Realized genome sharing in random effects models for quantitative genetic traits

Bowen Wang

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

Doctor of Philosophy

University of Washington

2019

Reading Committee:

Elizabeth A. Thompson, Chair

Timothy A. Thornton

Ellen M. Wijsman

Program Authorized to Offer Degree:
Statistics

University of Washington

Abstract

Realized genome sharing in random effects models for quantitative genetic traits

Bowen Wang

Chair of the Supervisory Committee:
Professor Elizabeth A. Thompson
Statistics

DNA copies inherited from the same ancestral copy by related individuals are said to be identical by descent (IBD). IBD gives rise to genetic similarities between related individuals. In quantitative genetics, two fundamental problems are heritability estimation and gene mapping for genetic traits. IBD plays a critical role in the study of both problems. When working with population-based samples where pedigree information is unavailable, it is essential to estimate IBD accurately from genetic marker data using pedigree-free methods. The estimated IBD can then be used in heritability estimation and gene mapping using random effects models.

For pedigree-free IBD estimation, we showed that it is important to use the fact that DNA is inherited in segments as opposed to independent loci. As the single nucleotide polymorphism (SNP) marker panels become increasingly dense, the impact of allelic association (or linkage disequilibrium, LD) on accuracy of IBD estimation also grows. Through simulation studies, we demonstrated that adjusting for LD in the marker panel can lead to improved IBD estimation accuracy. For heritability estimation and gene mapping using random effects models, a difficult task is to specify the correlation structures of the random genetic effects, which are typically functions of IBD sharing over the putative causal genomic region. We provided formulas for the asymptotic bias and sampling error of heritability estimates, when the ge-

netic correlation structures are potentially mis-specified. Mis-specification of the genetic correlation structures can occur due to inaccurate IBD estimation or mis-identification of the causal genome. We showed that such mis-specification can lead to substantial downward bias in heritability estimation, or loss of power in gene mapping.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

This dissertation was made possible by my advisor, Elizabeth Thompson, who has provided me invaluable guidance and mentoring since my second year at the University of Washington. I would also like to thank my committee members, Ellen Wijsman, Timothy Thornton, and Jonathan Wakefield, for their support throughout this process. In particular, Ellen offered me the opportunity to work on the Alzheimer's Disease Sequencing Project, a great experience for me to get involved with applied work.

I have benefitted greatly from interactions with members of the Thompson group, including Serge Sverdlov, Fiona Grimson, Aaron Baraff, Steven Lewis, Jesse Raffa, John Ronola, and Saonli Basu, who has visited the group during her sabbatical. In addition, I want to thank members of UW's statistical genetics and population genetics seminars, from whom I have learnt a great deal.

Thank you to all my friends in UW's statistics and biostatistics departments. The journey would not have been so memorable without them.

Lastly, I want to thank my wife Manlin for her unwavering support throughout the process, and my son Nathan for giving me that extra energy and determination to push on.

DEDICATION

to my dear wife, Manlin, and my son, Nathan

Chapter 1
INTRODUCTION

1.1 Identity by descent

DNA is passed on from parents to offspring through meiosis. For diploid species such as humans, there is variation in meiosis due to Mendelian segregation. At any given point in the genome, Mendel's first law (Mendel 1866) states that a random copy of the two parental alleles will be passed on to the offspring with 50:50 probability. This provides the basis for analyzing coancestry among individuals at a single locus. Within a pedigree, the inheritance pattern at locus j is represented by the inheritance vector $S_{\cdot,j}$, whose elements are indicators for the grandparental origins of gametes passed on to offspring at that locus in all relevant meioses. For meiosis i, the meiosis vector $S_{i,\cdot}$ captures the grandparental origins of gametes passed on to the offspring at all loci in that meiosis. Together, $\boldsymbol{S} = \{S_{i,j}\}$ provides the complete inheritance pattern at all loci in a pedigree. The rows of $\boldsymbol{S}$ are independent of each other due to independence of meioses. Under the assumption of no genetic interference, there is first order Markov dependence between elements within each row of $\boldsymbol{S}$. Change of values (between 0 and 1) in $S_{i,\cdot}$ are due to recombination events in meiosis i. Recombination is a coarse process. In one meiotic event, the expected number of recombinations across all 22 autosomes is about 36.

DNA copies that are inherited from the same ancestral copy are said to be identical by descent (IBD). There is no absolute measure of IBD. IBD is always relative to some ancestral reference population. In the general case, this reference may be the population at some past time point, with the implication that coancestry beyond the reference time point

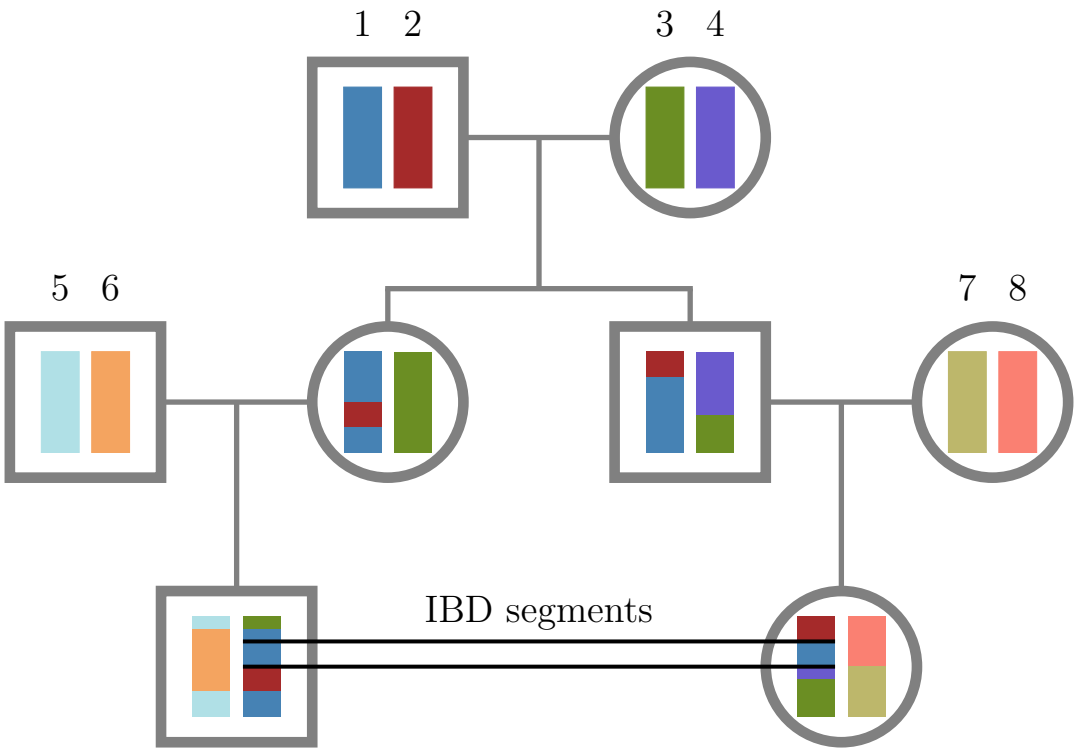


Figure 1.1: Example of DNA transmission over a chromosome segment in a three-generation pedigree. Distinct founder haplotypes are shown in different colors, and numbered 1 through 8. Segments that have the same colors at the same location on different chromosomes are IBD.

is ignored. In the case of a defined pedigree, the pedigree founders (whose parents are not present in the pedigree) form the reference population. Figure 1.1 shows an example of DNA transmission over a chromosome segment in a three-generation pedigree. As the reference population for determining IBD, the founders of the pedigree carry distinct chromosomes labeled 1 through 8. Recombination events occurred in some of the meioses, leading to non-founders having different ancestral origins at different positions on the same chromosome (represented by different colors).

Two copies of DNA are either IBD or not. The number of possible IBD states grows more than exponentially as we consider IBD sharing between more DNA copies (Thompson 1974).

IBD state	IBD partition	Jacquard state	IBD copies	ϕ
1111	(a, b, c, d)	1	—	1
1122	$(a, b)(c, d)$	2	—	0
1112	$(a, b, c)(d)$	3	—	0.5
1121	$(a, b, d)(c)$	3	—	0.5
1123	$(a, b)(c)(d)$	4	—	0
1211	$(a, c, d)(b)$	5	—	0.5
1222	$(a)(b, c, d)$	5	—	0.5
1233	$(a)(b)(c, d)$	6	—	0
1212	$(a, c)(b, d)$	7	2	0.5
1221	$(a, d)(b, c)$	7	2	0.5
1213	$(a, c)(b)(d)$	8	1	0.25
1231	$(a, d)(b)(c)$	8	1	0.25
1223	$(a)(b, c)(d)$	8	1	0.25
1232	$(a)(b, d)(c)$	8	1	0.25
1234	$(a)(b)(c)(d)$	9	0	0

Table 1.1: Measures of IBD sharing between four copies of DNA carried by a pair of individuals at a single locus. When parental origins of the DNA copies are known, they are represented by a and b for individual 1, and c and d for individual 2 respectively.

IBD states can also be viewed as partitions, where DNA copies that are mutually IBD belong to the same partition. Between the four copies of DNA carried by a pair of individuals at a single locus, there are 15 possible IBD states/partitions. Pairwise IBD sharing has been studied extensively in the literature that additional measures of IBD sharing have been defined. When the parental origins of the DNA carried by the pair of individual are unknown, the 15 IBD states reduce to 9 genotypically distinct classes (Jacquard 1974). If the pair of individuals are both non-inbred (parents are not related), the number of possible classes reduces

further to 3, where the pair share $0, 1$ or 2 copies of DNA IBD respectively. The probabilities that two non-inbred individuals share 0, 1 or 2 copies of DNA at a randomly chosen locus are referred to as the k coefficients, k_0, k_1 and k_2 (Cotterman 1940). The kinship coefficient, ϕ, is defined as the probability that a randomly chosen DNA copy from each individual in a pair are IBD. Table 1.1 shows the relationship between these IBD measures on the four DNA copies carried by a pair of individuals. Within a defined pedigree, the inheritance vector $S_{\cdot,j}$ determines the IBD state/partitions at locus j, which in turn determines the Jacquard state and kinship in the case of pairwise IBD sharing.

At a single locus, the four copies of DNA carried by a pair of individuals belong to exactly one of the possible states under the respective measures of IBD sharing shown in Table 1.1. The pedigree provides the null distribution over those possible states. For a pair of non-inbred full siblings, for example, they have IBD state 1212, 1213, 1232 and 1234 with probabilities 0.25 each. In terms of kinship, they have $\phi = 0$, 0.25 and 0.5 with probabilities 0.25, 0.5 and 0.25 respectively. Since ϕ is a numeric random variable, it makes sense to take expectations (over realizations on pedigree) and averages (across loci). The pedigree kinship, Ψ, is the pedigree expectation of kinship. The genome-wide realized kinship, Φ, is the average of (local) kinship across genome-wide loci. Ψ is a deterministic value given pedigree relationship, whereas Φ is a random variable that varies widely around its expectation Ψ (Hill and Weir 2011). 2Ψ and 2Φ represent the expected and actual proportion of genome-shared IBD between a pair of individuals respectively. The focus of this thesis is on pairwise IBD sharing and its use in the analyses of quantitative traits. The various measures of kinship (ϕ, Φ and Ψ) will be central in subsequent chapters.

1.2 Pedigree-free estimation of pairwise IBD

We do not observe IBD directly. Instead, we infer IBD from observed marker data. Advances in genomic technologies have enabled us to infer IBD using dense single nucleotide polymorphism (SNP) marker data. DNA that is IBD has very high probability of hav-

ing the same allelic type. Conversely, similarity in allelic types provides evidence for IBD. While each SNP marker is relatively uninformative, IBD segments typically contain many SNPs that collectively provide enough information for accurate estimation of IBD. Correlation of allelic types has long been used to measure relatedness between individuals (Wright 1922). In practice, a commonly used assumption in estimation of IBD is that IBD DNAs are of the same allelic types, whereas non-IBD DNAs are of independent allelic types. Allele frequencies are usually estimated from large reference population samples such as the 1000 Genomes Data (The 1000 Genomes Project Consortium 2015). With the additional assumption of Hardy-Weinberg Equilibrium (HWE), one can relate observed marker genotypes (e.g., counts of reference alleles) of a pair of individuals to IBD sharing through conditional probabilities of the form $\mathbb{P}\{\text{genotypes}|\text{inheritance}\}$, or through conditional expectations of the form $\mathbb{E}\{\text{genotypes}|\text{inheritance}\}$. This provides the basis for inferring IBD from genetic marker data.

A complicating factor in IBD estimation is allelic association (or linkage disequilibrium, LD) between linked loci. LD can arise due to mutation or population substructure. When LD is present, the allelic types carried by different individuals at linked loci are no longer independent conditional on the underlying IBD states, violating the conditional independence assumption used by many methods. Modeling for LD is difficult due to the sample size required to obtain accurate estimates of haplotype frequencies. It is an area of active research (see, for example, Albrechtsen *et al.* (2009); Browning (2008); Browning and Browning (2010)). We do not explicitly model for LD in this thesis, but we discuss the impact of LD on IBD estimation and the potential improvement that can result from adjusting for LD in IBD estimation (Chapters 2 and 3).

This thesis focuses on pedigree-free estimation of pairwise IBD. The pedigree, if known, provides a strong prior for estimating IBD with genetic marker data. The impact of this prior on IBD estimation reduces as marker data become more informative. Methods that esti-

mate IBD without pedigree information are important in population-based studies, where the pedigree is not available. They are also very useful when pedigree information is incomplete or inaccurate. For pairwise IBD, we are primarily interested in estimation of genome-wide realized kinship Φ, and local kinship ϕ. Since $\Phi \approx \frac{1}{L}\sum_{j=1}^{L}\phi_j$ across genome-wide markers, the estimated genome-wide realized kinship, $\widehat{\Phi}$, is naturally a by-product of local kinship estimation over the whole genome. For local kinship estimation, the order of markers along a chromosome is important. This information can be specified either in the form of basepair positions of markers, or genetic distance of markers relative to a genetic map. Φ can also be estimated without information on marker order/positions. In that case, the selected markers are treated as a permutable set of independent and random samples from the genome. We will discuss estimation of Φ and ϕ in detail in Chapters 2 and 3.

1.3 Use of pairwise IBD in analyses of quantitative traits

The use of IBD to explain phenotypic similarities between relatives has had a long history. Fisher (1918) considered phenotypic correlations among relatives within the framework of Mendelian segregation. That was long before SNP marker data became available, so that one could only work with probabilities or expectations of realized kinship given pedigree (e.g., pedigree kinship Ψ). For any trait that has a genetic basis, it is reasonable to argue that the actual amount of genetic sharing, as opposed to the expected amount would lead to phenotypic similarities between relatives. In the past two decades, there has been an abundance of research interests in IBD estimation and the use of estimated IBD to study genetic traits. For estimated pairwise IBD from population samples, two important use cases are heritability estimation and gene mapping.

1.3.1 Heritability estimation

The narrow sense heritability of a trait, h^2, is defined as the proportion of phenotypic variance (σ^2) that is explained by additive genetic effect (with additive genetic variance σ_a^2), *i.e.*, $h^2 = \sigma_a^2/\sigma^2$. Fisher (1918) laid the foundation for earlier approaches for heritability estimation, as

phenotypic covariance of relatives is decomposed into variance components. Traditionally, heritability was estimated from simple and balanced designs, as functions of correlation between parent-offspring pairs or sibling pairs, and the difference in correlation between monozygotic and dizygotic twin pairs (Falconer and Mackay 1996). More recently, it has become popular to estimate heritability from unbalanced designs using linear mixed models, where the additive genetic and environmental variances can be estimated more efficiently (e.g. Kruuk 2004; Yang *et al.* 2010). Ignoring fixed effects such as age and sex typically used in the linear mixed models, the resulting random effects models have the basic form

$$\boldsymbol{y} = \boldsymbol{g} + \boldsymbol{e}, \tag{1.1}$$

where $\boldsymbol{y}$ is the vector of mean-adjusted trait values, $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \sigma_a^2 \boldsymbol{G})$ are additive genetic random effects, and $\boldsymbol{e} \sim \mathcal{N}(\boldsymbol{0}, \sigma_e^2 \boldsymbol{I})$ are individual environmental random effects. The additive correlation matrix, $\boldsymbol{G}$, takes the form of twice the matrix of some kinship measure, such as the pedigree-based numerator relationship matrix (Henderson 1976), or the SNP-based Genomic Relationship Matrix (VanRaden 2008).

1.3.2 Gene mapping

The variance component model for linkage analysis (Amos 1994; Almasy and Blangero 1998) in its basic form assumes that the quantitative trait value is the sum of a grand mean, independent random effects attributed to quantitative trait loci (QTL), and an individual environmental random effect. In practice, it is essential to consider one QTL at a time so that the total number of statistical tests needed to perform is manageable. This means when testing for the effect of a QTL at locus l, effects of all other (unknown) QTL are combined into one residual genetic random effect. The fitted model has the form

$$\boldsymbol{y} = \boldsymbol{a}_l + \boldsymbol{g} + \boldsymbol{e}, \tag{1.2}$$

where $\boldsymbol{y}$ are the mean-adjusted trait values, $\boldsymbol{a}_l \sim \mathcal{N}(\boldsymbol{0}, \sigma_l^2 \boldsymbol{G}_l)$ are the genetic effects attributed to locus l, $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \sigma_g^2 \boldsymbol{G}_g)$ are the residual genetic random effects from all other QTL, and

$\boldsymbol{e} \sim \mathcal{N}(\mathbf{0}, \sigma_e^2 \boldsymbol{I})$ are the unique environmental effects. The correlation matrix $\boldsymbol{G}_l$ is twice the matrix of estimated kinship at locus l, whereas $\boldsymbol{G}_g$ can be twice the matrix of genome-wide realized kinship. The goal is to test if $\sigma_l^2 > 0$.

1.4 Data used throughout this thesis

1.4.1 Marker data

Two marker panels will be used for simulation studies in later chapters. The first will be referred to as the K170 marker panel, where a total of 169,751 SNP markers were selected from the 22 autosomes based on spacing (∼50 markers/cM), minor allele frequency (≥ 0.05) and complete genotype information in the 1000 Genomes Project Phase 3 data (The 1000 Genomes Project Consortium 2015). The second is the OmniExpress24 marker panel, where we retained 670,810 autosomal markers from the genome wide association study (GWAS) chip that can be found in the 1000 Genomes data. An additional selection criteria for both marker panels is that the genetic positions (computed under the assumption of the Haldane map function) of the markers are available in the Rutgers Map v.3a (Matise *et al.* 2007).

1.4.2 Pedigrees

Three special pedigrees will be used throughout the thesis. Figure 1.2 shows the JV pedigree of Goddard *et al.* (1996). Consanguineous marriages occurred in the third and fourth generations, leading to individuals 41, 42 and 51 being inbred. The 14-member three-generation pedigree shown in Figure 1.3 represents a more typical pedigree structure that can be found in studies involving pedigree samples. Figure 1.4 shows the highly inbred pedigree of Cleopatra VII Philopator (Wikipedia 2018), also having 14 members.

1.5 Thesis outline

This main theme of this thesis is on pedigree-free estimation of pairwise IBD, and the use of estimated pairwise IBD in analyses of quantitative trait through random effects models.

Chapter 2 focuses on estimation of genome-wide realized kinship, Φ. We introduce a generalized framework for a class of estimators, propose two new estimators within the class, and compare performance of various estimators in a simulation study. Chapter 3 delves into local kinship estimation. We investigate factors that affect local kinship estimation accuracy, such as marker density and level of LD present in the marker panel. Chapter 4 discusses the topic of heritability estimation. We investigate the impact of mis-specifying the genetic correlation matrix on heritability estimation, and how this could be one of the factors contributing to the "missing heritability" problem (Maher 2008). Chapter 5 looks at gene mapping through population-based linkage analysis. We discuss how mis-specifying either the local or residual genetic correlation matrix could affect power to detect linkage at a causal locus. Chapter 6 concludes the thesis with a discussion.

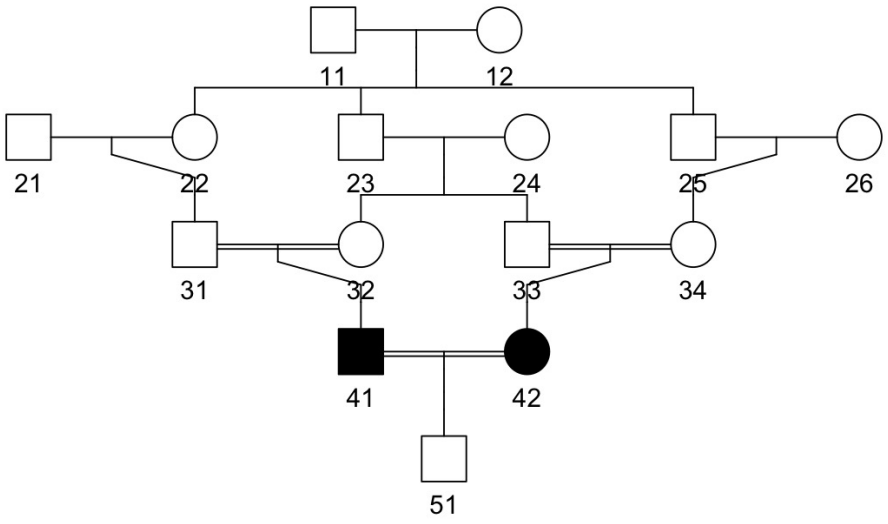


Figure 1.2: The JV pedigree. Individual 41 and 42 are the inbred quadruple cousins (IN) considered in the simulation study. Double lines indicate consanguineous marriages.

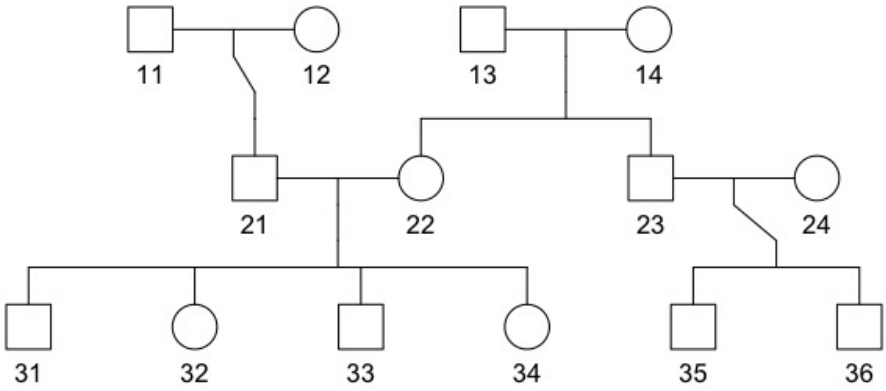


Figure 1.3: The three-generation pedigree with 14 members.

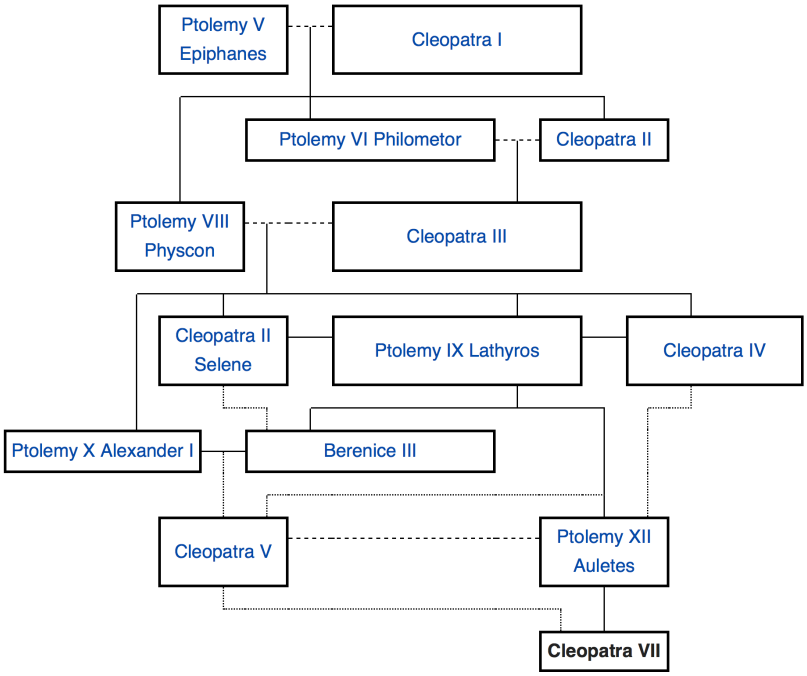


Figure 1.4: The Cleopatra pedigree.

Chapter 2

ESTIMATION OF GENOME-WIDE REALIZED KINSHIP

In this chapter we focus on pedigree-free estimation of genome-wide realized kinship. We first provide an overview of current methods. Then we introduce a generalized framework for a class of estimators. We show that several existing estimators are special cases within this class, and propose two new estimators from this class. In a simulation study, we demonstrate that improved estimators of realized kinship can be obtained (1) by optimal weighting of markers, (2) by taking physical contiguity of genome into account, and (3) by weighting on the basis of LD.

2.1 Overview of current methods

One group of existing estimators of realized kinship makes use only of the population allele frequencies at each SNP marker, and not of additional information such as the ordering of markers along the chromosome. Estimators in this group require only dense SNP genotypes and reliable sources of marker allele frequencies as input. The PLINK (Purcell *et al.* 2007) method-of-moments estimator ($\widehat{\Phi}_P$) estimates realized kinship from the k coefficients; the proportion of genome at which two non-inbred individuals share 0, 1 or 2 genes IBD (see Section 1.1). Choi *et al.* (2009) adopted a maximum likelihood estimator ($\widehat{\Phi}_M$) that estimates the k coefficients using an EM algorithm. The classic Genomic Relationship Matrix (GRM) estimator ($\widehat{\Phi}_C$) estimates kinship through the empirical correlation of genotypes; see, for example Hayes *et al.* (2009). An alternative version of the GRM estimator ($\widehat{\Phi}_R$) is more robust to presence of rare alleles (VanRaden 2008). Day-Williams *et al.* (2011) proposed a method of moment estimator ($\widehat{\Phi}_D$) that estimates kinship by exploring the relationship between identity by state (IBS) and IBD.

Other estimators make use of various sources of additional information to improve accuracy of kinship estimation. A number of methods have been developed to estimate IBD sharing at the locus level; see for example Moltke *et al.* (2011) and methods cited in Brown *et al.* (2012). These location specific IBD estimates in turn provide estimates of kinship; note that location-specific kinships are constrained to the values 0, 1/4, 1/2, and 1(Day-Williams *et al.* 2011). Here we consider estimators from two of such local IBD methods that have not been previously used to estimate genome-wide realized kinship. The local method of Day-Williams *et al.* (2011) with resulting kinship estimator $\widehat{\Phi}_L$ (to be distinguished from $\widehat{\Phi}_D$), requires information on the ordering of markers along the chromosomes. It predicts the amount of sharing at each marker by first estimating a neighborhood kinship for each marker and then applying a constrained smoothing algorithm on each chromosome. The hidden Markov model (HMM) proposed by Brown *et al.* (2012), with resulting kinship estimator $\widehat{\Phi}_H$, estimates probabilities of IBD states between two or more individuals at each marker location. A genetic map is needed to compute transition probabilities along the Markov chain.

Finally, some estimators take linkage disequilibrium (LD) into account. An increase in the density of SNP panels, by itself, will not improve precision without limit in the presence of LD), as additional SNPs are not independent sources of information. Speed *et al.* (2012) developed LDAK ($\widehat{\Phi}_K$), a weighted version of the GRM that takes LD into account. By analogy with methods introduced in the population structure context by Patterson *et al.* (2006) and Zou *et al.* (2010), LDAK equalizes contributions of linked SNPs by down-weighting SNPs which make redundant contribution to the GRM as evidenced by off-diagonal terms in the (squared) SNP correlation matrix.

2.2 A generalized framework for the GRM estimators

We consider a general class of GRM estimators of the following form:

$$\widehat{\Phi}(\boldsymbol{a}, \boldsymbol{w}) = \sum_{l=1}^{L} w_l \cdot \frac{x_l y_l - a_l(x_l + y_l) + 4a_l p_l - 4p_l^2}{4p_l(1-p_l)}, \tag{2.1}$$

where L is the total number of marker loci, $\boldsymbol{x} = (x_1, ..., x_L)^T$ and $\boldsymbol{y} = (y_1, ..., y_L)^T$ are counts of reference alleles at each marker for the pair of individuals, $\boldsymbol{p} = (p_1, ..., p_L)^T$ are the population frequencies of the reference alleles, $\boldsymbol{a} = (a_1, ..., a_L)^T$ are multiplicative factors, and $\boldsymbol{w} = (w_1, ..., w_L)^T$ are non-negative weights satisfying $\sum_{l=1}^{L} w_l = 1$. The vectors $\boldsymbol{a}$ and $\boldsymbol{w}$ are parameters that distinguish different GRM estimators, while $\boldsymbol{x}$, $\boldsymbol{y}$ are the data random variables and $\boldsymbol{p}$ are assumed known (in practice, allele frequencies are typically estimated from the sample). Note that the denominator in equation (2.1) is $4p_l(1-p_l)$ as opposed to $2p_l(1-p_l)$, reflecting the difference between kinship coefficients and the relatedness coefficients of the numerator relationship matrix (Henderson 1976).

The classic GRM estimator is a special case of equation (2.1) with $a_l = 2p_l$ and $w_l = 1/L$ for all l:

$$\begin{aligned}\widehat{\Phi}_C &= \frac{1}{L}\sum_{l=1}^{L} \frac{(x_l - 2p_l)(y_l - 2p_l)}{4p_l(1-p_l)} \\ &= \sum_{l=1}^{L} \frac{1}{L} \cdot \frac{x_l y_l - 2p_l(x_l + y_l) + 4p_l^2}{4p_l(1-p_l)}.\end{aligned} \tag{2.2}$$

The robust GRM estimator is a special case with $a_l = 2p_l$ and $w_l = 4p_l(1-p_l)/(\sum_{m=1}^{L} 4p_m(1-p_m))$ for all l:

$$\begin{aligned}\widehat{\Phi}_R &= \frac{\sum_{l=1}^{L}(x_l - 2p_l)(y_l - 2p_l)}{\sum_{l=1}^{L} 4p_l(1-p_l)} \\ &= \sum_{l=1}^{L} \frac{4p_l(1-p_l)}{\sum_{m=1}^{L} 4p_m(1-p_m)} \cdot \frac{x_l y_l - 2p_l(x_l + y_l) + 4p_l^2}{4p_l(1-p_l)}.\end{aligned} \tag{2.3}$$

The global Day-Williams estimator is defined in Day-Williams *et al.* (2011) as

$$\widehat{\Phi}_D = \frac{e - \sum_{l=1}^{L}\left[p_l^2 + (1-p_l)^2\right]}{L - \sum_{l=1}^{L}\left[p_l^2 + (1-p_l)^2\right]}. \tag{2.4}$$

Here e is a measure of identity by state matches between the two individuals across the genome. Let $\mathbf{1}_{x_{1,l}}$ and $\mathbf{1}_{x_{2,l}}$ be the indicator functions that the two alleles of individual 1 at marker l are the reference allele respectively. Define $\mathbf{1}_{y_{1,l}}$ and $\mathbf{1}_{y_{2,l}}$ similarly for individual 2. e is defined as

$$e = \sum_{l=1}^{L} \frac{4 - |\mathbf{1}_{x_{1,l}} - \mathbf{1}_{y_{1,l}}| - |\mathbf{1}_{x_{1,l}} - \mathbf{1}_{y_{2,l}}| - |\mathbf{1}_{x_{2,l}} - \mathbf{1}_{y_{1,l}}| - |\mathbf{1}_{x_{2,l}} - \mathbf{1}_{y_{2,l}}|}{4}. \tag{2.5}$$

Note that $x_l = \mathbf{1}_{x_{1,l}} + \mathbf{1}_{x_{2,l}}$ and $y_l = \mathbf{1}_{y_{1,l}} + \mathbf{1}_{y_{2,l}}$. It is easy to verify that the numerator of the summand in equation (2.5) is simply $2 + 2(x_l - 1)(y_l - 1)$. Make this substitution in equation (2.4) and we have

$$\begin{aligned}\widehat{\Phi}_D &= \frac{\sum_{l=1}^{L} x_l y_l - (x_l + y_l) + 4p_l - 4p_l^2}{\sum_{l=1}^{L} 4p_l(1-p_l)} \\ &= \sum_{l=1}^{L} \frac{4p_l(1-p_l)}{\sum_{m=1}^{L} 4p_m(1-p_m)} \cdot \frac{x_l y_l - (x_l + y_l) + 4p_l - 4p_l^2}{4p_l(1-p_l)},\end{aligned} \tag{2.6}$$

which fits into the form of (2.1) with $a_l = 1$ and $w_l = 4p_l(1-p_l)/(\sum_{m=1}^{L} 4p_m(1-p_m))$ for all l.

In the general form of equation (2.1), let

$$Z_l(a_l) = \frac{x_l y_l - a_l(x_l + y_l) + 4a_l p_l - 4p_l^2}{4p_l(1-p_l)} \tag{2.7}$$

so that $\widehat{\Phi}(\boldsymbol{a}, \boldsymbol{w}) = \sum_{i=1}^{L} w_l \cdot Z_l(a_l)$. Note that $\boldsymbol{Z}(\boldsymbol{a}) = (Z_1(a_1), ..., Z_L(a_L))^T$ depends on the parameters $\boldsymbol{a}$ but not on $\boldsymbol{w}$. We make two basic assumptions throughout the rest of this chapter. First, we assume IBD genes have the same allelic types and non-IBD genes have independent allelic types. In addition, we assume exchangeability of parental lineage, so that either of the two genes from the first individual is equally likely to be IBD to either of the two genes of the second individual at the same locus.

Under the above assumptions, we have

$$\begin{aligned}\mathbb{E}[x_l] &= 2\mathbb{E}[\mathbf{1}_{x_{1,l}}] = 2p_l,\\ \mathbb{E}[x_l y_l] &= \mathbb{E}[(\mathbf{1}_{x_{1,l}} + \mathbf{1}_{x_{2,l}})(\mathbf{1}_{y_{1,l}} + \mathbf{1}_{y_{2,l}})]\\ &= 4\mathbb{E}[\mathbf{1}_{x_{1,l}}\mathbf{1}_{y_{1,l}}]\\ &= 4\Phi p_l + 4(1-\Phi)p_l^2,\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}[Z_l(a_l)] &= \mathbb{E}\left(\frac{x_l y_l - a_l(x_l + y_l) + 4a_l p_l - 4p_l^2}{4p_l(1-p_l)}\right)\\ &= \frac{4\Phi p_l + 4(1-\Phi)p_l^2 - a_l \cdot 4p_l + 4a_l p_l - 4p_l^2}{4p_l(1-p_l)}\\ &= \Phi.\end{aligned}$$

Since

$$\mathbb{E}[\widehat{\Phi}(\boldsymbol{a}, \boldsymbol{w})] = \mathbb{E}\left[\sum_{l=1}^{L} w_l Z_l(a_l)\right] = \sum_{l=1}^{L} w_l \cdot \Phi = \Phi,$$

$\widehat{\Phi}(\boldsymbol{a}, \boldsymbol{w})$ is unbiased for any $\boldsymbol{a}$ and $\boldsymbol{w}$.

Performance of unbiased estimators depend on their variances. For a GRM estimator in the general form of equation (2.1),

$$\mathrm{Var}[\widehat{\Phi}(\boldsymbol{a}, \boldsymbol{w})] = \sum_{l=1}^{L} w_l^2 V_l(a_l) + \sum_{l \neq m} w_l w_m \mathrm{Cov}[Z_l(a_l), Z_m(a_m)], \tag{2.8}$$

where $V_l(a_l) = \mathrm{Var}[Z_l(a_l)]$. Computation of $\mathrm{Var}[\widehat{\Phi}(\boldsymbol{a}, \boldsymbol{w})]$ is intractable without simplifying assumptions. We next derive the $\boldsymbol{a}$ and $\boldsymbol{w}$ that minimize $\mathrm{Var}[\widehat{\Phi}(\boldsymbol{a}, \boldsymbol{w})]$ under different assumptions.

2.3 Minimizing variance: two-step GRM

We first assume linkage equilibrium. Although in reality there is LD, empirical results show that the relative values of variances of estimators are well approximated by those derived

under this assumption (see Section 2.5.3). When markers are in linkage equilibrium, the allelic types at different markers on the same haplotype are independent. Equation (2.8) then reduces to

$$\mathrm{Var}[\widehat{\Phi}(\boldsymbol{a}, \boldsymbol{w})] = \sum_{l=1}^{L} w_l^2 V_l(a_l). \tag{2.9}$$

To find the GRM estimator with minimal variance, equation (2.9) suggests that one should first (for each l) choose a_l that minimizes $V_l(a_l)$, and then choose $\boldsymbol{w}$ to minimize $\mathrm{Var}[\widehat{\Phi}(\boldsymbol{a}, \boldsymbol{w})]$.

We now make an additional assumption of no inbreeding. To derive an expression for $V_l(a_l)$ defined in (2.8), note that

$$\begin{aligned}
\mathbb{E}[x_l^2] &= \mathbb{E}[\mathbf{1}_{x_{1,l}}^2 + 2\mathbf{1}_{x_{1,l}}\mathbf{1}_{x_{2,l}} + \mathbf{1}_{x_{2,l}}^2] \\
&= 2p_l + 2p_l^2, \\
\mathrm{Var}[x_l] &= \mathbb{E}[x_l^2] - (\mathbb{E}[x_l])^2 \\
&= 2p_l + 2p_l^2 - 4p_l^2 \\
&= 2p_l(1 - p_l), \\
\mathrm{Cov}[x_l, y_l] &= \mathbb{E}[(x_l - 2p_l)(y_l - 2p_l)] \\
&= \mathbb{E}[x_l y_l - 2p_l x_l - 2p_l y_l + 4p_l^2] \\
&= 4\Phi p_l(1 - p_l), \\
\mathbb{E}[x_l^2 y_l] &= \mathbb{E}[(\mathbf{1}_{x_1} + \mathbf{1}_{x_2})^2(\mathbf{1}_{y_1} + \mathbf{1}_{y_2})] \\
&= 4\mathbb{E}[\mathbf{1}_{x_1}^2\mathbf{1}_{y_1}] + 4\mathbb{E}[\mathbf{1}_{x_1}\mathbf{1}_{x_2}\mathbf{1}_{y_1}] \\
&= 4\Phi p_l + 4(1 - \Phi)p_l^2 + 4[2\Phi p_l^2 + (1 - 2\Phi)p_l^3] \\
&= 4\Phi p_l + 4p_l^2 + 4\Phi p_l^2 + 4p_l^3 - 8\Phi p_l^3,
\end{aligned}$$

IBD State	(1,2,1)	(1,2,2)	(1,2,3)
Probability	Φ	Φ	$1-2\Phi$

Table 2.1: IBD states and probabilities for the two genes of individual 1 and one random gene from individual 2.

$$\begin{aligned}
\mathrm{Cov}[x_l y_l, x_l] &= \mathbb{E}[(x_l y_l - 4\Phi p_l - 4(1-\Phi)p_l^2)(x_l - 2p_l)] \\
&= \mathbb{E}[x_l^2 y_l] - 2p_l \mathbb{E}[x_l y_l] \\
&= 4\Phi p_l + 4p_l^2 + 4\Phi p_l^2 + 4p_l^3 - 8\Phi p_l^3 - 8\Phi p_l^2 - 8p_l^3 + 8\Phi p_l^3 \\
&= 4\Phi p_l + 4p_l^2 - 4\Phi p_l^2 - 4p_l^3 \\
&= 4(\Phi + p_l)p_l(1-p_l).
\end{aligned}$$

Calculation of the term $\mathbb{E}[\mathbf{1}_{x_1}\mathbf{1}_{x_2}\mathbf{1}_{y_1}]$ involves probabilities of the underlying IBD states between the two genes of individual 1 and one gene of individual 2. Given the assumption of no inbreeding, there are only three possible IBD states with probabilities given in Table 2.1. It follows that

$$\begin{aligned}
\mathbb{E}[x_l^2 y_l^2] &= \mathbb{E}[(\mathbf{1}_{x_1} + \mathbf{1}_{x_2})^2(\mathbf{1}_{y_1} + \mathbf{1}_{y_2})^2] \\
&= 4\mathbb{E}[\mathbf{1}_{x_1}^2\mathbf{1}_{y_1}^2] + 8\mathbb{E}[\mathbf{1}_{x_1}\mathbf{1}_{x_2}\mathbf{1}_{y_1}^2] + 4\mathbb{E}[\mathbf{1}_{x_1}\mathbf{1}_{x_2}\mathbf{1}_{y_1}\mathbf{1}_{y_2}] \\
&= 4\Phi p_l + 4(1-\Phi)p_l^2 + 8[2\Phi p_l^2 + (1-2\Phi)p_l^3] + 4[k_2 p_l^2 + k_1 p_l^3 + k_0 p_l^4] \\
&= 4\Phi p_l + 4p_l^2 + 12\Phi p_l^2 + 8p_l^3 - 16\Phi p_l^3 + 4[k_2 p_l^2 + k_1 p_l^3 + k_0 p_l^4], \\
\mathrm{Var}[x_l y_l - a_l(x_l + y_l)] &= \mathrm{Var}[x_l y_l] + a_l^2 \mathrm{Var}[x_l + y_l] - 2a_l \mathrm{Cov}[x_l y_l, x_l + y_l] \\
&= \mathbb{E}[x_l^2 y_l^2] - (\mathbb{E}[x_l y_l])^2 + 2a_l^2 \mathrm{Var}[x_l] + 2a_l^2 \mathrm{Cov}[x_l, y_l] - 4a_l \mathrm{Cov}[x_l y_l, x_l] \\
&= 4\Phi p_l + 4p_l^2 + 12\Phi p_l^2 + 8p_l^3 - 16\Phi^2 p_l^2 - 48\Phi p_l^3 - 16p_l^4 + 32\Phi^2 p_l^3 + 32\Phi p_l^4 - 16\Phi^2 p_l^4 \\
&\quad + 4[k_2 p_l^2 + k_1 p_l^3 + k_0 p_l^4] + [4a_l^2 + 8a_l^2\Phi - 16a_l(\Phi + p_l)]p_l(1-p_l)
\end{aligned}$$

$$
\begin{aligned}
&= 4\Phi p_l + 4p_l^2 + 12\Phi p_l^2 + 8p_l^3 - 16\Phi^2 p_l^2 - 48\Phi p_l^3 - 16p_l^4 + 32\Phi^2 p_l^3 + 32\Phi p_l^4 - 16\Phi^2 p_l^4 \\
&\quad + 4[k_2 p_l^2 + (4\Phi - 2k_2)p_l^3 + (1 - k_2 - 4\Phi + 2k_2)p_l^4] + [4a_l^2 + 8a_l^2\Phi - 16a_l(\Phi + p_l)]p_l(1 - p_l) \\
&= 4\Phi p_l + 4p_l^2 + 12\Phi p_l^2 + 8p_l^3 - 16\Phi^2 p_l^2 - 48\Phi p_l^3 - 16p_l^4 + 32\Phi^2 p_l^3 + 32\Phi p_l^4 - 16\Phi^2 p_l^4 \\
&\quad + 4[k_2 p_l^2(1 - p_l)^2 + 4\Phi p_l^3 + p_l^4 - 4\Phi p_l^4] + [4a_l^2 + 8a_l^2\Phi - 16a_l(\Phi + p_l)]p_l(1 - p_l) \\
&= 16\Phi(1 - \Phi)p_l^2(1 - p_l)^2 + 4p_l^2(1 + 3p_l)(1 - p_l) + 4\Phi p_l(1 - p_l) + 4k_2 p_l^2(1 - p_l)^2 \\
&\quad + [4a_l^2 + 8a_l^2\Phi - 16a_l(\Phi + p_l)]p_l(1 - p_l), \\
V_l(a_l) &= \mathrm{Var}\left[\frac{x_l y_l - a_l(x_l + y_l) + 4a_l p_l - 4p_l^2}{4p_l(1 - p_l)}\right] \\
&= \frac{\mathrm{Var}[x_l y_l - a_l(x_l + y_l)]}{16p_l^2(1 - p_l)^2} \\
&= \frac{4\Phi(1 - \Phi)p_l(1 - p_l) + (k_2 + 1)p_l(1 - p_l) + (a_l - 2p_l)^2 + 2\Phi(a_l - 1)^2 - \Phi}{4p_l(1 - p_l)}, \qquad (2.10)
\end{aligned}
$$

where k_2 is the realized proportion of the genome that the pair of individuals share both genes IBD. Note that the value of $V_l(a_l)$ is asymmetric about $p_l = 0.5$ for general choices of a_l, and thus is sensitive to the choice of reference allele. The part of $V_l(a_l)$ that is responsible for this asymmetry is

$$(a_l - 2p_l)^2 + 2\Phi(a_l - 1)^2,$$

where a_l can be a function of p_l. Suppose however that a_l is a weighted average of $2p_l$ and 1: $a_l = b \cdot 2p_l + (1 - b)$ for $b \in [0, 1]$. Then

$$
\begin{aligned}
(a_l - 2p_l)^2 + 2\Phi(a_l - 1)^2 &= (b \cdot 2p_l + 1 - b - 2p_l)^2 + 2\Phi(b \cdot 2p_l + 1 - b - 1)^2 \\
&= (2p_l - 1)^2[(b - 1)^2 + 2\Phi b^2],
\end{aligned}
$$

and $V_l(a_l)$ takes the value

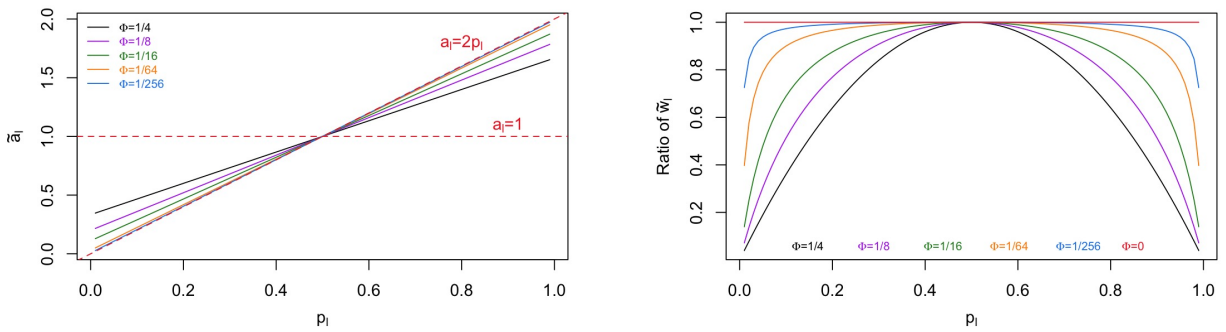
$$4\Phi(1 - \Phi) + k_2 + 1 + \frac{(2p_l - 1)^2[(b - 1)^2 + 2\Phi b^2] - \Phi}{4p_l(1 - p_l)},$$

which is symmetric about $p_l = 0.5$, and it attains its minimum at $p_l = 0.5$ conditional on Φ and k_2.

Returning to the general form of $V_l(a_l)$ and setting its derivative with respect to a_l equal to 0, leads to

$$\tilde{a}_l = \arg\min_{a_l} V_l(a_l) = \frac{1}{1+2\Phi} \cdot 2p_l + \frac{2\Phi}{1+2\Phi}. \tag{2.11}$$

Thus the optimal $\tilde{a}_l$ is a weighted average of $2p_l$ and 1, and $V_l(\tilde{a}_l)$ is invariant to the choice of reference allele. Interestingly, $2p_l$, $2p_l$ and 1 are the choices of a_l used by $\widehat{\Phi}_C$, $\widehat{\Phi}_R$ and $\widehat{\Phi}_D$ respectively (compare equation (2.2), (2.3) and (2.6)). Figure 2.1a shows how $\tilde{a}_l$ varies with p_l for different Φ's. We see that $a_l = 2p_l$ is optimal when $\Phi = 0$, whereas $a_l = 1$ is far from optimal even when $\Phi = \frac{1}{4}$. Since $\widehat{\Phi}_R$ and $\widehat{\Phi}_D$ use the same weights, $\widehat{\Phi}_R$ is more efficient than $\widehat{\Phi}_D$ for small Φ under the assumptions of this section.



(a) (b)

Figure 2.1: **(a)** $\tilde{a}_l$ as a function of p_l for different Φ's, with $a_l = 2p_l$ and $a_l = 1$ as reference lines. **(b)** Ratio of weights between a marker with allele frequency p_l and a reference marker with allele frequency 0.5, under the optimal weighting scheme $\tilde{\boldsymbol{w}}(2\boldsymbol{p})$ for different combinations of (Φ, k_2). In calculation of $\tilde{\boldsymbol{w}}(2\boldsymbol{p})$, we used $k_2 = \frac{1}{4}$ when $\Phi = \frac{1}{4}$ and 0 otherwise. Note that these combinations of (Φ, k_2) correspond to pedigree expectations of a pair of individuals that are full siblings, half siblings, 1st cousins, 2nd cousins or 3rd cousins.

For fixed $\boldsymbol{a}$ (and therefore fixed $\boldsymbol{V}(\boldsymbol{a})$), we can find the set of optimal weights $\tilde{\boldsymbol{w}}(\boldsymbol{a})$ by

solving the following minimization problem,

$$\min_{\boldsymbol{w}} \sum_{l=1}^{L} w_l^2 V_l(a_l) \qquad : \qquad \boldsymbol{w}^T \mathbf{1} = 1,\ w_l \geq 0\ \forall\ l.$$

With Lagrange multipliers λ,

$$\begin{aligned}
\mathcal{L}(\boldsymbol{w}, \lambda) &= \sum_{l=1}^{L} w_l^2 V_l(a_l) - \lambda(\sum_{l=1}^{L} w_l - 1) \\
\frac{\partial \mathcal{L}}{\partial w_l} &= 2 w_l V_l(a_l) - \lambda = 0 \\
\tilde{w}_l(\boldsymbol{a}) &= \frac{\lambda}{2} \cdot V_l(a_l)^{-1}.
\end{aligned}$$

The above expression holds true for all l, implying

$$\tilde{w}_l(\boldsymbol{a}) = \frac{V_l(a_l)^{-1}}{\sum_{m=1}^{L} V_m(a_m)^{-1}}, \qquad l = 1, ..., L. \tag{2.12}$$

The non-negativity constraint is automatically satisfied as $V_l(a_l) > 0$ for all l. The weights in equation (2.12) can be equivalently specified by the ratios

$$\frac{\tilde{w}_l(\boldsymbol{a})}{\tilde{w}_m(\boldsymbol{a})} = \frac{V_m(a_m)}{V_l(a_l)}, \qquad l, m = 1, ..., L,$$

which are functions of allele frequencies at the corresponding pairs of markers, conditional on Φ, k_2 and the choice of $\boldsymbol{a}$. For $\boldsymbol{a} = 2\boldsymbol{p}$, Figure 2.1b shows how a marker with frequency p_l is weighted relative to a marker with frequency 0.5 under the optimal weighting scheme $\tilde{\boldsymbol{w}}(2\boldsymbol{p})$ for different combinations of (Φ, k_2). The optimal solution weights markers very differently, especially for large Φ. In this case ($\boldsymbol{a} = 2\boldsymbol{p}$), the uniform weighting of $\widehat{\Phi}_C$ is optimal when $\Phi = 0$, whereas the weighting of $\widehat{\Phi}_R$ is optimal when $\Phi = \frac{1}{4}$ and $k_2 = \frac{1}{4}$.

The estimator $\widehat{\Phi}(\tilde{\boldsymbol{a}}, \tilde{\boldsymbol{w}}(\tilde{\boldsymbol{a}}))$ would be an obvious choice if we knew $\tilde{\boldsymbol{a}}$ and $\tilde{\boldsymbol{w}}(\tilde{\boldsymbol{a}})$. However, elements of $\tilde{\boldsymbol{a}} = (\tilde{a}_1, ..., \tilde{a}_L)^T$ are functions of the unknown Φ, and elements of $\tilde{\boldsymbol{w}}(\tilde{\boldsymbol{a}}) = (\tilde{w}_1(\tilde{\boldsymbol{a}}), ..., \tilde{w}_L(\tilde{\boldsymbol{a}}))^T$ are functions of the unknown Φ, k_2 and $\tilde{\boldsymbol{a}}$. The closed form expressions given in equation (2.10), (2.11) and (2.12) motivate a two-step estimator, $\widehat{\Phi}_T$, that approximates $\widehat{\Phi}(\tilde{\boldsymbol{a}}, \tilde{\boldsymbol{w}}(\tilde{\boldsymbol{a}}))$ following these two steps:

1. Obtain initial estimates of Φ and k_2 using an existing method;

2. Compute $\tilde{\boldsymbol{a}}^*$ (equation (2.11)) and then $\tilde{\boldsymbol{w}}^*(\tilde{\boldsymbol{a}}^*)$ (equation (2.10), (2.12)) using these estimates of Φ and k_2.

Then $\widehat{\Phi}_T = \widehat{\Phi}(\tilde{\boldsymbol{a}}^*, \tilde{\boldsymbol{w}}^*(\tilde{\boldsymbol{a}}^*))$.

In practice, existing kinship estimators such as $\widehat{\Phi}_C$, $\widehat{\Phi}_R$ and $\widehat{\Phi}_D$ may produce negative estimates of Φ in Step 1. Our implementation uses $\widehat{\Phi}_C$ to obtain an initial estimate of Φ. When this initial estimate is negative, we simply retain it as the $\widehat{\Phi}_T$ estimate. For simplicity, we set $k_2 = 0$ when computing $V_l(\tilde{a}_l^*)$ for all l. In principle, this affects the calculation of $\tilde{\boldsymbol{w}}^*(\tilde{\boldsymbol{a}}^*)$ for bilateral relatives, but it makes no practical difference (see Section 2.5.3).

2.4 Taking LD into account: LD weighted GRM

We now drop the assumptions of linkage equilibrium and absence of inbreeding. To calculate variances, we follow the approach of Sverdlov (2014), and consider the case where the pair of individuals are unrelated, so that $x_l \perp y_m$ for all l and m. Under these assumptions, all of the results from Section 2.2 still hold, but most of the results from Section 2.3 do not. As in Section 2.2, write $x_l = \mathbf{1}_{x_{1,l}} + \mathbf{1}_{x_{2,l}}$. Let the inbreeding coefficients of the two individuals be F_x and F_y respectively. Then

$$\begin{aligned}
\mathbb{E}[x_l^2] &= \mathbb{E}[\mathbf{1}_{x_1}^2 + 2\mathbf{1}_{x_1}\mathbf{1}_{x_2} + \mathbf{1}_{x_2}^2] \\
&= p_l + 2[F_x \cdot p_l + (1 - F_x) \cdot p_l^2] + p_l \\
&= 2p_l[1 + F_x + (1 - F_x)p_l],
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}[x_l] &= \mathbb{E}[x_l^2] - (\mathbb{E}[x_l])^2 \\
&= 2p_l[1 + F_x + (1 - F_x)p_l] - 4p_l^2 \\
&= 2p_l(1 - p_l)(F_x + 1).
\end{aligned}$$

Analogous results hold for y_l. It can be verified (expression of $\text{Var}[x_l y_l - a_l(x_l + y_l)]$ in Section 2.3) that

$$\tilde{a}_l = \arg\min_{a_l} V_l(a_l) = \frac{\text{Cov}[x_l y_l, x_l + y_l]}{\text{Var}[x_l + y_l]}.$$

This time

$$\begin{aligned}\tilde{a}_l &= \frac{\text{Cov}[x_l y_l, x_l] + \text{Cov}[x_l y_l, y_l]}{\text{Var}[x_l] + \text{Var}[y_l]} \\ &= \frac{(\mathbb{E}[x_l^2] - 2\mathbb{E}[x_l]\mathbb{E}[y_l] + \mathbb{E}[y_l^2]) \cdot 2p_l}{\mathbb{E}[x_l^2] - (\mathbb{E}[x_l])^2 + \mathbb{E}[y_l^2] - (\mathbb{E}[y_l])^2}, \\ &= 2p_l,\end{aligned}$$

since

$$2\mathbb{E}[x_l]\mathbb{E}[y_l] = (\mathbb{E}[x_l])^2 + (\mathbb{E}[y_l])^2 = 8p_l^2.$$

Now

$$\begin{aligned}&\text{Cov}[x_l y_l - a_l(x_l + y_l), x_m y_m - a_m(x_m + y_m)] \\ =\ &\text{Cov}[x_l y_l, x_m y_m] - a_m \text{Cov}[x_l y_l, x_m + y_m] - a_l \text{Cov}[x_m y_m, x_l + y_l] \\ &+ a_l a_m \text{Cov}[x_l + y_l, x_m + y_m],\end{aligned}$$

where

$$\begin{aligned}\text{Cov}[x_l y_l, x_m + y_m] &= \text{Cov}[x_l y_l, x_m] + \text{Cov}[x_l y_l, y_m] \\ &= \mathbb{E}[x_l x_m]\mathbb{E}[y_l] - \mathbb{E}[x_l]\mathbb{E}[x_m]\mathbb{E}[y_l] \\ &\quad + \mathbb{E}[y_l y_m]\mathbb{E}[x_l] - \mathbb{E}[y_l]\mathbb{E}[y_m]\mathbb{E}[x_l] \\ &= 2p_l \left(\mathbb{E}[x_l x_m] + \mathbb{E}[y_l y_m] - 8p_l p_m\right), \\ \text{Cov}[x_m y_m, x_l + y_l] &= 2p_m \left(\mathbb{E}[x_l x_m] + \mathbb{E}[y_l y_m] - 8p_l p_m\right), \\ \text{Cov}[x_l + y_l, x_m + y_m] &= \text{Cov}[x_l, x_m] + \text{Cov}[y_l, y_m] \\ &= \mathbb{E}[x_l x_m] + \mathbb{E}[y_l y_m] - 8p_l p_m.\end{aligned}$$

Unfortunately, there generally does not exist an $\boldsymbol{a}$ that jointly minimizes each of the unweighted covariance terms in equation (2.8). Thus, we conveniently set $\boldsymbol{a} = 2\boldsymbol{p}$ in the

remainder of this section. Setting $\boldsymbol{a} = 2\boldsymbol{p}$,

$$\begin{aligned}
&\mathrm{Cov}[x_l y_l - 2p_l(x_l + y_l), x_m y_m - 2p_m(x_m + y_m)] \\
&= \mathrm{Cov}[x_l y_l, x_m y_m] - 4p_l p_m(\mathbb{E}[x_l x_m] + \mathbb{E}[y_l y_m] - 8p_l p_m) \\
&= \mathbb{E}[x_l x_m]\mathbb{E}[y_l y_m] - 4p_l p_m(\mathbb{E}[x_l x_m] + \mathbb{E}[y_l y_m]) + 16p_l^2 p_m^2 \\
&= (\mathbb{E}[x_l x_m] - 4p_l p_m)(\mathbb{E}[y_l y_m] - 4p_l p_m) \\
&= \rho_{lm}\sqrt{\mathrm{Var}[x_l] \cdot \mathrm{Var}[x_m]} \cdot \rho_{lm}\sqrt{\mathrm{Var}[y_l] \cdot \mathrm{Var}[y_m]} \\
&= 4p_l(1-p_l)p_m(1-p_m)(F_x+1)(F_y+1)\rho_{lm}^2.
\end{aligned}$$

It then follows

$$\begin{aligned}
&\mathrm{Cov}[Z_l(2p_l), Z_m(2p_m)] \\
&\qquad = \frac{\mathrm{Cov}[x_l y_l - 2p_l(x_l + y_l), x_m y_m - 2p_m(x_m + y_m)]}{4p_l(1-p_l) \cdot 4p_m(1-p_m)} \\
&\qquad = \frac{1}{4}(F_x+1)(F_y+1)\rho_{lm}^2, \qquad (2.13) \\
&V_l(2p_l) = \mathrm{Var}[Z_l(2p_l)] = \frac{1}{4}(F_x+1)(F_y+1), \qquad (2.14)
\end{aligned}$$

where ρ_{lm} is the genotype dosage correlation between locus l and locus m. Conveniently, ρ_{lm} only enters the expression in a squared form, so the covariance is invariant to the choice of reference allele. Combining (2.14) and (2.13), equation (2.8) becomes

$$\mathrm{Var}[\widehat{\Phi}(2\boldsymbol{p}, \boldsymbol{w})] = \sum_l \sum_m w_l w_m \cdot \frac{1}{4}(F_x+1)(F_y+1)\rho_{lm}^2, \qquad (2.15)$$

where $\rho_{lm} = 1$ when $l = m$. We assume that the matrix of squared LD correlations, $\mathbf{R} = [\rho_{lm}^2]$, is known. The goal is to find the $\boldsymbol{w}$ that minimizes $\mathrm{Var}[\widehat{\Phi}(2\boldsymbol{p}, \boldsymbol{w})]$. Note that the inbreeding coefficients F_x and F_y are part of a fixed scaling factor. The optimization problem reduces to

$$\min_{\boldsymbol{w}} [\boldsymbol{w}^T \mathbf{R} \boldsymbol{w} - \boldsymbol{w}^T \mathbf{1}] \qquad : \qquad w_l \geq 0 \; \forall \; l. \qquad (2.16)$$

Presence of the second term in the objective function forces a solution that satisfies $\boldsymbol{w}^T \mathbf{1} = c$ for some $c > 0$, which can be rescaled by $1/c$ to obtain the final solution $\tilde{\boldsymbol{w}}$. The LD weighted

GRM estimator is then $\widehat{\Phi}_W = \widehat{\Phi}(2\boldsymbol{p}, \tilde{\boldsymbol{w}})$.

Correlation matrices are positive semi-definite. By the Schur product theorem, $\mathbf{R}$, the Hadamard product of a correlation matrix and itself must also be positive semi-definite. The above minimization problem can be solved using standard quadratic programming procedures. In practice, it will be necessary to divide the large set of genome-wide SNPs into blocks. In the case where $\mathbf{R}$ has a block-diagonal structure (e.g. each chromosome forms a block), we can rewrite the minimization problem in terms of subvectors and submatrices,

$$\min_{\boldsymbol{w}} \sum_{i=1}^{n} [\boldsymbol{w}_{(i)}^T \mathbf{R}_{(i)} \boldsymbol{w}_{(i)} - \boldsymbol{w}_{(i)}^T \mathbf{1}] \qquad : \qquad w_l \geq 0 \; \forall \; l. \tag{2.17}$$

Since the $\boldsymbol{w}_{(i)}$'s form a partition of $\boldsymbol{w}$ and the $\mathbf{R}_{(i)}$'s are independent submatrices of $\mathbf{R}$, minimization can be implemented for each block independently. We can solve for $\boldsymbol{w}_{(i)}$ in block i using the corresponding submatrix $\mathbf{R}_{(i)}$. If $\boldsymbol{w}_{(i)}^T \mathbf{1} = c_i$ for the block solutions, the final solution $\tilde{\boldsymbol{w}}$ will be a concatenation of the block solutions $\boldsymbol{w}_{(i)}$'s rescaled by $1/\sum_i c_i$.

2.5 Simulation study

2.5.1 Setup

In the simulation study, we considered six relationship types: full siblings (FS), half siblings (HS), first cousins (C1), second cousins (C2), third cousins (C3) and inbred cousins (IN) from the complex JV pedigree shown in Figure 1.2. Dense SNP genotypes were generated for 1000 independent pairs of each relationship type as follow:

1. Simulate recombination breakpoints and Mendelian sampling for all meioses in the smallest complete pedigree that contains the pair of relatives;

2. Assign founder genome labels (FGLs) (Sobel and Lange 1996) to founder haplotypes, and determine the inherited FGLs at all marker positions for all non-founders with respect to the inheritance pattern simulated in Step 1;

3. Sample founder haplotypes from a reference pool, and assign alleles to non-founders with respect to both the sampled founder haplotypes and the inherited FGLs determined in Step 2.

4. At each locus, combine the two alleles inherited by the related pair of individuals to create genotype data.

Data generation was implemented using the `ibd_create` program of MORGAN v.3.3.1 (MORGAN Tutorial 2016). The K170 marker panel described in Section 1.4.1 was used in the main analysis. All 5008 phased haplotypes from the 1000 Genomes (The 1000 Genomes Project Consortium 2015) combined population were made available for sampling of founder haplotypes. This use of real haplotypes preserves the natural patterns of LD in the combined population. The genetic positions of markers were obtained from the Rutgers Map v.3a (Matise *et al.* 2007). The distribution of allele frequencies in the marker panel reflects real studies using common SNPs. The marker density was chosen to be dense enough to show patterns of LD, yet sparse enough to be attainable by older SNP arrays in human genetics and by modern SNP arrays in animal genetics.

All the kinship estimators were evaluated on the simulated data. We used PLINK v1.07 to implement the PLINK estimator, `ibd_haplo` program (Brown *et al.* 2012) of MORGAN v3.3.1 to implement the HMM estimator, and our own code to implement all other estimators. Marker allele frequencies used in analysis were estimated by PLINK v1.07 for each relationship type separately. For the MLE estimator, we adopted a very stringent convergence criterion (order of 10^{-8} of the final log-likelihood) and used pedigree k coefficients as the starting configurations of the EM algorithm. The GRM estimators were computed both unconstrained to the range $[0, 1]$ (consistent with the theory in Section 2.2 through 2.4), and also constrained to this range.

The local Day-Williams method imputes local kinship directly, whereas the HMM method es-

timates probabilities of local IBD states. For ease of comparison, we used the most probable state from the HMM output to impute local kinship. For these two local methods, estimates of genome-wide realized kinships were calculated as average of imputed local kinship across all marker loci. We simulated genotype data on an additional 300 independent pairs of cousins (100 first cousins, 100 second cousins and 100 third cousins) for tuning purposes. For each of the two local IBD methods, a sparse grid search was implemented to find the set of tuning parameters that maximizes local kinship imputation rate over the tuning data set. The selected sets of tuning parameters were subsequently used in the actual analysis.

All LD-adjusted methods used the reference genotypes of all 2504 individuals from the 1000 Genomes Project Phase 3 data to obtain LD information. A naive LD-pruned GRM estimator ($\widehat{\Phi}_N$) is included as a baseline for comparison. For this estimator, LD pruning was done using PLINK v1.07, where we sequentially threw out one of each pair of markers that had genotypic dosage correlation $\rho^2 > 0.2$ (option --r2 in PLINK). This pruning step reduced the marker set to about half of the original size, and the classic GRM was applied to the reduced marker set to obtain $\widehat{\Phi}_N$. For the LD weighted GRM estimator ($\widehat{\Phi}_W$), weights were computed in blocks of 2000 SNPs using the optimization package JuMP v0.14.1 (Lubin and Dunning 2015) in the Julia language. Weights of the LDAK estimator ($\widehat{\Phi}_K$) were computed using LDAK v4.9 with the default options.

Using a single processor on a standard desktop, the PLINK estimates or any of the GRM estimates can be computed for 1000 pairs of individuals with the selected genome-wide marker panel in a couple of minutes. The MLE estimates can take many hours to compute depending on run conditions (see Section 2.5.2). Either of the two local IBD estimates takes several hours to compute, but this is still computationally feasible. With reference genotypes from 2504 individuals as the basis for LD adjustment, LD pruning takes several minutes to implement. Weights for the LD weighted GRM take less than 15 minutes to compute, whereas weights for LDAK take more than 15 hours to compute.

The performances of estimators were compared on the simulated data. In order to compare performance both across estimators and across relationship types, we summarize estimation accuracy by the ratio of the root mean squared error (RMSE) and the average realized kinship,

$$\frac{\sqrt{\frac{1}{1000}\sum_{m=1}^{1000}(\widehat{\Phi}_m - \Phi_m)^2}}{\frac{1}{1000}\sum_{m=1}^{1000}\Phi_m}, \tag{2.18}$$

where Φ_m's are realized kinship and $\widehat{\Phi}_m$'s are their estimates.

2.5.2 Main results

Table 2.2 summarizes the performance of different kinship estimators. The estimators are divided into two groups: Group-1 uses only marker allele frequencies, whereas those in Group-2 use additional information. First, every estimator works better on closer relatives. This is expected given we are measuring estimation accuracies relative to the average amount of sharing (equation (2.18)), and the coefficients of variation are higher for remote relatives. For each estimator, the raw MSE are in fact smaller on remote relatives (Table 2.3). Second, estimators that make use of additional sources information (Group-2) generally do better than estimators that do not (Group-1). This is also expected as chromosomes are inherited as segments. Information such as marker order, genetic positions and LD pattern are informative of the joint inheritance across markers.

Relative performances of the GRM estimators in Group-1 match our expectations under the assumption of linkage equilibrium (Figure 2.1a and 2.1b). The two-step GRM estimator ($\widehat{\Phi}_T$) compares favorably to others on all relationship types. The classic GRM estimator ($\widehat{\Phi}_C$) works well when $\Phi \approx 0$ (e.g., third cousins). The robust GRM estimator ($\widehat{\Phi}_R$) is preferred to ($\widehat{\Phi}_C$) on close relatives, and it dominates the global Day-Williams estimator ($\widehat{\Phi}_D$) on all relationship types.

Additional information	Estimator		Relationship					
			FS	HS	C1	C2	C3	IN
None	$\widehat{\Phi}_P$	PLINK	1.64	4.66	10.81	50.76	178.78	30.03
	$\widehat{\Phi}_M$	MLE	1.51	3.55	7.26	27.00	78.66	11.81
	$\widehat{\Phi}_D$	global Day-Williams	3.12	6.38	12.34	49.55	187.68	6.31
	$\widehat{\Phi}_C$	classic GRM	3.00	4.85	8.32	30.83	116.22	4.81
	$\widehat{\Phi}_R$	robust GRM	2.42	4.33	7.99	31.64	120.56	4.31
	$\widehat{\Phi}_T$	two-step GRM	2.19	4.26	7.96	30.84	116.24	4.17
LD pattern	$\widehat{\Phi}_N$	naive LD pruned GRM	3.01	4.70	7.87	28.20	105.46	4.68
LD pattern	$\widehat{\Phi}_K$	LDAK	2.64	4.10	7.21	23.74	85.57	5.03
LD pattern	$\widehat{\Phi}_W$	LD weighted GRM	1.59	2.60	4.44	17.33	65.60	3.09
marker order	$\widehat{\Phi}_L$	local Day-Williams	2.04	2.76	6.52	10.79	24.26	6.84
LD pattern + marker order	$\widehat{\Phi}_{L_W}$	LD weighted local Day-Williams	2.10	2.85	6.70	11.27	23.56	7.00
genetic position	$\widehat{\Phi}_H$	HMM	1.77	3.36	6.04	17.90	64.97	3.82
LD pattern + genetic position	$\widehat{\Phi}_{H_W}$	LD weighted HMM	1.57	2.83	5.16	14.55	52.19	3.33
pedigree	Ψ	pedigree kinship	7.80	10.49	17.63	37.30	70.16	13.24

Table 2.2: Estimation accuracies from simulation study as measured by the ratio ($\times 10^2$) of the RMSE and the average realized kinship (equation (2.18)). All estimators compared require dense SNP genotypes and reliable sources of marker allele frequencies as input. Pedigree kinship for the six relationship types are: 0.25 (FS), 0.125 (HS), 0.0625 (C1), 0.0156 (C2), 0.0039 (C3) and 0.1094 (IN) respectively.

	Relationship					
Estimator	FS	HS	C1	C2	C3	IN
$\widehat{\Phi}_P$	4.09	5.83	6.71	7.84	6.88	32.70
$\widehat{\Phi}_M$	3.76	4.43	4.51	4.17	3.03	12.87
$\widehat{\Phi}_D$	7.78	7.98	7.66	7.65	7.22	6.87
$\widehat{\Phi}_C$	7.49	6.07	5.17	4.76	4.47	5.23
$\widehat{\Phi}_R$	6.05	5.42	4.96	4.88	4.64	4.70
$\widehat{\Phi}_T$	5.46	5.32	4.94	4.76	4.47	4.54
$\widehat{\Phi}_N$	7.52	5.88	4.89	4.35	4.06	5.10
$\widehat{\Phi}_K$	6.58	5.13	4.47	3.66	3.29	5.48
$\widehat{\Phi}_W$	3.97	3.25	2.76	2.67	2.53	3.36
$\widehat{\Phi}_L$	5.10	3.45	4.05	1.67	0.93	7.44
$\widehat{\Phi}_{L_W}$	5.26	3.56	4.16	1.74	0.91	7.62
$\widehat{\Phi}_H$	4.41	4.21	3.75	2.76	2.50	4.17
$\widehat{\Phi}_{H_W}$	3.91	3.54	3.21	2.25	2.01	3.63
Ψ	19.47	13.12	10.95	5.76	2.70	14.42

Table 2.3: RMSE ($\times 10^3$). The GRM estimates are unconstrained.

The MLE estimator ($\widehat{\Phi}_M$) stands out among estimators in Group-1. Likelihood estimators are known to be often more accurate than method of moment estimators (Milligan 2003; Anderson and Weir 2007). However, accuracy and computational efficiency of $\widehat{\Phi}_M$ are extremely sensitive to the starting configurations and the convergence criterion of the EM algorithm. Our implementation of $\widehat{\Phi}_M$ adopted very favorable conditions (see Section 2.5.1). Otherwise, the results were much less accurate and computation time much longer (results not shown). For full siblings, the PLINK estimator ($\widehat{\Phi}_P$) is more accurate than any Group-1 GRM esti-

mator. Since $\widehat{\Phi}_P$ estimates the k coefficients directly, this provides higher resolution on full siblings who share two genes IBD over, on average, 25% of the genome. When inbreeding is present, $\widehat{\Phi}_P$ and $\widehat{\Phi}_M$ perform poorly relative to the other Group-1 estimators. These two estimators estimate the zero-inbreeding k coefficients directly and are thus more sensitive to violation of the no-inbreeding assumption.

Among the estimators of Group-2, the local Day-Williams estimator ($\widehat{\Phi}_L$) performs better than the others on remote relatives. However, the smoothing algorithm of $\widehat{\Phi}_L$ tends to produce downward bias (discussed in greater detail in Section 3.2), so that $\widehat{\Phi}_L$ must perform well on remote relatives where there is not much room for underestimation. In contrast, the HMM estimator ($\widehat{\Phi}_H$) generally overestimates IBD in the presence of LD (Brown *et al.* 2012). The naive LD-pruned GRM estimator ($\widehat{\Phi}_N$) shows slight improvement over the classic GRM estimator ($\widehat{\Phi}_C$), but loses to the LDAK ($\widehat{\Phi}_K$) on all occasions. The LD weighted GRM estimator ($\widehat{\Phi}_W$) dominates both $\widehat{\Phi}_H$ and $\widehat{\Phi}_K$ in performance and loses to $\widehat{\Phi}_L$ only on remote relatives. This reflects the amount of LD present in the selected marker panel in the combined population, and shows that appropriately adjusting for patterns of LD can significantly improve the accuracy of kinship estimation.

When inbreeding is present, performance of $\widehat{\Phi}_L$ is the most affected among the Group-2 estimators. Glazner and Thompson (2015) noted that in their example this local-IBD Day-Williams method failed to pick up short segments of complex (autozygous) IBD; the varying kinship levels across short distances seem to challenge this method (see Section 3.2). The performance of $\widehat{\Phi}_W$ is also affected by inbreeding. Perhaps the higher IBD levels in inbred individuals conflict with the assumption of unrelatedness in the estimator derivation.

When the LD weights are combined with the local IBD methods, there is clear improvement in performance for the HMM method (compare $\widehat{\Phi}_H$ to $\widehat{\Phi}_{H_W}$), but less so for the Day-Williams local method (compare $\widehat{\Phi}_L$ and $\widehat{\Phi}_{L_W}$). As noted above, the HMM overestimates IBD in high-

	Unconstrained						Constrained					
	FS	HS	C1	C2	C3	IN	FS	HS	C1	C2	C3	IN
$\widehat{\Phi}_D$	3.12	6.38	12.34	49.55	187.68	6.31	3.12	6.38	12.34	47.90	144.48	6.31
$\widehat{\Phi}_C$	3.00	4.85	8.32	30.83	116.22	4.81	3.00	4.85	8.32	30.42	95.98	4.81
$\widehat{\Phi}_R$	2.42	4.33	7.99	31.64	120.56	4.31	2.42	4.33	7.99	31.19	99.20	4.31
$\widehat{\Phi}_T$	2.19	4.26	7.96	30.84	116.24	4.17	2.19	4.26	7.96	30.43	96.01	4.17
$\widehat{\Phi}_N$	3.01	4.70	7.87	28.20	105.46	4.68	3.01	4.70	7.87	27.98	88.58	4.68
$\widehat{\Phi}_K$	2.64	4.10	7.21	23.74	85.57	5.03	2.64	4.10	7.21	23.69	72.58	5.03
$\widehat{\Phi}_W$	1.59	2.60	4.44	17.33	65.60	3.09	1.59	2.60	4.44	17.33	56.70	3.09

Table 2.4: Estimation accuracies from simulation study as measured by the ratio ($\times 10^2$) of the RMSE and the average realized kinship. For each GRM estimator, the left panel shows results from unconstrained estimates. The right panel shows results from estimates constrained to the natural range of $[0, 1]$.

LD regions, so that the LD weights are beneficial.

Pedigree kinship (Ψ) is included in Table 2.2; it may be considered an estimator based only on the pedigree, and not on genetic data. The RMSE here represents the variation in realized kinship among pairs of individuals in the same pedigree relationship. The advantage of using genetic-data-based estimates of realized kinship instead of pedigree kinship is clear. Only on remote relatives (C2 and C3) do some of the marker-based estimates differ from the realized kinship by more than do the pedigree values. In fact, the results of Table 2.2 underplay the performance of the GRM estimators on remote relatives, since, for all the other estimators, estimates are constrained to be within the $[0, 1]$ range. A comparison of constrained and unconstrained performance of the GRM estimators is shown in Table 2.4.

	Relationship		
k_2	FS	DFC	QHFC
set to 0	2.19	3.74	3.73
sim. value	2.19	3.74	3.73

Table 2.5: Estimation accuracies of $\widehat{\Phi}_T$ from simulation study as measured by the ratio ($\times 10^2$) of the RMSE and the average realized kinship (equation (2.18)). Optimal weights were computed using either $k_2 = 0$ or the simulated k_2 values.

2.5.3 Additional results

In this section we provide additional results to complement our findings in Section 2.5.2. We commented in Section 2.3 that the effect of k_2, the realized proportion of the genome that the pair of individuals share both genes IBD, is ignorable in deriving the two-step GRM estimates. Bilateral relatives may share two genes IBD over parts of the genome. Recall that we need initial estimates of Φ and k_2 in the implementation of $\widehat{\Phi}_T$. In practice, we ignore estimation of k_2 by simply setting $k_2 = 0$. This simplification only affects calculation of $\tilde{\boldsymbol{w}}^*(\tilde{\boldsymbol{a}}^*)$ for bilateral relatives. We evaluate the effect of setting $k_2 = 0$ in terms of estimation accuracy on three bilateral relative types: full siblings (FS), double first cousins (DFC) and quadruple half first cousins (QHFS).

Table 2.5 shows estimation accuracies in the same fashion as Table 2.2. The best possible initial estimates of k_2 are simply the simulated values themselves. It is clear that estimation accuracies are indifferent to the two choices of k_2 values used in computing $\tilde{\boldsymbol{w}}^*(\tilde{\boldsymbol{a}}^*)$. This is because the effect of k_2 on $\tilde{\boldsymbol{w}}^*(\tilde{\boldsymbol{a}}^*)$ (and in turn on $\mathrm{Var}[\widehat{\Phi}(\tilde{\boldsymbol{a}}^*, \tilde{\boldsymbol{w}}^*(\tilde{\boldsymbol{a}}^*))]$) is negligible.

Recall that differences between sets of weights can be characterized by how markers with different allele frequencies are weighted relative to a marker with frequency 0.5 (see Section

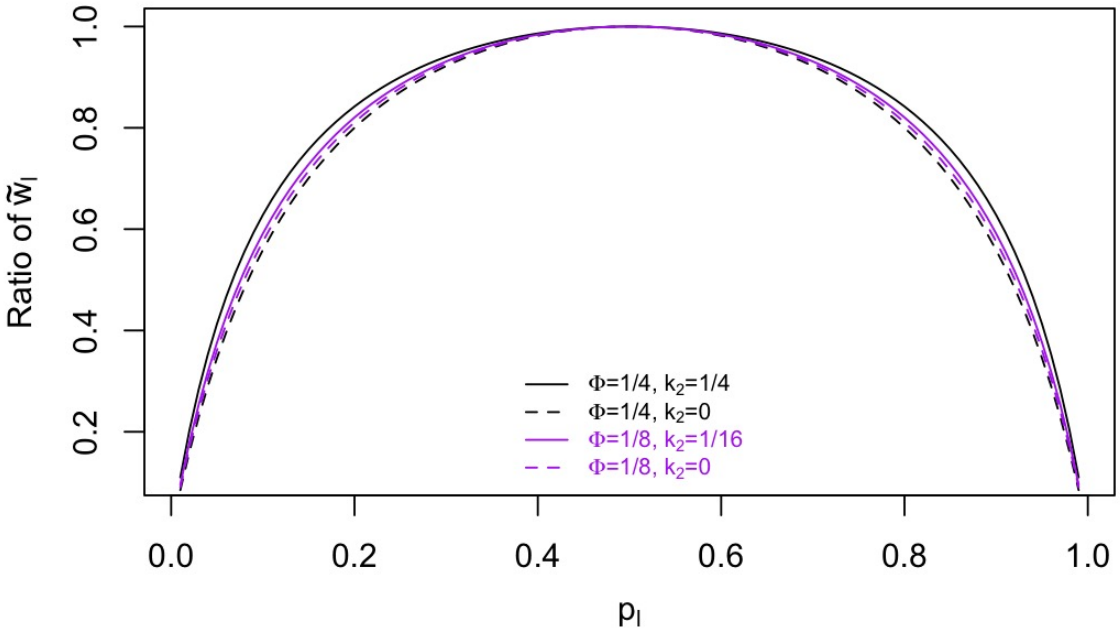


Figure 2.2: Ratio of weights between a marker with frequency p_l and a marker with frequency 0.5, under $\tilde{\boldsymbol{w}}^*(\tilde{\boldsymbol{a}}^*)$. Four sets of (Φ, k_2) were used as initial estimates to compute $\tilde{\boldsymbol{a}}^*$ and $\tilde{\boldsymbol{w}}^*(\tilde{\boldsymbol{a}}^*)$. Note that (Φ, k_2) of the solid lines correspond to pedigree expectations for full siblings (black) and double first cousins (purple) respectively.

2.3). Figure 2.2 shows that setting $k_2 = 0$ has little effect on the relative weights of $\tilde{\boldsymbol{w}}^*(\tilde{\boldsymbol{a}}^*)$, when initial estimates of (Φ, k_2) reflect the expected amount of sharing between full siblings or double first cousins. Results on quadruple half first cousins are indistinguishable from that of double first cousins, and were thus not plotted.

In Section 2.5.2 we used real haplotypes and a dense SNP marker panel. Thus LD is present in the simulated data. However, in assuming the variance formula (2.9) the covariances due to LD are ignored (see equation (2.8)). To investigate the impact of LD on the relative performance of estimators we compared the empirical and theoretical (no-LD) standard deviations for several estimators and relationship types. For any GRM estimator described in

	Relationship				
Estimator	FS	HS	C1	C2	C3
$\widehat{\Phi}_D$	4.11	4.10	3.82	3.75	3.53
$\widehat{\Phi}_C$	4.30	4.12	3.81	3.80	3.66
$\widehat{\Phi}_R$	3.94	3.89	3.63	3.66	3.50
$\widehat{\Phi}_T$	4.00	4.04	3.79	3.82	3.66

Table 2.6: Ratio of empirical root mean squared error and standard error computed using Equation 2.9 under the no-LD assumption.

Section 2.3 and any relationship type (except the inbred cousins) listed in Table 2.2, the true variance is well approximated by empirical MSE. The theoretical (no-LD) variance (2.9) can be computed using the simulation values of Φ and k_2 and averaged across all 1000 pairs of that relationship type. Table 2.6 summarizes the results. We see that factor by which the standard deviation is underestimated by ignoring LD is generally smaller on remote relatives. More importantly, for a given relationship type, it is fairly consistent across estimators. This suggests that between pairs of GRM estimators that do not adjust for LD, relative efficiency computed under the assumption of linkage equilibrium can be a good approximation to the true relative values.

We have assumed population homogeneity so that allele frequencies and LD weights can be estimated once and applied to all pairs of relatives. This simulation choice has the advantage of having a bigger pool of founder haplotypes available for sampling, and thus lower correlation among estimates from different simulation replicates. It also induces more complex population structure in the simulated data, which is likely to influence performance of some estimators more than others. In limited additional experiments, we investigated performance of a subset of estimators on data simulated under either European ancestry or

	European ancestry						African ancestry					
Estimator	FS	HS	C1	C2	C3	IN	FS	HS	C1	C2	C3	IN
$\widehat{\Phi}_D$	1.40	2.87	6.10	24.76	101.21	3.58	1.32	2.68	5.63	22.28	90.47	3.29
$\widehat{\Phi}_C$	1.31	2.25	4.36	17.24	72.83	2.65	1.26	2.02	3.76	14.12	59.52	2.37
$\widehat{\Phi}_R$	1.21	2.22	4.49	18.49	78.06	2.68	1.00	1.80	3.66	14.82	62.81	2.23
$\widehat{\Phi}_T$	1.11	2.14	4.32	17.32	72.89	2.58	0.92	1.77	3.63	14.10	59.56	2.19
$\widehat{\Phi}_W$	1.40	2.24	4.26	15.14	63.24	2.74	1.26	1.98	3.76	12.97	54.25	2.51
$\widehat{\Phi}_H$	2.39	4.46	7.60	22.59	81.02	5.06	1.09	1.75	2.63	6.27	21.74	2.00

Table 2.7: Estimation accuracies from simulation study (European or African ancestry) as measured by the ratio ($\times 10^2$) of the square root of MSE and the average realized kinship.

African ancestry (Table 2.7), and compared results to those of Table 2.2. Unsurprisingly, the GRM estimators generally benefit from the lesser structure in these more homogeneous pools of haplotypes. Relative performance among the GRM estimators that do not adjust for LD remains unchanged. The LD weighted GRM estimator ($\widehat{\Phi}_W$) loses to the two-step GRM estimator ($\widehat{\Phi}_T$) on close relatives, suggesting the amount of LD adjustment (given the choice of marker panel and ancestry) is not enough to offset the effect of deviation from the assumption of unrelatedness. However, these results also suggest $\widehat{\Phi}_W$ is robust to population substructure and admixture; compared to other GRM estimators, its performance is much less affected by the complex structure of the combined population (compare Table 2.2 to Table 2.7). The HMM estimator ($\widehat{\Phi}_H$) does better under African ancestry (than under combined ancestry), and worse under European ancestry. This is likely due to the higher level of LD in the European population; $\widehat{\Phi}_H$ is sensitive to LD (Brown *et al.* 2012).

For convenience, we selected SNPs based on even genetic distance spacing and minor allele frequency (MAF). The relationship between genetic distance and LD is far from uniform,

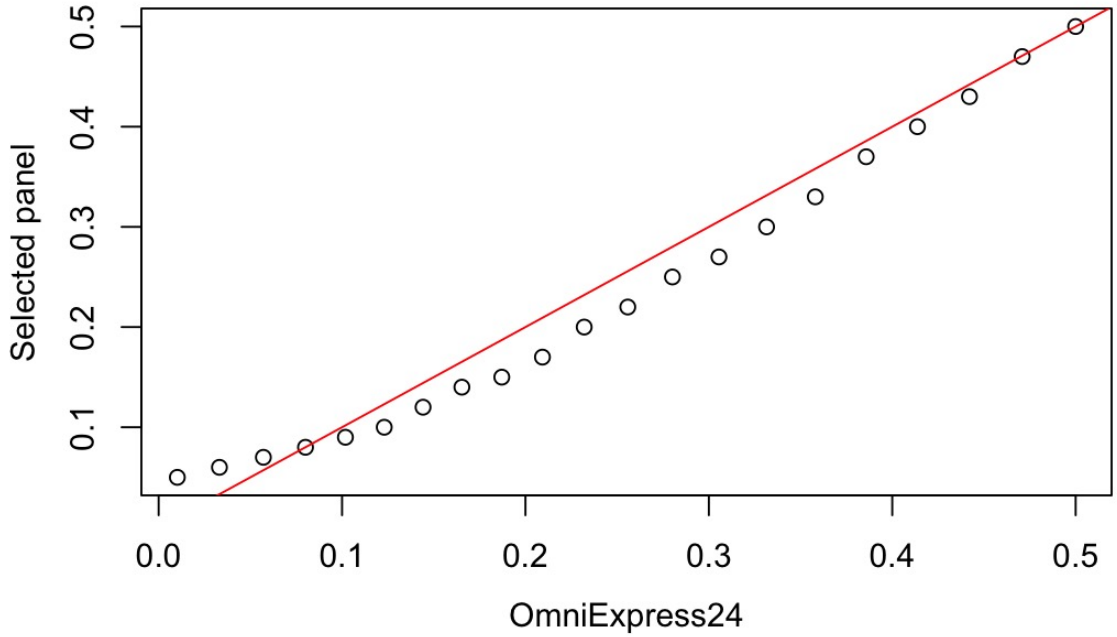


Figure 2.3: QQ-plot of distributions of minor allele frequencies in the selected marker panel versus OmniExpress24 autosomal markers.

and our results on the impact of adjusting for LD show that it is a significant factor in our marker panel. The distribution of allele frequencies in the K170 marker panel used here is quite similar to that in the commonly used OmniExpress24 GWAS chip (see Figure 2.3). Compared to the OmniExpress24 autosomal markers, the K170 marker panel has a slight underrepresentation for markers with MAF ≥ 0.1. However, the K170 panel was selected with a threshold of MAF ≥ 0.05, whereas about 9% of the OmniExpress24 autosomal markers fall below this threshold. We found that naive pruning of markers by MAF generally reduces accuracy of the two-step GRM estimator ($\widehat{\Phi}_T$), but can sometimes improve accuracy of other Group-1 GRM estimators depending on relationship types. Even on MAF-pruned marker sets, $\widehat{\Phi}_T$ always stands out among the Group-1 GRM estimators (results not shown). Overall, our selection of SNPs by genetic spacing and MAF create no strong biases in our estimator comparisons.

2.6 Discussion

We have shown that improved estimators of realized kinship can be obtained (1) by optimal weighting of markers, (2) by taking physical contiguity of genome into account, and (3) by weighting on the basis of LD. In practice, the choice of estimator largely depends on the availability of information. When one only has SNP genotypes and marker allele frequencies to work with, the two-step GRM estimator ($\widehat{\Phi}_T$) is both accurate and computationally efficient. If large genotyped samples from the relevant population are available, the LD weighted GRM estimator ($\widehat{\Phi}_W$) is an attractive alternative. The LD weights can be computed very efficiently using existing optimization software, and the computation needs to be done only once for a population. If information on marker order or genetic positions is available, either the local Day-Williams estimator ($\widehat{\Phi}_L$) or the HMM estimator ($\widehat{\Phi}_H$) can offer substantial increase in accuracy at the cost of longer computation time (see Section 2.5.1). Once computed, these local IBD estimates across the genome can also be used in other analyses that use location-specific IBD: for example, in gene mapping (see Chapter 5).

In the derivation of optimal estimators, assumptions such as absence of inbreeding, linkage equilibrium or unrelatedness were necessary to keep computations tractable. However, the proposed estimators applied well outside the initially assumed context. The two-step GRM estimator ($\widehat{\Phi}_T$) does not seem to be affected by presence of inbreeding. It compares favorably to other estimators that use the same amount of information even when the assumption of linkage equilibrium is violated. The LD weighted GRM estimator ($\widehat{\Phi}_W$) performed well on all relationship types considered, and is only slightly affected when related individuals happen to be inbred.

In our simulation study, relative pairs are generated as random draws from each relationship type. In practice, non-random sampling may create biases causing realized kinship to differ from the pedigree values. For instance, ascertainment by traits in human genetics and

artificial selection in animal and plant genetics can result in sampled relatives being more (or less) genetically related than expected under the pedigree structure (Purcell *et al.* 2007; Liu *et al.* 2003). The comparison of pedigree kinship versus estimated realized kinships in Table 2.2 is thus an idealized best-case scenario (in favor of pedigree kinship).

Even as SNP panels become denser, or sequence data become available, the issue of LD remains. Additional typed loci do not provide additional relatedness information without limit. Although methods that adjust for LD will gain from the use of additional markers, other methods may not benefit, and can be adversely affected. In particular, non-LD-adjusted local methods such as the HMM estimator ($\widehat{\Phi}_H$) should be applied on an LD-pruned marker set to avoid over-estimation of IBD (see Chapter 3). In human applications with nominally unrelated individuals, haplotypic similarities due to LD must be distinguished from cryptic relatedness. In animal and plant breeding applications, high levels of LD are a regular feature of artificially selected populations with a small effective founder population size. Therefore, methods for efficient kinship estimation in the presence of LD remain relevant even in the age of full genome sequencing.

Chapter 3

ESTIMATION OF LOCATION-SPECIFIC KINSHIP

In this chapter we focus on pedigree-free estimation of local kinship. We start by comparing two of such methods: the local method of Day-Williams *et al.* (2011), and the method proposed by Brown *et al.* (2012). In Chapter 2 we have compared their performance as estimators of genome-wide realized kinship. Here we take an in-depth look at their abilities to estimate local kinship, and show why one method might be preferred to the other. Next, we explore factors that are important for estimating local kinship, such as marker density and the level of linkage disequilibrium (LD) present in the marker panel. Lastly, we investigate the potential performance gain (or loss) by incorporating pedigree information, should it be available.

3.1 Overview of current methods

We provide an overview of three methods that will feature in the chapter. The local method of Day-Williams *et al.* (2011), referred to as DWL in the rest of this chapter, is a two-stage method that estimates local kinships. In stage one, we obtain initial local kinship estimates, z_l's, by applying a global kinship estimator (topic of Chapter 2) over sliding windows that center at each marker location. This means the ordering of markers is required in addition to marker genotypes and allele frequencies. In stage two, a smoothing algorithm is applied to obtain final local kinship estimates, $\hat{\phi}_l$'s. Several key ideas are incorporated in the smoothing algorithm: **(a)** final local kinship estimates are constrained, *i.e.*, $\hat{\phi}_l \in \{0, 0.25, 0.5, 1\}$; **(b)** changes in $\hat{\phi}_l$'s along a chromosome segment are rare; **(c)** patterns of z_l's and $\hat{\phi}_l$'s on the same chromosome should be similar; and **(d)** difference between $\hat{\phi}_l$ and $\hat{\Phi}$, estimates of realized kinship over the chromosome that contains marker l, should be small. These considerations

lead to the following minimization objective function,

$$f(\hat{\phi}_1, ..., \hat{\phi}_L) = \sum_{l=1}^{L}(z_l - \hat{\phi}_l)^2 + \lambda_1 \sum_{l=1}^{L}(\hat{\phi}_l - \hat{\Phi})^2 + \lambda_2 \sum_{l=2}^{L}(\hat{\phi}_l - \hat{\phi}_{l-1})^2, \tag{3.1}$$

where λ_1 and λ_2 are tuning parameters for penalty weights. Equation (3.1) can be solved with dynamic programming. Day-Williams *et al.* (2011) suggested that setting $\lambda_1 = 0$ and $\lambda_2 = 100$ generally work well. In addition, the authors proposed to use observed genotypes to constrain values of $\hat{\phi}_l$. For example, observing two homozygotes with different alleles at marker l is highly suggestive of $\phi_l = 0$, so we set $\hat{\phi}_l = 0$. If this observation is made at two markers less than 1 million basepairs (Mb) apart, the proposed algorithm sets $\hat{\phi}_l = 0$ at all intervening markers.

The method of Brown *et al.* (2012), referred to as `ibd_haplo` for the computer program that implements it, uses a hidden Markov model (HMM) to infer IBD partition (Section 1.1) at each marker location. At any point in the genome, IBD state between multiple DNA copies can be represented as a partition into subsets that are IBD. The Ewen's sampling formula (ESF) (Ewens 1972; Balding and Nicolas 1994) provides a useful one-parameter model for IBD partition at a single locus. The parameter β represents the pointwise probability of IBD between two DNP copies (Thompson 2008). The HMM of `ibd_haplo` has ESF as its limiting distribution, and uses a modified "Chinese restaurant process" (Tavare and Ewens 1997) for transition of IBD partitions across loci. Transition rate is governed by another parameter, α, that specifies the rate at which recombination occurs. The data model assumes that IBD DNA is of the same allelic type, and that non-IBD DNA are of independent allelic types. Genetic positions of markers are required for this method. Although our focus is on pairwise kinship estimation, `ibd_haplo` can be applied to infer IBD involving more than four DNA copies.

When the pedigree is known, it provides the null distribution for the inheritance vector at any point in the genome. Under the assumption of no genetic interference, the inheri-

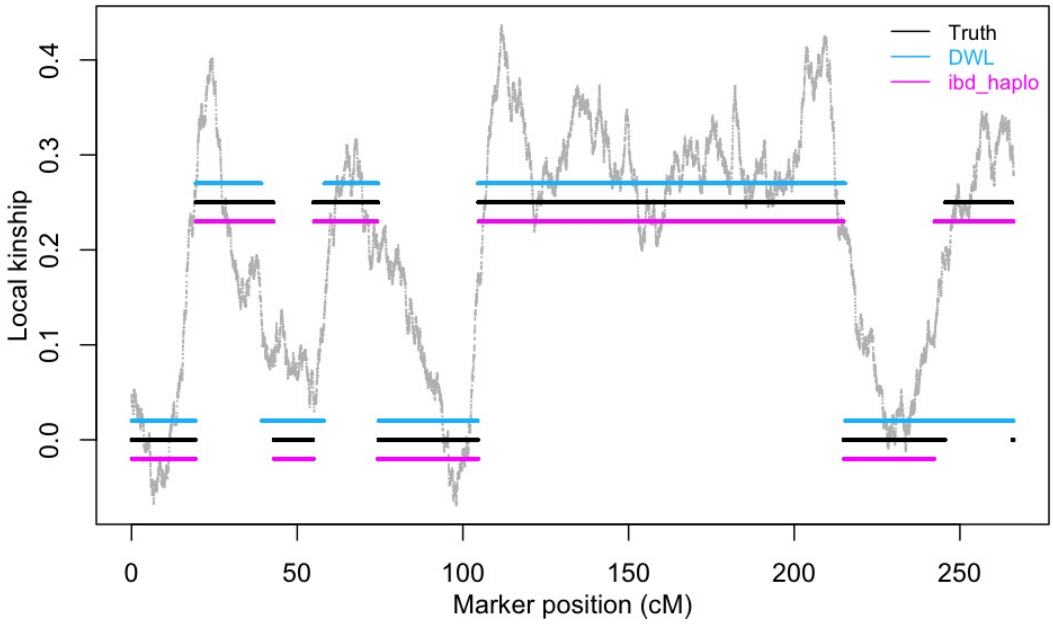


Figure 3.1: Local kinship estimation by DWL and `ibd_haplo` on chromosome 1 of a full sibling pair. For visual clarity, estimates from the two methods are plotted with $\pm$ 0.02 offsets from the truth respectively. Grey wiggly line in the background represents "global" kinship estimates in sliding windows in the intermediate step of DWL.

tance vectors along the chromosome form a Markov chain (Section 1.1). The widely used computer program Merlin (Abecasis *et al.* 2002) uses a HMM to calculate the probabilities of inheritance vector configurations (the hidden states) at each locus, given pedigree information and observed marker genotypes. This method will be referred to as Merlin in the rest of this chapter.

3.2 Local kinship estimation: DWL and `ibd_haplo`

In Chapter 2 we compared the performance of DWL and `ibd_haplo` as estimators of global realized kinship. Estimation accuracy was measured by the root mean squared error scaled by the expected kinship (equation (2.18)). While this measure remains a useful one-number summary for performance of a local method, it does not provide insights on how DWL and

`ibd_haplo` perform differently at the local level. To better understand the difference between the two methods, it is helpful to look at the local kinship estimates on a single chromosome.

Under the same experimental conditions described in Section 2.5.1, Figure 3.1 shows local kinship estimates produced by DWL and `ibd_haplo` on chromosome one of a first cousin pair. For DWL, the initial estimates z_l's (grey) are generally very noisy. When the smoothing algorithm is applied on top of this noisy background, it is likely that DWL will have trouble at edges of IBD segments (e.g., around 50cM mark). At the end of the chromosome, DWL failed to pick up a long segment of IBD even when the background initial estimates are suggestive of IBD sharing. This is likely due to a combination of (a) penalization on the number of recombination events; and (b) observations of two homozygotes of different alleles are taken as signs of no IBD sharing. On the other hand, `ibd_haplo` is generally able to identify edges of IBD segments very well. It failed to do so only once near the 250cM mark, over-estimating IBD possibly due to LD. It also missed the very short non-IBD segment at the end of the chromosome, which is probably a challenge to any local method.

We noted in Chapter 2 that DWL tends to under-estimate IBD, whereas `ibd_haplo` tends to over-estimate IBD. This statement is further supported by Table 3.1. For the marker density used in the simulation study, both methods tend to under-estimate IBD more often for distant relatives, whereas they tend to over-estimate IBD more often for close relatives. This is exactly what we would expect given the average amount of IBD sharing varies with remoteness of relationship. When there is IBD, `ibd_haplo` has a higher chance of finding it. When there is no IBD, DWL does only slightly better in confirming the absence of IBD. For more complex relationship types such as the JV inbred cousins, `ibd_haplo` continues to estimate local kinship with high level of accuracy, whereas DWL consistently under-estimates the amount of IBD sharing. Having observed that DWL has trouble especially at edges of segments (Figure 3.1), it is no surprise that the complex IBD patterns of the JV inbred cousins pose a strong challenge to this method.

		DWL				ibd_haplo			
	$\phi \backslash \hat{\phi}$	0	0.25	0.5	1	0	0.25	0.5	1
FS	0	0.968	0.032	0.000	0.000	0.960	0.038	0.001	0.000
	0.25	0.043	0.909	0.048	0.000	0.006	0.975	0.018	0.001
	0.5	0.000	0.100	0.900	0.000	0.000	0.001	0.998	0.001
HS	0	0.983	0.017	0.000	0.000	0.977	0.022	0.001	0.000
	0.25	0.050	0.937	0.014	0.000	0.007	0.981	0.011	0.001
C1	0	0.988	0.012	0.000	0.000	0.983	0.016	0.001	0.000
	0.25	0.101	0.889	0.010	0.000	0.016	0.971	0.012	0.001
C2	0	0.994	0.006	0.000	0.000	0.990	0.008	0.001	0.000
	0.25	0.158	0.835	0.008	0.000	0.029	0.958	0.012	0.001
C3	0	0.996	0.004	0.000	0.000	0.992	0.006	0.001	0.000
	0.25	0.215	0.779	0.006	0.000	0.043	0.945	0.012	0.001
IN	0	0.983	0.017	0.000	0.000	0.977	0.021	0.002	0.000
	0.25	0.100	0.876	0.024	0.000	0.016	0.969	0.014	0.001
	0.5	0.006	0.226	0.766	0.001	0.002	0.004	0.990	0.004
	1	0.000	0.001	0.259	0.740	0.000	0.000	0.007	0.993

Table 3.1: Proportions of local kinship estimates ($\hat{\phi}$) conditional on truth (ϕ), computed over the genome and over simulation replicates. Proportions that correspond to correct local kinship estimates are shown in blue.

Overall, `ibd_haplo` is a more accurate method than DWL. The parameters of DWL do not hold intuitive meanings. Rather, they need to be tuned for the dataset under study to achieve best results. For results of DWL we have seen so far, we used parameters tuned

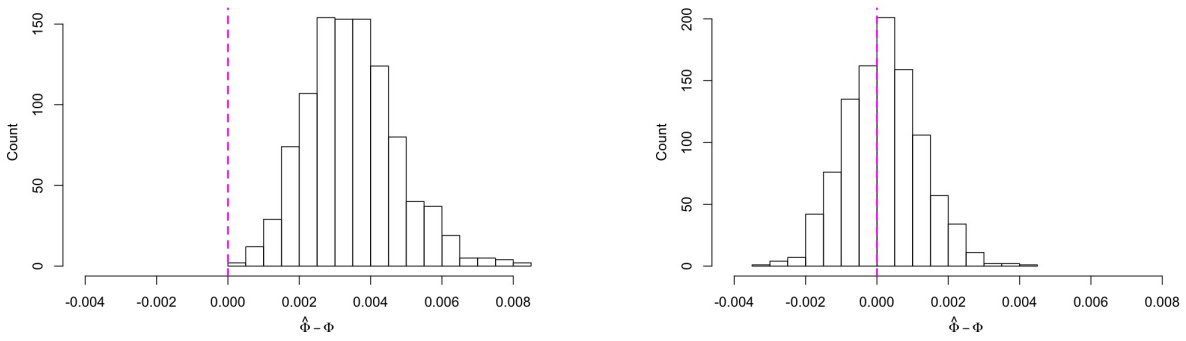
on an independently simulated dataset (see Section 2.5.1 for details). Performance of DWL appeared to be highly sensitive to the choices of parameters in the tuning process (results not shown). As a contrast, parameters of `ibd_haplo` have intuitive meanings (see Section 3.1). These parameters specify a weak prior distribution for the HMM, whose influence diminishes as information from marker data accrues. This is exactly what we have seen in the tuning process: performance of `ibd_haplo` is not very sensitive to choices of parameters (results not shown).

Another big advantage of `ibd_haplo` over DWL is its ability to easily handle potential data error. In the HMM of `ibd_haplo`, a small error probability can be (and is by default) included in the emission stage to account for possible mis-matches between observed genotypes and true IBD states at each marker location. This is, however, not easily done in the case of DWL. The initial kinship estimates in sliding windows produced in the intermediate step should be robust to data error, but the subsequent smoothing algorithm relies on observed genotypes to limit the possible values of local kinship (see description in Section 3.1). This feature of the algorithm improves accuracy and efficiency in absence of data error, but it can be detrimental when the assumption of perfect data is violated.

Lastly, `ibd_haplo` is more general than DWL in that it estimates local IBD state as opposed to local kinship. The higher resolution of `ibd_haplo` is particularly useful when inbreeding is of concern. Thus, we will focus primarily on `ibd_haplo` in the rest of this chapter, where we explore factors that are important for local IBD estimation.

3.3 Effects of marker density and LD pruning

In Chapter 2 we showed that adjusting for LD in the marker panel can lead to more accurate estimation of global realized kinship. This is because additional markers do not provide additional information without limit. As marker density increases, at some point the noise due to LD outweighs the benefits of additional markers. This is particularly true in the



(a) (b)

Figure 3.2: Distributions of global realized kinship estimation errors by `ibd_haplo` for 1000 first cousin pairs, using **(a)** the full K170 marker set, and **(b)** the LD-pruned K170 marker set that contains about 50% of the original markers. The pruning process is described in Section 2.5.1 in the setup for the LD-pruned GRM estimator $\widehat{\Phi}_N$.

case of `ibd_haplo`, where the underlying HMM assumes independence of genotypes at different markers conditional on IBD states. The performance of `ibd_haplo` (its derived global kinship estimator $\widehat{\Phi}_H$) improved noticeably when LD weights from another estimator (the LD weighted GRM $\widehat{\Phi}_W$) were applied to the local kinship estimates (see Table 2.2). We commented in Section 2.6 that a more natural way to adjust for LD in the case of `ibd_haplo` is to apply the method on a LD-pruned marker set, since the HMM assumes absence of LD. This point is demonstrated in Figure 3.3, where applying `ibd_haplo` on a LD-pruned marker set has effectively eliminated the positive bias in estimation of global realized kinship for first cousins. It is of interest to know how to strike a good balance between marker density and LD for local IBD methods such as `ibd_haplo`.

We attempt to answer this question using a simulation study. For a given marker panel, pruning for LD leads to reduction in marker density. Pruning for the same level of LD (e.g., by measure of r^2 (or ρ^2), the squared genotypic dosage correlation between two loci) in different marker panels typically leads to different marker densities in the resulting LD-pruned

panels. For this reason, we use both the K170 and the OmniExpress24 panels (see Section 1.4.1) so that different combinations of marker density $\times$ LD can be compared. For the K170 panel, different extents of LD pruning are done (as described in Section 2.5.1 for LD-pruned GRM) based on $r^2 \in \{0.01, 0.04, 0.1, 0.2, 0.36, 0.5, 0.8\}$. For the OmniExpress24 panel, LD pruning is done only for $r^2 \in \{0.005, 0.01, 0.04, 0.1\}$, to ensure the range of resulting marker densities overlaps with that of the K170 panel.

We consider five pairwise relationship types in this simulation: full sibling (FS), half sibling (HS), first cousins (C1), second cousins (C2) and third cousins (C3). For each relationship type, inheritance patterns are simulated for 1000 independent pairs on all 22 autosomes using the R package `rres` (Wang 2018). Separately, genotypes for both the K170 and the OmniExpress24 panels are generated given each simulated inheritance pattern, using the same process described in Section 2.5.1. For each set of simulated genotypes, `ibd_haplo` is run for each LD-pruned marker set. Given the large number of experimental conditions for the same local method, we use the root mean squared error (RMSE) scaled by expected kinship (equation (2.18))to summarize performance of `ibd_haplo` on each LD-pruned marker set.

Figure 3.3 shows results of the simulation. For LD-pruned marker sets of the K170 panel (shown in magenta), it is clear that increasing marker density beyond $\sim$30 markers/cM, or LD level beyond the threshold of $r^2 = 0.2$ starts to hurt the accuracy of the estimator. This observation is consistent across relationship types. For LD-pruned marker sets of the OmniExpress24 panel (shown in blue), we see that accuracy of the estimator increases at a decreasing rate as marker density (and level of LD) increases, even beyond that of 30 markers/cM. In fact, the curve flattens out right after the $\sim$43 markers/cM, or $r^2 = 0.1$ mark for LD-pruned marker sets of the OmniExpress24 panel (results not shown for clarity). It stays flat utill $\sim$82 markers/cM , or $r^2 = 0.2$, after which it is expected to start increasing again at some point as level of LD increases, just as we saw with the K170 curve.

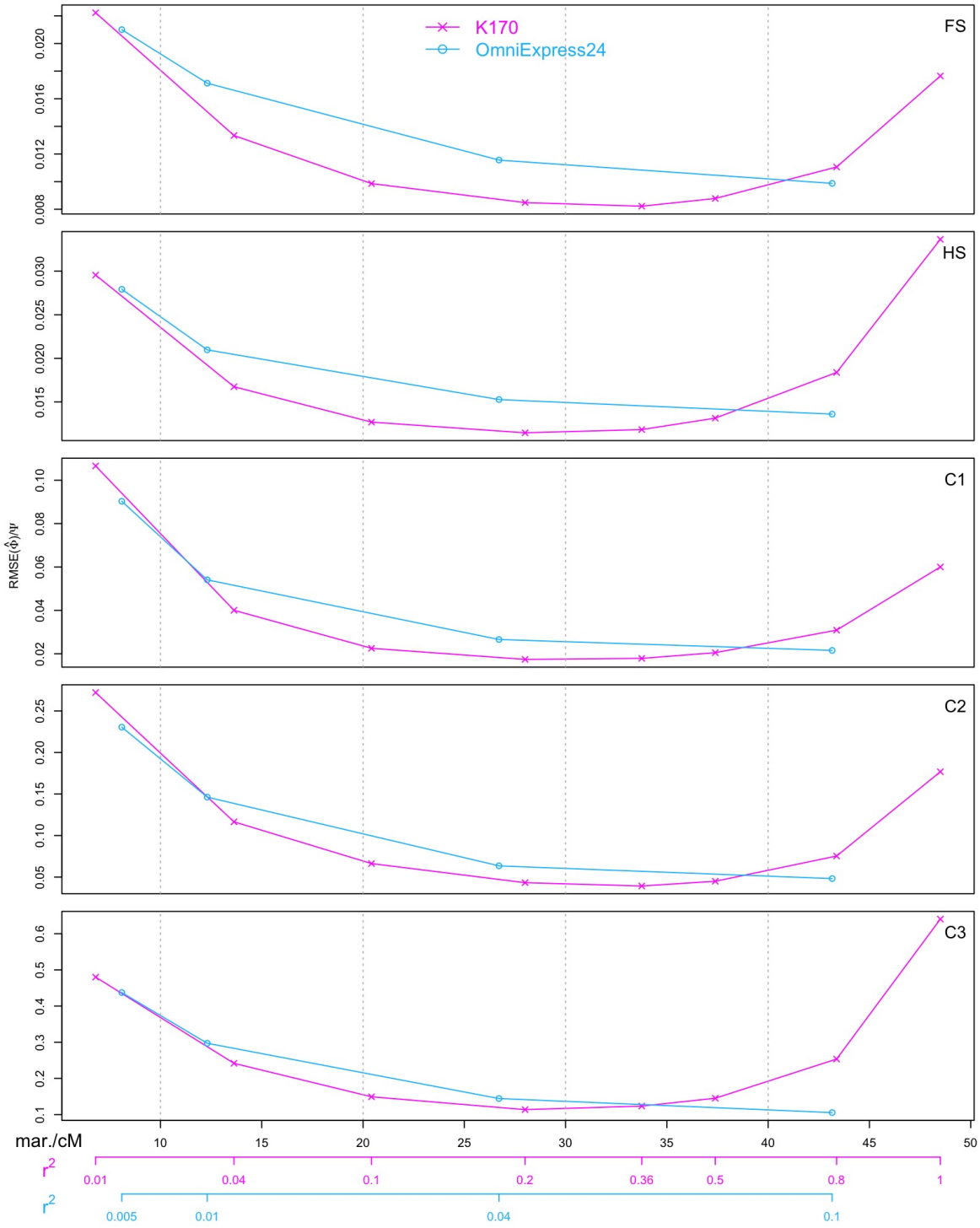


Figure 3.3: Accuracy of `ibd_haplo` for different relationship types, applied to LD-pruned marker sets. Results are summarized as the root mean squared error of global kinship estimates scaled by expected kinship. Marker densities correspond to different level of LD pruning in K170 (magenta) and OmniExpress24 (blue). Level of LD pruning is represented by threshold on r^2, where $r^2 = 1$ indicates the original marker panel without pruning.

	K170	$r^2 = 0.01$				$r^2 = 0.2$				full panel			
	$\phi\backslash\hat{\phi}$	0	0.25	0.5	1	0	0.25	0.5	1	0	0.25	0.5	1
FS	0	0.855	0.144	0.000	0.000	0.961	0.039	0.000	0.000	0.960	0.038	0.001	0.000
	0.25	0.042	0.938	0.020	0.000	0.010	0.984	0.006	0.000	0.006	0.975	0.018	0.001
	0.5	0.001	0.038	0.962	0.000	0.000	0.004	0.996	0.000	0.000	0.001	0.998	0.001
HS	0	0.940	0.060	0.000	0.000	0.982	0.018	0.000	0.000	0.977	0.022	0.001	0.000
	0.25	0.059	0.941	0.000	0.000	0.010	0.989	0.000	0.000	0.007	0.981	0.011	0.001
C1	0	0.973	0.027	0.000	0.000	0.989	0.011	0.000	0.000	0.983	0.016	0.001	0.000
	0.25	0.169	0.831	0.000	0.000	0.029	0.971	0.000	0.000	0.016	0.971	0.012	0.001
C2	0	0.995	0.005	0.001	0.000	0.997	0.003	0.000	0.000	0.990	0.008	0.001	0.000
	0.25	0.306	0.694	0.000	0.000	0.055	0.944	0.001	0.000	0.029	0.958	0.012	0.001
C3	0	0.999	0.001	0.000	0.000	0.999	0.001	0.000	0.000	0.992	0.006	0.001	0.000
	0.25	0.428	0.572	0.000	0.000	0.082	0.917	0.001	0.000	0.043	0.945	0.012	0.001

Table 3.2: Proportions of local kinship estimates ($\hat{\phi}$) conditional on truth (ϕ), computed over the genome and over simulation replicates. Proportions that correspond to correct local kinship estimates are shown in blue. `ibd_haplo` is applied on three LD-pruned marker sets of the K170 panel.

The differences between the magenta and blue curves in each plot of Figure 3.3 illustrate several interesting points. When marker density is low (≤ 10/cM), the two curves roughly coincide, as the difference in LD levels is very small. As marker density increases beyond 10 markers/cM, `ibd_haplo` is more accurate on the LD-pruned K170 markers. This is likely due to the OmniExpress24 panel having more markers with low minor allele frequencies (see Section 1.4.1), and the difference in LD levels is still not big. As marker density increases beyond 30 markers/cM, LD levels in the LD-pruned K170 marker sets increase rapidly, outweighing the benefits of additional markers. On the other hand, LD levels in the LD-pruned OmniExpress24 marker sets increase at a much slower pace. Thus, the two curves cross paths.

Table 3.2 shows the proportions of local kinship estimates ($\hat{\phi}$) conditional on truth (ϕ), when `ibd_haplo` was applied to three LD-pruned marker sets of the K170 marker panel. These three marker sets correspond to the most sparse marker set, the best performing marker set, and the full panel shown by the magenta curves in Figure 3.3. As expected, insufficient marker information in the most sparse marker set results in frequent under-estimation of IBD, whereas high level of LD in the full panel often leads to over-estimation of IBD. The moderately pruned marker set with threshold $r^2 = 0.2$ enjoys the best overall performance for having a good balance between marker information and level of LD.

3.4 Known pedigree structure

When pedigree structure is known, it can be used as additional information in local IBD estimation. As we described in Section 3.1, a widely used pedigree-based HMM is that of Merlin. Unlike the parameters α and β of `ibd_haplo` (Section 3.1), pedigree structure acts as a strong prior in the HMM of Merlin. This strong prior is expected to be helpful when marker data is not very informative, but can be a problem when assumptions of the model are violated. In this section we assume pedigree structure is correctly specified, and focus on the effect of changing LD levels on performance of Merlin.

3.4.1 Pairwise IBD estimation

In order to make a direct comparison to `ibd_haplo` studied in the previous section, we first consider Merlin in pairwise IBD estimation. That is, marker data are only available for the relative pair under study, not for all other members of the pedigree. We use the same simulation setup as in Section 3.3, and apply Merlin on various LD-pruned marker sets. Results are shown in Figure 3.4.

We observe some similar patterns in the results of `ibd_haplo` (solid lines) and of Merlin (dashed lines). For the K170 panel (magenta), both curves of `ibd_haplo` and Merlin have a convex shape, indicating the trade off between marker information and level of LD. For the

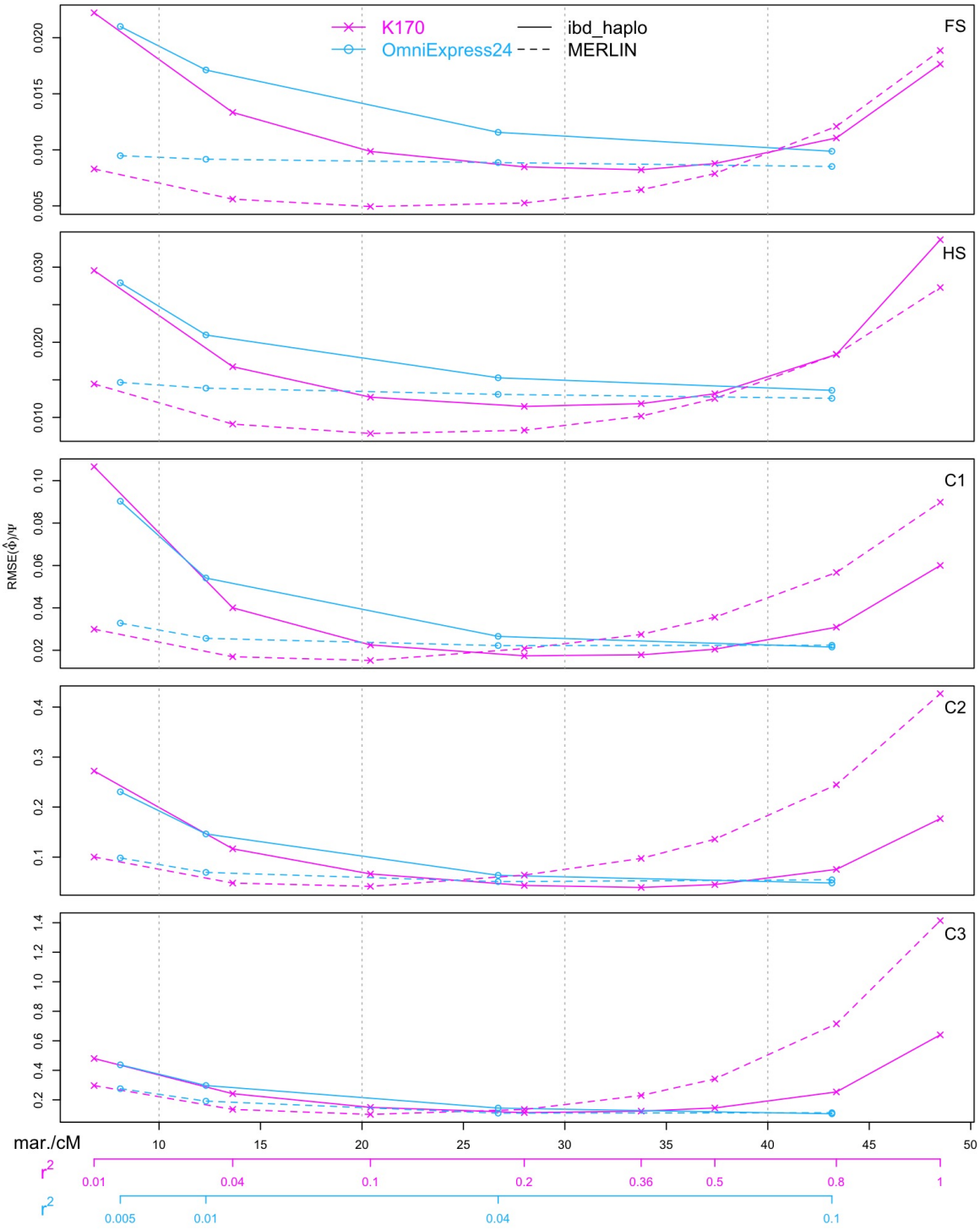


Figure 3.4: Accuracy of `ibd_haplo` (solid) and Merlin (dashed) for different relationship types, applied to LD-pruned marker sets. Results are summarized as the root mean squared error of global kinship estimates scaled by expected kinship. Marker densities correspond to different level of LD pruning in K170 (magenta) and OmniExpress24 (blue). Level of LD pruning is represented by threshold on r^2, where $r^2 = 1$ indicates the original marker panel without pruning.

OmniExpress24 panel (blue), both curves drop at a decreasing rate as marker density and LD level increases, and almost flatten out towards the end of the spectrum as they approach their respective inflection points. When marker density is low, both `ibd_haplo` and Merlin performed better on LD-pruned marker sets of the K170 panel than that of the OmniExpress panel, possibly due to difference in minor allele frequency distributions in the two panels. At higher densities, effects of LD kicks in and performance is better with the OmniExpress pruned marker sets.

There are also interesting differences between results of `ibd_haplo` and of Merlin. At lower marker densities, knowing the pedigree structure can be very helpful. This is most obvious in the case of close relatives, as there are fewer intervening pedigree members without genetic data. The benefit of pedigree information diminishes as marker density increases. When LD level is high, performance of Merlin can suffer much more than that of `ibd_haplo`. Even though the pedigree is correctly specified, the additional constraints imposed by the pedigree means that violation of other model assumptions (in this case, the assumption of independence across loci of genotypes conditional on IBD state) can be more detrimental. The effect of LD is stronger on remote relationship types. This is likely because it is more difficult to distinguish LD from IBD for remote relatives. Additional constraints imposed by pedigree information aggravate the effect of LD so that LD is mistaken to be IBD more often.

3.4.2 Joint IBD estimation

As described in Section 3.1, Merlin calculates the probabilities of inheritance vector configurations at all marker loci, given pedigree and observed marker genotypes. In general, the higher the availability of marker data among members of the pedigree, the more accurately inheritance vectors can be estimated. In turn, IBD sharing between any pair of individuals in the pedigree can be more accurately estimated. In this sense, the performance of Merlin shown in Section 3.4.1 represents a lower bound. Here we consider joint IBD estimation with

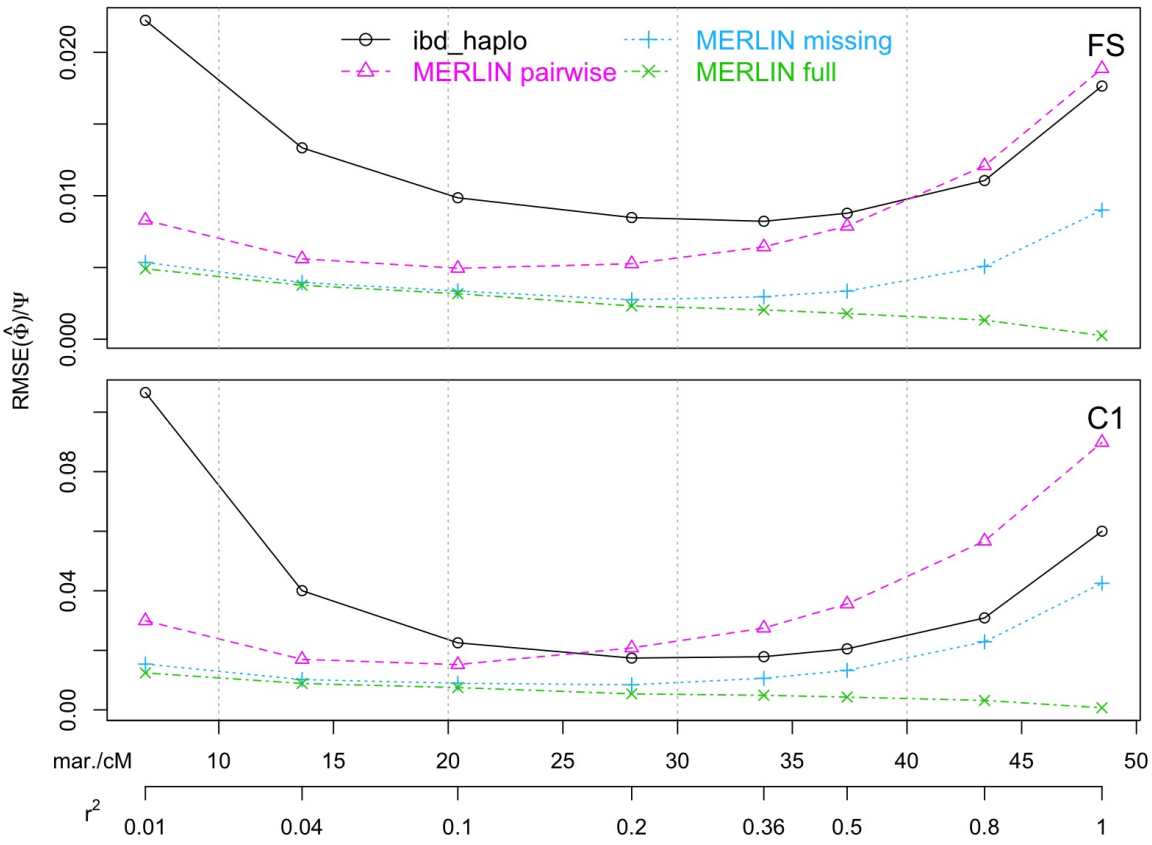


Figure 3.5: Accuracy of `ibd_haplo`, Merlin pairwise estimation, and Merlin joint estimation with two levels of marker data availability, applied to LD-pruned marker sets of the K170 panel. Results are summarized as the root mean squared error of global kinship estimates scaled by expected kinship. Level of LD pruning is represented by threshold on r^2, where $r^2 = 1$ indicates the original marker panel without pruning.

Merlin, where marker genotypes are available for additional members of the pedigree. This contrasts with pairwise IBD estimation in Section 3.4.1, where marker data is only available on the pair of individuals under investigation. The extent to which joint IBD estimation with Merlin is more accurate than pairwise IBD estimation depends on the pedigree structure as well as the availability of marker data on additional members of the pedigree. Our focus in this section is on how accuracy of joint IBD estimation with Merlin is affected by changing marker densities and levels of LD.

For simplicity, we only consider LD-pruned marker sets of the K170 panel in this section. We use the three-generation pedigree (Figure 1.3) as the full pedigree, in which there are multiple full sibling pairs (FS) and first cousin pairs (C1). We measure estimation accuracy only for those two relationship types. Within this framework, we consider joint IBD estimation at two different levels of marker data availability: (a) marker data available for all members of the pedigree (full); and (b) marker data only available for the six individuals in the last generation (missing). In each setup, estimation accuracy for each of the two relationship types is measured as an average over all pairs in the pedigree that have marker data available. For example, the full sibling pair "22" and "23" do not have marker data in case (b), so they are excluded from calculation of estimation accuracy for full sibling pairs in case (b).

Figure 3.5 shows results of the current simulation study, together with some results we showed earlier (Figure 3.4) on pairwise IBD estimation using `ibd_haplo` and Merlin. As expected, Merlin joint IBD estimation shows significant improvements over pairwise IBD estimation, and having full marker data is preferred to having marker data on the last generation only. The accuracy curve of joint IBD estimation with missing data (blue) has a similar shape to that of pairwise IBD estimation (magenta), where the trade off between marker information and level of LD is apparent. On the other hand, estimation accuracy of joint IBD estimation with full data (brown) increases steadily as marker density and level of LD increase. This is likely because the availability of marker data on all members of the pedigree provides very strong information for estimating inheritance vector configurations. As marker density and LD level increases, information provided by additional markers outweighs uncertainty caused by higher LD level.

3.5 Discussion

We have investigated the problem of local kinship estimation in this chapter. The primary focus is on methods that do not require pedigree information, where we compared the performance of DWL and `ibd_haplo` on common relationship types, as well as in a special case of inbreeding (JV inbred cousins). `ibd_haplo` is the more accurate method of the two. Apart from having higher overall estimation accuracy as shown in our simulation studies, `ibd_haplo` has additional advantages such as the ease of handling genotyping errors, low sensitivity to parameter specification, and the ability to infer joint IBD states among more than four chromosomes. These two methods present a good contrast between non-parametric and likelihood-based approaches for local kinship estimation. While they are not the only methods out there (see also, e.g., Purcell *et al.* (2007); Browning and Browning (2011); Moltke *et al.* (2011)), they do well to demonstrate the competitiveness of likelihood-based methods.

As SNP marker panels increase in density, the problem of LD becomes more relevant. For local kinship estimation by `ibd_haplo`, we showed that there is a trade off between marker information and level of LD in the marker panel. This is due to the HMM assumption of no LD. With a high density marker panel, `ibd_haplo` works best when applied on a LD-pruned marker set. A good rule of thumb is to prune the marker panel so that $r^2 \leq 0.2$ between markers. This pruning process can be implemented on a large population dataset, such as that of the The 1000 Genomes Project Consortium (2015).

We also extended our simulation study to include the pedigree-based HMM implemented in Merlin. Pedigree structure acts as a strong prior in local kinship estimation. It can be very helpful when marker density is low, but its benefits erode as marker density increases. More importantly, Merlin is also subject to the trade off between marker information and level of LD. When genetic marker data is only available for the pair of individuals under study, high level of LD is more detrimental to Merlin than to `ibd_haplo`. The effect of LD

on Merlin can be reduced by collecting genetic marker data on other members of the pedigree.

Local kinship estimation is often used as an intermediate step in different types of genetic analyses. We will return to this topic in Chapter 5 where we discuss gene mapping of quantitative traits.

Chapter 4

HERITABILITY ESTIMATION OF QUANTITATIVE TRAITS

In this chapter we investigate the problem of heritability estimation using a random effects model of two components, where the trait values are sums of normally distributed additive genetic random effects and unique environmental random effects. The correlation structure of the additive genetic random effects is captured by twice the matrix of kinship. We explore how study designs and measures of kinship affect outcome of heritability estimation, when the kinship matrix is potentially mis-specified.

4.1 Basic polygenic model

The basic polygenic model described in Section 1.3.1 was used by Thompson and Shaw (1990) to show how eigen-transformation can significantly improve efficiency of EM algorithm to find the maximum likelihood estimates (MLEs) of the variance parameters. More recently, Visscher and Goddard (2015) used the same model to study asymptotic sampling variance (ASV) of heritability estimates. Raffa and Thompson (2016) considered hypothesis testing and construction of confidence intervals for heritability. These authors focused on the use of pedigree kinship under the assumption of correct model specification.

On the applied side, it has become popular to estimate heritability from population samples, where the genetic correlation structure takes the form of the classic Genomic Relationship Matrix (GRM) (e.g., Yang *et al.* 2010). A population-based design avoids confounding due to shared environmental effects in close relatives. However, we have shown in Chapter 2 that the classic GRM is not an accurate estimator of realized kinship, especially for remote relatives. It is of interest to know the potential impact of the choice of kinship estimator on

heritability estimation.

The basic polygenic model (equation (1.1) of Section 1.3.1) is reproduced here

$$\boldsymbol{y} = \boldsymbol{g} + \boldsymbol{e}, \tag{4.1}$$

where $\boldsymbol{y}$ are trait values after adjustment of fixed effects, $\boldsymbol{g} \sim N(\boldsymbol{0}, \sigma_g^2 \boldsymbol{G})$ are additive genetic effects and $\boldsymbol{e} \sim N(\boldsymbol{0}, \sigma_e^2 \boldsymbol{I})$ are residuals. σ_g^2 and σ_e^2 are unknown variance parameters, and $\boldsymbol{G}$ is the appropriate genetic correlation matrix. Total phenotypic variance is $\sigma^2 = \sigma_g^2 + \sigma_e^2$. The goal is to estimate heritability, $h^2 = \sigma_g^2/\sigma^2$. We can parametrize the trait distribution in terms of $\boldsymbol{\theta} = (h^2, \sigma^2)$ as

$$\boldsymbol{y} \sim N\Big(\boldsymbol{0}, [h^2 \boldsymbol{G} + (1 - h^2)\boldsymbol{I}]\sigma^2\Big). \tag{4.2}$$

We assume throughout this chapter that the model described in (4.1) and (4.2) is true, but the fitted correlation matrix $\boldsymbol{G}_f$ may differ from the true correlation matrix $\boldsymbol{G}_t$. This type of model mis-specification can arise for a number of reasons (see Section 4.3 and 4.4). To investigate the asymptotic distributions of the MLEs and the effect of pedigree structure on the MLEs, we assume there are m mutually independent pedigrees of the same structure with finite pedigree size n, so that the overall correlation matrices $\boldsymbol{G}_t$ and $\boldsymbol{G}_f$ have block-diagonal structures where each block corresponds to a pedigree (e.g., $\boldsymbol{G}_t = \text{ diag}\{\boldsymbol{G}_{t,i}\}_{i=1}^m$).

4.2 Asymptotic distribution of heritability estimates

4.2.1 Correct model specification

When $\boldsymbol{G}_t = \boldsymbol{G}_f$, it follows from likelihood theory that the MLEs are consistent. Let the eigen-decomposition of $\boldsymbol{G}_f$ be $\boldsymbol{T}\boldsymbol{D}\boldsymbol{T}^T$, where $\boldsymbol{T}$ is the orthogonal matrix of eigenvectors and $\boldsymbol{D}$ is the diagonal matrix of eigenvalues of $\boldsymbol{G}_f$. The transformed trait

$$\boldsymbol{y}^* = \boldsymbol{T}^T \boldsymbol{y} \sim N\Big(\boldsymbol{0}, [h^2 \boldsymbol{D} + (1 - h^2)\boldsymbol{I}]\sigma^2\Big). \tag{4.3}$$

The covariance matrix $\text{Var}(\boldsymbol{y}^*)$ is diagonal, so that the log-likelihood function without the constant term is

$$\ell(h^2, \sigma^2; \boldsymbol{y}^*) = -\frac{1}{2}\Big[N\log(\sigma^2) + \sum_{i=1}^{m}\sum_{j=1}^{n}\log(h^2\lambda_{ij} + 1 - h^2) + \frac{1}{\sigma^2}\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{y_{ij}^{*2}}{h^2\lambda_{ij} + 1 - h^2}\Big], \quad (4.4)$$

where $N = mn$ is the total sample size, λ_{ij} is the jth eigenvalue of the correlation matrix of the ith pedigree, $\boldsymbol{G}_{t,i}$. $\text{ASV}(\hat{h}^2)$ is given by, among others, Visscher and Goddard (2015) as

$$\text{ASV}(\hat{h}^2) = \frac{2}{\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(\lambda_{ij}-1)^2}{(h^2\lambda_{ij}+1-h^2)^2} - \frac{1}{N}\Big[\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\lambda_{ij}-1}{h^2\lambda_{ij}+1-h^2}\Big]^2}. \quad (4.5)$$

Note in (4.4) and (4.5) that we grouped eigenvalues of $\boldsymbol{G}_f$ by pedigrees. This is possible because we assumed $\boldsymbol{G}_t$ (same as $\boldsymbol{G}_f$ in this case) to have a block-diagonal structure. Eigenvalues of the correlation matrix from pedigree i, $\boldsymbol{G}_{f,i}$, are also eigenvalues of $\boldsymbol{G}_f$.

4.2.2 Model mis-specification

When $\boldsymbol{G}_t \neq \boldsymbol{G}_f$, let the matrix of differences be $\boldsymbol{\Delta} = \boldsymbol{G}_t - \boldsymbol{G}_f$. The same eigen-decomposition and transformation based on the fitted correlation matrix $\boldsymbol{G}_f$ leads to the same log-likelihood function (4.4), but the true distribution of the transformed trait is now

$$\boldsymbol{y}^* = \boldsymbol{T}^T\boldsymbol{y} \sim N\Big(\boldsymbol{0}, [h_0^2\boldsymbol{D} + (1-h_0^2)\boldsymbol{I}]\sigma_0^2 + h_0^2\sigma_0^2\boldsymbol{T}^T\boldsymbol{\Delta}\boldsymbol{T}\Big), \quad (4.6)$$

where $\boldsymbol{\theta}_0 = (h_0^2, \sigma_0^2)$ are the true parameter values. Let

$$\boldsymbol{V} = \{V_{ij}\} \coloneqq \text{Var}(\boldsymbol{y}^*) = [h_0^2\boldsymbol{D} + (1-h_0^2)\boldsymbol{I}]\sigma_0^2 + h_0^2\sigma_0^2\boldsymbol{T}^T\boldsymbol{\Delta}\boldsymbol{T}$$

It follows that $\boldsymbol{y}^*\boldsymbol{y}^{*T} \sim W_N(\boldsymbol{V}, 1)$, a Wishart distribution with scale matrix $\boldsymbol{V}$ and 1 degree of freedom. Consequently,

$$\mathbb{E}[y_i^{*2}] = \text{Var}(y_i^*) = V_{ii}, \quad (4.7)$$

$$\mathbb{E}[y_i^{*2}y_j^{*2}] = \mathbb{E}[y_i^*y_j^*]^2 + \text{Var}(y_i^*y_j^*) = 2V_{ij}^2 + V_{ii}V_{jj}. \quad (4.8)$$

Let the true trait distribution be $P(\boldsymbol{\theta}_0)$, and the fitted model space as $\mathcal{Q} = \{Q(\boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta} = [0,1] \times [0,\infty)\}$. If we let $m \to \infty$ (consequently $N = mn \to \infty$), the MLEs from fitting the

wrong model, $\hat{\boldsymbol{\theta}} = (\hat{h}^2, \hat{\sigma}^2)$, will converge in probability to $\boldsymbol{\theta}_1 = (h_1^2, \sigma_1^2)$ that minimizes the Kullback-Leibler divergence between $P(\boldsymbol{\theta}_0)$ and $Q(\boldsymbol{\theta})$ over the parameter space $\boldsymbol{\Theta}$. That is,

$$\hat{\boldsymbol{\theta}} \to_p \boldsymbol{\theta}_1 = \arg\min_{\boldsymbol{\theta}} \frac{1}{N} KL\Big(P_m(\boldsymbol{\theta}_0), Q_m(\boldsymbol{\theta})\Big) \qquad \text{as } m \to \infty, \tag{4.9}$$

where $P_m(\cdot)$ and $Q_m(\cdot)$ represent joint distributions over all m pedigrees.

Since $\boldsymbol{G}_f$ may not have a block-diagonal structure, we can no longer group eigenvalues of $\boldsymbol{G}_f$ by pedigree. The first order partial derivatives of the log-likelihood function (4.4) are

$$\begin{aligned}
\frac{\partial \ell}{\partial h^2} &= -\frac{1}{2}\Big[\sum_{i=1}^{N} \frac{\lambda_i - 1}{h^2\lambda_i + 1 - h^2} - \frac{1}{\sigma^2}\sum_{i=1}^{N} \frac{y_i^{*2}(\lambda_i - 1)}{(h^2\lambda_i + 1 - h^2)^2}\Big], \\
\frac{\partial \ell}{\partial \sigma^2} &= -\frac{1}{2}\Big[\frac{N}{\sigma^2} - \frac{1}{\sigma^4}\sum_{i=1}^{N} \frac{y_i^{*2}}{h^2\lambda_i + 1 - h^2}\Big].
\end{aligned}$$

The second order partial derivatives are

$$\frac{\partial^2 \ell}{\partial h^4} = \frac{1}{2}\Big[\sum_{i=1}^{N} \frac{(\lambda_i - 1)^2}{(h^2\lambda_i + 1 - h^2)^2} - \frac{2}{\sigma^2}\sum_{i=1}^{N} \frac{y_i^{*2}(\lambda_i - 1)^2}{(h^2\lambda_i + 1 - h^2)^3}\Big], \tag{4.10}$$

$$\frac{\partial^2 \ell}{\partial h^2 \partial \sigma^2} = -\frac{1}{2\sigma^4}\sum_{i=1}^{N} \frac{y_i^{*2}(\lambda_i - 1)}{(h^2\lambda_i + 1 - h^2)^2}, \tag{4.11}$$

$$\frac{\partial^2 \ell}{\partial \sigma^4} = \frac{1}{2}\Big[\frac{N}{\sigma^4} - \frac{2}{\sigma^6}\sum_{i=1}^{N} \frac{y_i^{*2}}{h^2\lambda_i + 1 - h^2}\Big]. \tag{4.12}$$

The covariance matrices of $\boldsymbol{y}$ (prior to eigen-transformation) under the true and the fitted models are

$$\begin{aligned}
\boldsymbol{\Omega}_P(\boldsymbol{\theta}_0) &= [h_0^2 \cdot \boldsymbol{G}_t + (1 - h_0^2)\boldsymbol{I}]\sigma_0^2 \\
&= [h_0^2 \cdot \boldsymbol{G}_f + (1 - h_0^2)\boldsymbol{I}]\sigma_0^2 + h_0^2\sigma_0^2\boldsymbol{\Delta}, \\
\boldsymbol{\Omega}_Q(\boldsymbol{\theta}) &= [h^2 \cdot \boldsymbol{G}_f + (1 - h^2)\boldsymbol{I}]\sigma^2.
\end{aligned}$$

The KL divergence of (4.9) is then

$$
\begin{aligned}
&\frac{1}{N} KL\Big(P_m(\boldsymbol{\theta}_0), Q_m(\boldsymbol{\theta})\Big) \\
=& -\frac{1}{2N}\ln\left|\boldsymbol{T}^T\boldsymbol{\Omega}_P(\boldsymbol{\theta}_0)\boldsymbol{T}\right| - \frac{1}{2N}\mathbb{E}_{P(\boldsymbol{\theta}_0)}\Big[\boldsymbol{y}^{*T}\big(\boldsymbol{T}^T\boldsymbol{\Omega}_P(\boldsymbol{\theta}_0)\boldsymbol{T}\big)^{-1}\boldsymbol{y}^*\Big] \\
&+ \frac{1}{2N}\ln\left|\boldsymbol{T}^T\boldsymbol{\Omega}_Q(\boldsymbol{\theta})\boldsymbol{T}\right| + \frac{1}{2N}\mathbb{E}_{P(\boldsymbol{\theta}_0)}\Big[\boldsymbol{y}^{*T}\big(\boldsymbol{T}^T\boldsymbol{\Omega}_Q(\boldsymbol{\theta})\boldsymbol{T}\big)^{-1}\boldsymbol{y}^*\Big] \\
=& \, c + \frac{1}{2}\ln(\sigma^2) + \frac{1}{2N}\sum_{i=1}^{N}\ln(h^2\lambda_i + 1 - h^2) \\
&+ \frac{1}{2N}\operatorname{tr}\left[\Big([h_0^2\boldsymbol{D} + (1-h_0^2)\boldsymbol{I}]\sigma_0^2 + h_0^2\sigma_0^2\boldsymbol{T}^T\boldsymbol{\Delta}\boldsymbol{T}\Big)\Big([h^2\boldsymbol{D} + (1-h^2)\boldsymbol{I}]\sigma^2\Big)^{-1}\right] \\
=& \, c + \frac{1}{2}\ln(\sigma^2) + \frac{1}{2N}\sum_{i=1}^{N}\ln(h^2\lambda_i + 1 - h^2) \\
&+ \frac{\sigma_0^2}{2N\sigma^2}\sum_{i=1}^{N}\frac{h_0^2\lambda_i + 1 - h_0^2}{h^2\lambda_i + 1 - h^2} + \frac{h_0^2\sigma_0^2}{2N\sigma^2}\sum_{i=1}^{N}\frac{\boldsymbol{t}_i^T\boldsymbol{\Delta}\boldsymbol{t}_i}{h^2\lambda_i + 1 - h^2},
\end{aligned}
\tag{4.13}
$$

where c is an ignorable constant, λ_i and $\boldsymbol{t}_i$ are the ith eigenvalue and eigenvector of the fitted correlation matrix $\boldsymbol{G}_f$. Minimizing (4.13) with respect to $\boldsymbol{\theta}$ for large m leads to

$$
(h_1^2, \sigma_1^2) \approx \underset{(h^2,\sigma^2)}{\arg\min}\, \frac{1}{2}\ln(\sigma^2) + \frac{1}{2N}\sum_{i=1}^{N}\left[\ln(h^2\lambda_i + 1 - h^2) + \frac{(h_0^2\lambda_i + 1 - h_0^2)\sigma_0^2 + h_0^2\sigma_0^2\boldsymbol{t}_i^T\boldsymbol{\Delta}\boldsymbol{t}_i}{(h^2\lambda_i + 1 - h^2)\sigma^2}\right],
\tag{4.14}
$$

It follows from likelihood theory that

$$
\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_1) \to_d N\Big(\boldsymbol{0}, \boldsymbol{J}^{-1}\boldsymbol{K}(\boldsymbol{J}^{-1})^T\Big),
\tag{4.15}
$$

with

$$
\boldsymbol{J} = \lim_{m\to\infty}\mathbb{E}_{P(\boldsymbol{\theta}_0)}\left[\frac{\partial^2\ell(\boldsymbol{\theta};\boldsymbol{y}^*)/N}{\partial\boldsymbol{\theta}^2}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_1}\right],
\tag{4.16}
$$

$$
\boldsymbol{K} = \lim_{m\to\infty}\mathbb{E}_{P(\boldsymbol{\theta}_0)}\left[\frac{\partial\ell(\boldsymbol{\theta};\boldsymbol{y}^*)/N}{\partial\boldsymbol{\theta}} \times \Big(\frac{\partial\ell(\boldsymbol{\theta};\boldsymbol{y}^*)/N}{\partial\boldsymbol{\theta}}\Big)^T\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_1}\right],
\tag{4.17}
$$

where $\ell(\boldsymbol{\theta};\boldsymbol{y}^*)$ is the log-likelihood over all data. Combining results from Equation (4.10),

(4.11) and (4.12), elements of the matrix $\boldsymbol{J}$ in (4.16) before taking limits are

$$\begin{aligned}
&\mathbb{E}_{P(\boldsymbol{\theta}_0)}\left[\frac{\partial^2 \ell(h^2,\sigma^2;\boldsymbol{y}^*)/N}{\partial h^4}\Big|_{h_1^2,\sigma_1^2}\right] \\
&\quad = \frac{1}{2N}\sum_{i=1}^{N}\left[\frac{(\lambda_i-1)^2}{(h_1^2\lambda_i+1-h_1^2)^2} - \frac{2(\lambda_i-1)^2\cdot\mathbb{E}[y_i^{*2}]}{\sigma_1^2(h_1^2\lambda_i+1-h_1^2)^3}\right], \qquad (4.18)\\
&\mathbb{E}_{P(\boldsymbol{\theta}_0)}\left[\frac{\partial^2 \ell(h^2,\sigma^2;\boldsymbol{y}^*)/N}{\partial h^2\partial\sigma^2}\Big|_{h_1^2,\sigma_1^2}\right] \\
&\quad = -\frac{1}{2N}\sum_{i=1}^{N}\frac{(\lambda_i-1)\cdot\mathbb{E}[y_i^{*2}]}{\sigma_1^4(h_1^2\lambda_i+1-h_1^2)^2}, \qquad (4.19)\\
&\mathbb{E}_{P(\boldsymbol{\theta}_0)}\left[\frac{\partial^2 \ell(h^2,\sigma^2;\boldsymbol{y}^*)/N}{\partial \sigma^4}\Big|_{h_1^2,\sigma_1^2}\right] \\
&\quad = \frac{1}{2N}\sum_{i=1}^{N}\left[\frac{1}{\sigma_1^4} - \frac{2\cdot\mathbb{E}[y_i^{*2}]}{\sigma_1^6(h_1^2\lambda_i+1-h_1^2)}\right]. \qquad (4.20)
\end{aligned}$$

Similarly, elements of the matrix $\boldsymbol{K}$ in (4.17) before taking limits are

$$\begin{aligned}
&\mathbb{E}_{P(\boldsymbol{\theta}_0)}\left[\left(\frac{\partial^2 \ell(h^2,\sigma^2;\boldsymbol{y}^*)/N}{\partial h^2}\right)^2\Big|_{h_1^2,\sigma_1^2}\right] \\
&\quad = \frac{1}{4N^2}\left[\Big(\sum_{i=1}^{N}\frac{\lambda_i-1}{h_1^2\lambda_i+1-h_1^2}\Big)^2 + \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{(\lambda_i-1)(\lambda_j-1)\cdot\mathbb{E}[y_i^{*2}y_j^{*2}]}{\sigma_1^4(h_1^2\lambda_i+1-h_1^2)^2(h_1^2\lambda_j+1-h_1^2)^2}\right. \\
&\qquad \left. - \frac{2}{\sigma_1^2}\Big(\sum_{i=1}^{N}\frac{\lambda_i-1}{h_1^2\lambda_i+1-h_1^2}\Big)\Big(\sum_{i=1}^{N}\frac{(\lambda_i-1)\cdot\mathbb{E}[y_i^{*2}]}{(h_1^2\lambda_i+1-h_1^2)^2}\Big)\right], \qquad (4.21)\\
&\mathbb{E}_{P(\boldsymbol{\theta}_0)}\left[\frac{\partial^2 \ell(h^2,\sigma^2;\boldsymbol{y}^*)/N}{\partial h^2}\times\frac{\partial^2 \ell(h^2,\sigma^2;\boldsymbol{y}^*)/N}{\partial \sigma^2}\Big|_{h_1^2,\sigma_1^2}\right] \\
&\quad = \frac{1}{4N^2}\left[\frac{N}{\sigma_1^2}\sum_{i=1}^{N}\frac{\lambda_i-1}{h_1^2\lambda_i+1-h_1^2} - \frac{1}{\sigma_1^4}\Big(\sum_{i=1}^{N}\frac{\lambda_i-1}{h_1^2\lambda_i+1-h_1^2}\Big)\Big(\sum_{i=1}^{N}\frac{\mathbb{E}[y_i^{*2}]}{h_1^2\lambda_i+1-h_1^2}\Big)\right. \\
&\qquad \left. - \frac{N}{\sigma_1^4}\sum_{i=1}^{N}\frac{(\lambda_i-1)\cdot\mathbb{E}[y_i^{*2}]}{(h_1^2\lambda_i+1-h_1^2)^2} + \frac{1}{\sigma_1^6}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{(\lambda_i-1)\cdot\mathbb{E}[y_i^{*2}y_j^{*2}]}{(h_1^2\lambda_i+1-h_1^2)^2(h_1^2\lambda_j+1-h_1^2)}\right], \qquad (4.22)\\
&\mathbb{E}_{P(\boldsymbol{\theta}_0)}\left[\left(\frac{\partial^2 \ell(h^2,\sigma^2;\boldsymbol{y}^*)/N}{\partial \sigma^2}\right)^2\Big|_{h_1^2,\sigma_1^2}\right] \\
&\quad = \frac{1}{4N^2}\left[\frac{N^2}{\sigma_1^4} + \frac{1}{\sigma_1^8}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\mathbb{E}[y_i^{*2}y_j^{*2}]}{(h_1^2\lambda_i+1-h_1^2)(h_1^2\lambda_j+1-h_1^2)} - \frac{2N}{\sigma_1^6}\sum_{i=1}^{N}\frac{\mathbb{E}[y_i^{*2}]}{h_1^2\lambda_i+1-h_1^2}\right], \qquad (4.23)
\end{aligned}$$

where $\mathbb{E}[y_i^{*2}]$ and $\mathbb{E}[y_i^{*2}y_j^{*2}]$ are given by (4.7) and (4.8). Given h_0^2, σ_0^2, $\boldsymbol{G}_t$ and $\boldsymbol{G}_f$ for any finite m, (4.14) can be solved numerically for $\boldsymbol{\theta}_1 = (h_1^2, \sigma_1^2)$. $\boldsymbol{J}$ and $\boldsymbol{K}$ in (4.16) and (4.17) can be computed subsequently to obtain the asymptotic variance covariance matrix of the MLEs.

For the special case where $\boldsymbol{G}_f = 2\boldsymbol{\Psi}$, twice the pedigree kinship matrix, elements of $\boldsymbol{\Delta}$ represent deviations of twice the realized kinship (over causal genome contributing to $\boldsymbol{G}_t$) from its pedigree expectation. Such deviations have expectation 0 when observations are not ascertained by trait or by IBD sharing. Since the pedigree kinship matrix is fixed given the pedigree structure, $\boldsymbol{G}_f$ will have a block diagonal structure with identical blocks. Each block has the same set of eigenvalues, λ_{ij}'s and eigenvectors, $\boldsymbol{t}_{ij}$'s. In addition, $\boldsymbol{\Delta}$ will also have a block diagonal structure with $\boldsymbol{\Delta}_i$'s on the diagonal. This means $\boldsymbol{\Delta}$ enters (4.14), (4.16) and (4.17) only in a factored form. For example, the term in (4.14) that involves $\boldsymbol{\Delta}$ becomes

$$\begin{aligned}
&\frac{1}{2N}\sum_{i=1}^{N}\frac{(h_0^2\lambda_i + 1 - h_0^2)\sigma_0^2 + h_0^2\sigma_0^2\boldsymbol{t}_i^T\boldsymbol{\Delta}\boldsymbol{t}_i}{(h^2\lambda_i + 1 - h^2)\sigma^2} \\
&= \frac{1}{2mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(h_0^2\lambda_{ij} + 1 - h_0^2)\sigma_0^2 + h_0^2\sigma_0^2\boldsymbol{t}_{ij}^T\boldsymbol{\Delta}_i\boldsymbol{t}_{ij}}{(h^2\lambda_{ij} + 1 - h^2)\sigma^2} \\
&= \frac{1}{2n}\sum_{j=1}^{n}\frac{(h_0^2\lambda_{ij} + 1 - h_0^2)\sigma_0^2 + h_0^2\sigma_0^2\boldsymbol{t}_{ij}^T(\sum_{i=1}^{m}\boldsymbol{\Delta}_i/m)\boldsymbol{t}_{ij}}{(h^2\lambda_{ij} + 1 - h^2)\sigma^2}.
\end{aligned}$$

Since $\boldsymbol{\Delta}_i$'s are $n \times n$ matrices whose elements have expectation 0 over realizations of descent on the pedigree structure, $\boldsymbol{t}_{ij}^T(\sum_{i=1}^{m}\boldsymbol{\Delta}_i/m)\boldsymbol{t}_{ij}$ converges to 0 in probability as $m \to \infty$. The same can be shown for (4.16) and (4.17): the impact of $\boldsymbol{\Delta}$ diminishes as $m \to \infty$. This implies fitting model (4.1) with $\boldsymbol{G}_f = 2\boldsymbol{\Psi}$, regardless what kinship information is captured in $\boldsymbol{G}_t$, produces consistent MLEs. In addition, the ASV of the MLEs will be the same as if $\boldsymbol{G}_t = \boldsymbol{G}_f = 2\boldsymbol{\Psi}$, which can be easily computed using (4.5).

4.3 Mis-identification of causal genome

One possibility for $\boldsymbol{G}_t \neq \boldsymbol{G}_f$ in practice is that we mis-identified the causal genome (part of genome that has an effect on the trait) for model fitting. For example, when $\boldsymbol{G}_t = 2\boldsymbol{\Phi}$, twice

the genome-wide realized kinship matrix but we fit $\boldsymbol{G}_f = 2\boldsymbol{\Psi}$, twice the pedigree kinship matrix. In this section we investigate the impact of such mis-specification on heritability estimation when using different study designs and combinations of $(\boldsymbol{G}_t, \boldsymbol{G}_f)$ that represent different extent of mis-specification. We consider four different types of designs: sibpairs, sibships of 14, three-generation pedigree of 14 members , and the Cleopatra pedigree of 14 members introduced in Section 1.4.2. Three types of kinship measures are used for $\boldsymbol{G}_t$ and $\boldsymbol{G}_f$: twice the pedigree kinship matrix ($2\boldsymbol{\Psi}$), twice the genome-wide realized kinship matrix ($2\boldsymbol{\Phi}$), and twice the realized kinship matrix on chromosome 22 ($2\boldsymbol{\Phi}^*$). These choices represent increasing amount of variation around the pedigree expectation. All 9 combinations of $(\boldsymbol{G}_t, \boldsymbol{G}_f)$ are considered and the correlation matrices are assumed known in the analysis. We set total phenotypic variance $\sigma_0^2 = 1$ throughout, and use three heritability values, $h_0^2 \in \{0.2, 0.5, 0.8\}$. In total, there are 108 simulation scenarios (4× designs, 9× $(\boldsymbol{G}_t, \boldsymbol{G}_f)$ combinations, and 3× heritability values).

The total sample size of the study is kept at $N = 1400$ (eg., 700 sibpairs, or 100 sibships of 14). 500 simulation replicates are used to capture the empirical distributions of the MLEs in each of the 108 scenarios. Within each simulation replicate, we simulate $(\boldsymbol{G}_t, \boldsymbol{G}_f)$ given the design, simulate trait data given $\boldsymbol{G}_t$ and h_0^2, and then fit model (4.1) with $\boldsymbol{G}_f$ to obtain the MLEs. Mean and standard deviation of the empirical distribution of $\hat{h}^2$ are compared to those computed from (4.14) and (4.15) based on $(\boldsymbol{G}_t, \boldsymbol{G}_f)$. Simulations were performed using the R package `rres` (Wang 2018). Twice the genome-wide realized kinship between each pair of individuals was measured as the proportion of shared genome (in genetic distances) over the 22 autosomes.

There is generally very good agreement between the empirical and analytical results. Table 4.1 displays the results from the 9 combinations of $(\boldsymbol{G}_t, \boldsymbol{G}_f)$, for the sibship design and $h_0^2 = 0.5$. We see that $\hat{h}^2$ appears to be unbiased when $\boldsymbol{G}_t = \boldsymbol{G}_f$, and when $\boldsymbol{G}_t \neq \boldsymbol{G}_f = 2\boldsymbol{\Psi}$. In addition, $\boldsymbol{G}_f = 2\boldsymbol{\Psi}$ resulted in very similar $\mathrm{SE}(\hat{h}^2)$ regardless of $\boldsymbol{G}_t$, as predicted in the

$\boldsymbol{G}_t$		$2\boldsymbol{\Psi}$			$2\boldsymbol{\Phi}$			$2\boldsymbol{\Phi}^*$		
$\boldsymbol{G}_f$		$2\boldsymbol{\Psi}$	$2\boldsymbol{\Phi}$	$2\boldsymbol{\Phi}^*$	$2\boldsymbol{\Psi}$	$2\boldsymbol{\Phi}$	$2\boldsymbol{\Phi}^*$	$2\boldsymbol{\Psi}$	$2\boldsymbol{\Phi}$	$2\boldsymbol{\Phi}^*$
$\hat{h}^2$	emp.	0.494	0.451	0.278	0.497	0.498	0.281	0.492	0.486	0.499
	analy.	0.500	0.453	0.279	0.500	0.500	0.282	0.498	0.490	0.500
$\mathrm{SE}(\hat{h}^2)$	emp.	0.063	0.052	0.032	0.065	0.060	0.032	0.064	0.059	0.032
	analy.	0.067	0.055	0.032	0.067	0.061	0.032	0.067	0.061	0.033

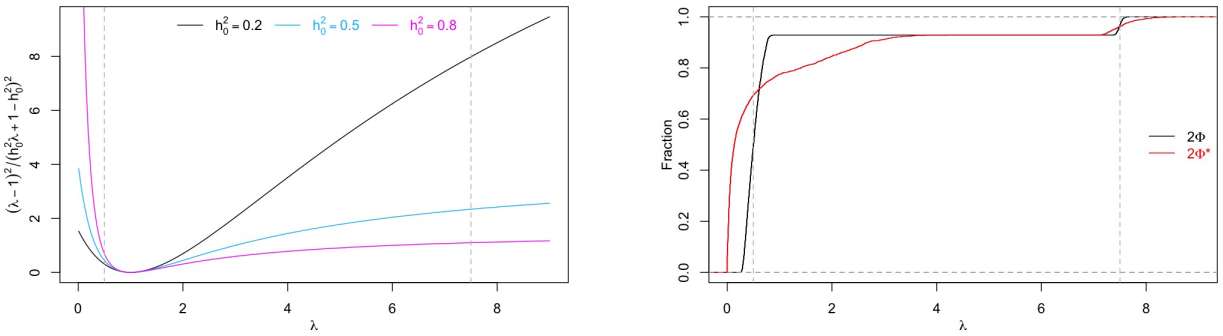
Table 4.1: Comparison of limits and asymptotic sampling standard errors of heritability estimates obtained by fitting model (4.1) and by using (4.14) and (4.15), with data on 100 independent sibships of 14. True parameters are $(h_0^2, \sigma_0^2) = (0.5, 1)$. Empirical distribution of $\hat{h}^2$ is constructed from 500 simulation replicates.

previous section. When $\boldsymbol{G}_t \neq \boldsymbol{G}_f$ and $\boldsymbol{G}_f$ represents a more variable kinship measure than $\boldsymbol{G}_t$ (in this case, realized kinship on a shorter causal genome), both the point estimates and sampling errors tend to be smaller than expected under correct model specification.

When $\boldsymbol{G}_t = \boldsymbol{G}_f$, equation (4.5) suggests that pedigree structure affects $\mathrm{ASV}(\hat{h}^2)$ only through eigenvalues of $\boldsymbol{G}_t$. For large N, the first term in the denominator of (4.5),

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(\lambda_{ij}-1)^2}{(h_0^2\lambda_{ij}+1-h_0^2)^2}, \tag{4.24}$$

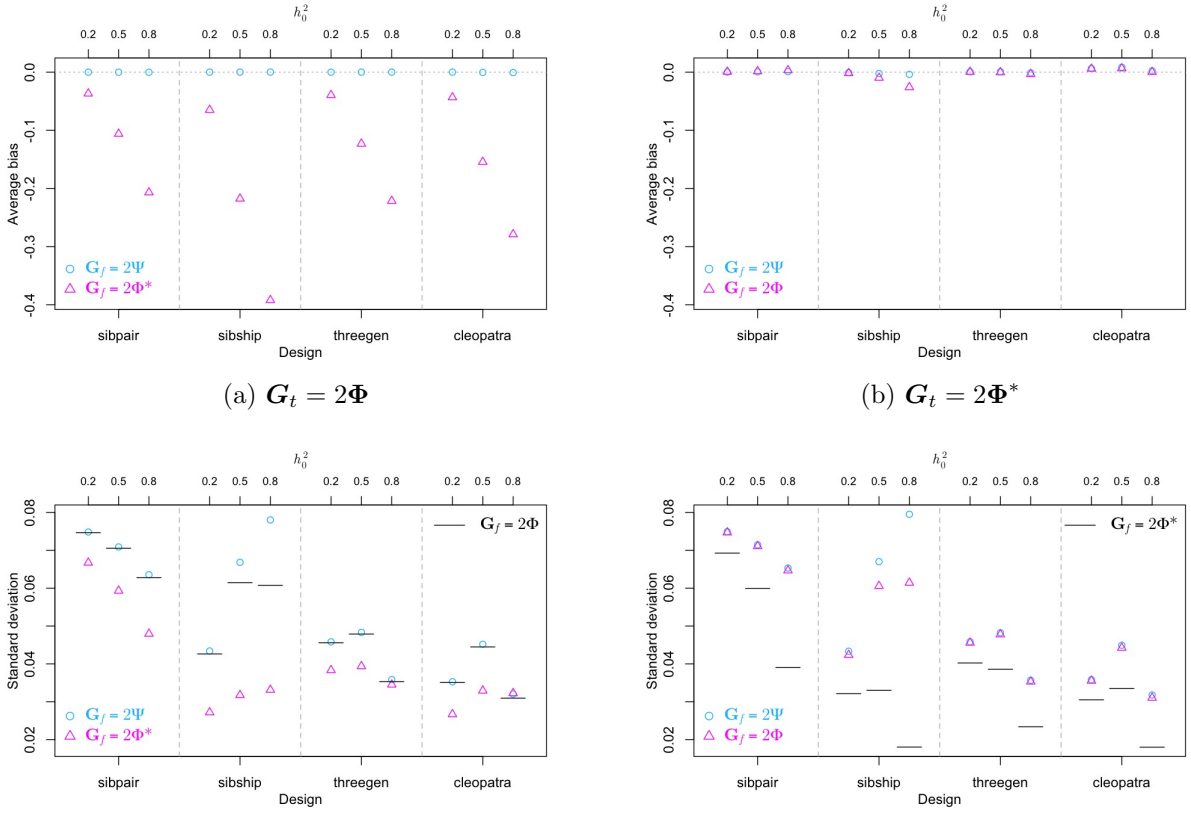
dominates the second term. Since the summand in (4.24) is non-negative, eigenvalues that lead to bigger summand values have a higher tendency to reduce $\mathrm{ASV}(\hat{h}^2)$. Figure 4.1a suggests that extreme eigenvalues have biggest impact in reducing sampling variance. Figure 4.1b shows that twice the realized kinship matrices on shorter genomic segments tend to produce more extreme eigenvalues. This provides intuition on why $\mathrm{SE}(\hat{h}^2)$ is much smaller when $\boldsymbol{G}_t = \boldsymbol{G}_f = 2\boldsymbol{\Phi}^*$, as compared to the other cases of correct model specification shown in Table 4.1.



(a) (b)

Figure 4.1: **(a)** $(\lambda-1)^2/(h_0^2\lambda+1-h_0^2)^2$ evaluated at a range of λ values and $h_0 \in \{0.2, 0.5, 0.8\}$. **(b)** Cumulative distributions of eigenvalues of $2\boldsymbol{\Phi}$ and $2\boldsymbol{\Phi}^*$ from 100 simulated sibships of 14. Vertical dashed lines in both plots represent the two distinct eigenvalues of $2\boldsymbol{\Psi}$: 0.5 and 7.5.

Figure 4.2 shows simulation results from use of other designs and heritability values. We omit scenarios where $\boldsymbol{G}_t = 2\boldsymbol{\Psi}$ since it cannot be true biologically. Most of the observations from earlier discussion of the sibship design also hold true for other designs. The one clear exception is that when $\boldsymbol{G}_t = 2\boldsymbol{\Phi}$ and $\boldsymbol{G}_f = 2\boldsymbol{\Psi}$, the impact of this specific model mis-specification on SE($\hat{h}^2$) is very small when using other designs (compare blue circles to black bars in each scenario in Figure 4.2c). This is likely because the distribution of eigenvalues of the true correlation matrix depends on the length of the causal genome as well as pedigree structure. For sibship design, a good proportion of eigenvalues associated with $2\boldsymbol{\Phi}$ are close to 0 (Figure 4.1b), which is not the case for other designs (results not shown). As a contrast, when $\boldsymbol{G}_t = 2\boldsymbol{\Phi}^*$ and $\boldsymbol{G}_f = 2\boldsymbol{\Psi}$, the impact of model mis-specification on SE($\hat{h}^2$) is clear for all designs (compare blue circles to black bars in each scenario in Figure 4.2d). This is because the greater variation in $2\boldsymbol{\Phi}^*$ due to shorter length of causal genome leads to the possibility of more extreme eigenvalues for each study design.



(a) $\boldsymbol{G}_t = 2\boldsymbol{\Phi}$ (b) $\boldsymbol{G}_t = 2\boldsymbol{\Phi}^*$

(c) $\boldsymbol{G}_t = 2\boldsymbol{\Phi}$ (d) $\boldsymbol{G}_t = 2\boldsymbol{\Phi}^*$

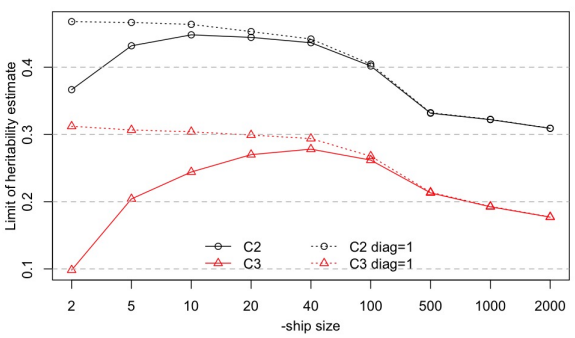
Figure 4.2: Simulation results for all four designs, three values of h_0^2 and various combinations of $(\boldsymbol{G}_t, \boldsymbol{G}_f)$. **(a)** and **(b)** show the average bias of heritability estimates from 500 simulation replicates, when $\boldsymbol{G}_t = 2\boldsymbol{\Phi}$ and $\boldsymbol{G}_t = 2\boldsymbol{\Phi}^*$ respectively. **(c)** and **(d)** show the standard deviation of heritability estimates from 500 simulation replicates, when $\boldsymbol{G}_t = 2\boldsymbol{\Phi}$ and $\boldsymbol{G}_t = 2\boldsymbol{\Phi}^*$ respectively. In both **(c)** and **(d)**, results obtained under correct model specification are shown in black bars as references.

Lastly, we note that for any given combination of $(\boldsymbol{G}_t, \boldsymbol{G}_f)$, $\mathrm{SE}(\hat{h}^2)$ varies with h_0^2 quite differently across designs. For example, when $\boldsymbol{G}_t = \boldsymbol{G}_f = 2\boldsymbol{\Phi}$ the sibship design has slightly smaller $\mathrm{SE}(\hat{h}^2)$ than the three-generation design at $h_0^2 = 0.2$, but it is the opposite when $h_0^2 = 0.8$.

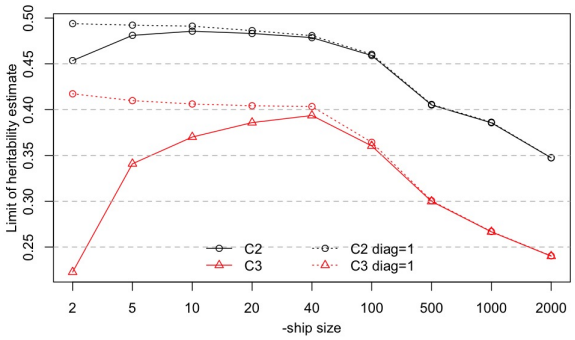
4.4 Errors in estimation of realized kinship

Another main reason for $\boldsymbol{G}_t \neq \boldsymbol{G}_f$ is that $\boldsymbol{G}_f$ is often an estimated quantity in practice. For example, $\boldsymbol{G}_t = 2\boldsymbol{\Phi}$, twice the genome-wide realized kinship matrix, and $\boldsymbol{G}_f = 2\widehat{\boldsymbol{\Phi}}$ using some choice of kinship estimator. We have shown in Chapter 2 that different estimators of realized kinship differ greatly in estimation accuracy. We have also seen in the previous section that $\boldsymbol{G}_t \neq \boldsymbol{G}_f$ often leads to downward bias in heritability estimation. These observations suggest that accuracy of kinship estimation, among many other factors, may contribute to the "missing heritability" problem (see, for example, Maher (2008)). It is of particular interest to investigate the impact of using the classic GRM as $\boldsymbol{G}_f$ in a population-based design, an approach that has become widely popular in recent years.

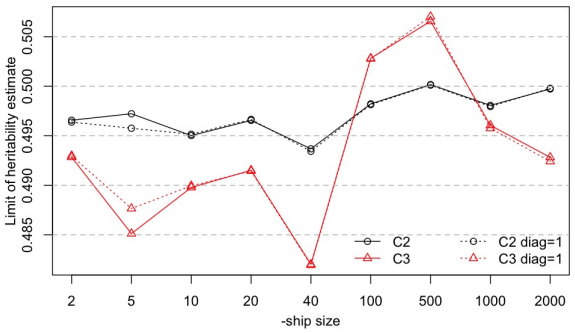
For ease of simulation, we use multiple independent pedigrees of second or third cousinships of various sizes. This block-diagonal structure might approximate that of a population sample that included clusters of more closely related individuals. This also allows us to use (4.14) to compute asymptotic bias. Total sample size is kept at 2000 across different simulation conditions, so that in one of the conditions, for example, the sample consists of 200 independent third cousinships of 10. We consider three different kinship measures: twice the genome-wide realized kinship matrix ($2\boldsymbol{\Phi}$), classic GRM ($2\widehat{\boldsymbol{\Phi}}_C$), and LD (linkage disequilibrium) weighted GRM ($2\widehat{\boldsymbol{\Phi}}_W$) introduced in Chapter 2. Here the LD weighted GRM represents a more accurate estimator of realized kinship matrix than the classic GRM. For the two GRMs, we look at the cases both with and without constraining the diagonal terms (twice the self-kinship) to 1. True parameter values are set to $h_0^2 = 0.5$ and $\sigma_0^2 = 1$. Inheritance as well as genetic marker data were simulated on pedigrees using the R package `rres`. We use the K170 marker panel introduced in Section 1.4.1 and the 1000 Genome (The 1000 Genomes Project Consortium 2015) haplotypes as founder haplotypes. Allele frequencies are assumed known for computing $2\widehat{\boldsymbol{\Phi}}_C$ and $2\widehat{\boldsymbol{\Phi}}_W$.



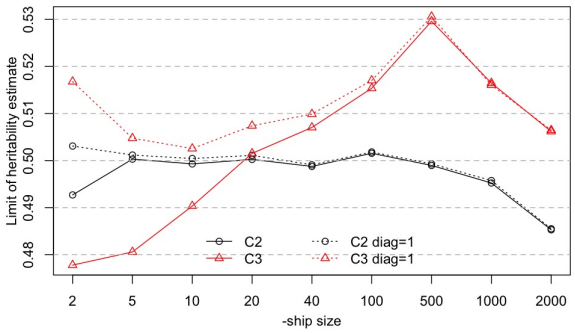
(a) $\boldsymbol{G}_t = 2\boldsymbol{\Phi}$, $\boldsymbol{G}_f = 2\widehat{\boldsymbol{\Phi}}_C$.



(b) $\boldsymbol{G}_t = 2\boldsymbol{\Phi}$, $\boldsymbol{G}_f = 2\widehat{\boldsymbol{\Phi}}_W$.



(c) $\boldsymbol{G}_t = 2\widehat{\boldsymbol{\Phi}}_C$, $\boldsymbol{G}_f = 2\boldsymbol{\Phi}$.



(d) $\boldsymbol{G}_t = 2\widehat{\boldsymbol{\Phi}}_C$, $\boldsymbol{G}_f = 2\widehat{\boldsymbol{\Phi}}_W$.

Figure 4.3: Limits of $\hat{h}^2$ when $h_0^2 = 0.5$ and $\boldsymbol{G}_t \neq \boldsymbol{G}_f$. Each sample contains multiple independent second (C2) or third (C3) cousinships of specific sizes. Three kinship measures are considered: genome-wide realized kinship matrix ($2\boldsymbol{\Phi}$), classic GRM ($2\widehat{\boldsymbol{\Phi}}_C$) and LD weighted GRM ($2\widehat{\boldsymbol{\Phi}}_W$). The estimated matrices ($2\widehat{\boldsymbol{\Phi}}_C$ and $2\widehat{\boldsymbol{\Phi}}_W$) are used with and without constraining the diagonal terms to 1.

There is generally more bias in heritability estimates when the sample consists of third cousinships as opposed to second cousinships (Figure 4.3). This is in good agreement with the findings of Chapter 2, where we showed it is relatively more difficult to estimate realized kinship between more remote relatives. This observed association between level of bias and remoteness of relationship is further verified by additional simulations using full sibships, half sibships and first cousinships (results not shown).

Another observation from Figure 4.3 is that the decision to constrain the diagonal terms of $2\widehat{\mathbf{\Phi}}_C$ or $2\widehat{\mathbf{\Phi}}_W$ to 1 makes a big difference when pedigree sizes are small, but that difference erodes when pedigree sizes increase. A possible explanation is that apart from its effect on eigenvalues and eigenvectors in (4.14), $\boldsymbol{G}_t \neq \boldsymbol{G}_f$ induces bias in heritability estimates most directly from the matrix of difference, $\mathbf{\Delta}$. When we constrain the diagonal terms of $2\widehat{\mathbf{\Phi}}_C$ or $2\widehat{\mathbf{\Phi}}_W$ to 1, $\mathbf{\Delta}$ will have 0's on the diagonal. This is expected to have a bigger impact when pedigree sizes are small and the diagonal terms make up a bigger proportion of non-zero terms of $\mathbf{\Delta}$.

Figure 4.3a and 4.3b shows that when $\boldsymbol{G}_t = 2\mathbf{\Phi}$, setting $\boldsymbol{G}_f = 2\widehat{\mathbf{\Phi}}_W$ produces less bias than $\boldsymbol{G}_f = 2\widehat{\mathbf{\Phi}}_C$ in all cases. This is expected since the LD weighted GRM is a more accurate estimator of genome-wide realized kinship. When constraining the diagonal terms of twice the estimated kinship matrices to 1, bias increases with pedigree size for both second and third cousinships. Again, this is likely the result of the composition of non-zero terms in $\mathbf{\Delta}$: larger pedigree sizes imply higher proportion of non-zero terms. When not constraining the diagonal terms of twice the estimated kinship matrices to 1, bias increases and then decreases with pedigree size. A possible explanation is that in smaller cousinships, the diagonal terms of $\mathbf{\Delta}$ take up a bigger proportion of all non-zero terms. While self-kinship can be estimated relatively accurately compare to kinship between remote cousins, the magnitude of deviations on the original scale, as captured in $\mathbf{\Delta}$, is greater for self-kinship (see Chapter 2). This means the diagonal terms of $\mathbf{\Delta}$ could be more influential than off-diagonal terms on a per-entry basis, which lead to the bigger bias observed for smaller pedigrees.

Figure 4.3c and 4.3d show limits of heritability estimates when $\boldsymbol{G}_t = 2\widehat{\mathbf{\Phi}}_C$, and $\boldsymbol{G}_f = 2\mathbf{\Phi}$ or $2\widehat{\mathbf{\Phi}}_W$. In both figures, the main observation is that the scale of bias is very small compare to that in Figure 4.3a and 4.3b. This matches our findings from the first simulation study: when $\boldsymbol{G}_t$ and $\boldsymbol{G}_f$ represent two kinship measures with zero expected differences (expectation

over realization of IBD on pedigree), we expect downward bias in heritability estimate if $\boldsymbol{G}_f$ represents a more variable kinship measure than $\boldsymbol{G}_t$, but not much bias the other way round (e.g., Figure 4.2a and 4.2b). We showed in Chapter 2 that both classic GRM and LD weighted GRM are unbiased estimators of genome-wide realized kinship under certain assumptions, and the unbiasedness property held very well in simulation studies. Since classic GRM is a less efficient estimator, it is a more variable measure of kinship compare to LD weighted GRM.

4.5 Discussion

For heritability estimation using a two-component random effects model, we have provided formulas for the limits and the asymptotic sampling variances of the MLEs. These formulas are applicable even when the wrong kinship matrix is used to capture additive genetic correlation. To simplify notations, we have assumed that the sample consists of multiple independent pedigrees of the same structure. This assumption is not required to derive formulas (4.4) through (4.18) in Section 4.2.2. Limiting results for the special case of $\boldsymbol{G}_f = 2\boldsymbol{\Psi}$ can be obtained alternatively by assuming a finite number of relationship types are observed repeatedly as $m \to \infty$.

Under correct specification of model (4.1), ASV of heritability estimates is a function of eigenvalues of the appropriate correlation matrix, which in turn depends on pedigree structure and variation in realized kinship. For any pedigree structure, realized kinship can be easily simulated and subsequently used in the formulas we presented to assess effectiveness of the pedigree in heritability estimation under the assumed model. This is beneficial to pedigree selection in study design.

There can be bias in heritability estimate if the correlation matrix has been mis-specified. The extent of the bias depends on both the true and the fitted correlation matrices, as well as the study design. When the true genetic correlation between individuals is captured by

genome-wide realized kinship, one should use an accurate estimator of realized kinship to compute the fitted correlation matrix in order to reduce downward bias in heritability estimate. This is especially important when the study sample contains many remotely related individuals.

The popular classic GRM is not an accurate estimator of realized kinship. When used in a population-based design, it can lead to substantial downward bias in heritability estimate as shown in our simulation study. The downward bias is expected to persist even if denser SNP panels are used to estimate realized kinship, as additional markers do not provide information without limit. In a follow up study, we found that increasing marker density by 4 times made little improvement in accuracy of the classic GRM estimator (results not shown). This choice of kinship estimator and study design can contribute to the "missing heritability" problem, among many other factors. On the other hand, using a more accurate kinship estimator to compute the fitted correlation matrix is more robust to mis-specification of correlation matrix.

Chapter 5

GENE MAPPING OF QUANTITATIVE TRAITS

In this chapter we look at the topic of gene mapping via variance component linkage analysis, where both local and genome-wide kinship between pairs of individuals are at the heart of the models we fit. This extends the discussion of the previous chapter in the sense that the trait model contains more than two random components. We are interested in how misspecification of kinship matrices affects the outcome of variance component linkage analysis, and how such impacts differ by study design.

5.1 Basic gene mapping model

The basic gene mapping model described (equation (1.2) of Section 1.3.2) is reproduced here

$$\boldsymbol{y} = \boldsymbol{a}_l + \boldsymbol{g} + \boldsymbol{e}, \tag{5.1}$$

where $\boldsymbol{y}$ are the mean-adjusted trait values, $\boldsymbol{a}_l \sim \mathcal{N}(\boldsymbol{0}, \sigma_l^2 \boldsymbol{G}_l)$ are the genetic effects attributed to locus l, $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \sigma_g^2 \boldsymbol{G}_g)$ are the residual genetic effects from all other QTLs, and $\boldsymbol{e} \sim \mathcal{N}(\boldsymbol{0}, \sigma_e^2 \boldsymbol{I})$ are the unique environmental effects. σ_l^2, σ_g^2 and σ_e^2 are unknown variance parameters to be estimated. At locus l, it is natural to use $\boldsymbol{G}_l = 2\boldsymbol{\phi}_l$, where $\boldsymbol{\phi}_l$ is the realized (actual) kinship matrix at that locus. To capture residual genetic effects, some choices for $\boldsymbol{G}_g$ include the twice the pedigree kinship matrix ($2\boldsymbol{\Psi}$), twice the genome-wide realized kinship matrix ($2\boldsymbol{\Phi}$), or twice the realized kinship matrix over the genome less a segment (e.g., a chromosome) containing locus l ($2\boldsymbol{\Phi}_{-l}$). The distribution of the trait is

$$\boldsymbol{y} \sim \mathcal{N}\Big(\boldsymbol{0}, \sigma_l^2 \boldsymbol{G}_l + \sigma_g^2 \boldsymbol{G}_g + \sigma_e^2 \boldsymbol{I}\Big) \tag{5.2}$$

Let $\boldsymbol{\theta} = (\sigma_l^2, \sigma_g^2, \sigma_e^2)$, the log-likelihood function is

$$\ell(\boldsymbol{\theta}; \boldsymbol{y}) = -\frac{1}{2} \ln |\boldsymbol{\Omega}| - \frac{1}{2} \boldsymbol{y}^T \boldsymbol{\Omega}^{-1} \boldsymbol{y}, \tag{5.3}$$

where

$$\boldsymbol{\Omega} = \mathrm{Var}(\boldsymbol{y}) = \sigma_l^2 \boldsymbol{G}_l + \sigma_g^2 \boldsymbol{G}_g + \sigma_e^2 \boldsymbol{I}.$$

In a typical genome scan, one will conduct the hypothesis test

$$H_n : \sigma_l^2 = 0 \qquad \text{vs.} \qquad H_l : \sigma_l^2 > 0,$$

at $l \in \{1, ..., L\}$ over the whole genome. Testing is typically done with a likelihood ratio (LOD score) test. If H_l is true, the likelihood ratio test statistic (LRT) will have an asymptotic non-central chi-squared distribution. When the true parameter $\sigma_{l,0}^2 > 0$,

$$\mathrm{LRT} = 2\Big[\ell(\hat{\boldsymbol{\theta}}_l; \boldsymbol{y}) - \ell(\hat{\boldsymbol{\theta}}_n; \boldsymbol{y})\Big] \rightarrow_d \chi_1^2(\xi), \tag{5.4}$$

where ξ is the non-centrality parameter. The expected likelihood ratio test statistic (ELRT) can be a useful statistic for analyzing the power of a linkage analysis (e.g. Raffa and Thompson 2016).

We assume throughout this chapter that the trait model described in (5.1) and (5.2) is true with variance parameters $\boldsymbol{\theta}_0 = (\sigma_{l,0}^2, \sigma_{g,0}^2, \sigma_{e,0}^2)$ all strictly positive. We assume there is a major causal linkage region centered at locus l_0 with local kinship matrix $\boldsymbol{\phi}_{l_0}$. The residual genetic effect is spread over the whole genome with corresponding genome-wide realized kinship matrix $\boldsymbol{\Phi}$. That is, $\boldsymbol{G}_{l,t} = 2\boldsymbol{\phi}_{l_0}$ and $\boldsymbol{G}_{g,t} = 2\boldsymbol{\Phi}$. When performing linkage tests across the genome, the fitted genetic correlation matrices $\boldsymbol{G}_{l,f}$ and/or $\boldsymbol{G}_{g,f}$ may be mis-specified, *i.e.*, $\boldsymbol{G}_{l,f} \neq \boldsymbol{G}_{l,t}$ and/or $\boldsymbol{G}_{g,f} \neq \boldsymbol{G}_{g,t}$. To study the effect of model mis-specification and pedigree structure on ELRT in a linkage analysis, we assume there are m mutually independent pedigrees of the same structure with finite pedigree size n.

5.2 Large sample approximation of ELRT

Under the model assumptions, the true distribution of the trait is

$$\boldsymbol{y} \sim \mathcal{N}\Big(\mathbf{0}, \sigma^2_{l,0} \cdot 2\boldsymbol{\phi}_{l_0} + \sigma^2_{g,0} \cdot 2\boldsymbol{\Phi} + \sigma^2_{e,0}\boldsymbol{I}\Big). \tag{5.5}$$

When fitting model (5.1) with the correct correlation matrices (*i.e.*, $\boldsymbol{G}_{l,f} = \boldsymbol{G}_{l,t}$, $\boldsymbol{G}_{g,f} = \boldsymbol{G}_{g,t}$), it follows from likelihood theory that the MLEs are consistent. If $\boldsymbol{G}_{l,f}$ and/or $\boldsymbol{G}_{g,f}$ are mis-specified, as discussed in Section 4.2.2, the MLEs from fitting a mis-specified model converge in probability to values that minimize the Kullback-Leibler divergence between the true model $P(\boldsymbol{\theta}_0)$ and the fitted model $Q(\boldsymbol{\theta})$ over the parameter space $\boldsymbol{\Theta}$. That is,

$$\hat{\boldsymbol{\theta}} \to_p \boldsymbol{\theta}_1 = \arg\min_{\boldsymbol{\theta}} \frac{1}{N} KL\Big(P_m(\boldsymbol{\theta}_0), Q_m(\boldsymbol{\theta})\Big) \qquad \text{as } m \to \infty, \tag{5.6}$$

where $N = mn$ is the total sample size, $P_m(\cdot)$ and $Q_m(\cdot)$ represent joint distributions over all m pedigrees.

Unlike in Chapter 4, having more than two random effects in the model means we can no longer use eigen-decomposition to simplify computations. Let

$$\begin{aligned}
\boldsymbol{\Omega}_P(\boldsymbol{\theta}_0) &= \sigma^2_{l,0}\boldsymbol{G}_{l,t} + \sigma^2_{g,0}\boldsymbol{G}_{g,t} + \sigma^2_{e,0}\boldsymbol{I}, \\
\boldsymbol{\Omega}_Q(\boldsymbol{\theta}) &= \begin{cases} \sigma^2_l\boldsymbol{G}_{l,f} + \sigma^2_g\boldsymbol{G}_{g,f} + \sigma^2_e\boldsymbol{I} & \text{under } H_l, \\ \sigma^2_g\boldsymbol{G}_{g,f} + \sigma^2_e\boldsymbol{I} & \text{under } H_n, \end{cases}
\end{aligned}$$

where it is understood that $\boldsymbol{\theta} = (0, \sigma^2_g, \sigma^2_e)$ under H_n. The formula of KL divergence in (5.6) takes the form

$$\begin{aligned}
&\frac{1}{N} KL\Big(P_m(\boldsymbol{\theta}_0), Q_m(\boldsymbol{\theta})\Big) \\
=& -\frac{1}{2N}\ln\big|\boldsymbol{\Omega}_P(\boldsymbol{\theta}_0)\big| - \frac{1}{2N}\mathbb{E}_{P(\boldsymbol{\theta}_0)}\Big[\boldsymbol{y}^T\boldsymbol{\Omega}_P(\boldsymbol{\theta}_0)^{-1}\boldsymbol{y}\Big] \\
&\quad + \frac{1}{2N}\ln\big|\boldsymbol{\Omega}_Q(\boldsymbol{\theta})\big| + \frac{1}{2N}\mathbb{E}_{P(\boldsymbol{\theta}_0)}\Big[\boldsymbol{y}^T\boldsymbol{\Omega}_Q(\boldsymbol{\theta})^{-1}\boldsymbol{y}\Big] \\
=& -\frac{1}{2N}\ln\big|\boldsymbol{\Omega}_P(\boldsymbol{\theta}_0)\big| - \frac{1}{2} + \frac{1}{2N}\ln\big|\boldsymbol{\Omega}_Q(\boldsymbol{\theta})\big| + \frac{1}{2N}\mathrm{trace}\big[\boldsymbol{\Omega}_P(\boldsymbol{\theta}_0)\cdot\boldsymbol{\Omega}_Q(\boldsymbol{\theta})^{-1}\big].
\end{aligned} \tag{5.7}$$

For large m, $\boldsymbol{\theta}_1$ can be approximated by numerically solving (5.6), and subsequently plugged into (5.7) in place of $\boldsymbol{\theta}$ to compute the last two terms. This leads to

$$\frac{1}{N}\mathbb{E}_{P(\boldsymbol{\theta}_0)}\Big[\ell_Q(\boldsymbol{\theta}_1;\boldsymbol{y})\Big] = -\frac{1}{2N}\ln\big|\boldsymbol{\Omega}_Q(\boldsymbol{\theta}_1)\big| - \frac{1}{2N}\text{trace}\big[\boldsymbol{\Omega}_P(\boldsymbol{\theta}_0)\cdot\boldsymbol{\Omega}_Q(\boldsymbol{\theta}_1)^{-1}\big], \tag{5.8}$$

the per individual contribution to the expected log-likelihood from fitting $Q(\boldsymbol{\theta})$, when $P(\boldsymbol{\theta}_0)$ is the true distribution of the trait. Equation (5.8) provides the basis for approximating ELRT in large sample. At any locus l, we can compute (5.8) for the fitted models both under H_l and H_n respectively. The difference of the two quantities multiplied by 2 is the per individual contribution to the ELRT (pELRT). Multiplying the pELRT by the sample size produces the ELRT for a study.

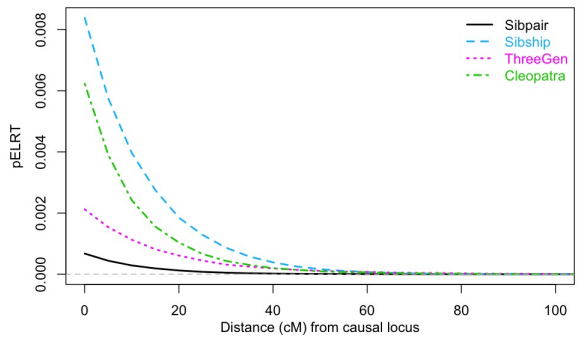
When the genetic correlation matrices are correctly specified (*i.e.*, testing at locus l_0 with perfect IBD information), the LRT in (5.4) has an asymptotic non-central chi-squared distribution, $\chi^2_1(\xi)$. The ELRT is related to the non-centrality parameter ξ by

$$\xi = \text{ELRT} - 1. \tag{5.9}$$

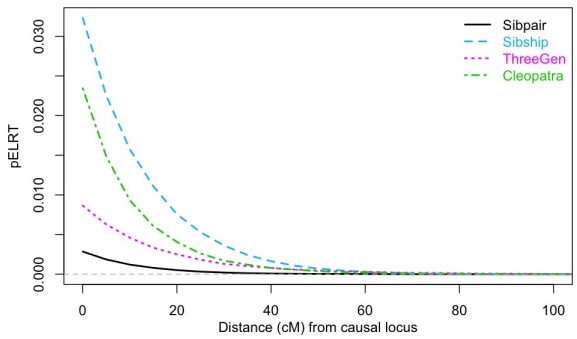
Given the ELRT, we can calculate the power of the linkage test as a function of the variance parameters for any pre-specified linkage detection threshold. When the correlation matrices are mis-specified, the LRT no longer has a limiting distribution of $\chi^2_1(\xi)$. We can still use $\chi^2_1(\xi)$ as an approximate limiting distribution. The simulation studies in Section 5.3 will show that the approximation is good if the mis-specification is mild.

5.3 Testing at non-causal loci

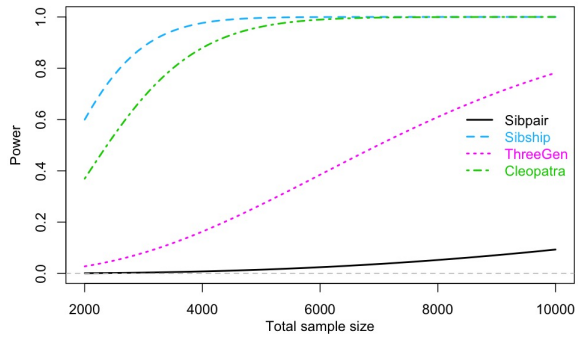
We first investigate how pELRT changes with distance to the true causal region centered at locus l_0. Test loci are chosen to be in increment of 5cM away from l_0. IBD patterns are simulated jointly at l_0 and all other linked loci. We assume perfect IBD information (not estimated) in this study, so that mis-specification of $\boldsymbol{G}_{l,f}$ occurs only when testing is done at non-causal loci. To focus on the effect of mis-specifying $\boldsymbol{G}_{l,f}$, we assume the residual genetic



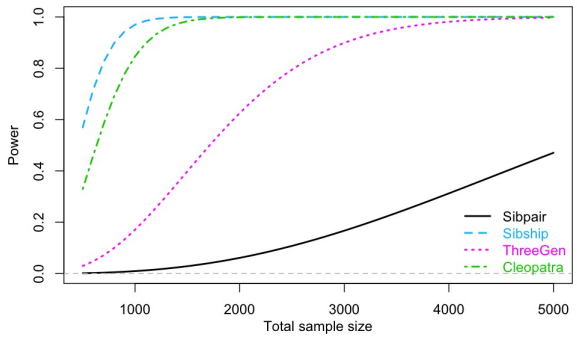
(a) $\boldsymbol{\theta}_0 = (0.1, 0.2, 0.7)$



(b) $\boldsymbol{\theta}_0 = (0.2, 0.2, 0.6)$



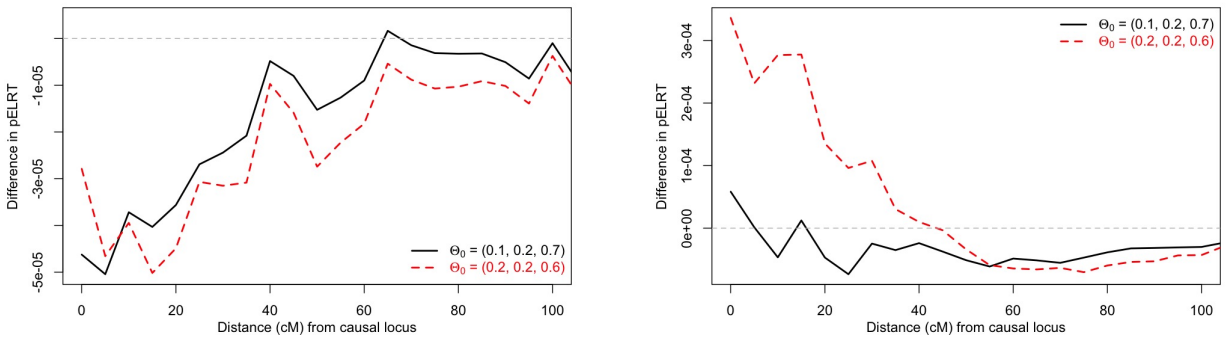
(c) $\boldsymbol{\theta}_0 = (0.1, 0.2, 0.7)$



(d) $\boldsymbol{\theta}_0 = (0.2, 0.2, 0.6)$

Figure 5.1: **(a) & (b)**: Per individual contribution to ELRT under different designs, when linkage tests are performed at loci linked to the true causal locus l_0. **(c) & (d)**: Power to detect linkage at the causal locus with a critical value of 13.815 (LOD score of 3). True variance parameters used are $\boldsymbol{\theta}_0 = (0.1, 0.2, 0.7)$ and $\boldsymbol{\theta}_0 = (0.2, 0.2, 0.6)$.

correlation matrix $\boldsymbol{G}_{g,f}$ is correctly specified. We consider four different types of designs: sibpairs, sibships of 14, three-generation pedigree of 14 members , and the Cleopatra pedigree of 14 members introduced in Section 1.4.2. To capture variation in realized kinship under each design, we use a sample size of $N = 14000$ for the 14-member pedigrees, and $N = 20000$ for sibpair. Two sets of true variance parameters are used: $\boldsymbol{\theta}_0 = (\sigma^2_{l,0}, \sigma^2_{g,0}, \sigma^2_{e,0})$ take values of either (0.1, 0.2, 0.7) or (0.2, 0.2, 0.6). Total sample size is in the number of study subjects.

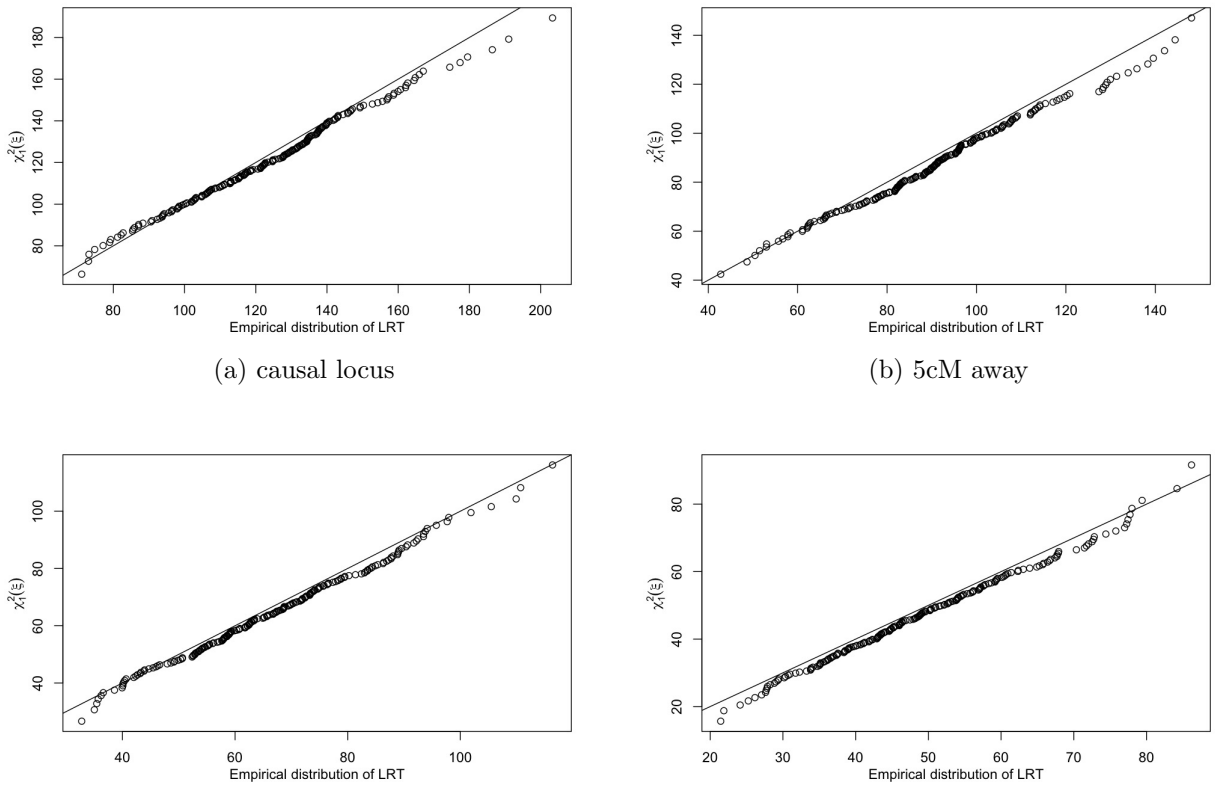


(a) (b)

Figure 5.2: Difference in pELRT obtained using equations (5.6) through (5.8) and from the empirical distribution of LRT, under **(a)** the sibpair design and **(b)** the sibship design.

Given $\boldsymbol{G}_{l,f}$ and $\boldsymbol{G}_{g,f}$ (from true IBD sharing) at each test locus and knowledge of the true trait model, we compute pELRT using equations (5.6) through (5.8). Figure 5.1a and 5.1b show how pELRT varies with distance to the causal locus under two different sets of variance parameters. For any design and distance from the causal locus, pELRT is larger when $\sigma^2_{l,0}$ is larger. Otherwise, the patterns shown in Figure 5.1a and 5.1b are almost identical. Within each design, pELRT decreases with distance away from the causal locus. Among the designs, sibpairs are the least powerful, whereas sibships of 14 are the most powerful. Figure 5.1c and 5.1d show the power to detect linkage at the causal locus with a critical value of 13.815 (corresponding to LOD score of 3). To achieve the same power, the sibpair design requires a much larger sample size than other designs. In general, designs with high level of inter-relatedness of sampled individuals seem to be more powerful. This is likely because there is higher variability in the difference between $\boldsymbol{G}_{l,t}$ and $\boldsymbol{G}_{g,t}$, leading to bigger contrast between the likelihood functions obtained under H_n and H_l.

Next, we verify the analytical results by simulating trait values under the trait model (5.2) with the two sets of $\boldsymbol{\theta}_0$ described earlier, and perform linkage tests at all test loci. Empirical



(a) causal locus

(b) 5cM away

(c) 10cM away

(d) 15cM away

Figure 5.3: QQ-plots of the empirical distribution of LRT against $\chi^2_1(\xi)$ **(a)** at the causal locus l_0; **(b)** 5cM away from l_0; **(c)** 10cM away from l_0; **(d)** 15cM away from l_0. The sample consists of 1000 independent three-generation pedigrees. $\xi = \text{pELRT} \times N - 1$ is computed using equation (5.6) through (5.8). True variance parameters are $\boldsymbol{\theta}_0 = (0.2, 0.2, 0.6)$.

distributions of LRT at each test locus are constructed from 200 simulation replicates. The mean of the LRT divided by sample size N can be compared to the analytically computed pELRT. Figure 5.2 compares the pELRT computed using the two approaches, under the sibpair design and the sibship design respectively. We see that the differences in pELRT obtained using the two approaches are generally very small. This agreement has also been observed under the three-generation and Cleopatra design (results not shown).

Figure 5.3 shows the QQ-plots of the empirical distributions of LRT and the corresponding non-central chi-squared distributions under the three-generation design. At the causal locus l_0, the pELRT computed using equations (5.6) through (5.8) can be used to obtain the non-centrality parameter ξ for any sample size N. As expected, the resulting non-central chi-squared distribution matches the empirical distribution of LRT for testing at l_0 (Figure 5.3a). Moving away from l_0, LRT no longer has $\chi^2_1(\xi)$ as the limiting distribution, but the non-central chi-squared distribution can still be a reasonable approximation when the test locus is not too far from the causal locus (Figure 5.3b through 5.3d). As we move further away from the causal locus, the approximation of limiting distribution of LRT using the non-central chi-squared distribution eventually breaks down. Across the four designs in our simulation study, the approximations generally seem reasonable when the test locus is within 10cM of the causal locus, for both sets of true variance parameters. Beyond 10cM, the approximation starts to break down for $\boldsymbol{\theta}_0 = (0.1, 0.2, 0.7)$ under the sibpair design (results not shown).

5.4 Error in local kinship estimation

Another main reason for $\boldsymbol{G}_{l,f} \neq \boldsymbol{G}_{l,t}$ is that there are errors in local kinship estimation. As we have shown in Chapter 3, accuracy of local kinship estimation can depend on factors such as the choice of local kinship estimator, remoteness of relationship under study, density of marker panel, and availability of accurate pedigree information. Beyond the investigation of Chapter 3, there are additional factors such as genotyping accuracy, availability of adequate allele frequency estimates, and potentially many more that can affect accuracy of local kinship estimation. While it is not possible to thoroughly study the effects of all those factors on detection of linkage, in this section we look at the potential impact on detection of linkage with (pedigree free) local kinship estimation accuracies attainable using `ibd_haplo` in real studies.

In this simulation study, we simulate IBD on the four designs as in previous sections, and assume that $\boldsymbol{G}_{l,f}$ is estimated without pedigree information. We use the three sets of local kinship estimation accuracies shown in Table 3.2 as references. They represent local kinship estimation accuracies of `ibd_haplo` obtained using relatively sparse, moderate and dense marker panels. Given true local kinship matrix $\boldsymbol{G}_{l,t}$, we create $\boldsymbol{G}_{l,f}$ by jittering elements of $\boldsymbol{G}_{l,t}$ according to conditional probabilities $(\hat{\phi}|\phi)$ given in the corresponding row of Table 3.2, for each of the three sets of estimation accuracies. Since the relationship types in Table 3.2 do not match perfectly with those in the four designs, we make the simplification that: (a) self-kinship can be perfectly estimated; and (b) local kinship estimation accuracy for a relationship type in one of the four designs, if not among those shown in Table 3.2, will be the same as the relationship type in Table 3.2 that is the closest in pedigree kinship. For example, a parent-offspring pair has a pedigree kinship of 0.25, so we introduce local kinship errors for it using the conditional probabilities for full siblings in Table 3.2. Since parent-offspring share 1 gene IBD at every locus, only the conditional probabilities for $(\hat{\phi}|\phi = 0.25)$ will be relevant. Again, we assume $\boldsymbol{G}_{g,f}$ is correctly specified. Sample size is kept at $N = 14000$ for the 14-member pedigrees, and $N = 20000$ for sibpairs to capture randomness in realized kinship. True parameters are either $\boldsymbol{\theta}_0 = (0.1, 0.2, 0.7)$ or $\boldsymbol{\theta}_0 = (0.2, 0.2, 0.6)$. Under each condition, pELRT at causal locus l_0 is computed using equations (5.6) through (5.8).

Table 5.1 shows pELRT ($\times 10^3$) computed with both known and estimated local kinship matrices. As expected, errors in local kinship estimation generally leads to reduction in pELRT at the causal locus. This reduction in pELRT is larger in some designs than others. For instance, the three generation pedigree has seen the largest reduction in pELRT among the four designs. This is potentially due to it having more (relatively) remote relationships for whom local kinship estimation is less accurate. Among the three sets of local kinship estimation accuracies, the one corresponding to use of a moderately dense marker panel is the most accurate (Figure 3.3 and Table 3.2), thus having the smallest reduction in pELRT. Between the sparse and the dense marker panels, local kinship estimation is less accurate

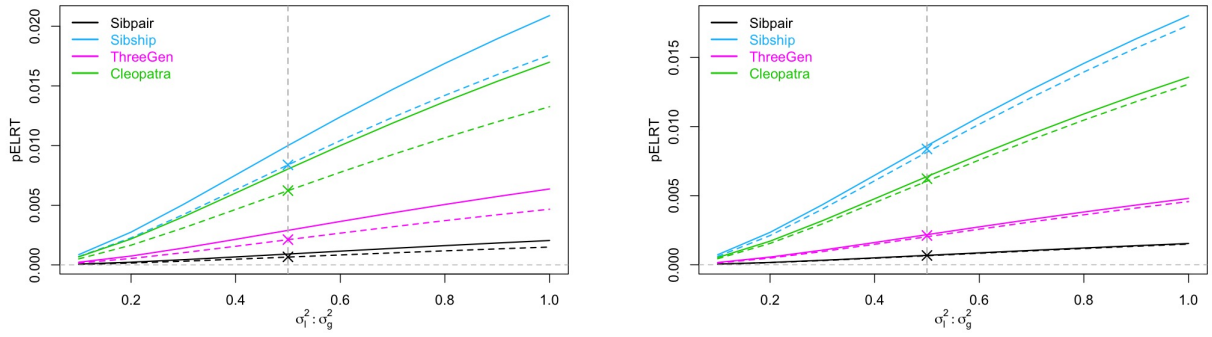
	Sibpair		Sibship		Threegen		Cleopatra	
σ_l^2	0.1	0.2	0.1	0.2	0.1	0.2	0.1	0.2
perfect IBD	0.674	2.854	8.382	32.314	2.121	8.643	6.227	23.443
sparse	0.572	2.429	6.957	26.212	1.105	4.431	5.115	18.624
moderate	0.649	2.750	8.003	30.543	1.749	7.032	5.949	22.147
dense	0.636	2.683	7.825	29.669	1.637	6.521	5.800	21.391

Table 5.1: Per individual contribution to ELRT ($\times 10^3$) at the causal locus when local kinship is either known or estimated with different levels of accuracy. $\sigma_l^2 = 0.1$ refers to $\boldsymbol{\theta}_0 = (0.1, 0.2, 0.7)$, whereas $\sigma_l^2 = 0.2$ refers to $\boldsymbol{\theta}_0 = (0.2, 0.2, 0.6)$. Sparse, moderate and dense refer to the relative density of the marker sets used to obtained the three sets of accuracies in Table 3.2.

with the sparse panel. In particular, use of the sparse marker panel often leads to under-estimation of local kinship, which is likely to be more detrimental for testing the presence of a local effect. As a result, we see the biggest reduction in pELRT when using the sparse marker panel. Use of the dense marker panel often leads to over-estimation of local kinship due to higher level of LD present. This is likely the reason for a bigger reduction in pELRT compared to the use of the moderately dense marker panel.

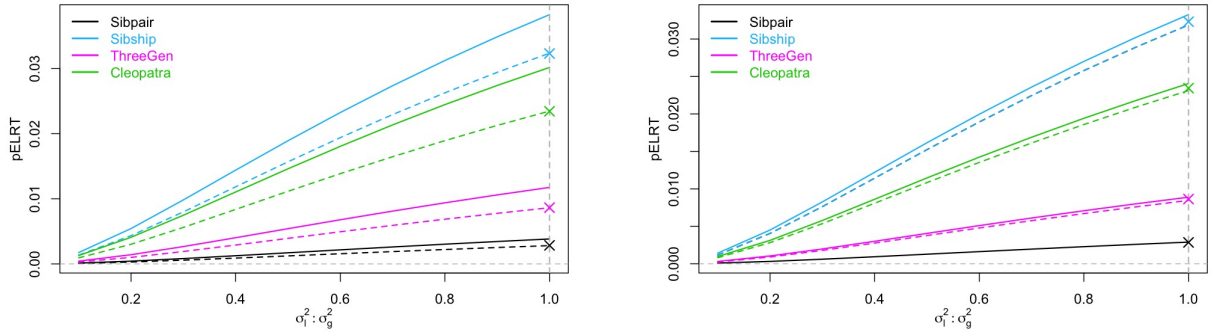
5.5 Mis-specifying residual genetic correlation

In previous sections we have focused on the impact of mis-specifying $\boldsymbol{G}_{l,f}$ on pELRT, and ignored the possibility that $\boldsymbol{G}_{g,f}$ may also be mis-specified. This is because the impact of mis-specifying $\boldsymbol{G}_{l,f}$ in a linkage analysis is likely to dominate that of $\boldsymbol{G}_{g,f}$. Being the genetic correlation matrix at a single locus, there is much higher variance in local kinship than in genome-wide realized kinship over descent for any pedigree relationship. In addition, there is much higher uncertainty in estimating local kinship than in estimating realized kinship over a long genomic segment. For example, if $\boldsymbol{G}_{g,t} = 2\boldsymbol{\Phi}$ and $\boldsymbol{G}_{g,f}$ is constructed as twice



(a) $\boldsymbol{G}_{g,t}$ from Chr 22, $\sigma_l^2 + \sigma_g^2 = 0.3$

(b) $\boldsymbol{G}_{g,t}$ from Chr 2∼22, $\sigma_l^2 + \sigma_g^2 = 0.3$



(c) $\boldsymbol{G}_{g,t}$ from Chr 22, $\sigma_l^2 + \sigma_g^2 = 0.4$

(d) $\boldsymbol{G}_{g,t}$ from Chr 2∼22, $\sigma_l^2 + \sigma_g^2 = 0.4$

Figure 5.4: Per individual contribution to ELRT computed at different ratios of $\sigma_l^2 : \sigma_g^2$, when $(\sigma_l^2 + \sigma_g^2)/\sigma^2 \in \{0.3, 0.4\}$. $\boldsymbol{G}_{g,t}$ corresponds to realized kinship on Chr 22 alone, or realized kinship from Chr 2 to Chr 22. Solid curves correspond to $\boldsymbol{G}_{g,f} = \boldsymbol{G}_{g,t}$, whereas dashed curves correspond to $\boldsymbol{G}_{g,f} = 2\boldsymbol{\Phi}$. Dashed vertical lines correspond to $\boldsymbol{\theta}_0 = (0.1, 0.2, 0.7)$ or $\boldsymbol{\theta}_0 = (0.2, 0.2, 0.6)$ used in previous sections. Crosses ($\times$) represent pELRT computed under respective $\boldsymbol{\theta}_0$ when $\boldsymbol{G}_{g,f} = \boldsymbol{G}_{g,t} = 2\boldsymbol{\Phi}$.

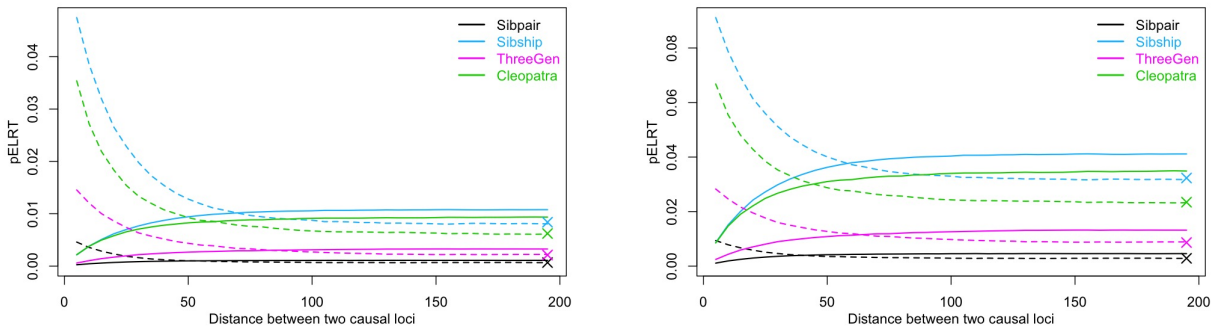
the estimates of genome-wide realized kinship using a local IBD method such as `ibd_haplo`, then the difference between $\boldsymbol{G}_{g,f}$ and $\boldsymbol{G}_{g,t}$ will generally be very small (Chapter 2), and on a much smaller scale than the difference between $\boldsymbol{G}_{l,f}$ and $\boldsymbol{G}_{l,t}$. Without knowing where the residual genetic effect come from, it is natural to associate $\boldsymbol{G}_{g,f}$ with genome-wide realized kinship. The main source of concern for mis-specifying $\boldsymbol{G}_{g,f}$ does not come from errors in genome-wide kinship estimation, but rather on the possibility that residual genetic effect is

explained by only part of the genome. In this section we use a simulation study to investigate how such mis-specification of $\boldsymbol{G}_{g,f}$ may impact pELRT.

In this simulation study, we assume $\boldsymbol{G}_{l,f}$ is correctly specified and l_0 lies on chromosome (Chr) 1. We assume that residual genetic effect is either captured by realized kinship on Chr 22 alone, or by realized kinship from Chr 2 to Chr 22 (genome less Chr 1). These realized kinship values are assumed known without error. We set $\sigma_l^2 + \sigma_g^2 = 0.3$ or 0.4, while total phenotypic variance $\sigma^2 = 1$. For each of the four designs used in previous sections, we compute pELRT from large samples ($N = 14000$ for 14-member pedigrees, and $N = 20000$ for sibpairs) at different ratios of $\sigma_l^2 : \sigma_g^2$. Results are shown in Figure 5.4.

When $\boldsymbol{G}_{g,f} = 2\boldsymbol{\Phi} \neq \boldsymbol{G}_{g,t}$, there is generally a drop in pELRT (compare solid and dashed curves of the same color). This reduction is more severe when the ratio $\sigma_l^2 : \sigma_g^2$ is high, when $\boldsymbol{G}_{g,t}$ correspond to a smaller portion of the genome (compare Figure 5.4a to 5.4b, and 5.4c to 5.4d), and when $\sigma_l^2 + \sigma_g^2$ is small (compare Figure 5.4a to 5.4c, and 5.4b to 5.4d). Another interesting observation is that when $\boldsymbol{G}_{g,f} = 2\boldsymbol{\Phi}$, the resulting pELRT is almost identical to that obtained under $\boldsymbol{G}_{g,f} = \boldsymbol{G}_{g,t} = 2\boldsymbol{\Phi}$ (marked by $\times$) with the same $\boldsymbol{\theta}_0$. This is similar to what we observed in Chapter 4 for heritability estimation: when the fitted genetic correlation matrix ($\boldsymbol{G}_{g,f}$) represents a broader measure of kinship than the truth ($\boldsymbol{G}_{g,t}$), asymptotic bias in pELRT tends to be small. This means that the pELRT we obtained by fitting $\boldsymbol{G}_{g,f} = 2\boldsymbol{\Phi}$ is still appropriate for approximating power of the study under the assumption of $\boldsymbol{G}_{g,t} = 2\boldsymbol{\Phi}$.

When fitting $\boldsymbol{G}_{g,f} = 2\boldsymbol{\Phi}$, the most extreme case of mis-specification occurs when residual genetic correlation $\boldsymbol{G}_{g,t}$ is in fact captured by realized kinship at a single locus (another causal locus). We investigate this scenario in the last simulation study of this chapter. We assume that the second causal locus is linked to the testing locus l_0 at some genetic distance away. Under each of the four designs and each of the two sets of variance parameters we used previously, we compute pELRT for testing at l_0 analytically both with $\boldsymbol{G}_{g,f} = \boldsymbol{G}_{g,t}$ and



(a) $\boldsymbol{\theta}_0 = (0.1, 0.2, 0.7)$ (b) $\boldsymbol{\theta}_0 = (0.2, 0.2, 0.6)$

Figure 5.5: Per individual contribution to ELRT at l_0, when residual genetic correlation is captured by realized kinship at a linked (second) causal locus. Solid curves correspond to $\boldsymbol{G}_{g,f} = \boldsymbol{G}_{g,t}$, whereas dashed curves correspond to $\boldsymbol{G}_{g,f} = 2\boldsymbol{\Phi}$. Crosses ($\times$) represent pELRT computed under respective $\boldsymbol{\theta}_0$ when $\boldsymbol{G}_{g,f} = \boldsymbol{G}_{g,t} = 2\boldsymbol{\Phi}$. They are placed at the right end of the figures for clarity,

with $\boldsymbol{G}_{g,f} = 2\boldsymbol{\Phi} \neq \boldsymbol{G}_{g,t}$. Both local and genome-wide realized kinship are assumed known without error. Results of the simulation are shown in Figure 5.5.

An interesting observation from Figure 5.5 is that pELRT increases with distance between the causal loci when $\boldsymbol{G}_{g,f} = \boldsymbol{G}_{g,t}$ (solid lines), whereas it decreases with distance when $\boldsymbol{G}_{g,f} = 2\boldsymbol{\Phi}$ (dashed lines). This is because when the two causal loci are very close, they behave similarly to a single causal locus with bigger effect. Fitting $\boldsymbol{G}_{g,f} = \boldsymbol{G}_{g,t}$ will capture genetic correlation at both loci, so the test for $\sigma_l^2 = 0$ is much less powerful. On the other hand, fitting $\boldsymbol{G}_{g,f} = 2\boldsymbol{\Phi}$ will not mask the effect at l_0. The combined effects of two loci in close proximity leads to higher power to detect either of them. When the two loci are more than 100cM apart, they behave similarly to two independent loci. This is evident as the curves correspond to both $\boldsymbol{G}_{g,f} = \boldsymbol{G}_{g,t}$ and $\boldsymbol{G}_{g,f} = 2\boldsymbol{\Phi}$ flatten out after 100cM. In this case, fitting $\boldsymbol{G}_{g,f} = \boldsymbol{G}_{g,t}$ leads to higher power in detecting linkage at l_0, as we have seen in Figure 5.4. In addition, fitting $\boldsymbol{G}_{g,f} = 2\boldsymbol{\Phi}$ (when $\boldsymbol{G}_{g,t}$ corresponds to a second causal locus)

produces pELRT that is almost identical to that obtained under $\boldsymbol{G}_{g,f} = \boldsymbol{G}_{g,t} = 2\boldsymbol{\Phi}$ (marked by $\times$).

5.6 Discussion

For linkage analysis of complex quantitative trait using the variance component models, the per individual contribution to the expected likelihood ratio test statistics (pELRT) is a useful statistic to analyze power to detect linkage. Given design choice and hypothesized covariance structure of the trait, pELRT can be computed numerically using simulated IBD information on large samples. It can be used to approximate the sample size needed to achieve certain power for any design, and can be compared between designs to help planning for big studies.

When testing for linkage at genome-wide markers, it is possible that both the local and residual genetic correlation matrices are mis-specified, leading to loss of power to detect linkage at the causal locus. Given the higher amount of variation in local kinship, the impact of mis-specifying the local genetic correlation matrix is likely to dominate. The extent of mis-specification can be reduced by using more accurate local IBD estimators, by carefully selecting the marker panel so as to increase marker informativeness and limit the amount of linkage disequilibrium at the same time (see Chapter 3), and by having better qualify genotyping procedures.

Without knowing which part of the genome is actually responsible for the residual genetic effects, it is reasonable to use twice the genome-wide realize kinship to capture residual genetic correlation. This quantity can usually be estimated accurately as a by-product of local kinship estimation. When the sample is not ascertained by either trait or marker, twice the genome-wide realized kinship on average captures the correct residual genetic correlation even if residual genetic effects are attributable to only part of the genome. Fitting a model with a broader measure of kinship for $\boldsymbol{G}_{g,f}$ has the additional benefit of having less asymptotic bias in pELRT when $\boldsymbol{G}_{g,f}$ is mis-specified, as we have shown in our simulation

studies.

Chapter 6

OVERVIEW

This chapter provides brief discussions on broader issues that have not been investigated in this thesis, and summarizes the original contributions.

6.1 Broader issues not discussed

For estimation of local and genome-wide realized kinship, we implicitly assumed that the subjects come from a single, homogeneous population. This assumption is often untenable in studies involving population-based samples. In the context of inferring genome-wide realized kinship, ignoring population substructure and admixture can lead to bias in estimates (Manichaikul *et al.* 2010, e.g.). Thornton *et al.* (2012) and Conomos *et al.* (2016) proposed different methods to estimate subject-specific allele frequencies that take population substructure and admixture into account, and subsequently use them to estimate genome-wide realized kinship. This idea has also been used by Nafikov *et al.* (2018) to obtain more accurate estimates of IBD states for pedigree-based linkage analysis. The same logic can be applied to both global and local kinship estimators discussed in Chapter 2 and 3.

We have assumed throughout the thesis that samples are obtained randomly from the population. In practice, however, samples are sometimes ascertained by trait values. For genetic traits, ascertainment changes the distribution of IBD states among related individuals at or near the causal loci. Apart from the fact that higher level of IBD sharing makes it easier to estimate IBD more accurately (Chapter 2 and 3), it is not expected to have an impact on accuracies of the pedigree-free IBD estimation methods discussed in Chapter 2 and 3. For gene mapping, changes in expected IBD sharing between related individuals at the causal

locus can be used to detect linkage. Ascertainment by extreme trait values has long been used with sibpair designs for linkage analysis (e.g. Risch and Zhang 1995). This approach works fine with IBD-sharing methods such as the Haseman-Elston regression (Haseman and Elston 1972), but can pose problems for the random effects model we discussed in Chapter 4, which models trait conditional on IBD sharing. In the case of ascertainment by extreme discordant sibpairs, Forrest and Feingold (2000) showed that the random effects model can be severely under-powered to detect linkage at the causal locus, due to distortion of the bivariate normal trait distribution. This is not to say that the random effects model will not work whenever samples are ascertained by trait values, as the impact on gene mapping performance depends heavily on the ascertainment strategy. For oligogenic traits, we found in our experiments (results not shown) that ascertainment by trait value can be helpful for detecting the major causal loci. However, it can also make detection of minor causal loci more difficult. Depending on the ascertainment strategy, the ascertained pedigrees may not be segregating variants at the minor causal loci, or they are segregating variants at the major as well as the minor causal loci, so that signals at the minor causal loci are masked by increased IBD sharing elsewhere (captured in the polygenic component).

In both Chapter 4 and Chapter 5, we have modeled genetic effects as random effects that follow multivariate normal distributions. This is not true biologically, because genes affect trait values through alleles, whose effects may not be independently normally distributed. Nevertheless, the random effects models provide a simple and yet flexible framework for modeling genetic correlation between related individuals, making it a useful tool for studying variance components. In the context of gene mapping, the likelihood ratio test on variance components has been shown to be robust to violation of multivariate normality (e.g. Allison *et al.* 1999).

For the purpose of investigating the effect of study design on heritability estimation and gene mapping, we have assumed that the samples consisted of multiple pedigrees with the

same structure. In practice, the model can incorporate pedigrees with different structures. The model can also include more random effects than discussed in this thesis. For example, there can be additional random effects to account for covariances due to dominance effect or shared environment both in heritability estimation and gene mapping.

6.2 Original contributions

The overall goal of this thesis is to study the use of realized kinship in modeling and analysis of quantitative trait data on related individuals. An important first step is to acquire accurate estimates of realized kinship, both locally and as a genome-wide average. The covariance matrix of quantitative trait values is often modeled as some function of some estimates of realized kinship. This covariance matrix may be mis-specified due to poor estimation of realized kinship or incorrect specification of the causal genome. Understanding the effects of such mis-specification can help with selecting the best approaches to construct the covariance matrix for more robust analysis outcomes.

Chapter 2 focused on pedigree-free estimation of genome-wide realized kinship. Existing estimators can be divided into two groups. One group of estimators treat markers as permutable, ignoring the fact that DNA is inherited in segments. Another group of estimators makes use of information that reflects joint inheritance of nearby markers, such as basepair position, genetic map and LD information. We introduced a generalized framework for a class of GRM type estimators, and showed that several existing estimators are special cases within this class. Under certain assumptions, we showed that the GRM type estimators are unbiased, and obtained formulas for variance of the estimators. We proposed two new estimators within this class. Through extensive simulation studies, we showed that improved estimators of genome-wide realized kinship can be obtained (1) by optimal weighting of markers, (2) by taking physical contiguity of genome into account, and (3) by weighting on the basis of LD.

Chapter 3 discussed pedigree-free estimation of local kinship. In a simulation study, we

compared local kinship estimation accuracy of the Hidden Markov model (HMM) used in the `ibd_haplo` program (Brown *et al.* 2012) to that of the non-parametric method proposed by Day-Williams *et al.* (2011). The HMM is generally more accurate than the non-parametric method, especially when there are short segments of IBD. For the HMM of `ibd_haplo`, we explored factors that are important for accurate estimation of local kinship. Denser SNP panels provide more information for IBD estimation, but higher levels of LD in denser panels hurt estimation accuracy by violating the conditional independence assumption that underlies the HMM. The trade off between marker information and LD for `ibd_haplo` is evident from our simulation studies. Presence of LD provides challenges to local IBD methods that do not adjust for LD. For a pedigree-based IBD method such as the HMM implemented in `MERLIN` (Abecasis *et al.* 2002), we have shown that high levels of LD can be even more harmful than they are to a pedigree-free method like `ibd_haplo`. For local IBD methods that do not adjust for LD, a good strategy is to apply the methods to LD-pruned marker set.

In Chapter 4 we investigated the problem of heritability estimation using a two-component random effects model. A key step in the estimation process is to specify the genetic correlation matrix. Under the multivariate normality assumption of the trait values, we provided formulas for the asymptotic bias and sampling variance of heritability estimate as functions of the true and the fitted genetic correlation matrices. This is important for understanding the potential impact of mis-specifying the genetic correlation matrix on heritability estimation. Genetic correlation between a pair of individuals is typically captured with twice some measure of kinship. When there is mis-specification, we showed that using a broader measure of kinship (e.g., realized kinship over longer segment, or more accurate kinship estimates) than the truth to capture genetic correlation is less likely to create bias in heritability estimates. Through simulation studies, we explored how study designs can influence the outcome of heritability estimation. In particular, we showed that using the classic GRM as the fitted genetic correlation matrix in a population-based design can lead to substantial downward

bias in heritability estimate, if the true genetic correlation is captured by genome-wide realized kinship. This choice of kinship estimator is one of many factors that can contribute to the "missing heritability" problem.

In Chapter 5 we extended the work of Chapter 4 by studying gene mapping via a three-component random effects model, where local kinship at the test locus and genome-wide realized kinship are used to capture local and residual genetic correlations between related individuals respectively. We provided a method for approximating the expected likelihood ratio test statistic in large samples, which can be useful for power analysis and selection of study designs. In practice, it is possible that both the local and residual genetic correlation matrices are mis-specified, leading to loss of power to detect linkage at the causal locus. The extent of mis-specifying the local genetic correlation matrix can be reduced by improving accuracy of local IBD estimation (Chapter 3). To capture residual genetic correlation, we showed in a simulation study that it is helpful to use a broader measure of kinship such as genome-wide realized kinship. This choice of kinship measure generally leads to small loss of power to detect linkage at a causal locus if residual genetic correlation is mis-specified, and it does not obscure linkage signal at the test locus when there are multiple linked causal loci.

BIBLIOGRAPHY

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

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

Allison, D. B., M. C. Neale, R. Zannolli, N. J. Schork, C. I. Amos, *et al.*, 1999 Testing the robustness of the likelihood-ratio test in a variance-component quantitative-trait loci-mapping procedure. Americal Journal of Human Genetics **65**: 531–544.

Almasy, L. and J. Blangero, 1998 Multipoint quantitative-trait linkage analysis in general pedigrees. American Journal of Human Genetics **62**: 1198–1211.

Amos, C. I., 1994 Robust variance-components approach for assessing genetic linkage in pedigrees. American Journal of Human Genetics **54**: 535–543.

Anderson, A. D. and B. S. Weir, 2007 A maximum-likelihood method for the estimation of pairwise relatedness in structured populations. Genetics **176**: 421–440.

Balding, D. J. and R. A. Nicolas, 1994 Dna profile match prob- ability calculations: how to allow for population stratification, relatedness, database selection, and single bands. Forensic Sci. Int. **64**: 125–140.

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

Browning, B. L. and S. R. Browning, 2011 A fast powerful method for detecting identity by descent. American Journal of Human Genetics **88**: 173–182.

Browning, S. R., 2008 Estimation of pairwise identity by descent from sense genetic marker data in a population sample of haplotypes. Genetics **178**: 2123–2132.

Browning, S. R. and B. L. Browning, 2010 High-resolution detection of identity by descent in unrelated individuals. American Journal of Human Genetics **86**: 526–539.

Choi, Y., E. M. Wijsman, and B. S. Weir, 2009 Case-control association testing in the presense of unknown relationships. Genetic Epidemiology **33**: 668–678.

Conomos, M. P., A. P. Reiner, B. S. Weir, and T. A. Thornton, 2016 Model-free estimation of recent genetic relatedness. American Journal of Human Genetics **98**: 127–148.

Cotterman, C. W., 1940 *A Calculus for Statistico-Genetics*. Ph.D. thesis, Ohio State University.

Day-Williams, A., J. Blangero, T. Dyer, K. Lange, and E. Sobel, 2011 Linkage analysis without defined pedigrees. Genetic Epidemiology **35**: 360–370.

Ewens, W. J., 1972 The sampling theory of selectively neutral alleles. Theoretical Population Biology **3**: 87–112.

Falconer, D. S. and T. F. C. Mackay, 1996 *Introduction to quantitative genetics*. Longman, Harlow.

Fisher, R. A., 1918 The correlation between relatives on the supposition of mendelian inheritance. Transactions of the Royal Society of Edinburgh **52**: 399–433.

Forrest, W. F. and E. Feingold, 2000 Composite statistics for qtl mapping with moderately discordant sibling pairs. American Journal of Human Genetics **66**: 1642–1640.

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

Goddard, K. A., C. E. Yu, J. Oshima, T. Miki, J. Nakura, *et al.*, 1996 Toward localization of the werner syndrome gene by linkage disequilibrium and ancestral haplotyping: lessons learned from analysis of 35 chromosome 8p11.1-21.1 markers. Am. J. Hum. Genet. **58**: 1286–1302.

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

Hayes, B. J., P. M. Visscher, and M. E. Goddard, 2009 Increased accuracy of artificial selection by using the realized relationship matrix. Genetical Research **91**: 47–60.

Henderson, C. R., 1976 A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. Biometrics **32**: 69–83.

Hill, W. G. and B. S. Weir, 2011 Variation in actual relationship as a consequence of mendelian sampling and linkage. Genetics Research **93**: 47–64.

Jacquard, A., 1974 *The Genetic Structure of Populations*. Springer-Verlag, New York.

Knowles, E. E. M., J. W. Kent, D. R. McKay, E. Sprooten, S. R. Mathias, *et al.*, 2016 Genome-wide linkage on chromosome 10q26 for a dimensional scale of major depression. Journal of Affective Disorders **191**: 123–131.

Kruuk, L. E., 2004 Estimating genetic parameters in natural populations using the "animal model". Philos. Trans. R. Soc. Lond. B. Biol. Sci. **359**: 879–890.

Lander, E. S. and P. Green, 1987 Construction of multilocus genetic linkage maps in humans. Proceedings of the National Academy of Sciences, USA **84**: 2363–2367.

Liu, K., M. Goodman, S. Muse, J. S. Smith, E. Buckler, *et al.*, 2003 Genetic structure and diversity among maize inbred lines as inferred from dna microsatellites. Genetics **165**: 2117–2128.

Lubin, M. and I. Dunning, 2015 Computing in operations research using julia. INFORMS Journal on Computing **27**: 238–248.

Lynch, M. and B. Walsk, 1998 *Genetics and analysis of quantitative traits*. Sinauer Associates, Sunderland, Massachusetts.

Maher, B., 2008 Personal genomes: The case of the missing heritability. Nature **456**: 18–21.

Manichaikul, A., J. C. Mychaleckyj, S. S. Rich, K. Daly, M. Sale, *et al.*, 2010 Robust relationship inference in genome-wide association studies. Bioinformatics **26**: 2867–2873.

Matise, T. C., F. Chen, W. Chen, F. M. De La Vega, M. Hansen, *et al.*, 2007 A second-generation combined linkage physical map of the human genome. Genome Research **17**: 1783–6.

Mendel, G., 1866 Experiments in plant hybridisation. Verhandlungen des naturforschenden Vereines in Brunn **4**: 3–47.

Milligan, B. G., 2003 Maximum-likelihood estimation of relatedness. Genetics **163**: 1153–1167.

Moltke, I., A. Albrechtsen, T. Hansen, F. C. Nielsen, and R. Nielsen, 2011 A method for detecting ibd regions simultaneously in multiple individuals: with applications to disease genetics. Genome Research **21**: 1168–1180.

MORGAN Tutorial, 2016 Monte carlo genetic analysis package. `https://www.stat.washington.edu/thompson/Genepi/MORGAN/morgan-tut_33_html/morgan-tut.html`.

Nafikov, R. A., A. Q. Nato, H. Sohi, B. Wang, L. Brown, *et al.*, 2018 Analysis of pedigree data in populations with multiple ancestries: strategies for dealing with admixture in caribbean hispanic families from the adsp. Genetic Epidemiology **42**: 500–515.

Patterson, N., A. L. Price, and D. Reich, 2006 Population structure and eigenanalysis. PLOS Genetics **2**: e190.

Pearson, K. S., 1904 Mathematical contributions to the theory of evolution. xii. on a generalised theory of alternative inheritance, with special reference to mendel's laws. Philos. Trans. R. Soc. Lond. **203**: 53–86.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, *et al.*, 2007 PLINK: a tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics **81**: 559–575.

Raffa, J. D. and E. A. Thompson, 2016 Power and effective study size in heritability studies. Statistics in Biosciences **8**: 264–283.

Risch, N. and H. Zhang, 1995 Extreme discordant sib pairs for mapping quantitative trait loci in humans. Science **268**: 1584–1589.

Sobel, E. and K. Lange, 1996 Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. American Journal of Human Genetics **58**: 1323–1337.

Speed, D., G. Hemani, M. R. Johnson, and D. J. Balding, 2012 Improved heritability estimation from genome-wide SNPs. The American Journal of Human Genetics **91**: 1011–1021.

Sprooten, E., C. N. Gupta, E. E. M. Knowles, D. R. McKay, S. R. Mathias, *et al.*, 2015 Genome-wide significant linkage of schizophrenia-related neuroanatomical trait to 12q24. Am J Med Genet B Neuropsychiatr Genet **168**: 678–686.

Sverdlov, S., 2014 *Functional Quantitative Genetics and the Missing Heritability Problem*. Ph.D. thesis, University of Washington.

Tavare, S. and W. J. Ewens, 1997 The multivariate ewens distribution. In *Discrete Multivariate Distributions*, edited by N. L. Johnson, S. Kotz, and N. Balakrishnan, pp. 232–246, Wiley, New York.

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

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

Thompson, E. A., 2000 *Statistical Inference from Genetic Data on Pedigrees*, volume 6. Institute of Mathematical Statistics, Beechwood OH and Alexandria VA.

Thompson, E. A., 2008 The ibd process along four chromosomes. Theoretical Population Biology **73**: 369–373.

Thompson, E. A. and R. G. Shaw, 1990 Pedigree analysis for quantitative traits: variance components without matrix inversion. Biometrics **46**: 399–413.

Thornton, T., H. Tang, T. J. Hoffman, H. M. Ochs-Balcom, B. J. Caan, *et al.*, 2012 Estimating kinship in admixed populations. American Journal of Human Genetics **91**: 122–138.

Truong, D. T., L. D. Shriberg, S. D. Smith, K. L. Chapman, A. R. Scheer-Cohen, *et al.*, 2016 Multipoint genome-wide linkage scan for nonword repetition in a multigenerational family further supports chromosome 13q as a locus for verbal trait disorders. Human Genetics **135**: 1329–1341.

VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. Journal of Dairy Science **91**: 4414–4423.

Visscher, P. M. and M. E. Goddard, 2015 A general unified framework to assess the sampling variance of heritability estimates using pedigree or marker-based relationships. Genetics **199**: 223–232.

Wang, B., 2018 *rres: Realized Relatedness Estimation and Simulation*. R package version 1.1.

Wikipedia, 2018 Cleopatra pedigree. `http://en.wikipedia.org/wiki/Cleopatra#Ancestry`.

Wright, S., 1922 Coefficients of inbreeding and relationship. The American Naturalist **56**: 330–338.

Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, *et al.*, 2010 Common SNPs explain a large proportion of the heritability for human height. Nature Genetics **42**: 565–569.

Zou, F., S. Lee, M. R. Knowles, and F. A. Wright, 2010 Quantification of population structure using correlated SNPs by shrinkage principal components. Human Heredity **70**: 9–22.