true and merged-estimate also increases as set size increases, due to the reduction in IBD estimate quality. In the set of 9 individuals the quality of population-IBD estimates produced by `ibd_stitch` were better than in the set of 57. A higher threshold on population-IBD also ensures IBD quality, at the risk of excluding true IBD segments that are not estimated with a high probability.

### *5.2.4 LOD scores on merged graphs*

Next we examine the change in LOD scores when the true population IBD is merged into estimated pedigree IBD. The merging method performance is thus examined without the population IBD estimate quality as a factor.

Figure 5.8 shows average LOD scores over 50 trait realizations for the Merge2 pedigrees. Four LOD scores are plotted. The black lines are LOD scores based on IBD relative to the pedigree founders - simulation truth and estimated by `gl_auto`. The green lines are LOD scores for merged IBD - the true population-IBD between the pedigree founders relative to the population founders merged in to the estimated pedigree IBD, and the simulation truth. The figure shows that the target true IBD state is able to be recovered well through the merging algorithm given the true population IBD.

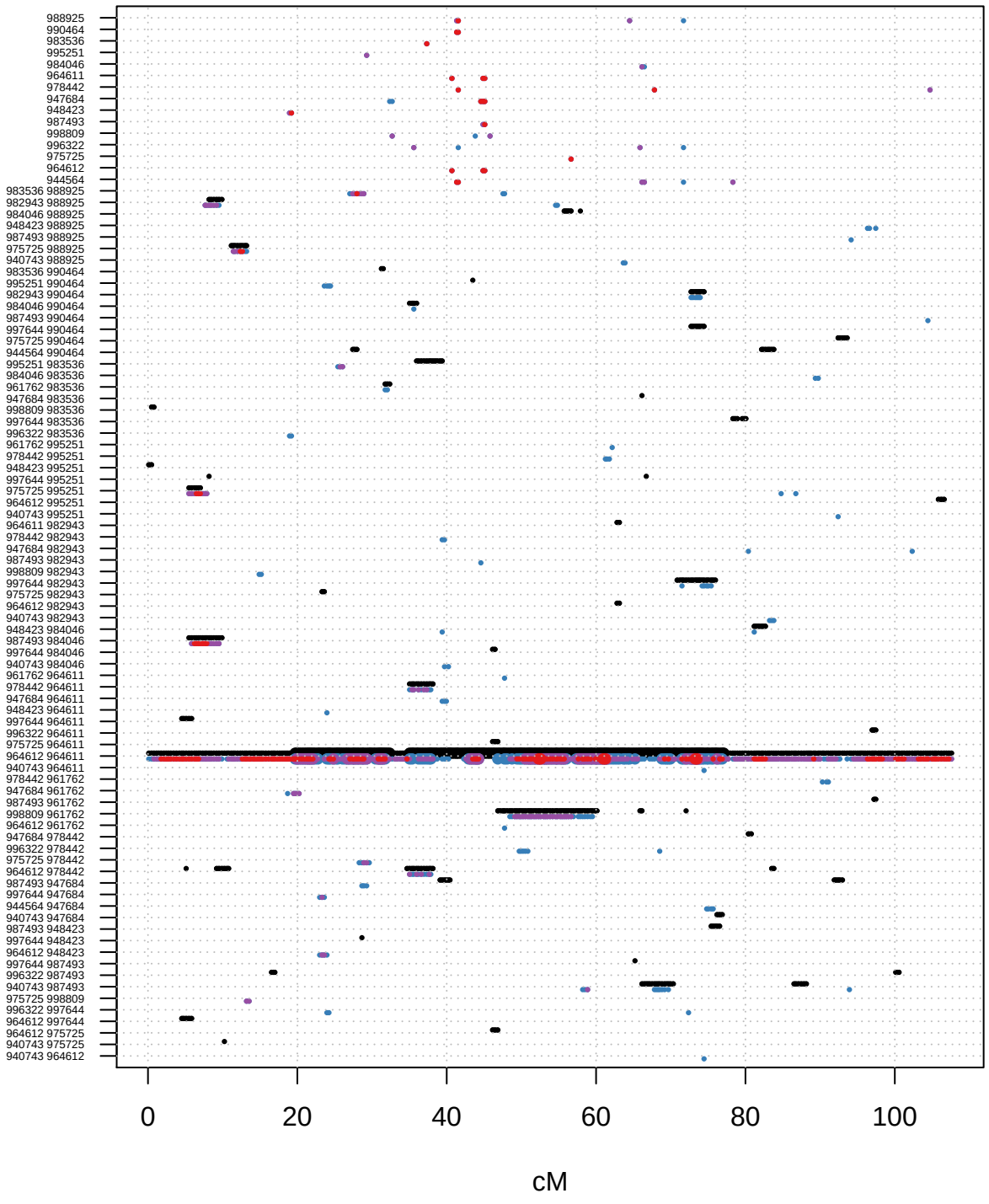



Figure 5.6: IBD segments set of 19 individuals. True IBD (black) compared to estimated IBD, colors indicate threshold of 80% (blue), 90% (purple) and 99% (red), and the pair share either one (thin line) or two (thick line) copies IBD.

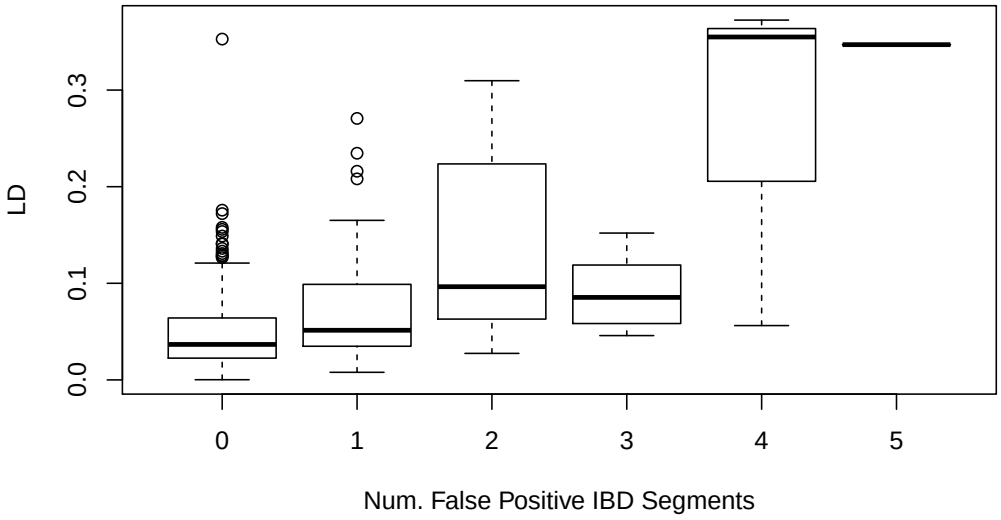



Figure 5.7: Amount of local LD where spurious, false positive IBD is estimated, in set of 57 individuals. LD is $R^2$ between allelic type of dense markers surrounding each sparse marker, number of spurious IBD segments is counted at each sparse locus.

| | | Largest Class | Single Copies | Avg Class Size | Number Classes |
|---|---|---|---|---|---|
| 2 indivs | Ped_Only | 0.10 | 0.09 | 0.04 | 0.04 |
| | Merged_True | 0.08 | 0.06 | 0.00 | 0.00 |
| | Merged_80 | 0.08 | 0.06 | 0.00 | 0.00 |
| | Merged_90 | 0.08 | 0.06 | 0.00 | 0.00 |
| | Merged_99 | 0.08 | 0.06 | 0.01 | 0.01 |
| 9 indivs | Ped_Only | 0.03 | 0.05 | 0.01 | 0.01 |
| | Merged_True | 0.04 | 0.05 | 0.00 | 0.00 |
| | Merged_80 | 0.04 | 0.05 | 0.01 | 0.01 |
| | Merged_90 | 0.04 | 0.05 | 0.01 | 0.01 |
| | Merged_99 | 0.03 | 0.05 | 0.01 | 0.01 |
| 19 indivs | Ped_Only | 0.11 | 0.12 | 0.07 | 0.07 |
| | Merged_True | 0.09 | 0.09 | 0.00 | 0.00 |
| | Merged_80 | 0.11 | 0.11 | 0.03 | 0.03 |
| | Merged_90 | 0.11 | 0.11 | 0.04 | 0.04 |
| | Merged_99 | 0.11 | 0.11 | 0.06 | 0.06 |
| 57 indivs | Ped_Only | 0.13 | 0.15 | 0.05 | 0.05 |
| | Merged_True | 0.14 | 0.07 | 0.00 | 0.00 |
| | Merged_80 | 0.11 | 0.13 | 0.05 | 0.05 |
| | Merged_90 | 0.12 | 0.14 | 0.05 | 0.05 |
| | Merged_99 | 0.13 | 0.14 | 0.05 | 0.05 |

Table 5.2: Merged IBD graphs with estimated pedigree IBD and either true or estimated population IBD at different thresholds compared to simulation truth. Statistics are RRMSE, and comparisons are made for varying numbers of individuals used for population IBD estimation.

The LOD score depends on the trait realization. Figure 5.9 has each trait realization plotted in addition to the average to show the variation among trait realizations. The trait model and IBD graphs were the same for each realization; only variation in the trait causes any differences. On average the simulated trait has an increase in LOD score at the causal locus, but this is not necessarily true for any given trait realization.

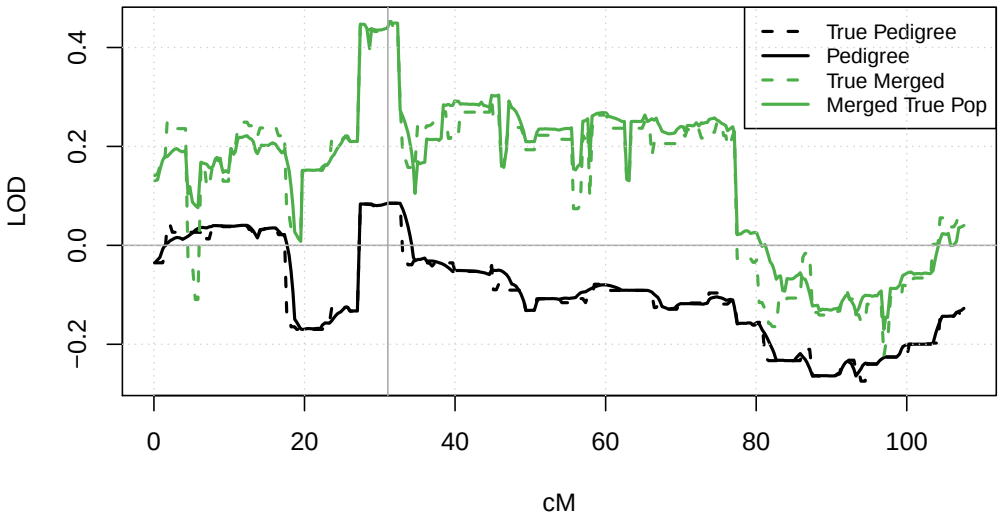



Figure 5.8: LOD scored for Merge 2 pedigrees, with the true IBD for all founders merged. LOD scores are averaged over 50 trait realizations.

The results obtained by the merging method are affected differently by false positive and false negative IBD estimates. Figure 5.10 has the average LOD score over trait realizations when using the 80% threshold `ibd_stitch` consensus estimates. The resulting LOD score is more jagged due to small IBD segments. The true population-IBD was in long segments over the chromosome, indicated by the target LOD score that is higher than the pedigree-only LOD score everywhere. False negative IBD occurs where no population-IBD is estimated but is present in the simulation truth, in this case the LOD score is not increased enough. False positive IBD occurs when population-IBD is estimated that was not present in the

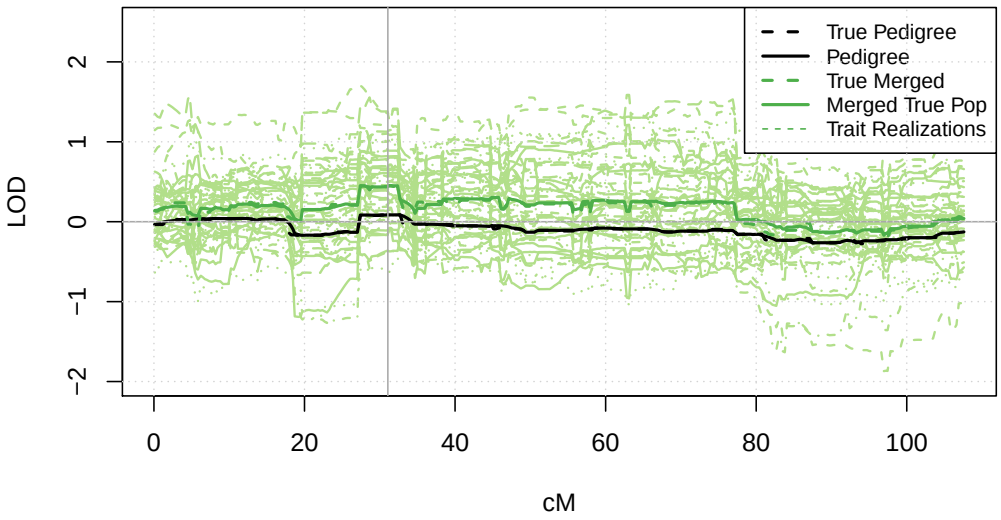



Figure 5.9: LOD scored for Merge 2 pedigrees, with all founders merged. LOD scores are averaged over 50 trait realizations, and each trait realization is also plotted.

simulation truth, in this case there is an excess of IBD compared to the truth and the LOD score is too high.

The effect of different thresholds is shown in Figure 5.11 for the 80%, 90% and 99% thresholds. If the estimates were perfect the corresponding colored area would reach precisely to the black line that indicates the change in LOD when true IBD is merged. The increasing thresholds contain less IBD so the 80% level has all the IBD of the 90% and 99%. The larger the threshold the less the change in LOD score on average meaning that the true linkage signal is less likely to be recovered. On the other hand, the higher thresholds have fewer false positive linkage signals.

Figure 5.12 shows the LOD score for an individual trait realization. In the upper panel the LOD score obtained from the true and estimated IBD state for the pedigree is compared to that when the true population IBD between all the founders is merged in. Also displayed are the 5% and 95% quantiles of the LOD score contributions from the 1000 merged graphs.

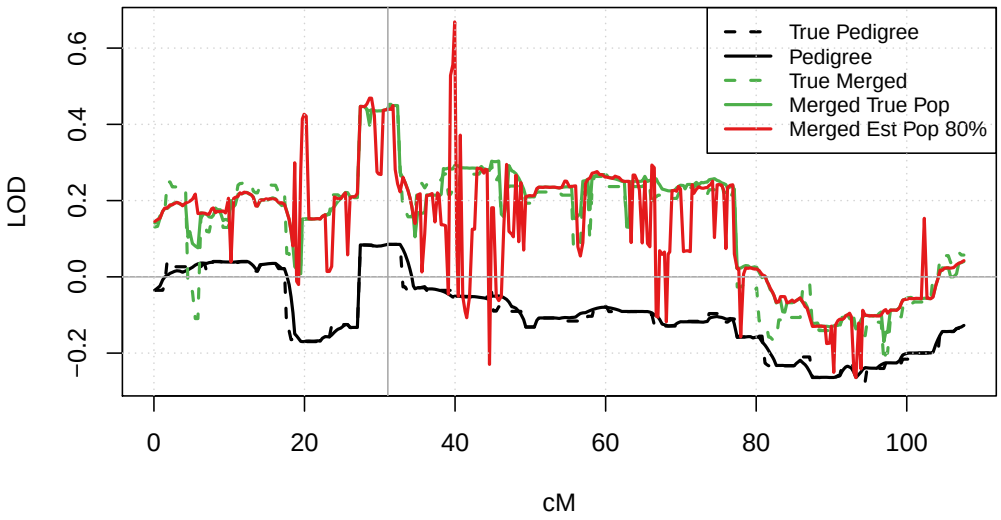



Figure 5.10: LOD scored for Merge 2 pedigrees, with the true and estimated IBD for all founders merged. LOD scores are averaged over 50 trait realizations.

The pedigree-IBD quantiles for LOD score contributions, plotted in grey, are narrower than the quantiles after merging. Although the merged IBD is the simulation truth, randomness in the merging procedure has increased the variability in the LOD score contributions. The averaged LOD score is close to the target and the quantiles cover the target LOD score for the majority of the chromosome.

The lower panel of Figure 5.12 is the corresponding figure with the estimated population IBD merged at the 80% threshold as well as the true population IBD. The LOD score from the averaged contributions is close to the target LOD score and the LOD score from the merged truth. They differ where false positive and false negative IBD segments have caused large peaks or troughs in the LOD score for merged estimated IBD, compared to the target. There is a large peak in the LOD score for merged estimated IBD at the 40cM position. The quantiles of the LOD score contributions from the 1000 merged graphs cover a large range of LOD score values including the true LOD score. The quantiles may be used to indicate

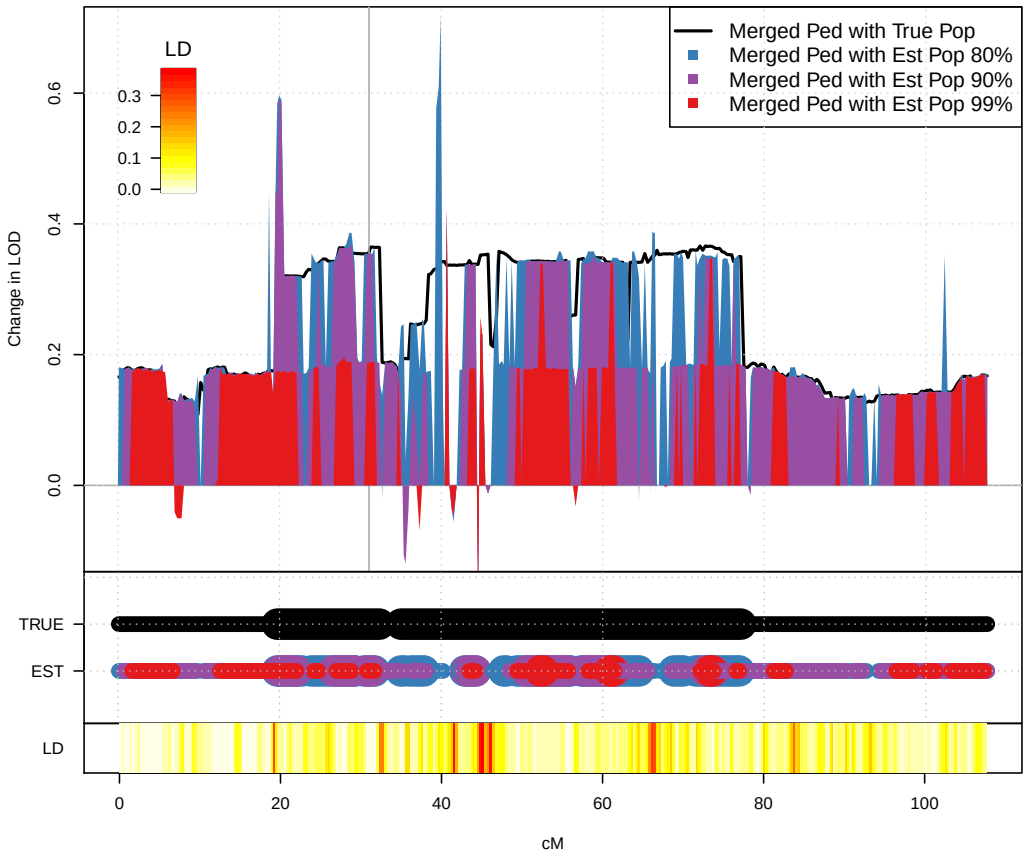



Figure 5.11: Change in LOD score for Merge 2 pedigrees when true IBD among all founders is merged versus estimated IBD at 80%, 90% and 99% thresholds. This is compared to the IBD segment that has the largest influence on the LOD change, both its truth and its estimate. Also shown is the background LD measured as the average pairwise $R^2$ among dense markers surrounding each sparse marker.

areas where the added IBD should be further investigated or used with caution. Although some false IBD was estimated by `ibd_stitch` with the same probability as the true IBD, the spread of the LOD score quantiles is much larger where the estimated IBD is false.

The quantiles of the LOD score contributions are compared in Figure 5.13. The upper panel compares the pedigree-only LOD against those with the true population-IBD merged

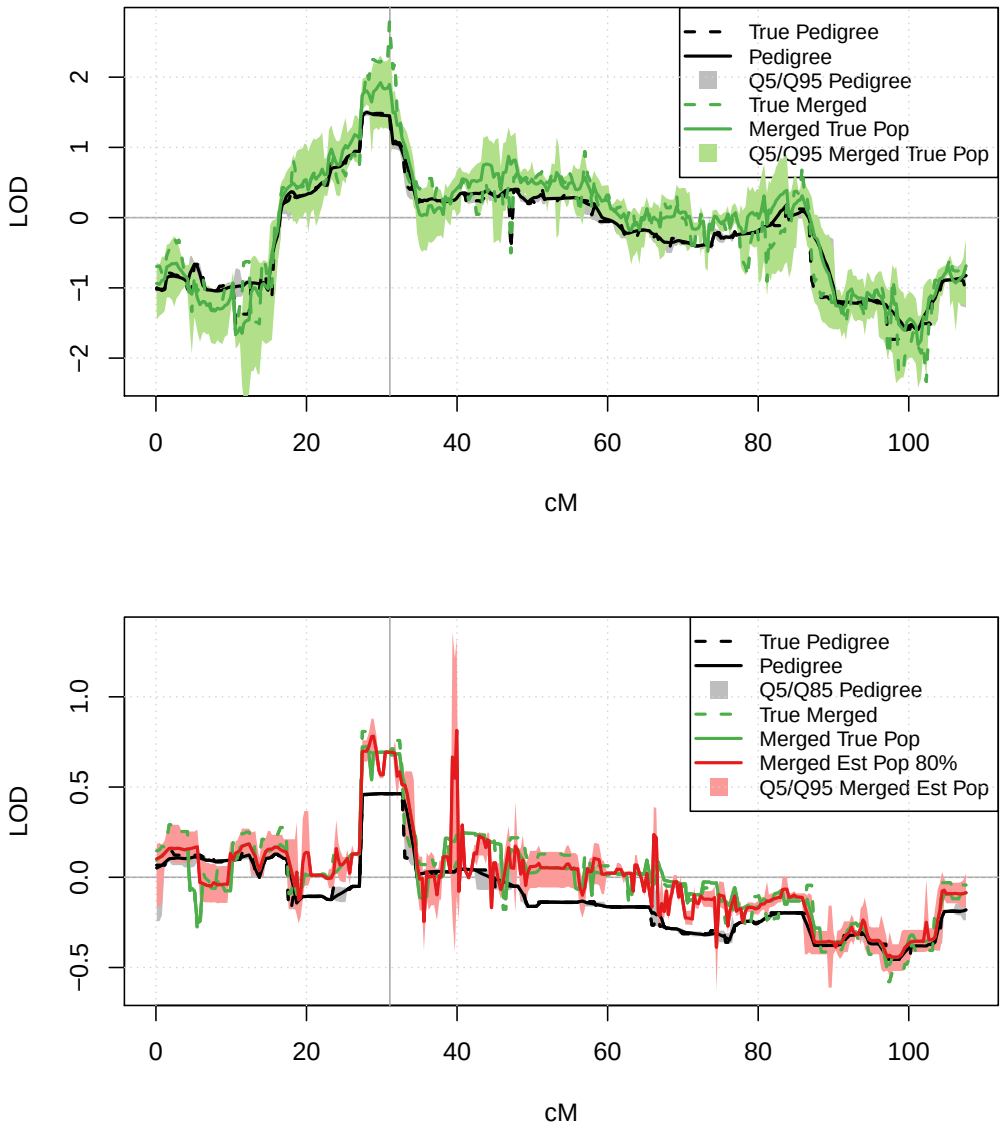



Figure 5.12: LOD scored for Merge 2 pedigrees, with the true and estimated IBD for all founders merged. LOD scores are for the 32nd trait realization, 5% and 95% quantiles are of the contributions to the LOD score from each IBD graph.

in. The lower panel shows the merged true population IBD against the merged estimated IBD. There is an increase in variability among LOD score contributions after merged compared to pedigree-only. The increase is due to the randomness in the merging procedure. The difference in variability between merging the truth versus merging estimate is small with both having larger variability at different loci. The variability in the estimate is removed when we take the consensus at a fixed threshold.

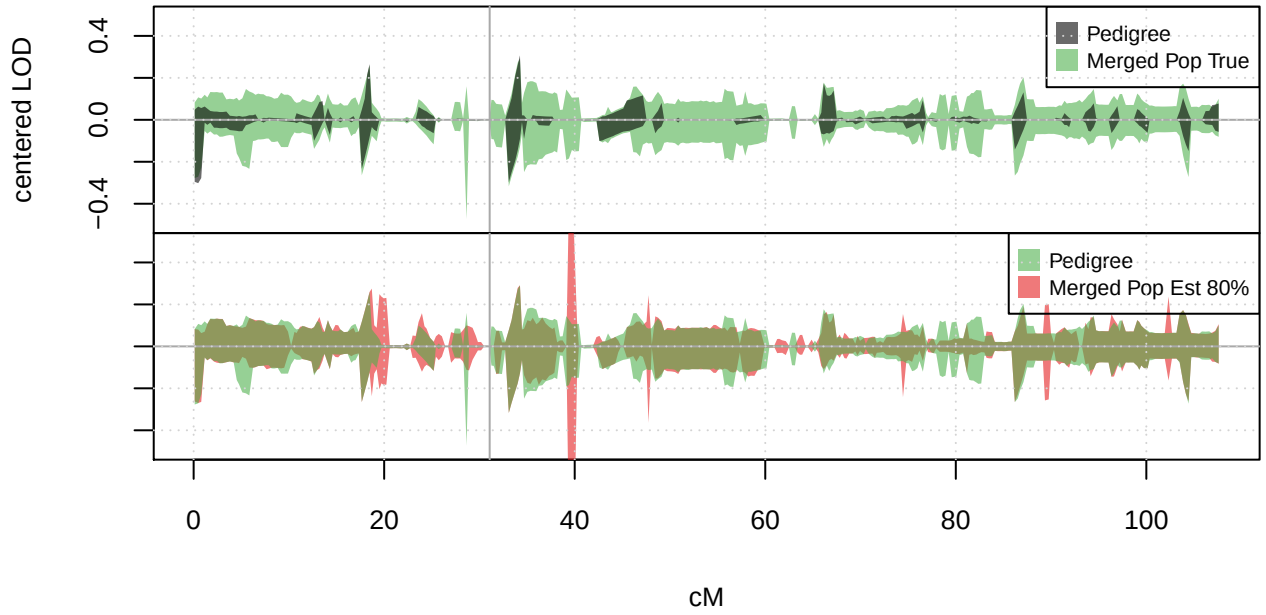



Figure 5.13: Centered Quantiles for LOD score contributions from pedigree-only, merged true, and merged estimated population IBD graphs. LOD scores are for the 32nd trait realization, 5% and 95% quantiles.

Finally, a larger data set comprising all six pedigrees from the Merge 2 and Merge 4 sets was tested. The merging was done using the true IBD graphs among the founders of the Merge2 and Merge4 respectively. The LOD scores for one trait realization is given in Figure

5.14. The LOD score overall is larger, as expected from a data set with more individuals. There is a small sharp peak of almost 3 at the trait locus and a larger segment of a LOD score just over 2 near the trait locus. The method fails to find the sharp peak even with the true population IBD provided. The majority of the available merges do not result in the exact true IBD graph that gives the full LOD score increase here. The larger segment surrounding the trait locus is identified and is just within the quantiles, although the quantiles are not centered on the truth. There is also increasing jaggedness in the merged line. There are a large number of possible random merges in such a large set and only 1000 random merges were performed. This causes variation between adjacent loci - more merges would be required for a smoother average LOD. The large variation in merges can also be seen in the spread of the quantiles.

### *5.2.5 Summary*

In the simulated pedigrees, the true IBD state between the individuals is known. Both pedigree-IBD relative to the pedigree founders and population-IBD relative to the population founders is known. The performance of both IBD estimation methods and the merging procedure in capturing the true IBD state can be assessed.

In Section 5.2.2 the performance of population-IBD estimation was examined by comparing the consensus population-IBD to the truth. The best estimates were between small numbers of individuals and at an 80% consensus threshold. In Sections 5.2.3 and 5.2.4 the performance of the merging procedure was tested. Both true and estimated population-IBD was merged into the estimated pedigree-IBD. The merging procedure itself added variability to the LOD score contributions from each IBD graph, but was unbiased. When estimated population-IBD was used the quality of the estimates had a large influence on the results. False positives caused large spikes in the LOD score that were much higher than the truth.

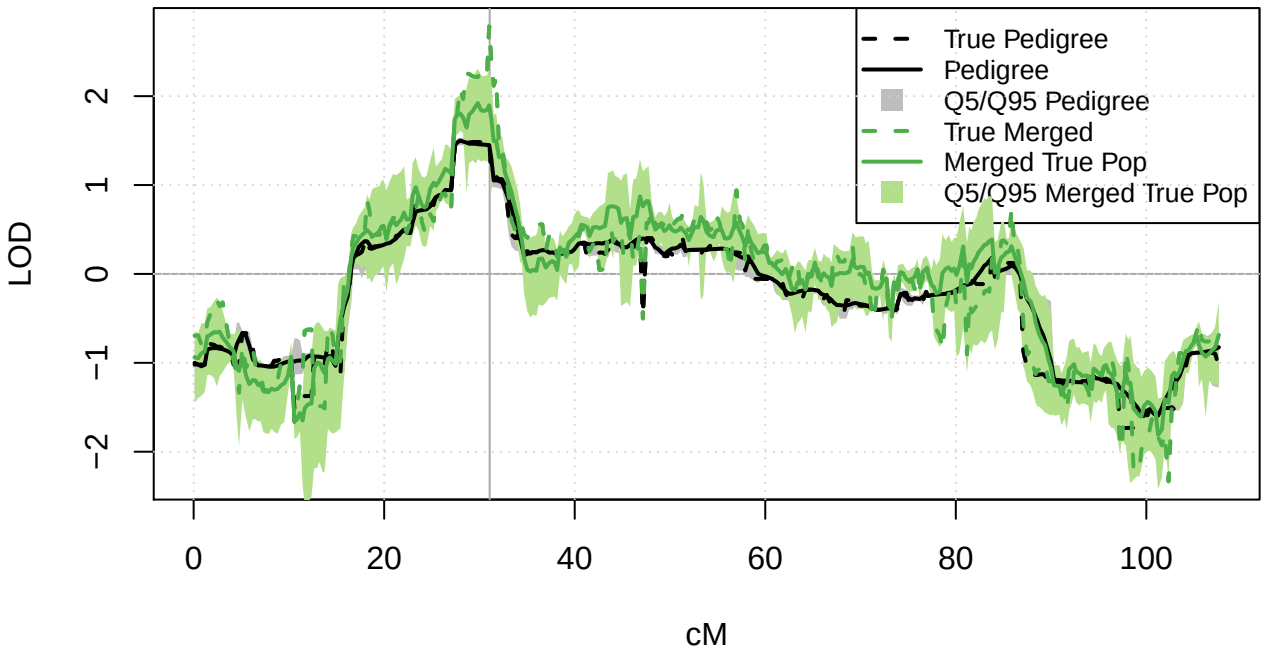



Figure 5.14: LOD scored for the combined Merge 2 and Merge 4 pedigrees, with the true IBD for all founders merged. LOD scores are for the 32nd trait realization, 5% and 95% quantiles are of the contributions to the LOD score from each IBD graph.

# Chapter 6

# MERGING ALGORITHM WITH REAL DATA

## *6.1 Pedigree-only Analysis*

The Alzheimer's dataset is described in Section 2.5. The merging analysis builds on a pedigree-only analysis on the data set, with `gl_auto` realizations and linkage analysis results obtained from E Wijsman (personal communication).

The pedigree-only linkage analysis was conducted on a sparse panel of markers over the 22 chromosomes. The panel was selected over the genome with PBAP version 1.0 [Nato et al., 2015]. The sparse panel is described in Table 6.2, and compared with a denser panel for IBD detection in Section 6.2. Realizations of IBD graphs were produced with MORGAN `gl_auto`, saving every 50th iteration for a total of 5000 realizations. On the larger ERF203 pedigree, a thinned panel of every second sparse marker was used due to computational cost; the computation time for chromosome 1 was almost 2 weeks. These realizations, $\mathbf{f}^{(r)}$ for $r = 1, ..., 5000$, on the half sparse panel were used in the merging analysis.

As described in Section 2.5, the individuals in the pedigrees were coded as either affected or unknown trait status. In the trait model the disease allele has frequency $P(D) = 0.05$ and the trait penetrance probabilities are $P(\mathbf{Y} = 1|\mathbf{t} = NN) = 0.001$ and $P(\mathbf{Y} = 1|\mathbf{t} = ND) = 0.9$ and $P(\mathbf{Y} = 1|\mathbf{t} = DD) = 0.9$. This trait model was selected to represent a dominant mode of inheritance, where the presence of at least one copy of the disease allele increases the risk of the disease almost 1000 fold. As all genotyped individuals in the two families are affected, under the model it is likely that they have at least one copy of the disease allele. Under Hardy-Weinberg equilibrium, the population prevalence is 8.8%. In the earlier linkage analysis of Liu et al. [2007] a dominant trait model was also used, with a trait allele frequency of 0.01 and an age-dependent penetrance model. Liu et al. [2007]

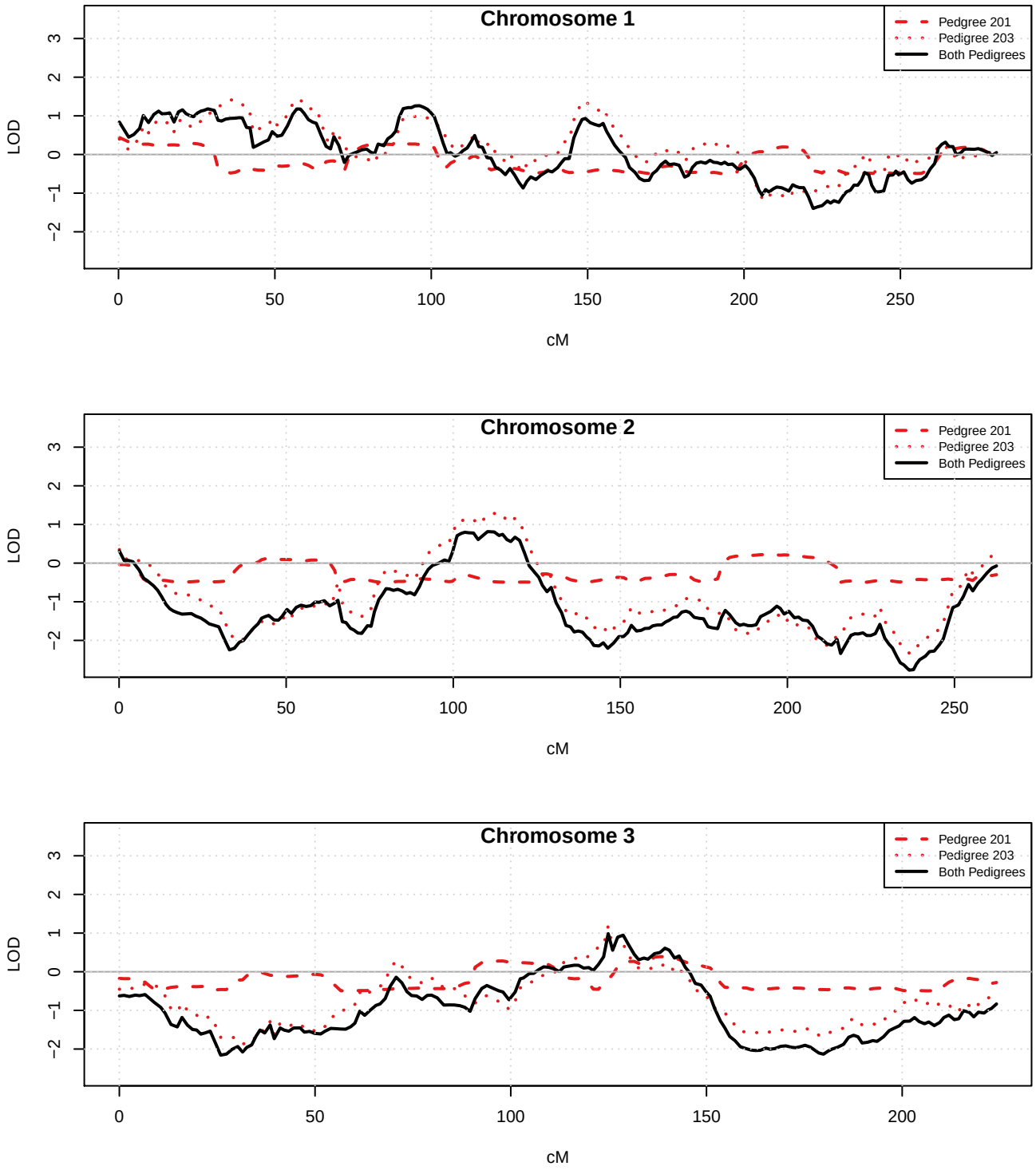
calibrated prevalence to The Rotterdam Study [Ott et al., 1995] with values ranging from 0.2% prevalence of Alzheimer's disease in the general population for ages 55-64 and 26% for ages 85+. The 8.8% prevalence of the Wijsman trait model is also in line with this study. The base log likelihoods calculated for each pedigree were -24.83718 for ERF203 and -4.766662 for ERF201.

LOD score plots produced for this thesis based on the pedigree-only analysis from the Wijsman analysis are in Figure 6.1. The red lines are the LOD scores from pedigree ERF201 and ERF203, the black line is the summation of the two. The largest LOD score signal was in ERF203 on chromosome 5 at just over 3 (see Figure 6.1b). The second highest signal was over 2 in ERF203 on chromosome 4. After these, the highest signals were around 1.5 in ERF203 on chromosome 8 (see Figure 6.1c), on chromosome 10 (see Figure 6.1d) and chromosome 16 (see Figure 6.1f). ERF201 did not show any notable LOD score signals.

## *6.2 Selection of Dense Markers*

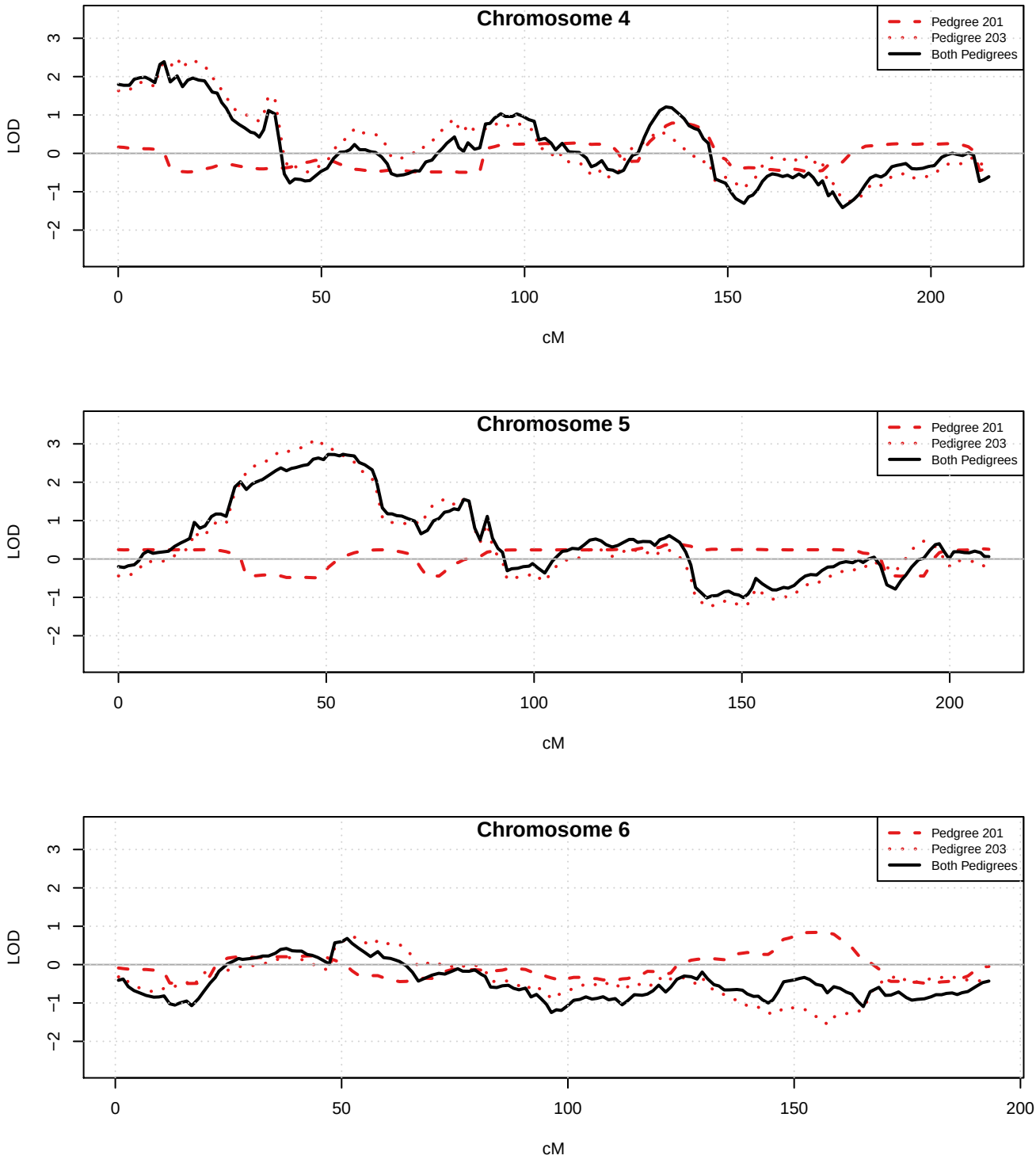
To perform the merged IBD analyses it was necessary to select a denser marker panel that includes all the markers in the sparse panel. The dense panel is used for IBD estimation. The dense panel for IBD estimation was selected using PBAP version 1.0 [Nato et al., 2015]. The subprogram `marker_subpanels` was used to select markers with the condition that the sparse panel markers be included in the dense panel. Marker data files were then prepared at the dense marker positions using the PBAP `setup_gl_auto` program.

To find appropriate PBAP marker selection parameters for use across the genome, chromosome 22 was used as a test case. Chromosome 22 has total length 51,304,566 bp [Kong et al., 2002]. In the provided data, markers were at positions between 16,050,408bp to 51,243,297bp which excludes the centromere at the beginning of the chromosome. There are 217,428 markers in this range from which a sparse panel of 117 markers for linkage analysis was selected using PBAP. The sparse marker panel had an average 0.29 Mbp or 0.69 cM intermarker distance. The cM to bp conversion was done using Rutger's map version 3, which specified an average of 0.437 Mbp/cM or 2.23 cM/Mb.



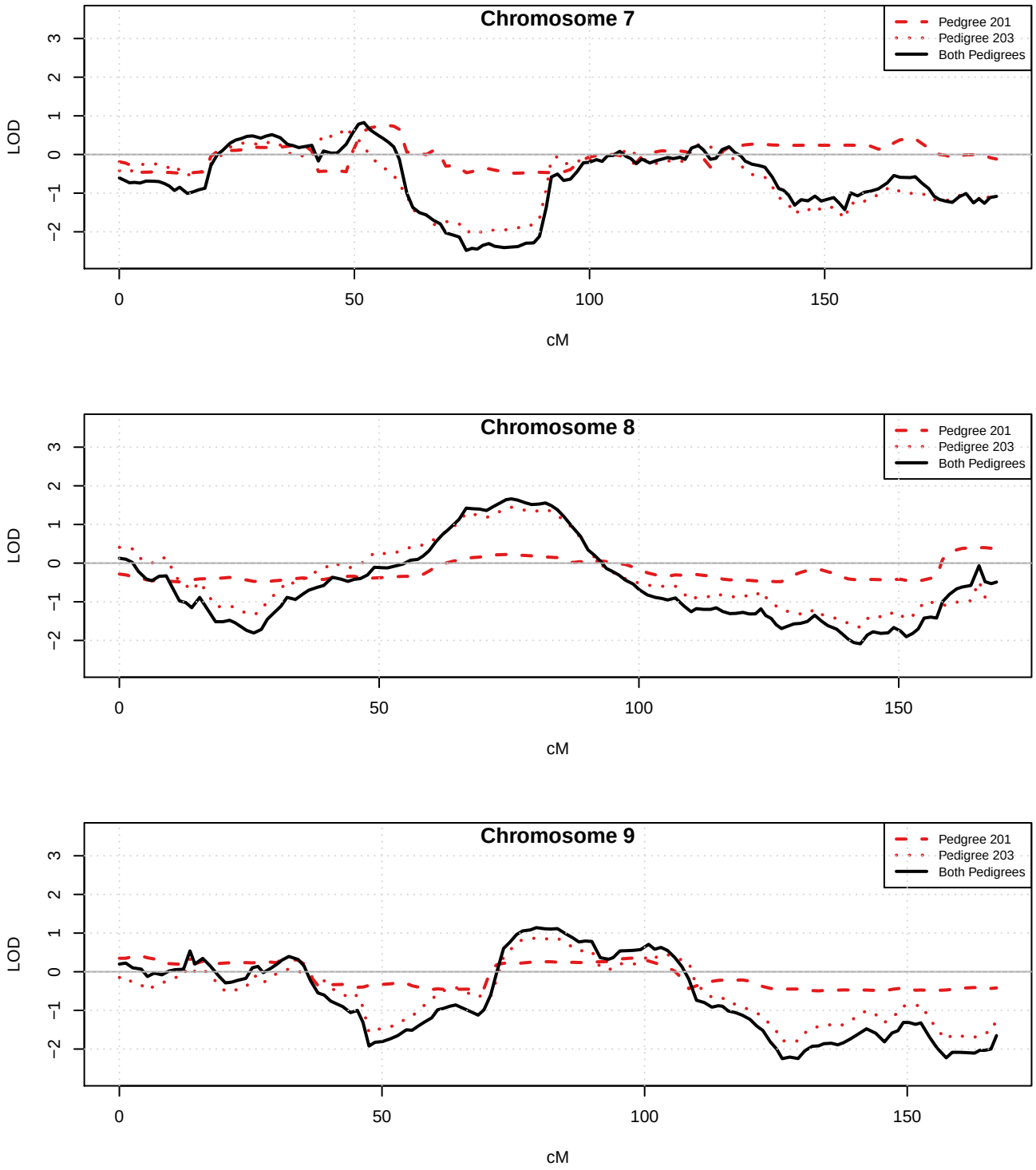

(a) Chromosomes 1,2,3

Figure 6.1: Pedigree-only LOD scores, comparing signals from ERF201 and ERF203.



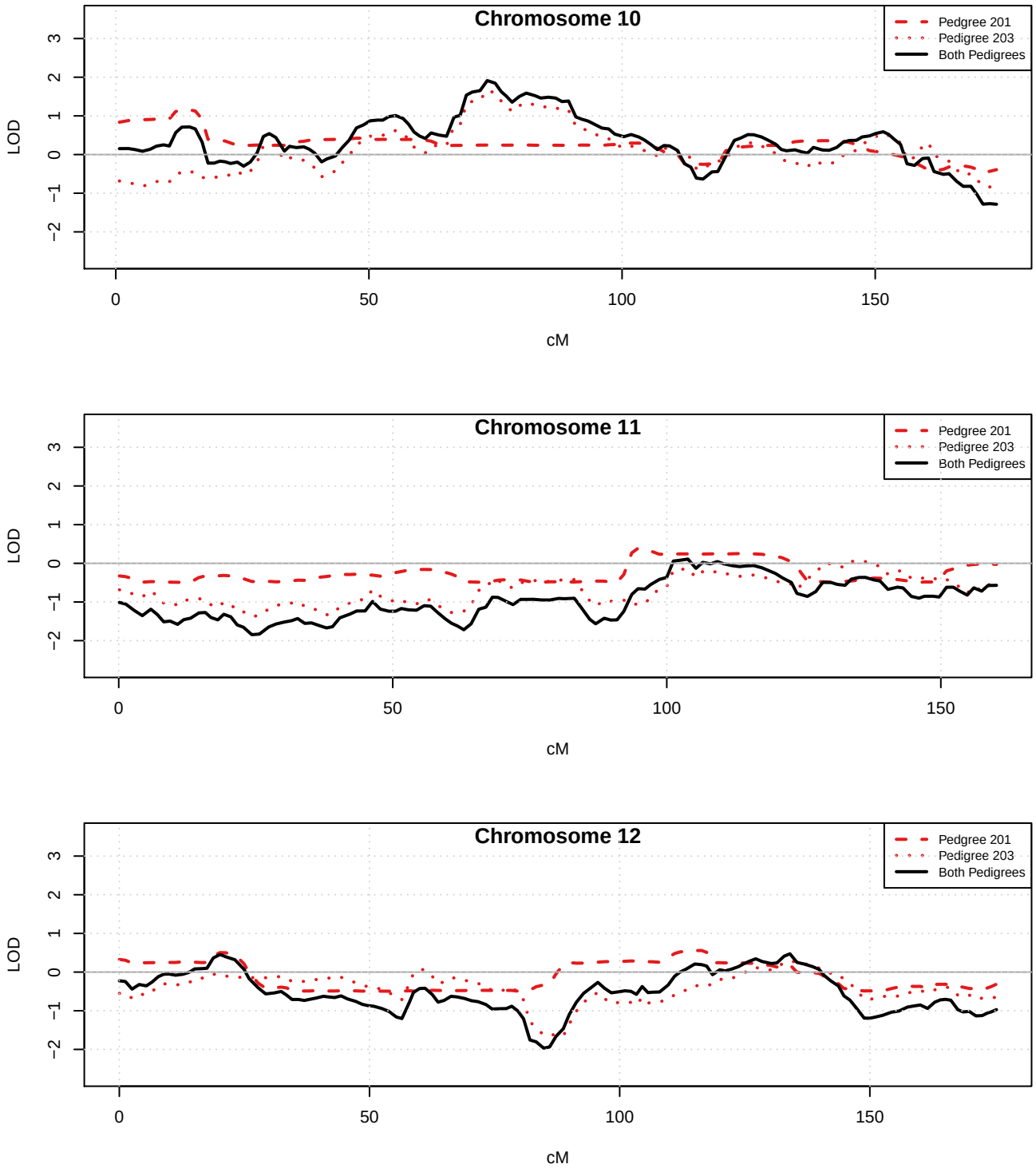

(b) Chromosomes 4,5,6

Figure 6.1: Pedigree-only LOD scores, comparing signals from ERF201 and ERF203.



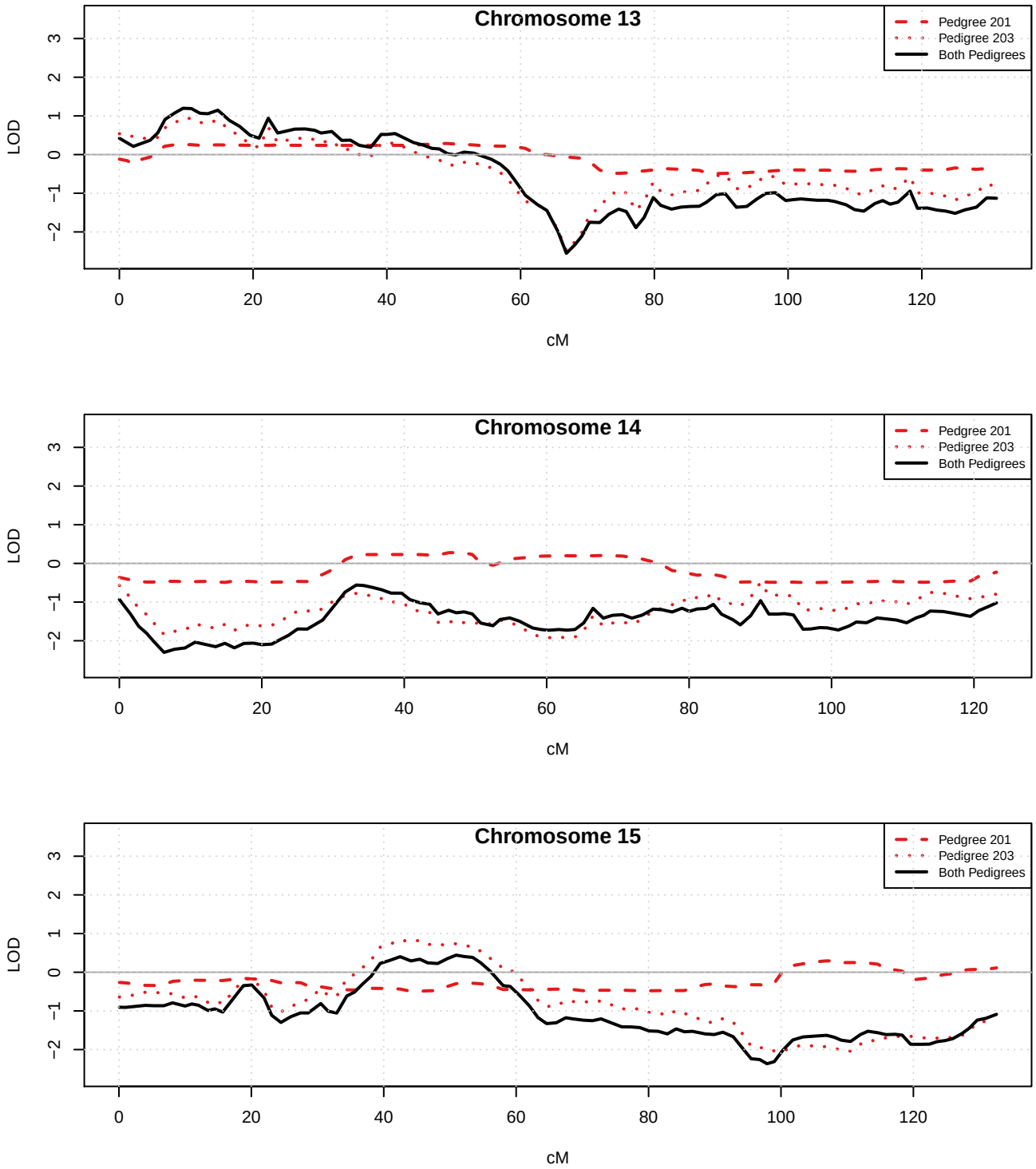

(c) Chromosomes 7,8,9

Figure 6.1: Pedigree-only LOD scores, comparing signals from ERF201 and ERF203.



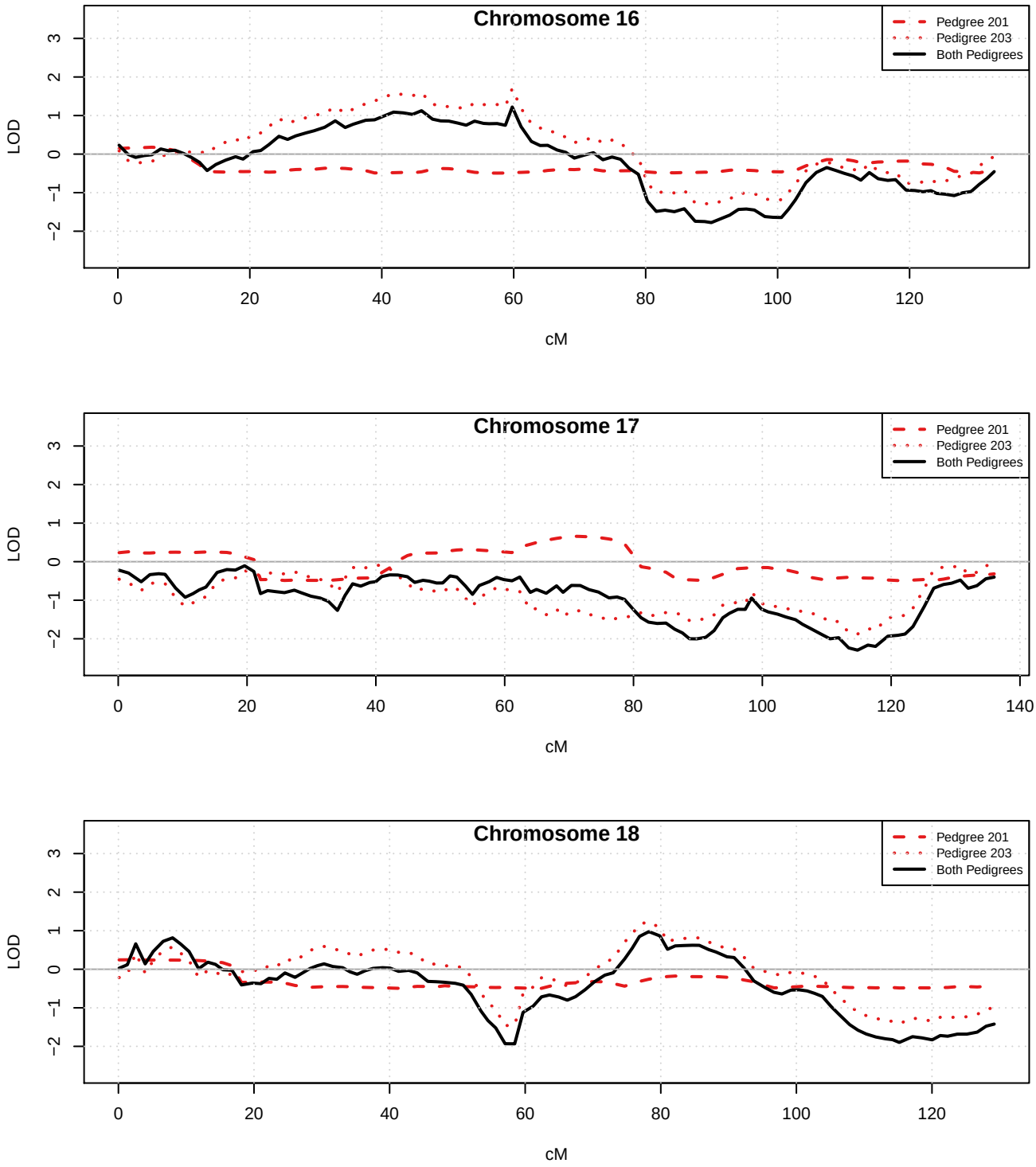

(d) Chromosomes 10,11,12

Figure 6.1: Pedigree-only LOD scores, comparing signals from ERF201 and ERF203.



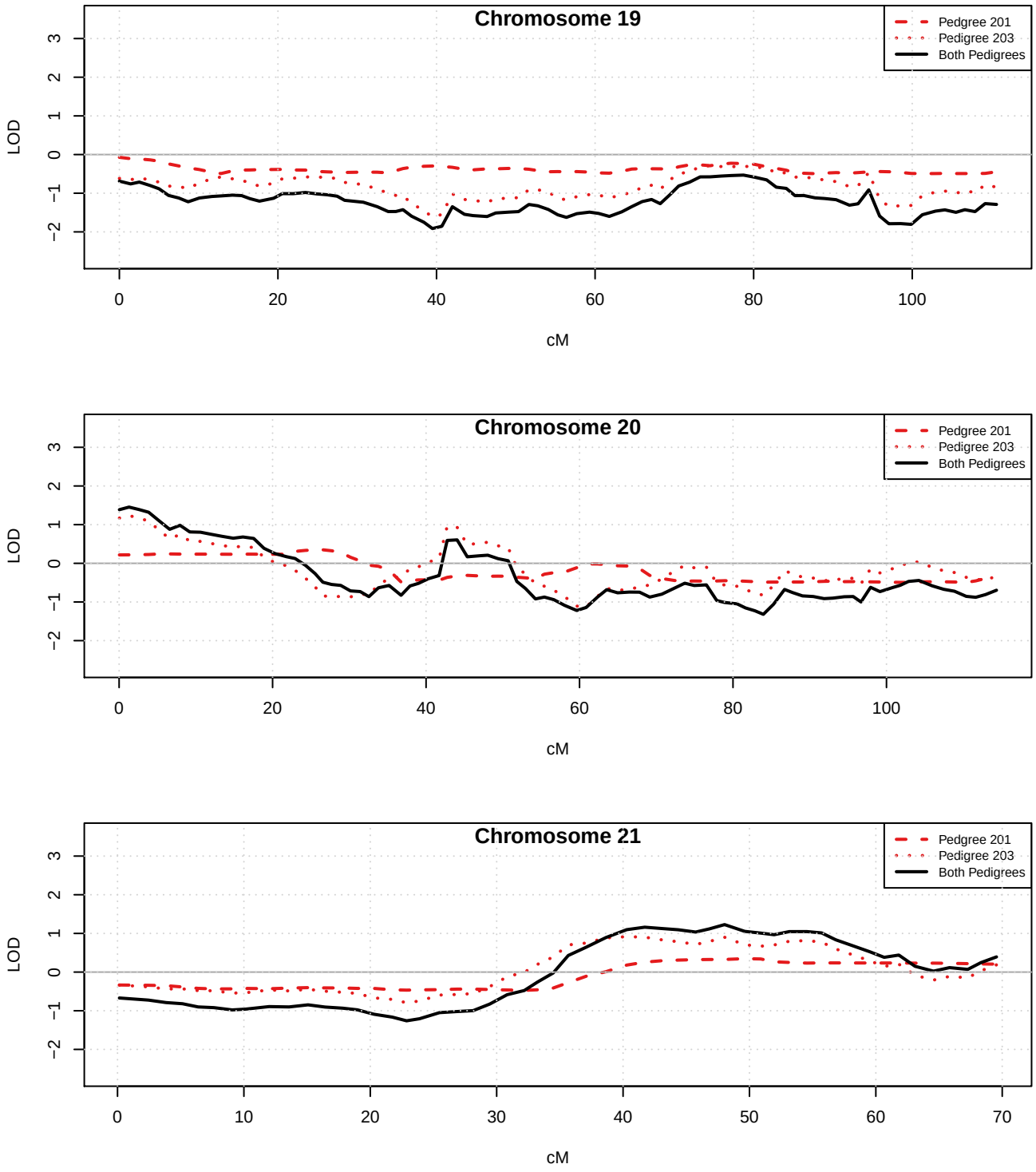

(e) Chromosomes 13,14,15

Figure 6.1: Pedigree-only LOD scores, comparing signals from ERF201 and ERF203.



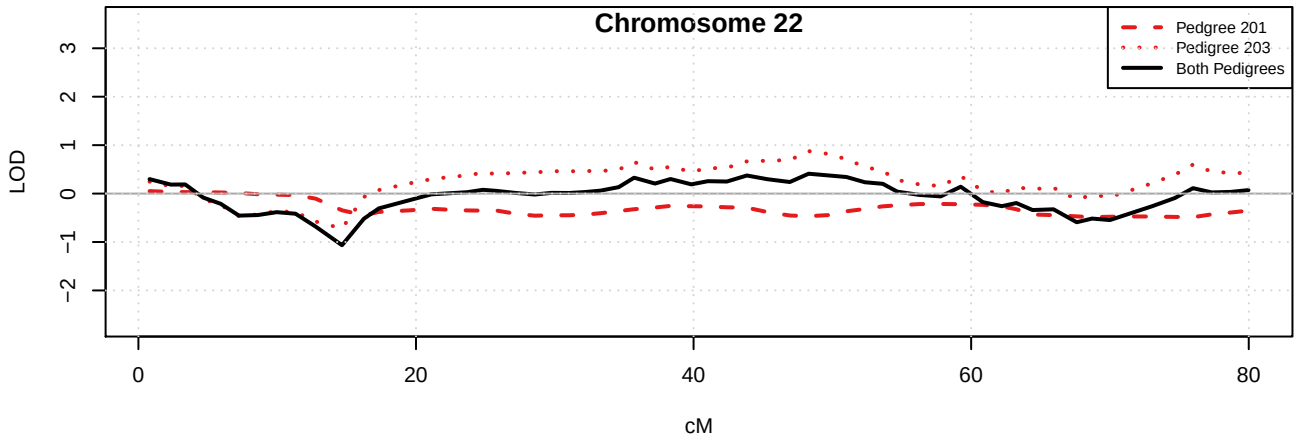

(f) Chromosomes 16,17,18

Figure 6.1: Pedigree-only LOD scores, comparing signals from ERF201 and ERF203.




(g) Chromosomes 19,20,21

Figure 6.1: Pedigree-only LOD scores, comparing signals from ERF201 and ERF203.




(h) Chromosome 22

Figure 6.1: Pedigree-only LOD scores, comparing signals from ERF201 and ERF203.

When selecting the dense set of markers PBAP requires specification of the minimum intermarker distance (min cM) and maximum LD (max LD) which is measured as the $r^2$ between adjacent markers in the panel. These were varied to find the best parameters to use - the final panel should have both low LD and be as dense as possible. The marker completion threshold was set to 80%, this is the minimum percentage of individuals with an observed genotype at the marker. In all cases the completion of the output panel was 100% and changing the marker completion threshold parameter did not affect the panel. Markers that are monomorphic in the population were excluded and the minimum minor allele frequency was set to 0.05 in the 1000 genomes European reference file. The methods were run in the forward direction along the chromosome, starting at the first marker on both the main panel and the pre-subpanel.

PBAP uses LD-based SNP pruning via PLINK [Purcell et al., 2007b] to create a prepanel which is then gap-filled - additional markers are placed in the gaps between markers in the prepanel. The max LD parameter value is used in gapfilling: at the two adjacent markers on each side of a potential gapfiller the pairwise $r^2$ is calculated. The marker is selected if the $r^2$ is lower than the specified max LD and its MAF is less than the threshold of 0.05 less a decrement value, set to 0.02. The MAF can be lower for gapfillers than the prepanel. Results of testing different combinations of min cM and max LD parameter values on chromosome 21 are shown in Table 6.1. The densest panel was found using a min cM of 0.03 and max LD of 0.3 resulting in 1277 markers spaced on average every 0.06cM. The set includes the sparse 117 markers which were spaced with an average inter-marker distance of 0.68cM with approximately 7 dense markers for every sparse marker.

Based on these tests, a minimum distance of 0.03 cM and a maximum pairwise $r^2$ of 0.3 between adjacent markers were used as panel selection parameters on all chromosomes. Summaries of the panels for all chromosomes are given in Table 6.2. The sparse panel of markers on each chromosome had already been selected for the previous linkage analysis. The half sparse panel is every other sparse marker, used for merging and LOD score locations. The dense panel is used for population IBD detection.

| Test Set | Min cM | Max LD | Num_mark | Mean cM | Mean Mb |
|---|---|---|---|---|---|
| Sparse | | | 117 | 0.683 | 0.292 |
| 1 | 0.04 | 0.05 | 780 | 0.103 | 0.044 |
| 2 | 0.04 | 0.025 | 626 | 0.128 | 0.055 |
| 3 | 0.02 | 0.05 | | | |
| 4 | 0.03 | 0.05 | 856 | 0.093 | 0.040 |
| 5 | 0.05 | 0.1 | 802 | 0.0998 | 0.043 |
| 6 | 0.03 | 0.2 | 1177 | 0.0680 | 0.029 |
| 7 | 0.02 | 0.3 | | | |
| 8 | 0.03 | 0.3 | 1277 | 0.0626 | 0.027 |
| 9 | 0.03 | 0.5 | | | |

Table 6.1: Markers Selected with varying PBAP parameters for Min cM and Max LD. Blank indicates that no such subset was able to be found.

### *6.3 Selection of Chromosomes*

To demonstrate the effect and potential advantage gained by using the merging procedure, a few chromosomes were selected to investigate closely. Chromosome 5 has the highest pedigree-only LOD score. It was also of interest to look at chromosomes where genotyped individuals share population-IBD around the linkage peak. A genome scan for IBD, described in this section, identified Chromosomes 21 and 8 as good candidates for merging.

The genome scan over all 22 chromosomes was performed using `ibd_stitch` to detect areas with a high level of IBD between pedigrees. A total of 22 individuals have marker data but the performance and speed of `ibd_stitch` is better for smaller groups. The 22 individuals were divided into four sets of 7 or 8 individuals. `ibd_stitch` was run on each of set for each chromosome to detect IBD among individuals within the set over that chromsome. The four sets are listed in Table 6.3 as set 1-4. Each set included the two individuals from

| | | Dense | | Sparse | | Half Sparse | |
|---|---|---|---|---|---|---|---|
| Chrom. | Total cM | Num Markers | Markers per cM | Num Markers | Markers per cM | Num Markers | Markers per cM |
| 1 | 284 | 5138 | 18.07 | 420 | 1.49 | 210 | 0.75 |
| 2 | 263 | 4976 | 18.88 | 398 | 1.51 | 199 | 0.75 |
| 3 | 224 | 4364 | 19.47 | 349 | 1.55 | 175 | 0.78 |
| 4 | 214 | 4060 | 18.95 | 327 | 1.52 | 164 | 0.76 |
| 5 | 209 | 3872 | 18.48 | 325 | 1.55 | 163 | 0.77 |
| 6 | 194 | 3863 | 19.90 | 300 | 1.55 | 150 | 0.77 |
| 7 | 187 | 3535 | 18.88 | 282 | 1.50 | 141 | 0.75 |
| 8 | 169 | 3270 | 19.32 | 261 | 1.54 | 131 | 0.77 |
| 9 | 167 | 3149 | 18.82 | 257 | 1.53 | 129 | 0.77 |
| 10 | 174 | 3292 | 18.91 | 263 | 1.51 | 132 | 0.76 |
| 11 | 161 | 3047 | 18.91 | 244 | 1.51 | 122 | 0.76 |
| 12 | 176 | 3278 | 18.62 | 271 | 1.54 | 136 | 0.77 |
| 13 | 132 | 2467 | 18.72 | 195 | 1.48 | 98 | 0.74 |
| 14 | 125 | 2207 | 17.63 | 192 | 1.53 | 96 | 0.77 |
| 15 | 133 | 2151 | 16.22 | 193 | 1.45 | 97 | 0.72 |
| 16 | 134 | 2334 | 17.43 | 200 | 1.49 | 100 | 0.75 |
| 17 | 137 | 2275 | 16.59 | 210 | 1.53 | 105 | 0.77 |
| 18 | 130 | 2278 | 17.56 | 200 | 1.54 | 100 | 0.77 |
| 19 | 111 | 1794 | 16.15 | 169 | 1.52 | 85 | 0.76 |
| 20 | 115 | 1988 | 17.33 | 175 | 1.52 | 88 | 0.76 |
| 21 | 69 | 1239 | 17.82 | 105 | 1.50 | 53 | 0.75 |
| 22 | 80 | 1281 | 16.02 | 117 | 1.47 | 59 | 0.73 |

Table 6.2: Summaries of panels selected for analysis of Alzheimer's Data

ERF201 and 5 or 6 individuals from ERF203. The sets were chosen by first separating the closest relatives, sibling and avuncular pairs, into different groups and randomly assigning the remaining individuals. Finally, the pedigree structure of ERF203 and kinship matrix show that it is possible to select 9 individuals all of whom have zero kinship with each other. These zero-kinship individuals make up set 5 in Table 6.3.

1. 201_44, 201_46, 203_175, 203_166, 203_188, 203_174, 203_177, 203_178
2. 201_44, 201_46, 203_176, 203_181, 203_173, 203_183, 203_169, 203_186
3. 201_44, 201_46, 203_180, 203_171, 203_189, 203_170, 203_184
4. 201_44, 201_46, 203_172, 203_182, 203_168, 203_167, 203_5.
5. 201_44, 201_46, 203_178, 203_186, 203_188, 203_168, 203_174, 203_189, 203_183, 203_171, 203_170

Table 6.3: Sets of individuals used in `ibd_stitch` runs. Sets 1-4 used in genome scan; Set 5 is made up of zero-kinship individuals.

For the genome scan for IBD, `ibd_stitch` was used on sets 1-4 for each of the 22 chromosomes, with 1000 iterations in each case. For all chromosomes, the dense markers selected in Section 6.2 were used. A kinship of $\beta = 0.05$ and rate of change $\alpha = 0.05$ were selected, the same values used on the simulated and GAW data sets in Chapter 5. The transition matrix null fraction was 0.05, and the genotyping error rate was 0.01. A 1% error rate is a lot higher than expected for this cleaned SNP data (data was described in Section 2.5) but having a high error rate allows for additional flexibility in the `ibd_stitch` model, increasing the probability of moving between IBD states that have a small transition probability.

Summary statistics from the genome scan were collected with R scripts. At each locus the

distribution over iterations (median, 10% and 90% percentiles) of the number of chromosomes in the largest IBD class, the total number of IBD classes, the number of singletons, the number of classes in intersection, and the size of the intersection between the pedigrees. Selected plots are shown in Figure 6.2. At each locus, a singleton is a chromosome that is not IBD with any other chromosome, and the intersection if non-empty consists of all the chromosomes with an FGL that appears in both pedigrees.

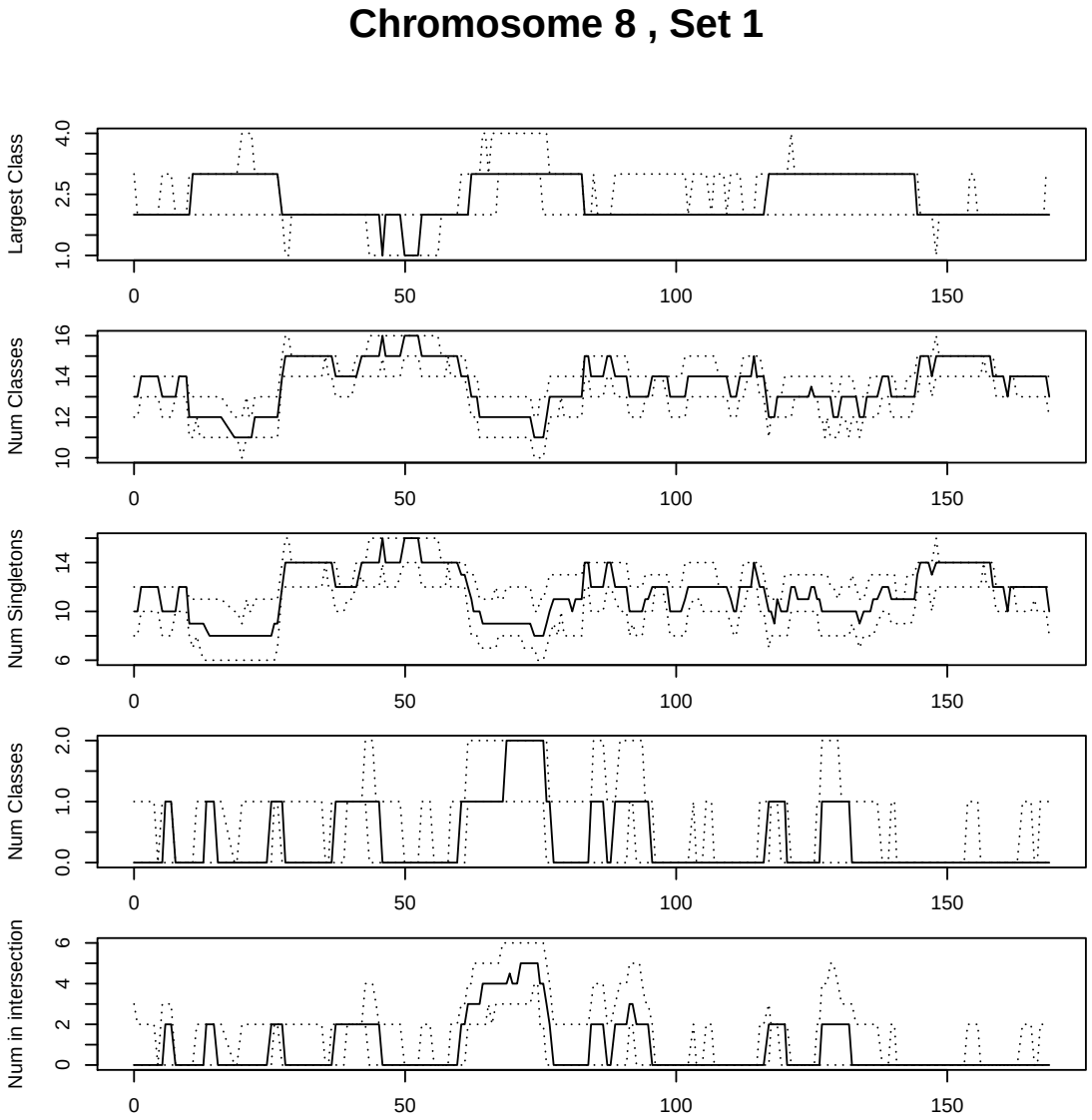
In Subfigures 6.2a, 6.2b, and 6.2c sets 1, 2 and 3 have a long intersection between the pedigrees in the region of the LOD score peak. There is some signal in the LOD score for both families so there is an opportunity to increase the LOD score here. Subfigure 6.2d describes chromosome 21; there is a long intersection around the region of the LOD score peak, but only in set 3. No significant IBD sharing was detected in Chromosome 5 - the chromosome with the highest linkage peak.

The APP gene, known to be implicated in Alzheimer's disease [Papassotiropopulos et al., 2008], is found on on chromosome 21 between bp 25,880,550 to 26,171,128 [National Library of Medicine (US), 2016a] which is centered around the 25 cM position. In the pedigree-only analysis there is no signal at this location on either pedigree.

Another gene that may contribute to Alzheimer's disease is HLA, the human version of the MHC complex which consists of more than 200 genes located close together on chromosome 6 [National Library of Medicine (US), 2016b]. The region spans from 29,677,984 to 33,485,635 and centered around the 54cM position. This could correspond to the peak in ERF203 that was not picked up in ERF201. The peak can be seen in Figure 6.1b which shows a LOD score of around 0.5 for pedigree 203 (dotted red line) at this position and a LOD score of around -0.5 for ERF 201 (dashed red line).

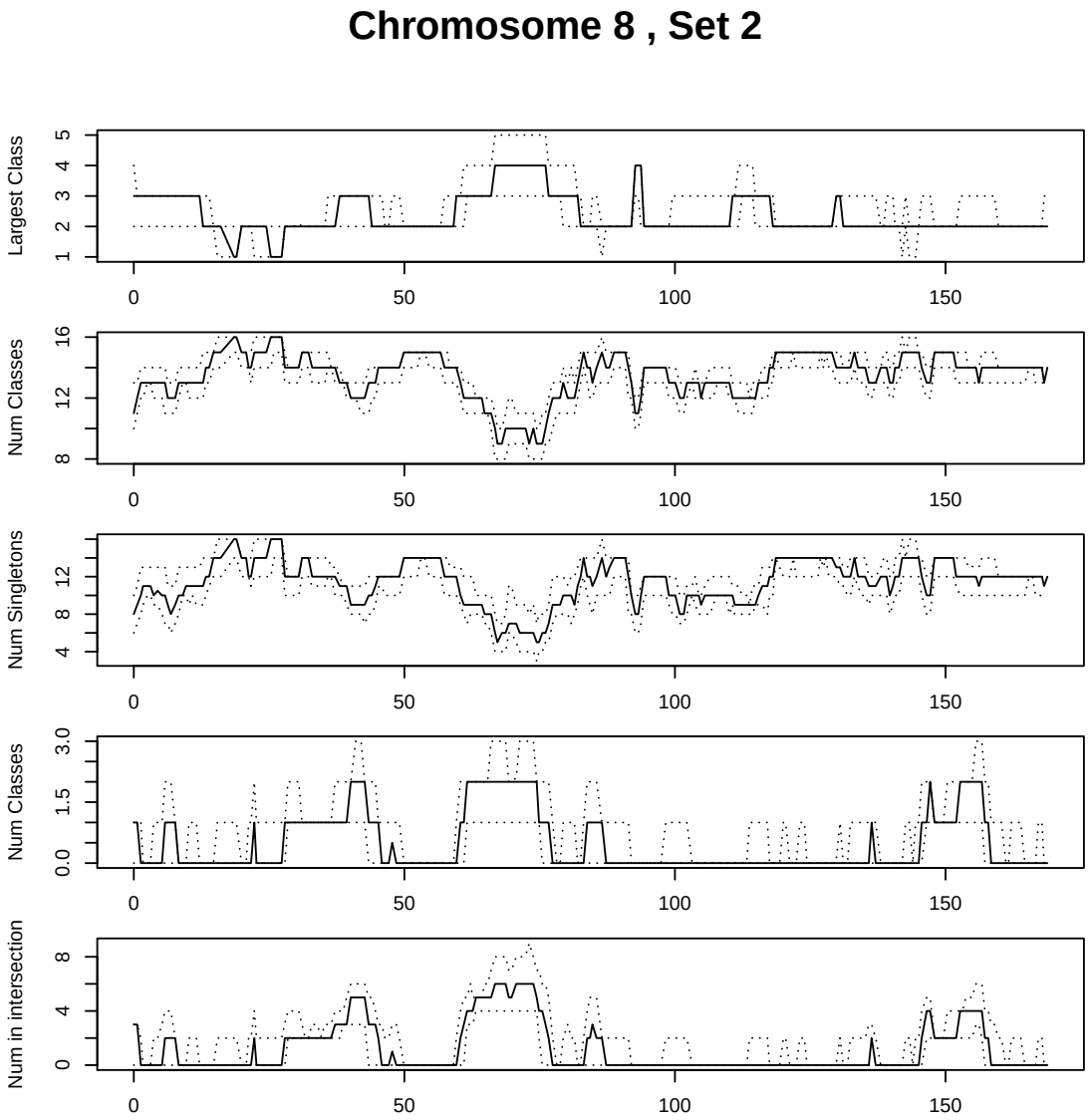
## *6.4 Merging IBD*

Merging was performed on Chromosomes 5, 8 and 21, the chromosomes selected in the genome scan described in Section 6.3. The genome scan was performed on sets 1-4 of individuals listed in Table 6.3. A merge set $\mathcal{M}$, set 5 in Table 6.3, was selected from all individuals



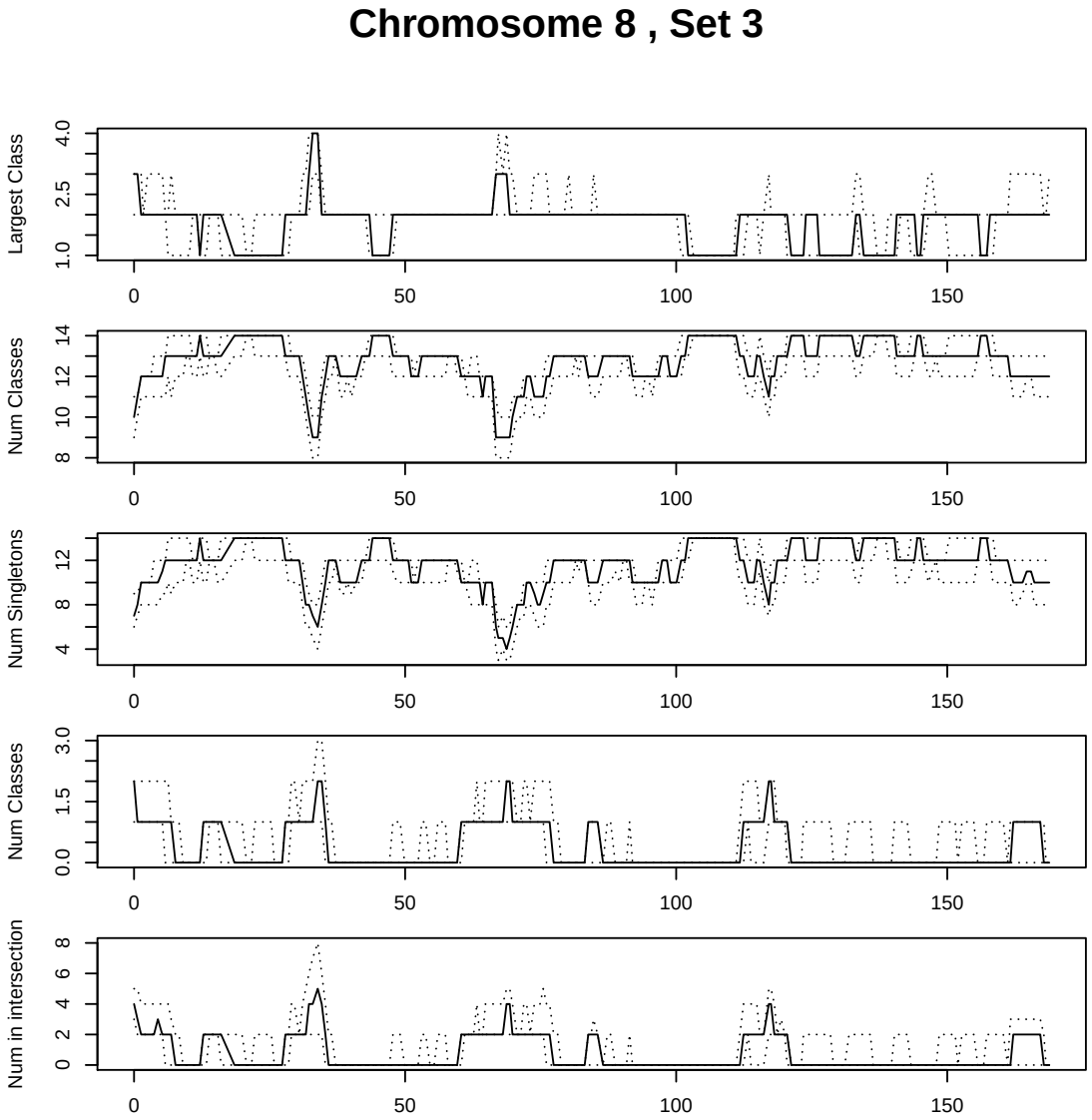

(a) Chromosome 8, LOD peak is around 50-100cM

Figure 6.2: IBD Classes in `ibd_stitch` genome scan



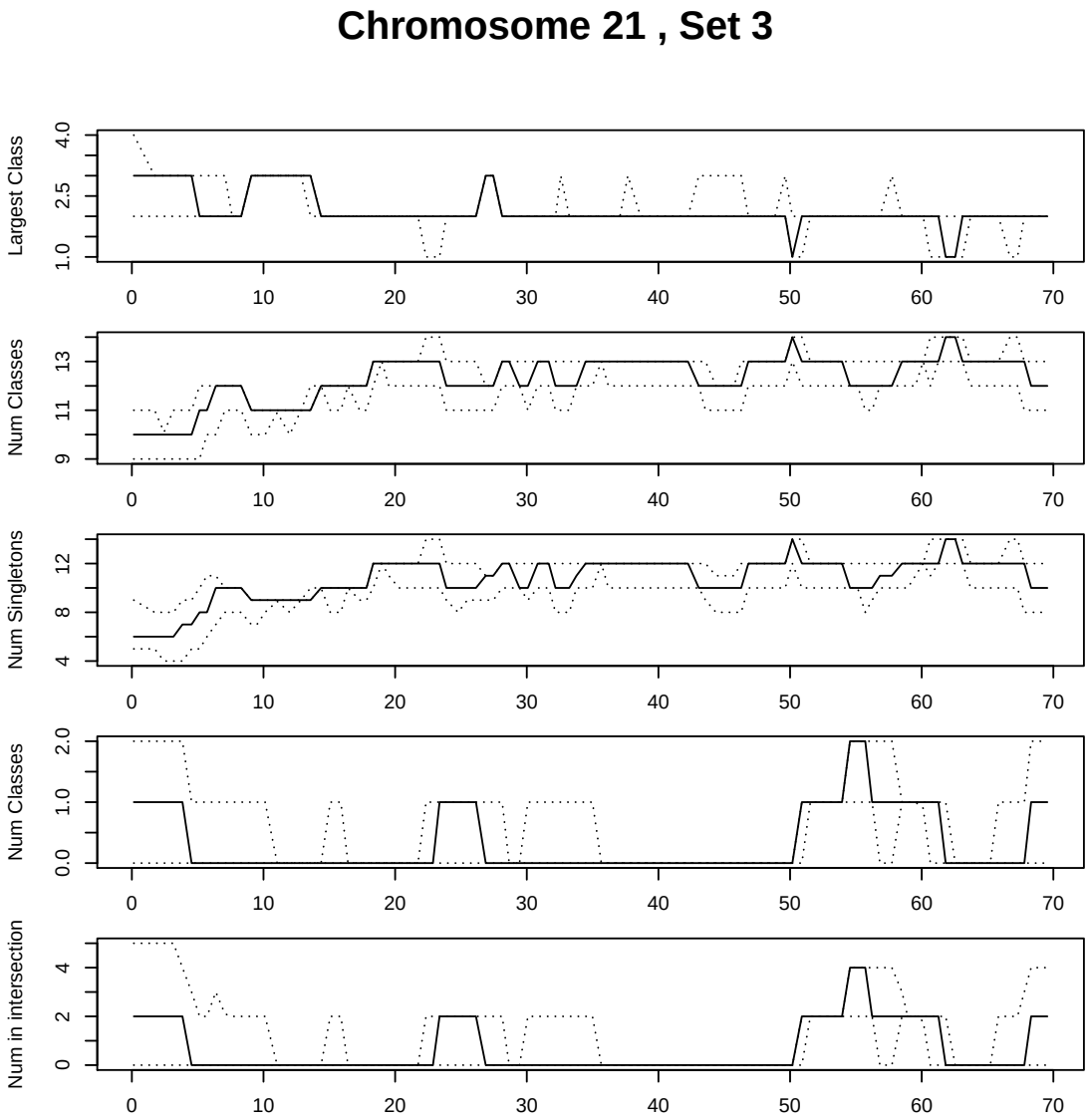

(b) Chromosome 8, LOD peak is around 50-100cM

Figure 6.2: IBD Classes in `ibd_stitch` genome scan




(c) Chromosome 8, LOD peak is around 50-100cM

Figure 6.2: IBD Classes in `ibd_stitch` genome scan




(d) Chromosome 21, LOD peak is around 35-55cM

Figure 6.2: IBD Classes in `ibd_stitch` genome scan

to use in the merging algorithm. Set 5 was used so that none of the individuals were related and to be small enough for good quality `ibd_stitch` joint IBD state estimates.

Realizations of population FGLs on $\mathcal{M}$ at all dense markers, denoted $\mathbf{m}^{(r)}, r = 1, ..., 1000$ were generated with `ibd_stitch`. Parameters were $\beta = 0.05$, $\alpha = 0.05$, null fraction 0.05 and error 0.01, the same as for the genome scan in Section 6.3. Consensus population-IBD was made after 5 runs of the consensus method at threshold levels from 80% to 99%. To calculate LOD scores, the same trait model was used as in previous analyses, described in Section 6.1. Base log likelihoods for the two pedigrees were added to give the base log likelihood for the combined graph.

### *6.4.1 Chromosome 21 Results*

This section describes the results of the merging analysis on Chromosome 21. Population-IBD detected among individuals in set 5 has been merged with pedigree-IBD on the chromosome at several $\theta$ values. The LOD scores before and after merging are compared.

The change in LOD score after merging at different population-IBD thresholds is shown in Figure 6.3. There is an increase in LOD score from around 1 to 2 that corresponds to IBD detected between individuals 201_44 and 203_180. The increase occurs in one area of the previous pedigree-only linkage peak that is located between the 40-60cM positions. The pedigree-IBD has increased the LOD score and narrowed the area of interest. The length of the detected segment depends on the threshold used, with a more stringent threshold resulting in a smaller detected segment - an approximately 10cM long segment at 80% and only around 2cM long at 99%. The length of the IBD segment indicates that it may be genuine IBD, despite the $< 99\%$ probabilty over the majority of the segment. In the simulated data analysis in Section 5.2 the 80% threshold gave the consensus closest to the true IBD. The 80% threshold seems appropriate for the estimated IBD segments on Chromosome 21; almost all 80% segments coincide with segments at a higher threshold. Another longer IBD segment located from 30-50cM has no influence on the LOD score, as this increase in IBD is not associated with the phenotype. This segment was estimated with a higher probabilty.

The smaller IBD segments located from 10-30 cM do affect the LOD score, but do not provide enough evidence to raise the LOD score much above zero.

The change in LOD score contributions is examined in Figure 6.4. The change in LOD score is plotted for just the 99% threshold, with corresponding changes in LOD score contributions from individual realizations. Only the 99% threshold is shown as there are fewer IBD segments and the corresponding changes in LOD score contributions are clearer. The detected segments do cause changes in the LOD score contributions. Segments that do not affect the overall value of the LOD score do cause changes in the LOD score contributions from IBD realizations, for example the segment positioned from 30-50cM between individuals 203_189 and 203_184.

### *6.4.2 Chromosome 8 Results*

This section describes the results of the merging analysis on Chromosome 8. Population-IBD detected among individuals in set 5 has been merged with pedigree-IBD on the chromosome, and the LOD scores before and after merging are compared. Like chromosome 21, chromosome 8 was selected from the IBD genome scan for having population-IBD around the linkage peak in Section 6.3.

Figure 6.5 shows both the LOD scores before merging and after merging at different thresholds, and the estimated population-IBD segments. The figure shows a particularly long segment of IBD, estimated at over the 99% threshold, between individuals 203_171 and 203_174 positioned from 70-110cM. This segment, combined with smaller overlapping ones was the reason for identification in the genome scan. Merging this segment results in an increase of the LOD score from around 1.5 to 3 when merged at the 99% level, and to 4 when merged at the 80% and 90% levels. The several short population IBD segments in this region cause a more jagged LOD score curve compared to the pedigree-only LOD. Another point of interest is the overlapping IBD segments positioned around 40cM. These result in a new linkage peak with a LOD score of around 2 that was not present in the pedigree-only analysis. At the 99% level, there are two roughly equally sized linkage peaks, the new

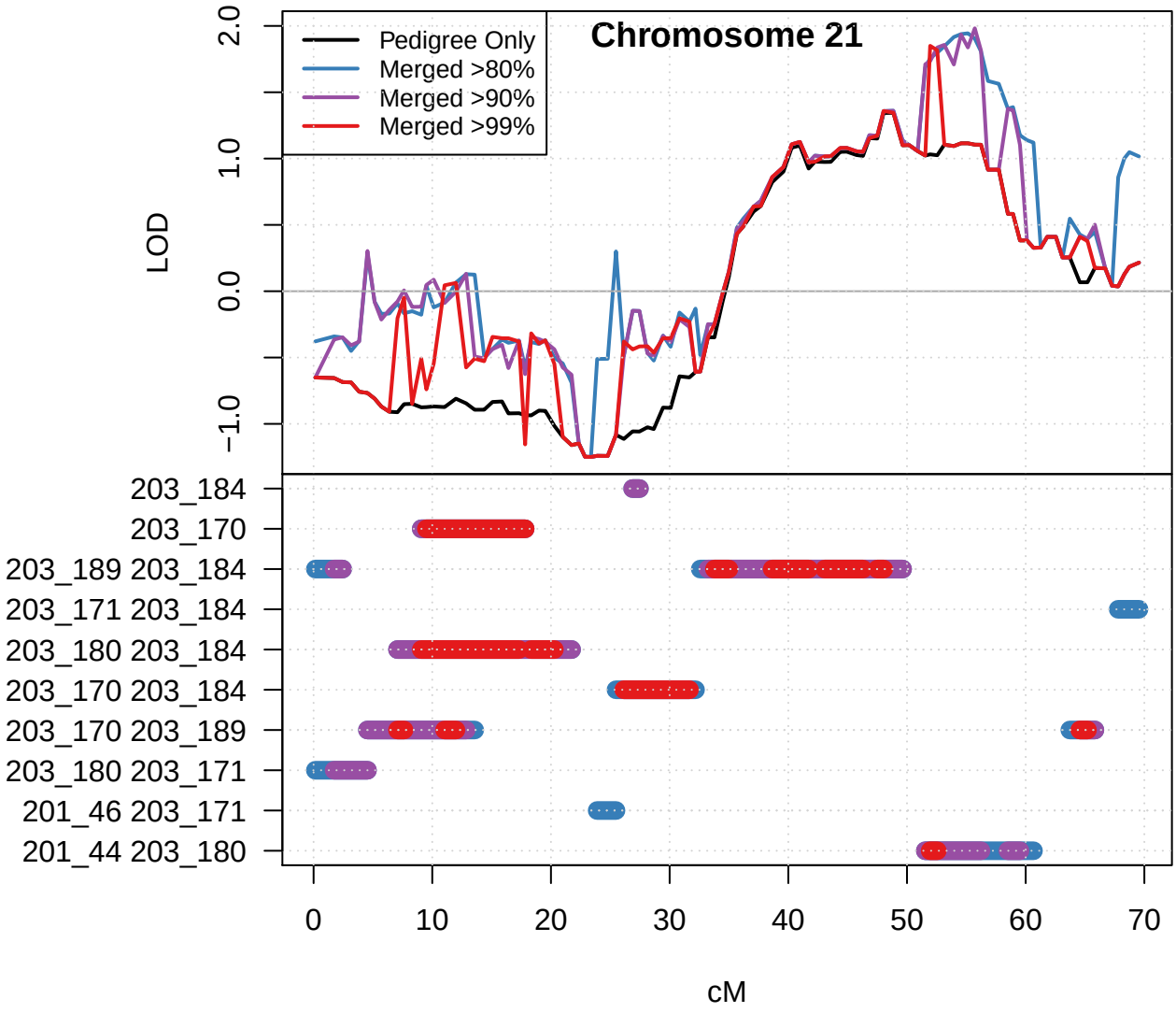



Figure 6.3: Chromosome 21 LOD scores before and after merging population-IBD at different thresholds, with detected IBD segments

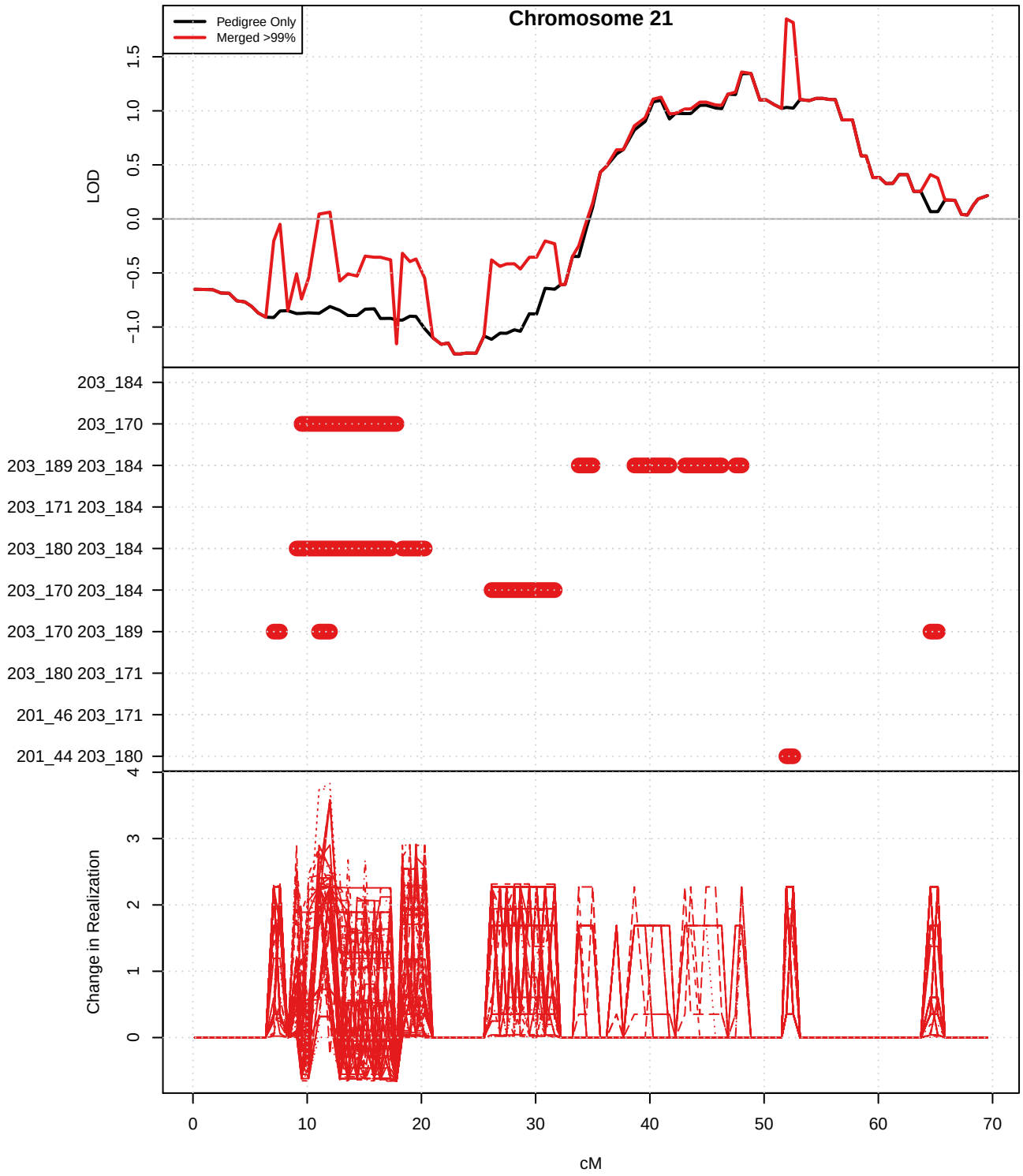



Figure 6.4: Chromosome 21 LOD scores before and after merging population-IBD at 99% threshold, with detected IBD segments and changes in LOD score contribution per realization

one positioned around 40cM and the increased one positioned around 70cM. The change in individual LOD score contributions for the 99% merge are given in Figure 6.6.

Many of the longer population-IBD segments were estimated with probabilities greater than 99% over almost the entire length of the segment. The shorter population-IBD segments, less than 5cM in length, were typically only estimated with probabilities between 80% and 90%. We can be more confident that the long segments are true population-IBD between the individuals, both due to the high probability and the length of the segment. False positive IBD can be caused by LD, the effect is typically short-range and only causes a short false positive IBD segment. The shorter segments may be false positives. A high threshold of 99% seems appropriate for this chromosome to be conservative with false positive IBD segments, although a lower threshold may be closer to the true IBD state as in the analysis of the simulated data set.

While 99% is appropriate for chromsome 8, in the analysis on chromosome 21 a lower threshold of 80% gave the best results. It is valid to select a different threshold for different chromsomes as the analyses were performed independently on each chromosome, and the important IBD in the analysis of each chromosome was between different pairs of individuals.

### *6.4.3 Chromosome 5 Results*

Chromosome 5 was selected as it has the highest pedigree-only LOD score, not through the genome scan. Population-IBD was estimated that when merged resulted in an increase at the pedigree-only linkage peak positioned around 50cM, shown in Figure 6.7. At the 99% threshold the LOD score increased from 2.5 to 4, at the 80% and 90% threshold the LOD increased to 6. There is another LOD score peak of a similar post-merging magnitude of around 4 positioned at 25cM. The pre-merging linkage peak located from 25-65cM has been isolated into the two smaller peaks, one positioned around 25cM and the other around 50cM. Each peak represents populaiton-IBD sharing between different pairs of individuals, and thus possibly different genetic variants. The LOD score contribution changes are given in Figure 6.8.

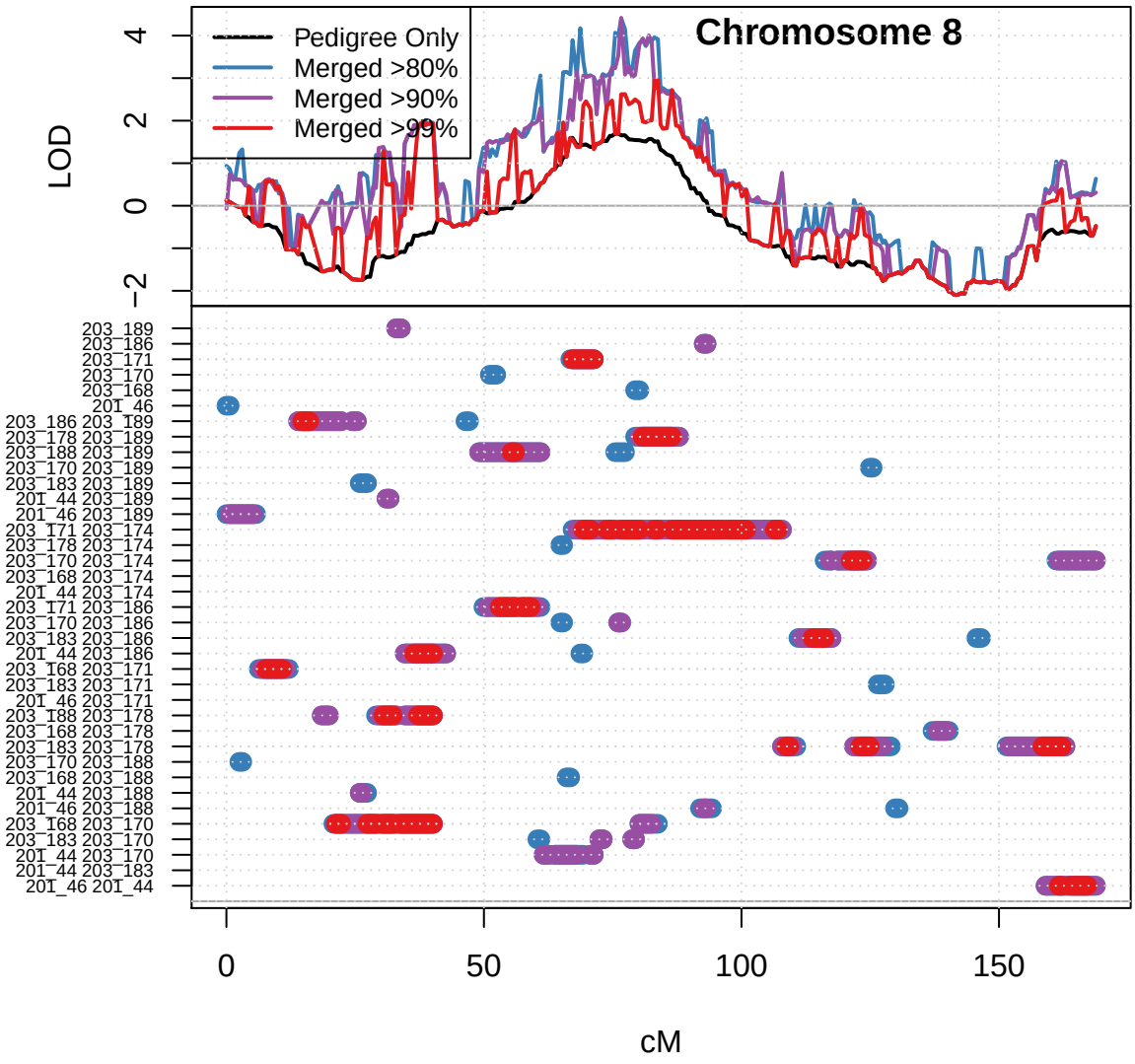



Figure 6.5: Chromosome 8 LOD scores before and after merging population-IBD at different thresholds, with detected IBD segments

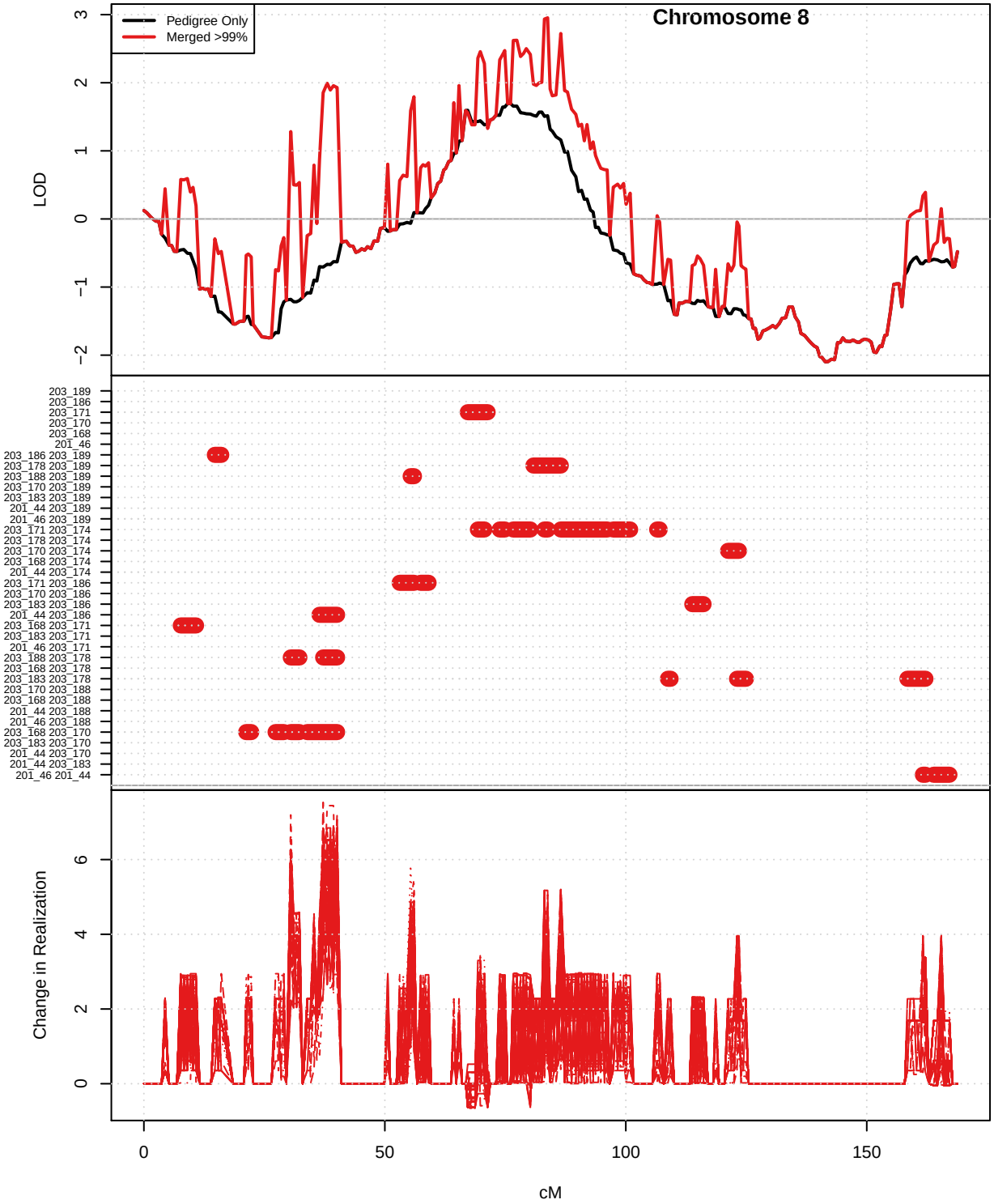



Figure 6.6: Chromosome 8 LOD scores before and after merging population-IBD at 99% threshold, with detected IBD segments and changes in LOD score contribution per realization

The majority of the 80% segments are not present at 99%. Some are likely real, such as the broken segment between 203_178 and 203_186 from 30-60cM that is present at the 90% and 80% threshold levels. Another interesting segment is between 203_178 and 203_189 from 10-40cM that is estimated at the 99% level everywhere but at a small gap in the middle of the segment. The IBD is likely to be in one segment without a break. In other segments, such as from 0-20cM between 203_183 and 203_171, IBD is present at the 99% level in the middle of the segment but tapers off towards each endpoint. In these segments the start and end points of the IBD segment are unclear but there is strong evidence for IBD.

## *6.5 Summary*

In this chapter the merging method was applied the Alzheimer's disease data introduced in Section 2.5. In this real data set the true IBD state is unknown. A pedigree-only linkage analysis had already been performed on the data, described in Section 6.1, and had identified a linkage signal on Chromosome 5. To extend the analysis using the merging method a dense set of markers was selected for population-IBD detection, described in Section 6.2. Next, in Section 6.3 a genome scan for population-IBD was performed to identify a small number of chromosomes to use to evaluate the merging method. Chromsome 21 and 8 were chosen, as population-IBD was detected among individuals near the pedigree-only linkage peaks on these chromsomes. The results of the merging analysis on chromsomes 21, 8 and 5 were described in Sections , and respectively. Population-IBD was detected between individuals in disjoint pedigree components, between individuals who were in the same pedigree component but not related by the pedigree structure, and between the two DNA copies of single individuals. The population-IBD was merged at different threshold levels.

On chromsome 21 the merging of population-IBD resulted in an increase in the LOD score in the area of the pedigree-only linkage peak. The LOD increased from 1 to 2 and the area of interest was narrowed. Population-IBD segments estimated at other positions on the chromosome either did not have an effect on the LOD score, or increase the LOD score above zero. On chromosome 8 the incorporation of population-IBD increased the LOD score

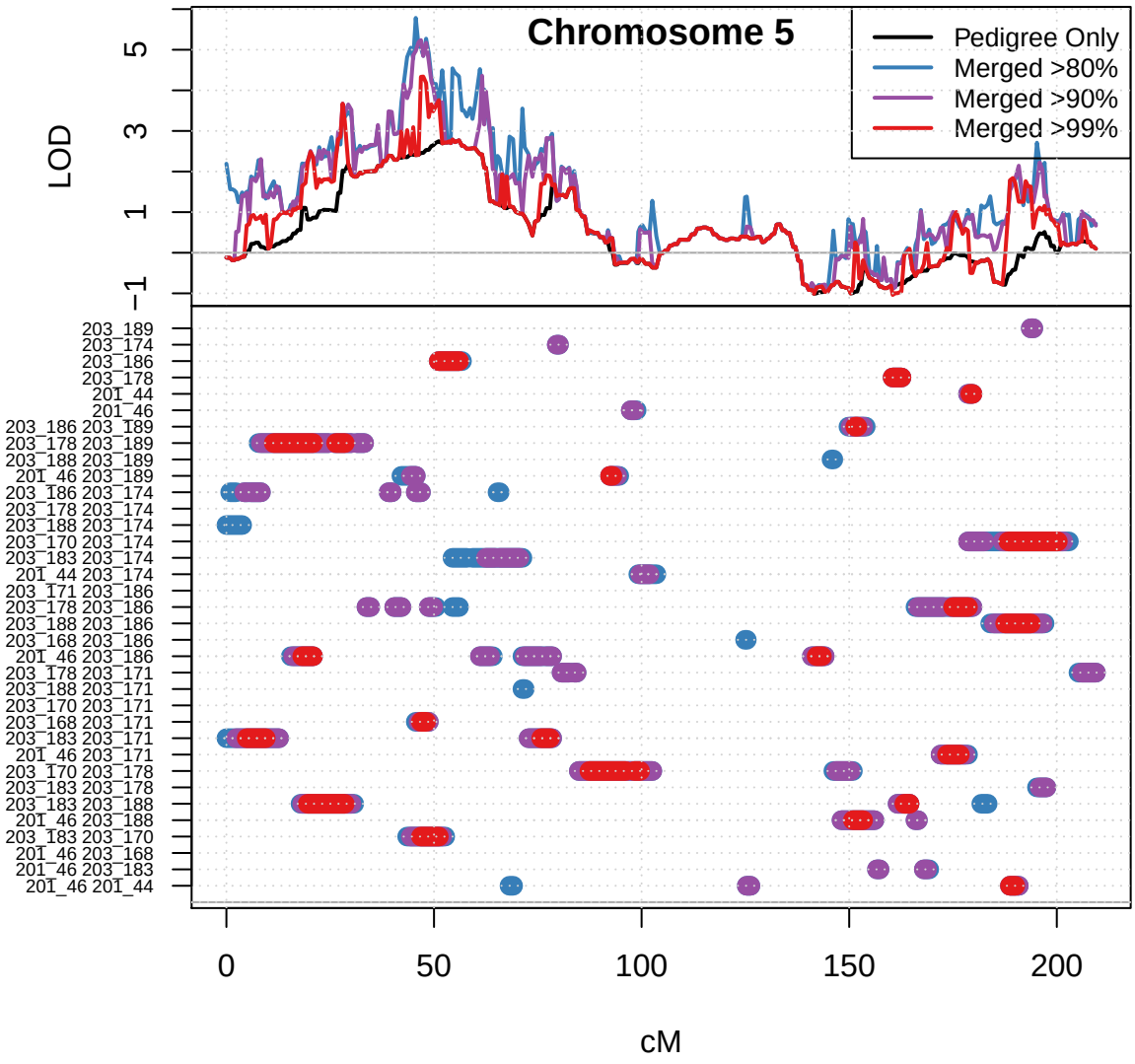



Figure 6.7: Chromosome 5 LOD scores before and after merging population-IBD at different thresholds, with detected IBD segments

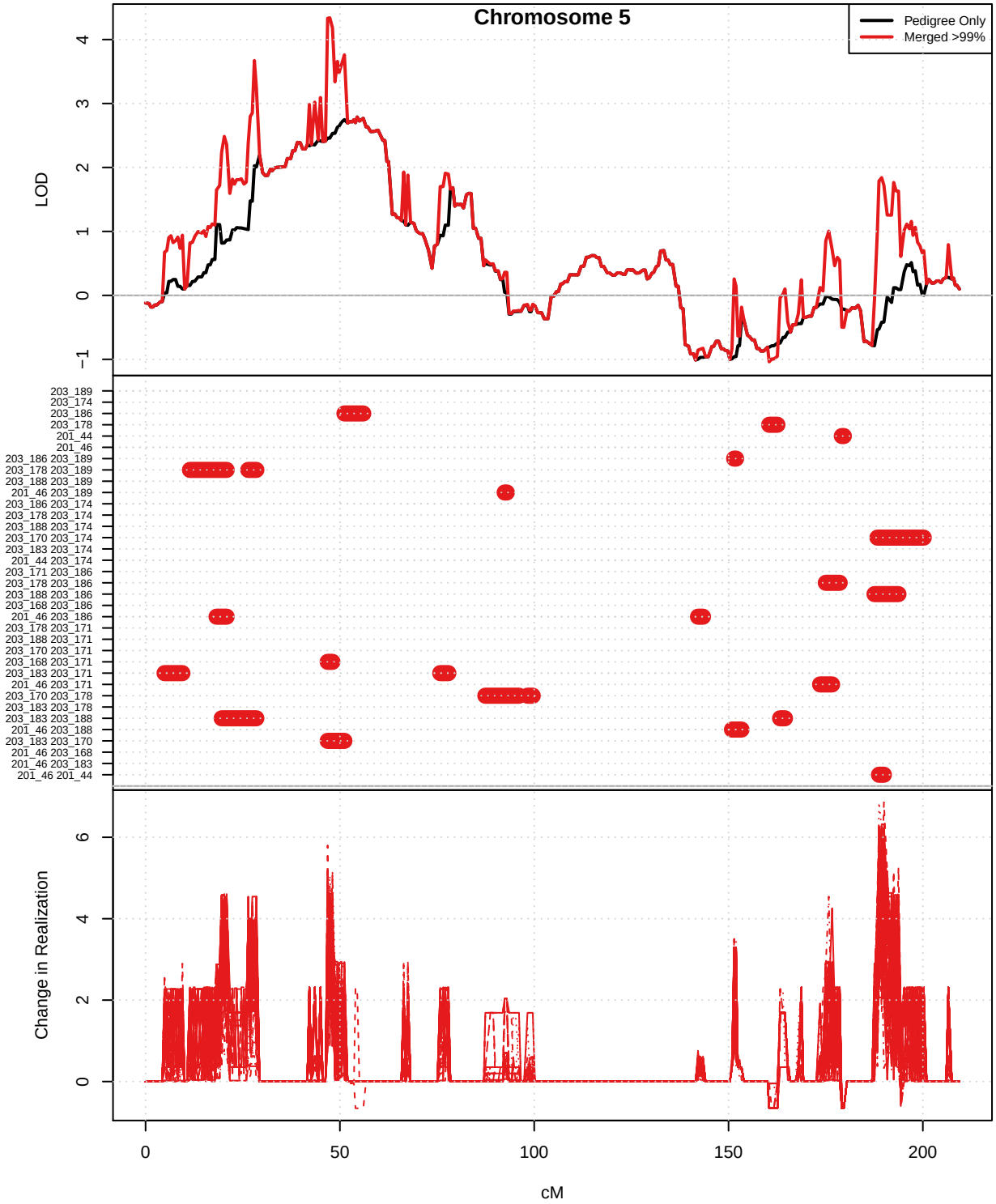



Figure 6.8: Chromosome 5 LOD scores before and after merging population-IBD at 99% threshold, with detected IBD segments and changes in LOD score contribution per realization

at the pedigree-only linkage peak. A second linkage peak of similar magnitude that was not present in the pedigree-only analysis was also discovered. On chromosome 5, the merging again resulted in an increase in the LOD score at the pedigree-only linkage peak, this time separating the linkage peak into two narrower regions. The new smaller linkage regions were caused by population-IBD between different pairs of individuals.

These results show that population-IBD detected among pedigree individuals, when merged with pedigree-IBD, can increase and refine the pedigree-only LOD score linkage peaks and to discover new linkage peaks. Most merged IBD resulted in an increase in the LOD score, so the quality of the IBD estimates was again found to be critical. The merging threshold was a useful but imperfect measure of IBD quality. A high threshold for merging of 99% gave the clearest results. At the 99% level fewer small and unreliable segments of IBD were merged, however, some IBD that is likely real was missed. An example of this was in long segments with a small gap caused by a slightly reduced IBD probability.

# Chapter 7

# CONCLUSIONS

## *7.1 Contributions*

This thesis demonstrates the importance and potential advantage of utilizing both population and pedigree information in IBD estimation. In a family study there are close relatives, $\mathcal{G}$, who can be genotyped in the most recent generations. The family pedigree on $\mathcal{F}$ where $\mathcal{G} \subset \mathcal{F}$ may also extend back a few generations and include deceased individuals who cannot be genotyped and have an unknown trait status. The larger population pedigree on $\mathcal{P}$ where $\mathcal{F} \subset \mathcal{P}$ is not observed. The further back in time individuals lived, the less likely it is that genotype or trait data are observed and the less likely it is that we know the true and complete pedigree relationships. In this thesis, pedigree structures are used to estimate IBD between close relatives in the most recent 1-4 generations of a pedigree relative to the founders of the pedigree. The IBD estimates are updated to be relative to the population founders by incorporating estimates of population-IBD between the pedigree founders.

In sib-pair studies, in Chapter 3, the only family pedigree structure used is the sibling relationship. Population-IBD between the parents of the siblings was incorporated by altering the IBD model in Section 3.1. The IBD state space between the siblings was expanded to allow for states that are only possible with related parents, and transitions probabilities around the expanded state space are integrated over all possible IBD states among the parents. The updated transition probabilities are used to estimate IBD states among siblings using an HMM model that did not model LD. As an alternative to HMM estimation of IBD states, the composite likelihood method (CM) was proposed to improve computational efficiency and robustness to the no-LD assumption. The CM and HMM method are described in Section 3.1.3.

In Section 3.1.4 the simulated sib-pair population was used to demonstrate the performance of IBD estimation with the CM method. The CM and HMM likelihoods were compared, using both the original and updated IBD transition probabilities. The updated IBD transitions allowed the estimation of IBD states that were previously not allowed in the model. The CM method was shown to have computational advantages over the HMM in some circumstances, and was more robust to LD, producing more accurate IBD estimates. In Section 3.2 the simulated sib-pair population was used to demonstrate the usefulness of IBD-based trait locus detection methods. IBD-based methods were compared with association-based methods for traits with allelic heterogeneity. IBD-based methods were able to detect rare variant traits that were not detectable in association tests.

In family studies, such as the ERF Alzheimer's disease study, the family pedigree structure used is a 4 generation pedigree. A merging algorithm was presented in Chapter 4 to combine pedigree-IBD and population-IBD. The pedigree-IBD was estimated with MCMC realizations. Guidelines for selecting a merging set of genotyped individuals for population-IBD estimation are given in Section 4.2. A method for summarizing population-IBD estimates with a consensus partition was presented in Chapter 4.3. A method for merging the population- and pedigree-IBD was presented in Section 4.4. The accuracy of the merging method was demonstrated on simulated data in Chapter 5 and applied to the ERF data in Chapter 6. The incorporation of population-IBD increased the strength of the linkage signal and narrowed the high-signal region, giving more power and resolution for trait detection. New linkage peaks were discovered, and existing linkage peaks were narrowed and increased in the ERF data.

It was demonstrated in the analyses of simulated and real data that power and precision for trait locus detection are gained by using population-IBD. The addition of population-IBD to pedigree-IBD forms new IBD groups, that if associated with trait expression, increase the LOD score signal. This was demonstrated on simulated examples in Sections 4.6 and 5.2.4. An increase in LOD score was also demonstrated on the ERF data in Section 6.4. The known family pedigree components were used as an approximate null likelihood for use in the LOD

likelihood ratio test. This was demonstrated on the simulated population in Section 4.5 to provide a good approximation.

### *7.2 Future Directions*

Improvements can be made to the merging methodology presented in Chapter 4. One improvement would be better treatment of the estimated IBD probabilities. The current method includes IBD if the probability exceeds a threshold. The threshold is used as useful but imperfect measure of IBD estimate quality. A probabilistic merging method would reflect the strength of evidence for IBD and eliminate the need for a user-defined threshold. Alternatively, more along-chromosome information could be used in deciding which IBD should be merged. For example, if there is a relatively long IBD segment with a small gap caused by a slight dip in IBD probability below the defined threshold. Currently along-chromosome information is used by IBD estimation methods but not in the merging method.

Another improvement would be in the estimation of the IBD probabilities from `ibd_stitch` realizations. IBD probabilities are obtained from the consensus membership matrix. Basic checks can be added to better ensure that the consensus obtained truly has the minimum distance over all realizations. The minimization of the distance function between the proposed consensus and the realizations uses numerical minimization. The numerical minimization runs can end in a local minimim, such as when IBD should be present between two individuals, but probabilities are split over the DNA copies, as shown in Section 4.3. Different strategies for selecting initial conditions should also be investigated. Instead of an initial random permutation, an initial median matrix could be used. With a median matrix a more informed initial condition could be selected.

More research is also required into the effect of the ascertainment of pedigrees on the analysis. Ascertainment has an effect in the selection of families, the selection of individuals in the families for the collection of genotype and phenotype data, and the selection of individuals for the estimation of population-IBD.

In real family studies the pedigrees are selected because the trait is known to be expressed

in several family members. In the simulated data, as described in Section 2.1, the pedigrees were not ascertained based on trait expression - the trait was simulated on the selected pedigrees from a sample of population FGLs. The sets of pedigrees were selected in an ad-hoc manner that resulted in long segments of IBD shared between the disjoint pedigrees. The Merge2 pedigrees were selected in part due to the sibling relationship between the two founders. A trait could be simulated in the population, and pedigrees ascertained based on trait expression without inspecting IBD. This would give simulated data results that were not biased towards finding long segments of IBD between the disjoint pedigrees more than when selected by trait expression. The shared IBD segments in pedigrees ascertained by trait expression may be smaller, thus harder to detect, but giving a larger impact on the LOD score when detected.

Furthermore, in real family studies the selection of a subset of individuals to collect genotype and phenotype data on is also based on trait expression. For example, in the Alzheimer's study only affected individuals were genotyped. In the simulated data it was assumed that all individuals had genotyped and trait data available. This level of availability of data is not realistic. In future research, availability of data can be restricted to individuals in the final two generations of the pedigree. Results can also be compared on sets of individuals where all individuals are affected versus where there is a mix of cases and controls. A pair of siblings that are both affected are more likely to share IBD than two randomly selected siblings with the same parents. Ascertainment based on trait expression biases the analysis towards the discovery of IBD segments, and results in longer IBD segments with less resolution for trait detection.

In the merging method described in Section 4.4, the selection of the subset of genotyped individuals for population-IBD estimation is also subject to ascertainment bias. Preference was given to founder individuals who were genotypically similar. This ascertainment bias made the discovery of population-IBD between the individuals more likely than in a random sample of pedigrees from the population, or in a random sample of founders from the pedigree. As the majority of population-IBD was found, this made the merged IBD state closer to the

true population-IBD state among the individuals. Further investigation is also needed into the selection of a merge set that includes non-founder and related individuals that may already share pedigree-IBD. If the merge set share pedigree-IBD, population-IBD estimates will include pedigree-IBD estimates. The effects of potentially including pedigree-IBD twice, and whether this can be prevented in the merge step, need to be investigated.

# BIBLIOGRAPHY

T2D-GENES consortium. Accessed 1 July 2016. URL `https://t2d-genes.sph.umich.edu/index.php`.

Pedfiddler version 0.6.1, 2010. URL `http://www.stat.washington.edu/thompson/Genepi/Pedfiddler.shtml`.

1000 Genomes. 1000 Genomes vcf files, 20110521 release, Version 3. ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/, April 2012a.

BEAGLE 1000 Genomes. 1000 Genomes haplotypes, Version 2 of 20110521 release. http://bochet.gcc.biostat.washington.edu/beagle/1000_Genomes.phase1_release_v3/, February 2012b.

G R Abecasis and J E Wigginton. Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers. *Am J Hum Genet.*, 77(5):754–767, 2005.

G R Abecasis, S S Cherny, W O Cookson, and L R Cardon. Merlin: rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet*, 30:97–101, 2002.

ADSP. Alzheimers disease sequencing project, 2016. URL `https://www.niagads.org/adsp`.

A. Albrechtsen, T. Sand Korneliussen, I. Moltke, T. van Overseem Hansen, F.C. Nielsen, and R. Nielsen. Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. *Genet Epidemiol*, 33:266–274, 2009.

L Almasy, T D Dyer, J M Peralta, G Jun, A R Wood, C Fuchsberger, M A Almeida, J W Kent Jr, S Fowler, T W Blackwell, S Puppala, S A Kumar, J E Curran, D Lehman, G Abecasis, R Duggirala, J Blangero, and The T2D-GENES Consortium. Data for genetic analysis

workshop 18: human whole genome sequence, blood pressure, and simulated phenotypes in extended pedigrees. *BMC Proc*, 8(S2), 2014.

W C Blackwelder and R C Elston. A comparison of sib-pair linkage tests for disease susceptibility loci. *Genet Epidemiol*, 2, 1985.

E Blue, C Cheung, C Glazner, M Conomos, S Lewis, and S Sverdlov. Identity-by-descent graphs offer a flexible framework for imputation and both linkage and association analyses. *BMC Proc*, 8(S19), 2014.

M Boehnke. Limits of resolution of genetic linkage studies: implications for the positional cloning of human desease genes. *Am J Hum Genet.*, 55(2):379–390, 1994.

K H Bonaa and D S Thelle. Association between blood pressure and serum lipids in a population. The Tromso Study. *Circulation*, 83:1305–1314, 1991. Try: http://circ.ahajournals.org/content/83/4/1305 Came from Elizabeth via email.

M D Brown, C G Glazner, C Zheng, and E A Thompson. Inferring coancestry in population samples in the presence of linkage disequilibrium. *Genetics*, 190:1447–1460, 2012.

B L Browning and S R Browning. A fast powerful method for detecting identity by descent. *Am J Hum Genet.*, 88(173-182), 2011.

S R Browning. Multilocus Association Mapping Using Variable-Length Markov Chains. *Am J Hum Genet.*, 78(6):903–913, 2006.

S R Browning. Estimation of pairwise identity by descent from dense genetic marker data in a population sample of haplotypes. *Genetics*, 178(4):2123–32, Apr 2008.

S R Browning and B L Browning. High-resoulution detection of identity by descent in unrelated individuals. *Am J Hum Genet.*, 86(4):526–539, 2010.

S R Browning and B L Browning. Identity by descent between distant relatives: detection and applications. *Annu Rev Genet*, 46:617–33, 2012.

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

S R Browning and E A Thompson. Detecting rare variant association by identity-by-descent mapping in case-control studies. *Genetics*, 190(4):1521–1531, Apr 2012.

W S Bush and J H Moore. Chapter 11: Genome-wide association studies. *PLoS Comput Biol*, 8(12), 2012.

C Cannings, E A Thompson, and E H Skolnick. Probability functions on complex pedigrees. *Adv Appl Prod*, 10:26–61, 1978.

C Cannings, E A Thompson, and M H Skolnick. *Current Developments in Anthropological Genetics*, chapter Pedigree analysis of complex models, pages 251–298. Plenum Press, New York, 1980.

R E Chandler and S Bate. Inference for clustered data using the independence log-likelihood. *Biometrika*, 94(167-183), 2007.

H J Cordell, G C Wedig, K B Jacobs, and R C Elston. Multilocus linkage tests based on affected relative pairs. *Am J Hum Genet.*, 66:1273–1286, 2000.

E H Corder and et al. Gene dose of apolipopstudiesE type 4 allele and the risk of Alzheimer's disease in late onset families. *Science*, 261:921–923, 1993.

JF Crow and M Kimura. *An introduction to population genetics theory*. Burgess Publishing, 1970.

G Csardi and T Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006. URL `http://igraph.org`.

AG Cudworth and JC Woodrow. Evidence for HLA-linked ggene in "juvelile" diabetes mellitus. *Br. Med. J.*, 3:133–135, 1975.

E Dimitriadou, A Weingessel, and K Hornik. A combination scheme for fuzzy clustering. *International Journal of Pattern Recognition and Artificial Intelligence*, 12(2):901, 2002.

S Dudoit and T P Speed. A score test for linkage using identity by descent data from sibships. *The Annals of Statistics*, 27(3):943–986, 1999.

S Dudoit and T P Speed. A score test for the linkage analysis of qualitative and quantitative traits based on identity by descent data from sib-pairs. *Biostatistics*, 1(1):1–26, 2000.

T D Dyer, J Blangero, J T Williams, H H Goring, and M C Mahaney. The effect of pedigree complexity on quantitative trait linkage analysis. *Genet Epidemiol*, 21(Suppl 1):S236–43, 2001.

R C Elston and J Stewart. A general model for the genetic analysis of pedigree data. *Hum Hered*, 21(6):523–42, 1971.

ERF. Erasmus Rucphen family study, 2016. URL `http://www.epib.nl/research/erf/erf_index.html`.

J Eu-ahsunthornwattana, EN Miller, M Fakiola, Wellcome Trust Case Control Consortium 2, SMB Jeronimo, JM Blackwell, and HJ Cordell. Comparison of methods to account for relatedness in genome-wide association studies with family-based data. *PLOS Genetics*, 2014.

W J Ewens. The sampling theory of selectively neutral alleles. *Theor Popul Biol*, 3(87-112), 1972.

D Falush, M Stevens, and J K Pritchard. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164(4):1567–87, 2003.

P Fearnhead and P Donnelly. Estimation recombination rates from population genetic data. *Genetics*, 159:1299–1318, 2001.

R A Fisher. On the interpretation of $\chi^2$ from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society*, 85(1):87–94, 1922.

C Francks, F Tozzi, A Farmer, J B Vincent, D Rujescu, D St Clair, and P Muglia. Population-based linkage analysis of schizophrenia and bipolar case-control cohorts identifies a potential susceptibility locus on 19q13. *Mol Psychiatry*, 15(3):319–25, Mar 2010.

D W Fulker and L R Cardon. A sib-pair approach to mapping of quantitative trait loci1. *Am J Hum Genet.*, 54(6):1092–103, 1994.

H N Gabow and R E Tarjan. A linear-time algorithm for a special case of disjoint set union. *Journal of Computer and System Sciences*, 30(2):209–221, 1985.

C J Geyer. Practical markov chain monte carlo. *Statistical Science*, 7(4):473–483, 1992.

S Ghosh, RM Watanabe, TT Valle, ER Hauser, VL Magnuson, CD Langefeld, DS Ally, KL Mohlke, K Silander, K Kohtamaki, P Chines, J Balow, G Birznieks, J Chang, W Eldridge, MR Erdos, ZE Karanjawala, JI Knapp, K Kudelko, C Martin, A Morales-Mena, A Musick, T Musick, C Pfahl, R Porter, JB Rayman, D Rha, L Segal, S Shapiro, R Sharaf, B Shurtleff, A So, J Tannenbaum, C Te, J Tovar, A Unni, C Welch, R Whiten, A Witt, J Blaschack-Harvan, JA Douglas, WL Duren, MP Epstein, TE Fingerlin, HS Kaleta, EM Lange, C Li, RE McEachin anf HM Stringham, E Trager, PP White, J Eriksson, L Toivanen, EH Ross, E Demirchyan, WA Hagpian, TA Buchanan, J Tuomilehto, RN Bergman, FS Collins, and M Boehnke. The Finland-United States investigation of non-insulin-dependent diabetes mellitus genetics (FUSION) study. 1. an autosomal genome scan for genes that predispose to type 2 diabetes. *Am J Hum Genet.*, 67(5):1174–1185, 2000.

M Gill, H Vallada, D Collier, P Sham, P Holmans, R Murray, P McGuffin, S Nanko, M Owen, S Antonarakis, D Housman, H Kazazian, G Nestadt, A E Pulver, R E Straub, C J MacLean, D Walsh, K S Kendler, L DeLisi, M Polymeropoulos, H Coon, W Byerley,

R Lofthouse, E Gershon, C M Read, and et al. A combined analysis of D22S278 marker alleles in affected sib-pairs: support for a susceptibility locus for schizophrenia at chromosome 22q12. Schizophrenia Collaborative Linkage Group (chromosome 22). *Am J Hum Genet*, 16(67):40–5, 1996.

C Glazner. *Monte Carlo estimation of identity by descent in populations*. PhD thesis, University of Washington, 2014.

C Glazner and E A Thompson. Improving pedigree-based linkage analysis by estimating coancestry among families. *Stat Appl Genet Mol Biol*, 11(2), Jan 2013.

C Glazner and E A Thompson. Pedigree-free descent based gene mapping from population samples. *Human Heredity*, 80:21–35, 2015.

F Grimson. *IBDLabels: Convert Between Different IBD-State Labelling Schemes.*, 2015. URL `http://R-Forge.R-project.org/projects/morgan-rtools/`. R package version 1.1/r36.

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: 318–326, 2009.

A Gusev, E E Kenny, J K Lowe, J Salit, R Saxena, S Kathiresan, D M Altshuler, J M Friedman, J L Breslow, and I Pe'er. DASH: A method for identical-by-descent haplotype mapping uncovers association with recent variation. *Am J Hum Genet.*, 88(6):706–717, Jun 2011.

A Gusev, M J Shah, E E Kenny, A Ramachandran, J K Lowe, J Salit, C C Lee, E C Levandowsky, T N Weaver, Q C Doan, H E Peckham, S F McLaughlin, M R Lyons, V N Sheth, M Stoffel, F M DeLaVega, J M Friedman, J L Breslow, and I Pe'er. Low-pass genome-wide sequencing and variant inference using identity-by-descent in an isolated human population. *Genetics*, 190(2):679–689, Feb 2012.

J Haldane. An exact test for randomness of mating. *J Genet*, 52:631–635, 1954.

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

Y Halperin, C Linhart, I Ulitsky, and R Shamir. Allegro: Analyzing expression and sequence in concert to discover regulatory programs. *Nucleic Acids Research*, pages 1–14, 2009.

J K Haseman and R C Elston. The investigation of linkage between a quantitative trait and a marker locus. *Behavior Genetics*, 2:3–19, 1972.

D He. IBD-Groupon: an efficient method for detecting group-wise identity-by-descent regions simultaneously in multiple individuals based on pairwise IBD relationships. *Bioinformatics*, 29(13):i162–i170, 2013.

S C Heath. Markov chain segregation and linkage analysis for oligogenic models. *Am J Hum Genet.*, 61:748–760, 1997.

W G Hill and A Robertson. Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics*, 38:226–231, 1968.

K Hornik. A CLUE for CLUster Ensembles. *Journ Stat Soft*, 14(12), 2005.

K Hornik. CLUE: CLUster Ensembles version 0.3-51, 2015. URL `http://CRAN.R-project.org/package=clue`.

I Ionita-Laza, S Lee, V Makarov, J D Buxbaum, and X Lin. Family-based associtation tests for sequence data, and comparisons with population-based association tests. *European Journal of Human Genetics*, 21:1158–1162, 2013.

SK Iyengar and RC Elston. The genetic basis of complex traits: rare variants or "common gene, common disease"? *Methods Mol Biol*, pages 376–84, 2007.

A Jacquard. *Structure génétique des populations*. Masson, 1970.

H M Kang, J H Sul, S K Service, N A Zaiten, S Y Kong, and et al. Variance component model to account for sample structure in genome-wide associaiton studies. *Nature genetics*, 42:348–54, 2010.

M J Khoury, W D Flanders, R B Lipton, and J S Dorman. Commentary: the affected sib-pair method in the context of an epidemiologic study design. *Genet Epidemiol*, 8:277–282, 1991.

M Knapp, SA Seuchter, and MP Baur. Linkage analysis in nuclear families 1: Optimality criteria for affected sib-pair tests. *Hum. Hered.*, 44:37–43, 1994.

E Koch, M Ristroph, and M Kirkpatrick. Long range linkage disequilibrium across the human genome. *PLoS ONE*, 8(12), 2013.

A Kong and N J Cox. Allele-sharing models: LOD scores and accurate linkage tests. *Am J Hum Genet*, 61(5):1179–88, 1997.

A Kong, D F Gudbjartsson, J Sainz, G M Jonsdottir, S A Gudjonsson, B Richardsson, S Sigurdardottir, J Barnard, B Hallbeck, G Masson, A Shlien, S T Palsson, M L Frigge, T E Thorgeirsson, J R Gulcher, and K Stefansson. A high-resolution recombination map of the human genome. *Nature Genetics*, 31:241–247, 2002.

A Kong, G Masson, M L Frigge, A Gylfason, P Zusmanovich, G Thorleifsson, P I Olason, A Ingason, S Steinberg, T Rafnar, P Sulem, M Mouy, F Jonsson, U Thorsteinsdottir, D F Gudbjartsson, H Stefansson, and K Steffansson. Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat Genet*, 40(9):1068–75, 2008.

L Kruglyak and E S Lander. Complete multipoint sib-pairs analysis of qualitative and quantitative traits. *Am J Hum Genet.*, 57(439-454), 1995.

L Kruglyak, M Daly, M P Reeve-Daly, and E S Lander. Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am J Hum Genet.*, 58:1347–1363, 1996.

M Kuhner and P Smith. Comparing likelihood and bayesian coalescent estimation of population parameters. *Genetics*, 175(1):155–165, 2007.

N Laird and C Lange. Family-based designs in the age of large-scale gene-association studies. *Genetics*, 7:385–394, 2006.

N Laird and C Lange. Family based methods for linkage and association analysis. *Advances in Genetics*, 60, 2008.

N Laird, S Horvath, and X Xu. Implementing a unified approach to family based tests of association. *Genet Epidemiol*, 19(Suppl 1):S36–S42, 2000.

E S Lander and P Green. Construction of multilocus genetic maps in humans. *Proc Natl Acad Sci USA*, 84(8):2363–7, 1987.

F Larribe and P Fearnhead. On composite likelihoods in statistical genetics. *Statistica Sinica*, 21:43–69, 2011.

G M Lathrop. Multilocus linkage analysis in humans: detection of linkage and estimation of recombination. *Am J Hum Genet.*, 37(3):482–98, 1985.

S L Lauritzen. Propagation of probabilities, means, and variances in mixed graphical association models. *Am Stat Assoc*, 87(420):1098–1108, 1992.

E. Letouze, A. Sow, F. Petel, R. Rosati, B. C. Figueiredo, N. Burnichon, A. P. Gimenez-Roqueplo, E. Lalli, and A. L. De Reynis. Identity by descent mapping of founder mutations in cancer using high-resolution tumor SNP data. *PLoS ONE*, 7(5), 2012.

A Leutenegger, B Prum, E Genin, C Verny, A Lemainque, F Clerget-Darpoux, and E A Thompson. Estimation of the inbreeding coefficient through use of genomic data. *Am J Hum Genet.*, 73(3):516–523, Sep 2003.

H Levene. On a matching problem arising in genetics. *Ann Math Stat*, 20, 1949.

B G Lindsay. Composite likelihood methods. *Contemporary Math.*, 80(221-239), 1988.

F Liu, A Arias-Vasquez, K Sleegers, Y S Aulchenko, M Kayser, P Sanchez-Juan, B Feng, A M Bertoli-Avella, J van Swieten, T I Axenovich, P Heutink, C van Broeckhoven, B A Oostra, and C M van Dujin. A genomewide screen for late-onset Alzheimer disease in a genetically isolated Dutch population. *Am J Hum Genet.*, 81:17–31, 2007.

T C Matise, F Chen, W Chen, F M De La Vega, M Hansen, C He, F C Hyland, G C Kennedy, X Kong, SS Murray, and et al. A second-generation combined linkage physical map of the human genome. *Genome Res*, 17(12):1783–1786, 2007.

G McVean, S Myers, S Hunt, P Deloukas, D Bentley, and P Donnelly. The fine-scale structure of recombination rate variation in the human genome. *Science*, 304(581-584), 2004.

G A McVean and N J Cardin. Approximating the ccoalescent with recombination. *Philos Trans R Soc Lond B Biol Sci.*, 360(1459):1387–93, 2005.

J Merriam. *Encyclopedia of Genetics*, chapter Independent Assortment, pages 1017–1018. Academic Press, 2001.

M F Moffat, P A Sharp, J A Faux, R P Young, W O Cookson, and J M Hopkin. Factors confounding genetic linkage between atopy and chromosome 11q. *Clin Exp Allergy*, 22 (12):1046–51, 1992.

I Moltke, A Albrechtsen, T Hansen, F C Nielsen, and R Nielsen. A method for detecting IBD regions simultaneously in multiple individuals with applications to disease genetics. *Genome Res.*, 21(1168-1180), 2011.

MORGAN. *Software for Markov Chain Monte Carlo in genetic analysis. Verson 3.3.1.* www.stat.washington.edu/thompson/Genepi/MORGAN/Morgan.shtml, Feb 2016.

National Library of Medicine (US). APP. In *Genetics Home Reference [Internet]*, http://ghr.nlm.nih.gov/gene/APP, February 2016a. Bethesda (MD): The Library. http://ghr.nlm.nih.gov/Citing.

National Library of Medicine (US). HLA. In *Genetics Home Reference [Internet]*, http://ghr.nlm.nih.gov/geneFamily/hla, February 2016b. Bethesda (MD): The Library.

A Q Nato, N H Chapman, H K Sohi, H D Nguyen, Z Brkanac, and E M Wijsman. PBAP: a pipeline for file processing and quality control of pedigree data with dense genetic markers. *Bioinformatics*, 1(31):3790–8, Dec 2015.

J R O'Connell and D E Weeks. The VITESSE algorithm for rapid exact multilocus linkage analysis via genotype set-recoding and fuzzy inheritance. *Nat Genet*, 11:402–408, 1995.

A Ott, M M Breteler, F van Harskamp, J J Claus, T J van der Cammen, D E Grobbe, and A Hoffman. Prevalence of alzheimer's disease and vascular dementia: associaiton with education. *BMJ*, 310:970–973, 1995.

J Ott. Estimation of the recombination fraction in human pedigrees: efficient computation of the likelihood for human linkage studies. *Am J Hum Genet.*, 26(5):588–97, 1974.

J Ott, Y Kamatani, and M Lathrop. Family-based designs for genome-wide association studies. *Nature Reviews Genetics*, 12:465–474, 2011.

P F Palamara and I Pe'er. Inference of historical migration rates via haplotype sharing. *Bioinformatics*, 29(13):180–188, 2013.

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

A Papassotiropopulos, M Fountoulakis, T Dunckley, D A Stephan, and E M Reiman. Ggenetic, transcriptomics and proteomics of Alzheimer's disease. *J Clin Psychiatry*, 64, 2008.

L S Penrose. The genetic background of common diseases. *Acta Genet*, pages 257–265, 1953.

M A Pericak-Vance, J L Bebout, P C Gaskell, L H Yamaoka, W Y Hung, M J Alberts, A P Walker, R J Bartlett, C A Haynes, and K A Welsh. Linkage studies in familial Alzheimer disease: evidence for chromosome 19 linkage‘. *Am J Hum Genet*, 48(6):1034–50, 1991.

A L Price, N J Patterson, R M Plenge, M E Weinblatt, and N A Shadick and. Principle components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38:904–909, 2006.

K Punera and J Ghosh. Consensus based ensembles of soft clusterings. *Applied Artificial Intelligence*, 22(7-8):3–9, 2006.

S Purcell. PLINK 1.07. http://pngu.mgh.harvard.edu/purcell/plink/, 2015.

S Purcell, B Neale, K Todd-Brown, L Thomas, M A Ferreira, D Bender, J Maller, P Sklar, P I de Bakker, M J Daly, and C Sham P. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.*, 81(3):559–75, Sep 2007a.

S Purcell, B Neale, K Todd-Brown, L Thomas, M A R Ferreira, D Bender, J Maller, P Sklar, PIW DeBakker, M J Daly, and PC Sham. PLINK: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics*, 81, 2007b.

Y Qian, B L Browning, and S R Browning. Efficient clustering of identity-by-descent between multiple individuals. *Bioinformatics*, 30(7):915–22, Apr 2014.

D Rabinowitz and N Laird. A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Hum Hered*, 50(4):211–23, 2000.

N Risch. Linkage strategies for genetically complex traits. 1. multilocus models. *Am J Hum Genet*, 46(2):222–228, 1990.

M Saad, A Q Nato, F L Grimson, S M Lewis, L A Brown, E M Blue, T Thornton, E A Thompson, and E M Wijsman. Identity-by-descent estimation with population- and pedigree-based imputation in admixed family data. *BMC Proc*, 10(7):295–301, 2016.

E Sobel and K Lange. Descent graphs in pedigree analysis: Applications to haplotyping, location scores and marker-sharing statistics. *Am J Hum Genet.*, 58:1323–1337, 1996.

P R Sosnay, K R Siklosi, F Van Goor, K Kaniecki, H Yu, N Sharma, A S Ramalho, M D Amaral, R Dorfman, J Zielenski, D L Masica, R Karchin, L Millen, P J Thomas, G P Patrinos, M Corey, M H Lewis, J M Rommens, C Castellani, C M Penland, and G R Cutting. Defining the disease liability of variants in the cystic fibrosis transmembrane conductance regulator gene. *Nat Genet.*, 45(19):1160–1167, 2013.

R S Spielman, R E McGinnis, and W J Ewens. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet.*, 1993.

A Strehl and J Ghosh. Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617, 2002.

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

The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449:851–861, 2007.

E A Thompson. Gene identities and multiple relationships. *Biometrics*, 30:667–680, 1974.

E A Thompson. *Statistical Inferences from Genetic Data on Pedigrees*, volume 6 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. IMS, Beachwood, OH, 2000.

E A Thompson. The IBD process along four chromosomes. *Theor Pop Biol*, 73(3):369–373, 2008.

E A Thompson. Inferring coancestry of genome segments in populations. In *Invited Proceedings of the 57th Session of the International Statistical Institute*, volume IPM13, Durban, South Africa, 2009. Paper 0325.

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

T Thornton, M P Conomos, S Sverdlov, E M Blue, C Y Cheung, C G Glazner, S M Lewis, and E M Wijsman. Estimating and adjusting for ancestry admixture in statistical methods for relatedness inference, heritability estimation, and association testing. *BMC Proc*, 8(Suppl 11)(S5), 2014.

I Tomlinson, E Webb, L Carvajal-Carmona, P Broderick, Z Kemp, S Spain, S Pengar, I Chandler, M Gormand, W Wood, and et.al.Kr. A genome-wide association scan of tag snps identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nat Genet.*, 39: 984–988, 2007.

L P Tong and E A Thompson. Multilocus lod scores in large pedigrees: combination of exact and approximate calculations. *Hum Hered*, 65:142–153, 2008.

C Varin, N Reid, and D Firth. An overview of composite likelihood methods. *Statistica Sinica*, 21:5–42, 2011.

D E Weeks and K Lange. The affected-pedigree-member method of linkage analysis. *Am J Hum Genet.*, 42(315-326), 1988.

E M Wijsman and C I Amos. Genetic analysis of simulated oligogenic traits in nuclear and extended ppedigree: Summary of GAW10 contributions. *Genet Epidemiol*, 14:719–735, 1997.

A F Wright, A D Carothers, and M Pirastu. Population choice in mapping genes for complex diseases. *Nat Genet*, 23:397–404, 1999.

C Zheng, M K Kuhner, and E A Thompson. Joint inference of identity by descent along multiple chromsomes from population samples. *J Comput Biol*, 21(3):185–200, 2014.

# Appendix A

# IBDLABELS PACKAGE

As discussed in Section 1.2, the joint IBD state between a group of individuals can be represented as a graphical model, a partition, or as a function of inheritance vectors. In the graphical representation nodes represent IBD classes and edges represent individuals. An equivalent representation is as a partition where each DNA copy is an object is assigned an arbitrary label. All DNA copies with the same label are IBD with one another.

IBD is most commonly described between a pair of individuals, or four DNA copies. In this case, there are 15 graphs that represent all possible IBD configurations between the pair. The 15 states are commonly referred to simply by numerical labels introduced by Jacquard [1970]. If the phase of the two DNA copies of an individual is unknown there are only 9 distinguishable states which also have Jacquard numbers.

For an arbitrary number of individuals, Thompson [1974] developed a numbering system for IBD states. This system converts between a vector of labels that describe the IBD partition to a positive integer. This number will be referred to as the IBD "Label" for a given IBD state. The ordering of IBD states from the state with the smallest integer label to the largest will be referred to as the "Lexicographic" ordering of the states.

The IBDLabels package Grimson [2015] was developed to convert between IBD labels (from Jacquard [1970] and Thompson [1974]) and vector representations of the IBD states. Tables A.1 and A.2 are taken from the documentation of the IBDLabels package. They show the equivalence between the numbering systems.

| | Label | Lex | 15-state Jacquard | 9-state Jacquard | State Vector | Num copies IBD |
|---|---|---|---|---|---|---|
|  | 0 | 1 | 1 | 1 | 1111 | 2 |
|  | 4 | 4 | 2 | 2 | 1122 | 0 |
|  | 1 | 2 | 3 | 3 | 1112 | 1 |
|  | 3 | 3 | 4 | 3 | 1121 | 1 |
|  | 5 | 5 | 5 | 4 | 1123 | 0 |
|  | 6 | 6 | 6 | 5 | 1211 | 1 |
|  | 10 | 10 | 7 | 5 | 1222 | 1 |
|  | 14 | 14 | 8 | 6 | 1233 | 0 |
|  | 7 | 7 | 9 | 7 | 1212 | 2 |
|  | 9 | 9 | 10 | 7 | 1221 | 2 |
|  | 8 | 8 | 11 | 8 | 1213 | 1 |
|  | 12 | 12 | 12 | 8 | 1231 | 1 |
|  | 11 | 11 | 13 | 8 | 1223 | 1 |
|  | 13 | 13 | 14 | 8 | 1232 | 1 |
|  | 15 | 15 | 15 | 9 | 1234 | 0 |

Table A.1: IBD state equivalences, ordered by Jacquard

| | Label | Lex | 15-state Jacquard | 9-state Jacquard | State Vector | Num Copies IBD |
|---|---|---|---|---|---|---|
|  | 0 | 1 | 1 | 1 | 1111 | 2 |
|  | 1 | 2 | 3 | 3 | 1112 | 1 |
|  | 3 | 3 | 4 | 3 | 1121 | 1 |
|  | 4 | 4 | 2 | 2 | 1122 | 0 |
|  | 5 | 5 | 5 | 4 | 1123 | 0 |
|  | 6 | 6 | 6 | 5 | 1211 | 1 |
|  | 7 | 7 | 9 | 7 | 1212 | 2 |
|  | 8 | 8 | 11 | 8 | 1213 | 1 |
|  | 9 | 9 | 10 | 7 | 1221 | 2 |
|  | 10 | 10 | 7 | 5 | 1222 | 1 |
|  | 11 | 11 | 13 | 8 | 1223 | 1 |
|  | 12 | 12 | 12 | 8 | 1231 | 1 |
|  | 13 | 13 | 14 | 8 | 1232 | 1 |
|  | 14 | 14 | 8 | 6 | 1233 | 0 |
|  | 15 | 15 | 15 | 9 | 1234 | 0 |

Table A.2: IBD state equivalences, ordered by Label