Use of theoretical and estimated identity-by-descent (IBD) allele sharing measures in genome-wide linkage and association studies, with application to large pedigrees

Jakris Eu-ahsunthornwattana

A thesis submitted for the degree of Doctor of Philosophy (Ph.D.) in the Faculty of Medical Sciences of the Newcastle University

Institute of Genetic Medicine

Newcastle University

April 2015

Abstract

Traditionally, identity-by-descent (IBD) sharing among related individuals is estimated on the basis of the assumed pedigree structure, possibly combined with genotyping information for some or all subjects at a series of genetic markers. Recently, there has been interest in using dense SNP genotype data to estimate both average (across the genome) and local (at particular locations) IBD sharing by pairs of individuals. Although originally intended for inference of pedigree relatedness, these genetically estimated IBDs can potentially replace the traditional IBD estimates used in various genetic data analysis methods. I compared IBD estimates from various software packages (PLINK, KING and linear mixed model (LMM) packages including EMMAX, FaST-LMM, GenABEL, GEMMA and MMM) with the theoretical estimates, and examined their utility in application to LMM association analysis of real and simulated qualitative and quantitative phenotypes from a Brazilian family-based study of visceral leishmaniasis (VL) and from the 18th Genetic Analysis Workshop (GAW) data. Generally, the results from the different software packages were highly concordant. When used to model correlations between individuals in LMM analysis, these approaches achieved good control of type 1 error (well beyond that attainable using theoretical IBD estimates), while also achieving superior power to comparable non-LMM methods. Furthermore, although technically misspecified, LMM methods were also successfully applied to simulated longitudinal data. In addition, a new non-parametric linkage analysis method, Regional IBD Analysis (RIA), is proposed, where theoretical IBD estimates are replaced with the average and local genetic IBD estimates. This method was compared with traditional methods for non-parametric linkage analysis (either exact methods using small pedigrees from a study of vesicoureteral reflux disorder (VUR) or simulation-based methods using large pedigrees from the VL study) and was found to perform at least equally well while taking less time.

แด่คุณยาย
ผู้จากไปเมื่ออายุ ๙๐ ปี
๓๐ มีนาคม ๒๕๕๖
หากหลานมิอาจจะได้ไปลา
ด้วยศึกษา ณ แดนไกล

IN·CARAM·MEMORIAM·AVIAE·MEAE
QVAE·ANNO·NONAGESIMO·AETATIS·SVAE
A·D·III·KAL·APR·MMDCCLXVI·A·V·C·OBIIT
CVM·HIC·NEPOS·ILLIVS
LONGINQVIVS·SCIENTIAM·PROSEQVERETVR

In loving memory of my grandmother,
who, at 90 years of age,
passed away on 30th March 2013,
while this grandson of hers
was so far away pursuing knowledge.

Acknowledgement

I am grateful to my supervisor, Prof Heather J Cordell, firstly for deciding to accept my rather unusual application (and for reading my unsolicited email in the first place!); for her guidance, encouragement and support throughout this project; and for her comments and suggestions on this thesis, as well as other related reports, abstracts and manuscripts.

I also wish to thank my co-supervisor, Dr Ian J Wilson, for his advice on both the scientific and administrative aspects of this project, as well as on the various issues of bringing children to and keeping them happy in Newcastle.

Two data sets used in this project were provided by our collaborators: *the Brazilian Visceral Leishmaniasis data set* was provided by Michaela Fakiola, E Nancy Miller, Selma M B Jeronimo, Jenefer M Blackwell for the LeishGEN Consortium; sample handling, genotyping, initial quality control and prior statistical analysis was done by the Wellcome Trust Case Control Consortium 2 investigators including Serge Dronov, Sarah Edkins, Emma Gray, Sarah E Hunt, Cordelia Langford, Amy Strange, Chris Spencer, Matti Pirinen, Heather Cordell and Peter Donnelly; and *the Vesicoureteral Reflux Disease data set*, which consists of data from two projects: the whole genome studies of primary, nonsyndromic vesicoureteric reflux in the UK and Slovenia, which were collected by Rajko B Kenda in Slovenia and by the UK VUR Study Group (Aisling Stewart, Ambrose Gullett, Heather Lambert, Sue Malcom, Sally Feather, Timothy Goodship, Adrian Woolf, Judith Goodship) in the UK; and the whole genome studies of primary, nonsyndromic vesicoureteric reflux in Dublin, Ireland, which were collected by John M Darlow, Mark G Dobson, Cliona M Molony, Manuela Hunziker, Andrew J Green, Prem Puri, David E Barton. I am grateful for their permission to use these data sets in my project.

I wish to thank Dr Mauro Santibáñez-Koref, who, apart from his various comments from his role as my assessor and during our group's meetings, also always tried to ensure that I have adequate computational capacity on our clusters for my needs. I am also grateful for his advice related to other aspects of life while I am here.

I also wish to thank our IGM IT support team (Bryan Hepworth and Arron Scott) for patiently handling my (often idiosyncratic) requests throughout this project. My thanks also go to Dr Ben Allen, in his capacity as the administrator of the new FMS cluster, for his enthusiastic help with my requests and with troubleshooting, which at times involved digging into the C++ code of a third party program (I have a feeling he might

actually be enjoying himself doing so, but thank you anyway, Ben!); and for his sage advice about cluster usage, which I always find witty, insightful and enjoyable to read.

I am grateful to my *alma mater* and employer, Faculty of Medicine Ramathibodi Hospital, Mahidol University, for granting me the scholarship and funding that allows me to pursue this project. I wish to thank my unit head, Assoc Prof Thanyachai Sura, for his guidance and support during the planning of my further education, throughout the scholarship application process, as well as while I am here. I am also grateful to my colleagues in the Medical Genetics Unit, Drs Atchara Tunteeratum and Objoon Trachoo, who, along with Prof Sura, have been 'holding the fort' while I am away, thus enabling me to take this extended study leave. My thanks also go to our departmental secretary, Ms Somchit Rattanajinda, and unit secretary, Miss Pailin Lengraksa, who have tirelessly dealt with the various administrative issues at the faculty on my behalf, and especially for ensuring I have a constant supply of funding while I am here.

I have had a pleasant time working on this project in the Computational and Statistical Genetics group (in absence of a formal name for the group, I guess I'll just use the name inscribed on our door!) thanks to my colleagues: Richard Howey, Kristin Ayers, Rebecca Darlay, Holly Ainsworth, Jo Elson, Valentina Mamasoula, Konstantinos 'Kostas' Douroudis, Helen Griffin, So-Youn Shin, Mikyung Jang, Yaobo Xu, Wei Wei, Darren Houniet, Marla Endriga, Matthieu Miosec and Ginikachukwu 'Ose' Izuogu.

I am grateful to my parents and parents-in-law for their encouragement and support in my pursuit of this degree, and to my wife, Nanthakorn, for sacrificing three years of her career to accompany me here and look after our children. I thank my daughter and son for keeping me company, and for the joy they brought while they were here.

Table of Contents

List of Abbreviations

ARP	affected relative pair
ASP	affected sib pair
BP	base pair
cM	centiMorgan
dBP	diastolic blood pressure
FASTA	family based score test approximation
FBAT	family-based association test
GAW18	18th Genetic Analysis Workshop—one of the data sets used
GC	genomic control
GWAS	genome-wide association study
HLA	human leucocyte antigen
HPC	high-performance computing
HWE	Hardy-Weinberg equilibrium
IBD	identity (or identical) by descent
IBS	identity (or identical) by state
LMM	linear mixed model
MAF	minor allele frequency
MCMC	Markov chain Monte Carlo
MDS	multidimensional scaling
ML	maximum likelihood (*cf.* REML)
MLS	maximum likelihood ratio statistic
MQLS	maximum quasi-likelihood statistic
PC	principal component
PCA	principal component analysis
QTL	quantitative trait locus
REML	restricted maximum likelihood (*cf.* ML)
RIA	regional IBD analysis (i.e. the genomic linkage analysis method proposed in this thesis)
ROADTRIPS	robust association-detection test for related individuals with population substructrue
RRM	realised relationship matrix—a type of kinship matrix used in FaST-LMM
SAFS	San Antonio Family Studies (from which the GAW18 data set was derived)

sBP	systolic blood pressure
SNP	single nucleotide polymorphism
SVD	singular value decomposition
VL	Brazilian visceral leishmaniasis—one of the data sets used
VUR	vesicoureteral reflux disease—one of the data sets used
WTCCC	Wellcome Trust Case-Control Consortium

Chapter 1. Introduction

1.1. Relatedness and IBD Sharing Estimation

Many aspects of life, such as family, marriage or inheritance rely on establishing genetic relatedness among individuals (Weir *et al.*, 2006). Although one's relatedness to another is usually intuitive, there are circumstances where formal quantification of relatedness is useful. Methods have been developed for this purpose since the early part of the last century.

A very simple and intuitive approach is perhaps to state that one individual is related to another if they both share at least one common ancestor (Malécot, 1969), but this may not be very useful. A more informative analytical approach to assess relatedness was developed by Sewell Wright in 1922, improving on an earlier idea on inbreeding by Pearl (1914) and Ellinger (1920). Under the then prevailing correlation analysis framework, Wright defined his *coefficient of relationship* as the coefficient of genetic correlation between two individuals (Wright, 1921; Wright, 1922). The approach appeared not so popular (at least among the biologists), however, perhaps because Wright's method requires familiarity with its mathematical framework, and appeared to be no longer appropriate for the genes that became available for study by the middle of the 20th century (C. C. Li and Sacks, 1954; Morton, 1969).

By that time, Gustav Malécot had created a different measurement of relatedness using a more intuitive probabilistic framework. In his 1948 seminal work, *Les mathématiques de l'hérédité* (translated into English as *The Mathematics of Heredity* in 1969), Malécot defined *coefficient of coancestry*—also known as *kinship coefficient* or *coefficient of consanguinity* (Blouin, 2003; Oliehock *et al.*, 2006; Weir *et al.*, 2006)—as the probability that two homologous alleles, each chosen randomly from each of the individuals in the pair of interest, are 'identical, i.e., are descended from the same [allele]' (Malécot, 1948, as translated by Yermanos in Malécot, 1969).

The word 'identical' on its own is in fact quite ambiguous. Eight years before Malécot published his work, Charles Cotterman identified three categories of genetic identity *sensu lato* in his thesis (Cotterman, 1940)—'identity' (*sensu stricto*), which seemed to refer to phenotypic effect in what could be loosely described as genocopy in modern terminology; 'derivatives', i.e. genes that are similar because they are 'derived' from a single ancestral gene; and 'alleles', defined as genes sharing identical locus (i.e. the similarity is in the locus, not the sequence). He further observed that these three aspects of identity are independent (in a logical sense, i.e. the state of one does not

imply the state of the other)—although some combinations could occur only under a very extraordinary circumstance (Cotterman, 1940). Cotterman's 'derivative' was the first time that what is now known, under Malécot's terminology, as *identity by descent* (IBD) was identified (Thompson, 1974); it continues with little change today: homologous alleles are now said to be IBD if they have descended from a single ancestral allele in a *recent* common ancestor (Blouin, 2003; Weir *et al.*, 2006; Astle and Balding, 2009; Powell *et al.*, 2010; Day-Williams *et al.*, 2011a; Ott *et al.*, 2011).

The word 'recent' in current the definition of IBD is worth a little discussion here. There is another type of identity in modern usage—*identity by state* (IBS). Homologous alleles that are apparently similar, regardless of their ancestry, are said to be IBS (Ott, 1999; Blouin, 2003; Powell *et al.*, 2010; Day-Williams *et al.*, 2011a). It follows that, in absence of mutation, IBD alleles will also be IBS, but the reverse is not necessarily true (C. C. Li and Sacks, 1954; McPeek and Sun, 2000). (The relationship between IBD and IBS indeed roughly follows the relationship between derivative and the other two identity categories in Cotterman's original work.) However, the distinction between IBD and IBS is in fact somewhat arbitrary: according to the coalescent theory, if one looks back far enough in time, most of the so-called IBS alleles would coalesce to some certain common ancestors and are therefore IBD; that is, unless they arose from separate mutation events, which would be rare (Cotterman, 1940; Blouin, 2003; Powell *et al.*, 2010). However, there is utility in making the distinction between the two (particularly in linkage analysis), and therefore, in practice, IBD is normally defined by recent common ancestors.

Yet, this definition itself may not resolve the ambiguity as it still begs another question—how recent is recent? If the definition of 'recent' is arbitrary, then the distinction between IBD and IBS will remain arbitrary. Indeed, Cotterman was aware of this predicament when he made the distinction between different types of identity. In his work, Cotterman (1940) suggested from a mainly biological viewpoint that the limit for 'recent' could be about 5-6 generations in human, with the main arguments that this timeframe is short enough for the chance of mutation to be negligible, but long enough that any effect of inbreeding prior to that is removed. He also pointed out a few additional practical advantages of this suggestion. Nevertheless, modern approaches seem to take an even more pragmatic and utilitarian stance that this should depend on the purpose of the estimation, or, in fact, may even be dictated by the data set being used (A. D. Anderson and Weir, 2007; Astle and Balding, 2009; Powell *et al.*, 2010).

It emerged that the idea of relatedness itself relies on the concept of IBD. Even our first, simplistic view of relatedness implies IBD sharing: two individuals can be said to be related if they share at least one allele IBD (Weir *et al.*, 2006)—the shared ancestry is

just a proxy for this. As for the coefficient of relationship and kinship coefficient, both, in fact, either explicitly or implicitly reflect the underlying degree of IBD—the kinship coefficient does so by definition; the coefficient of relationship, although not apparent from its original formulation, can be reformulated under a probabilistic framework as the *proportion of IBD alleles shared* between two individuals (C. C. Li and Sacks, 1954; Blouin, 2003).

Instead of specifying IBD sharing using one of the above summary coefficients, probability can alternatively be assigned to each of the possible IBD sharing classes between two individuals. To do so fully in a pair of diploid organisms requires 15 such classes (Jacquard, 1972). However, if the two chromosomes in each individual are treated as unordered (i.e. disregarding the parent of origin), then 6 redundant IBD sharing classes can be removed. Furthermore, if inbreeding is assumed to be absent, these can further be collapsed to just three classes, representing 0, 1 and 2 IBD alleles shared among the individuals (Jacquard, 1972; Thompson, 1974; A. D. Anderson and Weir, 2007; Astle and Balding, 2009).

There are many use of IBD sharing estimates (Oliehock *et al.*, 2006; Weir *et al.*, 2006; A. D. Anderson and Weir, 2007; Browning and Browning, 2011; Han and Abney, 2011). In context of genetic data analysis (or genetic mapping), IBD probabilities are central to linkage analyses (Day-Williams *et al.*, 2011a), and can also be used to control for the effect of *population substructure* in association studies (Bacanu *et al.*, 2000; Balding, 2006; Purcell *et al.*, 2007; Kang *et al.*, 2008; Thornton and McPeek, 2010). Being able to accurately measure IBD sharing probabilities is therefore very useful.

However, the IBD allele sharing probabilities (and their related coefficients) cannot be 'measured' directly, and have to be estimated (Weir *et al.*, 2006). Traditionally, they are estimated analytically from the pedigree structure; alternatively, if genotype data are available for a particular genetic location, these can be used in conjunction with the pedigree information to estimate the IBD. However, the IBD estimates in the latter case will be *local* to that location, whereas the IBD estimates in the former, which do not rely on any genotype data, will be *global*, i.e. will correspond to the theoretically expected IBD at any locus in that pair of individuals (Day-Williams *et al.*, 2011a). These two types of relatedness estimates need not be equal. In fact, non-parametric linkage analysis is only possible precisely because of the disparity between local and global IBD probabilities that can be expected under the conditions of linkage (Elston, 1998; Shih and Whittemore, 2001).

With increasing availability of genetic data, it becomes possible to estimate the IBD based only on the genotype data without having to relying on pedigrees (Milligan, 2003). These 'empirical' estimates can be based on maximum likelihood estimators

(MLE, e.g. Thompson, 1975; Milligan, 2003; A. D. Anderson and Weir, 2007), method of moments estimators (MME, e.g. Ritland, 1996; Purcell *et al.*, 2007; Manichaikul *et al.*, 2010), or some other methods (e.g. Queller and Goodnight, 1989; Lynch and Ritland, 1999; Wang, 2002; Day-Williams *et al.*, 2011a). These methods have their advantages and disadvantages, and there does not seem to be one that is best in all situations (Milligan, 2003; Astle and Balding, 2009). Nevertheless, it is not the aim of this thesis to compare the relative merits of empirical IBD estimation methods; rather, some method-of-moments empirical IBD estimates will be used (mainly for practical reasons) to assess the relative merits of analytic methods of interest.

Originally, genetic data analysis methods that use IBD information were designed for use with theoretical IBD estimates. However, there may be some advantages of using empirical instead of theoretical IBD probabilities. For example, the pedigrees of the samples may not be known, or may not be accurate. Even when the pedigree is known, there could still be some advantages in using empirical IBD. Traditionally, the founders of the pedigrees are treated as unrelated in this analytical approach (technically, the founders form the *base population* from which the relatedness is measured; this makes them unrelated by definition), but there would inevitably be some degree of relatedness among them, which could result in some inaccuracies (Weir *et al.*, 2006; Astle and Balding, 2009; Powell *et al.*, 2010; Day-Williams *et al.*, 2011a). Furthermore, with complex pedigrees, it may be impractical to calculate their theoretical IBD; in which case, empirical, genetically estimated IBD may be useful (Han and Abney, 2011). In this thesis, the merit of using empirical IBD estimates in various genetic data analysis methods will be investigated.

1.2. Application of IBD sharing estimation in genome-wide association studies

Genetic association studies examine the association between a particular allelic variant and the trait of interest. This association can occur not only when the variant is truly causal, but also when it is in linkage disequilibrium with the nearby causal allele (Lander and Schork, 1994), so the method could also naturally extend to mapping the locus of interest (Astle and Balding, 2009). Technical limitations in the past restricted the practicability of this class of methods to a limited set of markers or variants within a predetermined set of candidate genes. However, recent ability to genotype a dense set of polymorphic markers combined with knowledge about their positions and linkage disequilibrium structures has enabled the genome-wide association (GWA) approach, which is based on the examination of hundreds of thousands of possible associations between the phenotype of interest and markers across the whole genome. Linkage disequilibrium association mapping based on genome-wide association consequently allows mapping of the causal locus to a much higher theoretical resolution than linkage mapping (Hirschhorn and Daly, 2005; Astle and Balding, 2009; Ott *et al.*, 2011;

Visscher *et al.*, 2012). Additionally, it has also been demonstrated to be more powerful than linkage analysis when investigating common variants with weaker effects (Risch and Merikangas, 1996). Association studies are therefore currently more predominant than linkage analyses (Astle and Balding, 2009; Ott *et al.*, 2011). However, they can be subject to certain biases including that due to population substructure, as will be discussed below.

1.2.1. Effect of population substructure in genome-wide association studies

The term *population substructure* or *population structure* is defined as 'sample structure due to differences in genetic ancestry among samples' (Price *et al.*, 2010). It appears to have two levels of meaning in the literature: in a narrow sense, the differences are limited to those originating from distant ancestry and the term is essentially synonymous with *population stratification* (Ewens and Spielman, 1995; Pritchard *et al.*, 2000b; Astle and Balding, 2009); in a broader sense, it also includes the sample structure caused by much more recent ancestry i.e. close relatedness among the samples (McCarthy *et al.*, 2008; Price *et al.*, 2010; Zheng *et al.*, 2010).

Population stratification refers to the situation where there are subgroups of individuals with different ancestry within the population. Usage in the context of genetic association studies tends to be consequential, i.e. it implies differences in allele frequencies due to systematic ancestry differences between the cases and the controls (Freedman *et al.*, 2004; Voight and Pritchard, 2005; McCarthy *et al.*, 2008; Astle and Balding, 2009).

Relatedness among the samples may be known to the researchers in advance through relationships within the family (*'familial relatedness'*). Alternatively, relatedness among some individuals may not be known to the researchers or even to themselves. *Cryptic relatedness* refers to the situation where, as a result of the samples' relatedness which is not known to the researchers, their genotypes are not independent (Devlin and Roeder, 1999; Voight and Pritchard, 2005; McCarthy *et al.*, 2008; Astle and Balding, 2009).

For clarity, the term *population substructure* will be used here in its broader sense; *population stratification* will refer specifically to the sample structure caused by differences in genetic ancestry within populations; and *relatedness* will refer to all not-too-distant (i.e. within the same subpopulation) genetic relatedness, whether apparent or cryptic.

In its simplest form, the design of a genetic association study follows a case-control approach, looking for differences in allele frequencies between the cases and the controls at each particular locus (Hirschhorn and Daly, 2005). In an outbred

population, in which linkage disequilibrium decays rapidly over distance, this association implies tight physical linkage between the marker and causative loci (Pritchard and Rosenberg, 1999; Pritchard and Donnelly, 2001). However, this is not necessarily the case in presence of population stratification (Lander and Schork, 1994; Ewens and Spielman, 1995).

Unlike linkage analyses, in which the subjects are explicitly required to be related and are modelled as such, case-control association analyses assume that the subjects are from a single, homogeneous population, and that they are not related. Violation of these assumptions occurs in the presence of population substructure: violation of the former in the presence of population stratification; and violation of the latter in the presence of cryptic relatedness (Devlin and Roeder, 1999; Astle and Balding, 2009). These two types of population substructure are in fact two extremes of the same problem: the unobserved relationships among the samples (Astle and Balding, 2009; Kang *et al.*, 2010). Ultimately, alleles in any pair of samples would coalesce into a certain ancestor. If that common ancestor is distant, the samples are said to be from different populations; if the common ancestor is recent, then the samples are more or less related (known or cryptic). The rather arbitrary nature of 'distant' and 'recent' aside, it would appear that, one way or another, there will always be a certain degree of population substructure within any data set. However, this is not necessarily consequential.

Any true association observed in an association study ultimately reflects different genetic ancestry between the groups at that locus, regardless of the actual linkage between that locus and the disease locus (Ewens and Spielman, 1995; Rosenberg and Nordborg, 2006; Astle and Balding, 2009). Because population stratification can also cause the cases and controls to cluster into different subpopulations, which also reflects different ancestry, it could cause apparent association in absence of a true genetic effect (Astle and Balding, 2009).

A famous example of spurious association due to population stratification is that between the Gm system haplotype of human immunoglobulin G and risk of type 2 diabetes in a study in Pima Indians (Knowler *et al.*, 1988). (For historical accuracy, it should be noted here that the issue was identified by Knowler *et al.* themselves, who also performed appropriate analyses to demonstrate this issue). This was, however, a rather extreme case. When one locus is considered at a time and known ancestry is accounted for, the impact of population stratification observed in most association studies has so far been rather modest even for relatively large studies (e.g. Clayton *et al.*, 2005; The Wellcome Trust Case Control Consortium, 2007). This has led to an assertion that the impact of population stratification would be minimal provided that

the cases and controls were appropriately matched for broad ethnicity, and the individuals whose genomic data subsequently revealed substantially different ethnicity were excluded (McCarthy *et al.*, 2008).

Nevertheless, the results from the theoretical analyses by Ewens and Spielman (1995) and Pritchard and Rosenberg (1999) and a simulation study by Marchini *et al.* (2004) and Zheng *et al.* (2010) showed that the effect of population stratification increased in the same direction as the sample size and the relative risks of the disease among the subpopulations. As the size of genome-wide association studies increases to enable the detection of smaller genetic effects, population stratification is more likely to become a problem (Price *et al.*, 2006; Astle and Balding, 2009).

In cryptic relatedness, samples are related without the researchers' necessarily being aware of it. Devlin and Roeder (1999) argued that the very act of selecting the cases and controls could introduce this error as the cases would necessarily share similar genotypes and therefore would be more closely related to each other than to the controls; however, Voight and Pritchard (2005) later pointed out that this could be mitigated by natural selection. The degree of relatedness increases in small, isolated populations as the chance of inbreeding increases, and the limited sample choice means that each sample is more likely to be related to the others (Newman *et al.*, 2001; Voight and Pritchard, 2005). Nevertheless, some degree of inbreeding exists even in large, apparently outbred populations (Broman and Weber, 1999).

The relatedness of samples implies that their genotypes are not independently sampled. Although this should not affect the allele frequency estimates in the cases or controls, the variances would be underestimated, leading to inflation of the test statistics (Devlin and Roeder, 1999; Voight and Pritchard, 2005; Zheng *et al.*, 2010). Devlin and Roeder (1999) suggested that this issue may in fact be more important than population stratification in causing the inflation of the results. A study in a founder population has empirically confirmed the impact of this (Newman *et al.*, 2001). However, in their more detailed study, Voight and Pritchard (2005) concluded that the effect of cryptic relatedness is important only in a small, rapidly expanding, or heavily inbred populations; it is negligible in large outbred populations.

1.2.2. Using LMM to mitigate the effect of population substructure

A large number of methods have been proposed to reduce the effect of population substructure in genetic data analyses (perhaps reflecting the importance of the issue). Some examples of these include: sample restriction (to similar ethnicity) (McCarthy *et al.*, 2008), family-based tests of linkage and association (FBTLA) (e.g. Falk and Rubinstein, 1987; Terwilliger and Ott, 1992; Spielman *et al.*, 1993; Rabinowitz and Laird, 2000), genomic control (GC) (Devlin and Roeder, 1999; Bacanu *et al.*, 2000;

Reich and Goldstein, 2001; Cardon and Palmer, 2003; Astle and Balding, 2009), structured association (SA) (Pritchard *et al.*, 2000a; Pritchard *et al.*, 2000b; Purcell *et al.*, 2007; Alexander *et al.*, 2009), principal component analysis (PCA) and related methods (e.g. Menozzi *et al.*, 1978; S. Zhang *et al.*, 2003; Price *et al.*, 2006; Aulchenko *et al.*, 2007b; Purcell *et al.*, 2007; Q. Li and Yu, 2008) and robust association-detection test for related individuals with population substructure (ROADTRIPS: see Section 4.1.2 for detail) (Thornton and McPeek, 2010).

A theoretically attractive class of methods which has become rather successful in controlling for population substructure in GWAS is linear mixed modelling (LMM). Mixed modelling derived from earlier works in animal breeding (Henderson *et al.*, 1959; Kennedy *et al.*, 1992; George *et al.*, 2000; Yu *et al.*, 2006; Kang *et al.*, 2008). It captures the effect of population stratification and relatedness in a way very specific to the data set being analysed by using the kinship matrix to model the random effect part of a standard mixed model, while the candidate SNP and any additional covariates are modelled as fixed effects (Astle and Balding, 2009; Price *et al.*, 2010; Yang *et al.*, 2014).

Generally, methods in this class attempt to fit the model:

$$\boldsymbol{Y} = X\boldsymbol{\beta} + Q + \varepsilon$$

where $\boldsymbol{Y} = (y_1, \ldots, y_n)^T$ is a vector of responses on n subjects, $X = (x_{ik})$ is the $n \times K$ matrix of predictor values for variables to be modelled as fixed effects (including covariates and genotypes at any SNPs currently under test), $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)^T$ are regression coefficients (to be estimated) representing the linear effect of the predictors on the response, Q are random effects which follow the multivariate normal distribution $Q \sim N(0, 2\sigma_g^2\phi)$, ε are normally distributed random errors, $\varepsilon \sim N(0, \sigma_e^2 I)$ where σ_g^2 and σ_e^2 are parameters (to be estimated) representing the genetic and environmental components of variance respectively, ϕ is the $n \times n$ matrix of pairwise kinship coefficients ('kinship matrix') and I is the $n \times n$ identity matrix. Originally, the kinship matrix used was pedigree-based (Astle and Balding, 2009), and would therefore only be useful for correcting for the effect of familial relatedness, or, with an explicit additional modelling, of population stratification. Recently, genomic-based kinship estimates have also been used (Yu *et al.*, 2006). This enables the correction of population stratification as well as both familial and cryptic relatedness without explicit modelling effort.

Apart from its ability to account for both types of population substructure without requiring prior knowledge, Yang *et al.* (2014) also noted that it can have higher power even in samples without structure due to the implicit inclusion of the effects of weakly associated SNPs, which would not otherwise be included, into the model through the

polygenic random effect. Its main disadvantage is the slow speed, as it is computationally intense (Kang *et al.*, 2010). Nevertheless, several implementations that use certain simplifications or approximations to address this issue are now available, such as GenABEL's `polygenic`/`mmscore` functions (Aulchenko *et al.*, 2007b), which implement the FASTA method proposed by Chen and Abecasis (2007), GRAMMAR-Gamma (Svishcheva *et al.*, 2012), EMMA (Kang *et al.*, 2008), EMMAX (Kang *et al.*, 2010), TASSEL (Z. Zhang *et al.*, 2010), GEMMA (Zhou and Stephens, 2012), MMM (Pirinen *et al.*, 2013) and FaST-LMM (Lippert *et al.*, 2011; Listgarten *et al.*, 2012; Lippert *et al.*, 2013).

In general, LMM methods tend to perform well when compared with methods from other classes (Kang *et al.*, 2010; Price *et al.*, 2010; Liu *et al.*, 2011; Peloso *et al.*, 2011; Sawcer *et al.*, 2011). However, it is more difficult to judge which method is better within the LMM class, or if there is any difference at all. Direct comparisons that have been made among LMM methods are as follows:

Lippert *et al.* (2011) used both synthetic data constructed from Genetic Analysis Workshop 14 (GAW14) data and real Crohn's disease data from the WTCCC to compare their program, FaST-LMM, and EMMAX and found the results were comparable although FaST-LMM required less resource and run time.

Using real and simulated data from a young isolated Dutch population (Pardo *et al.*, 2005), Svishcheva *et al.* (2012) compared their proposed method, GRAMMAR-Gamma, with EMMAX, FASTA (as implemented in GenABEL's `mmscore` function), FaST-LMM and FMM (Astle, 2009, cited by Svishcheva *et al.*, 2012); and also compared GRAMMAR-Gamma with FASTA using an *Arabidopsis thaliana* data set. They found that results from all methods were comparable, but GRAMMAR-Gamma required much less run time.

Zhou and Stephens (2012) compared their LMM method, GEMMA, to EMMA, EMMAX, FaST-LMM and what they called 'GRAMMAR' (in reference to the `grammar` function in GenABEL, although it is not entirely clear if this means GRAMMAR-Gamma or the original GRAMMAR), using mouse high-density lipoprotein-cholesterol (HDL-C) level data from the Hybrid Mouse Diversity Panel (HMDP) (Bennett *et al.*, 2010) and the Crohn's disease data from the WTCCC. They found that, in the highly related HMDP data set containing strongly associated SNPs, the approximation methods (EMMAX and 'GRAMMAR') showed deflation of test statistics (which was particularly remarkable in 'GRAMMAR'). In the WTCCC Crohn's data set, GEMMA and EMMAX results were comparable, while results from GRAMMAR showed a slight deflation. In terms of speed, GEMMA was the fastest among the exact methods, whereas 'GRAMMAR' was the fastest among the approximation methods.

Pirinen *et al.* (2013) reported high concordance between the heritability and variance estimated from their software, MMM, with EMMA using simulated data sets. In terms of run time, MMM, using either the exact method or the GLS approximation, was faster than EMMA and FMM.

At least a few observations can be made here. Firstly, different sets of comparisons were using different data sets, so comparison across studies is not possible, except for Lippert *et al.* (2011) and Zhou and Stephens (2012) which both use the WTCCC Crohn's data set as part of their analyses; however, since the methods used in Lippert *et al.* (2011) are also used in Zhou and Stephens (2012), no additional information can be gained. It would be useful if all methods could be compared based on a single, similar data set.

Secondly, the data sets with any non-trivial degree of (known) relatedness were all rather unusual in some ways: the GAW14 synthetic data set created by Lippert *et al.* (2011) consisted of up to a hundredfold copies of the original GAW14 familial data, totalling about 120,000 individuals with high degree of redundancy; however, this data set was used only for evaluation of computational resources consumption, so the actual redundancy is probably not relevant; the mice in the HMDP (Bennett *et al.*, 2010) were heavily related and inbred; the isolated Dutch population were descended from a small number of founders and had a substantial inbreeding coefficient (Pardo *et al.*, 2005) and the simulated kinships in Pirinen *et al.* (2013) were completely random, and may not necessarily be biologically plausible (by way of example, the simulation did not require that the relatedness between individuals A and B be biologically consistent with that between A and C and between B and C).

Furthermore, although the comparisons among the LMM methods generally showed good correlation among the results, most did not conduct formal assessment of power or type I error rates.

Finally, none of these studies addressed the issue of usability. Apart from accuracy of results and resource requirements, the decision to use one method over the others could also be influenced by its usability.

I shall explore these issues in Chapters 3-5 through the use of real and simulated extended family data sets from outbred populations. Additionally, the advantage (or not) of using empirical relatedness estimation over theoretical estimation in LMM will also be explored.

1.3. Application of IBD sharing estimation in non-parametric linkage analysis

By analysing the cosegregation pattern between the disease and marker loci within each family, linkage analysis has successfully been used to locate the causal loci of many

genetic disorders (Nyholt, 2008; Visscher *et al.*, 2012). This type of study is the most suitable for studying a rare Mendelian trait in a large family with unambiguous relationships, in which sufficient recombinations can occur, and the genetic effect of the variant is strong and can be ascertained with high accuracy. When these conditions are satisfied, this approach can be very efficient, requiring only a relatively small set of markers to cover the whole genome and no prior knowledge of the likely location or type of the causal variant or the disease mechanism (Lander and Schork, 1994; Sham, 1997; Astle and Balding, 2009). Consequently, such studies have been instrumental in the identification of the causes of many high-penetrance genetic disorders (Astle and Balding, 2009; Visscher *et al.*, 2012). However, when applied to the study of more common complex traits caused by multiple genetic and environmental factors, this class of methods was less successful because the effect of each individual locus is generally too weak to be detected (Risch and Merikangas, 1996; Hirschhorn and Daly, 2005).

The relatively low number of recombinations that normally occur in a given meiosis has two important implications in linkage analysis. Historically, this means that relatively few markers are required to map the locus of interest to a chromosomal region even without any prior knowledge. However, it also means that, despite advances in technology which have led to the availability of massively increased number of genetic markers over the past decade, the potential for improving the resolution of the mapping is limited: once every recombination that occurred in the samples has been identified, genotyping of additional markers will not improve the mapping resolution any further (Risch and Merikangas, 1996; Astle and Balding, 2009; Visscher *et al.*, 2012).

With these two disadvantages, and with increasing power of GWAS, the popularity of linkage analysis as a tool for genetic mapping has declined, and it is currently less favoured than GWAS. However, the move to study rare disease-causing variants has again revived interest in linkage analysis, as these rare variants would be likely to cluster within particular families (Astle and Balding, 2009; Ott *et al.*, 2011), in which case linkage analysis is more advantageous.

Whilst linkage analyses may not benefit from the advances in technology which have increased the availability of genetic markers in terms of increasing their resolution, they may still benefit from such advances in another way: by increasing the accuracy and practicability of identity-by-descent (IBD) sharing estimation.

Unlike their parametric cousins, the various forms of traditional non-parametric linkage analyses generally compare the observed IBD sharing pattern within a pair or group of affected relatives at a particular locus with those expected from individuals with similar type of relatedness under the null hypothesis of no linkage between that

locus and the trait of interest. This approach has the advantage of not requiring the mode of inheritance to be specified (Whittemore and Halpern, 1994; Kruglyak *et al.*, 1996; Kong and Cox, 1997; Elston, 1998; Ott, 1999; Basu *et al.*, 2008; Nyholt, 2008). Nevertheless, these methods require accurate specification of IBD sharing to obtain valid results (Boehnke and Cox, 1997; Shih and Whittemore, 2001; Albers *et al.*, 2008; Day-Williams *et al.*, 2011a). Traditionally, these methods rely on the theoretical IBD estimates based on either the pedigree information alone, or pedigree information combined with genotypic information at a series of markers. However, in theory, it should be possible to either enhance or completely replace the theoretical IBD estimates by using genotype data.

Indeed, since the late 1990's, various methods based on an earlier work by Elizabeth Thompson (1975) have allowed the use of genotype data in conjunction with pedigree data to improve the accuracy of IBD estimation for linkage analysis (e.g. Boehnke and Cox, 1997; Ehm and Wagner, 1998; Epstein *et al.*, 2000; McPeek and Sun, 2000; Sun *et al.*, 2002). More recently, methods that allow estimation of relatedness based solely on genotype data have also been developed (see Section 1.1). Being able to do so is appealing as it can bypass the computational problem in complex pedigrees.

A popular class of non-parametric linkage analyses is *affected sib pairs* (ASP) methods (e.g. Cudworth and Woodrow, 1975; Suarez *et al.*, 1978; Risch, 1990), which has the benefits of having good power and not being affected by incomplete penetrance, unlike earlier methods, at the expense of not being able to use other type of relatives (Ott, 1999). For methods in this class, theoretical estimates worked well, as the expected IBD probabilities were known with certainty (provided that the relationship had been correctly ascertained in the first place), and the estimation of the observed IBD sharing probabilities was relatively straightforward.

A natural extension of ASP methods is to use any arbitrary pairs of affected relatives (*affected relative pair* (ARP) methods (e.g. Weeks and Lange, 1988; Risch, 1990; Kruglyak *et al.*, 1996; Kong and Cox, 1997; McPeek, 1999; Shih and Whittemore, 2001)). This is especially useful in linkage studies of complex disease because the lower penetrance makes finding affected sib pairs more difficult (Albers *et al.*, 2008). However, accurate IBD estimation in this type of study is more difficult as it requires complete knowledge of the pedigree structure, and is computationally more complex especially in large pedigrees or when some information is missing (Kong and Cox, 1997; Albers *et al.*, 2008; Bellenguez *et al.*, 2009; Day-Williams *et al.*, 2011b).

The problem arises because the Lander-Green algorithm (Lander and Green, 1987) used by many programs for exact enumeration of inheritance vector because of its ability to handle large amount of markers from dense SNP chips has a limitation on the

size of family it can handle (Lander and Green, 1987; Abecasis *et al.*, 2002; Albers *et al.*, 2008). For larger pedigrees, approximation methods using Markov chain Monte Carlo (MCMC) sampling are normally used (Shih and Whittemore, 2001; Albers *et al.*, 2008; Day-Williams *et al.*, 2011b).

Since the problem lies with the difficulty in estimating IBD in presence of complex pedigree structures, a method that can estimate the IBD without using pedigree structure and use the estimates in linkage analysis can potentially bypass this. In fact, in the last few years, two quantitative non-parametric linkage analysis methods that are conceptually similar to this (although in a slightly different context) have been proposed. Both involve estimating the local IBD sharing probabilities, which are specific to a small area of the chromosome and function as the 'observed' IBD sharing in the traditional sense of linkage analysis, and the global IBD sharing probabilities, which are calculated across the whole genome of each pair of individuals and function as the 'expected' IBD sharing in the traditional linkage analysis. These are then fed into a variance component model as random effects to perform quantitative linkage analysis.

The first method (Day-Williams *et al.*, 2011a) implements fast estimators for global and local kinship coefficients (as opposed to the full three IBD states) based solely on genomic data. The estimated genetic kinship coefficients are then used in a variance component model for quantitative trait locus (QTL) mapping. In this method, the kinship coefficient Φ between two individuals, u and v, is expressed in terms of the expected number of IBS matches between the two individuals, e_{uv}, over m SNPs, given the two allele frequencies, p_i and q_i, at each SNP i (Day-Williams *et al.*, 2011a):

$$\Phi_{uv} = \frac{e_{uv} - \sum_{i=1}^{m}(p_i^2 + q_i^2)}{m - \sum_{i=1}^{m}(p_i^2 + q_i^2)}$$

This requires the knowledge of the expected number of IBS matches, e_{uv}, which Day-Williams *et al.* equated to the observed number of IBS matches over all SNPs. The latter is the sum of the observed proportion of IBS matches in each SNP, i, which, for an autosomal SNP, is defined as:

$$o_{uv}^{i} = \frac{1}{4}[1_{(I_i=K_i)} + 1_{(I_i=L_i)} + 1_{(J_i=K_i)} + 1_{(J_i=L_i)}]$$

where the subscripted condition takes the value of 1 if an allele (I_i or J_i) in individual u is similar to an allele (K_i or L_i) in individual v, and 0 otherwise. An analogous relationship for the sex chromosomes was also described.

Unlike the global kinships, the local kinships in this method are imputed using a dynamic programming procedure which produces long stretches of uniform local IBD

states (resembling linkage structure within the chromosomes; see Day-Williams *et al.* (2011a) for details) instead of the method of moments estimator analogous to that used in global kinships as it was observed that the latter gave very noisy estimates when used with small SNP windows.

Since the primary motivation of this method was to avoid the difficulty in collecting pedigree information, a graph-based clustering algorithm is employed to automatically group the samples into families without having to rely on pedigree information. Day-Williams *et al.* noted that the use of this clustering algorithm is not strictly necessary (the analysis becomes population-based if it is not used), but opted to do so for computation efficiency.

These kinship calculation and clustering algorithms were successfully used in a variance component model to perform the QTL mapping of vannin 1 (*VNN1*) expression levels in the San Antonio Family Heart Study (SAFHS) data set (Mitchell *et al.*, 1996) without using the pedigree information, although they noted some reduction in the maximum LOD scores, which they attributed to the loss of information on specific relationship between each pair of individuals (Day-Williams *et al.*, 2011a).

The second method (Nagamine *et al.*, 2012) was actually developed as a way to perform association analysis while also incorporating 'regional' effects, although the underlying model is quite similar to that used in quantitative trait linkage analysis. Similar to the previous method, this method estimates the global and local ('regional') kinship coefficients from the genotype data. Both of these are calculated using a method of moments estimator:

$$f_{ij} = \frac{1}{n}\sum_{k}\frac{(g_{ik} - p_k)(g_{jk} - p_k)}{p_k(1 - p_k)}$$

where f_{ij} is the estimated global or local kinship coefficient between individuals i and j, calculated from the total of n SNPs, which is the number of genome-wide SNPs for global kinship calculation, or the number of SNPs within a small, local window (set to 100 in their article) for local kinship calculation. For a particular k-th SNP, g_{ik} (or g_{jk}) represents the genotype of individual i (or j), and p_k the major allele frequency at that SNP.

Since this method was conceived as a GWAS method, no attempt was made to group the samples into families, making it effectively a population-based linkage analysis. A variance component analysis was performed on the global and local kinship coefficients from all pairs of individuals by feeding them as random effects into a mixed effect model, which included other non-genetic covariates as fixed effects. Nagamine *et al.*

(2012) demonstrated the success of this method in the analyses of both simulated and real data sets.

Whilst these methods are generally suitable for quantitative trait linkage analysis, using a single kinship coefficient instead of full IBD states implicitly ignores the dominance effect and may not be suitable when dominance may be possible or when analysing qualitative data.

Additionally, the advantage of methods using genetically estimated IBD over those using theoretical IBD is potentially even greater in qualitative trait data analysis based on affected relative pairs methods than in quantitative trait data analysis, as the former involves the estimation of IBD sharing only among the affected individuals, whereas the latter requires the estimation of IBD sharing among all individuals.

In Chapters 6 and 7 of this thesis, I shall investigate the use of empirical IBD estimates in affected relative pairs (ARP) analysis based on maximum likelihood framework, which should be suitable for qualitative data.

Chapter 2. Material and Methods

This chapter describes the data sets used in this thesis and also the methods common among the remaining chapters, in particular, the phenotype simulations. It also gives a brief overview of statistical methods being investigated. However, methods specific to a particular part of the thesis will be described in detail in the relevant chapter.

2.1. GAW18 Data Set

The 18th Genetic Analysis Workshop (GAW18) data set was used in the early phase of this project to assess the performance of various GWAS methods in presence of family structure. The strength of this data set was due to its true longitudinal nature and readily available simulated phenotypes. This was a smaller data set compared to the Brazilian visceral leishmaniasis data set (VL; described in section 2.2), and, although it resulted in less power, had the advantage of taking less time and computational resources to analyse.

2.1.1. The GAW18 GWAS data set

The GAW18 data set was originally provided for use in the 18th Genetic Analysis Workshop (GAW18) in 2012, and also subsequently used in the 19th Genetic Analysis Workshop (GAW19) in 2014 with slight corrections. It was derived from an earlier version of a larger set of data collected as part of project 2 of the Type 2 Diabetes Genetic Exploration by Next-Generation Sequencing in Ethnic Samples (T2D-GENES) Consortium, and has been described in detail in a recent publication (Almasy *et al.*, 2014). To summarise, the T2D-GENES project 2 aims to study the low-frequency or rare susceptibility variants for type 2 diabetes through the application of whole genome sequencing (WGS) in 1,043 Mexican American individuals from 20 families from San Antonio, Texas. It is a subset of the larger San Antonio Family Studies (SAFS), chosen for the number of potential founder alleles, sequencing efficiency and number of individuals with type 2 diabetes. Individuals in this project had been genotyped on a variety of Illumina Infinium Beadchips platforms including HumanHap550v3 with HumanExon510Sv1, Human660W-Quadv1, Human1Mv1 and Human1M-Duov3 (although not everybody was genotyped on the same platform). Additionally, whole genome sequencing was done in about 600 strategically chosen individuals, with the sequences in the remaining individuals imputed based on their high-density SNP data. After quality control, 959 individuals (464 sequenced, 495 imputed) remained in the data set, from which the sequence (8,348,674 locations) and GWAS (472,049 SNPs) data on odd-numbered autosomes were extracted and provided for use in GAW18

(Almasy *et al.*, 2014). The analyses presented here used only the GWAS data from this data set.

Phenotype data provided in the GAW18 data set included age, gender, blood pressure, use of antihypertensive medications, hypertension status and current smoking status. Type-2 diabetes phenotype, which was the main phenotype of the T2D-GENES project, was not provided in this data set due to the agreement with the data provider. During the study period, participants in the SAFS were examined up to four times, resulting in up to four longitudinal measurements for each individual. Additionally, the GAW18 data provider generated 200 replicates of three measurements of simulated blood pressure phenotypes based on over 1,000 'causal' variants from more than 200 genes, selected based on real SAFS data, along with gender, age and medication status (Almasy *et al.*, 2014). Detail of the simulation model was intentionally withheld from general users of the data before the workshop, and the analyses presented in this thesis were conducted without this knowledge. Only the real and the first replicate of simulated phenotypes were used here.

The supplied data set was converted into PLINK's transposed file format (Purcell *et al.*, 2007) by Richard Howey. These were then converted into PLINK's binary file format for further analysis.

2.1.2. Quality control

The GAW18 data set was reported to have already undergone extensive quality control procedures (Almasy *et al.*, 2014). However, upon closer inspection, a few issues were identified in the GAW18 GWAS data, and a decision was made to conduct another full quality control procedure on this data set prior to further analyses.

The procedure used here was quite similar to that described by Anderson *et al* (2010), with an obvious exception that none of the related individuals was excluded. Also, because genotyping data on sex chromosomes were not provided, sex verification could not be done. The remaining quality control procedures can be divided broadly into two steps: per individual and per SNP quality control.

The per individual quality control steps consist of missingness and heterozygosity checks and ethnicity checks. These resulted in exclusion of four individuals due to their total lack of genotype data, and a further individual whose ethnicity seemed different from the others (more similar to the Hapmap JPT ('Japanese in Tokyo, Japan') or CHB ('Han Chinese in Beijing, China') populations rather than CEU ('CEPH [Utah residents with ancestry from northern and western Europe]') according to principal component analysis which included these and the YRI ('Yoruba in Ibadan, Nigeria') reference populations). Thus, 954 individuals remained in the final data set.

An interesting problem also arose during the missingness check: since the genotyping was done on different platforms among the sample, the maximum number of SNPs that can be called necessarily varied among them; however, as the information on the genotyping platform used for each individual was not provided, the individual 'call rate' had to be calculated against the number of SNPs across all platforms. This resulted in some individuals having apparently low call rates, only because they were genotyped on the less dense platforms. They would have been unnecessarily excluded had the standard, globally stringent, threshold been used to exclude individuals with apparently low call rates. On the other hand, lowering the threshold to accommodate this effect carries the risk of retaining too many individuals who had been genotyped on the denser platforms but whose genotyping quality would otherwise have been deemed unacceptable.

Of the 959 individuals in the original data set, 40 appeared to have uniformly 'low' call rate of approximately 0.56. They can be seen clustering into their own group, with the remaining 'high' call rate individuals also clustered into another group (Figure 2.1).

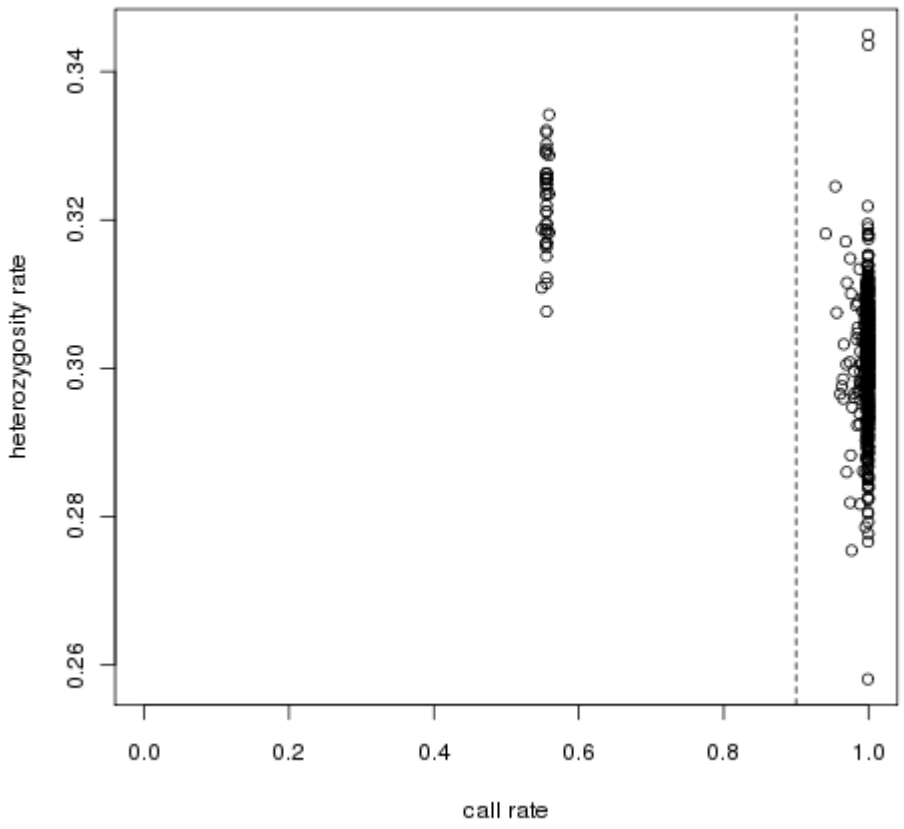

Figure 2.1 Heterozygosity rate vs call rate for each individual sample in the GAW18 data set, excluding the four individuals who have no GWAS data.

The per SNP missing rates also seem to corroborate this. The missing rates of the SNPs in each chromosome fell into two modes: just above 0.00 and slightly below 0.05 (Figure 2.2), the latter corresponds well with those SNPs being in the higher density platforms and therefore not genotyped in 40 'low call' individuals (44/959 = 0.046; this calculation included the 4 non-genotyped individuals).

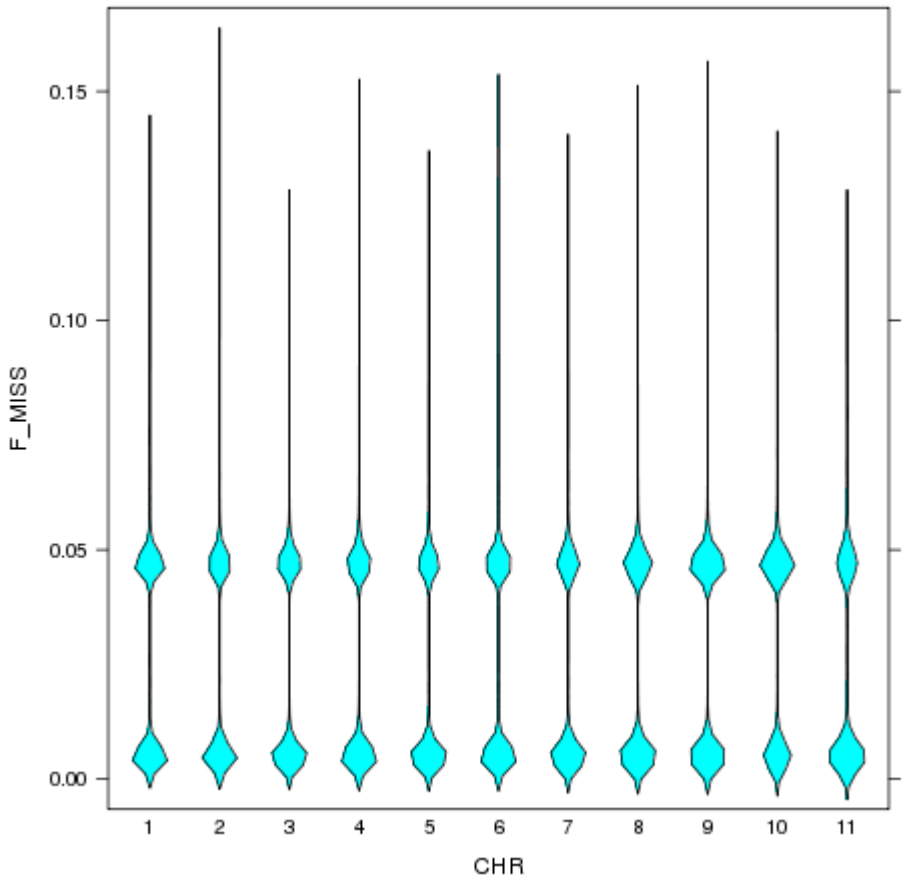

Figure 2.2 Violin plot showing the kernel density of missing rates (F_MISS) for SNPs in each chromosome (CHR). The width of the plot represents the kernel density at a particular missing rate.

Further checking confirmed that a very large proportion of the missing genotypes in these high-missing SNPs occurred in the 'low-call' individuals, thus confirming that these were the results of differences in genotyping platforms rather than a problem during the genotyping stage. It was therefore decided that the appropriate thresholds for exclusion should be determined separately between the two groups of individuals. Eventually, no sample was excluded on the basis of low call rate.

The per SNP quality control procedure involved checking for excessive missing data, violation of Hardy-Weinberg equilibrium, differential missingness and violation of Mendelian inheritance. The difference in genotyping platforms again posed a potential problem here due to the apparently high missing rates in certain SNPs in the denser platforms. However, it appeared that hardly any SNP had particularly high missing rate compared to their group. An overall missingness threshold of 0.10 was therefore chosen just to capture the relatively extreme cases, which resulted in exclusion of 109 SNPs.

In addition to these, 43,987 SNPs which either are monomorphic among the samples or have the minor allele frequency (MAF) of less than 1% were also excluded. These leave 427,953 SNPs in the final data set.

2.1.3. SNP reduction

In addition to the full genome-wide set of SNPs, the analyses performed also required a 'pruned' set of SNPs for IBD estimation. This was a set of 21,151 SNPs with MAF > 0.4, missingness <5% and in approximate linkage equilibrium, which was obtained by 'pruning' the full genotype data using the command '`--indep 50 5 2`' in PLINK (Purcell *et al.*, 2007).

The benefit of doing so, apart from reducing computational time, is that some methods for IBD estimation assume absence of linkage disequilibrium between the markers (Purcell *et al.*, 2007; Han and Abney, 2011), and would benefit from using SNPs that have been pruned so that no SNPs are in linkage disequilibrium.

2.2. The Brazilian Family Study of Visceral Leishmaniasis Data Set

The Brazilian Visceral Leishmaniasis (VL) data set was the main data set used in both the GWAS and linkage analysis parts of this thesis. It had much larger sample size compared to the GAW18 data set, and with larger families, making it ideal for this project.

2.2.1. The data set

This data set was collected in a family-based study in the cities of Belém and Natal in the north east part of Brazil. Access to this data set has been provided by Professor Jenefer Blackwell (University of Cambridge and University of Western Australia).

The sample collection and genotyping process of this data set has been described in detail by Fakiola *et al* (2013). To recap, 348 Brazilian families (65 from sites around Belém and 283 from sites around Natal) containing multiple members who had been diagnosed with clinical visceral Leishmaniasis were ascertained. The resulting pedigrees consist of 3,626 individuals in total; 2,159 of these (from 308 medium to

large families, 64 of which were from Belém and 244 from Natal) were genotyped at the Wellcome Trust Sanger Institute as part of the Wellcome Trust Case Consortium 2 (WTCCC2) project, using the Illumina Human660W-Quad chip. Extensive quality control procedures were performed on this genotype data to ensure only high quality samples were retained, and to exclude individuals whose apparent relatedness—assessed based on average genome-wide IBD, estimated using 11,177 high-quality autosomal SNPs via PLINK's `--z-genome` command (Purcell *et al.*, 2007)—was not compatible with their known pedigree relationship, and could not be resolved on further investigation (Fakiola *et al.*, 2013).

In the Online Methods section of their article, Fakiola *et al.* (2013) reported that 189 genotyped individuals were removed, leaving 1,970 Brazilian individuals for their analysis. However, in the demographic description of their samples (Supplementary Table 1 of the same article), the total number of samples was given as 1,972. The cause of this small discrepancy is not clear: it may be due to the removal of the two individuals whose phenotype was missing, or the removal of the two individuals whose heterozygosity was unusually low (see also Section 2.2.3 below). Regardless of the reason for the discrepancy, the data set being used here is inclusive of these individuals and therefore also consists of 1,972 individuals. Of these, 357 were affected, 1,613 unaffected and 2 were with missing phenotype.

The data have also undergone marker-wise quality control procedure to select only the SNPs that can be expected to be of high quality. In Fakiola *et al* (2013), SNPs were excluded if: their minor allele frequency was <0.01, the Fisher information for the allele frequency was <0.98, the call rate was <0.98, or they very clearly deviated from the Hardy-Weinberg equilibrium as demonstrated by p-value of $<10^{-20}$. This resulted in 553,323 autosomal SNPs being used in that article, out of the original 580,030. However, the data I received had undergone a further, slightly more stringent quality control by my supervisor, namely, excluding SNPs with call rate of <0.99 or having p-value of Hardy-Weinberg equilibrium of $<10^{-6}$. This resulted in my original data set containing 545,433 autosomal SNPs.

A few different subsets of individuals were used in the different analyses described here to satisfy the design and computational needs of the various methods. For GWAS or alternative methods requiring fully specified pedigree relationship, the entire set of 3,626 individuals was used, regardless of their genotyping status. For most LMM GWAS methods which do not require full pedigree specification, the set of all 1,972 genotyped individuals who had passed the quality control was used. For power comparisons between LMM methods, a subset of 462 'founder' individuals was also used, in addition to the previous sets of individuals. These were individuals who were

not known to be related and whose estimated kinships were also approximately unrelated. For the affected related pair linkage method (i.e. our proposed method RIA, described in Section 6.1), a subset of 1,960 genotyped individuals who were from families with at least two genotyped individuals who had passed quality control was used; for comparison, a subset of 3,370 individuals from those families were used in conventional linkage analysis methods.

2.2.2. Ethics statement

The local ethics committee at the Instituto Evandro Chagas, Belém, Paras, Brazil, granted the original ethical approval for the Belem Family Study. Continued use of these samples as well as collection and use of the Natal samples was approved by the local Institutional Review Board at the Universidade Federal do Rio Grande do Norte (CEP-UFRN 94-2004), and nationally by the Commisão Nacional de Ética em Pesquisa (CONEP: 11019). Shipping of samples out of Brazil was approved by the Ministerios Cencia e Tecnologia (portaria 617; 28 September 2005). Informed consent for sample collection was obtained in writing from adults and from parents of children under 18 years old (Fakiola *et al.*, 2013).

2.2.3. Quality control

As the data have previously passed a very stringent quality control procedure, the quality control steps done in this project are primarily to confirm the integrity of the data. It was intended that all samples and SNPs would be retained unless a significant deviation from any quality control criterion is encountered.

The procedure again broadly followed that of Anderson *et al.* (2010). However, sex verification was omitted as there was no genotype data on the sex chromosomes. The low call rate and outlying heterozygosity check revealed two samples with low heterozygosity (Figure 2.3), but this was confirmed to be due to consanguineous family history; they were thus retained.

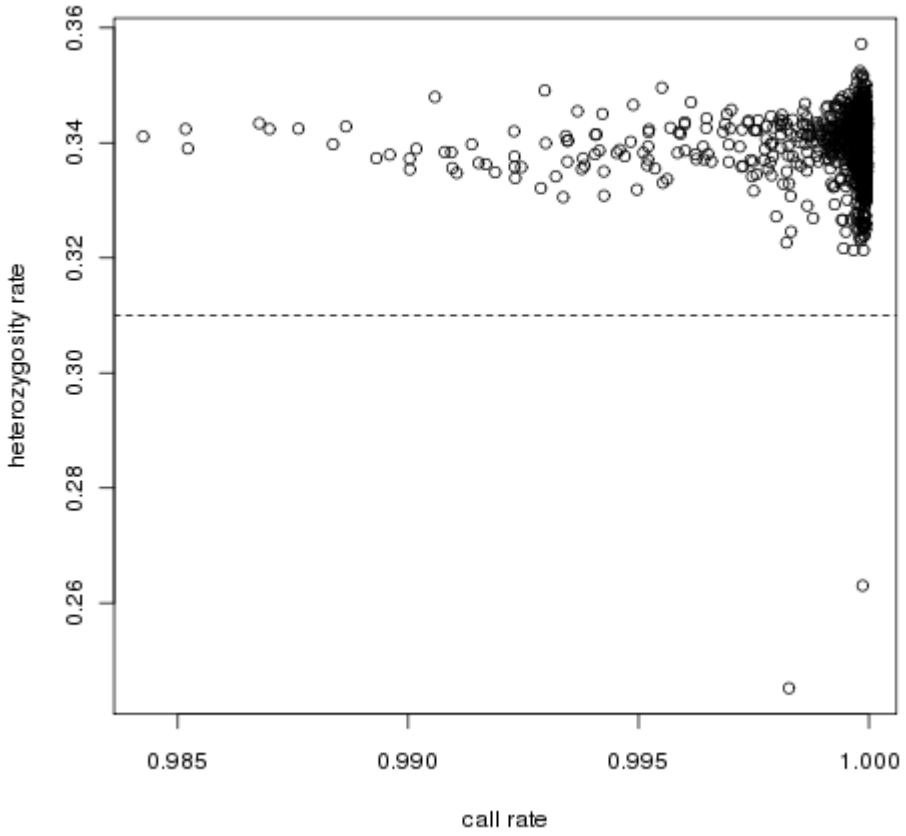

Figure 2.3 Heterozygosity rate vs call rate for each individual sample in the VL data set. Note the two outlying individuals with low heterozygosity due to consanguinity in their families.

The IBD and relatedness check was performed using 50,129 non-LD SNPs that were pruned down from the 100,488 SNPs with minor allele frequency of at least 0.4. This revealed that many samples were, as expected, related (Figure 2.4). Of note, five sample pairs appeared to be genetically identical. Verification with the pedigree data indicated that they were indeed twins, and were thus retained for further analysis.

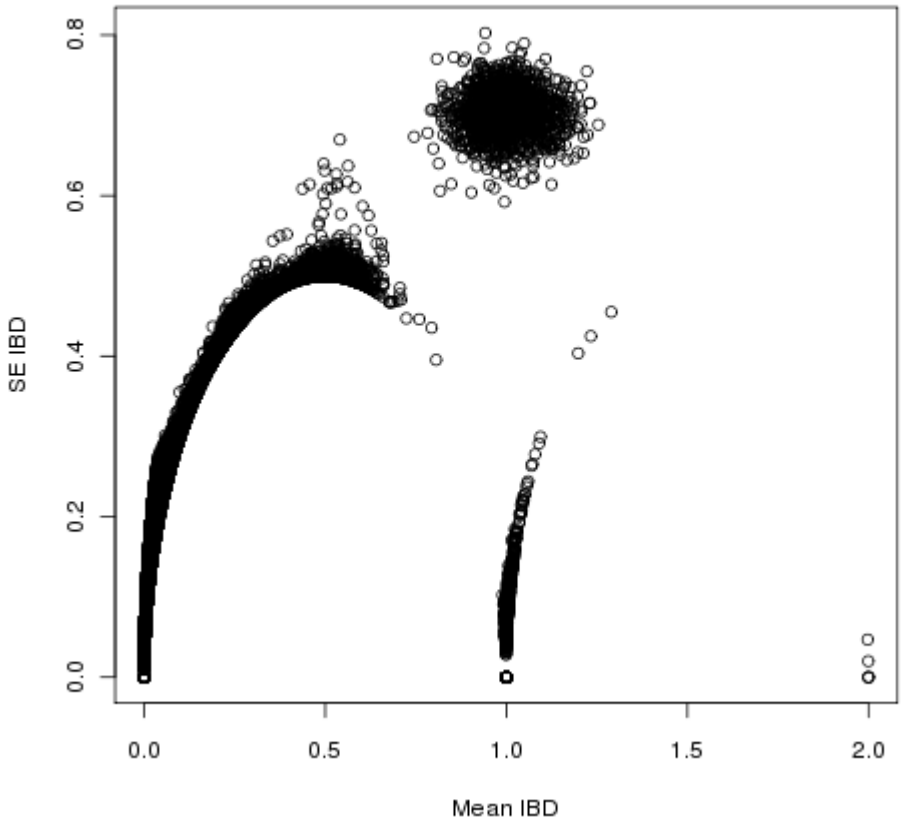

Figure 2.4 Mean and standard error of IBD sharing among the sample pairs in the VL data set.

The samples' ancestry was checked using 2-dimensional scaling similar to that described in the GAW18 data quality control (section 2.1.2 above). As shown in Figure 2.5, the ethnicity of the samples was between the European (CEU) and African (YRI) populations. No outlier was identified.

The ancestry was also checked using only the founders instead of all samples, as it is possible that the samples relatedness could cause problem with the PCA-like methods (Patterson *et al.*, 2006; Tian *et al.*, 2008). The results (not shown here) were in fact very similar to those in Figure 2.5.

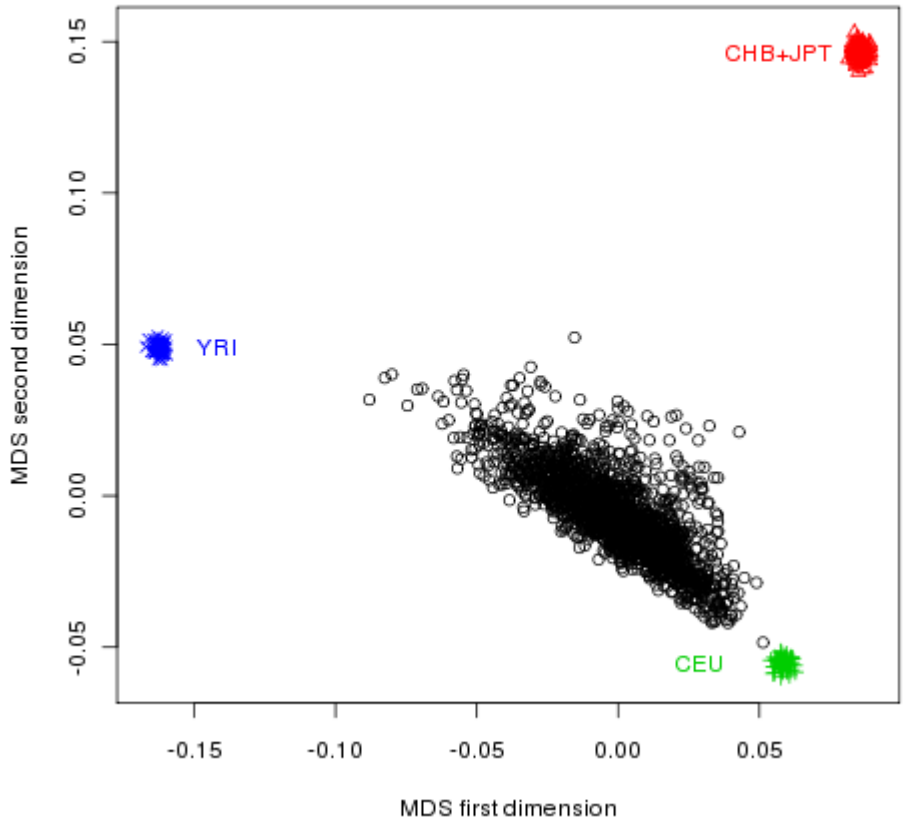

Figure 2.5 Two-dimensional scaling analysis of the VL genotype data with HapMap populations. (CEU = Utah residents with northern and western European ancestry, CHB = Han Chinese in Beijing, China, JPT = Japanese in Tokyo, Japan, YRI = Yoruba in Ibadan, Nigeria).

For the per-SNP quality control, PLINK was used to identify the founders' SNPs with allele frequency below 0.01, missing rate above 5%, or Hardy-Weinberg equilibrium (HWE) p-value below 10^{-8}. None failed the missingness or HWE test, although 172 SNPs were identified as having minor allele frequency of below 0.01, with the lowest being 0.0053. Comparison of the genotype missing rate between the affected and unaffected samples revealed one SNP with significantly different missing rates.

As this is a family-based dataset, a Mendelian error check using PLINK was also performed. This identified 79,457 errors which appeared to be inconsistencies caused by random genotyping errors averaging 0.1% over 30,928 SNPs, with rs7648971 having the highest error rate of 2.6%.

As all the problems flagged were relatively minor, and probably just reflected the slightly different calculations used within an already cleaned dataset, no SNP was

excluded. A list of the potentially problematic SNPs has been made so that they can be scrutinised if they are later found to have significant association.

2.2.4. SNP reduction

In addition to the full genome-wide set of SNPs, two reduced sets of SNPs were created for use in relatedness estimation. One was the 'pruned' set of SNPs, similar to those created in Section 2.1.3; the other was a further reduced 'thinned' set of SNPs, which was intended for investigating the performance of one of the linear mixed-model GWAS software packages, FaST-LMM, which has been described as being most efficient when the number of SNPs used are less than the number of samples (Lippert *et al.*, 2011).

The common SNPs (minor allele frequency > 0.4) were pruned using PLINK command '`--indep 50 5 2`', so that within each overlapping window of 50 SNPs, recurring at every 5 SNP interval, the variance inflation factor never exceeded 2. This reduced the number of SNPs down to 50,129. The pruned set of SNPs were then 'thinned' down using MapThin (Howey and Cordell, 2011) to create a further subset of 1,900 SNPs—this number was chosen so that it was less than the total number of samples (1972) and would thus allow FaST-LMM to operate at maximum efficiency. The three sets of SNPs (full, pruned, and 'thinned'—i.e. pruned then thinned) were then used for calculation of kinship measures in GWAS analyses.

In the linkage analysis part of this project, the pruned SNP set was also used in our new method RIA to estimate of the expected ('prior') IBD sharing probabilities among the affected individuals, whilst the thinned SNP set was used in standard linkage analyses.

2.3. The Vesicoureteral Reflux Disease Data Set

The main difference between the vesicoureteral reflux disease (VUR) data set and the other two data sets above was that this was a collection of nuclear families rather than extended families. This makes it suitable for use in the early phase of linkage analysis method development.

2.3.1. The data set

The VUR data set was a combination of data from two projects: the whole genome studies of primary, nonsyndromic vesicoureteric reflux in the UK and Slovenia (Cordell *et al.*, 2010) and in Dublin, Ireland (Darlow *et al.*, 2014). Both collected DNA samples from affected siblings and their parents from families in which at least two siblings had been diagnosed and radiologically confirmed with primary, nonsyndromic vesicoureteric reflux (Cordell *et al.*, 2010; Darlow *et al.*, 2014).

The DNA samples from the UK and Slovenia study were genotyped using the 262,264 SNPs Affymetrix NspI array (Cordell *et al.*, 2010); for the Dublin study, the 834,482

SNPs Affymetrix Genome-Wide Human SNP Array 6.0 was used (Darlow *et al.*, 2014). The raw fluorescence data from the two studies underwent slightly different genotype calling and quality control procedures. Generally, these involved filtering of call rates and heterozygosity rates, and checking for violation of Hardy-Weinberg equilibrium, incompatibility between pedigree and genomically-estimated relatedness, outlying ethnicity and high Mendelian inheritance error rates (Cordell *et al.*, 2010; Darlow *et al.*, 2014). The two data sets have since been combined for an aggregate analysis, which resulted in a final data set of 2,343 individuals from 555 families (Table 2.1), with 119,548 high-quality SNPs in common between the two original data sets retained.

Sample set	Number of families	Affected individuals	Total number of individuals
UK	172	303	722
Slovenia	148	353	614
Dublin	235	500	1,007
Total	*555*	*1,156*	*2,343*

Table 2.1 Number of samples from each subset in the VUR data set.

2.3.2. Ethics statement

The UK and Slovenia study was approved by the UK Research Ethics Committees and the Slovenian National Ethics Committee (Cordell *et al.*, 2010). The Dublin study was approved by the ethics committees of two hospitals in Dublin (Our Lady's Children's Hospital Crumlin and the National Children's Hospital, Tallaght) where the samples were collected (Darlow *et al.*, 2014). Both studies confirmed that informed consent had been obtained prior to sample collection (Cordell *et al.*, 2010; Darlow *et al.*, 2014).

2.3.3. SNP reduction

Similar to the VL data set, two reduced sets of SNPs were created from this data set.

The 'pruned' SNP set was created from common SNPs (minor allele frequency > 0.4) using PLINK command '`--indep 50 5 2`', reducing the number of SNPs to 13,258. This set of SNPs was used for the calculation of the expected ('global' or 'prior') IBD sharing probabilities among individuals for use in RIA and the kinship matrix for use in FaST-LMM.

The 'thinned' SNP set was created by my supervisor in a similar manner to the other thinned SNP set, that is, it was a thinned down version of an already pruned set of SNPs, again using the program MapThin. There were 6,586 independent SNPs in this set, which was used for standard linkage analysis in Merlin (Abecasis *et al.*, 2002).

2.4. Phenotype Simulations within the VL data set for the purpose of GWAS

To study the performance of the GWAS programs in identifying true association signals under different conditions, several sets of phenotypes were simulated. These include: cross-sectional qualitative (binary) traits, cross-sectional quantitative traits and longitudinal (repeated-measurement) quantitative traits. The traits were generated for the 1,972 individuals from the VL data set who had genotype information, except for the longitudinal traits which were generated from 498 individuals further drawn from these 1,972 individuals using stratified sampling method (see 2.4.3 below). The parameters in each simulation were manually adjusted until clear (but not exceedingly strong) association signals could be seen in the LMM software being used for initial evaluation of the traits, while still maintaining approximately the same numbers of affected and unaffected individuals to the original data set. The software used for this initial evaluation was usually FaST-LMM for its shortest computational time under optimally parallelised conditions as will be described in section 5.5. Detail of each simulation is as follows:

2.4.1. Cross-sectional qualitative traits

These are single measurements of qualitative, binary traits, reflecting, for example, the disease or non-disease status of an individual. Two different traits—one corresponding to a 'strong' genetic effect (sim-D1), and the other to a 'weak' genetic effect (sim-D2)—were generated from two similar models, each governed by two SNPs: rs9271252 and rs233722, located on chromosomes 6 and 12 respectively. These two SNPs were selected from the signal regions previously identified in GWAS studies: rs9271252 from this VL data set (Fakiola *et al.*, 2013), and rs233722 from a GWAS study of Tetralogy of Fallot in the Europeans (Cordell *et al.*, 2013).

Furthermore, 22 weaker 'polygenic' effects were also modelled based on the genotype of the 100th genotyped SNP on each autosome. The selection of these 22 polygenic SNPs was mostly arbitrary; the only requirement was that they were sufficiently distant from the two main effect SNPs and from one another, so that they were not linked to any other SNP within the model. Each SNP in the model contributed multiplicatively to the probability of developing disease ('penetrance'), according to the mathematical model:

$$\text{Penetrance} = \alpha \prod_{j=1}^{24} \beta_j^{x_j}$$

where j represented each causal SNP, with $j = 1$ corresponding to rs9271252, $j = 2$ corresponding to rs233722 and $j = 3, \ldots 24$ corresponding to the 100th ('polygenic') SNP from chromosomes 1 to 22, respectively; x_j was a variable coded (0, 1, 2) according to the number of copies of risk allele presented at the causal SNP j (for convenience as

well as for biological plausibility, the risk alleles were assumed to be the minor alleles); α was the baseline penetrance and β_j was the multiplicative effect for each copy of risk allele at SNP j. Penetrances exceeding one were set to one.

For the 'strong' scenario, α was set to 0.017, β_1 to 2, β_2 to 1.8 and β_j $(j = 3, \dots 24)$ to 1.1.

For the 'weak' scenario, α was set to 0.022, β_1 to 1.6 and β_2 to 1.55, while β_j $(j = 3, \dots 24)$ remained at 1.1.

Individuals were then assigned 'disease' or 'non-disease' phenotype through binomial sampling with success probabilities (i.e. the 'disease' probabilities) set to their calculated penetrances.

2.4.2. Cross-sectional quantitative traits

The cross-sectional (single measurement) quantitative traits (sim-Q) were based on a similar set of SNPs to section 2.4.1 above. Again, rs9271252 and rs233722 were chosen as the two strong effect SNPs, with additional polygenic effects from the remaining 22 SNPs from each chromosome. The traits were generated as a linear combination of the effect from each SNP with a normally distributed error component:

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \sum_{j=3}^{24} \beta_j x_{ij} + \varepsilon_i$$

where x_{ij} was a genotype variable for person i at SNP j, coded as described in section 2.4.1 above, α represents the baseline trait and was set to 100, β_j was the additive effect due to SNP j, with β_1 set to 3, β_2 to 2 and the remaining polygenic effect $\beta_j (j = 3, \dots 24)$ set to 1, ε_i was a randomly generated variable following a normal distribution with mean 0 and standard deviation 5. These resulted in a heritability of 0.34. (These values were chosen so that the simulated traits resembled adult blood pressure, although this is not particularly required in this study.)

2.4.3. Longitudinal quantitative traits

To make the analyses feasible while still maintaining the overall degree of relatedness, a longitudinal data set was constructed based on a smaller subset of individuals drawn using stratified sampling from the 1,972 genotyped VL individuals used in the above phenotype simulations. From each extended family, a number of individuals were randomly chosen approximately proportional to the family size. This process yielded a data set of 498 individuals whose phenotypes were then generated 20 times to create the longitudinal phenotypes. In addition, the genotype data for these individuals were repeated 20 times to create a set of 9,960 'individual' genotypes as required by most

software. This, in effect, means that the software will treat the observations from each individual as having come from 20 monozygotic twins (or 'vigintuplets').

The generation of longitudinal phenotype did not incorporate systematic change over time: the traits were assumed to be randomly distributed over each individual's mean, and the models used for their generation were quite similar to the model used in the cross-sectional quantitative traits, with the addition of another error term δ_i to account for individuals' non-genetic variation.

Two longitudinal quantitative traits were generated (the reason for this will become apparent in section 5.4). The first (sim-L20) used a model with similar set of SNPs to the cross-sectional model:

$$y_{i_k} = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \sum_{j=3}^{24} \beta_j x_{ij} + \delta_i + \varepsilon_{i_k}$$

where $k = 1, \dots 20$ denotes each measurement in individual i, the baseline trait α remained 100, β_1 was set to 5, β_2 to 4 and $\beta_j (j = 3, \dots 24)$ set to 1.5, δ_i was a random variable following a normal distribution with mean 0 and standard deviation 4, generated once for each individual. The residual error term ε_{i_k} was a randomly generated variable following a normal distribution with mean 0 and standard deviation 2.

The second trait (sim-P20) was purely polygenic and was contributed to equally by 402 small effect SNPs $j = 3, \dots 404$ (400 randomly chosen, and the two former strong effects, rs9271252 and rs233722, which no longer had strong effects in this simulation). In other words, using the above model, β_1 and β_2 did not exist, and the term $\sum_{j=3}^{24} \beta_j x_{ij}$ was replaced with $\sum_{j=3}^{404} \beta_j x_{ij}$ where all polygenic effects $\beta_j (j = 3, \dots 404)$ were set to 0.75. The background risk α was set to 20, δ_i followed a normal distribution with mean 0 and standard deviation 16 and ε_{i_k} followed a normal distribution with mean 0 and standard deviation 1.

2.4.4. Replication of Simulated Phenotypes

For power, type I error and concordance analysis, 1,000 replicates of each of the cross-sectional phenotypes were generated. Technically, for each trait, 1,972,000 phenotypic values were generated from a single random number stream before being split into 1,000 replicates with 1,972 individuals each. Since the random number seeds similar to the original simulations were used for the replication, the first replicate was always exactly the same as the original.

2.5. Phenotype Simulations within the VL data set for Linkage Analysis

To study the performance of the proposed linkage analysis method in comparison with other linkage and/or association software in identifying true linkage signals, two sets of phenotypes were simulated for the 1,960 individuals from the VL data set who had genotype information and in whose families there were at least two genotyped individuals. The phenotypes for the remaining 1,410 non-genotyped individuals from these families were set to missing for use in the conventional linkage analysis programs.

2.5.1. SNP-based qualitative trait simulation

This was a simple simulation based on the association between one SNP, in this case rs9271252 on chromosome 6, and a binary phenotype. This association would also give rise to a linkage signal, although this could potentially be weak as similar alleles in a single SNP are not necessarily IBD. It was used here as a relatively quick initial step to evaluate the proposed method before a more complicated (and time-consuming) simulation was implemented.

Initially, the simulated 'strong' qualitative trait from section 2.4.1 was used. However, this did not give a satisfactory result when used in linkage analysis, especially for linkage methods in which only the affected individuals are informative such as those using affected relative pairs, due to the low number of affected individuals.

In attempt to improve the power for the linkage methods, a multiplicative model analogous to the cross-sectional qualitative GWAS simulation (2.4.1 above), but with only one strong effect SNP was then tried, namely:

$$\text{Penetrance} = \alpha\beta^{x}$$

where x represented the number of disease alleles presented in rs9271252, and β its multiplicative effect. The disease allele was again set to be the minor allele which, rather than being merely a matter of convenience (and to a certain extent, biological plausibility) as in the previous simulations, was rather a requirement here in order to lessen the problem with disease alleles being IBS but not IBD, which could attenuate the linkage signal.

However, this model did not perform well either. Further analysis showed that this was because of the conflicting requirements on the value of β, which can never be fully satisfied. For the linkage methods to achieve adequate power from this data set, a substantial number of individuals carrying disease allele(s) are required to be affected while few, if any, of the individuals not carrying disease allele should be so. This implies a low α and a relatively high β. As the disease allele was set to be the minor allele, very few individuals in the data set would carry two disease alleles, and most individuals

who carried the disease allele would carry just one. To achieve adequate power, it was therefore necessary that the penetrance in the heterozygous individuals, $\alpha\beta$, was relatively high (in the region of 0.7 to 0.8). On the other hand, the penetrance in the homozygous disease individual, $\alpha\beta^2$, should preferably be slightly lower than 1 to create a realistic complex disease trait. These two requirements are contradictory in presence of a low α. (To satisfy both conditions, β needs to be < 1.43 while $\alpha > 0.49$, which makes the disease far too common in normal population and thus also results in loss of power.)

To satisfy all the requirements above, a slightly different model was used. This was based on a logistic relationship, expressed in exponential form:

$$\frac{\text{Penetrance}}{1 - \text{Penetrance}} = \alpha\beta^x$$

where α then became the background odds, and was set to 0.02; β was the marginal odds contributed by each disease allele, set to 117. This resulted in the following genotype-specific penetrance:

Number of alleles (x)	Model odds	Calculated odds	Penetrance
0	α	0.02	0.020
1	$\alpha\beta$	2.34	0.701
2	$\alpha\beta^2$	273.78	0.996

Table 2.2 Genotype-specific penetrance for qualitative trait simulation.

After the penetrance was calculated, the phenotypes were then obtained through binomial sampling, similar to the GWAS quantitative traits simulation (section 2.4.1).

2.5.2. Haplotype-based qualitative trait simulation

To simulate the situation where there is a strong linkage signal, but weak or no association signal, a haplotype-based simulation was used.

Firstly, the haplotypes within a 10 cM range on chromosome 6 (from 47 cM to 57 cM) were estimated using the command '`--rsq 0.1 --cfreq`' in Merlin (Abecasis *et al.*, 2002). This clustered SNPs within that range such that the pairwise correlation between SNPs in each cluster was above 0.1, and then estimated the haplotype frequencies of these SNP clusters. A specific cluster containing the SNP rs9271252 was chosen to be the cluster carrying the true causal SNP. The estimated haplotype frequencies in this cluster are as shown in Table 2.3:

Haplotype ID	Frequency	Haplotype
1	0.0007	GAGAAACGGGCAC**d**AAACAA
2	0.2700	GAGAAACGGGCAC**d**AAACAA
3	0.0745	GAGAAACGGACAC**d**AAACAA
4	0.0028	GAGAAACGGACAC**d**AAACAA
5	0.0090	GAGAAAAGGGAGA**d**AAGCAA
6	0.0007	GAGAAAAGGGAGA**d**AGGAGG
7	0.1414	GAGAAGCGGGAGA**d**AGGAGA
8	0.0291	GAGAAGCGGACAC**d**AAACAA
9	0.0193	GAGAAGCAGACAC**d**AAACAA
10	0.1906	GAGAAGCAAGAGA**d**AAGCAA
11	0.0007	GAGAAGCAAGAGA**d**AAGCAG
12	0.0007	GAGAAGCAAGAGA**d**AGGCAA
13	0.0009	GAGAAGCAAAAGA**d**AAGCAA
14	0.0602	GAGACAAGGGAGA**d**AGGAGA
15	0.0007	GAAGAACGGACAC**d**AAACAA
16	0.0007	CAGAAGCAAGAGA**d**CAGCAA
17	0.0007	CGGAAACGGGCAC**d**CAACAA
18	0.0007	CGGAAAAGGGAGA**d**AAGAGG
19	0.0771	CGGAAAAGGGAGA**d**AGGAGG
20	0.0097	CGAAAACGGACAC**d**AAACAA
21	0.0185	CGAAAACAAGAGA**d**AAGCAA
22	0.0808	CGAGAACGGACAC**d**AAACAA
23	0.0007	CGAGAAAGGAAGC**d**AGGAGG

Table 2.3 Estimated haplotypes containing rs9271252 and their frequencies. SNPs in this cluster were: rs9271252, rs9271255, rs9271256, rs9271366, rs34846487, rs9271522, rs17533090, rs3129763, rs3135003, rs9271640, rs9271842, rs9271858, rs9271891, rs9272070, rs35242582, rs9272130, rs17211510, rs28407322 and rs28693734. The bold letter 'd' marks the location where the disease locus was simulated.

Some of the rare haplotypes (frequency = 0.0007) from this haplotype cluster table were assigned to be the disease haplotype (depending on the simulation set). This was done through the modification of the haplotype cluster table by creating a dummy SNP, positioned at 32,599,771 BP (52.39 cM) on chromosome 6, between the SNPs rs9272070 and rs35242582. The allele on this SNP was set to A if it was on the disease haplotype, and to C if it was not. The genotype in this 10cM region was then replaced with the genotype simulated using Merlin's '`--simulate`' and '`--cluster`' command, which performed gene dropping simulation within the families using the haplotypes and haplotype frequencies from the new haplotype cluster table.

To ensure that the resulting genotypes can be used for linkage analyses, an iterative procedure was used to ensure that there were at least two individuals who carry at least one disease allele in each family, as the probability of having at least two affected

individuals within that family would be very low otherwise. For each family in the data set, the genotypes at the disease locus of each individual were assessed after the Merlin gene dropping simulation was completed; if there were less than two individuals carrying at least one disease allele, the replicate would be discarded and the simulation repeated with a different random number seed until this requirement was satisfied, in which case the resulting genotypes would be incorporated into the final genotype data set for that simulation, which would then be passed on to the phenotype simulation phase.

The phenotype simulation also ensured that there were at least two affected individuals in each family. So in a family with only two genotyped individuals, who would also carry at least one disease allele because of the way the genotype simulation was done, both would need to be assigned affected status. For families with more than two genotyped individuals, exactly the same model as in the SNP-based simulation (2.5.1 above) was used. This was done iteratively until there were at least two affected individuals in each family, in a similar manner to the genotype simulation.

The combined effects of this genotype and phenotype simulation is conceptually quite similar to the real-life process of recruiting appropriate families for linkage analysis from the population, although the family structure in this case came from a more restricted set of samples.

2.6. Statistical Methods/Software

2.6.1. Methods for association analysis

There are two groups of methods used in the association analyses performed in this thesis: linear mixed-model (LMM) methods and 'alternative' (i.e. non-LMM) methods which were designed to handle family data. Table 2.4 below summarises the association analysis methods used—some of these analyses, mainly the alternative methods, were performed by my supervisor (HJC). The methods will be described in detail in Chapter 4.

Package/method and version	Approach	Kinship estimation method	Reference(s)
EMMAX emmax-intel-binary-20120210.tar.gz	LMM (approximate)	Estimated internally using user-supplied set of SNPs, or set to theoretical/estimated values calculated externally	(Kang *et al.*, 2010)
FaST-LMM v2.04	LMM (approximate or exact)	Estimated internally using user-supplied set of SNPs, using SNPs selected through FaST-LMM-Select procedure, or set to theoretical/estimated values calculated externally	(Lippert *et al.*, 2011; Listgarten *et al.*, 2012; Lippert *et al.*, 2013)
GEMMA v0.91	LMM (exact)	Estimated internally using user-supplied set of SNPs, or set to theoretical/estimated values calculated externally	(Zhou and Stephens, 2012)
GenABEL v1.7-6 (FASTA)	LMM (approximate)	Estimated internally using user-supplied set of SNPs, or set to theoretical/estimated values calculated externally	(Aulchenko *et al.*, 2007b; Chen and Abecasis, 2007)
GenABEL v1.7-6 (GRAMMAR-Gamma)	LMM (approximate)	Estimated internally using user-supplied set of SNPs, or set to theoretical/estimated values calculated externally	(Aulchenko *et al.*, 2007b; Svishcheva *et al.*, 2012)
GTAM* (implemented in MASTOR v0.3)	LMM (approximate)	Calculated externally (assumed to reflect 'known' (theoretical) pedigree relationship)	(Abney *et al.*, 2002)
Mendel* v13.2	LMM (approximate or exact)	Estimated internally using theoretical pedigree relationships, or estimated using all SNPs, either within estimated pedigree clusters or fully estimated	(K. Lange *et al.*, 2013)
MMM v1.01	LMM (approximate or exact)	Estimated internally using user-supplied set of SNPs, or set to theoretical/estimated values calculated externally	(Pirinen *et al.*, 2013)
FBAT* v2.0.4	Transmission of alleles within pedigrees	Method by definition uses 'known' (theoretical) pedigree relationships	(Laird *et al.*, 2000; Horvath *et al.*, 2001)

Package/method and version	Approach	Kinship estimation method	Reference(s)
MASTOR* v0.3	Retrospective quantitative trait version of MQLS	Calculated externally (assumed to reflect 'known' (theoretical) pedigree relationship)	(Jakobsdottir and McPeek, 2013)
MQLS* v1.5	Adjusted version of retrospective case/control test	Calculated externally (assumed to reflect 'known' (theoretical) pedigree relationship)	(Thornton and McPeek, 2007)
ROADTRIPS* v1.2 (RM test)	Adjusted version of retrospective case/control test	Calculated externally (assumed to reflect 'known' (theoretical) pedigree relationship). Further correction based on genome-wide set of SNPs applied internally	(Thornton and McPeek, 2010)

Table 2.4 Summary of methods/software packages used in the chapters on association analysis.

* indicates that the analysis was performed by my supervisor (HJC).

2.6.2. Methods for linkage analysis

Three linkage analysis methods were used in this thesis as summarised in Table 2.5 below. Again, a detailed description of these methods will be provided in the relevant chapter (Chapter 6).

Package/method and version	Approach	Expected IBD sharing estimation method	Reference(s)
Merlin v1.1.2 (option `--exp`)	Kong and Cox multipoint exponential likelihood model	Calculated internally using 'known' (theoretical) pedigree relationship	(Abecasis *et al.*, 2002; Abecasis and Wigginton, 2005)
MORGAN v3.2 (lm_ibdtests program, using the S_{pairs} statistics under normality assumption)	MCMC estimation of S_{pairs} statistics	Calculated internally using 'known' (theoretical) pedigree relationship	(Basu *et al.*, 2008)
RIA (using PLINK v1.07 with `--Z-genome` option or KING v1.4 with `--homo` option for IBD estimation, and modified version of onelocarp for MLS calculation)	MLS-like statistics	Estimated externally	(Cordell *et al.*, 2000; Purcell *et al.*, 2007; Manichaikul *et al.*, 2010) RIA itself has not yet been published.

Table 2.5 Summary of methods/software packages used in the chapters on linkage analysis.

2.7. Computing Facilities

The data manipulation and computation was done on either stand-alone linux servers or one of the high-performance computing (HPC) clusters.

The stand-alone linux servers each consists of between 8-12 2.59 GHz CPU cores and has about 32-64 GB of memory. These were used mainly for data manipulation and some simpler calculations. The more complex calculations were done on one of the two HPC clusters.

Most of the GWAS analyses and all formal runtime measurements were done on the older HPC cluster, which consists of 20 worknodes, each has 8 2.67 GHz CPU cores. Sixteen of the worknodes were older, each with 47 GB of memory; the remaining four were the newer 'high-memory' worknodes, each of which has 95 GB of memory.

Most of the linkage analyses were done on the newer HPC cluster, which consists of 20 worknodes, each has 20 2.8 GHz CPU cores. Two of these are 'high-memory' and have 504 GB of memory each, while the remaining 18 have 126 GB of memory each.

2.8. Measurement of Computational Time

Formal computational time was measured for certain analyses (cross-sectional GWAS and linkage analyses). These were done by requesting an exclusive execution of a dedicated timing script on a whole worknode of the older HPC cluster to prevent interference from other tasks. Under the timing condition, tasks were not parallelised unless they were natively multi-threaded, in which case they would be allowed to run using the maximum available cores (i.e. 8). Run time measurements made by my supervisor (alternative GWAS methods) also used a similar method.

Approximate run times are sometimes given. These are based on normal running conditions without exclusive use of the worknode (unless so required due to the program's resource demands), and with parallelisation as appropriate. If parallelisation was used, the total run time would be calculated from the sum of the (possibly approximate) run time of each task.

Chapter 3. Analysis of GAW18 Data

In the last few years, a bewildering number of different methods/software packages implementing linear mixed model approaches to account for population structure and relatedness among samples in genome-wide association studies has been proposed, but no detailed comparison between them has previously been made. Indeed, when a new method/package is developed, it is often quite unclear whether or how it differs substantially from the methods/software implementations that are already available. This and the next two chapters will attempt to address this question by exploring the performance of various implementations of such methods in familial and/or longitudinal data sets.

The analysis of GAW18 data in this chapter was done at about the same time as the early analyses of the VL data (Chapter 4). It therefore benefited from some of the findings from the early analyses of the VL data (e.g. the optimal set of SNPs to use for relatedness calculation). On the other hand, because of the smaller size and the innately longitudinal nature of the GAW18 data set, the longitudinal analysis in this chapter also functioned as a pilot for the more advanced analyses of the VL data set involving longitudinal data.

The GAW18 data set and quality control process has been described in detail in Chapter 2 (Section 2.1). This chapter will comment on the statistical methods used and the results. Although the results presented here are slightly different from those in the published article describing this part of the thesis (Eu-ahsunthornwattana *et al.*, 2014a) due to an initial data processing error that was later corrected, the main conclusions regarding the performance of each LMM method remain the same.

3.1. Statistical Methods

A two-step procedure was used to adjust for the effect of the covariates and for familial and intra-individual correlations. For each of the two sets of GAW18 phenotypes used in this project (the real phenotypes and the first replicate of the simulated phenotypes), linear regression of systolic blood pressure (sBP) and diastolic blood pressure (dBP) at each time point on age, medication and smoking status was conducted, except for the real dBP—which seemed to have a nonlinear relationship with age, as could be physiologically expected—for which a quadratic regression including age and age squared as predictors was used. The phenotype data from all individuals were used for these regressions regardless of their genotyping status. Residuals from these regressions in subjects who also have genotype data were then used in the next step.

The second step was genome-wide association analysis. To account for the longitudinal nature of the data, two approaches were used in this step.

The first approach (which will be referred to as *'longitudinal'*) was to model the residual from each individual observation without regard to its true longitudinal nature in the genome-wide association analysis, treating the multiple observations from the same individual as if they came from separate individuals. In this approach, genomic data was used to adjust for familial as well as intra-individual correlation through the use of an estimated kinship matrix, effectively treating the multiple observations from the same individual as having been collected from identical twins (or triplets or quadruplets).

The other approach (*'mean'*) was to calculate the mean of the residuals for each subject and then use each individual as a single observation. The genomically estimated kinship matrix in this case adjusts for familial relatedness only.

This analysis itself was performed using a variety of linear mixed model approaches. The approaches vary with respect to precise details of the calculation of kinship or 'relatedness', and with respect to whether an exact method or a fast approximation is used. In each case, the 21,151 pruned SNPs (see Section 2.1.3) were used for the relatedness calculations. The pruned set of SNPs was chosen based on prior work in the VL data set, which showed little difference between results when using such a pruned set of SNPs for calculating relatedness compared to using the full set of SNPs (see Section 4.2.2 for more detail).

The methods considered were:

1. EMMAX (Kang *et al.*, 2010), which implements two methods for relatedness calculations: one based on IBS sharing, and one based on the 'Balding-Nichols' model (Balding and Nichols, 1995; Rakovski and Stram, 2009).
2. FaST-LMM (Listgarten *et al.*, 2012), which also implements two methods to adjust for relatedness: one using a standard covariance matrix, and one using the realised relationship matrix (RRM). The GWAS stage of FaST-LMM was conducted using the 'approximate' calculation ('`-simLearnType Once`', see Section 4.1.1 for further detail).
3. the polygenic/mmscore functions in GenABEL (Aulchenko *et al.*, 2007b), which implement the FASTA method (Chen and Abecasis, 2007).
4. the polygenic/grammar functions in GenABEL, which implement the GRAMMAR-Gamma approximation (Svishcheva *et al.*, 2012).
5. GEMMA (Zhou and Stephens, 2012), which uses an efficient exact method (see also Section 4.1).

Additionally, simple linear regression without any relatedness adjustment was also performed in FaST-LMM. All analyses were performed using both the longitudinal and the mean approaches.

For each analysis, genomic inflation factors (λ) were calculated as proposed by Devlin and Roeder (1999). Since this factor was originally based on χ^2 values, the equivalent 1 degree of freedom χ^2 values derived from the p-values were used for programs that gave only p-values (and not χ^2 values).

3.2. Results

All LMM methods performed reasonably well in both mean and longitudinal approaches (Figure 3.1), controlling the λ to 1.01-1.07 (mean) and 0.98-1.07 (longitudinal). These values were much less inflated compared with the λ values of 1.39-1.87 (mean) and 2.27-3.81 (longitudinal) seen in the unadjusted analyses (not shown in the plot). In general, the longitudinal analysis tends to be slightly more inflated compared with the mean analysis of the same phenotype using the same method, with the exception of EMMAX (IBS) in which this trend is reversed. However, this reversal seems to be due to the deflation in the longitudinal analyses using EMMAX (IBS) rather than the inflation in the mean analyses, as the λ values in the latter are quite comparable to those in the other methods (particularly EMMAX (BN)).

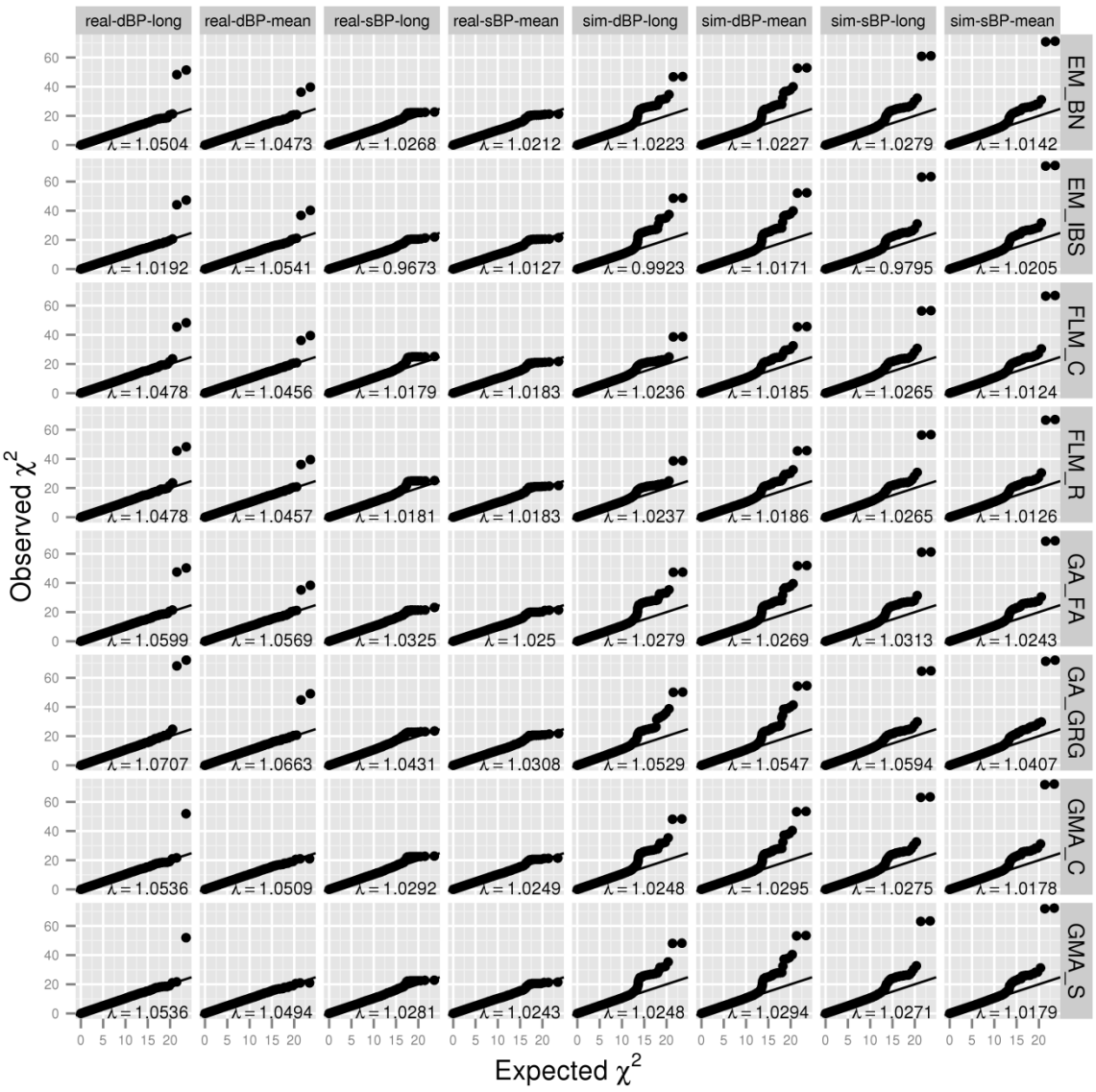

Figure 3.1 Q-Q plots of χ^2 statistics and genomic inflation factors (λ) for different LMM methods. These were calculated for each phenotype (real diastolic blood pressure [dBP], real systolic blood pressure [sBP], simulated dBP and simulated sBP), using either longitudinal ('long') or average ('mean') residuals. EM_BN = EMMAX using Balding-Nichols matrix, EM_IBS = EMMAX using IBS matrix, FLM_C = FaST-LMM using standard covariance matrix, FLM_R = FaST-LMM using realised relationship matrix, GA_FA = GenABEL/FASTA, GA_GRG = GenABEL/GRAMMAR-Gamma, GMA_C = GEMMA using centralised covariance matrix, GMA_S = GEMMA using standardised covariance matrix. The black, straight line represents the identity line in each panel. The missing of one and two top SNP(s) in both GEMMA methods (in longitudinal and mean analysis of real dBP phenotype, respectively) was because the genotype missing rates for these SNPs (one from chromosome 5, the other from chromosome 13) reached GEMMA's default missingness threshold for exclusion from its analysis (5%).

Comparisons of individual $-\log_{10}$ p-values (Figure 3.2) also showed highly concordant results among the methods, particularly between EMMAX (BN) and GEMMA, while the two GenABEL methods were also quite similar to these but not to the same degree. In general, the analyses using the mean values were more concordant than those using

longitudinal values, and variants of the same methods tended to give more concordant results.

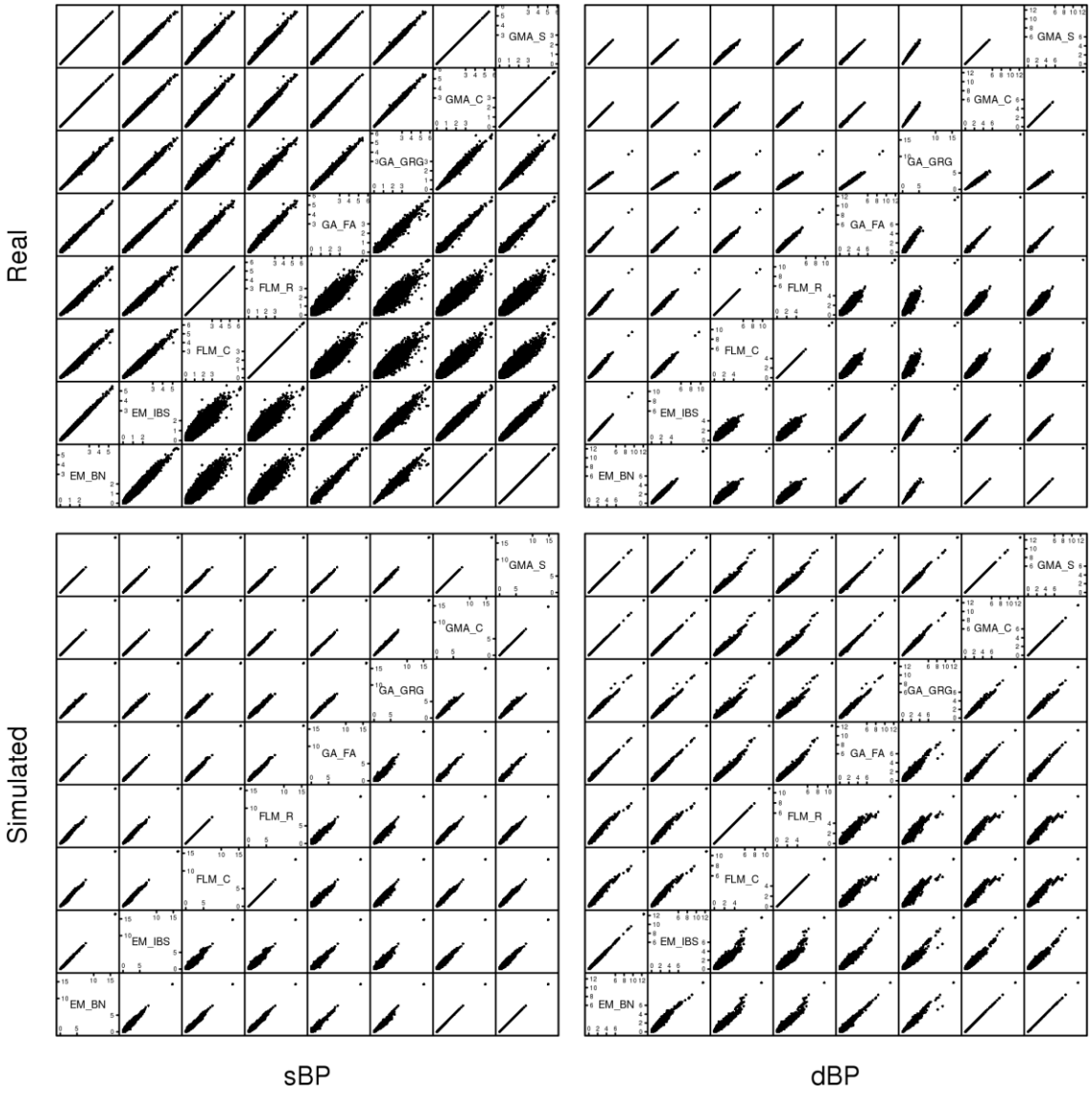

Figure 3.2 Comparison of –log$_{10}$ p-values calculated using different methods, based on mean (upper triangles) or longitudinal values (lower triangles). EM_BN = EMMAX using Balding-Nichols matrix, EM_IBS = EMMAX using IBS matrix, FLM_C = FaST-LMM using standard covariance matrix, FLM_R = FaST-LMM using realised relationship matrix, GA_FA = GenABEL/FASTA, GA_GRG = GenABEL/GRAMMAR-Gamma, GMA_C = GEMMA using centralised covariance matrix, GMA_S = GEMMA using standardised covariance matrix.

The Manhattan plots from all methods were quite similar for each phenotype, although the longitudinal data tended to show stronger signals (a selection of these plots is shown in Figure 3.3). All methods detected a clear, strong signal at the SNP rs11711953 in the *MAP4* gene in chromosome 3 in the analyses of both simulated phenotypes. This

was indeed the SNP used to simulate the strongest effect in both phenotypes in this data set.

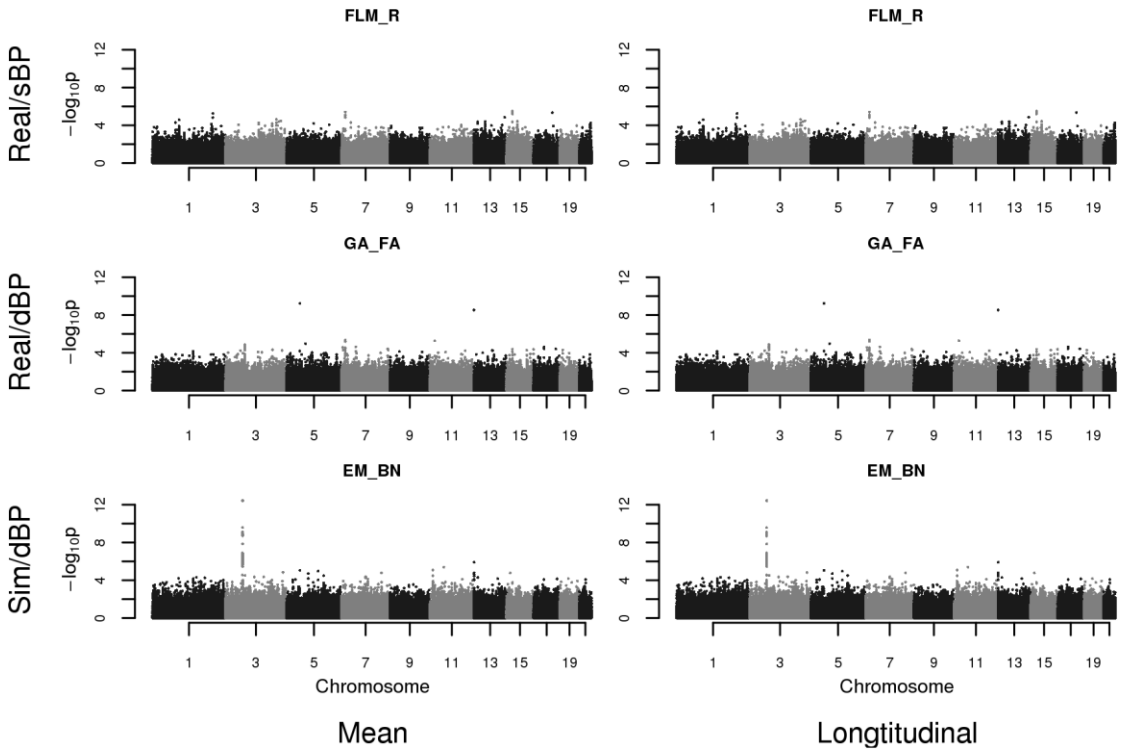

Figure 3.3 A selection of Manhattan plots showing p-values calculated using various methods. EM_BN = EMMAX using Balding-Nichols matrix, FLM_R = FaST-LMM using realised relationship matrix, GA_FA = GenABEL/FASTA, GA_GRG = GenABEL/GRAMMAR-Gamma, GMA_C = GEMMA using centralised covariance matrix, GMA_S = GEMMA using standardised covariance matrix.

Although the results from all packages considered here were similar, and all packages completed the analysis in reasonable time (less than one day) on our system, the differences in speed were substantial. Precise timings will depend on the computer resources and architecture available, but as a rule of thumb, FaST-LMM and GRAMMAR-Gamma were found to be the fastest (taking just a few hours), followed by EMMAX and GEMMA which took around 12-16 hours and GenABEL/FASTA which took around 18-20 hours (see also Section 5.5 for more formal comparison using the VL data set, as well as discussion about the various factors affecting speed).

3.3. Discussion

It is well known that population substructure and relatedness will cause an inflated distribution of genome-wide association test statistics ($\lambda > 1.00$) if not appropriately modelled (Yu *et al.*, 2006). All methods performed well in this regard, being able to control the genomic inflation to an acceptable level under most circumstances.

The higher inflation in longitudinal analyses, even when adjusting for relatedness, could be expected from the fact that additional (non-genetic) within-subject correlation was not allowed for in these analyses. This was because all methods considered attempted to fit a mixed-model with only one individual-specific source of variance, that is, the genetically-determined random effect component. This effectively disregards the fact that different observations from the same individual also share the same individual-specific environmental contribution. Although some of the variance from the individual's environmental component is absorbed in to the genetic component, there would still be some correlation left in the residuals, therefore some degree of inflation can be expected, albeit much lower than the unadjusted analysis in which neither the individuals' genetic relatedness nor the environmental contribution was accounted for. In fact, one may argue that GRAMMAR-Gamma may actually have shown the 'most correct' statistical behaviour (although this may not necessarily be desirable), in that it resulted in the highest inflation (note, however, that the mean analyses in GRAMMAR-Gamma were also quite inflated compared with other methods). Interestingly, EMMAX using the IBS matrix seemed to have the opposite behaviour—that is, *deflation* rather than *inflation* was observed in all longitudinal analyses using this method, resulting in consistently lower genomic inflation compared to the mean analyses. The reason for this is not currently known (but see also the results and discussion in Chapter 5).

That no clearly significant SNP was found in any analysis of the real phenotypes was not surprising, given the relatively small size of the GAW18 data set which would be under-powered for detecting, at genome-wide levels of significance, anything other than strong genetic effects. When the effect was strong enough, as was the case in the simulated phenotypes, all methods were equally successful in identifying the true signal. The high concordance in significance levels at any given SNP achieved by the different software packages (Figure 3.2) indicates that no package is substantially more powerful than another, as expected from the fact that all packages implement slightly different versions of essentially the same statistical model. Nevertheless, the differences in how the methods implement the model would explain the observed increase in discrepancies of longitudinal analysis results compared to mean analyses, as specific implementation could affect how the environmental variance is absorbed into the genetic component and residuals. A more detailed exploration of these differences would be an interesting topic for further investigation, although it is beyond the scope of this thesis.

Since all methods performed well and results were similar, particularly at the most significant SNPs, it makes little difference to the results—at least for non-longitudinal

traits—which method/software package is used. The user can make the choice of package on the basis of personal taste, speed or computational convenience.

Chapter 4. Application of Genomic IBD Estimates to Account for Relatedness in Genome-Wide Association Analyses of the Brazilian Visceral Leishmaniasis Data

Continuing the theme of comparison of LMM GWAS methods introduced in Chapter 3, various specific issues pertinent to LMM GWAS analysis will be explored in this chapter using the real phenotypes from the Brazilian visceral leishmaniasis (VL) data set. These include the differences in the IBD estimates and the resulting test statistics when different methods or SNP sets are used for IBD estimation, the effect of using externally estimated (and not necessarily correct) IBD probabilities in LMM programs and the performances of various LMM and alternative methods in analysing the real phenotype data.

These analysis methods utilise the IBD estimates as summarised into a single 'kinship measure' for each pair of individuals—which, depending on the method, could be either the kinship coefficient or the proportion of alleles shared (which is equivalent to the coefficient of relationship and is twice the kinship coefficient)—to model the relatedness between individuals. The discussion of IBD in this chapter will therefore be in terms of these 'kinship measures'.

As the VL data set and quality control process has been described in detail in Chapter 2 (Section 2.2), this chapter will comment only on the statistical methods used and the results.

4.1. Description of Software/Methods Being Compared

4.1.1. LMM-based methods

As previously mentioned, the LMM methods considered here attempt to fit the mixed effect model:

$$\boldsymbol{Y} = X\boldsymbol{\beta} + Q + \varepsilon$$

(see Section 1.2 for description of the variables). In theory, this can be done using a variety of generic LMM programs that were not specifically written for genome-wide data analysis, such as the nlme (Pinheiro *et al.*, 2013) or lme4 (Bates *et al.*, 2014) packages in R. In practice, however, several issues arise when these are used for genome-wide LMM analysis incorporating genetically estimated kinships.

Firstly, unlike linear models, LMMs do not have closed form solutions and therefore have to be solved numerically. This is computationally demanding, especially when

analysing a data set with a large number of individuals, as the required run time is a cubic function of the number of individuals (Kang *et al.*, 2010; Lippert *et al.*, 2011). Furthermore, this computationally expensive procedure needs to be repeated for each SNP under investigation, because full fitting of a mixed model requires complete re-estimation of the model parameters (Chen and Abecasis, 2007; Svishcheva *et al.*, 2012). In context of modern GWAS, where data on a very large number of SNPs need to be analysed, this could result in a prohibitively long run time, and is the main motivation for the development of the various LMM approximation/simplification methods described here.

Secondly, generic LMM programs may not allow the use of externally constructed variance-covariance matrices. Both nlme and lme4 internally construct their variance-covariance matrices following a set of pre-defined forms and do not provide a means to incorporate an externally constructed matrix. Interestingly, another 'generic' program, lmekin (from R package coxme (Therneau, 2012)), allows the use of an externally constructed, fully specified variance-covariance matrix; however, this was in fact because it was written primarily for genetic data analysis, although it is generic enough to be used in other situations as well.

Even for programs that permit the use of an external variance-covariance matrix, the externally constructed kinship matrices can still pose a problem. This is because standard LMM requires the variance-covariance matrix to be positive semidefinite, which may not necessarily be satisfied with standard genetic-based kinship estimation methods (Kang *et al.*, 2008; Astle and Balding, 2009). This tends to trigger a fatal error in most generic LMM programs, which is appropriate as the results in this case will be ill-defined.

Specialised programs for LMM GWAS analysis employ various techniques to circumvent these limitations. For example, most use (or permit) two stage approximation whereby the more time-consuming estimation of certain model parameters is done only once, before using them in subsequent simplified SNP-wise analyses. Some software packages (FaST-LMM (Lippert *et al.*, 2011), GEMMA (Zhou and Stephens, 2012) and MMM (Pirinen *et al.*, 2013)) also implement a speed up of the exact calculation through spectral decomposition. Furthermore, most programs considered here provide ways to make kinship estimation quicker (or even bypass it altogether such as in FaST-LMM), some of these also result in a kinship matrix that is always positive semidefinite (Kang *et al.*, 2008; Kang *et al.*, 2010). Programs that do not guarantee their estimated kinship matrices to be positive semidefinite seem to be implemented in such a way that this is handled without causing a fatal error. For

example, FaST-LMM sets any negative eigenvalue to zero, which is equivalent to forcing the kinship matrix to be positive semidefinite, thus eliminating the problem.

The description of each LMM software package considered in this thesis and its method(s) for kinship estimation is provided below:

GenABEL (FASTA)

The `mmscore` and `polygenic` functions of the GenABEL package (Aulchenko *et al.*, 2007b) together allow implementation of the **FA**mily based **S**core **T**est **A**pproximation (FASTA) method proposed by Chen and Abecasis (2007). Although the FASTA method is also implemented in the `--fast-Assoc` option of the MERLIN package (Abecasis *et al.*, 2002), MERLIN's kinship matrix is calculated internally on the basis of known (theoretical) kinships constructed from known pedigree relationships rather than allowing the pairwise kinship coefficients to be estimated using genome-wide SNP genotype data (Amin *et al.*, 2007) as is the case in GenABEL's kinship calculation. GenABEL was therefore the preferred software for the FASTA method in this thesis.

Rather than fitting the full linear mixed model $\boldsymbol{y} = X\beta + Q + \varepsilon$ and estimating β, σ_g^2 and σ_e^2 by maximum likelihood for each SNP across the genome, FASTA implements an 'approximate' two-stage approach. At the first stage a reduced model is fitted, where the regression coefficient β_1 (corresponding to the effect at the SNP currently under test) is assumed to equal 0, but all other covariates (if desired) are included. At the second stage, a score statistic for testing the null hypothesis that β_1 does indeed equal 0 is constructed as:

$$T_{\mathrm{FA}} = \frac{([\boldsymbol{x}_1 - E(\boldsymbol{x}_1)]^T \Omega^{-1} [\boldsymbol{y} - E(\boldsymbol{y})])^2}{[\boldsymbol{x}_1 - E(\boldsymbol{x}_1)]^T \Omega^{-1} [\boldsymbol{x}_1 - E(\boldsymbol{x}_1)]}$$

where $E(\boldsymbol{y})$ refers to an n-dimensional vector of fitted values of the response from the reduced model, $E(\boldsymbol{x}_1)$ refers to an n-dimensional vector of unconditional expectations of genotype scores at the test SNP (each element of which equals twice the allele frequency of the particular allele being counted, as it is the expected allele count in a pair of individuals under the assumption of Hardy-Weinberg equilibrium), and Ω refers to the estimated variance/covariance matrix $\Omega = 2\Phi\sigma_g^2 + \sigma_e^2 I$, with σ_g and σ_e taking their maximum likelihood estimates as calculated under the reduced model. The score statistic is calculated repeatedly using the appropriate n-dimensional vector $\boldsymbol{x}_1$ for each test SNP (typically between 500,000 and several million SNPs) across the genome, but the time-consuming maximum likelihood step for estimating σ_g^2, σ_e^2 and $(\beta_2, \ldots, \beta_j)$ need only be performed once, at the start.

GenABEL's `polygenic` function, which performs the first stage of FASTA analysis, can read in any user-specified kinship matrix as long as the matrix is conformed to its input format. In practice, the `ibs` function in GenABEL package can readily be used to calculate a kinship matrix based on average pairwise IBS for use with the `polygenic` function. This can be done with or without allele frequency weighting. The method used in this thesis, which is also the default method in GenABEL, is to estimate the pairwise IBS with allele frequency weighting:

$$f_{i,j} = \frac{1}{N}\sum_{k=1}^{N}\frac{(x_{i,k} - p_k)(x_{j,k} - p_k)}{p_k(1 - p_k)}$$

where $f_{i,j}$ is the average pairwise IBS (and therefore the estimated kinship coefficient) between individuals i and j; N is the number of SNPs used in the estimation; $x_{i,k}$ is the genotype of the i-th individual at the k-th SNP, coded as 0, 0.5 and 1; and p_k is the frequency of the allele being assessed. This is equivalent to excess allele-sharing estimator of kinship coefficient, which is more precise and is closer to true IBD sharing than the unweighted estimator (Astle and Balding, 2009).

Recently, Fabregat-Traver *et al.* (2014) proposed OmicABEL, an improvement to GenABEL which allows efficient LMM analysis of multiple phenotypes. However, the use of OmicABEL is beyond the scope of this thesis.

GenABEL (GRAMMAR-Gamma)

The grammar function of the GenABEL package (Aulchenko *et al.*, 2007b) implements the GRAMMAR-Gamma method proposed by Svishcheva *et al.* (2012), which can be considered as an extension of the original GRAMMAR method (Amin *et al.*, 2007; Aulchenko *et al.*, 2007a) to produce a test that is essentially a fast approximation to FASTA.

Similar to FASTA, the first step of GRAMMAR is to fit a reduced version of the full linear mixed model in which β_1 is set to 0. Phenotype residuals $\tilde{\boldsymbol{y}} = (\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_n)^T$ may be constructed as $\tilde{y}_i = y_i - E(y_i)$ where $E(y_i)$ refers to the fitted value of the response for individual i from the reduced model. These residuals are then used as the independent trait in a simple linear regression model:

$$\tilde{y}_i = \mu + \tilde{\beta}_1 x_{i1} + e_i$$

where the error term e_i is assumed to be independently normally distributed. Estimation of $\tilde{\beta}_1$ and testing of the null hypothesis that $\tilde{\beta}_1 = 0$ can be accomplished through maximum likelihood or least squares approaches. Alternatively, a rapid test of

$\tilde{\beta}_1 = 0$ can be achieved (Amin *et al.*, 2007; Svishcheva *et al.*, 2012) through construction of a score statistic:

$$T_{\mathrm{GR}} = \frac{n([\boldsymbol{x}_1 - E(\boldsymbol{x}_1)]^T[\tilde{\boldsymbol{y}}^*])^2}{[\boldsymbol{x}_1 - E(\boldsymbol{x}_1)]^T[\boldsymbol{x}_1 - E(\boldsymbol{x}_1)][\tilde{\boldsymbol{y}}^*]^T[\tilde{\boldsymbol{y}}^*]}$$

where $\tilde{\boldsymbol{y}}^* = (\tilde{y}_1^*, \tilde{y}_2^*, \ldots, \tilde{y}_n^*)$ are transformed version of the residuals $\tilde{\boldsymbol{y}}^* = \sigma_e^2 \Omega^{-1} \tilde{\boldsymbol{y}}$. Again, the time-consuming maximum likelihood step for estimating σ_g^2, σ_e^2 and $(\beta_2, \ldots, \beta_j)$ need only be performed once.

In the original GRAMMAR publication (Aulchenko *et al.*, 2007a), the assumption was that pedigree relationships between individuals would be known and so Φ would be constructed on the basis of theoretical kinship coefficients. Subsequently it was suggested that the use of estimated kinship coefficients (estimated on the basis of genome-wide SNP data) could perform as well or better (Amin *et al.*, 2007). Regardless of which kinship coefficients are used, GRAMMAR was found to be conservative and to result in biased regression coefficients representing the SNP effects of interest (Amin *et al.*, 2007). It was therefore suggested that the final χ^2 test statistics should be 're-inflated' by multiplying by an appropriate estimated correction factor (in a procedure analogous to the 'deflation' of χ^2 test statistics via genomic control (Devlin and Roeder, 1999)) to result in a final test statistic with the appropriate null distribution. This 'genomic control corrected' version of GRAMMAR was denoted GRAMMAR-GC (Amin *et al.*, 2007).

The GRAMMAR-Gamma method (Svishcheva *et al.*, 2012) improves on the original GRAMMAR so that it produces unbiased SNP effect estimates and test statistics that do not require any deflation. This is achieved through rewriting the FASTA score test statistic as:

$$T_{\mathrm{FA}} = \frac{([\boldsymbol{x}_1 - E(\boldsymbol{x}_1)]^T \Omega^{-1}[\boldsymbol{y} - E(\boldsymbol{y})])^2}{[\boldsymbol{x}_1 - E(\boldsymbol{x}_1)]^T[\boldsymbol{x}_1 - E(\boldsymbol{x}_1)]} \Big/ \frac{[\boldsymbol{x}_1 - E(\boldsymbol{x}_1)]^T \Omega^{-1}[\boldsymbol{x}_1 - E(\boldsymbol{x}_1)]}{[\boldsymbol{x}_1 - E(\boldsymbol{x}_1)]^T[\boldsymbol{x}_1 - E(\boldsymbol{x}_1)]}$$

The numerator then becomes a new statistic, which is similar to the GRAMMAR statistic:

$$T_{\mathrm{NEW}} = \frac{([\boldsymbol{x}_1 - E(\boldsymbol{x}_1)]^T \Omega^{-1}[\boldsymbol{y} - E(\boldsymbol{y})])^2}{[\boldsymbol{x}_1 - E(\boldsymbol{x}_1)]^T[\boldsymbol{x}_1 - E(\boldsymbol{x}_1)]}$$

This can be calculated from a standard linear regression analysis of $\Omega^{-1}[\boldsymbol{y} - E(\boldsymbol{y})]$ on $[\boldsymbol{x}_1 - E(\boldsymbol{x}_1)]$.

The denominator is effectively the ratio between the new score test and the FASTA score test, which, when averaged across all markers, becomes a constant known as the GRAMMAR-Gamma factor, γ, and can be simplified to:

$$\gamma = \frac{1}{n-1}\sum_{i,j=1}^{n} \omega_{ij}^{-1} r_{ij}$$

where i and j refer to a relative pair, ω^{-1} is an element of Ω^{-1} and r_{ij} refers to the genomic kinship between the pair. This needs to be calculated only once at the beginning, and is used to adjust subsequent T_{NEW} statistic for each marker to obtain the GRAMMAR-Gamma score statistic:

$$T_{\mathrm{GRG}} = \frac{T_{\mathrm{NEW}}}{\gamma}$$

which is approximately equivalent to the FASTA statistic T_{FA} (Svishcheva *et al.*, 2012).

Svishcheva *et al.* (2012) argue that their GRAMMAR-Gamma method has similar computational complexity to alternative methods such as FASTA, EMMAX and FaST-LMM at stage 1, while achieving computational savings over these methods at stage 2 (achieving a stage 2 computational complexity of $O(sn)$ where n is the sample size and s the number of SNPs to be tested).

Similar to GenABEL's FASTA implementation, GenABEL's GRAMMAR-Gamma implementation also requires the use of the `polygenic` function, and therefore shares the same kinship calculation step through the `ibs` function with GenABEL FASTA.

EMMAX

Kang *et al.* (2010) proposed a method that appears to be essentially equivalent to the FASTA method proposed by Chen and Abecasis (2007), except for the following caveats:

1. In the approach of Kang *et al.* (2010), there is no expectation that the individuals will be closely related. Indeed, the method is motivated as an alternative to principal component based approaches when adjusting for population substructure in genome-wide association studies of unrelated individuals. Thus, the kinship coefficients used to construct Φ are not based on any 'known' pedigree relationships but are estimated based on genome-wide SNP data (using either a simple estimated based on the proportion of alleles idenical-by-state (IBS) measure, or else an estimated that Kang *et al.* (2010) describe as a Balding-Nichols (BN) estimate, which, in practice, is equivalent to

FaST-LMM's covariance matrix (Rakovski and Stram, 2009)), resulting in a procedure essentially identical to that proposed by Amin *et al.* (2007).

2. In the approach of Kang *et al.* (2010), rather than applying the method solely to quantitative traits as had been done previously (Amin *et al.*, 2007; Aulchenko *et al.*, 2007a; Chen and Abecasis, 2007), the method is also proposed to apply to case/control data (with the response coded as 0 or 1, but analysed as if it were, in fact, a quantitative trait, i.e. assuming a normally distributed random environmental/error term ε). Kang *et al.* argue that this is computationally more convenient than the usual way to analyse binary response data by fitting a generalised linear mixed model with a logit or probit link function, and should not result in increased type 1 error for testing the null hypothesis.
3. Although not entirely clear from the description in Kang *et al.* (2010), it appears that, at the second stage, in contrast to Chen and Abecasis (2007), any covariates other than the SNP currently under test are re-estimated i.e. the entire vector of fixed effect predictors $\beta = (\beta_1, \beta_2, \dots, \beta_j)$ is estimated, rather than fixing $(\beta_2, \dots, \beta_j)$ at their estimated values from the first stage.

The method of Kang et al. (2010) has been implemented in the software package EMMAX. As pointed out by Lippert *et al.* (2011), over and above the computational efficiency achieved by simply estimating parameters σ_g^2 and σ_e^2 only once, EMMAX, along with its predecessor EMMA (Kang *et al.*, 2008), achieves additional computational efficiency by reparameterising the likelihood in terms of a parameter $\delta = \sigma_e^2/\sigma_g^2$, which is estimated only once, and by making clever use of spectral decompositions. This results in a computational complexity of $O(n^3 + rn)$ at stage 1 (where r is the number of iterations i.e. the number of evaluations of the likelihood required) together with a computational complexity of $O(sn^2)$ at stage 2, resulting in a total computational complexity of $O(n^3 + sn^2 + rn)$.

A similar approach to EMMAX and FASTA was proposed by Z. Zhang *et al.* (2010) and implemented in a software package TASSEL. The main focus of the paper by Z. Zhang *et al.* (2010) was to describe a clustering algorithm that results in an approximation to the kinship matrix with lower effective dimensionality, which can be used in place of the full known or estimated kinship matrix. Similarly to EMMAX, in TASSEL the values of σ_g^2 and σ_e^2 (as well as a cluster membership variable C) are estimated under the null hypothesis that $\beta_1 = 0$ at stage 1 and are then held fixed while estimating $\beta = (\beta_1, \beta_2, \dots, \beta_j)$ at stage 2. The motivation for the clustering approximation is to reduce computation time. However, existing software packages (e.g. EMMAX and the `mmscore` and `polygenic` function in GenABEL) that address the problem without making such an approximation are not computationally prohibitively time consuming;

therefore, the practical advantage of this approximation is not clear. Given the extreme similarity between the methods implemented in EMMAX and TASSEL when no clustering is performed, comparison with TASSEL is not included in this thesis.

FaST-LMM

Lippert *et al.* (2011) developed FaST-LMM, a fast 'exact' LMM implementation which utilises factorisation and spectral decomposition (thus the 'Fa' and 'ST' in its name) to reduce the calculation complexity. In common with EMMAX, FaST-LMM reparameterises the likelihood in terms of a parameter $\delta = \sigma_e^2/\sigma_g^2$, which, due to the factorisation of the likelihood calculation, is the only parameter that need to be optimised. It requires only a single spectral decomposition at the first stage of the algorithm (rather than for each tested SNP as in EMMA, and without the need to assume that the variance parameters are constant as in EMMAX), resulting in a total time complexity of $O(n^3 + sn^2 + rsn)$. This exact method is the default in the current versions of FaST-LMM (from at least version 2.04).

FaST-LMM also provides an 'approximate' method through the `-simLearnType Once` option, in which δ is fixed to its value from fitting a null model containing no fixed SNP effects, as is done in EMMAX, TASSEL and FASTA, which further reduces the complexity to $O(n^3 + sn^2 + rn)$. This used to be the default method in the earlier versions of FaST-LMM.

FaST-LMM can base its calculation of maximum likelihood (ML) or restricted maximum likelihood (REML). The default option used to be the former (ML) in earlier versions, but has since been replaced with REML. After some experimentation, the ML option seemed to be more reliable than REML in the presence of strong genetic effects. All results presented in this thesis are therefore based on ML estimation.

Two types of kinship estimation are implemented in FaST-LMM: realised relationship matrix (RRM) (Goddard *et al.*, 2009; Hayes *et al.*, 2009) and EIGENSTRAT 'covariance' matrix (Price *et al.*, 2006). The difference between these is that the latter uses the mean-centred and standardised genotype data for calculation, and should be quite similar to GenABEL's weighted IBS calculation.

An interesting feature of these kinship matrices is that they were chosen because they are constructed as a product of a genotype-based matrix, which also means that the kinship matrices can always be factorised to the form $K = WW^T$. Because of this, the spectral decomposition products of a kinship matrix K can be obtained directly from singular value decomposition (SVD) of the genotype-based matrix W without the need to calculate the kinship matrix first (Lippert *et al.*, 2011). Since FaST-LMM uses these spectral decomposition products rather than the actual kinship matrix in its GWAS

calculation, it can bypass the calculation of the kinship matrix altogether if this will be more efficient—a unique feature among the LMM packages considered here. As the time required for kinship matrix calculation is $O(sn^2)$, and for spectral decomposition of the matrix is generally $O(n^3)$, whereas the time required for SVD of the genotype matrix is $O(s^2n)$, FaST-LMM will bypass the computation of kinship matrix whenever the number of SNPs s is less than the number of samples n (Lippert *et al.*, 2011).

Because of FaST-LMM's computational advantage when $s \ll n$, it is quite natural to attempt to use the smallest set of SNPs that could still yield accurate results. In their original article, Lippert *et al.* (2011) used just 200 SNPs, selected based on their association with the phenotype, to successfully control the analysis of Wellcome Trust Case Control Consortium (WTCCC) data for Crohn's disease. This idea seems to have developed further in subsequent versions of FaST-LMM (version 2.00 and later), in which a class of methods ("FaST-LMM-Select") is implemented for selection of a small number of SNPs for kinship calculation. The actual implementation of these seems to differ among different versions of FaST-LMM. In an earlier version (2.00), SNPs were first ordered according to their linear regression p-values, after which kinship matrices were constructed iteratively with an increasing number of the top-ranking SNPs for use in LMM analysis, until the first minimum genomic control factor λ is obtained (Listgarten *et al.*, 2012). In a later version (2.05), a fully automated but slightly different procedure was implemented. This involves k-fold cross validation (Lippert *et al.*, 2013), with the ordering of SNPs and calculation of genomic control factors as varying numbers of SNPs are included in the kinship calculation carried out within the training data (and then used to predict the test data) within each cross-validation fold. The final number of SNPs to be used in the kinship calculation for the entire data set is that which minimises the mean-squared error summed over all folds. Both of these procedures were investigated in this thesis.

Another unique feature of FaST-LMM is that it is implemented as a multithreaded program. This allows parallelisation without needing explicit intervention from the user, thus gaining further advantage when used on a multi-core system.

GEMMA

Zhou and Stephens (2012) implemented an exact approach extremely similar to that of FaST-LMM in their package GEMMA. Indeed, they point out that GEMMA should give essentially identical inference to FaST-LMM in the same time complexity $O(n^3 + sn^2 + rsn)$, but note that the number of iterations r required to reach convergence in GEMMA is expected to be slightly smaller than in FaST-LMM, owing to the use of a more efficient optimisation method. GEMMA also has an attractive practical advantage

of allowing the input of imputed (Marchini *et al.*, 2007) genotype data, rather than real measured genotype data, if desired.

GEMMA provides two methods for relatedness matrix calculation: that based on centred genotype, and that based on (centred and) standardised genotypes. The latter is mathematically similar to GenABEL's weighted IBS, EMMAX's Balding-Nichols and FaST-LMM's covariance matrices.

Mendel

An approximate (score test) LMM implementation, suitable for analysis of GWAS data, has also been implemented in the software package Mendel (K. Lange *et al.*, 2013) (versions 13.0 and higher). A slower (exact) LMM implementation is also available, but only the approximate test is considered here. The resulting tests should be conceptually extremely similar to the LMM tests implemented in other software packages such as EMMAX and FaST-LMM.

For kinship estimation, Mendel can:

- calculate kinship coefficients on the basis of known pedigree relationships
- use the full set of genome-wide SNP data to cluster people into apparent pedigrees and then estimate kinship coefficients within those pedigree clusters; or
- use kinship coefficients estimated for all pairs of genotyped individuals on the basis of their full set of genome-wide SNPs.

The results presented in this thesis are based on the last option (kinship estimated from all genotyped individuals using full set of SNPs), and the analysis was performed by my supervisor. Results based on the other options are not presented here, but are available in our published article (Eu-ahsunthornwattana *et al.*, 2014b).

MMM

Pirinen *et al.* (2013) have implemented approximate and exact approaches similar to the approximate and exact approaches of FaST-LMM (and the exact approach of GEMMA) in their package MMM. An advantage of MMM in comparison to the other packages is that it allows the output of regression coefficients and standard errors for the SNP effects on the (log) odds ratio scale, making it convenient to compare or combine the results with results from traditional case/control studies analysed via logistic regression. In addition, MMM allows the input of imputed genotype data rather than real measured genotype data, if desired. MMM was used in the original analysis of the Brazilian VL family data described in Fakiola *et al.* (2013). For more details on the methodology implemented in MMM, see Pirinen *et al.* (2013).

MMM can read any positive semi-definite kinship matrix or its spectral decomposed products (likely to be from the previous run) for its calculation. Within the MMM package, the program 'generateR' provided can be used to calculate such matrix based on standardised genotype data in a similar manner to EMMAX's Balding-Nichols matrix.

4.1.2. Alternative methods

The results from the above LMM analyses were compared with those from the non-LMM methods that are designed specifically for analysis of family-based data or to allow for relatedness described below. (Analyses in this section were conducted by my supervisor.)

FBAT

Traditional approaches for family-based association analysis focus on the transmission of high-risk alleles through pedigrees, in an approach that is closely related to traditional linkage analysis. Indeed, the well-known transmission disequilibrium test (TDT) (Spielman *et al.*, 1993), which tests whether a particular allele is transmitted preferentially from heterozygous parents to affected offspring, was originally developed as a test of linkage in the presence of association, rather than as a test of association per se. In this context, 'linkage' means the transmission from parent to offspring of alleles in coupling at a test (marker) locus and an unobserved causal locus, i.e. the phenomenon whereby alleles that are in coupling (on the same haplotype) in the parent tend to be transmitted together to the offspring, whereas 'association' means population-level correlation between alleles at the two loci (usually referred to as linkage disequilibrium (LD)), i.e. the tendency for alleles at the two loci to occur in coupling in the founders of a pedigree.

The TDT was originally designed for the analysis of case/parent trios (i.e. units consisting of an affected child together with their parents) but has been extended to allow analysis of nuclear families and larger pedigrees (Laird *et al.*, 2000; Martin *et al.*, 2000; Rabinowitz and Laird, 2000; Horvath *et al.*, 2001; C. Lange *et al.*, 2004; Dudbridge, 2008; Dudbridge *et al.*, 2011). The focus here is on the family-based association test (FBAT) (Laird *et al.*, 2000; Horvath *et al.*, 2001), as implemented in the FBAT software package. FBAT can be thought of as a general class of test statistics of the form:

$$\frac{S - E(S)}{\sqrt{\mathrm{Var}(S)}}$$

where $S = \sum_{ij} T_{ij} X_{ij}$ and X_{ij} is some genotype variable and T_{ij} some trait variable for offspring i in nuclear family j. The exact form of FBAT thus depends on the genotype

and trait coding used. Genotype is generally coded in allelic fashion with a variable coded (0, 1, 2) according to the number of copies of the high-risk allele possessed. The trait variable is constructed as $T_{ij} = Y_{ij} - \mu_{ij}$ where Y_{ij} is coded 0/1 (for binary traits such as disease status) and μ_{ij} is an offset that can be chosen to consider transmissions to affected offspring only (the default), or else to contrast transmissions to affected offspring with transmissions to unaffected offspring, either weighted equally ($\mu_{ij} = 0.5$) or with μ_{ij} chosen to minimise the variance of test statistic. For quantitative traits, Y_{ij} would generally correspond to the measured trait for offspring i in nuclear family j, with μ_{ij} set to equal the mean trait value or else chosen to minimise the variance of test statistic.

Although, for binary traits, contrasting transmissions to affecteds with transmissions to unaffecteds seems an attractive idea, in practice this results in comparing the probability of transmission of high-risk alleles to affected individuals (which is expected, under the alternative hypothesis, to exceed 0.5) with an *estimate* of the probability of transmission of high-risk alleles to unaffected individuals (which is expected, under both null and alternative hypotheses, to approximately equal 0.5, unless the effect of the risk allele is large), rather than comparing the transmission probability to affecteds with an assumed fixed value of 0.5. For complex diseases, where the effects of risk alleles are likely to be modest (allelic odds ratios in the order of 1.2-1.5), this means that greater power would be expected from the default offset that considers transmissions to affected offspring only, without paying a penalty for (imperfect) estimation of the expected 0.5 transmission probability (along with a measure of uncertainty in the estimate) from the data at hand.

By default, FBAT divides larger pedigrees into nuclear families and constructs a test that corresponds to testing 'linkage in the presence of association' (Horvath *et al.*, 2001). The '`-e`' option in FBAT allows the alternative construction of a test for 'association in the presence of linkage' (Lake *et al.*, 2000) through the use of an empirical variance/covariance estimator that adjusts for the correlation among sibling genotypes and for different nuclear families within a single pedigree. Use of the '`-e`' option is expected to give smaller test statistics (larger p-values) than the default analysis, since it accounts for the fact that the effective sample size is smaller when considering FBAT as a test of association than as a test of linkage. Since, for complex diseases, one is more interested in maximising the power for detection of an effect, rather than in ensuring that the detection is genuinely driven by association (rather than linkage) between alleles at the test locus and the underlying unobserved causal locus, the default option was used in all analyses presented here. From a practical point of view, this means that any signal detected may in fact be marking a true effect that

lies some distance away, rather than necessarily being located in the immediate vicinity of the detected signal.

ROADTRIPS and MQLS

Thornton and McPeek (2010) implemented a '**RO**bust **A**ssociation-**D**etection **T**est for **R**elated **I**ndividuals with **P**opulation **S**ubstructure' in a package called ROADTRIPS. ROADTRIPS can be thought of as an extension of their previously-proposed Maximum Quasi-Likelihood Statistic (MQLS) (Thornton and McPeek, 2007). Both MQLS and ROADTRIPS construct adjusted versions of standard case/control χ^2 (or Armitage Trend) tests, adjusting for the known relatedness between individuals (that would ordinarily cause an inflation in standard case/control tests) through a kinship matrix that models the known pedigree relationships. ROADTRIPS (but not MQLS) additionally makes use of a covariance matrix based on estimated kinships (as estimated from genome-wide SNP data) to further correct for additional unknown relatedness and population stratification.

The ROADTRIPS test statistic takes the form:

$$\frac{(\boldsymbol{V}^T\boldsymbol{Y})^2}{\hat{\sigma}^2\boldsymbol{V}^T\hat{\Psi}\boldsymbol{V}} \sim \chi_1^2$$

Thornton and McPeek note that many commonly-used case/control statistics can be coerced into this form. Here $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ is genotype vector at a test SNP for n individuals (coded using an allelic coding), $\boldsymbol{V}$ is a vector of length n coding for phenotype information (disease status) and known (or externally estimated) relationships (see Thornton and McPeek (2010) for details of its construction), $\hat{\sigma}^2\hat{\Psi}$ is an estimate of the null variance/covariance matrix of $\boldsymbol{Y}$ (so that $\hat{\sigma}^2\boldsymbol{V}^T\hat{\Psi}\boldsymbol{V}$ is an estimate of null variance/covariance of $(\boldsymbol{V}^T\boldsymbol{Y})^2$), $\hat{\sigma}^2$ is an estimate of $\mathrm{Var}(\boldsymbol{Y})$ in an outbred population and $\hat{\Psi}$ is an internally estimated matrix used to simultaneously adjust for unknown relatedness/pedigree relationship errors and population stratification.

MASTOR and GTAM

Jakobsdottir and McPeek (2013) proposed a retrospective approach (MASTOR) for analysis of quantitative traits that can be considered essentially as a quantitative trait version of MQLS. Jakobsdottir and McPeek compared MASTOR to a previously-proposed LMM method, GTAM (Abney *et al.*, 2002), and found MASTOR to have some advantages. The main advantage of MASTOR over GTAM (and many other approaches) is that, in common with MQLS and ROADTRIPS, MASTOR allows information to be gained from individuals who are phenotyped but not genotyped. Both MASTOR and GTAM are implemented within the MASTOR software package. Although designed for analysis of quantitative (rather than binary) traits, given that the spirit of recent LMM

approaches has been to apply approaches originally designed for quantitative traits to binary traits (coded as 0 and 1), the performance of MASTOR and GTAM when applied to both binary and quantitative traits was investigated here.

In common with MQLS, kinships used in the MASTOR package are assumed to be estimated on the basis of known pedigree relationships. Although in principle kinships estimated from genome-wide SNP data could be read in instead, the results presented in this thesis were analysed using pedigree relationships.

4.1.3. Methods used only for kinship calculation

Unlike the native methods for kinship calculation used in the above LMM packages, which attempt to estimate a kinship measure (which would also reflect the extent of IBD sharing among the individuals) from genotype data for the purpose of subsequent LMM analysis, methods presented in this section explicitly attempt to infer the IBD sharing among individuals based on genomic data. This was intended for use, for example, in GWAS quality control (Purcell *et al.*, 2007), or to infer pedigree relationships (Manichaikul *et al.*, 2010), but can also be used to calculate the kinship matrix for use in LMM analysis.

The kinships thus estimated were fed into an LMM software package (FaST-LMM) to investigate the effect of different types of kinship estimation on the LMM results.

PLINK

The `--genome` (and its variant `--Z-genome`) command in PLINK (Purcell *et al.*, 2007) estimates the pairwise IBD among homogeneous samples given the IBS information, which can be readily estimated from genotype data. Purcell *et al.* (2007) start by calculating the probability that a pair of individuals share 0 allele IBD:

$$P(Z = 0) = \frac{N(I = 0)}{N(I = 0|Z = 0)}$$

where $N(I = 0)$ is the count of SNPs with IBS state $I = 0$ (which can only occur if the two individuals are opposite homozygotes at that particular locus) and $N(I = 0|Z = 0)$ is the expected count of SNPs with IBS state $I = 0$ given that the pair share 0 allele IBD at each locus, which, under the assumption of Hardy-Weinberg equilibrium, depends only on the allele counts in the samples. Having obtained this, the probability that the pair share 1 allele IBD can then be estimated as:

$$P(Z = 1) = \frac{N(I = 1) - P(Z = 0)N(I = 1|Z = 0)}{N(I = 1|Z = 1)}$$

The remaining IBD state $P(Z = 2)$ can then be analogously estimated once $P(Z = 0)$ and $P(Z = 1)$ are known.

PLINK automatically bounds and constrains the resulting IBD probabilities to biologically plausible values, which can then be used to calculate the proportion of alleles shared IBD (which equals twice the kinship coefficient).

KING

Manichaikul *et al.* (2010) proposed two alternative methods for kinship inference which are implemented in their **K**inship-based **IN**ference for **G**enome-wide association studies (KING) software package. One of these, 'KING-homo', assumes that the samples come from a single, homogenous population, in a similar manner to PLINK; the other, 'KING-robust', does not make that assumption and is therefore robust against population structure. Instead of sequentially estimating each IBD state probability, KING estimates the IBD sharing probability $P(Z = 0)$ using similar algorithm to PLINK, but then proceeds to estimating the kinship coefficient directly from the allele counts using a simplified and optimised algorithm, after which the probabilities of the two remaining IBD states can be derived. This results in a much faster calculation than that used in PLINK (Manichaikul *et al.*, 2010).

Unlike PLINK, the results from KING do not appear to be bounded to biologically plausible values, which could cause problems when fed into certain programs.

4.2. Comparison of Different SNP Sets and Different Methods/Software for Kinship Measure Estimation

Most software packages considered here allow a separation between the kinship matrix estimation and the actual GWAS analysis incorporating the desired kinship matrix. This is useful, perhaps in a rather unintended way, as it allows comparisons of the kinship matrices obtained from various LMM methods, as well as those from the software used only for kinship calculation, and based on various sets of SNPs; it also allows comparisons of the GWAS analysis results from the same LMM methods, using kinship matrices estimated using different sets of SNPs, or even estimated using different methods of kinship estimation.

The ability to do this is crucial in addressing two important questions that needed to be answered before attempting further exploration of the LMM methods in GWAS analysis: whether there is any significant difference in kinship measures estimated from various methods; and what would be the optimal set of SNPs, if any, to use for kinship estimation for the purpose of LMM GWAS analysis. Conclusions made from this section were used in all (GAW18 and VL) subsequent LMM GWAS analyses.

4.2.1. Comparison of different kinship measure estimation methods using similar sets of SNPs

To answer the first question, kinship measures were estimated using the various LMM and kinship calculation methods being considered, using each set of SNPs (full, pruned and thinned; see Section 2.2.4 for detailed description). The kinship measures estimated from each method using a similar set of SNPs were then compared.

Although the scale on which the kinship estimates were measured differed between different packages, the measures themselves were highly correlated for each SNP set (Figures 4.1-4.3). Nevertheless, the estimates based on the thinned SNP set appear to be slightly less correlated when compared to those based on the other two sets.

The correlation was particularly high among kinship measures from EMMAX (Balding-Nichols), FaST-LMM (both methods), GenABEL, GEMMA and MMM, as could be theoretically expected. Also quite correlated, but slightly different from those in the previous group, were the kinship measures from EMMAX (IBS) and PLINK. However, the calculated correlation coefficients were somewhat lower between these two compared to between each of these and the other methods; this was despite the correlation plots of the kinship measures from these two methods showing high degree of concordance. This was due to the discrepancy among the more distantly related pairs of individuals, which can be seen near the origin of the plots.

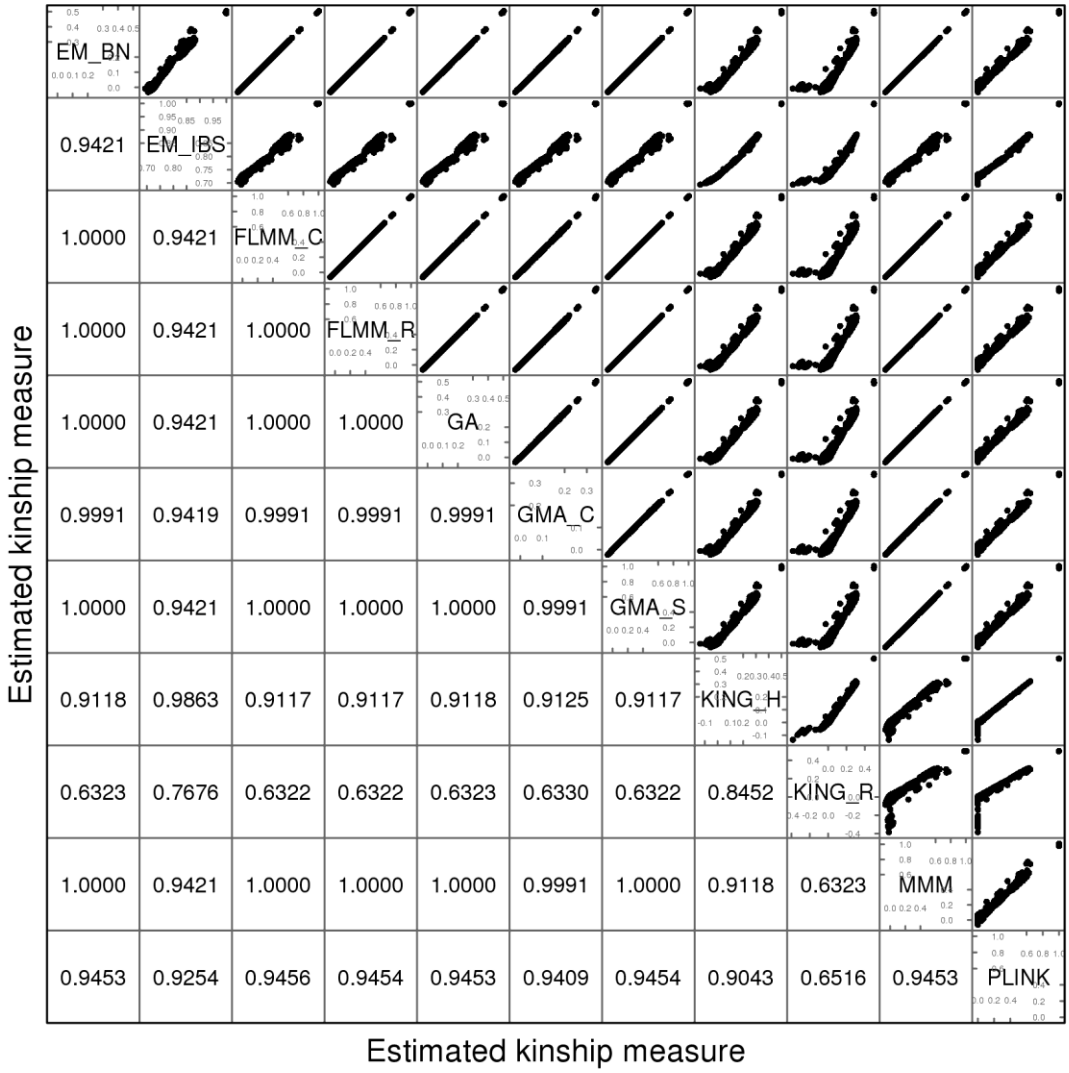

Figure 4.1 Comparison of kinship measures estimated from full genome-wide SNP set using different software packages. Plots above the diagonal show a comparison of the kinship measures estimated by two of the methods being compared, with correlation between the kinship measure estimates indicated below the diagonal. EM_BN = EMMAX (Balding-Nichols), EM_IBS = EMMAX (IBS method), FLMM_C = FaST-LMM using covariance matrix, FLMM_R = FaST-LMM using realised relationship matrix, GA = GenABEL, GMA_C = GEMMA using centred genotypes, GMA_S = GEMMA using standardised genotypes, KING_H = KING with homogeneous population assumption, KING_R = KING with robust estimation. Unlike the other methods, KING did not constrain negative values to zero, which resulted in apparently low correlation coefficients, particularly for KING_R.

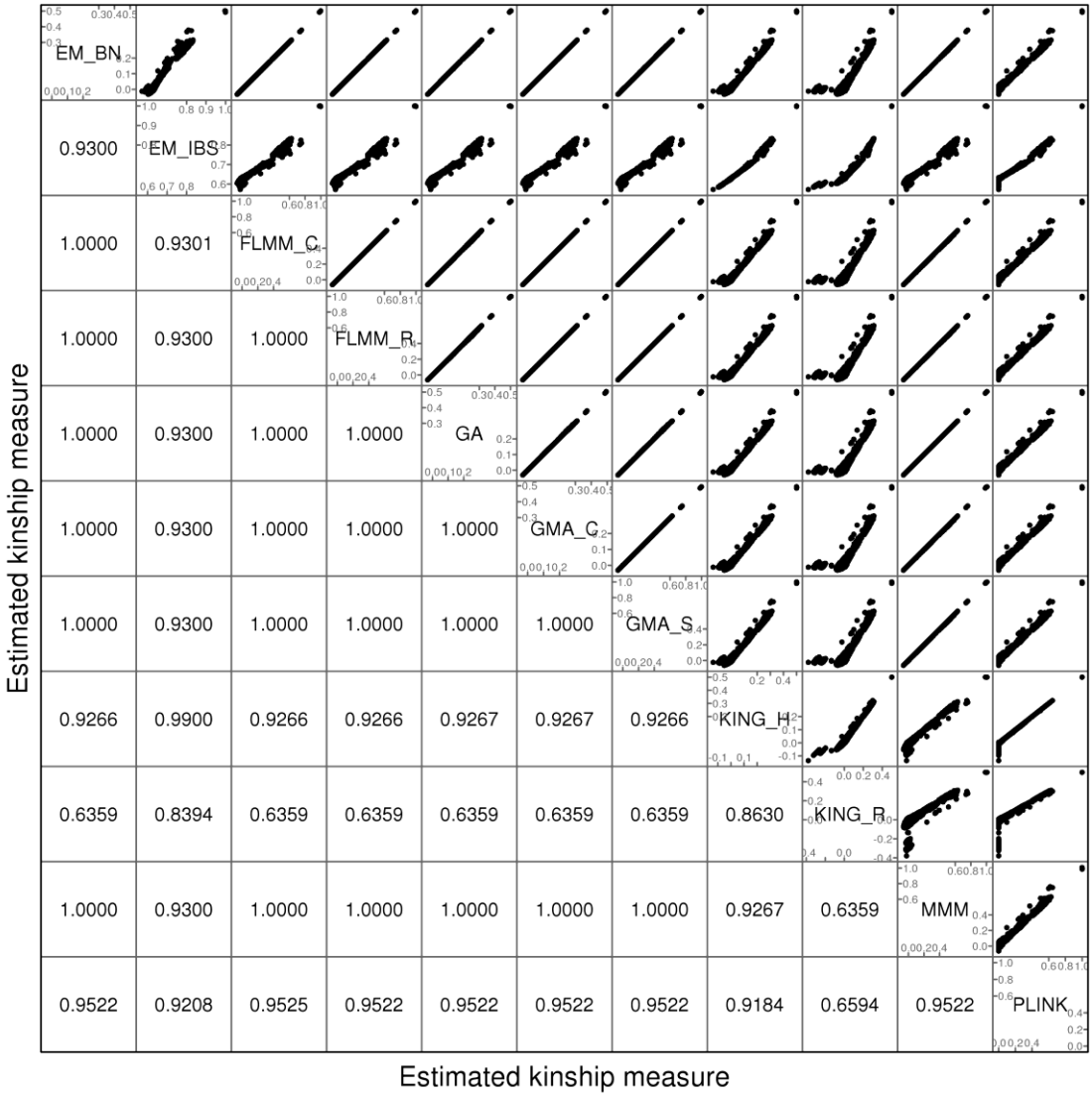

Figure 4.2 Comparison of kinship measures estimated from pruned SNP set using different software packages. Plots above the diagonal show a comparison of the kinship measures estimated by two of the methods being compared, with correlation between the kinship measure estimates indicated below the diagonal. EM_BN = EMMAX (Balding-Nichols), EM_IBS = EMMAX (IBS method), FLMM_C = FaST-LMM using covariance matrix, FLMM_R = FaST-LMM using realised relationship matrix, GA = GenABEL, GMA_C = GEMMA using centred genotypes, GMA_S = GEMMA using standardised genotypes, KING_H = KING with homogeneous population assumption, KING_R = KING with robust estimation. Unlike the other methods, KING did not constrain negative values to zero, which resulted in apparently low correlation coefficients, particularly for KING_R.

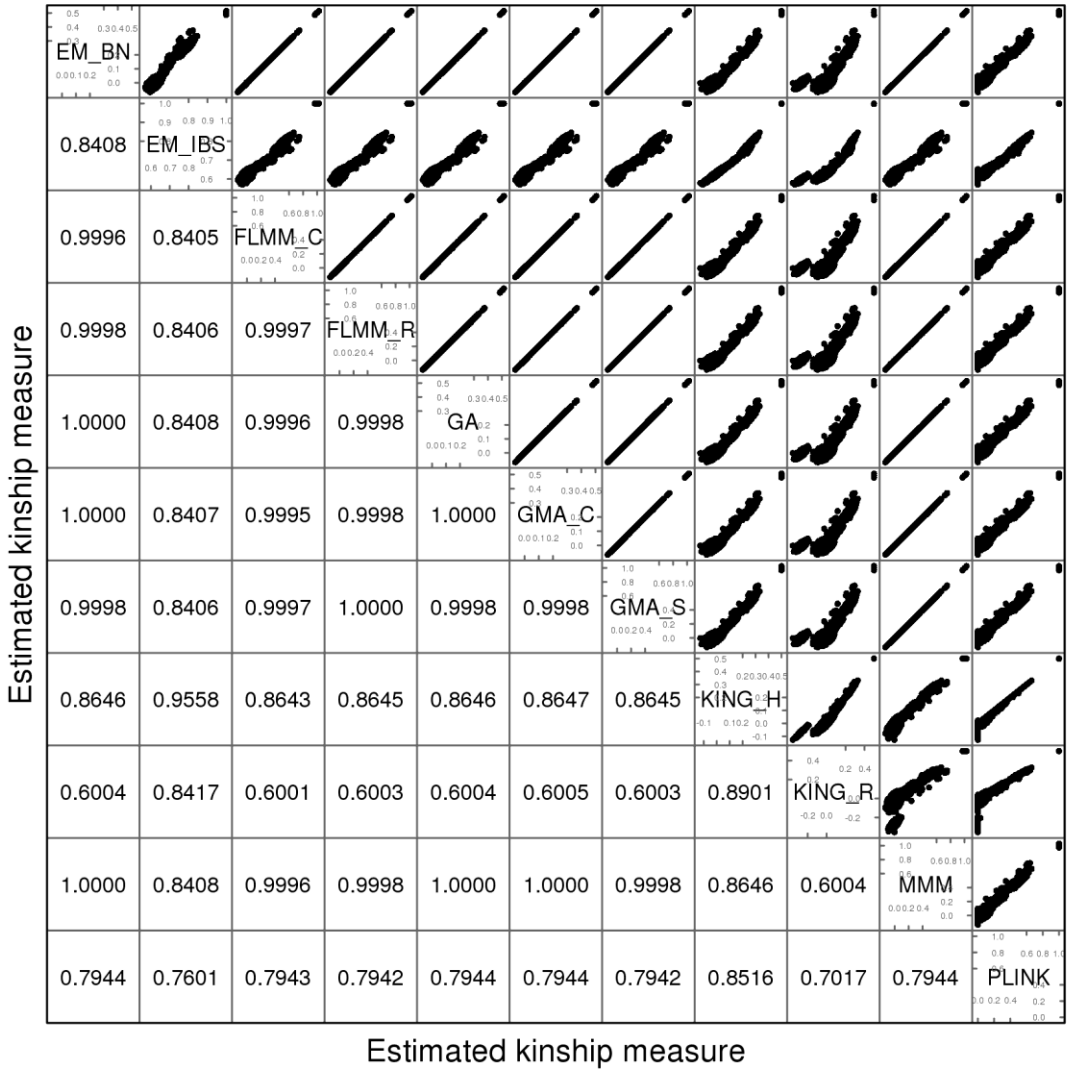

Figure 4.3 Comparison of kinship measures estimated from thinned SNP set using different software packages. Plots above the diagonal show a comparison of the kinship measures estimated by two of the methods being compared, with correlation between the kinship measure estimates indicated below the diagonal. EM_BN = EMMAX (Balding-Nichols), EM_IBS = EMMAX (IBS method), FLMM_C = FaST-LMM using covariance matrix, FLMM_R = FaST-LMM using realised relationship matrix, GA = GenABEL, GMA_C = GEMMA using centred genotypes, GMA_S = GEMMA using standardised genotypes, KING_H = KING with homogeneous population assumption, KING_R = KING with robust estimation. Unlike the other methods, KING did not constrain negative values to zero, which resulted in apparently low correlation coefficients, particularly for KING_R.

KING tended to give kinship estimates that differ most from the other methods, with frequent output of negative kinship estimates among the less related individuals (these were bounded at 0 in most other methods). This was more pronounced for KING (robust) than for KING (homogeneous). The possible implications of this for LMM analysis will be considered in Section 4.4.

4.2.2. Comparison of kinship measures estimated based on different sets of SNPs

To identify a robust set of SNPs for use in subsequent kinship measure estimation, each of the kinship measure estimation methods was applied to the three sets of SNPs (full, pruned and thinned; see Section 2.2.4 for detailed description). The results of these are shown in Figure 4.4.

The kinships estimated by any method, using any of the three SNP sets, correlated well with the theoretical kinship coefficients calculated using the pedigree relationship, considering the discrete nature of theoretical kinship coefficients (Figure 4.4 A). The kinships estimated using the full, genome-wide SNP set and the pruned SNP set were highly concordant in all methods, whereas the estimates based on the thinned SNP set tended to be slightly different from these (Figure 4.4 B). The implication of this for LMM GWAS analysis will be addressed in Section 4.2.3.

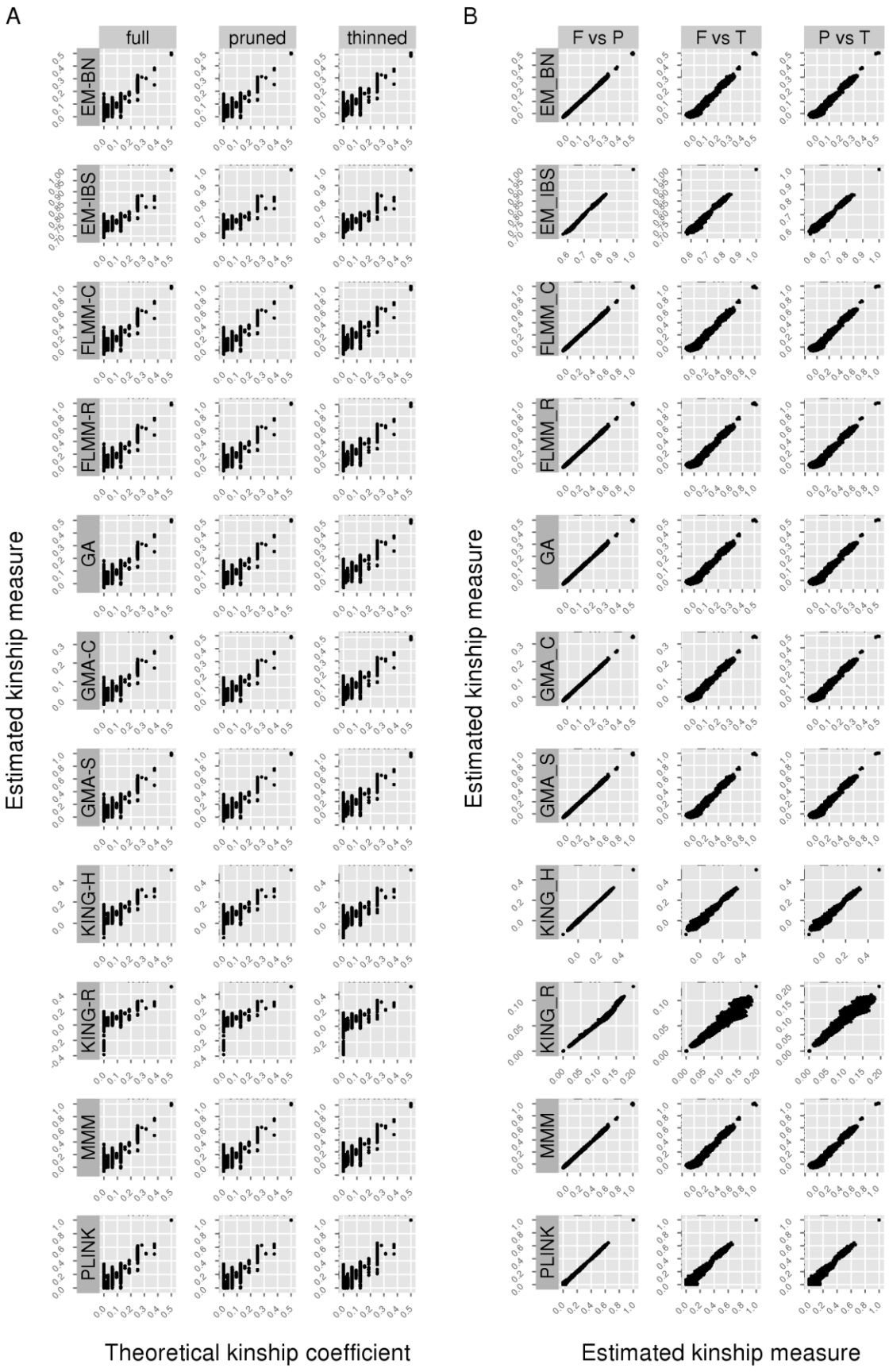

Figure 4.4 Comparison of estimated kinship measures based on full, pruned and thinned SNP sets against theoretical (pedigree-based) kinship coefficients (A) and against each other (B). F = full set, P = pruned set, T = thinned set. EM_BN = EMMAX (Balding-Nichols), EM_IBS = EMMAX (IBS method), FLMM_C = FaST-LMM using covariance matrix, FLMM_R = FaST-LMM using realised relationship matrix, GA = GenABEL, GMA_C = GEMMA using centred genotypes, GMA_S = GEMMA using standardised

genotypes, KING_H = KING with homogeneous population assumption, KING_R = KING with robust estimation.

4.2.3. Comparison of LMM results based on kinship measures estimated using different sets of SNPs

When the kinship measures estimated using different sets of SNPs were used in LMM GWAS analysis, the resulting p-values appeared to follow a similar pattern to that observed in the previous section (4.2.2) in relation to the measures themselves, that is, the results based on kinships estimated using the pruned SNP set were very similar to those based on kinships estimated using full, genome-wide SNP set, whereas the results based on kinships estimated using the thinned SNP set, whilst still highly correlated with the other two, differ somewhat from them (Figure 4.5).

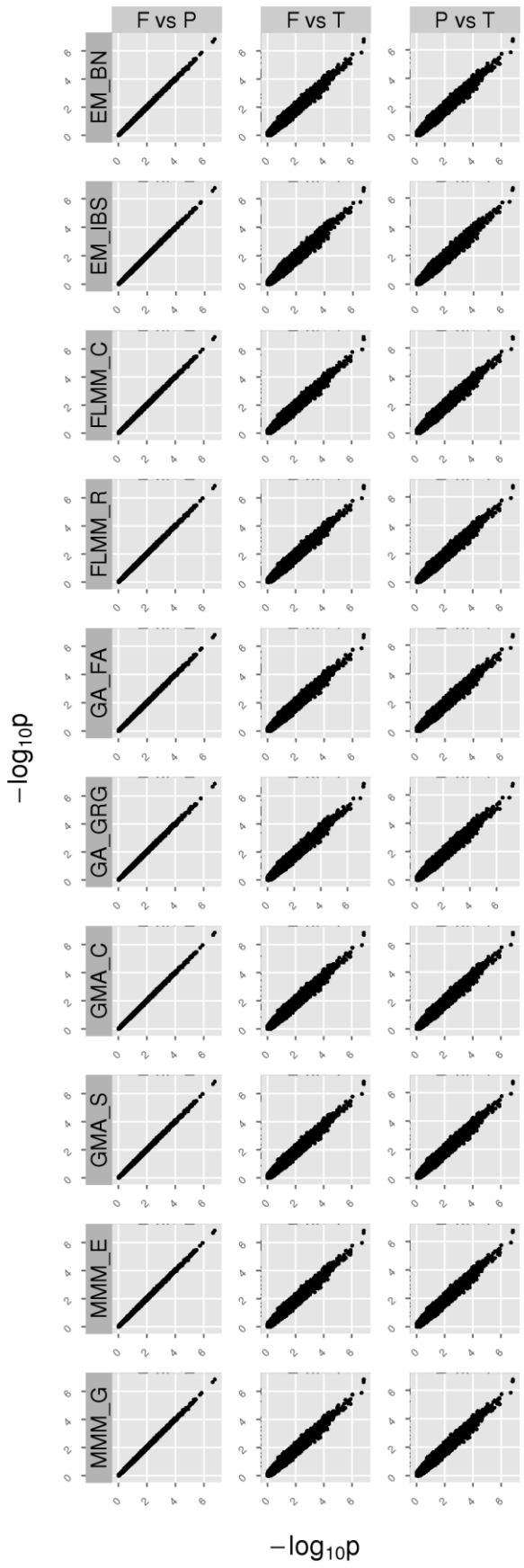

Figure 4.5 Comparison of –log10 (p-values) obtained based on full, pruned and thinned SNP sets against theoretical (pedigree-based) kinship coefficients (A) and against each other (B). F = full set, P = pruned set, T = thinned set. EM_BN = EMMAX (Balding- Nichols), EM_IBS = EMMAX (IBS method), FLMM_C = FaST-LMM using covariance matrix, FLMM_R = FaST-LMM using realised relationship matrix,

GA_FA = GenABEL (FASTA), GA_GRG = GenABEL (GRAMMAR-Gamma), GMA_C = GEMMA using centred genotypes, GMA_S = GEMMA using standardised genotypes, MMM_E = MMM using full mixed model (exact) calculation, MMM_G = MMM using GLS approximation.

In terms of controlling the genome-wide type 1 error rate, i.e. controlling the genomic inflation factor λ (Devlin and Roeder, 1999) to the desired level of λ = 1, all methods performed well when using full or pruned set of SNPs, with λ of 0.99-1.00, but less so when the thinned set was used (Figure 4.6). The λ achieved when the thinned set of SNPs was used were mostly between 1.08-1.10, with the exception of GenABEL (GRAMMAR-Gamma), which had the most inflated λ of 1.16.

Nevertheless, even when the thinned set of SNPs was used, the genomic inflation control achieved was still superior to when the theoretical kinship was used (λ = 1.11), which in turn was substantially superior to when no adjustment was made, in which case the λ was 1.23 (Figure 4.6).

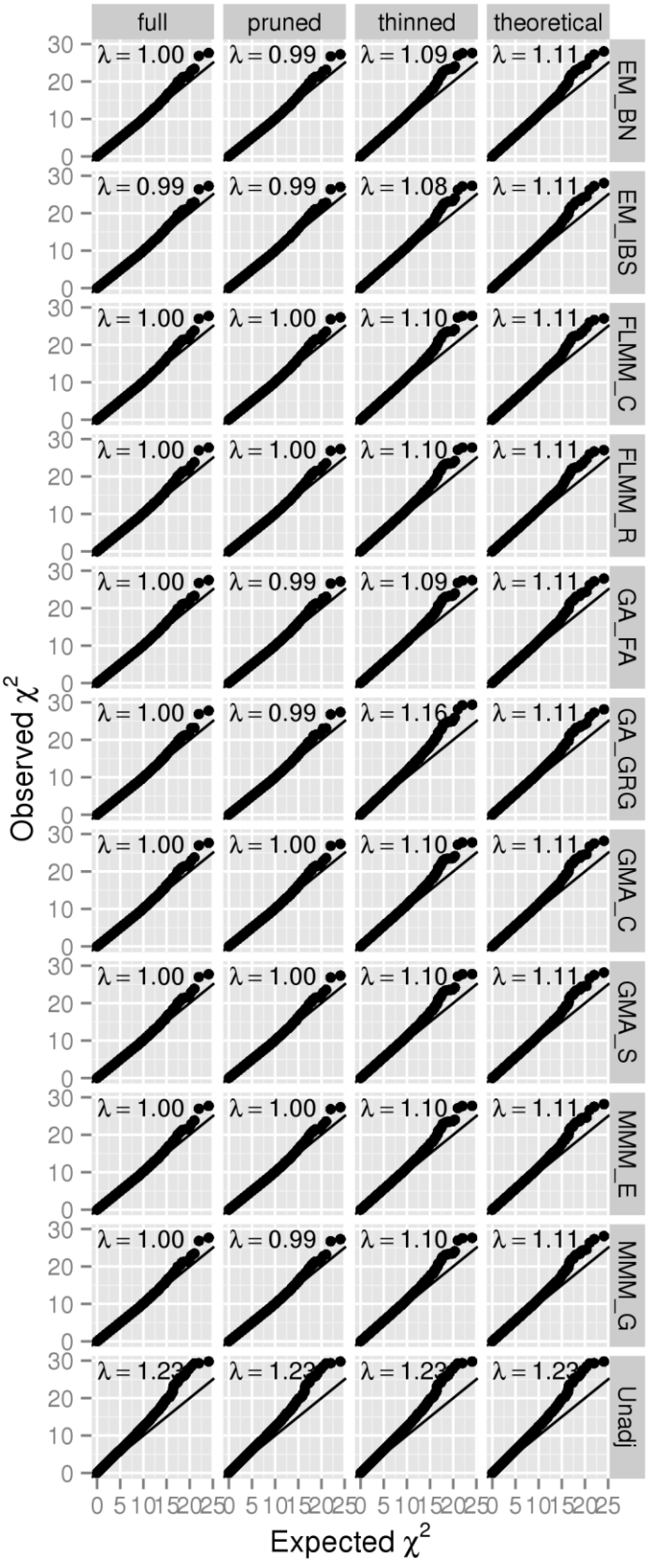

Figure 4.6 Q-Q plots of real VL phenotype GWAS results, using different LMM software packages and different SNP sets for kinship estimation. The black diagonal lines represent the line of equality. Where a method gave only the p-values, the equivalent 1-degree of freedom χ^2 values were used. The 'theoretical' set used pedigree structure to derive theoretical kinship coeffi cients. EM_BN = EMMAX (Balding- Nichols), EM_IBS = EMMAX (IBS method), FLMM_C = FaST-LMM using covariance matrix, FLMM_R = FaST-LMM using realised relationship matrix,

GA_FA = GenABEL (FASTA), GA_GRG = GenABEL (GRAMMAR-Gamma), GMA_C = GEMMA using centred genotypes, GMA_S = GEMMA using standardised genotypes, MMM_E = MMM using full mixed model (exact) calculation, MMM_G = MMM using GLS approximation, Unadj = unadjusted analysis. For methods with two ways to estimate the kinships, the same 'theoretical' results were plotted twice. Unadjusted analysis results were plotted once in each column only for comparison, and did not use the kinship estimates for adjustment.

The above observation that the adjustment using either the full or pruned SNP sets was mostly equivalent can be seen graphically in Figure 4.7. Although adjustment using the thinned SNP set tended to be better than using pedigree information alone (with the exception of GenABEL GRAMMAR-Gamma), marked improvement was seen when the pruned set was used instead of the thinned set.

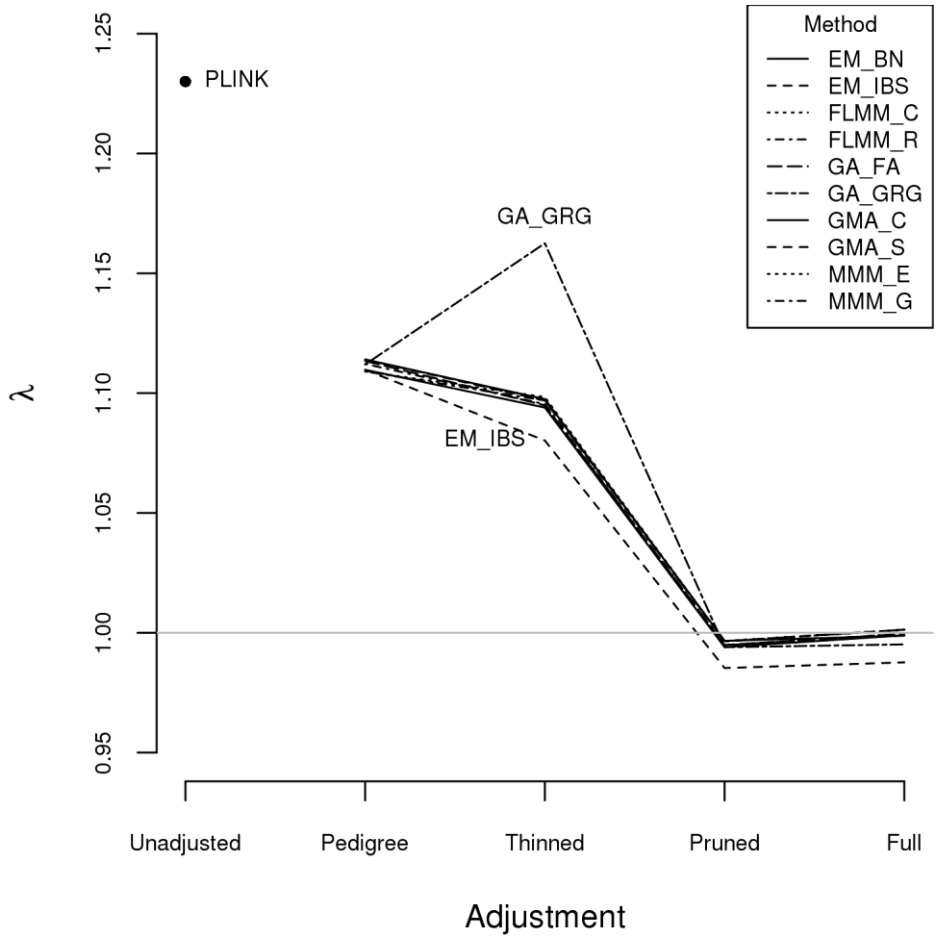

Figure 4.7 Genomic control factors obtained using different software packages, different strategies for modelling kinships and different sets of SNPs. PLINK = analysis in PLINK with no adjustment made for relatedness, EM_BN = EMMAX (Balding- Nichols), EM_IBS = EMMAX (IBS method), FLMM_C = FaST-LMM using covariance matrix, FLMM_R = FaST-LMM using realised relationship matrix, GA_FA = GenABEL (FASTA), GA_GRG = GenABEL (GRAMMAR-Gamma), GMA_C = GEMMA using centred genotypes, GMA_S = GEMMA using standardised genotypes, MMM_E = MMM using full mixed model (exact) calculation, MMM_G = MMM using GLS approximation.

The more recent versions of FaST-LMM provide a class of methods to select the most appropriate set of SNPs to use for kinship estimation when testing for association in a LMM framework ('FaST-LMM-Select'; see Section 4.1.1 for description). The common scheme among the methods in this class is to successively introduce SNPs, according to their nominal, unadjusted association with the phenotype, into the kinship estimation until an optimal inflation control has been reached. However, neither version of this approach seems to have performed satisfactorily in our data set.

The older version of this approach (implemented in FaST-LMM version 2.0), which involves systematic search for the set of SNPs (and therefore kinship matrix) that results in the first minimum of the genomic control factor λ (Listgarten *et al.*, 2012), gave λ that remained substantially higher than 1 at the first minimum (λ = 1.14 achieved with 3 ordered SNPs when starting from full SNP set, and λ = 1.11 achieved with 6 SNPs when starting from pruned SNP set). To explore this further, more SNPs were added to the kinship calculation after the first minimum had been reached, which resulted in subsequent decreasing of the λ to considerably less than 1, and then increasing back, eventually to 1, when all (pruned or full) SNPs had been included (Figure 4.8).

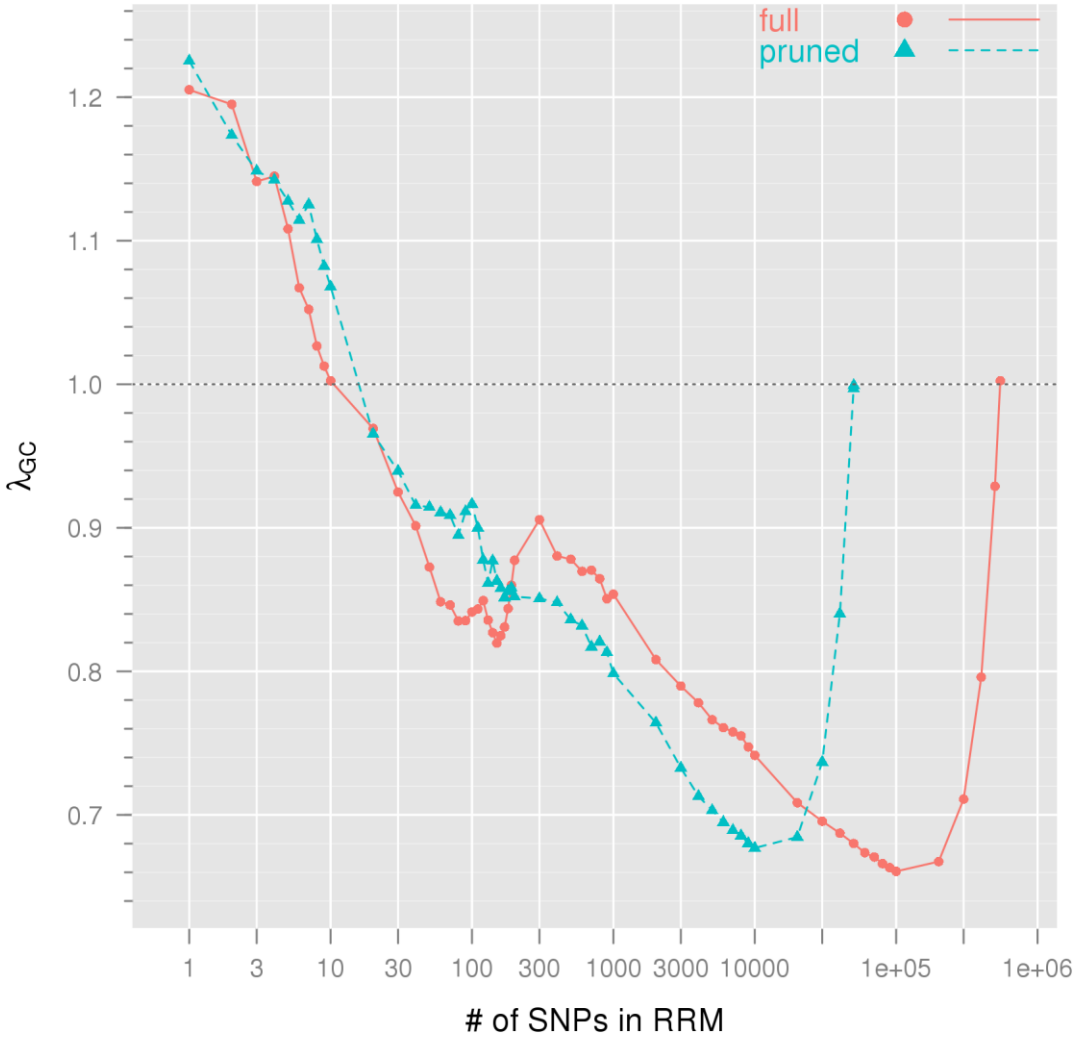

Figure 4.8 Performance of FaST-LMM-Select (v2.0). Genomic control factor (λ_{GC}) achieved in analysis of the real disease phenotype as different numbers of ordered SNPs are added in when calculating the kinship matrix (= realised relationship matrix, RRM). Method performed manually in FaST-LMM v2.0.

The newer version of FaST-LMM-Select, which is fully automated and involves minimizing the mean-squared error summed over *k* cross-validation folds (Lippert *et al.*, 2013), resulted in no SNP being selected for adjustment when starting from the full SNP set (and therefore would result in the maximum, unadjusted λ of 1.23), and 2 SNPs being selected for adjustment when starting from the pruned SNP set, which resulted in the genomic control value of 1.17.

Since these procedures seem to work less well than simply using all pruned or full SNPs for estimating pairwise kinships, while being practically more complicated, the remaining analyses involving FaST-LMM will be focused on the results obtained using the pruned SNP set for kinship estimation. Similarly, the pruned SNP set was also used in other methods, because of the substantially shorter computational time and the theoretical superiority compared with using the full SNP set due to the absence of

linkage disequilibrium between markers (although in practice the observed difference was minimal, LMM analyses using the pruned SNP set never performed worse than those using the full set).

4.3. Comparison of Association Analysis Results from LMM and Alternative Methods

The next questions to be addressed are how similar the results from the various LMM methods are when the same set of SNPs is used for controlling the relatedness among the samples, and how similar or different they are to the alternative methods.

The success (or otherwise) of the various methods in controlling the overall genome-wide type 1 error rate, as indicated by the ability to control the genomic inflation factor (Devlin and Roeder, 1999) λ to the desired level of $\lambda = 1$, is shown in Figures 4.9 and 4.10. All methods that made use of estimated kinships apart from MQLS ('MQLS_E' in Figure 4.10) performed well, being able to reduce the genomic inflation factor to around 1, compared to 1.23 in unadjusted analysis. For MQLS, the use of estimated kinships from 1972 genotyped individuals appeared to result in a slightly deflated genomic inflation factor (0.94).

Apart from FBAT, methods that used only theoretical kinships based on 'known' pedigree information (MASTOR and the other two MQLSs) tended not to be as successful in controlling the genomic inflation factor, resulting in a genomic inflation factor of about 1.15. Although they appeared to be quite well controlled for inflation, results from FBAT suffered from a different problem: they were very much in line with the theoretical distribution right up to the very top SNPs, suggesting little power to detect true effects. This will be more clearly demonstrated in the Manhattan plots.

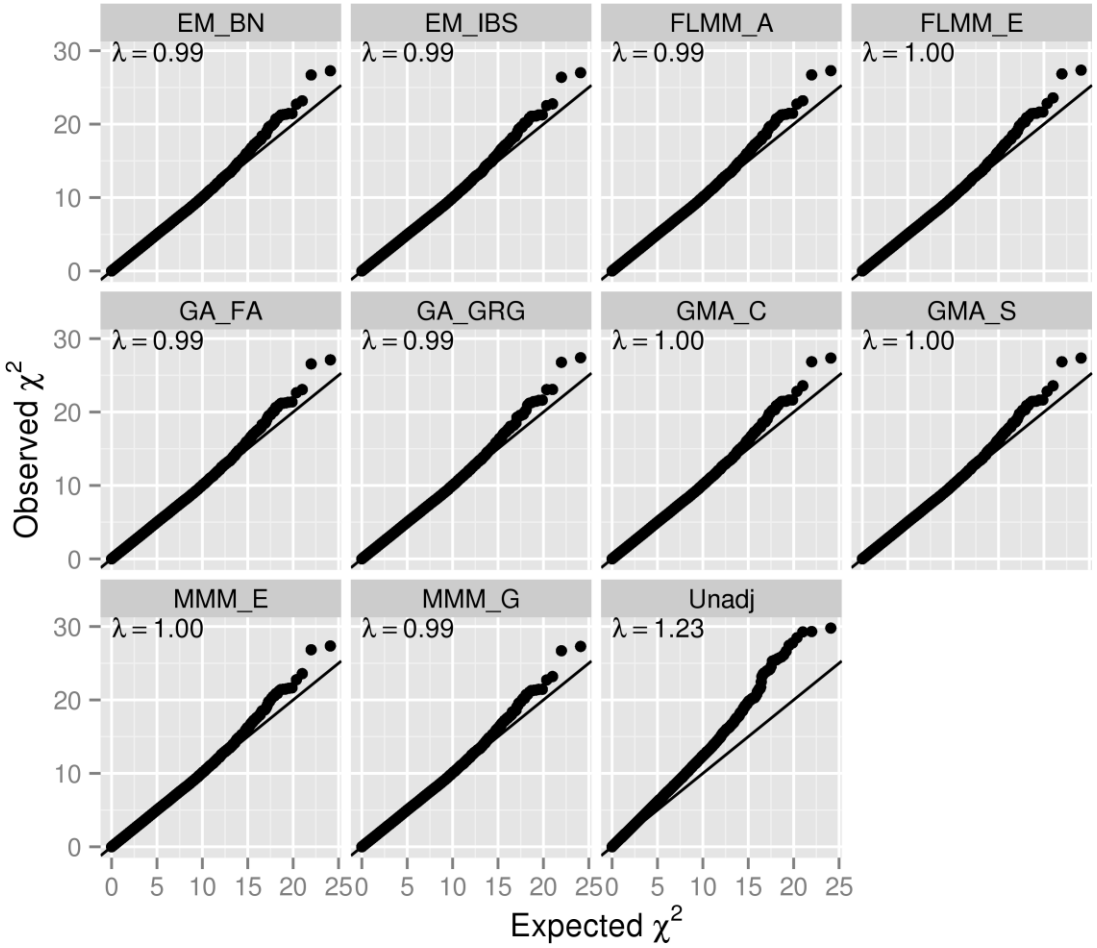

Figure 4.9 Q-Q plots of real VL phenotype GWAS results and genomic inflation factors (λ) for different LMM methods. EM_BN = EMMAX (Balding- Nichols), EM_IBS = EMMAX (IBS method), FLMM_A = FaST-LMM using approximate calculation with RRM, FLMM_E = FaST-LMM using exact calculation with RRM, GA_FA = GenABEL (FASTA), GA_GRG = GenABEL (GRAMMAR-Gamma), GMA_C = GEMMA using centred genotypes, GMA_S = GEMMA using standardised genotypes, MMM_E = MMM using full mixed model (exact) calculation, MMM_G = MMM using GLS approximation, Unadj = unadjusted analysis.

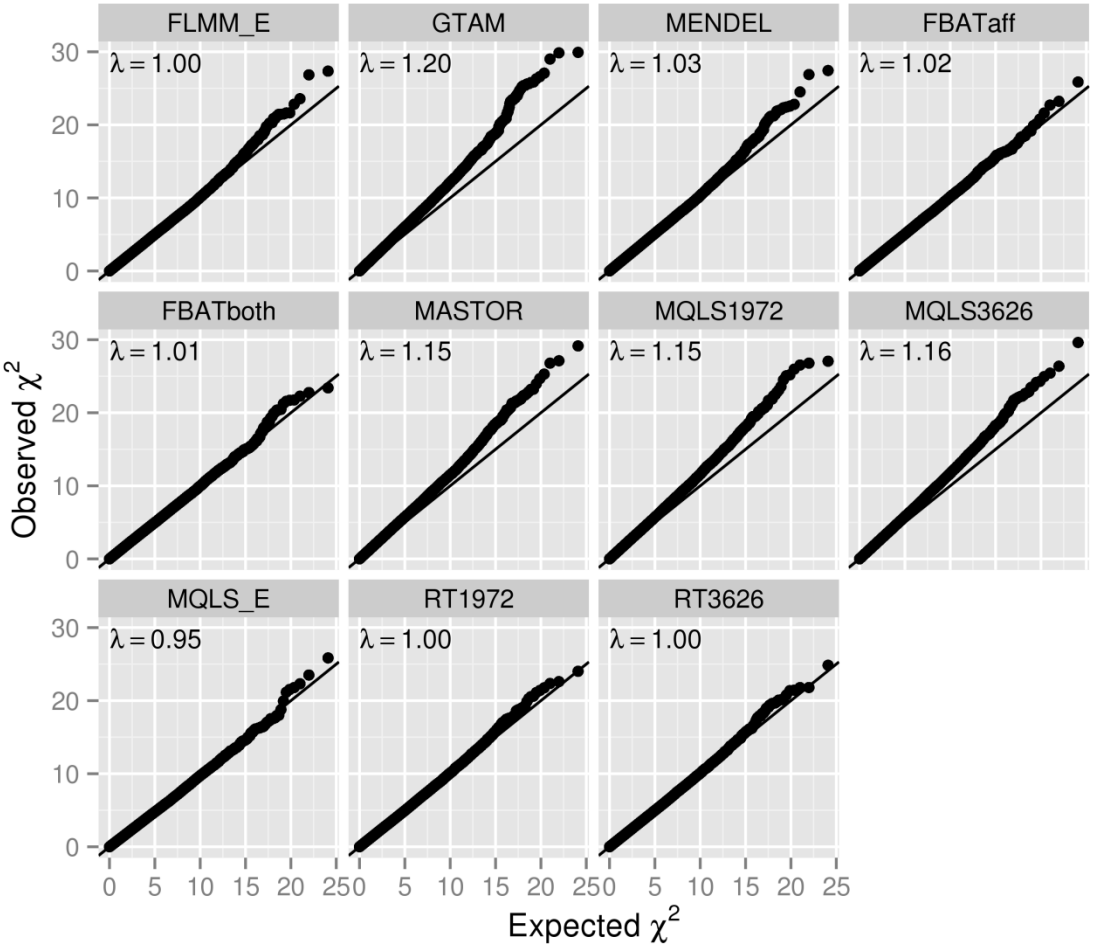

Figure 4.10 Q-Q plots of real VL phenotype GWAS results and genomic inflation factors (λ) for different LMM/alternative methods. FLMM_E = FaST-LMM using exact calculation with RRM, FBATaff = FBAT using transmissions to affecteds only, FBATboth = FBAT using transmissions to both affecteds and unaffecteds, MQLS1972/3626 = MQLS using theoretical kinships of either the 1,972 genotyped individuals or all 3,626 individuals in the pedigree, MQLS_E = MQLS using estimated kinships (1,972 individuals), RT1972/3626 = ROADTRIPS using 1,972 or 3,626 individuals. FaST-LMM is an LMM method and is included here for comparison; GTAM and Mendel are also LMM methods, but included here due to their unique characteristics (see Section 4.1)

The Manhattan plots of the results from LMM and alternative methods (Figures 4.11 and 4.12) appear to be quite similar for most methods, with a noticeable signal in the HLA region on chromosome 6, consistent with the main finding in the previous publication of these data (Fakiola *et al.*, 2013). Obvious exceptions to this are the plots from FBAT analyses, which show no association signal at all, consistent with the above observation. Furthermore, it can be seen from these plots that the results from the other alternative methods also produced much weaker signals than those from the LMM methods.

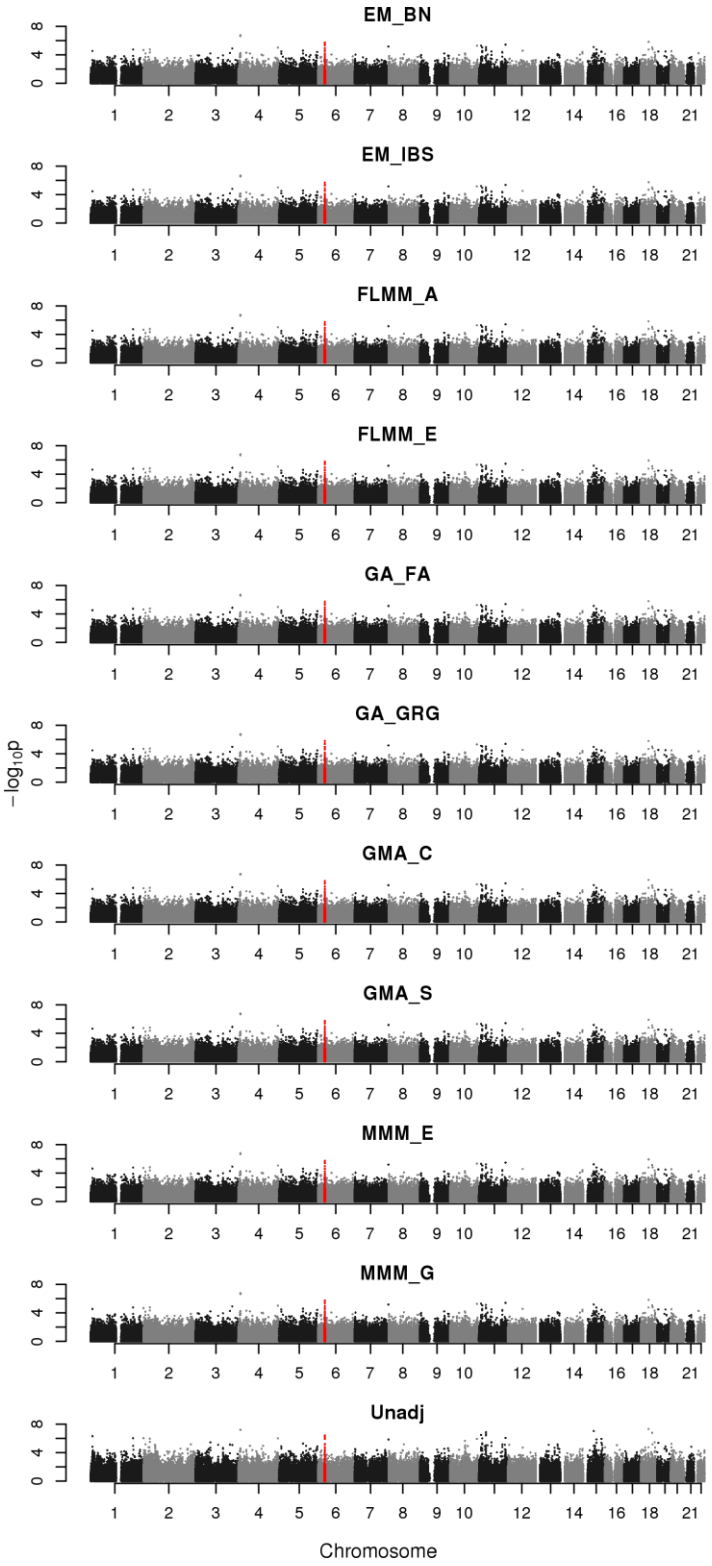

Figure 4.11 Manhattan plots for real VL data set using various LMM methods. The points marked in red (appear as dark grey area near the beginning of chromosome 6 if printed in black and white) denote the confirmed significant region form Fakiola *et al.* (Fakiola *et al.*, 2013). EM_BN = EMMAX (Balding- Nichols), EM_IBS = EMMAX (IBS method), FLMM_A = FaST-LMM using approximate calculation with RRM, FLMM_E = FaST-LMM using exact calculation with RRM, GA_FA = GenABEL (FASTA), GA_GRG = GenABEL (GRAMMAR-Gamma), GMA_C = GEMMA using centred genotypes, GMA_S = GEMMA using standardised genotypes, MMM_E = MMM using full mixed model (exact) calculation, MMM_G = MMM using GLS approximation, Unadj = unadjusted analysis.

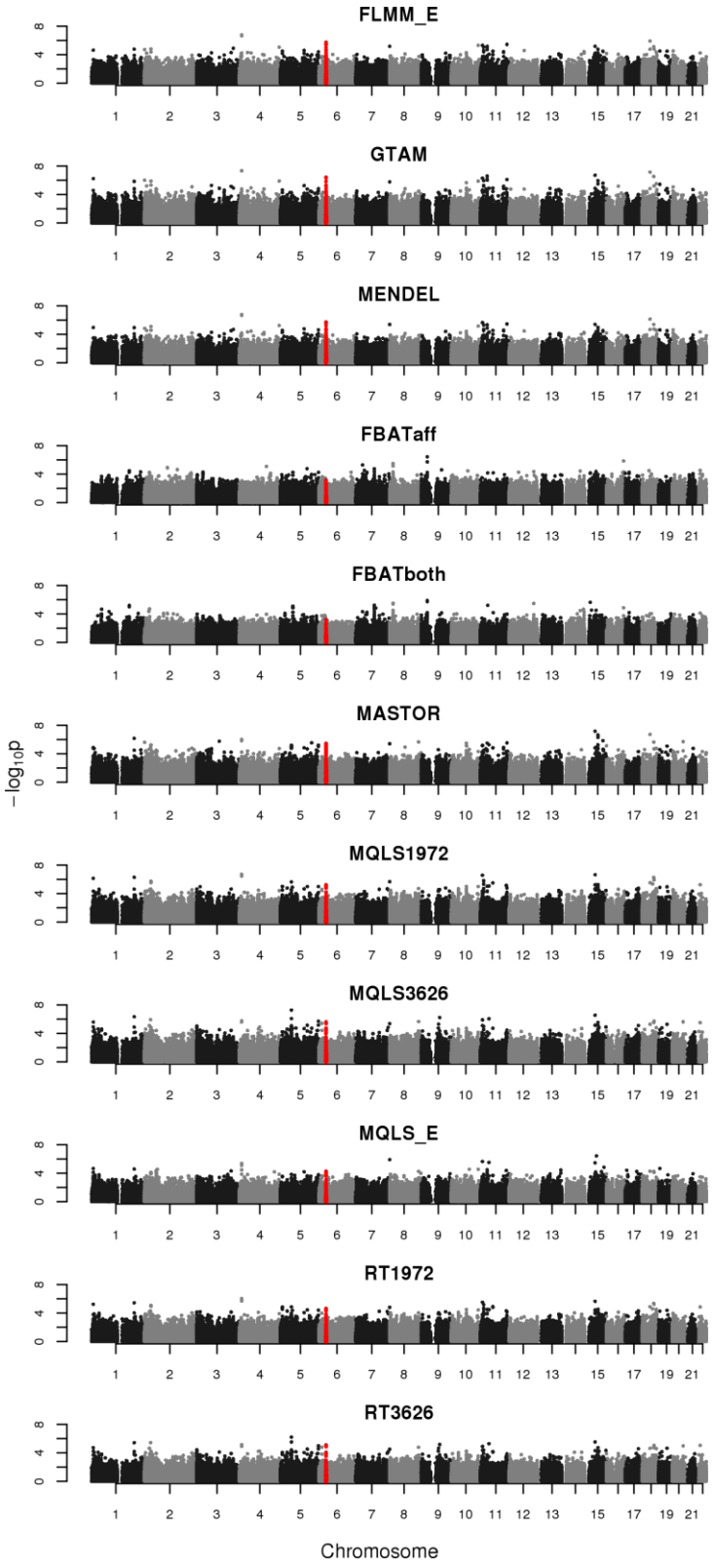

Figure 4.12 Manhattan plots for real VL data set using various LMM/alternative methods. The points marked in red (appear as dark grey area near the beginning of chromosome 6 if printed in black and white) denote the confirmed significant region form Fakiola *et al.* (Fakiola *et al.*, 2013). FLMM_E = FaST-LMM using exact calculation with RRM, FBATaff = FBAT using transmissions to affecteds only, FBATboth = FBAT using transmissions to both affecteds and unaffecteds, MQLS1972/3626 = MQLS using theoretical kinships of either the 1,972 genotyped individuals or all 3,626 individuals in the pedigree, MQLS_E = MQLS using estimated kinships (1,972 individuals), RT1972/3626 = ROADTRIPS using 1,972 or 3,626 individuals.

Although the LMM (and several alternative) approaches seem to show similar overall levels of power, an interesting separate question is the degree of concordance between the different methods with respect to the association signals detected. As can be seen from Figure 4.13 and the top-left part of Figure 4.14, GWAS results from all LMM methods using the pruned set of SNPs to estimate the pairwise kinships were highly concordant.

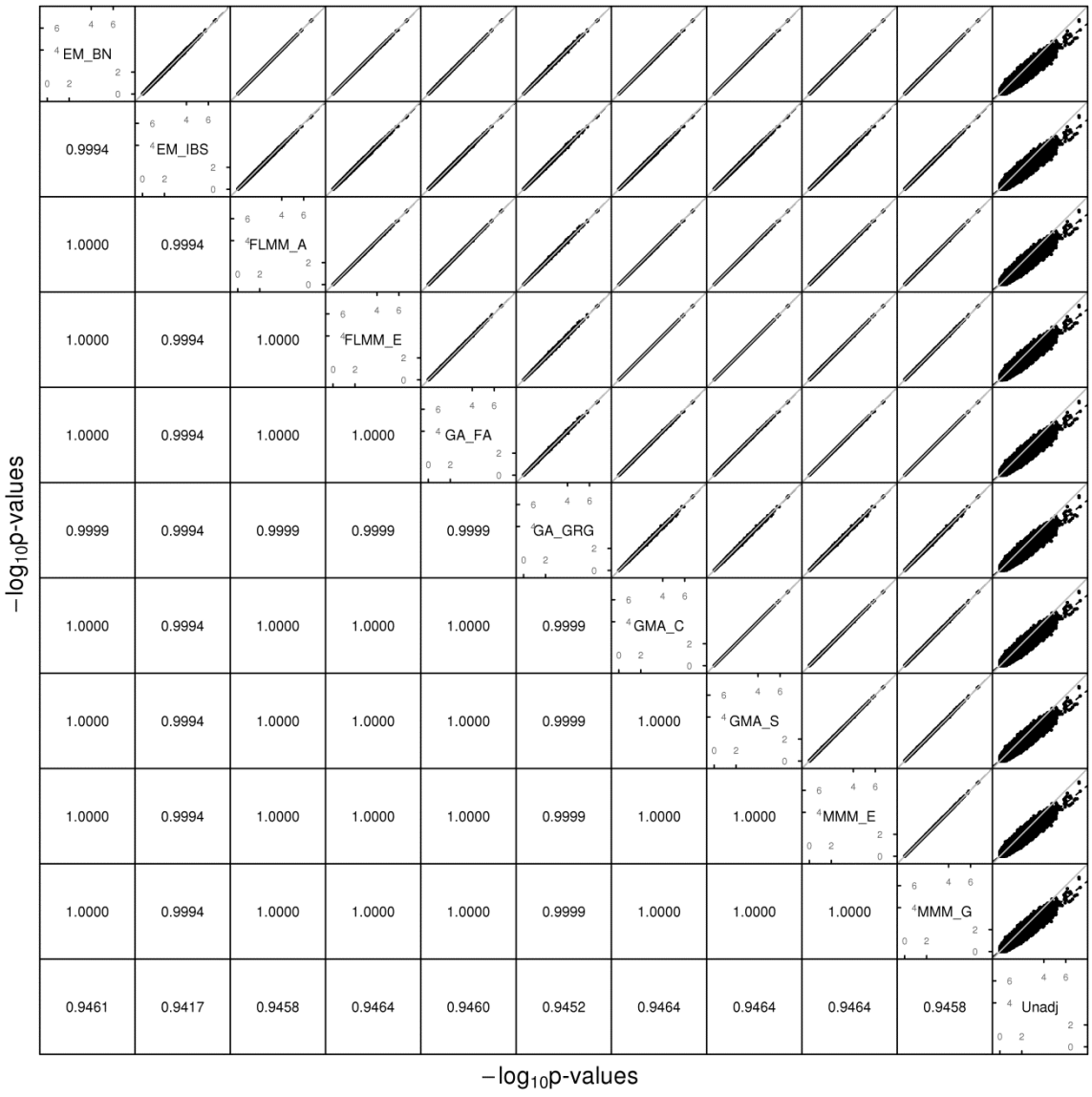

Figure 4.13 Comparison of -$\log_{10}$(p-values) using different LMM software packages, real disease phenotypes, and using pruned set of SNP for adjustment. Plots above the diagonal show a comparison of -$\log_{10}$(p-values), with correlation between the -$\log_{10}$(p-values) indicated below the diagonal. The grey solid lines represent the line of equality; the black dashed lines the linear regression line of the variable on the y asix on the variable on the x axis. EM_BN = EMMAX (Balding- Nichols), EM_IBS = EMMAX (IBS method), FLMM_A = FaST-LMM using approximate calculation, FLMM_E = FaST-LMM using exact calculation, GA_FA = GenABEL (FASTA), GA_GRG = GenABEL (GRAMMAR-Gamma), GMA_C = GEMMA using centred genotypes, GMA_S = GEMMA

using standardised genotypes, MMM_E = MMM using full mixed model (exact) calculation, MMM_G = MMM using GLS approximation, Unadj = unadjusted analysis.

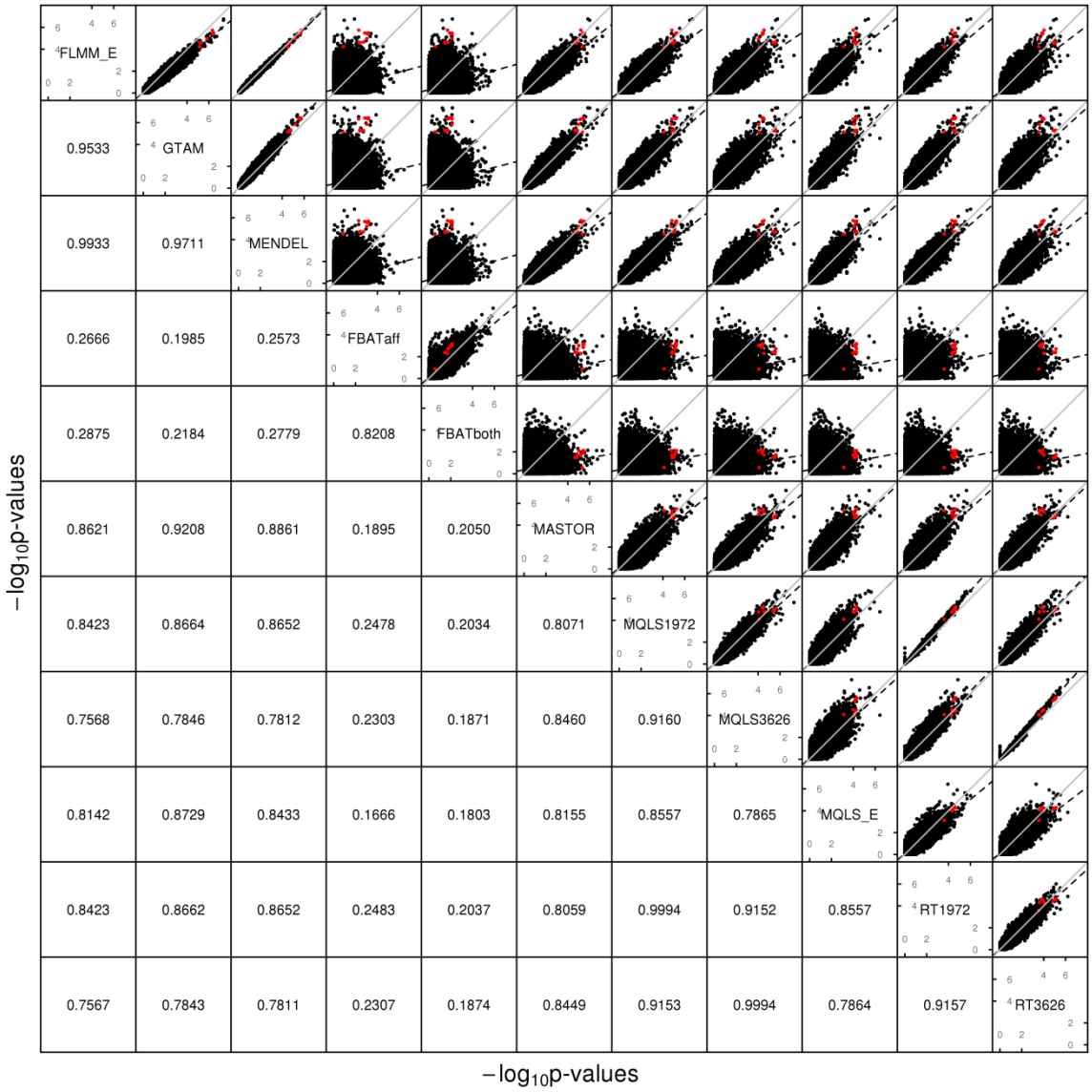

Figure 4.14 Comparison of -log10(p-values) using different LMM/alternative software packages, real disease phenotypes. Plots above the diagonal show a comparison of -$\log_{10}$(p-values), with correlation between the -$\log_{10}$(p-values) indicated below the diagonal. The red dots represent the top 12 SNPs ($p < 10^{-4}$) in the HLA region in chromosome 6. The grey solid lines represent the line of equality; the black dashed lines the linear regression line of the variable on the y asix on the variable on the x axis. FLMM_E = FaST-LMM using exact calculation with RRM, FBATaff = FBAT using transmissions to affecteds only, FBATboth = FBAT using transmissions to both affecteds and unaffecteds, MQLS1972/3626 = MQLS using theoretical kinships of either the 1,972 genotyped individuals or all 3,626 individuals in the pedigree, MQLS_E = MQLS using estimated kinships (1,972 individuals), RT1972/3626 = ROADTRIPS using 1,972 or 3,626 individuals. FaST-LMM is an LMM method and is included here for comparison; GTAM and Mendel are also LMM methods, but included here due to their unique characteristics (see Section 4.1)

GTAM, which itself is also an LMM method, produced slightly less concordant results to the other LMM methods, which probably reflects the fact that theoretical rather than genetically estimated kinship matrix was used. As for the results from alternative methods, most were also concordant with the LMM results (but to a lesser degree than had been seen for methods within the LMM class including GTAM), with the exception of FBAT which showed little concordance to the other methods at the vast majority of (presumably null) SNPs.

Figure 4.14 also shows that methods that use phenotype information from non-genotyped family members (MQLS3626 and RT3626, which use all 3,626 individuals regardless of whether or not they have genotype data) are most similar to each other and less similar to methods that use information only from the genotyped individuals.

4.4. Feeding Externally Estimated Kinship Measures into LMMs

The separation between the 'kinship estimation' and 'association testing' steps in most LMM packages enable the user to read in theoretical or estimated kinships as desired, and to consider using an alternative package for estimating kinships to the one used for the actual association testing. This raises the issue of to what extent this would affect the final outcome, and, perhaps more interestingly, what would happen if the externally estimated kinships are substantially less accurate than the native ones.

In this section, various sets of kinships are read into FaST-LMM for use in GWAS analysis using an exact calculation. These include: FaST-LMM's own realised relationship matrix (RRM), KING using homogeneous population assumption ('KING-homo'), KING using robust estimation ('KING-robust'), PLINK IBD estimation, 'theoretical' pedigree-based relatedness (calculated using KinInbcoef version 1.1 (Bourgain and Zhang, 2009)) and 'wrong' kinship calculated as an inverse of the theoretical kinship i.e. $0.5 - k$ (but with intra-individual 'kinship' set to the correct outbred value of 0.5). The results were also compared with results from simple, unadjusted linear regression in FaST-LMM, as shown in Figure 4.16.

Use of the 'wrong' kinship estimates resulted in very similar results to unadjusted analysis ($\lambda = 1.23$). Results based on kinship estimates from the two KING methods were very similar to those obtained using FaST-LMM's own RRM, and provided good control of the genome-wide error rate ($\lambda \approx 1$) in spite of the unusual pattern in KING's estimated kinship that had been noted in Section 4.2.1. Although still better than the unadjusted analysis, estimation of kinships using PLINK was less satisfactory, leading to inflated genomic control factor of 1.18, which was substantially worse than RRM or KING. Nevertheless, there was high degree of concordance in the association analysis results from all types of kinship matrices, especially between FaST-LMM and the two KING matrices, and between the 'wrong' kinship and unadjusted analysis (Figure 4.16).

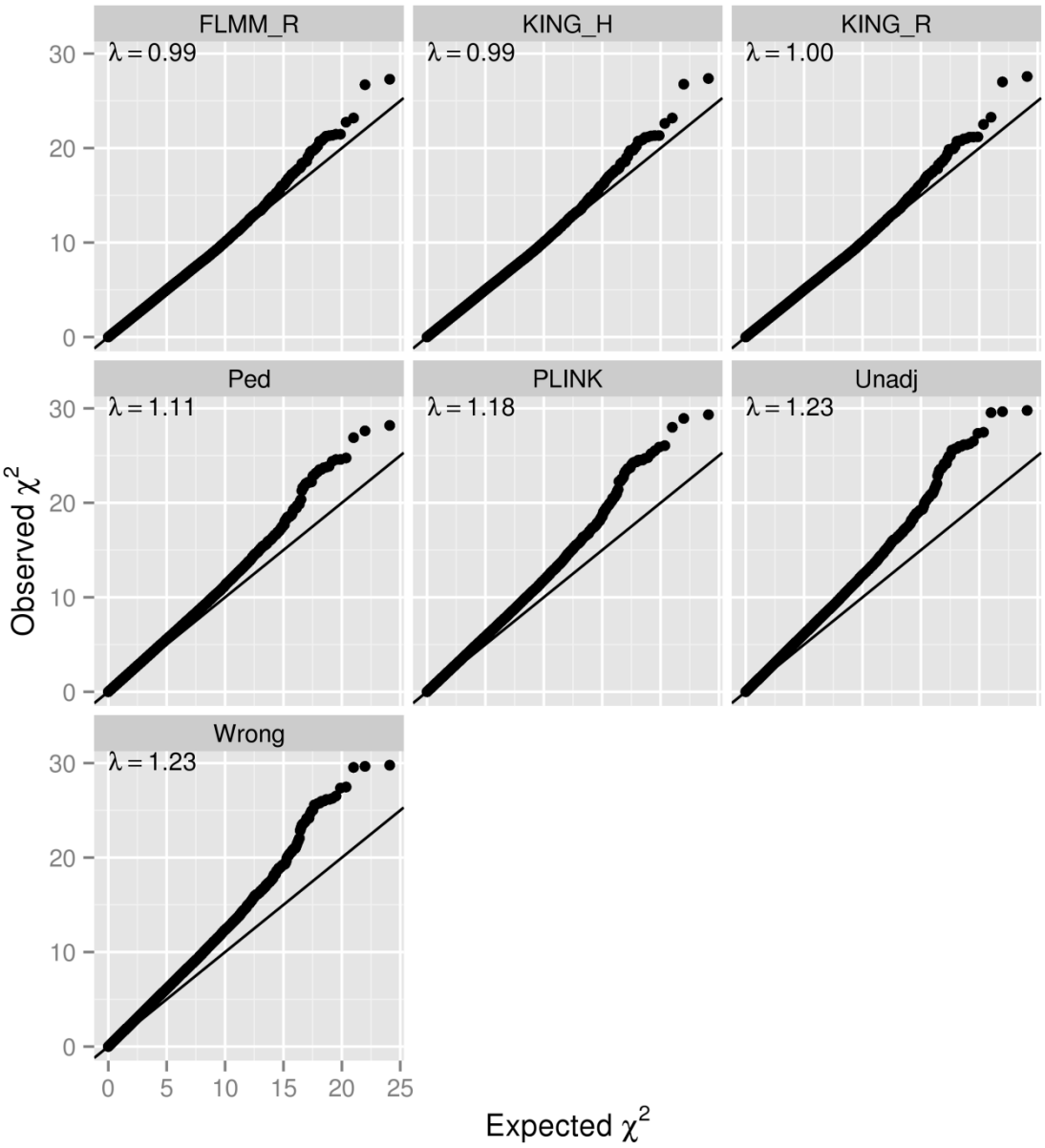

Figure 4.15 Q-Q plots of real VL phenotype GWAS results and genomic inflation factors (λ) obtained from FaST-LMM using alternative kinship estimates. FLMM_R = FaST-LMM's own realised relationship matrix, KING_H = KING homogeneous method, KING_R = KING robust method, Ped = theoretical kinship estimates based on pedigree information, Unadj = unadjusted, Wrong = misspecified kinships, chosen to be inversely related to the true kinship value.

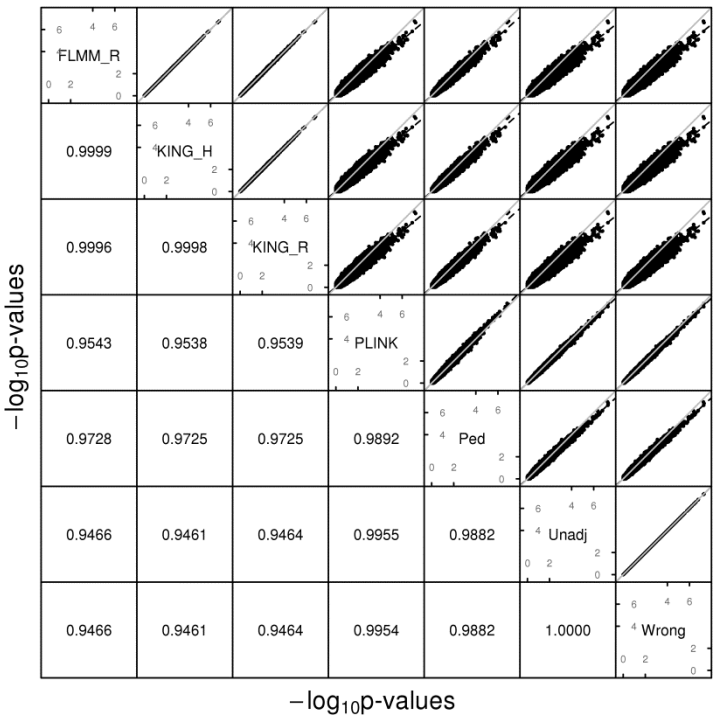

Figure 4.16 Comparison of -$\log_{10}$(p-values) obtained from FaST-LMM using alternative kinship estimates, real disease phenotypes. Plots above the diagonal show a comparison of –$\log_{10}$(p-values), with correlations between the –$\log_{10}$(p-values) indicated below the diagonal. The grey solid lines represent the line of equality; the black dashed lines the linear regression line of the variable on the y axis on the variable on the x axis. FLMM_R = FaST-LMM's own realised relationship matrix, KING_H = KING homogeneous method, KING_R = KING robust method, Ped = theoretical kinship estimates based on pedigree information, Unadj = unadjusted, Wrong = misspecified kinships, chosen to be inversely related to the true kinship value.

4.5. Discussion

This chapter demonstrates that various LMM methods can be used in family-based GWAS of binary trait to control the overall genomic inflation factor while also offering higher power than traditional family-based association analysis approaches such as those implemented in FBAT. Similar inference is also provided by related and alternative approaches implemented in the software packages Mendel, ROADTRIPS, MQLS and MASTOR. The inferior power in FBAT is likely to be caused by the smaller effective sample size (357 cases and 357 'pseudo' controls in FBAT, versus 357 cases and 1613 genuine control in the other approaches) due to the way the FBAT test statistics are constructed.

All LMM GWAS methods considered here as well as the alternative methods such as MQLS, ROADTRIPS, MASTOR and GTAM model the relatedness between individuals based on one or more kinship matrices, constructed either on the basis of known (hypothesised) pedigree relationships between individuals, or through estimating kinships on the basis of genome-wide SNP data (or their subset). Most methods allow

separation between the kinship matrix estimation and the analysis step. This is convenient for several reasons. Firstly, it allows the set of SNPs used for estimating the kinship matrix to be different from that used for genome-wide association testing. It also means that kinships estimated using one package can potentially be used in another package at the analysis stage, if desired. Furthermore, this allows better parallelisation as the kinship matrix needs to be calculated only once and subsequently (or concurrently) read into multiple association analysis tasks, without each having to calculate its own kinship matrix.

The ability to use a different set of SNPs to estimate the kinship matrix can improve the performance of a method in the situation where the set of SNPs being tested for association is not suitable for kinship estimation. An obvious example would be when the data are from a very dense GWAS chip, calculating the kinship matrix based on the whole data set would require much more computational time than would otherwise be required if a smaller set of SNPs can be used, provided that the use of estimations based on the smaller set does not cause significant deterioration of the results. It is also useful in the reverse situation in which only a small subset of SNPs needs to be analysed. If the kinship is also calculated based on this subset of SNPs, this may not be very accurate and could result in higher inflation of results than would otherwise be achievable (analogous to the situation where the thinned set of SNPs was used, Figures 4.6-4.7).

As there was not much difference in performance between kinships estimated from the pruned and the full SNP sets, the choice of one or the other may depend on the data already on hand. It should also be noted that the time required to estimate kinships from a pruned SNP set is significantly shorter than from a full SNP set.

The significant performance deterioration when the thinned set of SNPs was used could be because there were too few SNPs in the thinned set to accurately model the relationships within the data set. Although 1,900 SNPs may be sufficient to accurately model close relationships such as full sib or parent-offspring, many more SNPs will be required to accurately model distant relationships within pedigrees (such as cousins, second cousins, third cousins etc) or even more distant relationships between pedigrees.

The inflation of results obtained using theoretical kinships suggests the presence of additional relatedness/population structure in these data that is not well accounted for by known family relationships. In this situation, genetically estimated kinships can be expected to perform better than theoretical kinships.

Traditional methods for family-based association analysis make use of pedigree relationships, either (e.g. FBAT) through direct use of known pedigree structure or else (e.g. MQLS, ROADTRIPS and all LMM methods) through use of a covariance matrix that involves the known kinship between each pair of individuals (which is the probability that a randomly chosen allele at a randomly chosen allele at a locus in each individual is identical by descent). The assumption that all founders in a pedigree share no alleles identical by descent is clearly unrealistic, given human population history, while the assumption that all pedigrees are correctly specified and unrelated to one another is also likely to be violated in most real studies. The use of estimated kinships based on SNP data rather than theoretical kinships based on known pedigree relationships removes the reliance on these untenable assumptions, and allows essentially the same analysis approaches to be applied to apparently unrelated individuals, who may nevertheless display distant levels of shared ancestry.

A key point when using estimated kinships to structure the covariance matrix in an association analysis is that the goal is not relationship estimation (whether close or distant) in its own right, but rather to adjust the analysis for phenotypic correlations between individuals due to genetic factors—usually assumed to be polygenic effects—that would otherwise result in inflated association test statistics. Therefore, the extent to which the estimated kinship measures reflect the genuine relationships between individuals is arguably irrelevant. The important issue here is whether or not the use of such kinships succeeds with respect to adequately modelling phenotypic correlations between individuals. On that note, in the analyses performed here there was no large difference between the results obtained using different kinship measures, although use of the kinship measures output by PLINK (as well as use of completely incorrect kinship measures) did perform worse than the other kinship measures investigated.

Although FaST-LMM-Select has been reported to show some advantage over using all SNPs when applied to simulations that included population stratification (but not familial relatedness) of quantitative phenotypes in randomly ascertained individuals (Lippert *et al.*, 2013), application of this procedure to this highly ascertained set of Brazilian pedigrees resulted in substantially worse inflation than the simpler method of using all pruned or full SNPs for estimating pairwise kinships. This may be because the procedure tends to identify only a very small subset of SNPs, which may be adequate for capturing population stratification (as this should require relatively few principal components, which could be adequately approximated using these SNPs), but may not be sufficient to model family relatedness.

Regardless of the method or SNP set used, adjustment always resulted in substantially lower inflation than was seen in unadjusted analysis. At worst, the adjusted results

would still be comparable to unadjusted analysis, as seen when the kinships were incorrectly estimated. So on the grounds of accuracy alone, there is little rationale for unadjusted analysis (however, in practice, the increase in complexity of the adjusted analysis will have to be taken into account).

Using kinships estimated from a different package may be beneficial if, for example, the estimation from one package is substantially superior (or inferior) to another, or if the calculation time is substantially shorter or longer in one package compared to another. The former is unlikely to be the case (at least in practice) as has been demonstrated in Sections 4.2.1 and 4.2.3; the latter point will be investigated in Section 5.5.

Although their precise algorithms vary (Aulchenko *et al.*, 2007b; Rakovski and Stram, 2009; Kang *et al.*, 2010; Lippert *et al.*, 2011; Zhou and Stephens, 2012; Pirinen *et al.*, 2013), the kinship measures calculated from the various LMM methods tended to be highly correlated, the main difference being the scale on which they are measured. This should not be too important for the purpose of LMM analysis, as the kinship measures are used within the LMM framework to structure the variance/covariance matrix of the genetic random effect. Any rescaling of this will be compensated for by a similar rescaling of the estimated genetic variance parameter σ_g^2 (see Section 1.2).

Although kinship estimates from both KING methods tended to differ from most other methods, this does not seem to affect the final results much. This could be because adequate control could genuinely be achieved with these estimates, but might also be because of the way FaST-LMM was designed to handle non-postive semidefnite covariance matrices which may have eliminated the differences. Regardless of the reason, the implication seems to be that KING's kinship estimates can be used successfully in LMM GWAS analysis—at least with software that can handle non-positive semidefinite matrix.

The same cannot be said for the kinship estimates from PLINK, which resulted in substantial inflation. This is consistent with previous results (Manichaikul *et al.*, 2010) suggesting that PLINK performs less well than KING for relationship estimation. Interestingly, although KING-robust has been shown to have an advantage over KING-homo in non-homogeneous populations when the goal is relationship estimation for its own sake (Manichaikul *et al.*, 2010), this advantage is not apparent here, where the goal is instead to adjust for potentially different levels of relatedness, from close family relationships to more distant relationships—perhaps mimicking population membership—while performing association testing.

One caveat in interpreting the results in this chapter is that there is no guarantee that the HLA region signal detected by all but one method is genuine: it is possible that this

signal was false, in which case FBAT would be the only method that gives correct results. However, given the strength of the signal, the consistency across most methods, the biological plausibility and the independent confirmation in a different (Indian) data set (Fakiola *et al.*, 2013), it is very likely that the observed signal is genuine. The next chapter will take a different approach and investigate the performance of these methods when the SNP effects are known through the use of simulations.

Chapter 5. Application of Genomic IBD Estimates to Account for Relatedness in Genome-Wide Association Analyses of the Simulated Brazilian Visceral Leishmaniasis Data Set

This chapter continues to investigate the performance of various LMM GWAS methods as applied to the family-based VL data set. However, instead of analysing the real binary phenotype, various types of simulated phenotypes (as described in Section 2.4) are used. This allows investigation of application of the LMM GWAS methods to different types of phenotypes, including longitudinal measurements, while also ensuring that the true effect locations are known. Furthermore, this also allows assessment of power and type I error which would not be possible with the real phenotype.

5.1. Performance with Simulated Strong Qualitative Phenotype

With the simulated strong qualitative phenotype, all LMM and alternative methods performed well in terms of controlling the genomic inflation, with $\lambda = 0.99$ for all LMM methods that use genetically estimated kinships, and ranging from 0.99 to 1.01 for other methods, compared with $\lambda = 1.12$ in unadjusted analysis (Figures 5.1-5.3). Nevertheless, it is quite clear from these Q-Q plots that, although successful in controlling the inflation, FBAT could not detect the effect of the two simulated SNPs. This is confirmed in the Manhattan plots (Figures 5.4-5.5), in which all methods but FBAT gave clear, strong signals at the simulated strong effect loci.

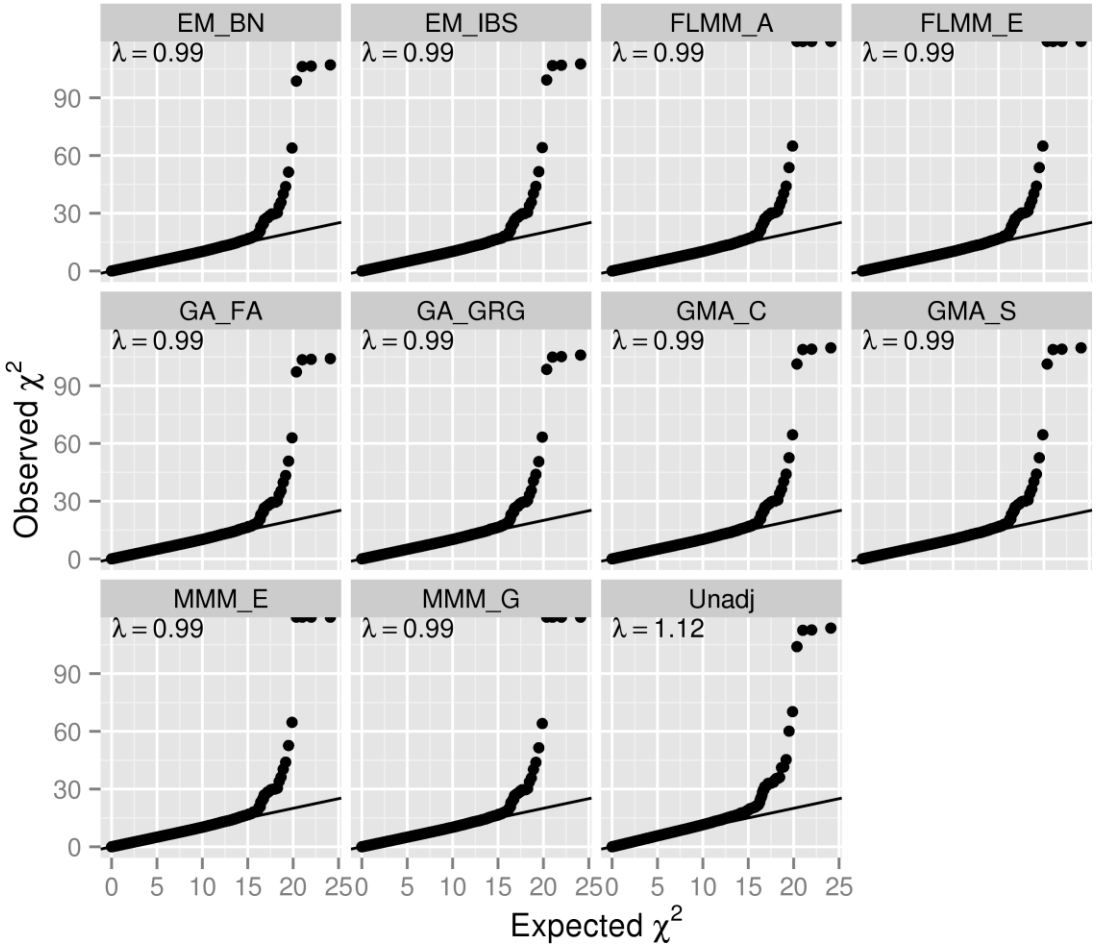

Figure 5.1 Q-Q plots of simulated strong qualitative phenotype GWAS results and genomic inflation factors (λ) for different LMM methods. EM_BN = EMMAX (Balding- Nichols), EM_IBS = EMMAX (IBS method), FLMM_A = FaST-LMM using approximate calculation with RRM, FLMM_E = FaST-LMM using exact calculation with RRM, GA_FA = GenABEL (FASTA), GA_GRG = GenABEL (GRAMMAR-Gamma), GMA_C = GEMMA using centred genotypes, GMA_S = GEMMA using standardised genotypes, MMM_E = MMM using full mixed model (exact) calculation, MMM_G = MMM using GLS approximation, Unadj = unadjusted analysis. The dots at the upper border of the panels (FLMM and MMM) represent the SNPs where the equivalent χ^2 values are ∞ (i.e. p-value = 0).

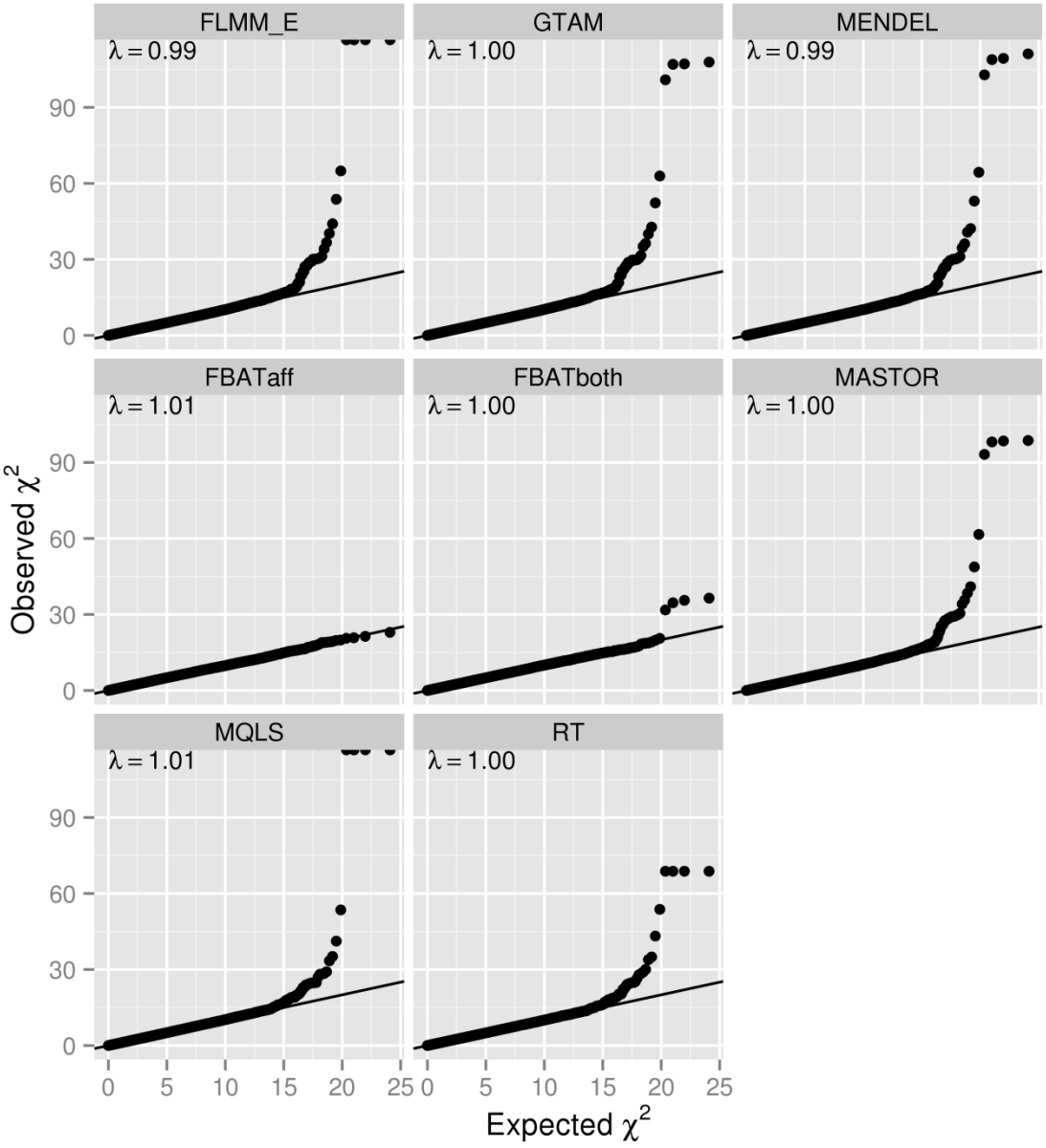

Figure 5.2 Q-Q plots of simulated strong qualitative phenotype GWAS results and genomic inflation factors (λ) for different LMM/alternative methods. FLMM_E = FaST-LMM using exact calculation with RRM, FBATaff = FBAT using transmissions to affecteds only, FBATboth = FBAT using transmissions to both affecteds and unaffecteds, MQLS = MQLS using theoretical kinships of the 1,972 genotyped individuals, RT = ROADTRIPS using 1,972 individuals. FaST-LMM is an LMM method and is included here for comparison; GTAM and Mendel are also LMM methods, but included here due to their unique characteristics. The dots at the upper border of the panels (FLMM_E and MQLS) represent the SNPs where the equivalent χ^2 values are ∞ (i.e. p-value = 0).

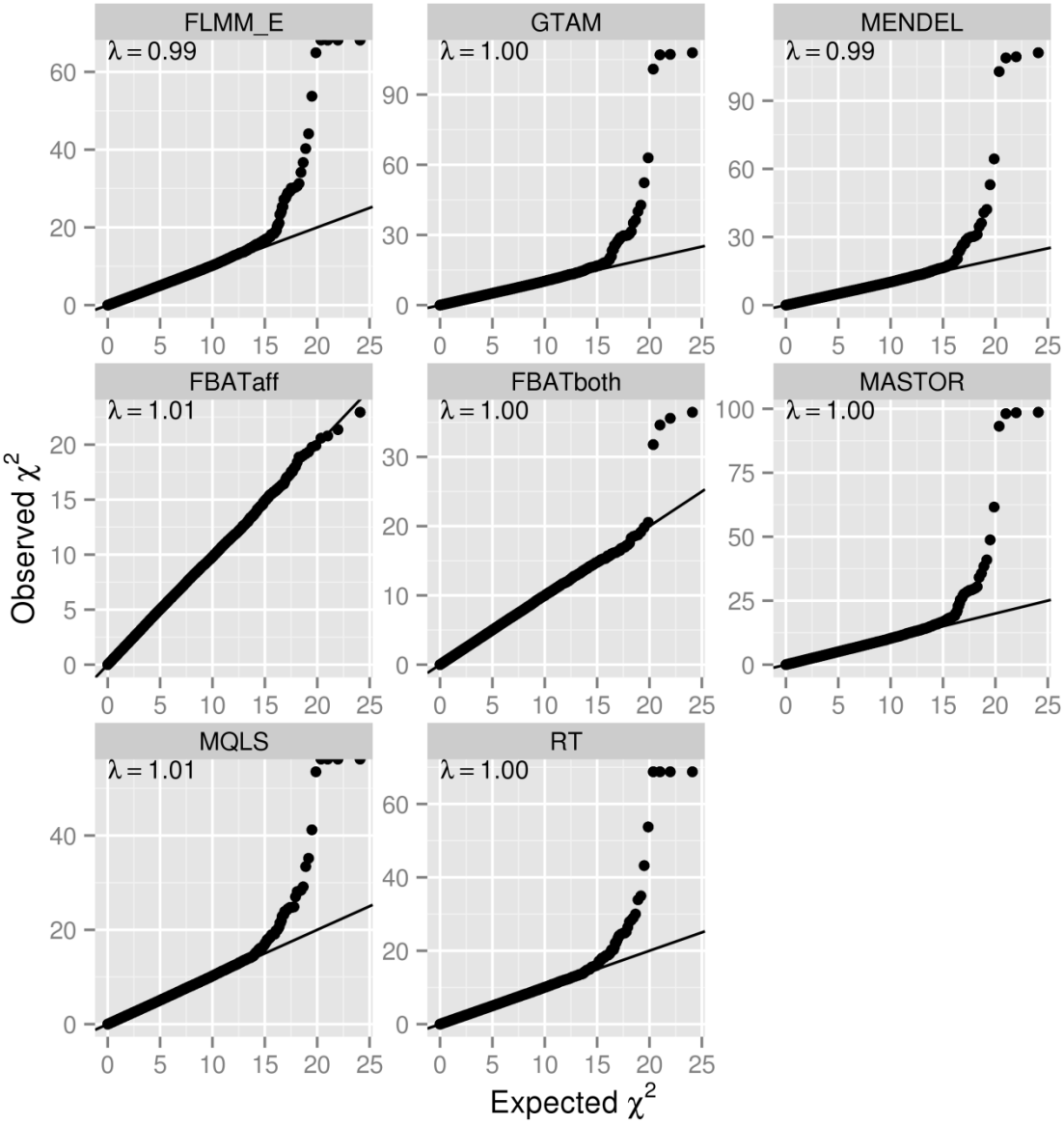

Figure 5.3 Q-Q plots of simulated strong qualitative phenotype GWAS results and genomic inflation factors (λ) for different LMM/alternative methods, with each panel plotted on its own scale. FLMM_E = FaST-LMM using exact calculation with RRM, FBATaff = FBAT using transmissions to affecteds only, FBATboth = FBAT using transmissions to both affecteds and unaffecteds, MQLS = MQLS using theoretical kinships of the 1,972 genotyped individuals, RT = ROADTRIPS using 1,972 individuals. FaST-LMM is an LMM method and is included here for comparison; GTAM and Mendel are also LMM methods, but included here due to their unique characteristics. The dots at the upper border of the panels (FLMM_E and MQLS) represent the SNPs where the equivalent χ^2 values are ∞ (i.e. p-value = 0). Unlike the previous plot, each panel in this plot has its own y-axis scale to better depict the distribution within its own panel.

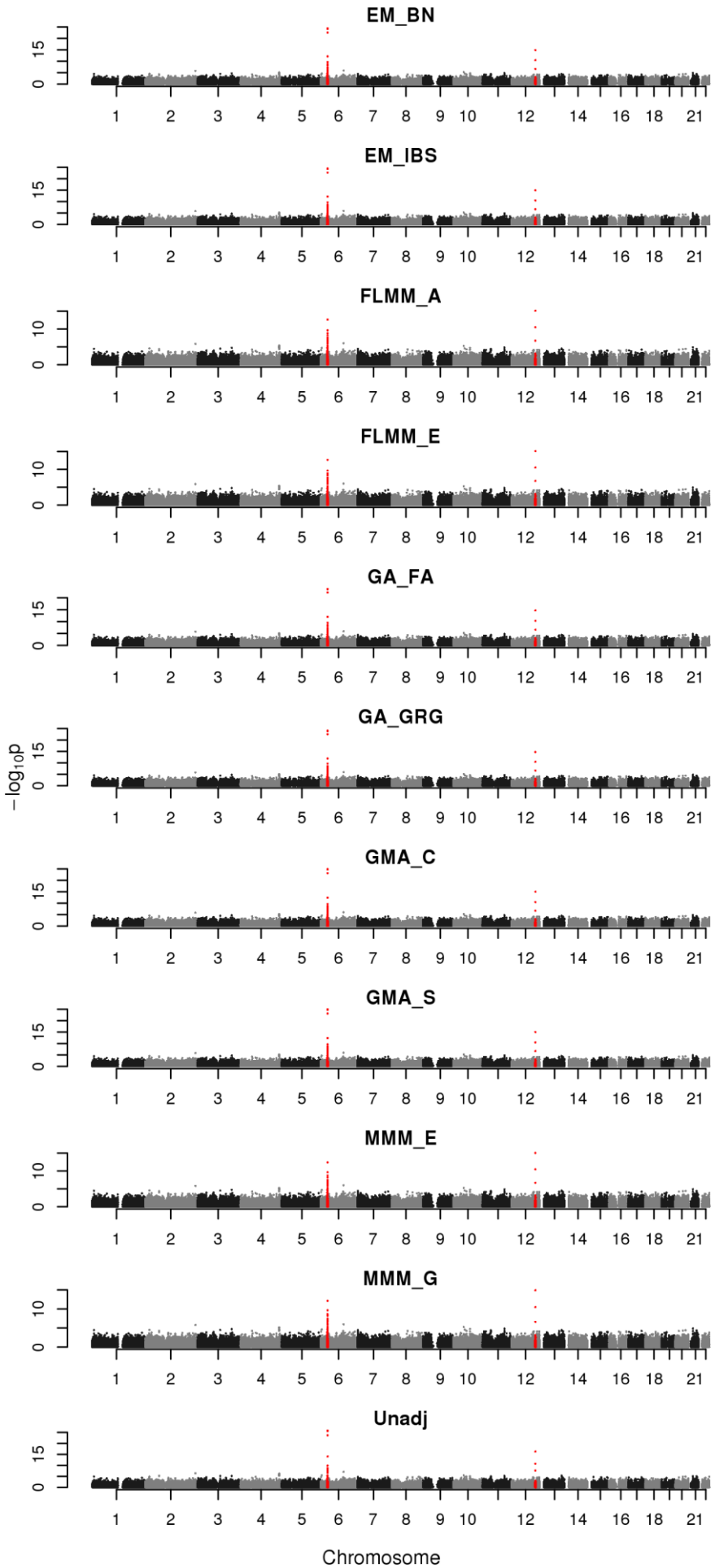
EM_BN
EM_IBS
FLMM_A
FLMM_E
GA_FA
GA_GRG
GMA_C
GMA_S
MMM_E
MMM_G
Unadj
−log10p
Chromosome
1 2 3 4 5 6 7 8 9 10 12 14 16 18 21

Figure 5.4 Manhattan plots for VL data set with simulated strong qualitative phenotype using various LMM methods. The points marked in red (appear as dark grey area near the beginning of chromosome 6 and the end of chromosome 12 if printed in black and white) denote the simulated strong effect loci. EM_BN = EMMAX (Balding- Nichols), EM_IBS = EMMAX (IBS method), FLMM_A = FaST-LMM using approximate calculation with RRM, FLMM_E = FaST-LMM using exact calculation with RRM, GA_FA = GenABEL (FASTA), GA_GRG = GenABEL (GRAMMAR-Gamma), GMA_C = GEMMA using centred genotypes, GMA_S = GEMMA using standardised genotypes, MMM_E = MMM using full mixed model (exact) calculation, MMM_G = MMM using GLS approximation, Unadj = unadjusted analysis.

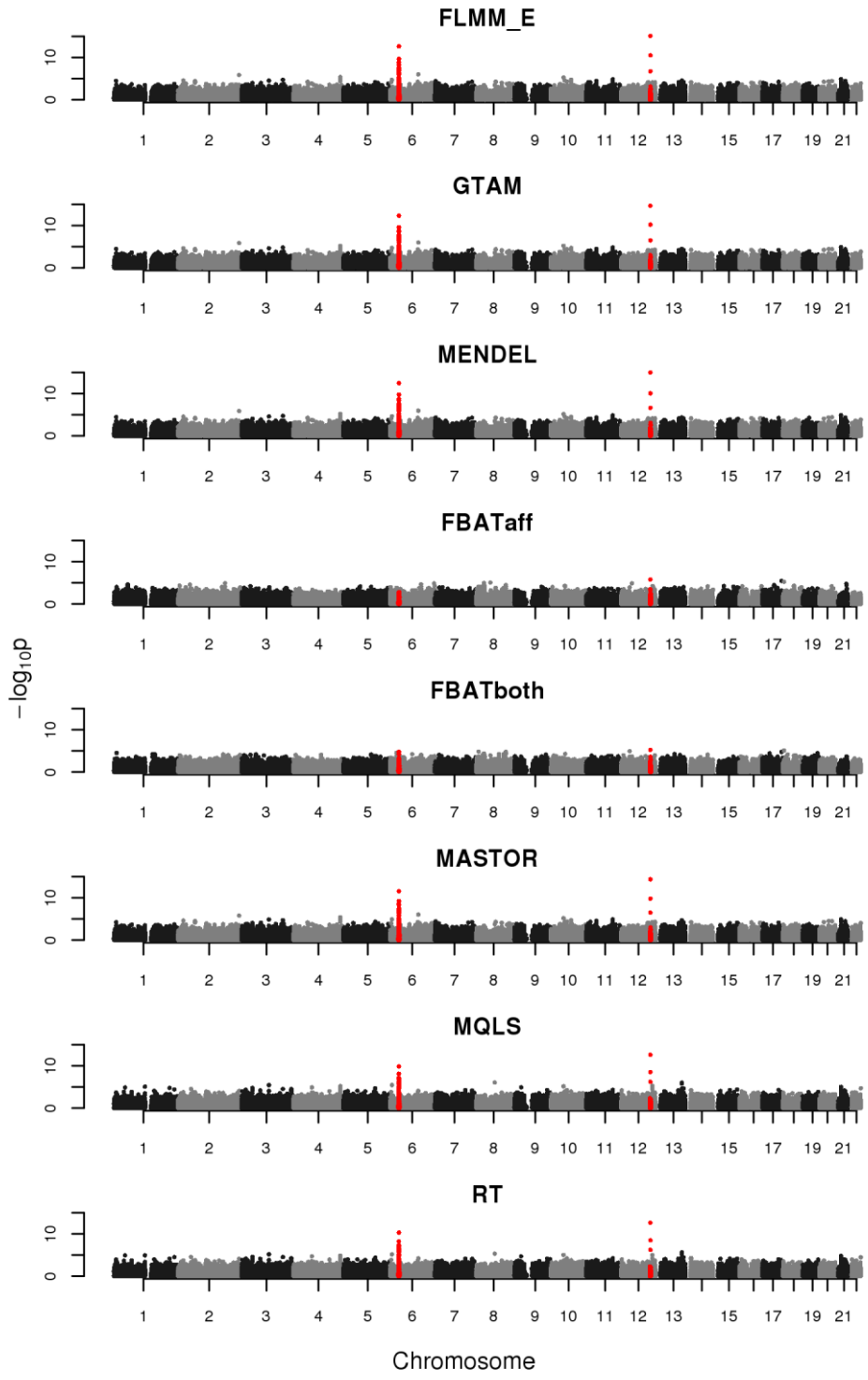

Figure 5.5 Manhattan plots for VL data set with simulated strong qualitative phenotype using various LMM/alternative methods. The points marked in red (appear as dark grey area near the beginning of chromosome 6 and the end of chromosome 12 if printed in black and white) denote the simulated strong effect loci. FLMM_E = FaST-LMM using exact calculation with RRM, FBATaff = FBAT using transmissions to affecteds only, FBATboth = FBAT using transmissions to both affecteds and unaffecteds, MQLS = MQLS using theoretical kinships of the 1,972 genotyped individuals, RT = ROADTRIPS using 1,972 individuals. FaST-LMM is an LMM method and is included here for comparison; GTAM and Mendel are also LMM methods, but included here due to their unique characteristics.

To formally compare the power and type I error for the different analysis methods, 1,000 simulation replicates as described in Section 2.4.4 were used. Technically, for each of these replicates, adjusted association analyses were conducted for each SNP from a limited set consisting of two groups of SNPs: the 'effect' SNPs, defined as any SNP within 40 kb from either of the two simulated disease loci; and the 'null' SNPs, defined as the 100th SNP *from the q-terminal* (that is, the 100th SNP before the last SNP available in each chromosome) of the remaining 20 chromosomes without the main effect SNPs. Power was defined as the proportion of replicates in which both simulated loci are detected, that is, at least one SNP within 40 kb of each simulated disease locus reaches the specified p-value threshold. Type I error rate was calculated in a simpler way—by just pooling the 20 null SNP results from all replicates (so this became a 20,000 SNPs sample) and calculating the proportion of these null SNPs which had achieved significance at the specified level.

The results of this power and type I error analysis is shown in Figure 5.6. All methods apart from an unadjusted analysis show acceptable levels of type 1 error—although that for FBAT appears to be slightly conservative. In terms of power, all LMM approaches—including GTAM and Mendel—and MASTOR show similar performance. ROADTRIPS and MQLS show slightly lower power than the LMM approaches, while the approaches implemented in FBAT appear to be considerably less powerful than those implemented in the LMM and other packages (even allowing for FBAT's slightly conservative levels of type I error). This appears to be in line with the findings from the previous chapter as well as the comparison shown in Figure 5.5.

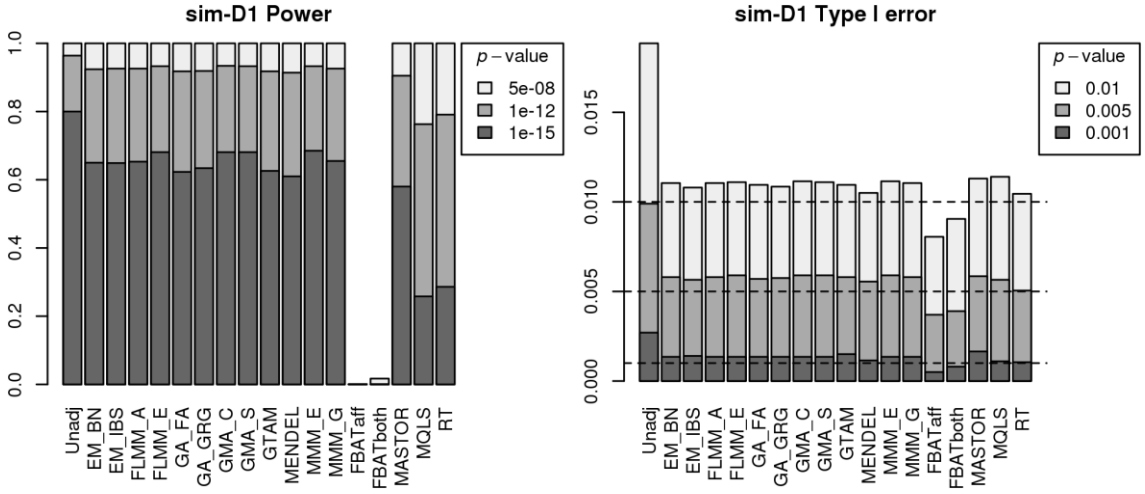

Figure 5.6 Power and type 1 error of different methods when applied to strong binary (disease) phenotype. Powers (left hand plot) are defined as the proportion of replicates (out of 1,000) in which both simulated disease loci are detected, with 'detection' corresponding to any SNP within 40 kb of the simulated disease locus reaching the specified p-value threshold. Type 1 errors (right hand plot) are defined as the proportion of null SNPs (out of 20,000 = 20 null SNPs times 1,000

simulation replicates) that reach the specified p-value threshold. Horizontal dashed lines indicate the target p-value thresholds (i.e. the expected type 1 error rates).

Similar to the findings from the analysis of the real VL phenotype (Section 4.3), all LMM methods and also MASTOR gave very concordant results (Figures 5.7-5.8). Interestingly, the concordance between GTAM (which used pedigree information) and other LMM software results was better than that seen in the real phenotype analysis. The results from FBAT and other alternative methods also seem to be more concordant with the LMM analyses, although still to a lesser extent than the concordance within the LMM class itself.

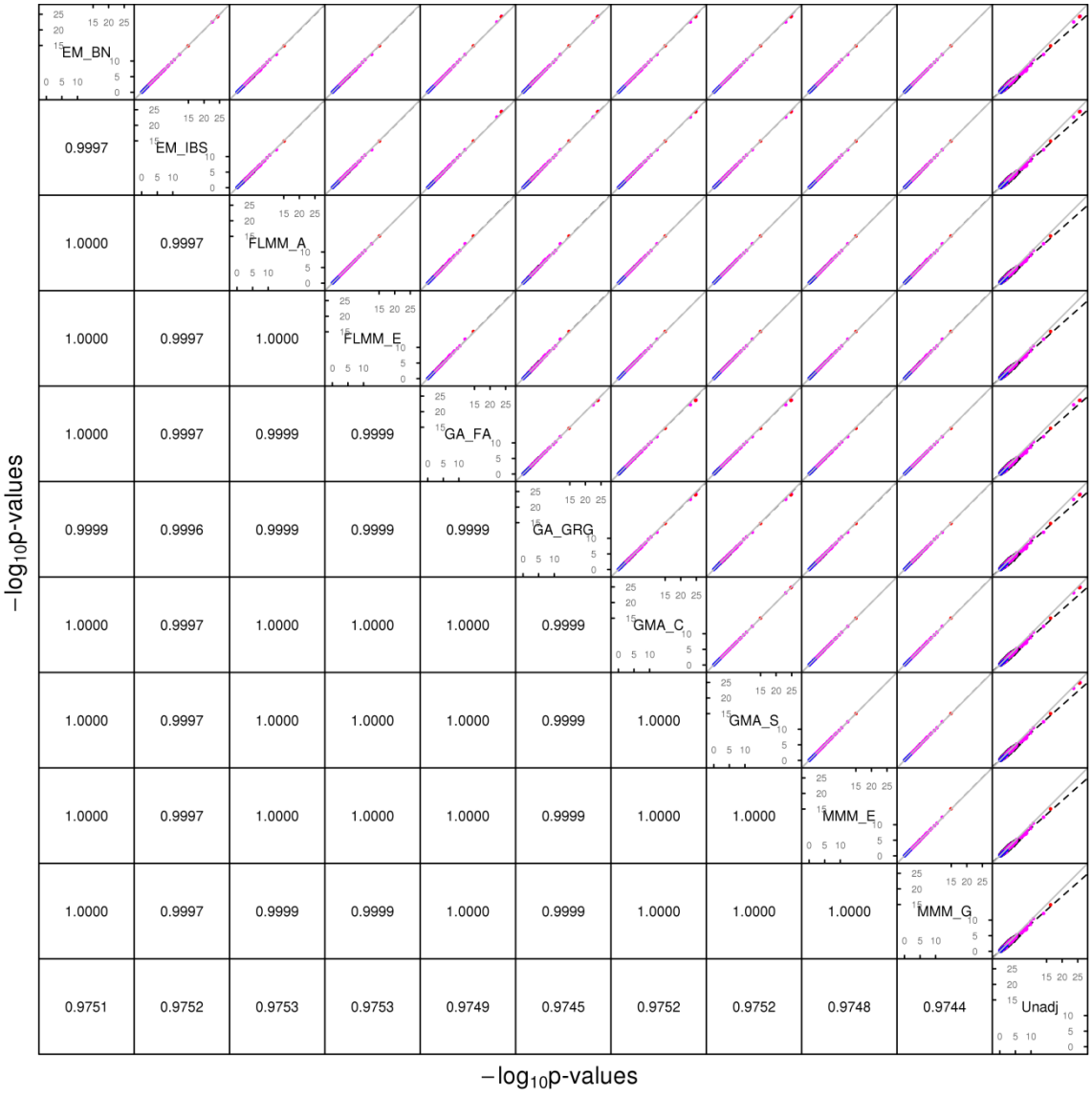

Figure 5.7 Comparison of -log(p-values) using various LMM software packages, simulated strong binary (disease) phenotype. Plots above the diagonal show a comparison of $-\log_{10}$(p-values), with correlations between the $-\log_{10}$(p-values) indicated below the diagonal. The grey solid lines represent the line of equality; the black dashed lines the linear regression line of the variable on

the y axis on the variable on the x axis. The colours denote: red = the two strong effect SNPs, magenta = SNPs within 2 Mb of the strong effect SNPs, blue = 22 polygenic SNPs, green = SNPs within 2 Mb of the polygenic SNPs, black = all other SNPs. Because the black/green/blue SNPs were plotted before the magenta/red SNPs, they may be obscured by the latter. EM_BN = EMMAX (Balding- Nichols), EM_IBS = EMMAX (IBS method), FLMM_A = FaST-LMM using approximate calculation with RRM, FLMM_E = FaST-LMM using exact calculation with RRM, GA_FA = GenABEL (FASTA), GA_GRG = GenABEL (GRAMMAR-Gamma), GMA_C = GEMMA using centred genotypes, GMA_S = GEMMA using standardised genotypes, MMM_E = MMM using full mixed model (exact) calculation, MMM_G = MMM using GLS approximation, Unadj = unadjusted analysis.

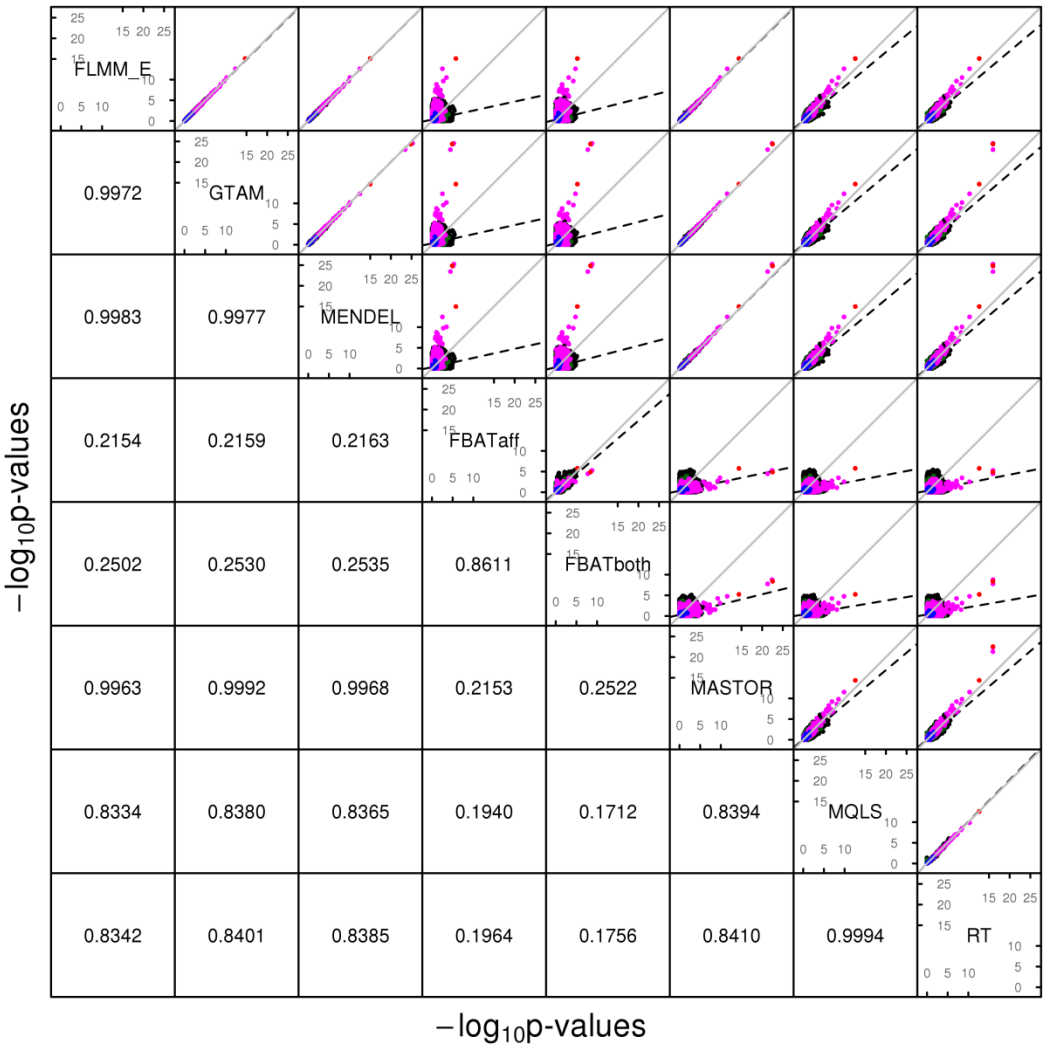

Figure 5.8 Comparison of -log(p-values) using various LMM software packages, simulated strong binary (disease) phenotype. Plots above the diagonal show a comparison of $-\log_{10}$(p-values), with correlations between the $-\log_{10}$(p-values) indicated below the diagonal. The grey solid lines represent the line of equality; the black dashed lines the linear regression line of the variable on the y axis on the variable on the x axis. The colours denote: red = the two strong effect SNPs, magenta = SNPs within 2 Mb of the strong effect SNPs, blue = 22 polygenic SNPs, green = SNPs within 2 Mb of the polygenic SNPs, black = all other SNPs. Because the black/green/blue SNPs were plotted before the magenta/red SNPs, they may be obscured by the latter. FLMM_E = FaST-LMM using exact calculation with RRM, FBATaff = FBAT using transmissions to affecteds only, FBATboth = FBAT using transmissions to both affecteds and unaffecteds, MQLS = MQLS using theoretical kinships of the 1,972 genotyped individuals, RT = ROADTRIPS using 1,972 individuals. FaST-LMM is

an LMM method and is included here for comparison; GTAM and Mendel are also LMM methods, but included here due to their unique characteristics.

A formal comparison of the concordance between 'top hits' identified by the different methods in the simulated data (1,000 simulation replicates, comparison restricted to true and null simulated regions) is shown in Table 5.1. Using EMMAX (Balding-Nichols) as reference (the choice of reference is quite arbitrary here, as there is no method that is innately a gold standard), the concordance between the top SNPs identified is seen to be extremely high for all methods except FBAT, suggesting again that all methods except FBAT provide essentially the same inference.

	Mean (standard deviation) in 1,000 replicates of proportion of top *t* SNPs within null and true regions that overlap with top *t* SNPs from EM_BN				
method	*t* = 5	*t* = 10	*t* = 15	*t* = 20	*t* = 25
Unadjusted	0.991 (0.042)	0.990 (0.030)	0.981 (0.033)	0.975 (0.032)	0.973 (0.027)
EM_IBS	0.999 (0.017)	0.999 (0.009)	0.997 (0.015)	0.996 (0.013)	0.996 (0.012)
FLMM_A	1.000 (0.009)	1.000 (0.003)	1.000 (0.007)	1.000 (0.004)	1.000 (0.003)
FLMM_E	0.998 (0.021)	1.000 (0.005)	0.999 (0.008)	0.999 (0.005)	1.000 (0.004)
GA_FA	0.998 (0.018)	1.000 (0.005)	0.999 (0.011)	0.999 (0.008)	0.998 (0.008)
GA_GRG	0.998 (0.021)	0.999 (0.011)	0.996 (0.017)	0.998 (0.010)	0.998 (0.008)
GMA_C	0.998 (0.021)	1.000 (0.004)	0.999 (0.009)	0.999 (0.005)	1.000 (0.004)
GMA_S	0.998 (0.021)	1.000 (0.005)	0.999 (0.008)	0.999 (0.005)	1.000 (0.004)
GTAM	0.998 (0.022)	0.995 (0.022)	0.990 (0.025)	0.988 (0.022)	0.987 (0.020)
MENDEL	0.997 (0.025)	0.996 (0.019)	0.991 (0.024)	0.989 (0.021)	0.989 (0.018)
MMM_E	0.991 (0.041)	1.000 (0.004)	0.999 (0.009)	0.999 (0.005)	1.000 (0.004)
MMM_G	0.993 (0.036)	1.000 (0.003)	1.000 (0.007)	1.000 (0.005)	0.999 (0.005)
FBATaff	0.684 (0.253)	0.790 (0.115)	0.773 (0.090)	0.771 (0.080)	0.760 (0.072)
FBATboth	0.859 (0.130)	0.844 (0.084)	0.811 (0.078)	0.795 (0.075)	0.777 (0.071)
MASTOR	0.993 (0.038)	0.994 (0.024)	0.989 (0.027)	0.985 (0.024)	0.985 (0.022)
MQLS	0.978 (0.062)	0.981 (0.040)	0.960 (0.043)	0.951 (0.041)	0.941 (0.038)
RT	0.981 (0.059)	0.984 (0.037)	0.962 (0.042)	0.952 (0.041)	0.942 (0.038)

Table 5.1 Concordance between top SNPs identified by different methods analysing simulated strong binary (disease) phenotype. EM_BN = EMMAX (Balding- Nichols), EM_IBS = EMMAX (IBS method), FLMM_A = FaST-LMM using approximate calculation with RRM, FLMM_E = FaST-LMM using exact calculation with RRM, GA_FA = GenABEL (FASTA), GA_GRG = GenABEL (GRAMMAR-Gamma), GMA_C = GEMMA using centred genotypes, GMA_S = GEMMA using standardised genotypes, MMM_E = MMM using full mixed model (exact) calculation, MMM_G = MMM using GLS approximation, FBATaff = FBAT using transmissions to affecteds only, FBATboth = FBAT using transmissions to both affecteds and unaffecteds, MQLS = MQLS using theoretical kinships of the 1,972 genotyped individuals, RT = ROADTRIPS using 1,972 individuals.

The genomic inflation factor of 1.12 in the unadjusted analysis was substantially lower that the value of 1.23 in the unadjusted analysis of the real VL phenotype. Furthermore, feeding alternative kinship estimations into a FaST-LMM (Exact) analysis of the simulated phenotype in a similar manner to Section 4.4 showed that, unlike in the analysis of the real phenotype, the inflation in this case can be well controlled using theoretical kinship estimates alone ($\lambda = 1.00$; Figure 5.9). These observations seem to further support the assertion in Section 4.5 that there may be additional relatedness/population structure in the real data set, and may explain the higher concordance between GTAM and other LMM software when analysing this phenotype data. Interestingly, analysis using PLINK's estimated IBD seems to have performed even worse here, with the degree of inflation exactly the same as that of the unadjusted analysis ($\lambda = 1.12$).

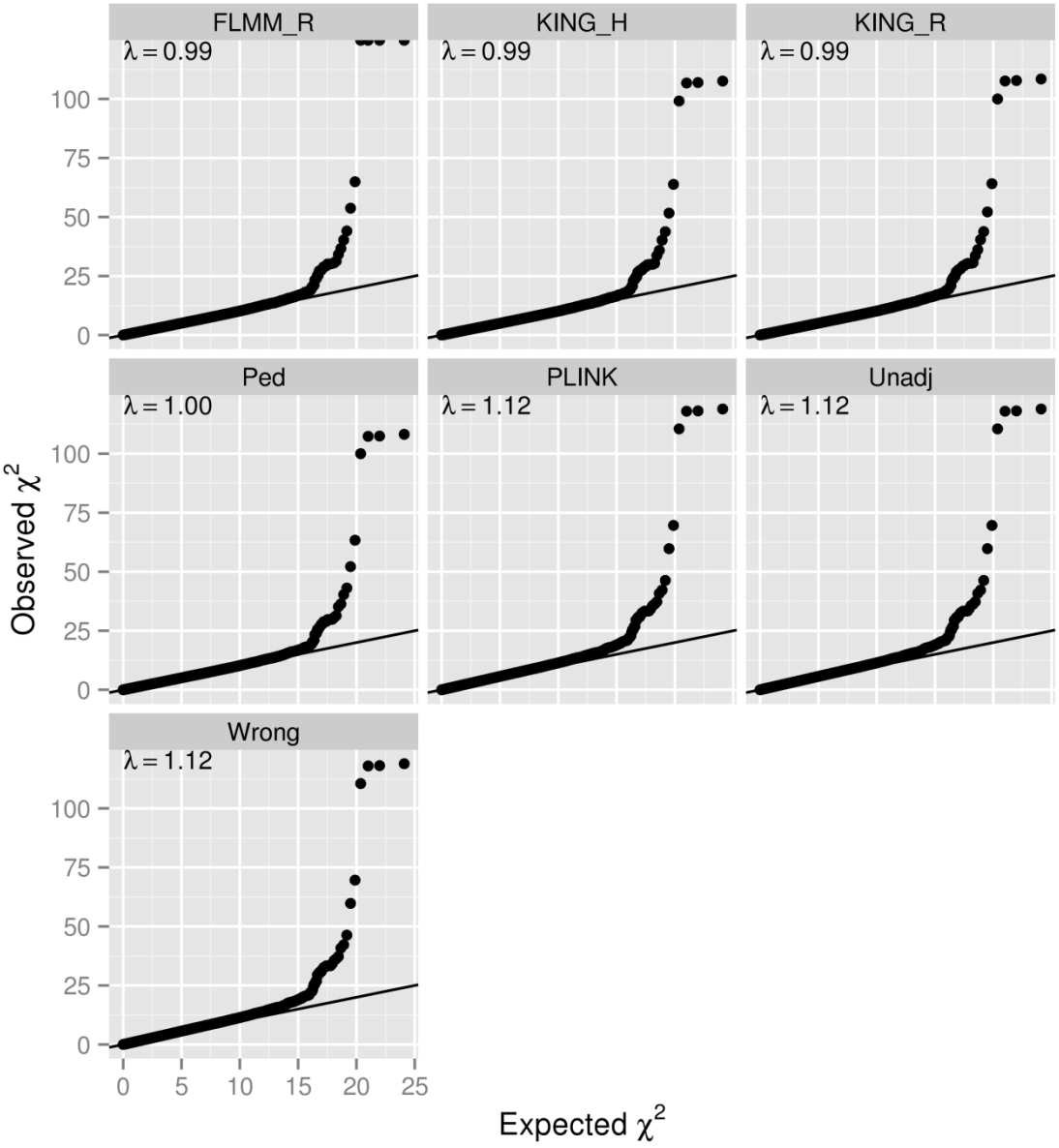

Figure 5.9 Q-Q plots of simulated strong qualitative phenotype GWAS results and genomic inflation factors (λ) obtained from FaST-LMM using alternative kinship estimates. FLMM_R = FaST-LMM's own realised relationship matrix, KING_H = KING homogeneous method, KING_R = KING robust method, Ped = theoretical kinship estimates based on pedigree information, Unadj = unadjusted, Wrong = misspecified kinships, chosen to be inversely related to the true kinship value. The dots at the upper border of the FLMM_R panel represent the SNPs where the equivalent χ^2 values are ∞ (i.e. p-value = 0).

5.2. Performance with Simulated Weak Qualitative Phenotype

The performance of LMM and alternative methods when applied to simulated weak qualitative phenotype did not differ much from when they were applied to the simulated strong qualitative phenotype. The genomic inflation factors from all methods were between 0.99 to 1.01, compared with 1.04 in the unadjusted analysis (Figures 5.10-5.11). Unsurprisingly, FBAT was again unable to detect the effect of the two simulated SNPs (Figures 5.12-5.13).

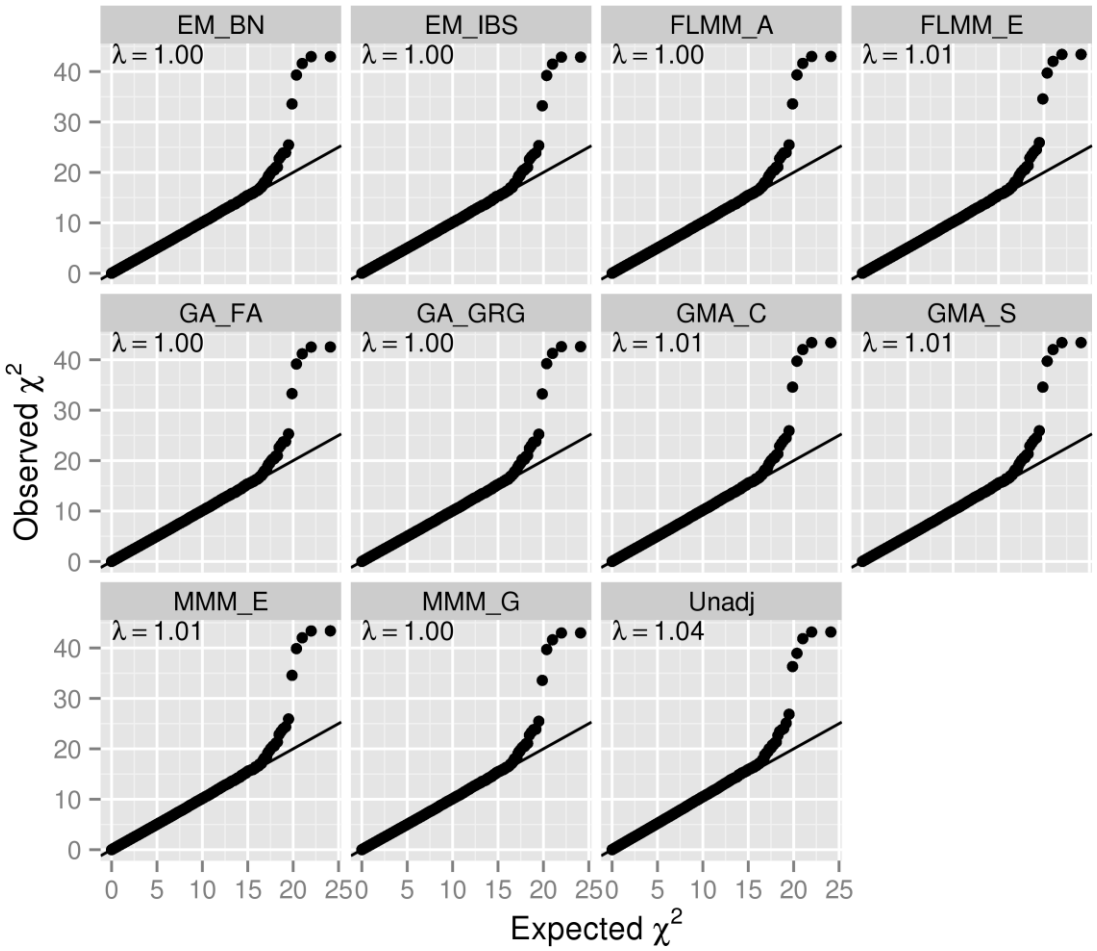

Figure 5.10 Q-Q plots of simulated weak qualitative phenotype GWAS results and genomic inflation factors (λ) for different LMM methods. EM_BN = EMMAX (Balding- Nichols), EM_IBS = EMMAX (IBS method), FLMM_A = FaST-LMM using approximate calculation with RRM, FLMM_E = FaST-LMM using exact calculation with RRM, GA_FA = GenABEL (FASTA), GA_GRG = GenABEL (GRAMMAR-Gamma), GMA_C = GEMMA using centred genotypes, GMA_S = GEMMA using standardised genotypes, MMM_E = MMM using full mixed model (exact) calculation, MMM_G = MMM using GLS approximation, Unadj = unadjusted analysis.

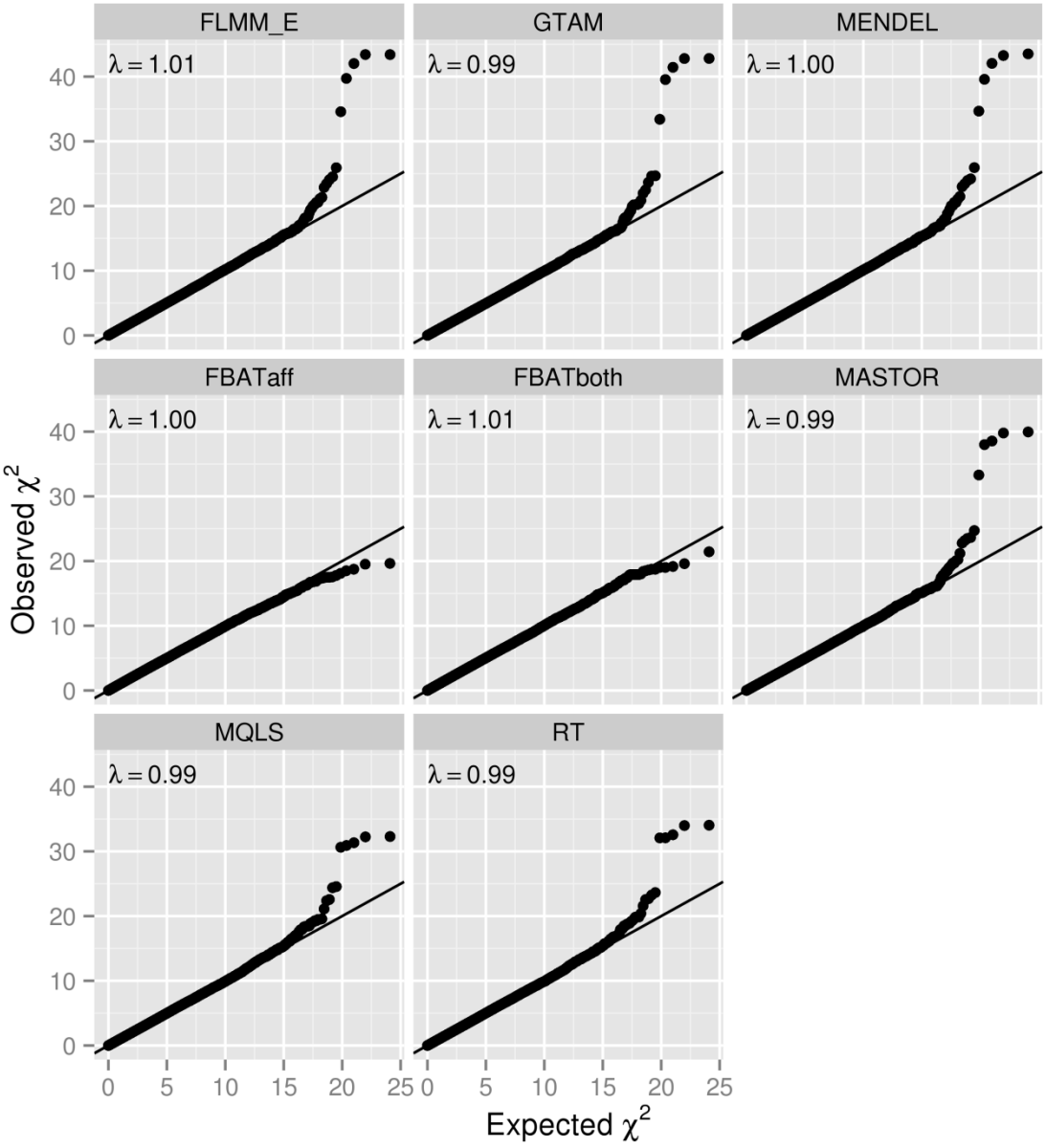

Figure 5.11 Q-Q plots of simulated weak qualitative phenotype GWAS results and genomic inflation factors (λ) for different LMM/alternative methods. FLMM_E = FaST-LMM using exact calculation with RRM, FBATaff = FBAT using transmissions to affecteds only, FBATboth = FBAT using transmissions to both affecteds and unaffecteds, MQLS = MQLS using theoretical kinships of the 1,972 genotyped individuals, RT = ROADTRIPS using 1,972 individuals. FaST-LMM is an LMM method and is included here for comparison; GTAM and Mendel are also LMM methods, but included here due to their unique characteristics.

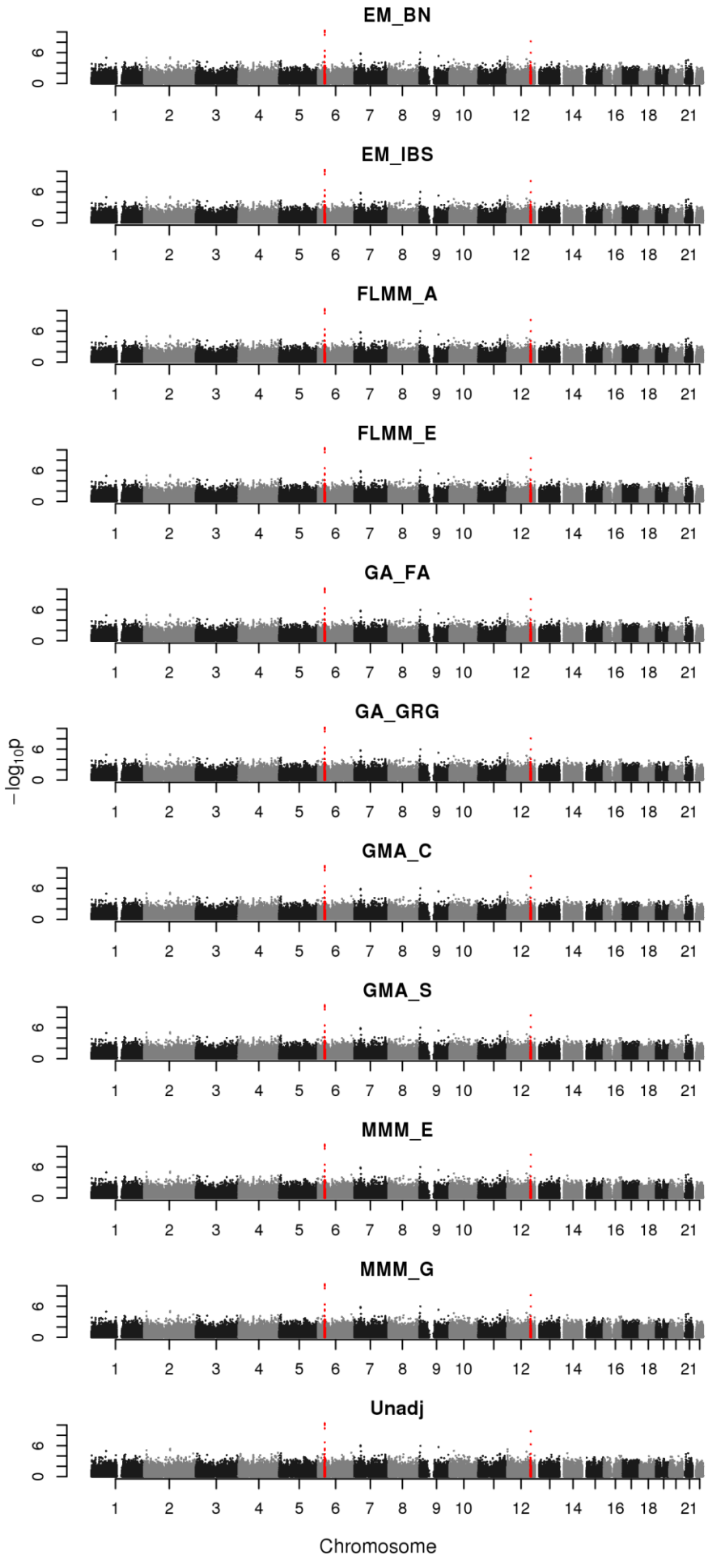
EM_BN
EM_IBS
FLMM_A
FLMM_E
GA_FA
GA_GRG
GMA_C
GMA_S
MMM_E
MMM_G
Unadj
0 6
1 2 3 4 5 6 7 8 9 10 12 14 16 18 21
−log10p
Chromosome

Figure 5.12 Manhattan plots for VL data set with simulated weak qualitative phenotype using various LMM methods. The points marked in red (appear as dark grey area near the beginning of chromosome 6 and the end of chromosome 12 if printed in black and white) denote the simulated weak effect loci. EM_BN = EMMAX (Balding- Nichols), EM_IBS = EMMAX (IBS method), FLMM_A = FaST-LMM using approximate calculation with RRM, FLMM_E = FaST-LMM using exact calculation with RRM, GA_FA = GenABEL (FASTA), GA_GRG = GenABEL (GRAMMAR-Gamma), GMA_C = GEMMA using centred genotypes, GMA_S = GEMMA using standardised genotypes, MMM_E = MMM using full mixed model (exact) calculation, MMM_G = MMM using GLS approximation, Unadj = unadjusted analysis.

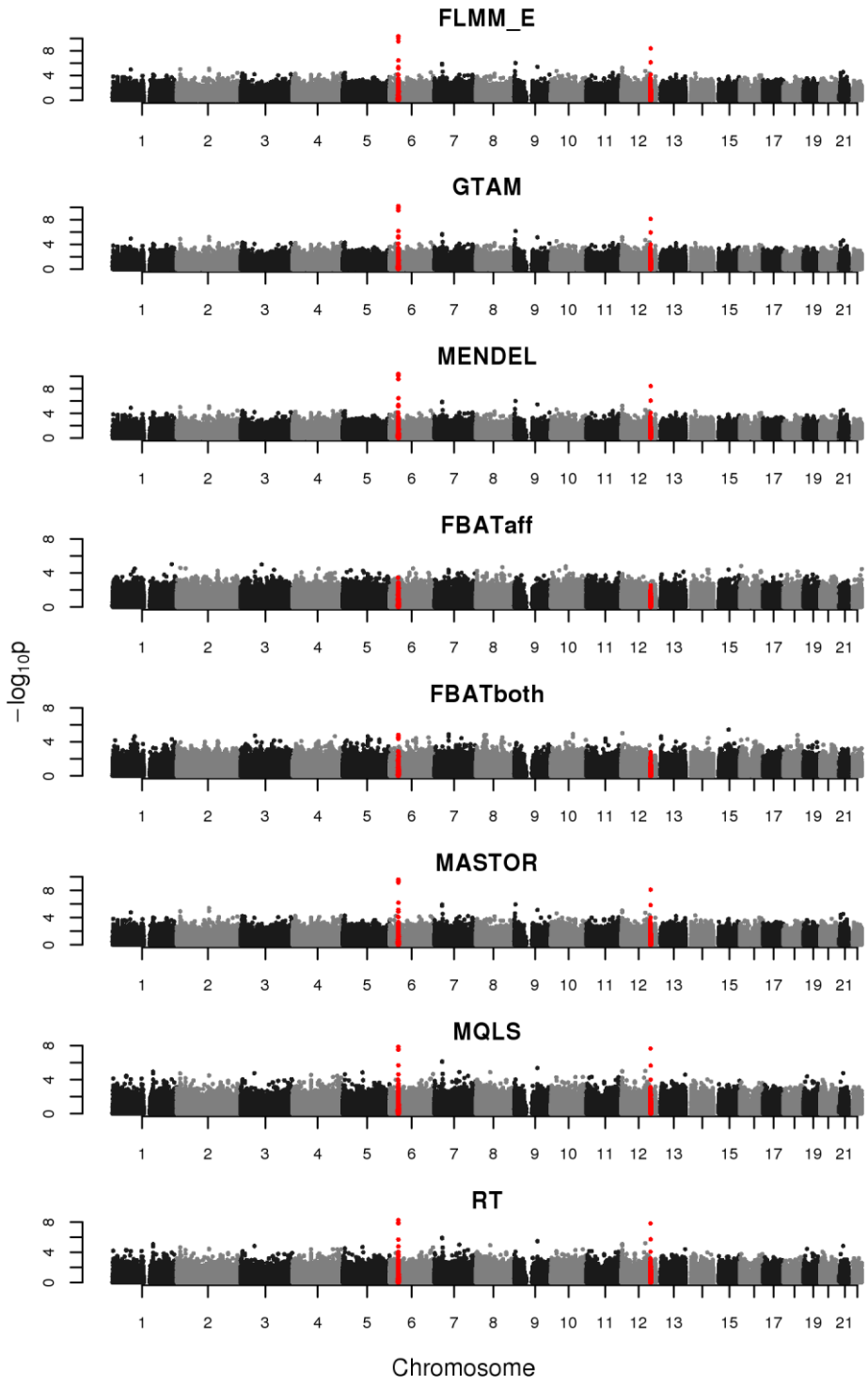

Figure 5.13 Manhattan plots for VL data set with simulated weak qualitative phenotype using various LMM/alternative methods. The points marked in red (appear as dark grey area near the beginning of chromosome 6 and the end of chromosome 12 if printed in black and white) denote the simulated weak effect loci. FLMM_E = FaST-LMM using exact calculation with RRM, FBATaff = FBAT using transmissions to affecteds only, FBATboth = FBAT using transmissions to both affecteds and unaffecteds, MQLS = MQLS using theoretical kinships of the 1,972 genotyped individuals, RT = ROADTRIPS using 1,972 individuals. FaST-LMM is an LMM method and is included here for comparison; GTAM and Mendel are also LMM methods, but included here due to their unique characteristics.

The formal power and type I error analysis (Figure 5.14) again show the similarity among the LMM methods and MASTOR, while ROADTRIPS and MQLS also show slightly less power. FBAT again appeared to be conservative and considerably less powerful than other methods.

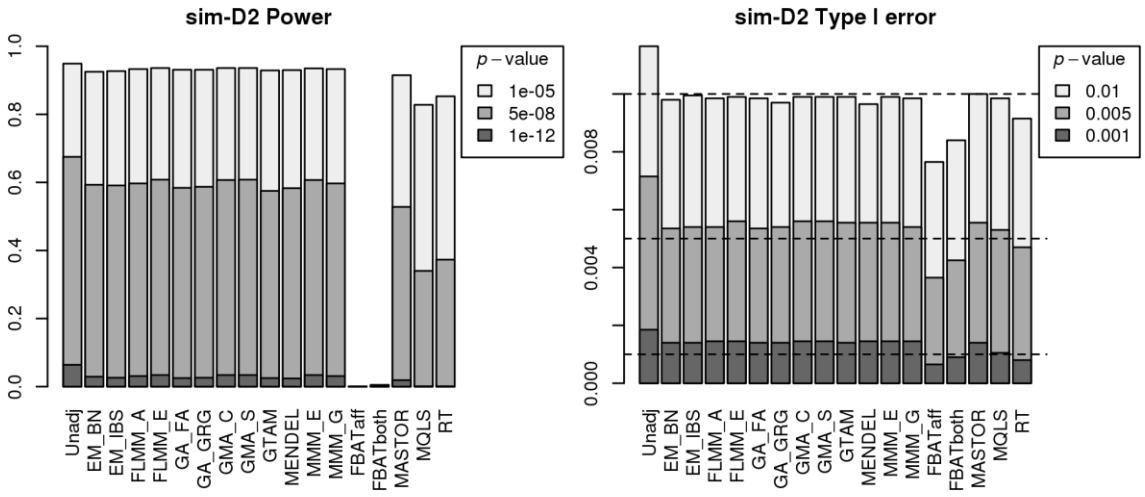

Figure 5.14 Power and type 1 error of different methods when applied to weak binary (disease) phenotype. Powers (left hand plot) are defined as the proportion of replicates (out of 1,000) in which both simulated disease loci are detected, with 'detection' corresponding to any SNP within 40 kb of the simulated disease locus reaching the specified p-value threshold. Type 1 errors (right hand plot) are defined as the proportion of null SNPs (out of 20,000 = 20 null SNPs times 1,000 simulation replicates)

Results from all LMM methods as well as MASTOR were highly concordant (Figures 5.15-5.16). MQLS and ROADTRIPS were also quite concordant with the LMM results, but to a lesser extent. FBAT again showed little concordance to the other methods.

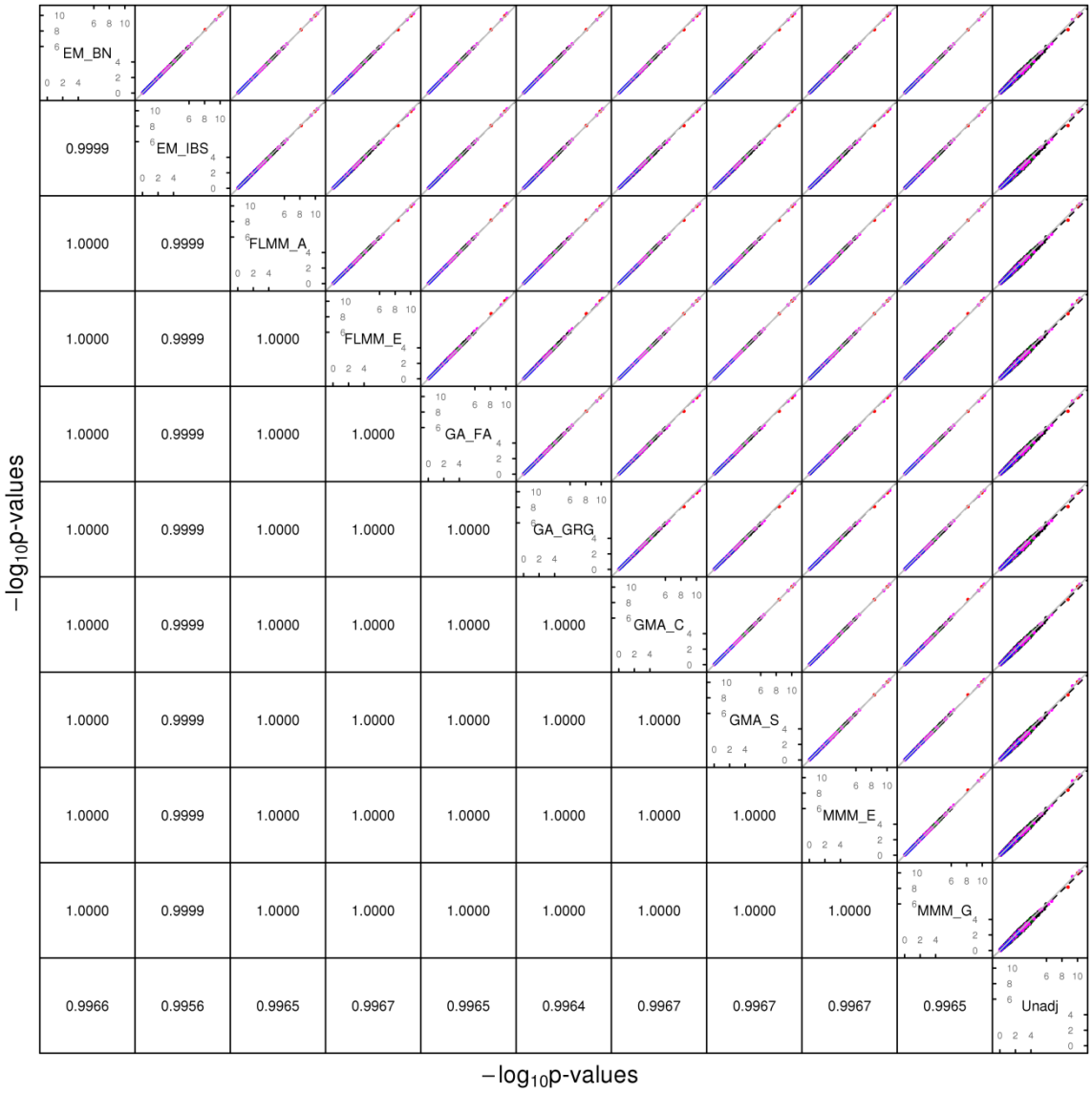

Figure 5.15 Comparison of -log(p-values) using various LMM software packages, simulated weak binary (disease) phenotype. Plots above the diagonal show a comparison of $-\log_{10}$(p-values), with correlations between the $-\log_{10}$(p-values) indicated below the diagonal. The grey solid lines represent the line of equality; the black dashed lines the linear regression line of the variable on the y axis on the variable on the x axis. The colours denote: red = the two weak effect SNPs, magenta = SNPs within 2 Mb of the weak effect SNPs, blue = 22 polygenic SNPs, green = SNPs within 2 Mb of the polygenic SNPs, black = all other SNPs. Because the black/green/blue SNPs were plotted before the magenta/red SNPs, they may be obscured by the latter. EM_BN = EMMAX (Balding- Nichols), EM_IBS = EMMAX (IBS method), FLMM_A = FaST-LMM using approximate calculation with RRM, FLMM_E = FaST-LMM using exact calculation with RRM, GA_FA = GenABEL (FASTA), GA_GRG = GenABEL (GRAMMAR-Gamma), GMA_C = GEMMA using centred genotypes, GMA_S = GEMMA using standardised genotypes, MMM_E = MMM using full mixed model (exact) calculation, MMM_G = MMM using GLS approximation, Unadj = unadjusted analysis.

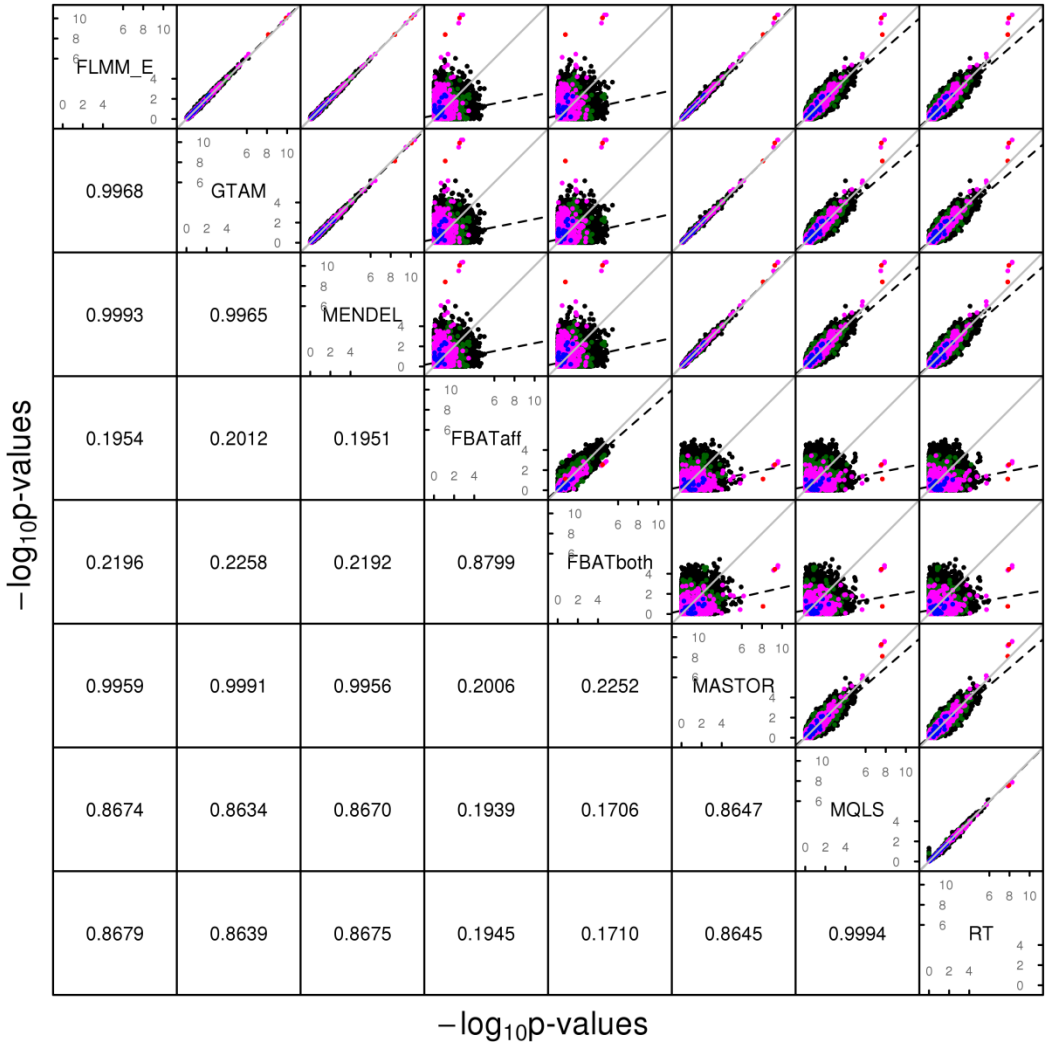

Figure 5.16 Comparison of -log(p-values) using various LMM software packages, simulated weak binary (disease) phenotype. Plots above the diagonal show a comparison of $-\log_{10}$(p-values), with correlations between the $-\log_{10}$(p-values) indicated below the diagonal. The grey solid lines represent the line of equality; the black dashed lines the linear regression line of the variable on the y axis on the variable on the x axis. The colours denote: red = the two weak effect SNPs, magenta = SNPs within 2 Mb of the weak effect SNPs, blue = 22 polygenic SNPs, green = SNPs within 2 Mb of the polygenic SNPs, black = all other SNPs. Because the black/green/blue SNPs were plotted before the magenta/red SNPs, they may be obscured by the latter. FLMM_E = FaST-LMM using exact calculation with RRM, FBATaff = FBAT using transmissions to affecteds only, FBATboth = FBAT using transmissions to both affecteds and unaffecteds, MQLS = MQLS using theoretical kinships of the 1,972 genotyped individuals, RT = ROADTRIPS using 1,972 individuals. FaST-LMM is an LMM method and is included here for comparison; GTAM and Mendel are also LMM methods, but included here due to their unique characteristics.

Formal comparison of concordance between the 'top hits' identified by each method again showed extremely high concordance in all methods except FBAT (Table 5.2), similar to that seen in the simulated strong binary phenotype.

	Mean (standard deviation) in 1,000 replicates of proportion of top *t* SNPs within null and true regions that overlap with top *t* SNPs from EM_BN				
method	*t* = 5	*t* = 10	*t* = 15	*t* = 20	*t* = 25
Unadjusted	0.982 (0.060)	0.984 (0.041)	0.979 (0.039)	0.974 (0.040)	0.973 (0.036)
EM_IBS	0.997 (0.029)	0.997 (0.024)	0.995 (0.025)	0.994 (0.028)	0.994 (0.024)
FLMM_A	0.998 (0.027)	0.998 (0.024)	0.997 (0.025)	0.997 (0.029)	0.997 (0.026)
FLMM_E	0.995 (0.035)	0.997 (0.025)	0.997 (0.025)	0.996 (0.030)	0.997 (0.026)
GA_FA	0.992 (0.044)	0.998 (0.024)	0.997 (0.026)	0.996 (0.030)	0.996 (0.026)
GA_GRG	0.994 (0.038)	0.997 (0.026)	0.996 (0.027)	0.995 (0.030)	0.996 (0.026)
GMA_C	0.995 (0.035)	0.997 (0.025)	0.997 (0.025)	0.996 (0.030)	0.997 (0.026)
GMA_S	0.995 (0.035)	0.997 (0.025)	0.997 (0.025)	0.996 (0.030)	0.997 (0.026)
GTAM	0.988 (0.050)	0.990 (0.036)	0.983 (0.037)	0.982 (0.036)	0.982 (0.032)
MENDEL	0.988 (0.051)	0.992 (0.033)	0.986 (0.035)	0.984 (0.036)	0.987 (0.031)
MMM_E	0.995 (0.037)	0.997 (0.025)	0.997 (0.025)	0.996 (0.030)	0.997 (0.026)
MMM_G	0.998 (0.028)	0.998 (0.024)	0.997 (0.025)	0.997 (0.029)	0.997 (0.026)
FBATaff	0.413 (0.255)	0.571 (0.201)	0.614 (0.157)	0.639 (0.128)	0.651 (0.102)
FBATboth	0.664 (0.246)	0.718 (0.146)	0.699 (0.111)	0.691 (0.099)	0.686 (0.088)
MASTOR	0.971 (0.075)	0.988 (0.038)	0.981 (0.038)	0.978 (0.039)	0.979 (0.033)
MQLS	0.934 (0.107)	0.962 (0.056)	0.942 (0.053)	0.928 (0.051)	0.917 (0.047)
RT	0.943 (0.099)	0.965 (0.055)	0.943 (0.053)	0.930 (0.052)	0.919 (0.047)

Table 5.2 Concordance between top SNPs identified by different methods analysing simulated weak binary (disease) phenotype. EM_BN = EMMAX (Balding- Nichols), EM_IBS = EMMAX (IBS method), FLMM_A = FaST-LMM using approximate calculation with RRM, FLMM_E = FaST-LMM using exact calculation with RRM, GA_FA = GenABEL (FASTA), GA_GRG = GenABEL (GRAMMAR-Gamma), GMA_C = GEMMA using centred genotypes, GMA_S = GEMMA using standardised genotypes, MMM_E = MMM using full mixed model (exact) calculation, MMM_G = MMM using GLS approximation, FBATaff = FBAT using transmissions to affecteds only, FBATboth = FBAT using transmissions to both affecteds and unaffecteds, MQLS = MQLS using theoretical kinships of the 1,972 genotyped individuals, RT = ROADTRIPS using 1,972 individuals.

5.3. Performance with Simulated Quantitative Phenotype

The findings from applying the LMM and alternative methods to simulated (strong) quantitative phenotype are similar to those observed in the two simulations above. All methods were very successful in controlling the genomic inflation to 0.99-1.00, compared with 1.43 in unadjusted analysis (Figures 5.17-5.18). All methods except FBAT detected clear signals at the simulated loci (Figures 5.19-5.20). FBAT appeared to have detected a weak signal at the stronger effect locus (chromosome 6) in this particular simulation set (Figure 5.20), but had almost no power in the formal power analysis (Figure 5.21), probably due to its failure to detect the weaker effect locus (chromosome 12), which resulted in its overall result being classed as non-detection.

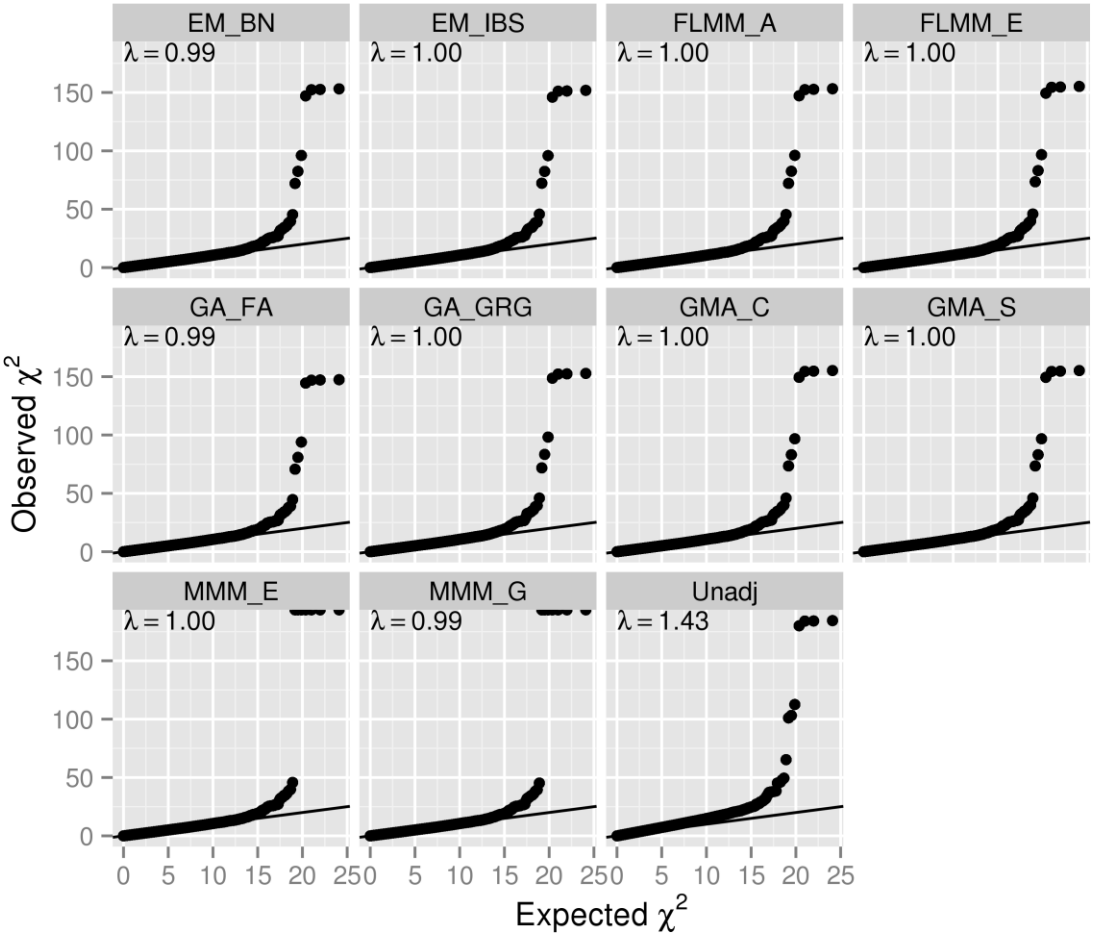

Figure 5.17 Q-Q plots of simulated quantitative phenotype GWAS results and genomic inflation factors (λ) for different LMM methods. EM_BN = EMMAX (Balding- Nichols), EM_IBS = EMMAX (IBS method), FLMM_A = FaST-LMM using approximate calculation with RRM, FLMM_E = FaST-LMM using exact calculation with RRM, GA_FA = GenABEL (FASTA), GA_GRG = GenABEL (GRAMMAR-Gamma), GMA_C = GEMMA using centred genotypes, GMA_S = GEMMA using standardised genotypes, MMM_E = MMM using full mixed model (exact) calculation, MMM_G = MMM using GLS approximation, Unadj = unadjusted analysis. The dots at the upper border of the MMM panels represent the SNPs where the equivalent χ^2 values are ∞ (i.e. p-value = 0).

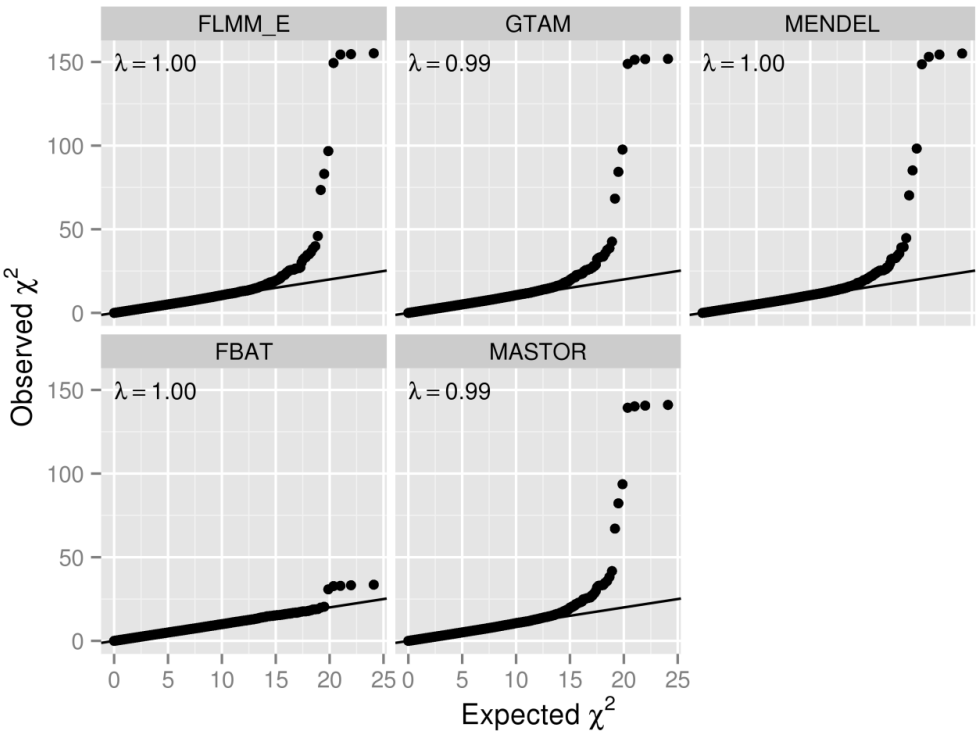

Figure 5.18 Q-Q plots of simulated quantitative phenotype GWAS results and genomic inflation factors (λ) for different LMM/alternative methods. FLMM_E = FaST-LMM using exact calculation with RRM, FBAT = FBAT using transmissions to all individuals. FaST-LMM is an LMM method and is included here for comparison; GTAM and Mendel are also LMM methods, but included here due to their unique characteristics.

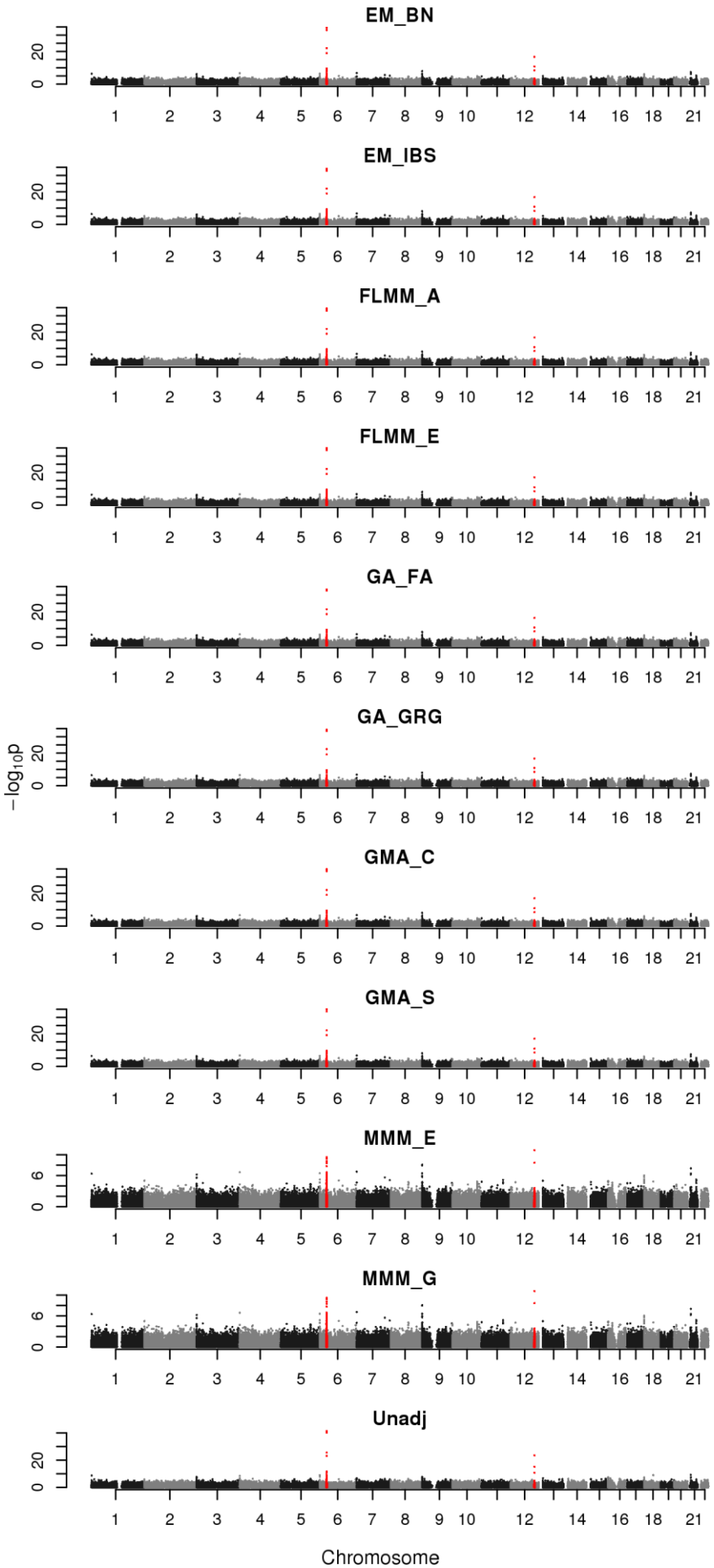
EM_BN
EM_IBS
FLMM_A
FLMM_E
GA_FA
GA_GRG
GMA_C
GMA_S
MMM_E
MMM_G
Unadj
−log10p
Chromosome
1 2 3 4 5 6 7 8 9 10 12 14 16 18 21
0 20
0 6

Figure 5.19 Manhattan plots for VL data set with simulated quantitative phenotype using various LMM methods. The points marked in red (appear as dark grey area near the beginning of chromosome 6 and the end of chromosome 12 if printed in black and white) denote the simulated strong effect loci. EM_BN = EMMAX (Balding- Nichols), EM_IBS = EMMAX (IBS method), FLMM_A = FaST-LMM using approximate calculation with RRM, FLMM_E = FaST-LMM using exact calculation with RRM, GA_FA = GenABEL (FASTA), GA_GRG = GenABEL (GRAMMAR-Gamma), GMA_C = GEMMA using centred genotypes, GMA_S = GEMMA using standardised genotypes, MMM_E = MMM using full mixed model (exact) calculation, MMM_G = MMM using GLS approximation, Unadj = unadjusted analysis.

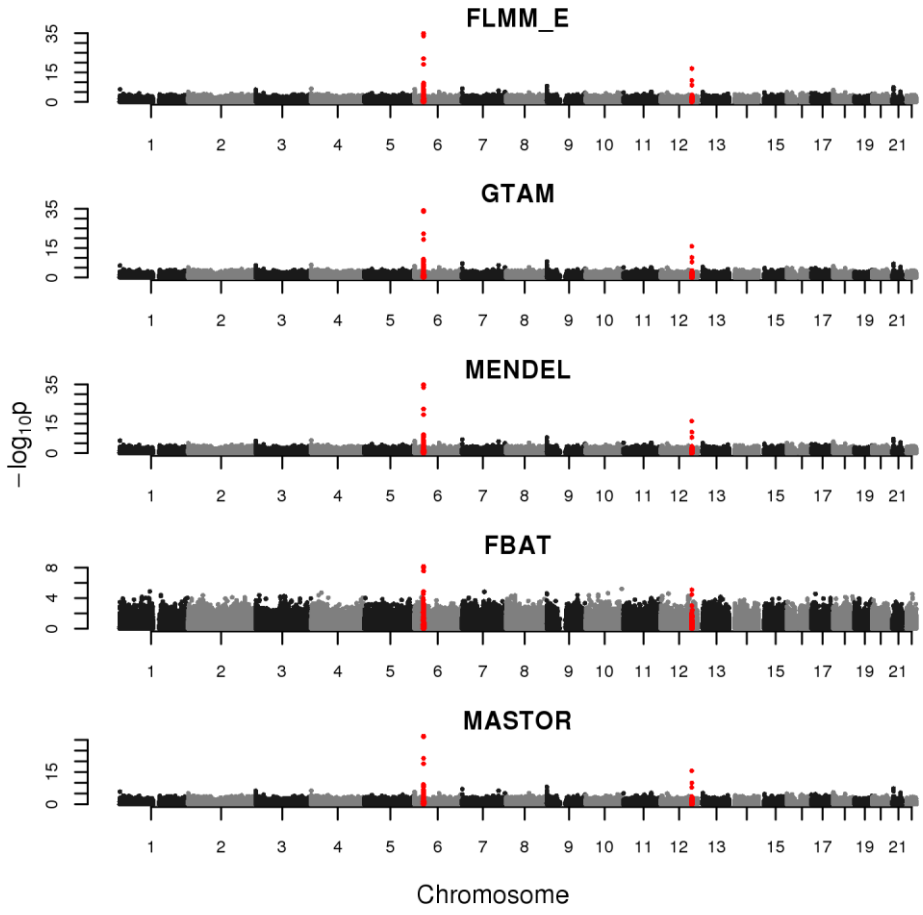

Figure 5.20 Manhattan plots for VL data set with simulated quantitative phenotype using various LMM/alternative methods. The points marked in red (appear as dark grey area near the beginning of chromosome 6 and the end of chromosome 12 if printed in black and white) denote the simulated strong effect loci. FLMM_E = FaST-LMM using exact calculation with RRM, FBAT = FBAT using transmissions to all individuals. FaST-LMM is an LMM method and is included here for comparison; GTAM and Mendel are also LMM methods, but included here due to their unique characteristics.

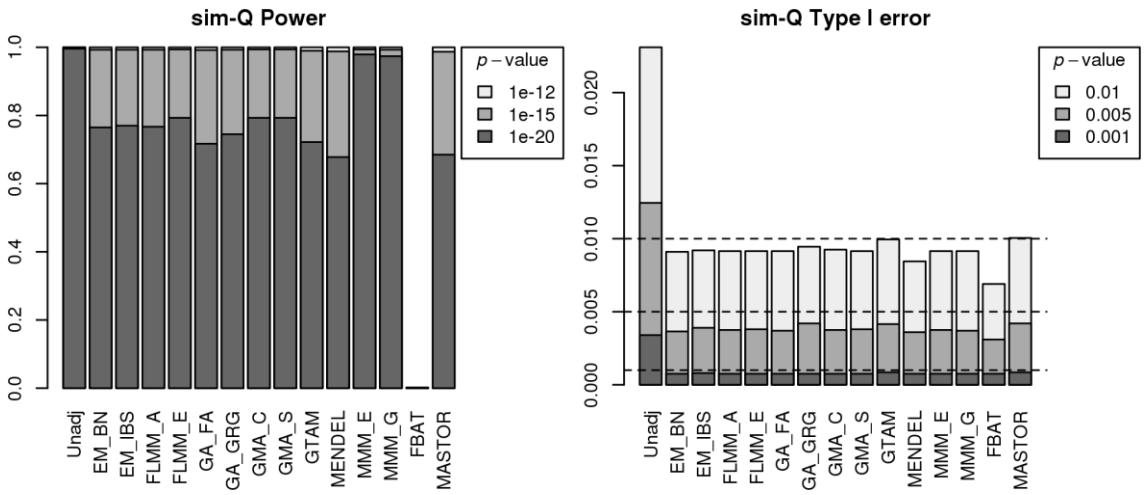

Figure 5.21 Power and type 1 error of different methods when applied to quantitative phenotype. Powers (left hand plot) are defined as the proportion of replicates (out of 1,000) in which both simulated disease loci are detected, with 'detection' corresponding to any SNP within 40 kb of the simulated disease locus reaching the specified p-value threshold. Type 1 errors (right hand plot) are defined as the proportion of null SNPs (out of 20,000 = 20 null SNPs times 1,000 simulation replicates)

The concordance of the results was extremely high across all methods except FBAT (Figures 5.22-5.23; but then MQLS and ROADTRIPS, which showed slightly less degree of concordance in the previous simulations, were not included in this simulation as they were not applicable to quantitative trait analysis). FBAT's results were actually quite concordant with other methods near the simulated loci, but less so at the other SNPs. This is also reflected in the top SNPs comparison (Table 5.3), where, unlike the previous simulations, FBAT showed a reasonable degree of concordance to EMMAX when only a small number (5) of the very top SNPs were compared, this then deteriorated when more SNPs were included for comparison, which again suggests higher discrepancies among the less associated SNPs. Interestingly, MMM also appears to behave slightly differently from the previous simulations, and in the opposite way to FBAT: with small number of the very top SNPs, the degree of concordance with EMMAX was lower than most other methods and was quite comparable with FBAT at about 0.90; this then quickly improved when more SNPs were included in the comparison and reached 1.00 with just 10 top SNPs. This suggests only a minor discrepancy of the p-values of some of the top SNPs.

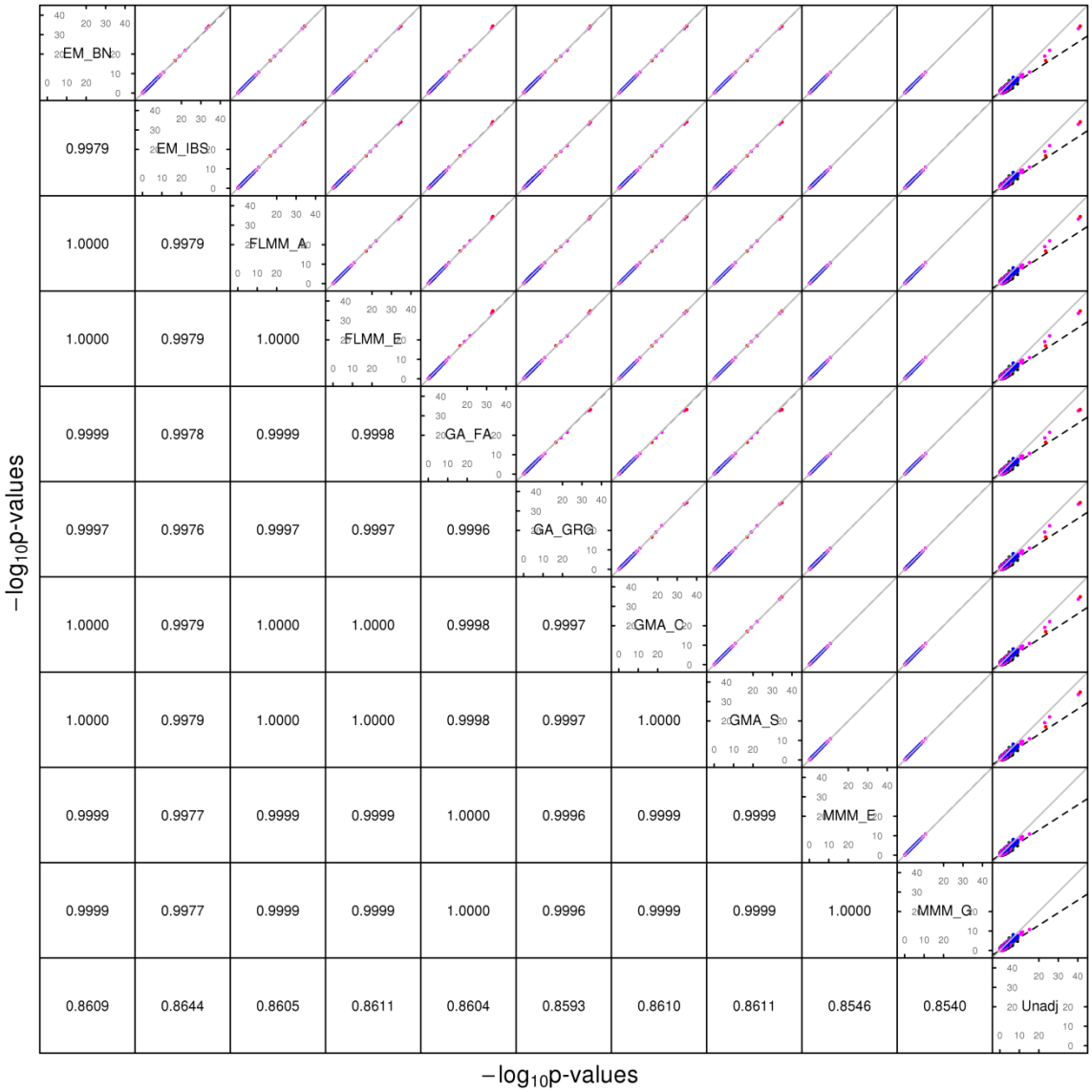

Figure 5.22 Comparison of -log(p-values) using various LMM software packages, simulated quantitative phenotype. Plots above the diagonal show a comparison of $-\log_{10}$(p-values), with correlations between the $-\log_{10}$(p-values) indicated below the diagonal. The grey solid lines represent the line of equality; the black dashed lines the linear regression line of the variable on the y axis on the variable on the x axis. The colours denote: red = the two strong effect SNPs, magenta = SNPs within 2 Mb of the strong effect SNPs, blue = 22 polygenic SNPs, green = SNPs within 2 Mb of the polygenic SNPs, black = all other SNPs. Because the black/green/blue SNPs were plotted before the magenta/red SNPs, they may be obscured by the latter. EM_BN = EMMAX (Balding- Nichols), EM_IBS = EMMAX (IBS method), FLMM_A = FaST-LMM using approximate calculation with RRM, FLMM_E = FaST-LMM using exact calculation with RRM, GA_FA = GenABEL (FASTA), GA_GRG = GenABEL (GRAMMAR-Gamma), GMA_C = GEMMA using centred genotypes, GMA_S = GEMMA using standardised genotypes, MMM_E = MMM using full mixed model (exact) calculation, MMM_G = MMM using GLS approximation, Unadj = unadjusted analysis.

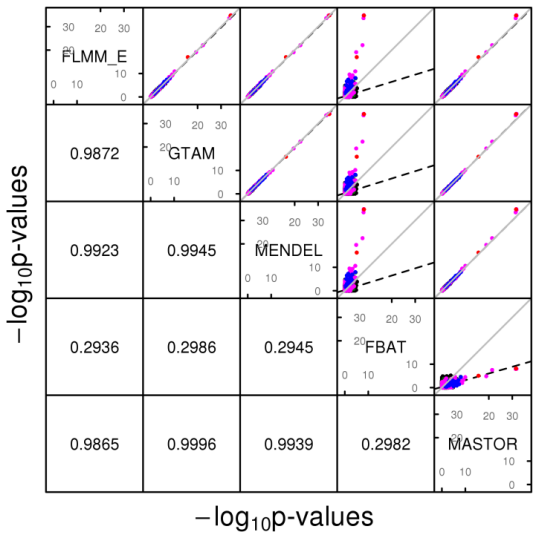

Figure 5.23 Comparison of -log(p-values) using various LMM software packages, simulated quantitative phenotype. Plots above the diagonal show a comparison of $-\log_{10}$(p-values), with correlations between the $-\log_{10}$(p-values) indicated below the diagonal. The grey solid lines represent the line of equality; the black dashed lines the linear regression line of the variable on the y axis on the variable on the x axis. The colours denote: red = the two strong effect SNPs, magenta = SNPs within 2 Mb of the strong effect SNPs, blue = 22 polygenic SNPs, green = SNPs within 2 Mb of the polygenic SNPs, black = all other SNPs. Because the black/green/blue SNPs were plotted before the magenta/red SNPs, they may be obscured by the latter. FLMM_E = FaST-LMM using exact calculation with RRM, FBAT = FBAT using transmissions to all individuals. FaST-LMM is an LMM method and is included here for comparison; GTAM and Mendel are also LMM methods, but included here due to their unique characteristics.

	Mean (standard deviation) in 1,000 replicates of proportion of top *t* SNPs within null and true regions that overlap with top *t* SNPs from EM_BN				
method	*t* = 5	*t* = 10	*t* = 15	*t* = 20	*t* = 25
Unadjusted	0.987 (0.049)	0.983 (0.038)	0.962 (0.040)	0.963 (0.034)	0.954 (0.033)
EM_IBS	0.998 (0.020)	0.998 (0.016)	0.993 (0.020)	0.994 (0.017)	0.993 (0.015)
FLMM_A	1.000 (0.000)	1.000 (0.000)	1.000 (0.004)	1.000 (0.005)	1.000 (0.004)
FLMM_E	1.000 (0.009)	0.999 (0.008)	1.000 (0.005)	1.000 (0.005)	0.999 (0.005)
GA_FA	1.000 (0.006)	0.999 (0.010)	0.998 (0.010)	0.998 (0.010)	0.996 (0.012)
GA_GRG	0.994 (0.034)	0.999 (0.010)	0.995 (0.018)	0.996 (0.014)	0.996 (0.012)
GMA_C	1.000 (0.009)	1.000 (0.007)	1.000 (0.004)	1.000 (0.004)	1.000 (0.004)
GMA_S	1.000 (0.009)	0.999 (0.008)	1.000 (0.005)	1.000 (0.005)	0.999 (0.005)
GTAM	0.995 (0.032)	0.991 (0.028)	0.984 (0.030)	0.985 (0.024)	0.984 (0.022)
MENDEL	0.998 (0.021)	0.996 (0.020)	0.987 (0.027)	0.988 (0.022)	0.988 (0.019)
MMM_E	0.899 (0.100)	0.999 (0.008)	1.000 (0.004)	1.000 (0.004)	1.000 (0.004)
MMM_G	0.903 (0.100)	1.000 (0.003)	1.000 (0.003)	1.000 (0.004)	1.000 (0.003)
FBAT	0.906 (0.101)	0.896 (0.067)	0.869 (0.059)	0.844 (0.067)	0.814 (0.066)
MASTOR	0.998 (0.020)	0.992 (0.027)	0.984 (0.030)	0.984 (0.025)	0.983 (0.023)

Table 5.3 Concordance between top SNPs identified by different methods analysing simulated quantitative phenotype. EM_BN = EMMAX (Balding- Nichols), EM_IBS = EMMAX (IBS method), FLMM_A = FaST-LMM using approximate calculation with RRM, FLMM_E = FaST-LMM using exact calculation with RRM, GA_FA = GenABEL (FASTA), GA_GRG = GenABEL (GRAMMAR-Gamma), GMA_C = GEMMA using

centred genotypes, GMA_S = GEMMA using standardised genotypes, MMM_E = MMM using full mixed model (exact) calculation, MMM_G = MMM using GLS approximation, FBAT = FBAT using transmissions to all individuals.

The pattern of inflation when various methods of kinship estimations were used was very similar to that seen in the strong binary phenotype simulation: FaST-LMM's RRM and both KING methods were very successful in controlling the inflation, as was the theoretical kinships calculated only from pedigree information (λ = 0.99-1.00; Figure 5.24). On the contrary, adjustment using PLINK's kinship estimation resulted inflation as high as in the unadjusted analysis or analysis using the 'wrong' kinship estimates (λ = 1.43).

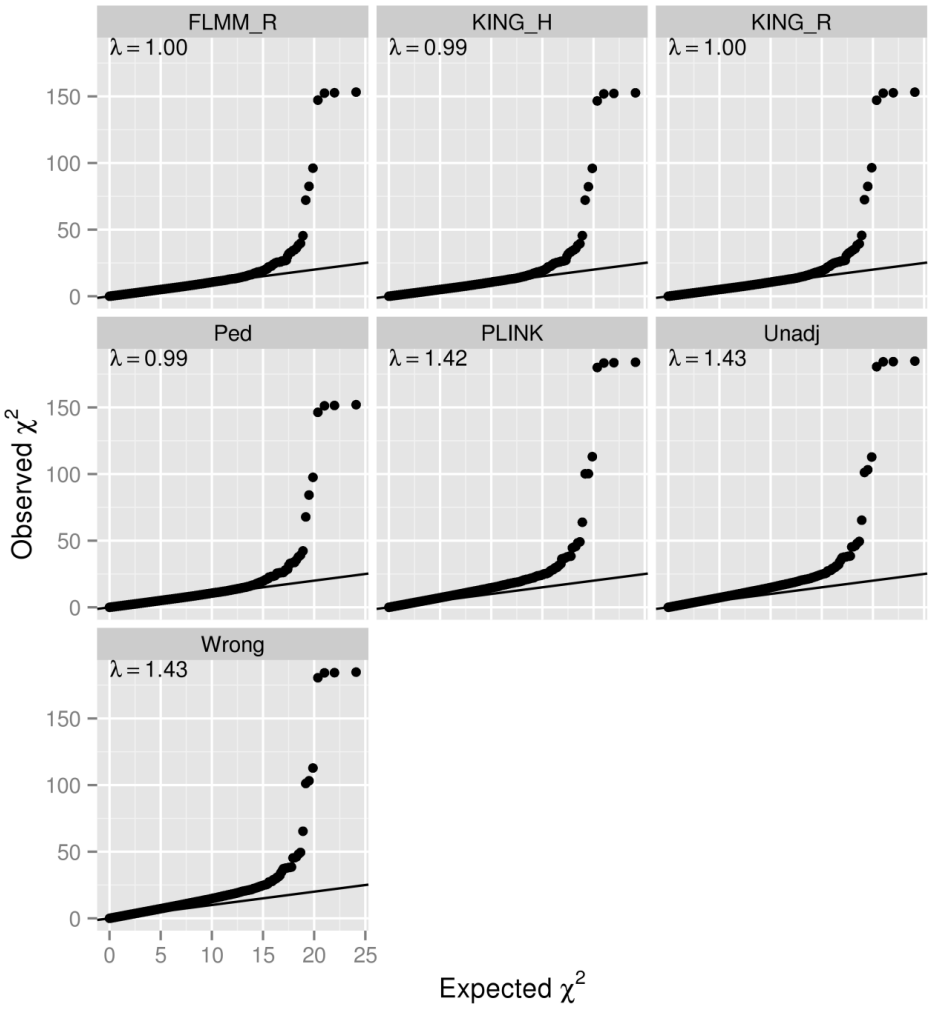

Figure 5.24 Q-Q plots of simulated strong qualitative phenotype GWAS results and genomic inflation factors (λ) obtained from FaST-LMM using alternative kinship estimates. FLMM_R = FaST-LMM's own realised relationship matrix, KING_H = KING homogeneous method, KING_R = KING robust method, Ped = theoretical kinship estimates based on pedigree information, Unadj = unadjusted, Wrong = misspecified kinships, chosen to be inversely related to the true kinship value.

5.4. Performance with Simulated Longitudinal Quantitative Phenotype

In Chapter 3, a strategy for analysing longitudinal traits (repeated measures) in a linear mixed model framework simply by treating each measurement as if it came from a different individual and expanding out the genetic data set accordingly—resulting in an expanded data set containing many apparent twins, triplets, quadruplets etc., depending on how many measurements are available for each person—was investigated using the GAW18 data. It will now be investigated in the current data set using a single replicate of data (498 individuals) simulated under either a longitudinal (sim-L20) or longitudinal polygenic (sim-P20) model (see Section 2.4.3 for details), which, at 20 measurements per person, have much higher degree of repetition than the GAW18 data.

The results from the longitudinal (sim-L20) simulation (Figures 5.25-5.26) showed that EMMAX, FaST-LMM, GEMMA and MMM were successful in maintaining the genomic inflation factor to about 1, whereas GenABEL (FASTA) showed some inflation (λ = 1.06) but was far better than the unadjusted analysis (λ = 20.82). Interestingly, GenABEL (GRAMMAR-Gamma) showed strong *deflation* (λ = 0.66), unlike other methods; in particular this was opposite to that seen in GenABEL (FASTA) to which it is supposed to be equivalent.

The Manhattan plots (Figure 5.27) showed that all LMM methods were successful in separating the true signals from background noise. Similarly, comparison of the concordance in $-\log_{10}$ p-values achieved by the different methods (Figure 5.28) indicated that the results from different methods were highly correlated. However, the actual p-values achieved were very different, consistent with the differences seen in overall distribution of test statistics.

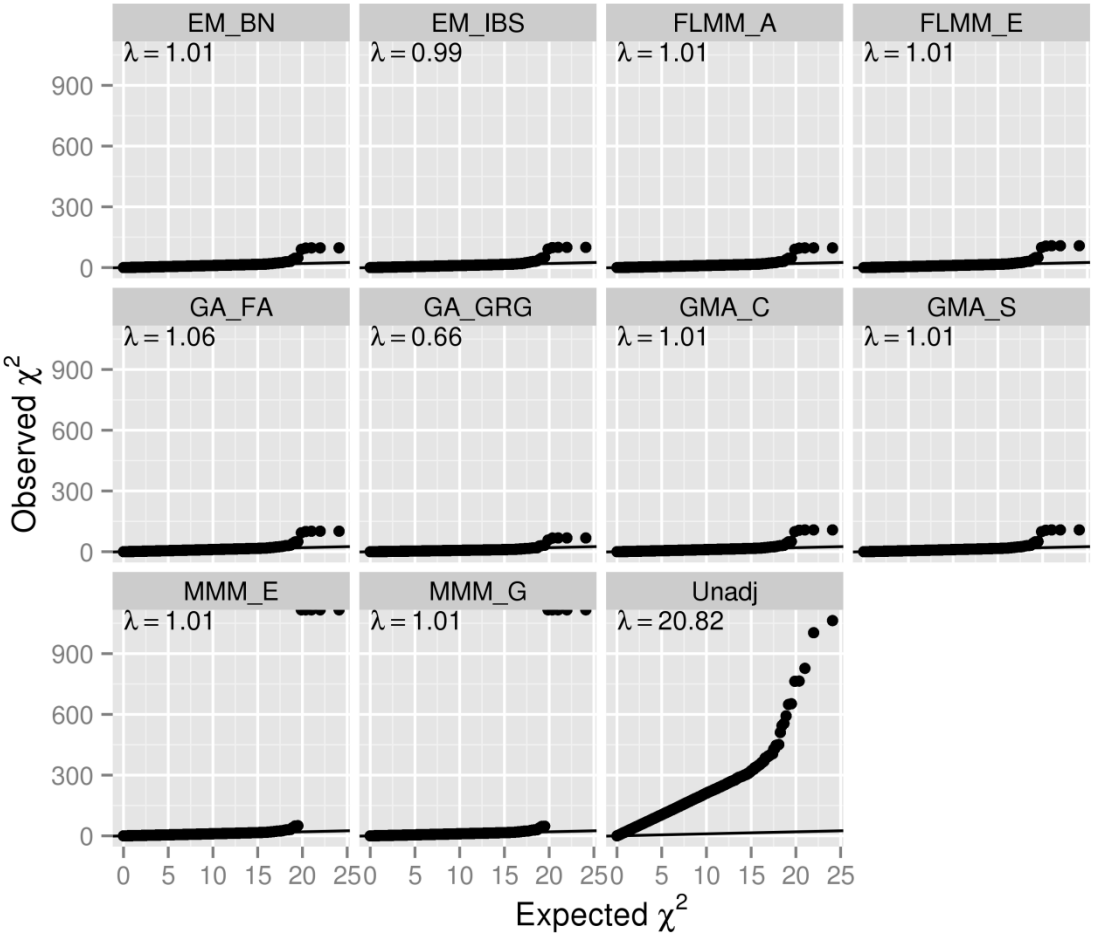

Figure 5.25 Q-Q plots of simulated longitudinal quantitative phenotype GWAS results and genomic inflation factors (λ) for different LMM methods. EM_BN = EMMAX (Balding- Nichols), EM_IBS = EMMAX (IBS method), FLMM_A = FaST-LMM using approximate calculation with RRM, FLMM_E = FaST-LMM using exact calculation with RRM, GA_FA = GenABEL (FASTA), GA_GRG = GenABEL (GRAMMAR-Gamma), GMA_C = GEMMA using centred genotypes, GMA_S = GEMMA using standardised genotypes, MMM_E = MMM using full mixed model (exact) calculation, MMM_G = MMM using GLS approximation, Unadj = unadjusted analysis. The dots at the upper border of the MMM panels represent the SNPs where the equivalent χ^2 values are ∞ (i.e. p-value = 0).

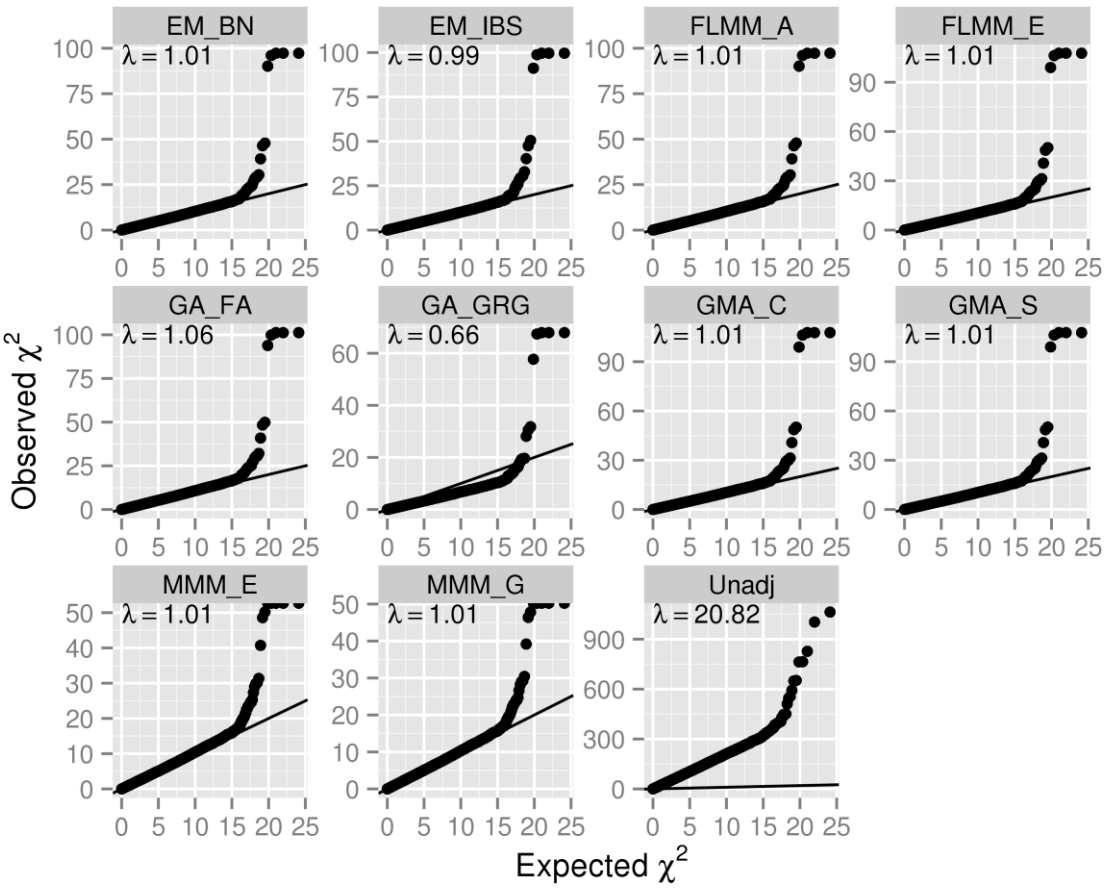

Figure 5.26 Q-Q plots of simulated longitudinal quantitative phenotype GWAS results and genomic inflation factors (λ) for different LMM methods, with each panel plotted on its own scale. EM_BN = EMMAX (Balding- Nichols), EM_IBS = EMMAX (IBS method), FLMM_A = FaST-LMM using approximate calculation with RRM, FLMM_E = FaST-LMM using exact calculation with RRM, GA_FA = GenABEL (FASTA), GA_GRG = GenABEL (GRAMMAR-Gamma), GMA_C = GEMMA using centred genotypes, GMA_S = GEMMA using standardised genotypes, MMM_E = MMM using full mixed model (exact) calculation, MMM_G = MMM using GLS approximation, Unadj = unadjusted analysis. The dots at the upper border of the MMM panels represent the SNPs where the equivalent χ^2 values are ∞ (i.e. p-value = 0). Unlike the previous plot, each panel in this plot has its own y-axis scale to better depict the distribution within its own panel.

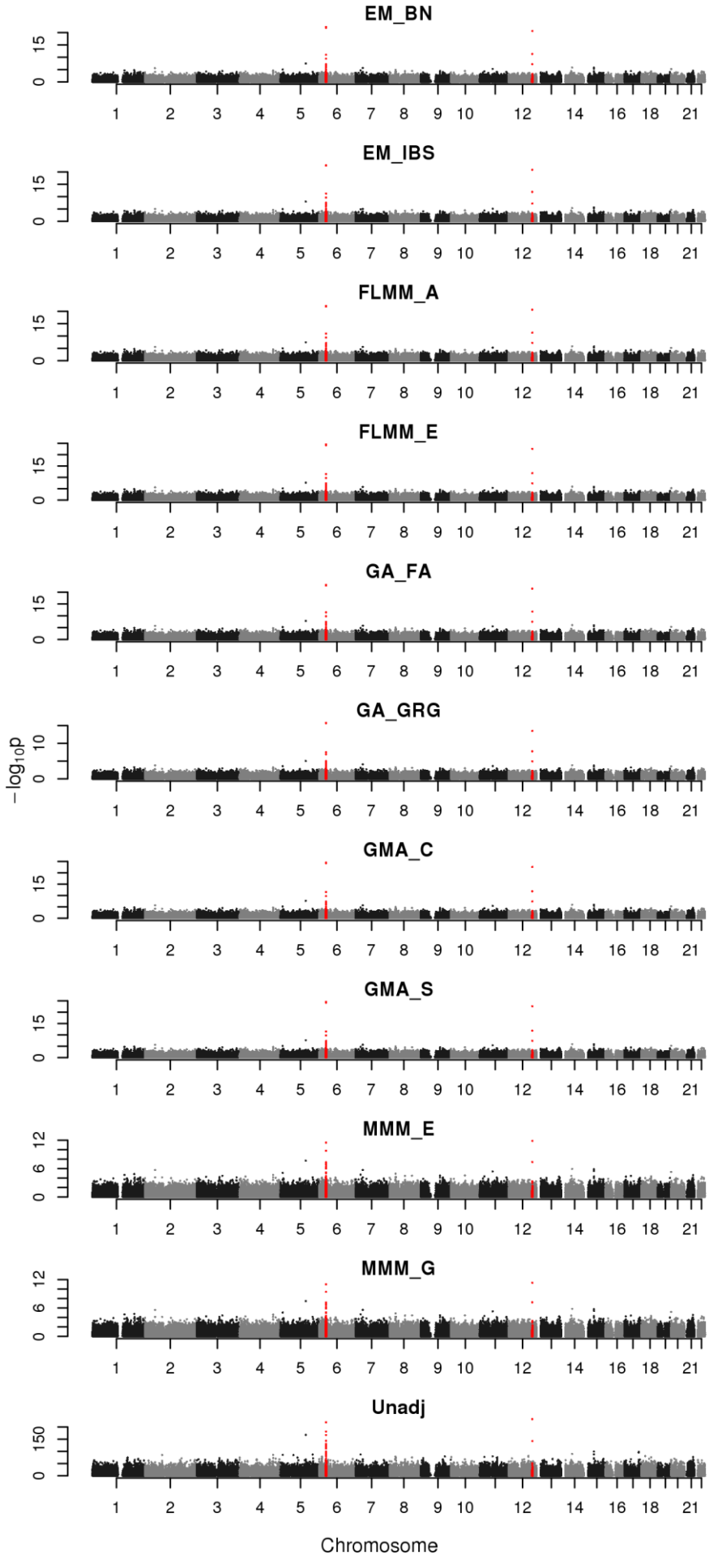
EM_BN
EM_IBS
FLMM_A
FLMM_E
GA_FA
GA_GRG
GMA_C
GMA_S
MMM_E
MMM_G
Unadj
−log10p
Chromosome
1 2 3 4 5 6 7 8 9 10 12 14 16 18 21

Figure 5.27 Manhattan plots for VL data set with simulated longitudinal qualitative phenotype using various LMM methods. The points marked in red (appear as dark grey area near the beginning of chromosome 6 and the end of chromosome 12 if printed in black and white) denote the simulated strong effect loci. EM_BN = EMMAX (Balding- Nichols), EM_IBS = EMMAX (IBS method), FLMM_A = FaST-LMM using approximate calculation with RRM, FLMM_E = FaST-LMM using exact calculation with RRM, GA_FA = GenABEL (FASTA), GA_GRG = GenABEL (GRAMMAR-Gamma), GMA_C = GEMMA using centred genotypes, GMA_S = GEMMA using standardised genotypes, MMM_E = MMM using full mixed model (exact) calculation, MMM_G = MMM using GLS approximation, Unadj = unadjusted analysis.

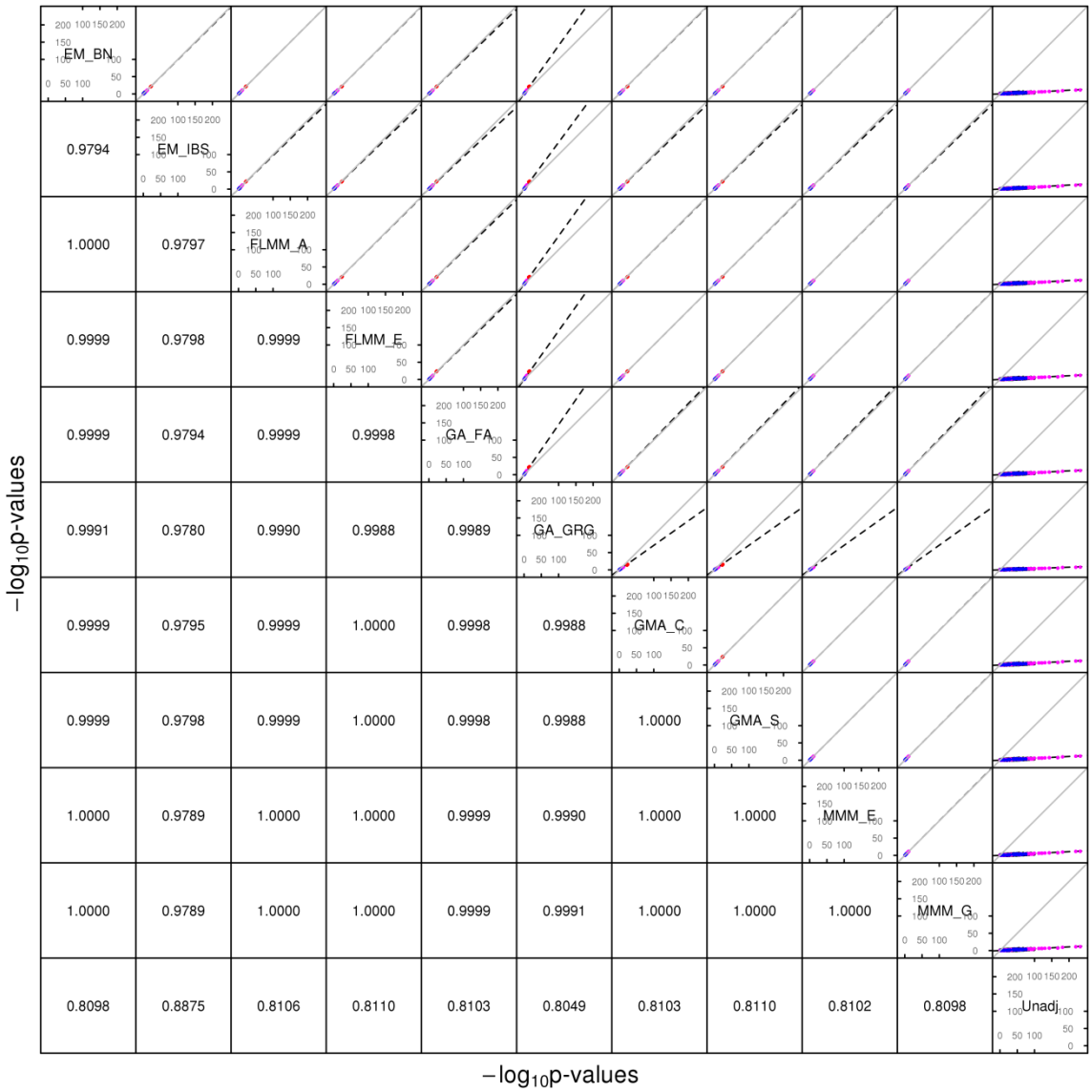

Figure 5.28 Comparison of $-\log_{10}$(p-values) using various LMM software packages, simulated longitudinal quantitative phenotype. Plots above the diagonal show a comparison of $-\log_{10}$(p-values), with correlations between the $-\log_{10}$(p-values) indicated below the diagonal. The grey solid lines represent the line of equality; the black dashed lines the linear regression line of the variable on the y axis on the variable on the x axis. The colours denote: red = the two strong effect SNPs, magenta = SNPs within 2 Mb of the strong effect SNPs, blue = 22 polygenic SNPs, green = SNPs within 2 Mb of the polygenic SNPs, black = all other SNPs. Because the black/green/blue SNPs were plotted before the magenta/red SNPs, they may be obscured by the latter. EM_BN = EMMAX

(Balding- Nichols), EM_IBS = EMMAX (IBS method), FLMM_A = FaST-LMM using approximate calculation with RRM, FLMM_E = FaST-LMM using exact calculation with RRM, GA_FA = GenABEL (FASTA), GA_GRG = GenABEL (GRAMMAR-Gamma), GMA_C = GEMMA using centred genotypes, GMA_S = GEMMA using standardised genotypes, MMM_E = MMM using full mixed model (exact) calculation, MMM_G = MMM using GLS approximation, Unadj = unadjusted analysis.

Because the longitudinal polygenic simulation (sim-P20) was constructed using a smaller effect size over a very large number of SNPs, the resulting phenotypes were expected to be highly correlated among family members, but much less so at any individual SNP. This is in contrast to the above longitudinal simulation (sim-L20) where the effects were spread over a more limited number of SNPs, including two with strong effects. The use of a larger number of SNPs would also mean that the distribution of the phenotypes more closely follows the genetic relatedness of the samples, and also is influenced by more distant relatedness and population substructure.

Despite these differences, a rather similar (but more extreme) pattern was also observed in this simulation (Figures 5.29-5.30): EMMAX (Balding-Nichols), FaST-LMM and GEMMA were again successful in maintaining the genomic inflation factor to about 1, whilst GenABEL (FASTA) showed even stronger inflation ($\lambda = 2.39$), and the deflation in GenABEL (GRAMMAR-Gamma) worsened ($\lambda = 0.47$), compared with the genomic inflation factor of 21.53 in the unadjusted analysis. However, some differences to the sim-L20 results were also noted: in this simulation, both MMM methods resulted in strong inflation, even exceeding that of GenABEL (FASTA) ($\lambda = 3.52$ compared with 2.39), whereas EMMAX (IBS) now showed a slight deflation ($\lambda = 0.97$), perhaps similar to that observed in the analysis of GAW18 data (Section 3.2).

Comparison of the concordance in $-\log_{10}$ p-values achieved by different methods (Figure 5.32) again showed high correlation among different methods, although the actual p-values were again different.

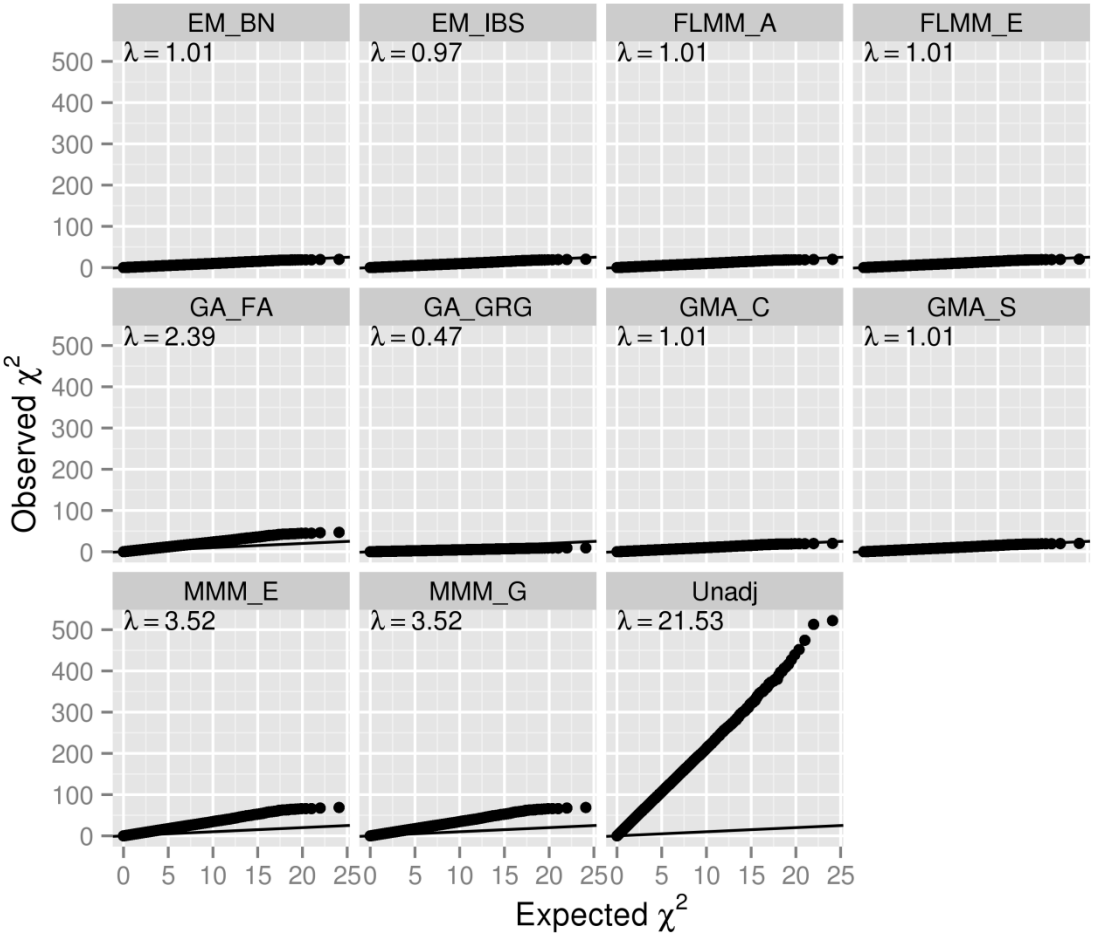

Figure 5.29 Q-Q plots of simulated longitudinal polygenic phenotype GWAS results and genomic inflation factors (λ) for different LMM methods. EM_BN = EMMAX (Balding- Nichols), EM_IBS = EMMAX (IBS method), FLMM_A = FaST-LMM using approximate calculation with RRM, FLMM_E = FaST-LMM using exact calculation with RRM, GA_FA = GenABEL (FASTA), GA_GRG = GenABEL (GRAMMAR-Gamma), GMA_C = GEMMA using centred genotypes, GMA_S = GEMMA using standardised genotypes, MMM_E = MMM using full mixed model (exact) calculation, MMM_G = MMM using GLS approximation, Unadj = unadjusted analysis.

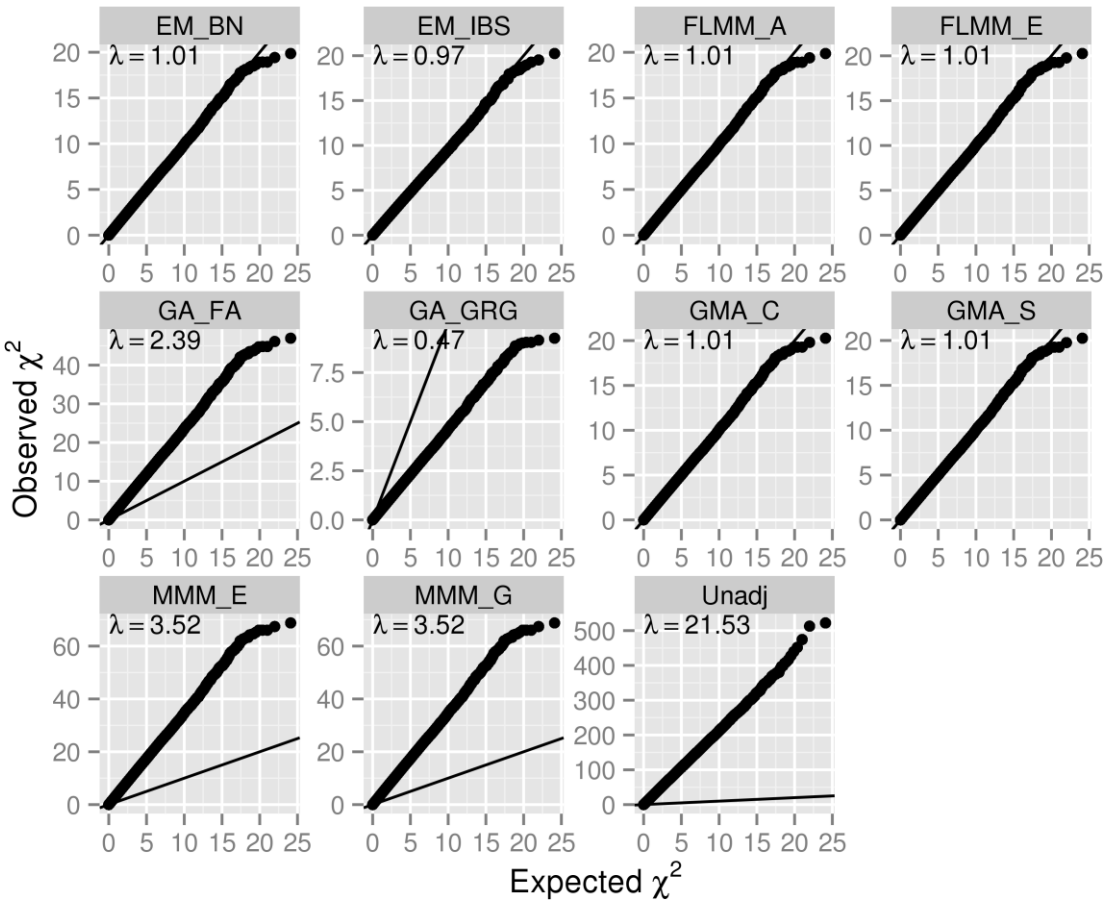

Figure 5.30 Q-Q plots of simulated longitudinal polygenic phenotype GWAS results and genomic inflation factors (λ) for different LMM methods, with each panel plotted on its own scale.
EM_BN = EMMAX (Balding- Nichols), EM_IBS = EMMAX (IBS method), FLMM_A = FaST-LMM using approximate calculation with RRM, FLMM_E = FaST-LMM using exact calculation with RRM, GA_FA = GenABEL (FASTA), GA_GRG = GenABEL (GRAMMAR-Gamma), GMA_C = GEMMA using centred genotypes, GMA_S = GEMMA using standardised genotypes, MMM_E = MMM using full mixed model (exact) calculation, MMM_G = MMM using GLS approximation, Unadj = unadjusted analysis. Unlike the previous plot, each panel in this plot has its own y-axis scale to better depict the distribution within its own panel.

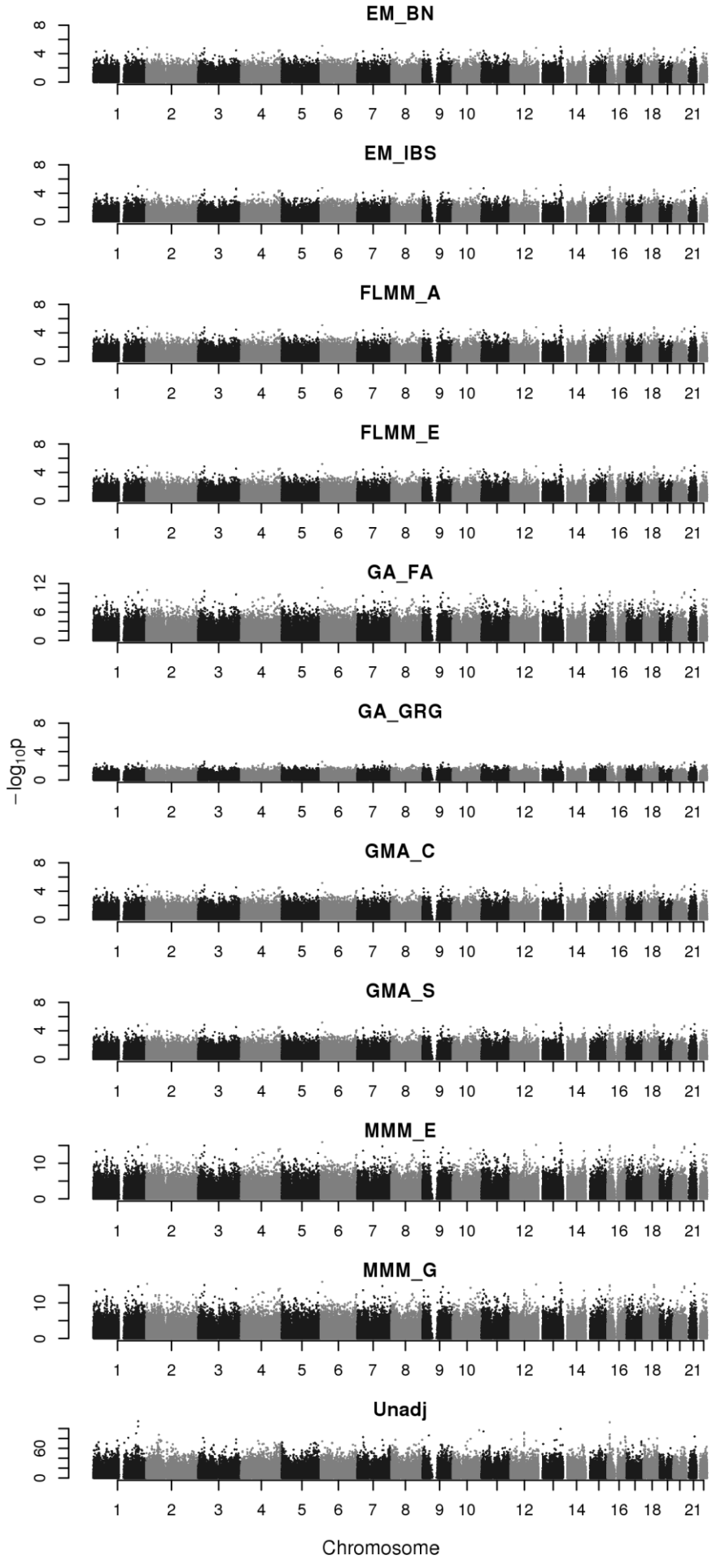
EM_BN
EM_IBS
FLMM_A
FLMM_E
GA_FA
GA_GRG
GMA_C
GMA_S
MMM_E
MMM_G
Unadj
−log10p
Chromosome
1 2 3 4 5 6 7 8 9 10 12 14 16 18 21

Figure 5.31 Manhattan plots for VL data set with simulated longitudinal polygenic phenotype using various LMM methods. EM_BN = EMMAX (Balding- Nichols), EM_IBS = EMMAX (IBS method), FLMM_A = FaST-LMM using approximate calculation with RRM, FLMM_E = FaST-LMM using exact calculation with RRM, GA_FA = GenABEL (FASTA), GA_GRG = GenABEL (GRAMMAR-Gamma), GMA_C = GEMMA using centred genotypes, GMA_S = GEMMA using standardised genotypes, MMM_E = MMM using full mixed model (exact) calculation, MMM_G = MMM using GLS approximation, Unadj = unadjusted analysis.

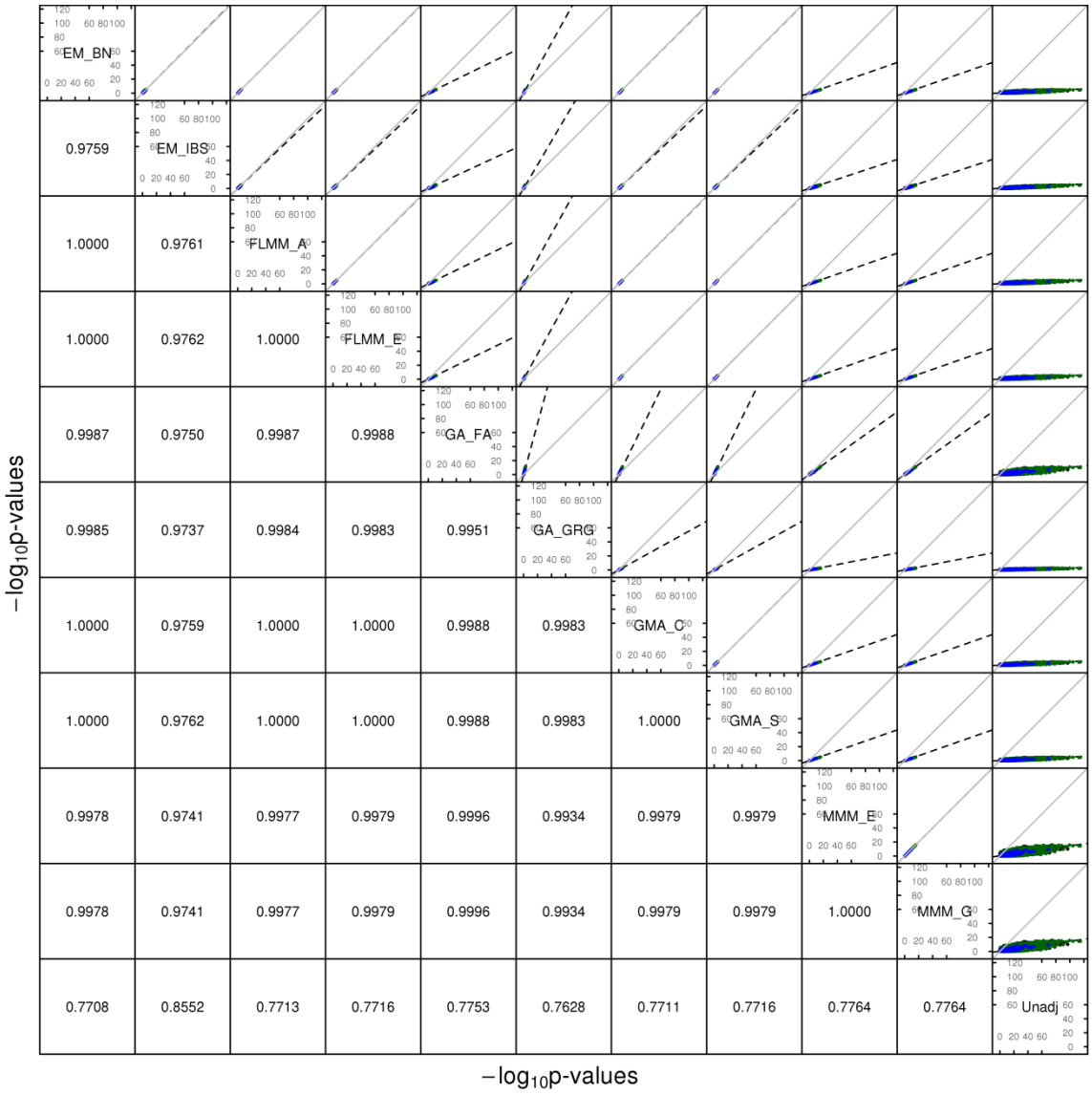

Figure 5.32 Comparison of -log(p-values) using various LMM software packages, simulated longitudinal quantitative phenotype. Plots above the diagonal show a comparison of $-\log_{10}$(p-values), with correlations between the $-\log_{10}$(p-values) indicated below the diagonal. The grey solid lines represent the line of equality; the black dashed lines the linear regression line of the variable on the y axis on the variable on the x axis. The colours denote: blue = 402 polygenic SNPs, green = SNPs within 2 Mb of the polygenic SNPs, black = all other SNPs. EM_BN = EMMAX (Balding- Nichols), EM_IBS = EMMAX (IBS method), FLMM_A = FaST-LMM using approximate calculation with RRM, FLMM_E = FaST-LMM using exact calculation with RRM, GA_FA = GenABEL (FASTA), GA_GRG = GenABEL (GRAMMAR-Gamma), GMA_C = GEMMA using centred genotypes, GMA_S = GEMMA using standardised genotypes, MMM_E = MMM using full mixed model (exact) calculation, MMM_G = MMM using GLS approximation, Unadj = unadjusted analysis.

To investigate the performance of these naive analyses against a 'proper' longitudinal analysis, the results from the FaST-LMM (Exact) analysis was compared with those from the R software package longGWAS (Furlotte *et al.*, 2012), which allows an extra, within-individual variance component to be fitted, while also making use of a two-stage approach (similar to FASTA) with a linear time search algorithm to estimate the components in the first stage so that the calculations can finish in a reasonable time for GWAS analysis. Results from both types of analysis appeared to be very similar (Figure 5.33), with longGWAS achieving the genomic inflation of 1.00 in the longitudinal phenotype, and 0.97 in the longitudinal polygenic phenotype. The marginal deflation in longGWAS's analysis of the longitudinal polygenic phenotype may in fact be in line with that observed in the EMMAX (IBS) analysis above.

Although the 'proper' analysis implemented in longGWAS may be considered theoretically most appealing, longGWAS was considerably slower than FaST-LMM, taking approximately 19 hours (in comparison to 5.5 minutes for FaST-LMM) when run in parallel for each of 22 chromosomes. If run as a single process (all chromosomes), this translates to about 9.5 days for longGWAS versus 7.6 hours for FaST-LMM. Thus, given the satisfactory performance of FaST-LMM, and the high correlation between the results obtained from FaST-LMM and those from longGWAS, from a practical point of view, FaST-LMM (or possibly EMMAX and GEMMA) would seem the more attractive option.

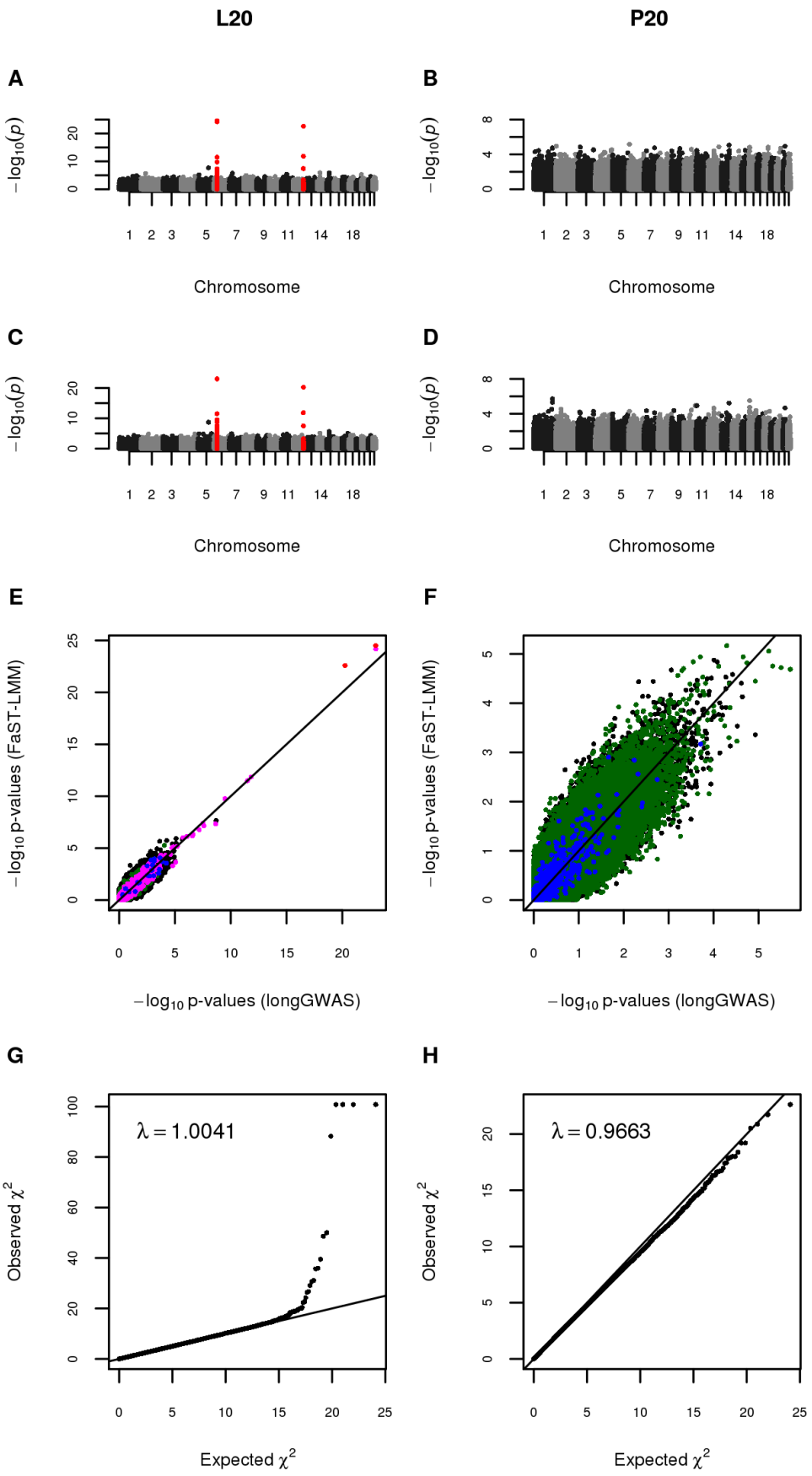

Figure 5.33 Comparison of results obtained from analyses of simulated longitudinal/longitudinal polygenic phenotypes using FaST-LMM (Exact) and longGWAS. A) Manhattan plot of results obtained from FaST-LMM (Exact) on longitudinal (sim-L20) phenotype, and B) on longitudinal polygenic (sim-P20) phenotype; C) Manhattan plot of results obtained from

longGWAS on longitudinal (sim-L20) phenotype, and D) on longitudinal polygenic (sim-P20) phenotype; E) comparison of $-\log_{10}$(p-values) using FaST-LMM and longGWAS on longitudinal (sim-L20) phenotype, and F) on longitudinal polygenic (sim-P20) phenotype; G) Q-Q plot of GWAS results and genomic inflation factor (λ) for longGWAS on longitudinal (sim-L20) phenotype, and H) on longitudinal polygenic (sim-P20) phenotype.

In addition to longGWAS, another program that can, in theory, implement a 'proper' longitudinal analysis is the lmekin function within the R package coxme (Therneau, 2012). In fact, lmekin is more generic than longGWAS as it can, in theory, handle any number of random effect components, which are not required to be polygenic or individual effects. The disadvantage of lmekin when applied to GWAS analysis is that, because it was designed as a generic mixed model method, it does not implement any speed up algorithm which can be found in methods designed for GWAS analysis. For this reason, it was found to be computationally infeasible for analysis of genome-wide data. For the purpose of comparing lmekin to other methods, a set of 2,423 SNPs of different effect sizes (2 strong/polygenic SNPs, 22 additional sim-L20 polygenic SNPs, 400 additional sim-P20 polygenic SNPs and 1,999 randomly chosen null SNPs) was extracted from the longitudinal data set. Application of lmekin to this set of SNPs in the sim-L20 data suggested that the results were very similar to those obtained from GenABEL (FASTA), EMMAX (Balding-Nichols), FaST-LMM, GEMMA and MMM (Figures 5.34-5.35). However, it did not give meaningful results (most were "NA") when applied to the sim-P20 data.

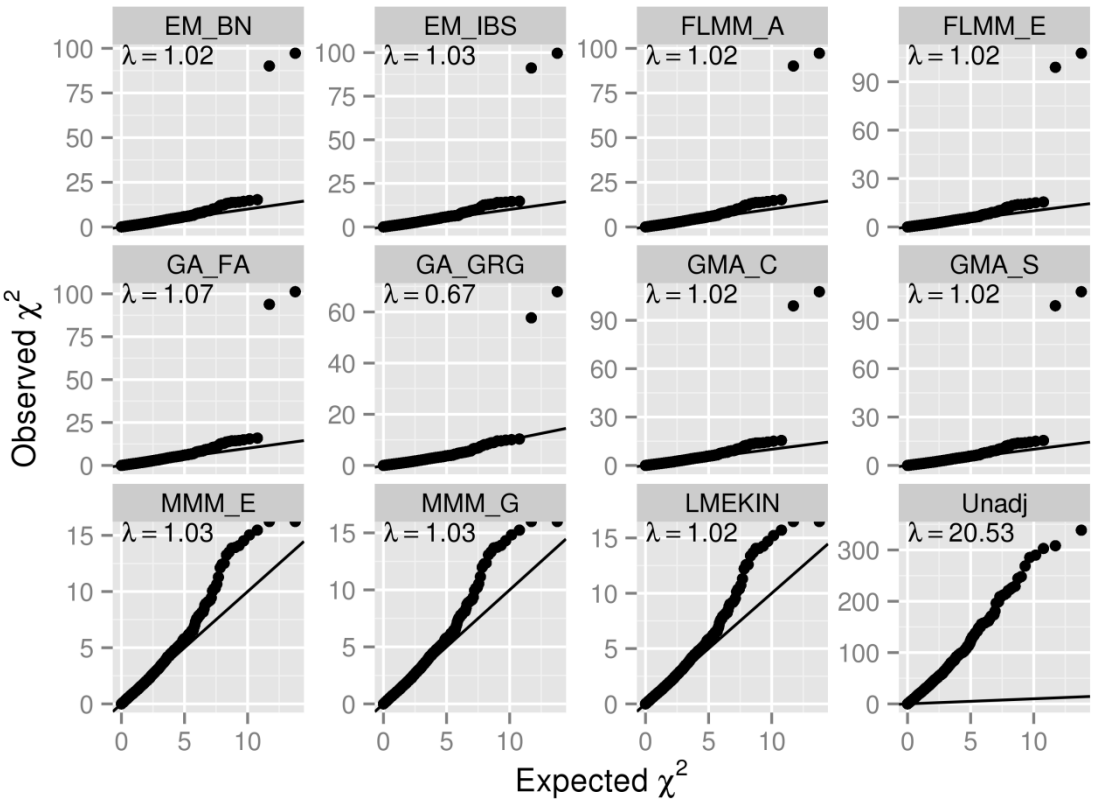

Figure 5.34 Q-Q plots of simulated longitudinal quantitative phenotype (sim-L20), restricted GWAS results and genomic inflation factors (λ) for different LMM methods compared with lmekin. EM_BN = EMMAX (Balding- Nichols), EM_IBS = EMMAX (IBS method), FLMM_A = FaST-LMM using approximate calculation with RRM, FLMM_E = FaST-LMM using exact calculation with RRM, GA_FA = GenABEL (FASTA), GA_GRG = GenABEL (GRAMMAR-Gamma), GMA_C = GEMMA using centred genotypes, GMA_S = GEMMA using standardised genotypes, MMM_E = MMM using full mixed model (exact) calculation, MMM_G = MMM using GLS approximation, Unadj = unadjusted analysis. The dots at the upper border of the MMM and LMEKIN panels represent the SNPs where the equivalent χ^2 values are ∞ (i.e. p-value = 0). The apparent genomic inflation factor in this case would be higher than usual due to the way the subset of 2,423 SNPs were chosen for analysis.

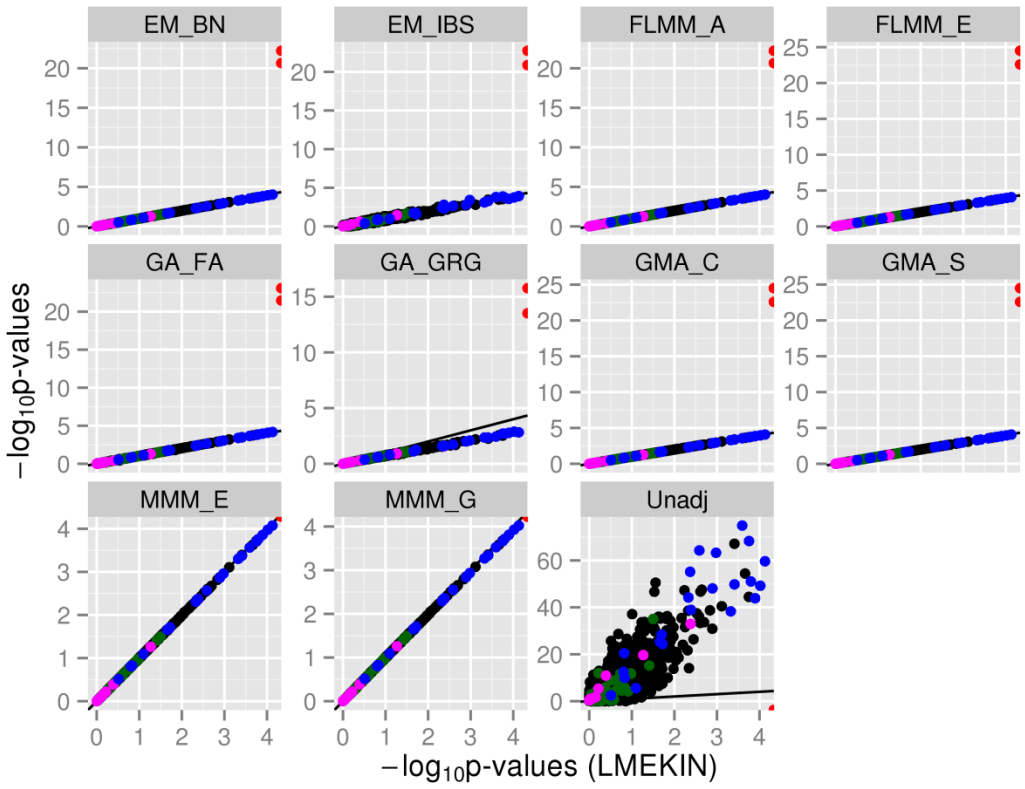

Figure 5.35 Comparison of $-\log_{10}$(p-values) from lmekin and various LMM software packages when applied to simulated longitudinal quantitative phenotype, on a restricted set of 4,323 SNPs. The black solid lines represent the line of equality. The colours denote: red = the two strong effect SNPs, magenta = SNPs within 2 Mb of the strong effect SNPs, blue = 22 polygenic SNPs, green = SNPs within 2 Mb of the polygenic SNPs, black = all other SNPs. Because the black/green/blue SNPs were plotted before the magenta/red SNPs, they may be obscured by the latter. EM_BN = EMMAX (Balding- Nichols), EM_IBS = EMMAX (IBS method), FLMM_A = FaST-LMM using approximate calculation with RRM, FLMM_E = FaST-LMM using exact calculation with RRM, GA_FA = GenABEL (FASTA), GA_GRG = GenABEL (GRAMMAR-Gamma), GMA_C = GEMMA using centred genotypes, GMA_S = GEMMA using standardised genotypes, MMM_E = MMM using full mixed model (exact) calculation, MMM_G = MMM using GLS approximation, Unadj = unadjusted analysis. Dots at the borders of the panels represent the SNPs where the $-\log_{10}$(p-values) are ∞ (i.e. p-value = 0), or in the unadjusted analysis, where the given p-values were ∞ (therefore their $-\log_{10}$ values were $-\infty$). Note that this means the red 'strong effect' dots are really much further to the right than were plotted here.

5.5. Computational Efficiency and Ease-of-use

Given that many of the software implementations investigated, and in particular all the various LMM implementations, showed similar levels of power and type 1 error, and gave rather similar inference in terms of localisation of signals and $-\log_{10}$ p-values achieved, an important practical consideration when deciding what implementation to use is the ease-of-use and computational efficiency. Ease-of-use is necessarily somewhat subjective as it depends on a user's prior experience and software/operating system preferences. Computational efficiency can, in theory, be examined more objectively. However, in practice, the total time required to perform an analysis is

dependent on the computer architecture available (in particular the ability of the system and of any given program to allow multithreading), demands of competing users and the availability of (and ability of any given program to make use of) facilities for parallel processing e.g. a multi-node compute cluster. These considerations make it hard to perform a genuine 'head-to-head' comparison between different packages. Table 5.4 presents an approximate comparison (carried out on the same machine, without use of parallel (i.e. multi-node) processing, but with multithreading allowed if native to that program) together with some comments concerning ease-of-use. Since PLINK (Purcell *et al.*, 2007) is commonly used to perform initial quality control of genome-wide association data, programs that could use PLINK files, either directly or with just a few easily-implemented transformation steps, were considered to be the easiest to use, while programs that required more extensive data transformation, creation of additional input files and/or external estimation of kinships were considered harder.

With respect to computational speed, as a rule of thumb, FaST-LMM (approximate) and GenABEL (GRAMMAR-Gamma) were found to be the fastest LMM implementations, taking between 2 minutes and a quarter of an hour to analyse 545,433 SNPs in 1,972 genotyped individuals. These were closely followed by EMMAX and MMM (approximate) which took around half an hour, GenABEL (FASTA), GEMMA, FaST-MMM (Exact) and MMM (Exact) which typically took 1-2 hours, Mendel (estimated kinships) which took around 2.5 hours (but see footnote of Table 5.4 and discussion), and GTAM which took around 4 hours. Of the non-LMM methods, FBAT, MQLS and MASTOR were the fastest, taking a few hours to perform the analysis, while ROADTRIPS was the slowest, taking several days.

Although slightly slower than FaST-LMM in absolute terms, it should be pointed out that GenABEL (GRAMMAR-Gamma) is a single-threaded application whereas FaST-LMM is natively multithreaded and, in this measurement, ran on all 8 CPU cores. This means that on a single core system, GenABEL (GRAMMAR-Gamma) may be marginally quicker than FaST-LMM.

The fastest LMM methods were all approximate. In practice, it should not matter if an exact or approximation method was used as the results should be very similar. However, if an exact LMM calculation is required, then GEMMA could potentially be the fastest program to run naively, i.e. without further user-enforced parallelisation.

Package/method	Time take to perform whole GWAS				Ease of use
	Data conversion from PLINK	Kinship calculation	Association analysis	Total	
EMMAX (BN)*	8m 19s	38s	14m 40s	23m 37s	Easy
EMMAX (IBS)*	8m 19s	43s	14m 04s	23m 06s	Easy
FaST-LMM (Approx)*		(7-9s)	14m 15s	14m 23s (2m 14s†)	Easy
FaST-LMM (Exact)*		(7-9s)	1h 53m 52s	1h 54m 00s (1h 51m 18s†)	Easy
GEMMA (GMA_C)		2m 49s	1h 06m 54s	1h 09m 43s	Easy
GEMMA (GMA_S)		2m 48s	1h 06m 54s	1h 09m 42s	Easy
GenABEL (FASTA)	4m 25s	11m 44s	41m 05s	57m 14s	Requires familiarity with R
GenABEL (GRAMMAR-Gamma)	4m 25s	11m 44s	25s	16m 34s	Requires familiarity with R
Mendel (Estimated kinships)‡				2h 27m 02s‡	Medium
MMM (Approx)	18m 01s	5m 31s	29m 33s	35m 05s	Medium
MMM (Exact)	18m 01s	5m 06s	1h 17m 24s	1h 40m 31s	Medium
FBAT (Affected only)	25m		1h 11m	1h 36m	Medium
FBAT (Both)	25m		1h 22m	1h 47m	Medium
GTAM (implemented in MASTOR v0.3)	Varies		3h 59m	3h 59m +conversion	File conversion fiddly
MASTOR	Varies		1h 02m	1h 02m +conversion	File conversion fiddly
MQLS (1972)	14m		26m	40m	Medium
MQLS (3626)	25m		36m	1h 01m	Medium
ROADTRIPS (1972)	Varies		15h 36m	15h 36m +conversion	File conversion fiddly
ROADTRIPS (3626)	Varies		39h 01m	39h 01m +conversion	File conversion fiddly

Table 5.4 Computational speed and ease of use of various packages in analysis GWAS data consisting of 545,433 SNPs in 1,972 individuals.

* These programs are either documented to be multithreaded (FaST-LMM) or observed to be multithreaded (EMMAX). FaST-LMM appeared to run single-threaded when using exact calculation.

† These numbers represent the total run time required in the FaST-LMM's default 'run-through' mode, in which kinship estimation and GWAS calculation were performed in a single run. There appeared to be substantial time saving, particularly for approximate calculation, compared with doing these in two separate steps, probably because it can use the genotype data directly without having to calculate the kinship matrix first.

‡ A new version of Mendel was released after the publication of the article describing this part of the thesis, which allows multithreading and also substantially improves the calculation efficiency. A comparable analysis in the new version of Mendel would now take 6 minutes and 38 seconds.

It is possible to further speed up, at least in theory, the computation in most LMM software packages by imposing another layer of parallelisation. Due to the two-stage nature of most programs, the kinship matrix calculation (or spectral decomposition of the genotype matrix), which is normally implemented in a way that would require a substantial effort to parallelise, needs to be performed only once, and its products can then be fed into the LMM analysis step, which can often be highly parallelised (depending on the computational resources available). However, the ease of parallelisation in the second stage also varies among different programs. FaST-LMM seems to facilitate this best by providing an option to automatically split the calculation into a specified number of tasks (through the specification of the task index and the total number of the tasks required, which can be easily implemented as an array job in the cluster environment), whereas GenABEL (both FASTA and GRAMMAR-Gamma) allows specification of SNPs to be used in each calculation, which still needs to be determined by the user (or the script). Neither EMMAX, GEMMA nor MMM provides a means to do this, and any attempt to parallelise these programs would require direct extraction of the required set of SNPs for each task, which could be time-consuming and likely to negate the benefit of parallelisation.

When parallelised, exact analysis of the equivalent data set to the above can typically be achieved by FaST-LMM in about 10 minutes. Exact analysis of a large longitudinal data set (equivalent to 19,720 individuals (not presented in this thesis)), which would normally take about a day for most programs, took just a few hours even in presence of moderate cluster load (the equivalent parallelised approximate analysis took about 50 minutes). This makes FaST-LMM, when optimally parallelised, the fastest exact LMM method in absolute terms.

5.6. Discussion

In general, all LMM programs were successful in controlling the inflation due to sample relatedness and gave very similar results in most simulations apart from the longitudinal simulations.

Analysing each repeated measure as if it comes from a different individual treats the data set as a larger 'pseudo data set' containing many apparent twins/triplets/quadruplets (or, in this case, vigintuplets (20-tuplets)). Although less satisfactory than a proper longitudinal analysis that takes into account correlations due to both relatedness between individuals and repeated measures within individuals (Furlotte *et al.*, 2012; Therneau, 2012), the LMM framework should intuitively be able to absorb the effect of repeated measures within individuals into the genetic component of variance estimated, resulting in an overall correct distribution of test statistics. For EMMAX (especially when using Balding-Nichols matrix), FaST-LMM and GEMMA,

this intuition appears to be correct. Although for GenABEL (FASTA) and MMM the resulting distribution of test statistics is inflated, the linear relationship between the observed and desired test statistics means that test statistics following the desired distribution could be obtained simply by dividing the observed χ^2 test statistics by the observed genomic control inflation factor, in an approach akin to standard genomic control (Devlin and Roeder, 1999). Similarly, for GenABEL (GRAMMAR-Gamma) which showed gross deflation in a similar manner to that found by Zhou and Stephens (2012) in their highly-related mouse data set, re-inflating the results with the observed genomic control inflation factor may also yield the desired distribution. (Of note, results from the GRAMMAR-Gamma method have actually been re-inflated once using the gamma correction factor. It may be that, in presence of repeated measurements in this data set, correction by gamma factor alone is not adequate.)

An interesting conclusion here, in common with the finding from Chapter 3, was that longitudinal traits (repeated measures) could be successfully analysed in an LMM framework simply by treating each measurement as if it came from a separate person and expanding out the genetic data set accordingly (resulting in an expanded data set containing many apparent twins, triplets, quadruplets etc). This led to the conclusion in our article describing this part of the thesis (Eu-ahsunthornwattana *et al.*, 2014b) that, from a practical point of view, this strategy is (or was) useful: analysis of an expanded data set in standard LMM software is computationally convenient, while a 'proper' analysis using software such as longGWAS (Furlotte *et al.*, 2012) or lmekin (Therneau, 2012) tends to be prohibitively slow (if at all feasible) when applied to this data set.

That said, this may no longer be the case, since while our article was being published, a new version (version 14.2+) of Mendel (K. Lange *et al.*, 2013) was released, which implements a more computationally efficient, parallelised version of LMM analysis, and also allows an additional variance component to be added to the analysis. Longitudinal data can therefore be directly modelled in this version using the extra variance component, and can potentially be analysed in a reasonable amount of time, making the approximation for longitudinal data analysis used here obsolete, unless these naive methods are significantly more resource-efficient than full analysis in Mendel, while still giving reasonably accurate results. Any conclusion in this regard cannot be made from this thesis, and could be a topic for further exploration.

Another feature which has been demonstrated in this chapter was the success of using estimated kinships to adjust for the inflation due to sample relatedness. Although the results in this chapter appeared to suggest that this is equivalent to using theoretical kinships, this seems to be due to the lack of the more distant relatedness/population structure in the simulation, and, taking into account the results from the previous

chapters, the conclusion should be that using estimated kinships is at least as good as using theoretical kinships.

Although all the results presented in this and the previous chapter relate to genotypes derived from a single data set (Fakiola *et al.*, 2013), high concordance between different LMM implementations seen here, as well as their performance from when applied naively to longitudinal data, should hold more generally for genetic studies of diverse phenotypes carried out in diverse human populations. Essentially the same pattern of results described here was observed when a more limited set of LMM implementations were applied to GWAS data from Genetic Analysis Workshop 18, as described in Chapter 3, and also when these approaches were applied to GWAS data from 402 Aboriginal Australian individuals that cluster loosely into 4 large nominal pedigrees (D. Anderson *et al.*, 2015). Therefore, although it is possible that highly structured populations such as those encountered in plant or animal breeding experiments may uncover subtle differences between the various LMM approaches, little difference is expected between the results seen from one approach over another for researchers carrying out complex genetic disease studies in human populations, and the choice of which method/software package to use is likely to be dictated by personal taste and convenience.

On this note, it should be pointed out that each package has its own particular advantages (as well as disadvantages). These include the ability of EMMAX, GEMMA and MMM to read in the dosages derived from imputed (in addition to real) genotypes; MMM has the advantage of allowing the output of regression coefficients and standard errors for the SNP effects on the (log) odds ratio scale, making it convenient to compare or combine the results with results from traditional case/control studies analysed via logistic regression; GenABEL (GRAMMAR-Gamma) has the advantage of scaling linearly with sample size, which makes it attractive for the analysis of very large data sets; FaST-LMM has the advantage, along with EMMAX and Mendel, of internally imputing missing data at any (genetic or non-genetic) covariates, which can make it convenient for implementing stepwise conditional analyses; and, unlike most LMM implementations, ROADTRIPS, MQLS and MASTOR have the advantage of using all phenotype information, including that for individuals that have not been genotyped, which can in theory generate a small increase in power.

One of the main differences between the different software implementations investigated was the time taken to perform the analysis (not including the time required to re-format data into an appropriate format for a given package). Although care was taken to measure the programs run time in similar circumstance (see Section 2.8), various factors that could not be totally controlled means that this was not a strict

head-to-head comparison. However, rough comparison in Section 5.5, assuming that kinships are to be estimated on the basis of SNP data, implicated FaST-LMM (approximate calculation) GenABEL (GRAMMAR-Gamma), EMMAX and Mendel as generally the fastest implementations.

In conclusion, linear mixed model approaches are convenient and powerful for family-based GWAS of quantitative or binary traits. They are successful in controlling the overall genome-wide error rate and perform well in comparison to competing approaches.

Chapter 6. Application of Genomic IBD Estimates in Non-parametric Linkage Analyses of the Brazilian Visceral Leishmaniasis Data

This chapter will continue with the overall theme of this thesis by investigating the use of genetically estimated IBD in non-parametric linkage analysis. A new method, Regional IBD Analysis (RIA) is proposed, and will be compared with methods implemented in other standard non-parametric analysis software.

6.1. Statistical Methods and Software

6.1.1. Regional IBD Analysis (RIA)

Following the core principle of comparing the observed and expected IBD sharing patterns in non-parametric linkage analysis, a new method, 'Regional IBD Analysis' ('RIA'), is proposed here. This method uses genetically estimated IBD sharing probabilities instead of the theoretical estimates, which should eliminate the aforementioned (Section 1.3) problems of IBD estimation in large, complex pedigrees, as well as potentially allowing the analysis to be extended to apparently unrelated individuals from different families (if so desired, but see also discussion in Section 6.5). This can be implemented as a two-stage approach using readily available software packages.

The first stage of RIA is the estimation of the IBD sharing probabilities, currently implemented using either PLINK (Purcell *et al.*, 2007) or KING (Manichaikul *et al.*, 2010). The estimated IBD probabilities are then fed into the second stage program for calculation of the non-parametric linkage statistic.

In theory, the second stage of RIA could be any non-parametric linkage analysis program that does not assume the expected IBD sharing probabilities to follow a pre-defined pattern (as is the case in the affected-sib-pair method); but in practice, most programs internally calculate theoretical IBD for use in their own analysis, thus precluding the implementation of RIA using those programs. The current implementation of RIA uses the program Onelocarp (Cordell *et al.*, 2000), which allows (in fact, requires) externally estimated IBD sharing probabilities in to be read in for analysis.

The methods for IBD estimation in PLINK and KING have been described in Chapter 4 (Section 4.1.3). Note, however, that only the 'homogeneous' estimations (KING-homo)

were used here as KING's robust method gives only IBS estimations. Onelocarp will be described in the next section.

6.1.2. Onelocarp

The actual program that calculates the non-parametric linkage statistic in RIA is Onelocarp, which was originally part of a program package that accompanied an article describing a multilocus affected relative pair linkage analysis method (Cordell *et al.*, 2000); however, unlike the two other programs in that package, Twolocarp and Threelocarp, Onelocarp was designed for single locus analysis (which is just a special case of the proposed multilocus method).

All programs in the package require externally calculated 'prior' and 'posterior' IBD sharing probabilities to be read in for their analyses. The prior IBD sharing probabilities are the expected probabilities of each IBD sharing state between each of the affected pairs of individuals based on their type of relatedness. In context of affected relative pair analysis, these are equivalent to the IBD sharing probabilities of the pair under the null hypothesis of no linkage, and were originally assumed to be derived from pedigree information. However, these can, in theory, be replaced with any other appropriate estimates. The posterior IBD sharing probabilities are the probabilities of each IBD sharing state between the pair, given the observed genotype at the locus (or loci) of interest. The estimation of these traditionally requires pedigree knowledge, but again can be replaced with other appropriate estimates.

Provided with the externally estimated prior and posterior IBD probabilities, Onelocarp calculates a non-parametric maximum-likelihood statistic (MLS)-like test of linkage of the form (Cordell *et al.*, 2000):

$$\mathrm{MLS} = \sum_{j} \log_{10}\left(\sum_{i=0}^{2} \frac{\hat{z}_{ij}\hat{f}_{ij}}{f_{ij}}\right)$$

where $\hat{z}_{ij}$ is population parameter (to be estimated) corresponding to the probability that an ARP of the same type as pair j shares i allele(s) IBD at that locus, $\hat{f}_{ij}$ is the posterior probability that pair j shares i allele(s) IBD at that locus given the observed marker data and f_{ij} the prior probability that pair j shares i allele(s) IBD.

Internally, Onelocarp simplifies the estimation of $\hat{z}_{ij}$, which is specific to each type of ARP, by instead parameterising in terms of overall additive and dominance variances which need to be estimated only once, given that the population prevalence of the disease is specified (see Cordell *et al.* (2000) for details). Additionally, if the effect is assumed to be purely additive, the dominance variance can *a priori* be set to zero, thus further simplifying the calculation.

The version of Onelocarp used here was slightly modified by my supervisor (HJC) from the original version to accommodate the use of empirical IBD estimates in the following ways:

1. Onelocarp handles the situation where the prior probability of any IBD sharing state i is 0 by setting the corresponding likelihood term $\hat{z}_{ij}\hat{f}_{ij}/f_{ij}$ to 0, and uses the remaining terms for MLS calculation. Since the theoretical posterior IBD probability will also be 0 when the prior is 0, this procedure was reasonable and worked successfully in the original version of Onelocarp. However, with genetically estimated IBD, the posterior IBD probabilities are now computationally independent of the prior probabilities, and may not necessarily be 0 when their corresponding prior probabilities are 0. The original procedure would then lead to a situation where the sum of posterior probabilities in the remaining likelihood terms is not 1, which results in incorrect MLS estimation. The procedure implemented in the modified version of Onelocarp is that, when it encounters a 0 prior, Onelocarp will set the corresponding likelihood term to 0 and also rescale the posterior probabilities in the remaining terms so that their sum is 1.
2. The maximum numbers of affected relative pairs and markers that can be analysed in a single run have been increased from 2,000 pairs and 300 markers to 80,000 pairs and 600 markers, respectively.

A variant of Onelocarp ('Onelocarp-ndv', for 'no dominance variance') was also created. In addition to the above modifications, this program fixes the dominance variance to 0 (so the genetic effect is purely additive). This is a reasonable assumption in complex diseases, and can speed up the calculation further. The performance of RIA was investigated here using both Onelocarp and Onelocarp-ndv (designated 'RIA' and 'RIA-ndv', respectively).

6.1.3. Global and local genomic IBD estimation

The IBD sharing probabilities used in RIA are estimated solely based on genomic data using either PLINK (Purcell *et al.*, 2007) or KING (Manichaikul *et al.*, 2010). Two types of estimation are used: 'global' and 'local' IBD estimates (Figure 6.1).

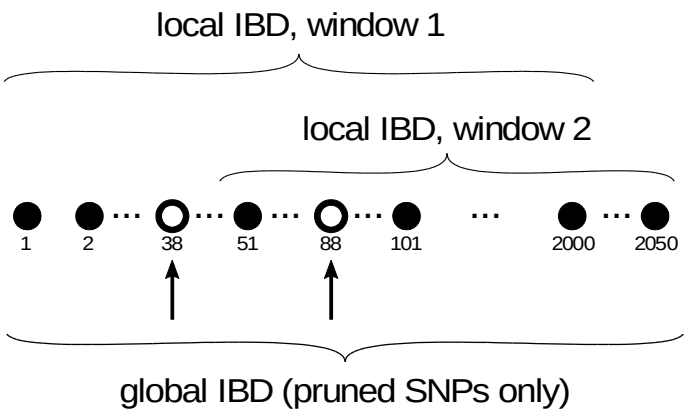

Figure 6.1 Example of how RIA selects SNPs for its IBD estimation. Circles (both black and white) represent SNPs in the data set. Number underneath each circle (SNP) denotes the order of that SNP in the data set for that chromosome. White circles (e.g. SNPs 38 and 88 here) are the pruned SNPs; only these are used in global IBD estimation (estimated from all available autosomes). In this example, the window size for local IBD estimation is 2,000 SNPs, moving 50 SNPs at a time.

The global IBD probabilities are estimated using genome-wide SNP data, which may have undergone a SNP reduction process (such as pruning or thinning) but still retain their genome-wide coverage. For both theoretical (due to independency among the SNPs) and practical reasons, pruned sets of SNPs were used to calculate the global IBD estimates in this thesis. Because of their global nature, the global IBD estimates reflect the overall degree of relatedness of each pair of individuals, and are suitable for use as the prior IBD probabilities in RIA.

The local IBD probabilities reflect the IBD sharing probabilities at the locus of interest and are used as the posterior probabilities in RIA. They are estimated using adjacent SNPs of a certain length ('window'), with the location of the window represented by its mid-point SNP. The window for IBD estimation moves along each chromosome at a certain pre-specified number of SNPs ('step'), except for the last window in each chromosome which may begin earlier than this so that the number of SNPs in it remains correct. Setting the step size less than the window size creates a series of overlapping windows on each chromosome.

The appropriate window size for local IBD estimation depends on the data set, particularly its number (or density) of SNPs, and I carried out a smaller scale trial run to optimise this. Empirically, a window size of 500 SNPs seems appropriate for a genome-wide data set of about 100,000 SNPs, and that of 2,000 SNPs seems appropriate for a data set of about 500,000-600,000 SNPs (see also Sections 6.2 and 6.4). The RIA analyses in this thesis therefore used a window size of 500 SNPs for the

all-sample VUR data set, and 2,000 SNPs for the Dublin-only VUR data set and for the VL data set. The step size was fixed at 50 SNPs in all analyses.

With a relatively small window size like this, it is possible that two individuals in a pair do not share any non-missing SNPs within a particular window. This can occur even in quality-controlled data set with reasonably low individual and marker missing rates. PLINK and KING respond differently to this.

When a pair which does not share any non-missing SNP is encountered during the pair-wise IBD estimation, PLINK immediately stops. This may in fact be the correct behaviour for its intended use (i.e. for data quality control, using genome-wide data for estimation), but causes a serious problem for RIA, as it means the remaining pairs in which IBD estimation should be possible do not have their IBD estimated, and this occurs at a rather unpredictable point in the calculation (in a sense that there is no fixed pattern in the data set; it only occurs as and when the condition is satisfied). To circumvent this problem, a dummy SNP was inserted into each local IBD window, with its value set to heterozygous (A/B) in all individuals. This allows PLINK to complete its calculation in all individuals, but could potentially introduce some bias into the analysis. The extent of this will be seen in the subsequent sections.

KING responds to this problem in a more desirable manner: it will just produce missing values for those pairs. In RIA analysis, any missing posterior probability is replaced by the corresponding prior probability, which would result in a slightly conservative analysis.

Unlike PLINK, the IBD probabilities from KING are not automatically constrained to biologically plausible ranges. This could also cause problem with Onelocarp, which, being originally written for pedigree-based data analysis, assumes that it will be given biologically valid probabilities. To prevent this issue, a simple constraining procedure was applied to KING's output: first constrain the kinship coefficient to the range of [0, 0.5] and the 0 IBD sharing probability to [0, 1] (these are the two output values from KING); then calculate the 1 and 2 IBD sharing probabilities from the constrained kinship coefficient and the 0 IBD probability; finally, constrain both 1 and 2 IBD sharing probabilities to [0, 1]. Incidentally, this leads to IBD probabilities that do not necessarily add to one. However, since the modified version of Onelocarp automatically rescales the IBD probabilities, no further adjustment is required here.

Another problem peculiar to KING is that, in its current implementation, it does not allow allele frequencies from external source to be used in its IBD estimation: the allele frequencies used in IBD estimation in KING is based solely on the data being fed in. This probably makes sense if one is also implementing a population-robust method, but

seems to be overly restrictive under homogeneity assumption, and could potentially lead to inaccuracies in RIA, where only genotype data from affected relative pairs are used. A possible method to circumvent this is to feed the full data set to KING, then select the IBD probabilities from only the relevant affected relative pairs. However, this is inefficient, less practical and, depending on the data set available, may not even be possible. In this thesis, the simple method of feeding the ARP data set to KING will be used.

6.1.4. Other linkage analysis software used

The results from RIA were compared to those from a set of traditional non-parametric linkage analysis methods, chosen to represent both the exact and simulation-based methods. These have different merits and are suitable for different situations, as described below.

Merlin

One approach to non-parametric linkage analysis is to view this as a test for excess IBD sharing (Kong and Cox, 1997). Whittemore and Halpern (1994) proposed two IBD scoring functions for use in affected relative pairs analysis: S_{pairs} and $S_{all.}$ The former is based on a simple count of allele pairs that are shared IBD in each affected relative pair, the latter takes into account the IBD sharing in all affected individuals within the same family (Whittemore and Halpern, 1994; Kruglyak *et al.*, 1996; Kong and Cox, 1997; McPeek, 1999). When these scores are standardised and weight-averaged across all pedigrees, the resulting statistic will be normally distributed with mean 0 and variance 1 under the null hypothesis of no linkage, given that the IBD sharing information is completely known. However, if the information on IBD sharing is incomplete, as would normally be the case, the test can become conservative (Kruglyak *et al.*, 1996; Kong and Cox, 1997; McPeek, 1999). (A side effect of this is that the test statistics tend to be lower *between* markers (since the information there will be less complete), which is in contrast to what would be expected in standard parametric linkage analysis (Cordell *et al.*, 2000)).

To correct for this behaviour, Kong and Cox (1997) proposed a class of likelihood ratio tests based on S_{pairs} and S_{all}, using either a linear or an exponential single-parameter likelihood model. Both of these can give LOD scores as well as a normally distributed Z_{lr} statistic, with the exponential model having an advantage over the linear model when the number of families is small and the excess IBD sharing is high, at the expense of higher computational cost.

Merlin (Abecasis *et al.*, 2002) uses sparse binary trees to allow the Lander-Green algorithm (Lander and Green, 1987) to estimate the inheritance vectors using a large

number of markers in relatively large families (compared with what can be achieved using the standard Lander-Green algorithm). These can then be used to produce several non-parametric linkage test statistics. Among these, the Kong and Cox exponential model LOD score (Kong and Cox, 1997) based on the S_{pairs} statistic was chosen here as it most resembles RIA's MLS method.

The Lander-Green algorithm, although capable of handling a large number of markers, is known to be resource-demanding and cannot handle large families (Lander and Green, 1987; Abecasis *et al.*, 2002; Albers *et al.*, 2008). The extension in Merlin relaxes this limitation, but in practice the family size that can be successfully analysed is still limited, as will be demonstrated.

MORGAN (lm_ibdtests)

The program lm_ibdtests (Basu *et al.*, 2010) in the software package MORGAN (http://www.stat.washington.edu/~thompson/Genepi/MORGAN/Morgan.shtml) uses a Markov chain Monte Carlo (MCMC) method to estimate the inheritance vectors in complex pedigrees, which can then be used for calculation of various IBD scoring statistics. The level of significance can be assessed using either by performing phenotype permutation or by using a standard normality assumption (Sieh *et al.*, 2005; Basu *et al.*, 2008; Basu *et al.*, 2010). For the purpose of comparison with RIA's results, the standardised S_{pair} statistics (Whittemore and Halpern, 1994; Kruglyak *et al.*, 1996) obtained under normality assumption was used here. (However, it should be noted that this is likely to be overly conservative, as described for the tests above, and under normal circumstances the permutation test is likely to be a better option.)

The use of MCMC sampling allows lm_ibdtests to handle complex pedigrees, but at the expense of not providing an exact calculation.

6.2. Comparison with Exact Non-parametric Linkage Analysis, Using a Pilot (VUR) Data Set

As a proof of concept before embarking on more complex analyses, RIA analyses using various IBD estimation methods were performed on the VUR data set. Because this is a data set of small nuclear families, aimed specifically for linkage analysis, it allows comparison with exact linkage analysis using Merlin. Additionally, the results were also compared with the standard transmission disequilibrium test (TDT) (Spielman *et al.*, 1993) as implemented in PLINK (Purcell *et al.*, 2007), and with LMM GWAS implemented in FaST-LMM (Lippert *et al.*, 2011).

Using all samples in this data set, the Manhattan plots (which, for genome-wide linkage analyses, tend to have much sparser data points than GWAS analyses, and function quite similarly to traditional linkage analysis plots) show that RIA was able to detect

linkage signals quite similar to those detected by Merlin (Figures 6.2 and 6.3). The results were especially similar when Merlin's theoretical IBD estimates were used in RIA, demonstrating the concordance of the MLS calculated using Onelocarp and the Kong and Cox exponential model LOD scores from Merlin. With RIA using genetically estimated IBD, the plots become more noisy. Although most of the 'true' signals (i.e. concordant with Merlin) could still be seen, particularly the top signals, these methods have also detected substantial 'extra' signals (which Merlin did not detect and could therefore potentially be false, but without the complete knowledge of the true effect locations, they could not be labelled as such with certainty). For example, in Figure 6.3 which focuses on chromosome 10, and contains the strongest signal peak in Merlin analysis, the peak toward the end of the chromosome detected by Merlin was also detected by all RIA methods; however, RIA using genetic IBD estimation also produced another peak before that, which was not detected by Merlin. Interestingly, this seems to coincide roughly with a (non-significant) peak seen in the TDT, as well as a rather extreme data point in FaST-LMM. Whether this is a genuine signal or whether RIA just gave a false signal is difficult to judge without complete knowledge of the genetic causality of VUR. There was no noticeable difference between RIA using IBD estimates from PLINK and KING (even with the problems discussed in the previous section), nor between the standard and the 'no dominance variance' ('NDV') versions of RIA. TDT and FaST-LMM results seem to be generally different from Merlin, although some of their peaks (or outlying points) seem to coincide with RIA's.

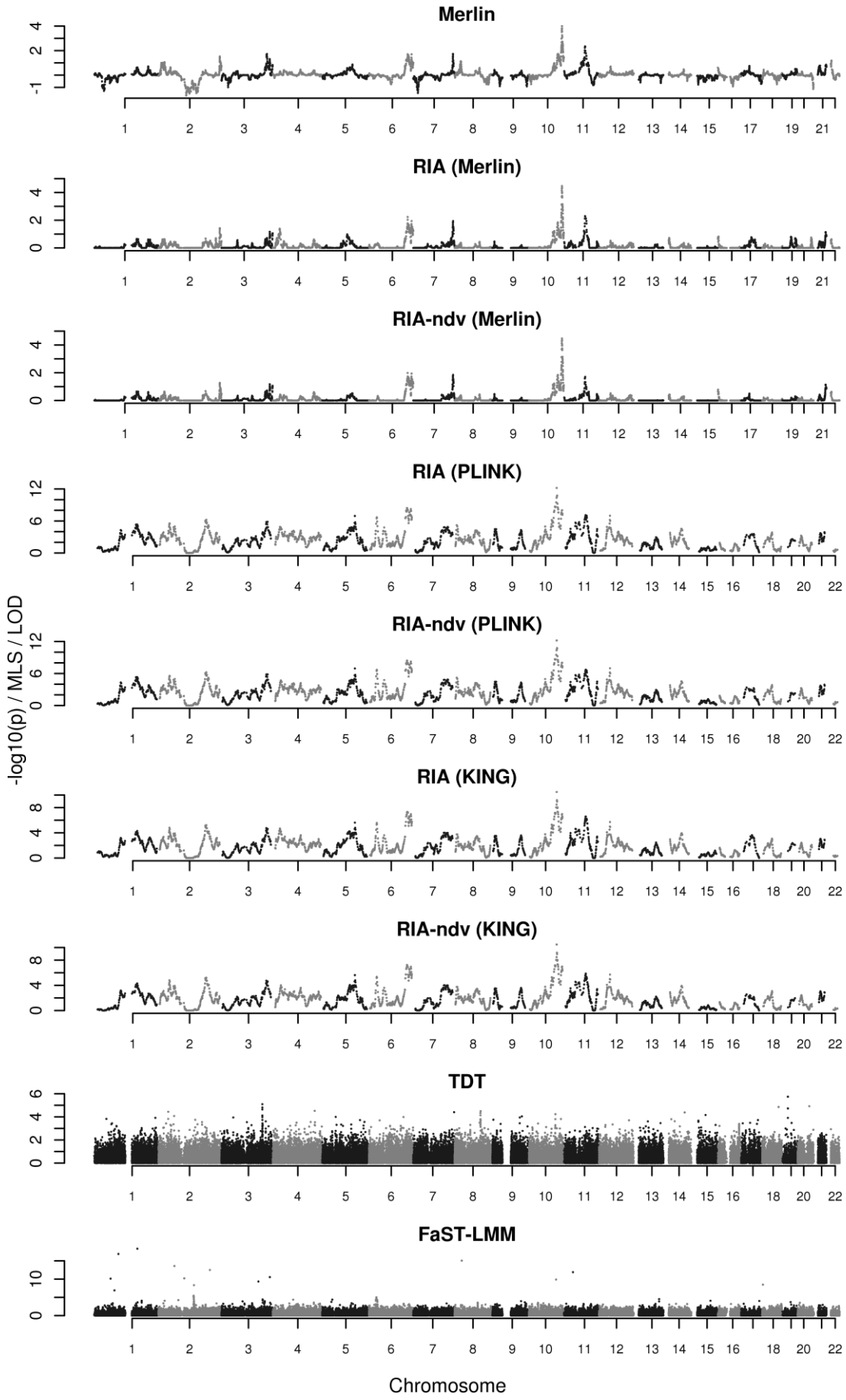

Figure 6.2 Manhattan plots for real VUR data set (all samples) using various non-parametric linkage analysis and association methods. RIA = Regional IBD Analysis, RIA-ndv = RIA with

dominance variance set to 0, (Merlin) = using theoretical IBD estimates from Merlin, (PLINK) = using IBD estimated by PLINK, (KING) = using IBD estimated by KING under homogeneous population assumption, TDT = transmission disequilibrium test implemented in PLINK.

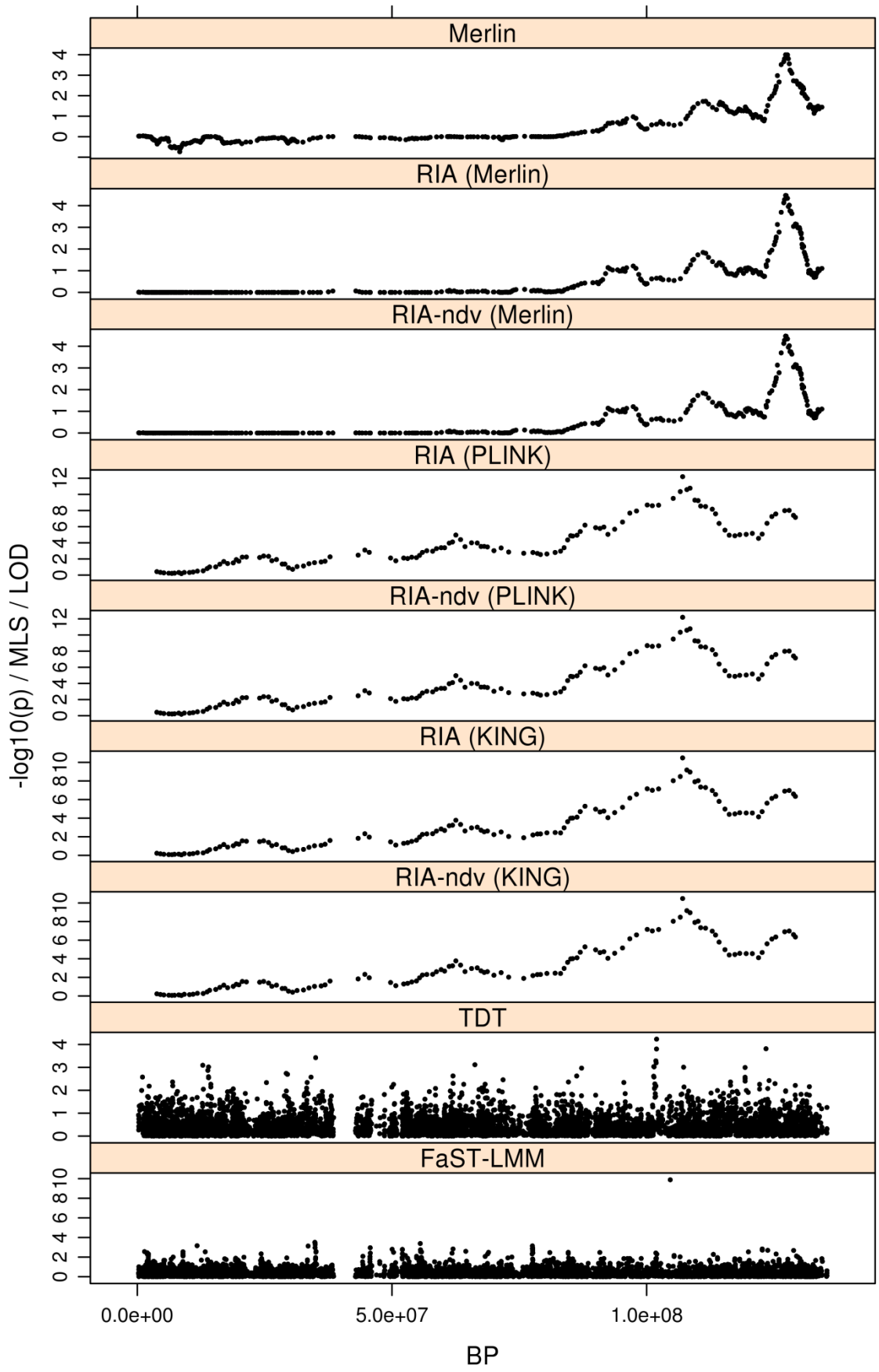

Figure 6.3 Comparisons of test statics for chromosome 10 of the real VUR data set (all samples) using various methods. RIA = Regional IBD Analysis, RIA-ndv = RIA with dominance variance set to 0, (Merlin) = using theoretical IBD estimates from Merlin, (PLINK) = using IBD estimated by

PLINK, (KING) = using IBD estimated by KING under homogeneous population assumption, TDT = transmission disequilibrium test implemented in PLINK.

The comparisons of the test score at each individual marker (Figure 6.4) also show similar pattern: high concordance between Merlin and RIA using Merlin's IBD estimates, and among RIA using various methods for genetic-based IBD estimation; and less concordance between Merlin and RIA using genetically estimated IBD. In fact, a pattern that can be seen here, and in subsequent similar comparisons, is that the lower values from methods using genetically-estimated IBD correspond well to those from the methods using theoretically estimated IBD, whereas the higher values from the genetic-based methods may not necessarily correspond to those from the theoretical methods.

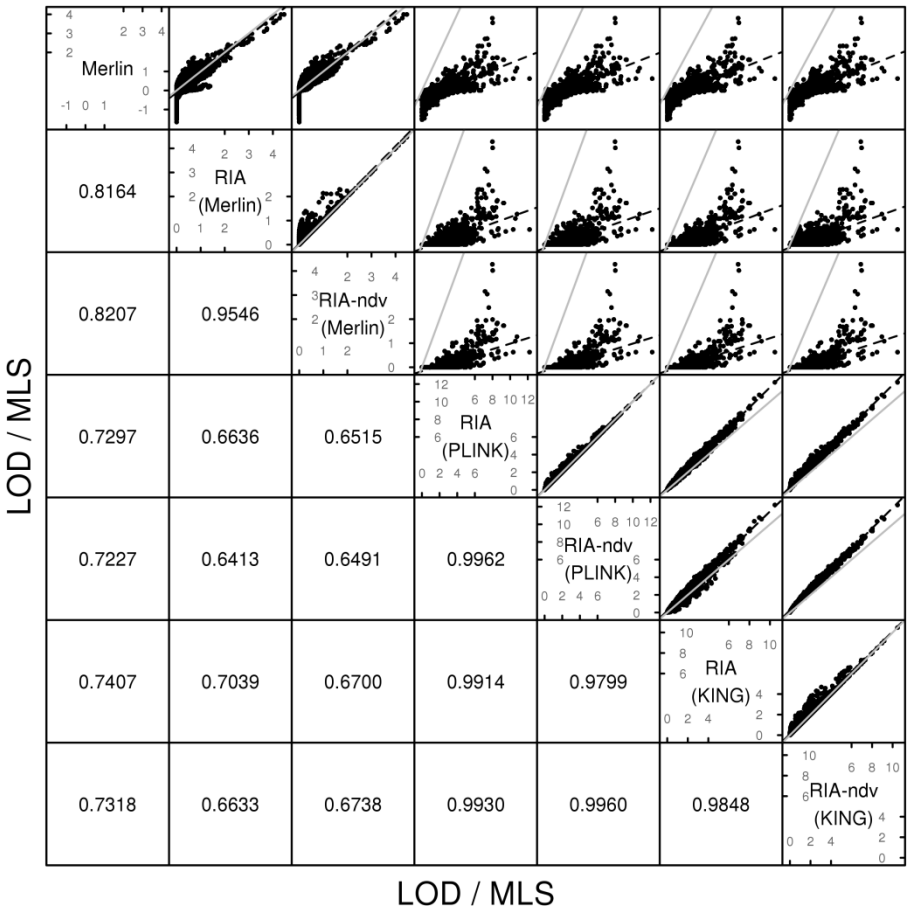

Figure 6.4 Comparison of MLS/LOD scores obtained from VUR data set (all samples) using RIA with various IBD estimation methods and using Merlin. Plot above the diagonal show a comparison of the scores, with correlation between them indicated below the diagonal. The grey solid lines represent the line of equality; the black dashed lines the linear regression line of the variable on the y axis on the variable on the x axis. RIA = Regional IBD Analysis, RIA-ndv = RIA with dominance variance set to 0, (Merlin) = using theoretical IBD estimates from Merlin, (PLINK) = using IBD estimated by PLINK, (KING) = using IBD estimated by KING under homogeneous

population assumption. Note that these are rather crude comparisons: because different sets of markers were used in different classes of analysis, these plots can only show the approximate matching between them, and will have additional discrepancies as a result of this.

A little caveat when looking at the plots similar to Figure 6.4 in this chapter is that, unlike this type of plots from the previous chapters, the marker locations in each class of methods may not correspond exactly to the locations in the other classes. This is because different sets of markers—or in RIA, 'pseudo-markers', defined as the mid-position of each window—were used in the analyses by different classes of methods. The comparison plots were then constructed by mapping each marker from the methods with the least number of markers to the nearest marker (in terms of map distance, which must also be within 1 cM) from the methods with more markers. This inevitably led to some further discrepancies in the plots due to the slight mismatch of markers alone. However, even in presence of these discrepancies, all methods still seem quite concordant.

To investigate the performance of RIA when a denser chip, more comparable to modern standards, is used, the methods were also applied to a subset of VUR samples, namely, those from the Dublin cohort, who had been genotyped on a higher resolution platform (see Table 2.1 on page 28 for details of this cohort). This increases the number of SNPs after quality control from the 119,548 common to all samples to 644,006, and would also mean the samples are more homogeneous, although at the expense of a smaller sample size.

As the number of SNPs on the chip increases, the size of the window for IBD estimation also has to be increased if the window is to span a similar distance. It appears that a window of 2,000 SNPs gives a reasonably smooth baseline in this denser data set without losing sensitivity. (Given that the number of SNPs increased by about 5.4 fold whilst the window size only by 4 fold, this actually means that each window is expected to span a slightly smaller distance than in the all-sample analysis.)

The findings in this analysis are quite similar to the all-sample analysis, although the concordance between the methods using theoretical IBD estimates and methods using genetically estimated IBD seems to be slightly lower (Figures 6.5-6.7). Interestingly, Merlin now produced further signals before the end of chromosome 10, which seem to correspond to the signals detected by RIA in the all-sample data set (Figure 6.6, *cf.* Figure 6.3 on page 151). As for RIA using genetic IBD estimation, these signals are now more prominent than that at the end of the chromosome. With these observations, it could be that the signals detected by RIA in both data sets are real, and perhaps contributed more by the Dublin group.

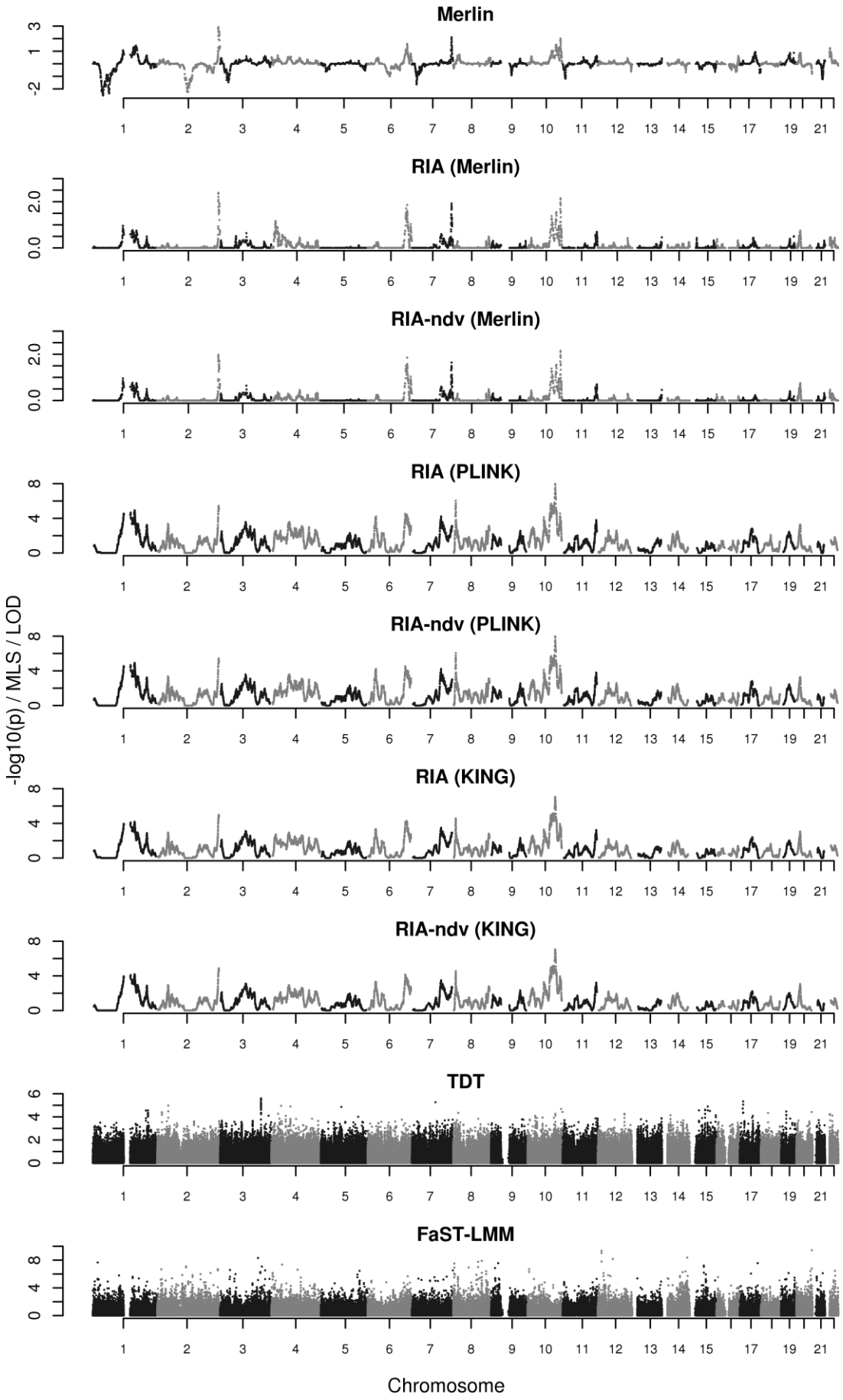

Figure 6.5 Manhattan plots for real VUR data set (Dublin samples) using various non-parametric linkage analysis and association methods. RIA = Regional IBD Analysis, RIA-ndv = RIA with

dominance variance set to 0, (Merlin) = using theoretical IBD estimates from Merlin, (PLINK) = using IBD estimated by PLINK, (KING) = using IBD estimated by KING under homogeneous population assumption, TDT = transmission disequilibrium test implemented in PLINK.

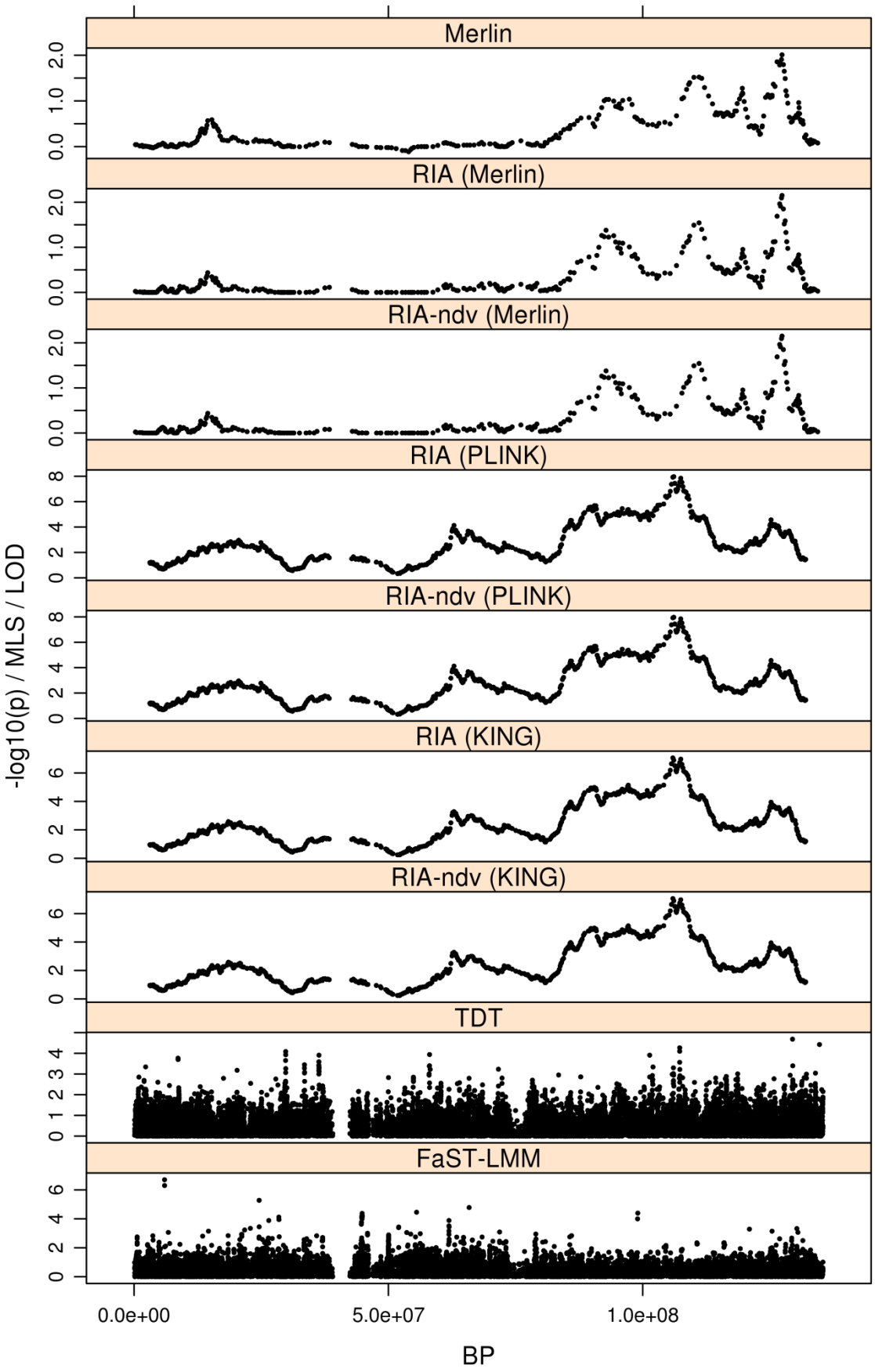

Figure 6.6 Comparisons of test statics for chromosome 10 of the real VUR data set (Dublin samples) using various methods. RIA = Regional IBD Analysis, RIA-ndv = RIA with dominance variance set to 0, (Merlin) = using theoretical IBD estimates from Merlin, (PLINK) = using IBD

estimated by PLINK, (KING) = using IBD estimated by KING under homogeneous population assumption, TDT = transmission disequilibrium test implemented in PLINK.

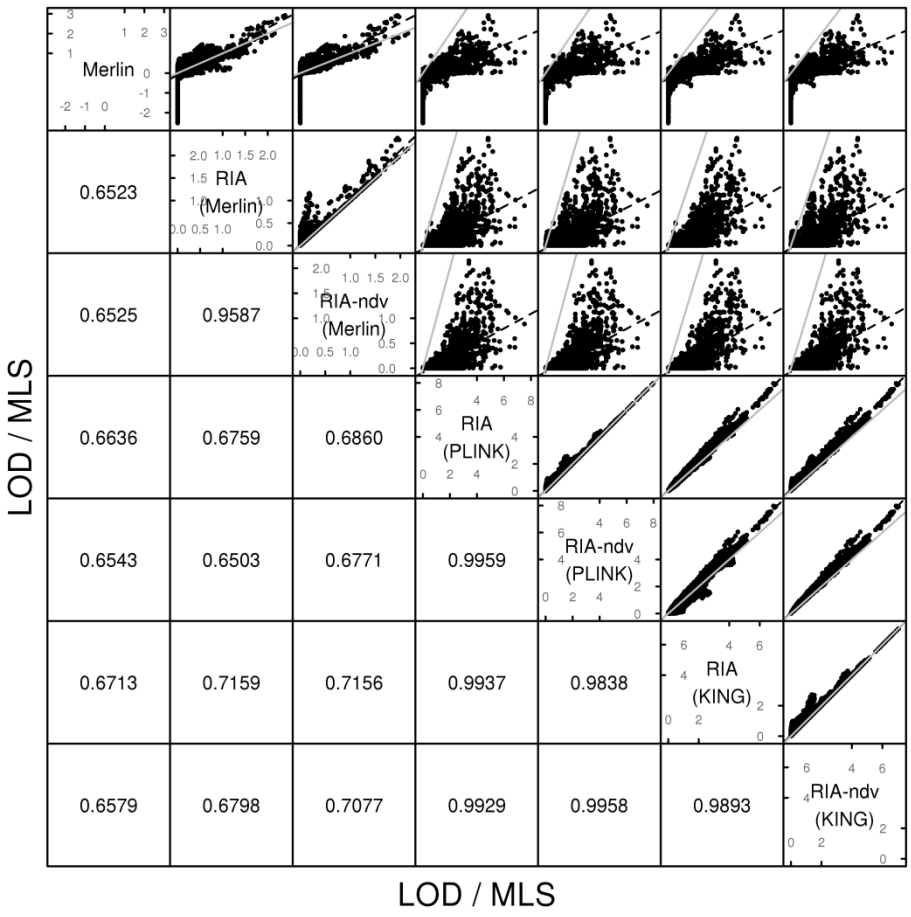

Figure 6.7 Comparison of MLS/LOD scores obtained from VUR data set (Dublin samples) using RIA with various IBD estimation methods and using Merlin. Plots above the diagonal show a comparison of the scores, with correlation between them indicated below the diagonal. The grey solid lines represent the line of equality; the black dashed lines the linear regression line of the variable on the y axis on the variable on the x axis. RIA = Regional IBD Analysis, RIA-ndv = RIA with dominance variance set to 0, (Merlin) = using theoretical IBD estimates from Merlin, (PLINK) = using IBD estimated by PLINK, (KING) = using IBD estimated by KING under homogeneous population assumption. Note that these are rather crude comparisons: because different sets of markers were used in different classes of analysis, these plots can only show the approximate matching between them, and will have additional discrepancies as a result of this.

Comparisons were also made between RIA using two different windows sizes on this data set. The finding was similar across all methods of estimation: background noise in RIA tends to increase when the window size for IBD estimation becomes smaller, to the point that it obscures the 'true' signals whilst at the same time gives out many 'false' signals. Interestingly, the magnitude of the test statistics also seems to be highly dependent on the window size in a similar fashion. An example of this using IBD estimates from PLINK in standard version of Onelocarp is shown in Figure 6.8.

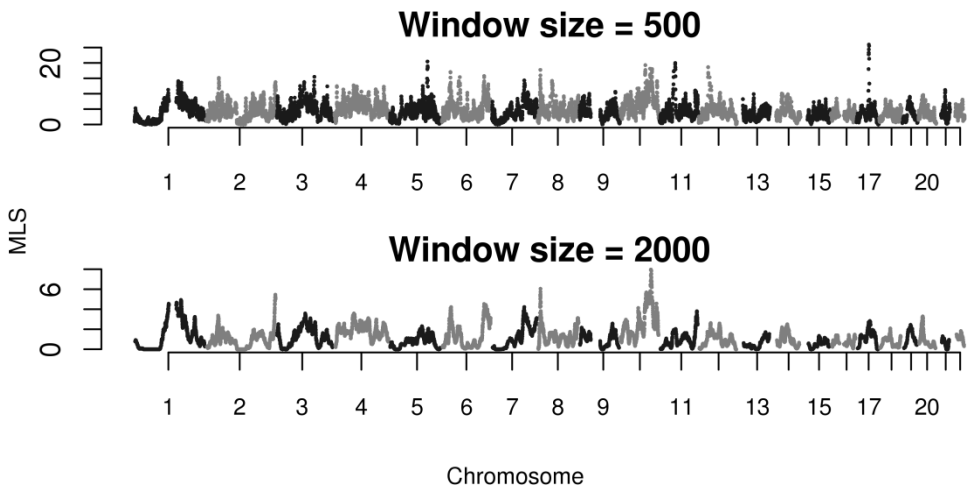

Figure 6.8 Manhattan plots for real VUR data set (Dublin samples) using RIA with IBD estimation from PLINK with different window sizes.

6.3. Comparison with Exact and Simulation-based Non-parametric Linkage Analysis, Using VL Data with Reduced Pedigree Complexity

With the VL data set, RIA is being applied to the type of data it is intended for: large, complex pedigrees, and with affected individuals who are less related. However, the attempt to use the full VL data set presented a complication—it was not possible (for comparison purposes) to perform the full analysis using the exact method implemented in Merlin.

Although using sparse trees instead of full binary trees in the Lander-Green algorithm should in theory allow Merlin to handle relatively large families, Merlin was not able to complete its analyses in many families in the VL data set. This was despite its having been compiled as a 64-bit application (so that it is not subjected to the 4 GB memory limit inherent to 32-bit applications) and supplied with more than adequate physical memory. The problem seems to be due to the actual program code, which, at least for the family tree construction module, is still coded using 32-bit data structure. Without significant modification, Merlin would not be able to analyse the full VL data set.

To enable comparisons between RIA and 'standard' non-parametric linkage analysis using the full VL data set, an alternative analysis using simulation method was used. The results of this will be presented in the next section. However, if RIA is to be compared to the exact method, then the complexity of the pedigrees in the VL data set needs to be reduced. Although this slightly defeats the purpose of using the VL data set in the first place, it was hoped that the reduced data set would still retain some of its

complex family structure—presumably, it should not be necessary to reduce the family size to the point that they effectively become nuclear families.

Merlin does in fact provide the `--trim` option which removes individuals without genotype or phenotype information who are not required to define pedigree relatedness between other individuals. However, this automated procedure did not completely resolve the problem as there remained several families that were still too complex.

Further reduction of pedigree complexity in these families was done through manual trimming. As an example, in family 245 (see Figure 6.9, next page, for annotated pedigree), there are 4 affected individuals (drawn in black, also marked A-D) in what appears to be a very large family. However, upon closer inspection, it is apparent that this family actually consists of 3 'subfamilies' (SF1-SF3) linked only through two marriages. Affected individual A from subfamily SF1 is only related to the other subfamilies (and affected individuals) through marriage. He is therefore biologically unrelated to them (as C. C. Li and Sacks (1954) remarked: "one's relatives are not necessarily still relatives."), except perhaps for cryptic relatedness, and cannot be used for (family-based) linkage analysis. This also means the whole SF1 subfamily is irrelevant to the analysis and can be excluded. Subfamily SF2 cannot be used either, but for a different reason. Although B who is a daughter of a member of this subfamily is also a cousin to C in subfamily SF3, she is not genotyped. And because neither of her parents or grandparents or in fact most members of SF2 have been genotyped, there is little to be gained from including B in the analysis while the memory cost is likely to be very large. With B excluded, the whole SF2 can also be excluded. This leaves only SF3 (without B's immediate family members) for the analysis. However, even this still caused problems with Merlin; and further removal of non-genotyped individuals who do not contribute to IBD estimation (R1 and R2 groups) was required before the family could be successfully analysed.

In some families, it was not possible to trim down without losing information. For example, in family 15 (Figure 6.10), there are 4 affected individuals who are all related and genotyped and the family appears to be a true single, large family. Most of the family members are also genotyped. In this type of family, removing any member potentially reduces the accuracy of theoretical IBD estimation, but is necessary for the reduction of pedigree complexity. Repeated 'random' trimming, while still attempting to keep all affected individuals and preserve overall family structure, was used in this situation, until the family could be analysed in Merlin.

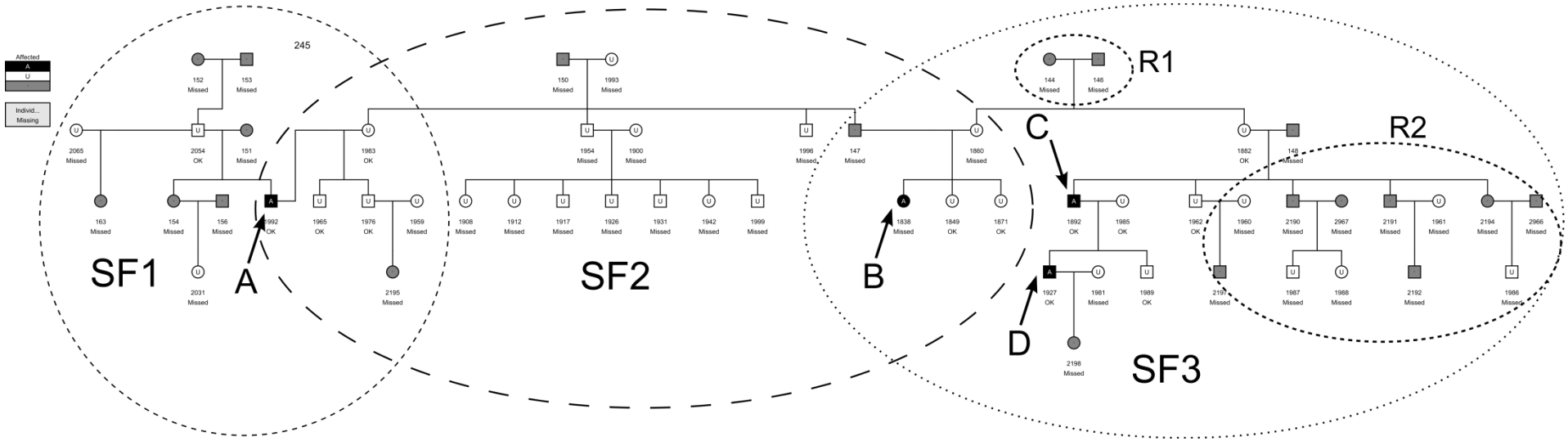

Figure 6.9 Pedigree of family 245 from the VL data set and example of manual trimming process. This pedigree was plotted using Madeline 2.0 PDE (Trager *et al.*, 2007) with slightly modified colour scheme to match conventional usage, i.e. black denotes affected individuals and white unaffected individuals. Individuals plotted in grey are those whose phenotype is missing. Although they may be too small to be read properly here, the first line of labels underneath each individual symbol is the individual ID, and the second line reflects the genotyping status: either 'OK' (appears as shorter text) for genotyped individuals or 'Missed' (appears as longer text) for non-genotyped individuals. See text for descriptions of the various annotations.

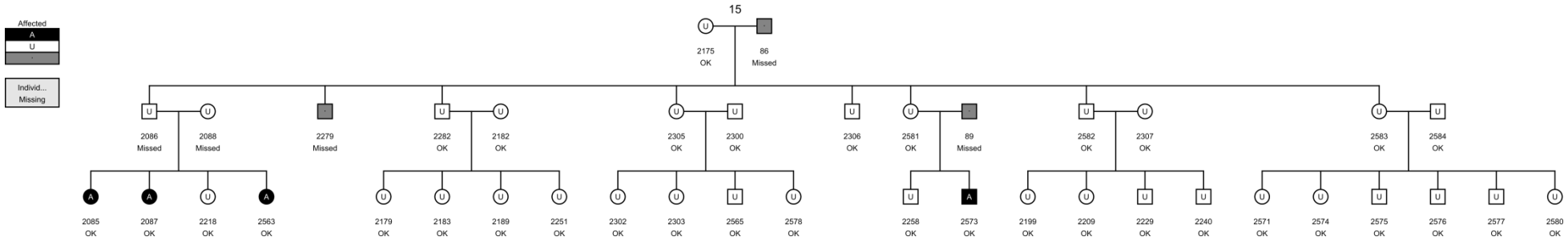

Figure 6.10 Pedigree of family 15 from the VL data set. This pedigree was plotted using Madeline 2.0 PDE (Trager *et al.*, 2007) with slightly modified colour scheme to match conventional usage, i.e. black denotes affected individuals and white unaffected individuals. Individuals plotted in grey are those whose phenotype is missing. Although they may be too small to be read properly here, the first line of labels underneath each individual symbol is the individual ID, and the second line reflects the genotyping status: either 'OK' (appears as shorter text) for genotyped individuals or 'Missed' (appears as longer text) for non-genotyped individuals.

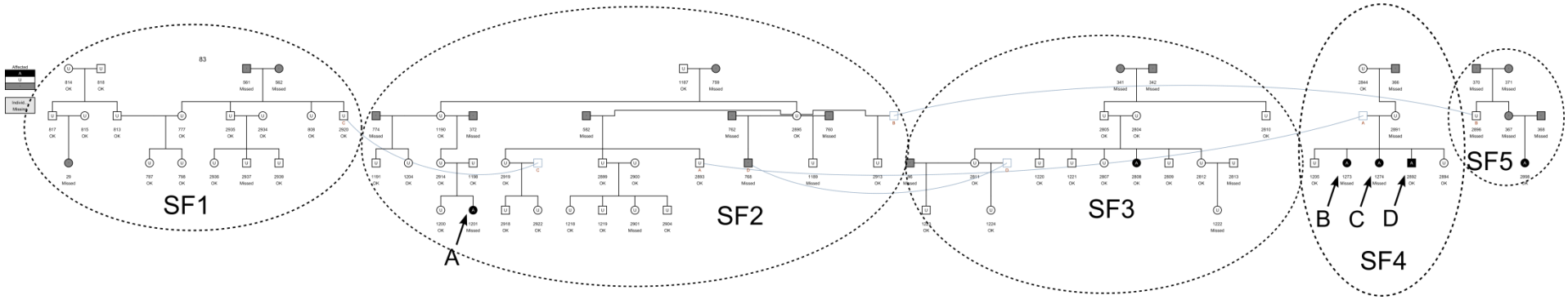

Figure 6.11 Pedigree of family 83 from the VL data set and example of manual trimming process. This pedigree was plotted using Madeline 2.0 PDE (Trager *et al.*, 2007) with slightly modified colour scheme to match conventional usage, i.e. black denotes affected individuals and white unaffected individuals. Individuals plotted in grey are those whose phenotype is missing. The curved lines linking two 'individuals' signify that they are in fact a single individual drawn twice in two separate locations to enable the pedigree to be plotted (as would be the case when one member of an extended pedigree marries into another extended pedigree). Although they may be too small to be read properly here, the first line of labels underneath each individual symbol is the individual ID, and the second line reflects the genotyping status: either 'OK' (appears as shorter text) for genotyped individuals or 'Missed' (appears as longer text) for non-genotyped individuals. See text for descriptions of the various annotations.

The situation with family 83 (Figure 6.11, previous page) was more complicated. This family consists of 5 subfamilies (SF1-SF5). However, whilst SF1, SF3 and SF5 are each connected to SF2 through a single marriage at a non-strategic branch and can be readily removed, the same cannot be said for the relationship between SF2 and SF4. The marriage between the two members of SF2 and SF4 connects the affected individual A in SF2 to the three other affected individuals (B-D) in SF4. This makes them potentially informative for linkage analysis. However, because among these four biologically-related affected individuals, only D is genotyped, the application of both Merlin and RIA to this pedigree is precluded. This pedigree (or any other pedigrees similar to it) was therefore excluded.

The manual trimming was done for every family that Merlin still reported as too complex after the `--trim` option was used. This resulted in a final reduced data set of 816 individuals from 82 families, 198 of whom are affected.

Although not strictly necessary in this reduced data set, a simulation-based non-parametric linkage analysis using lm_ibdtests was also performed, so that its results could also be compared with those from Merlin before being used as a sole reference in the next section.

The results from these analyses again show rough similarities among all methods (Figures 6.12-6.13), although Merlin and RIA using Merlin's theoretical IBD estimates appear to have better discriminatory power than RIA using genetically estimated IBD and the simulation-based lm_ibdtests (which also uses theoretical IBD estimates). In fact, results from RIA using genetically estimated IBD appear to be quite similar to those from lm_ibdtests, taking into account the apparently more random nature of the latter.

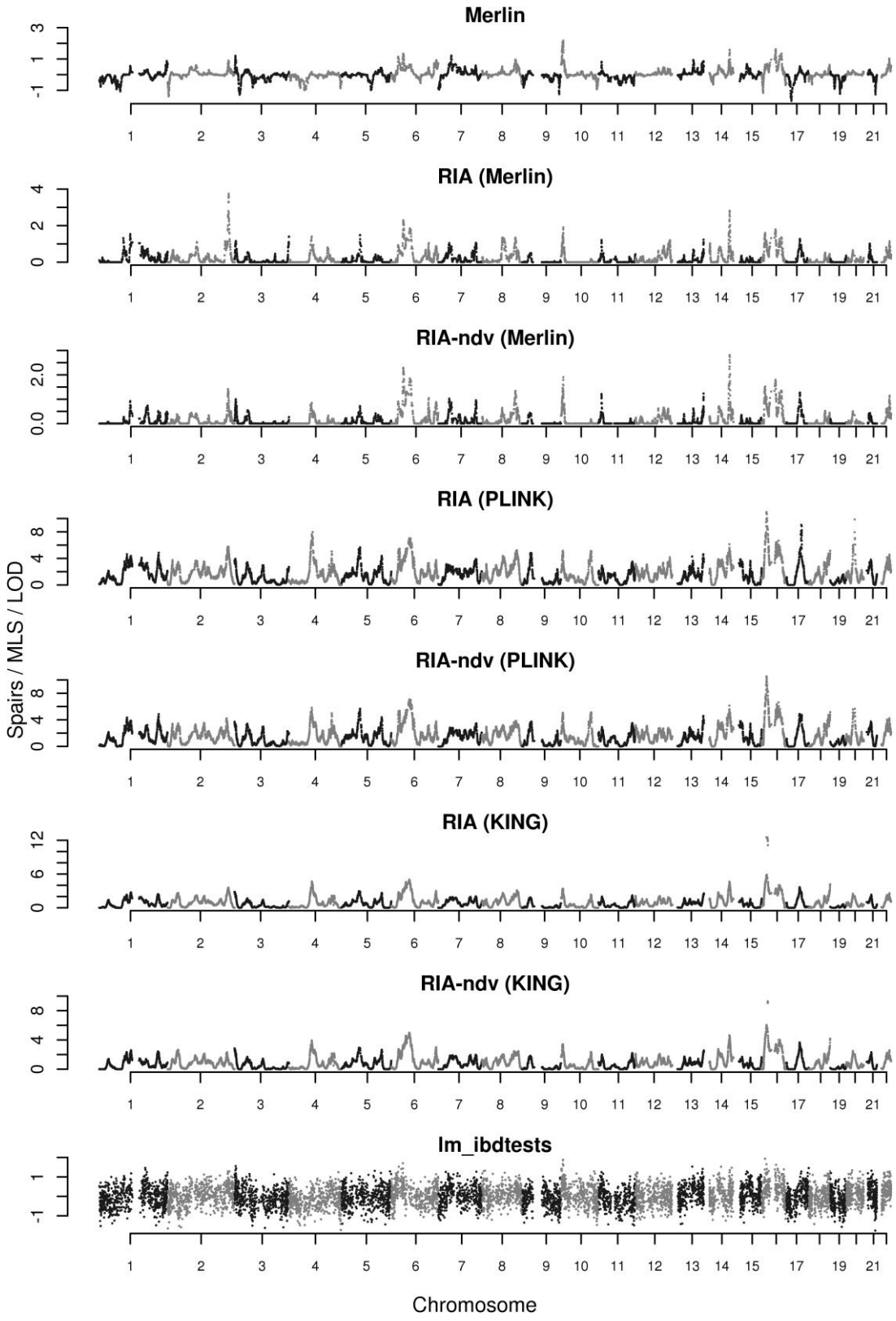

Figure 6.12 Manhattan plots for the reduced-complexity VL data set with real phenotype using various non-parametric linkage analysis methods. RIA = Regional IBD Analysis, RIA-ndv = RIA with dominance variance set to 0, (Merlin) = using theoretical IBD estimates from Merlin, (PLINK) = using IBD estimated by PLINK, (KING) = using IBD estimated by KING under homogeneous population assumption.

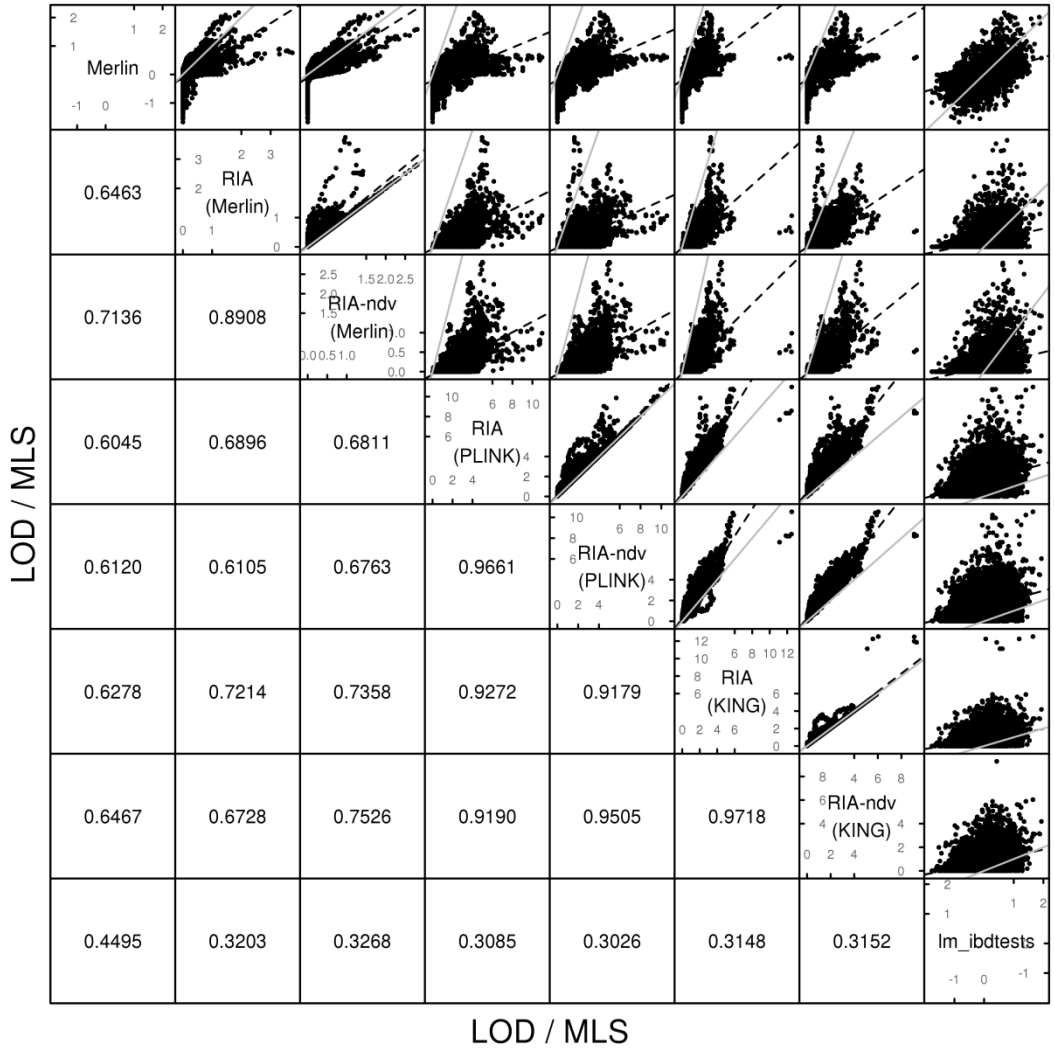

Figure 6.13 Comparison of MLS, LOD and S_{pairs} scores obtained from the reduced-complexity VL data set with real phenotype using RIA with various IBD estimation methods and using Merlin and lm_ibdtests. Plot above the diagonal show a comparison of the scores, with correlation between them indicated below the diagonal. The grey solid lines represent the line of equality; the black dashed lines the linear regression line of the variable on the y axis on the variable on the x axis. RIA = Regional IBD Analysis, RIA-ndv = RIA with dominance variance set to 0, (Merlin) = using theoretical IBD estimates from Merlin, (PLINK) = using IBD estimated by PLINK, (KING) = using IBD estimated by KING under homogeneous population assumption. Note that these are rather crude comparisons: because different sets of markers were used in different classes of analysis, these plots can only show the approximate matching between them, and will have additional discrepancies as a result of this.

6.4. Comparison with Simulation-based Non-parametric Linkage Analysis, Using VL Data with Full Pedigree Complexity

Most families in the VL data set are included in this 'full complexity' data set. The exceptions are the families with less than two individuals that are both affected and genotyped, which cannot be used in family-based linkage analysis methods. After the exclusion of these families, 1114 individuals from 84 families remain in the final data set, 203 of them affected and genotyped.

With full pedigree complexity, Merlin can no longer be used. Instead, the simulation-based lm_ibdtests was used as the reference method. This poses a slight problem as results from lm_ibdtests tend to be rather noisy (as can be seen from the previous analysis) and conservative (see Section 6.1.4). Furthermore, since neither of the methods detected any signal in this data set (Figure 6.14), it could be said that the comparisons were being made here only with regard to the degree of concordance of the background noise. Nevertheless, as can be seen from Figures 6.14 and 6.15, results from these methods appear to be at least roughly concordant, with RIA appearing to have slightly better discriminatory 'power'.

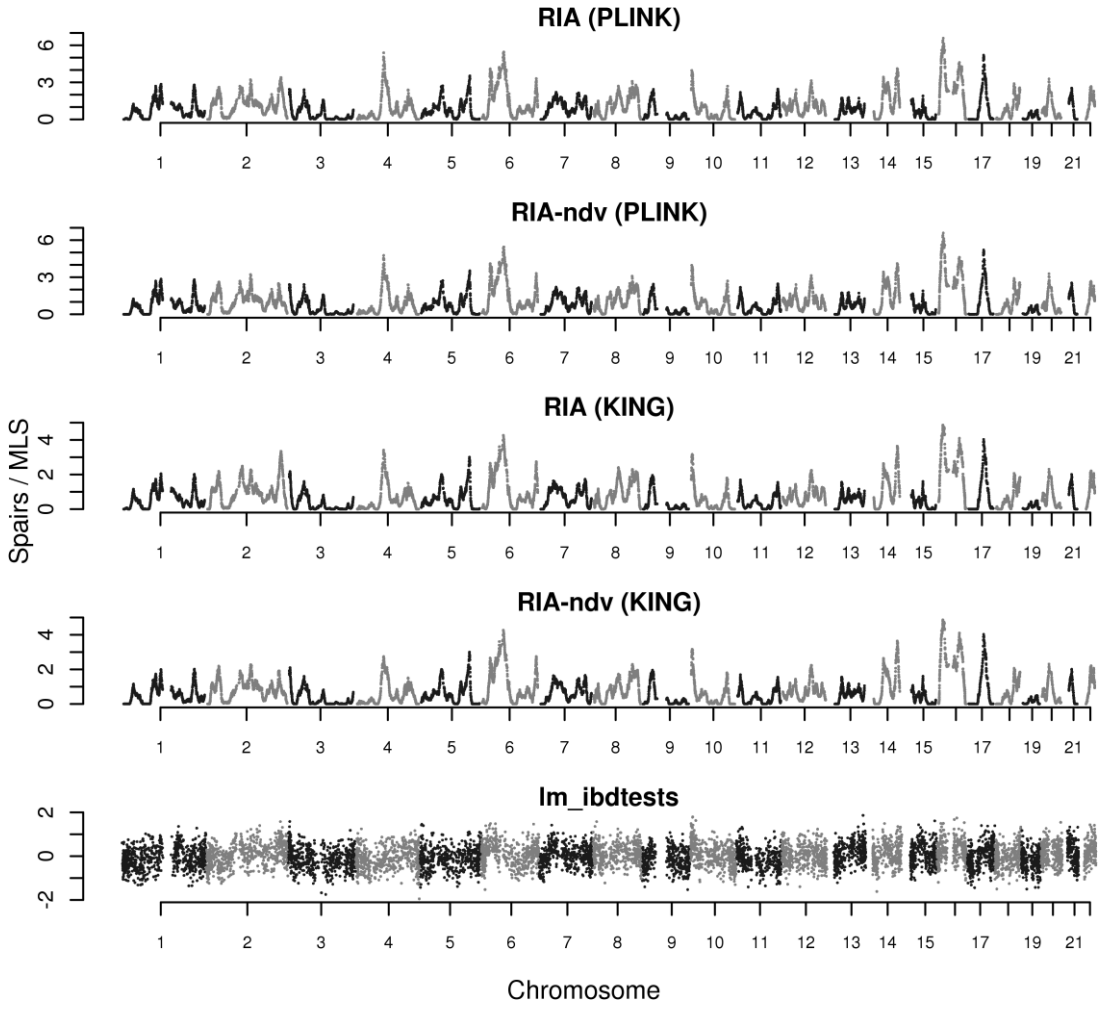

Figure 6.14 Manhattan plots for the full complexity VL data set with real phenotype using various non-parametric linkage analysis methods. RIA = Regional IBD Analysis, RIA-ndv = RIA with dominance variance set to 0, (PLINK) = using IBD estimated by PLINK, (KING) = using IBD estimated by KING under homogeneous population assumption.

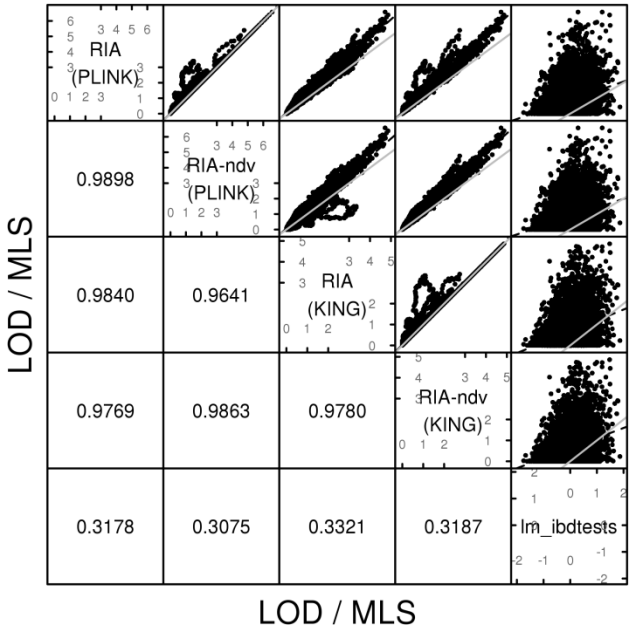

Figure 6.15 Comparison of MLS and LOD scores obtained from the full complexity VL data set with real phenotype using RIA with various IBD estimation methods and using lm_ibdtests. Plot above the diagonal show a comparison of the scores, with correlation between them indicated below the diagonal. The grey solid lines represent the line of equality; the black dashed lines the linear regression line of the variable on the y axis on the variable on the x axis. RIA = Regional IBD Analysis, RIA-ndv = RIA with dominance variance set to 0, (PLINK) = using IBD estimated by PLINK, (KING) = using IBD estimated by KING under homogeneous population assumption. Note that these are rather crude comparisons: because different sets of markers were used in different classes of analysis, these plots can only show the approximate matching between them, and will have additional discrepancies as a result of this.

The effect of different window size for IBD estimation on this data set is demonstrated in Figure 6.16. Again, the larger window size resulted in less noisy results as well as in the decrease in magnitude of test statistics, with 2,000 SNPs appearing to be an optimal window size for this data set, similar to the Dublin VUR data set.

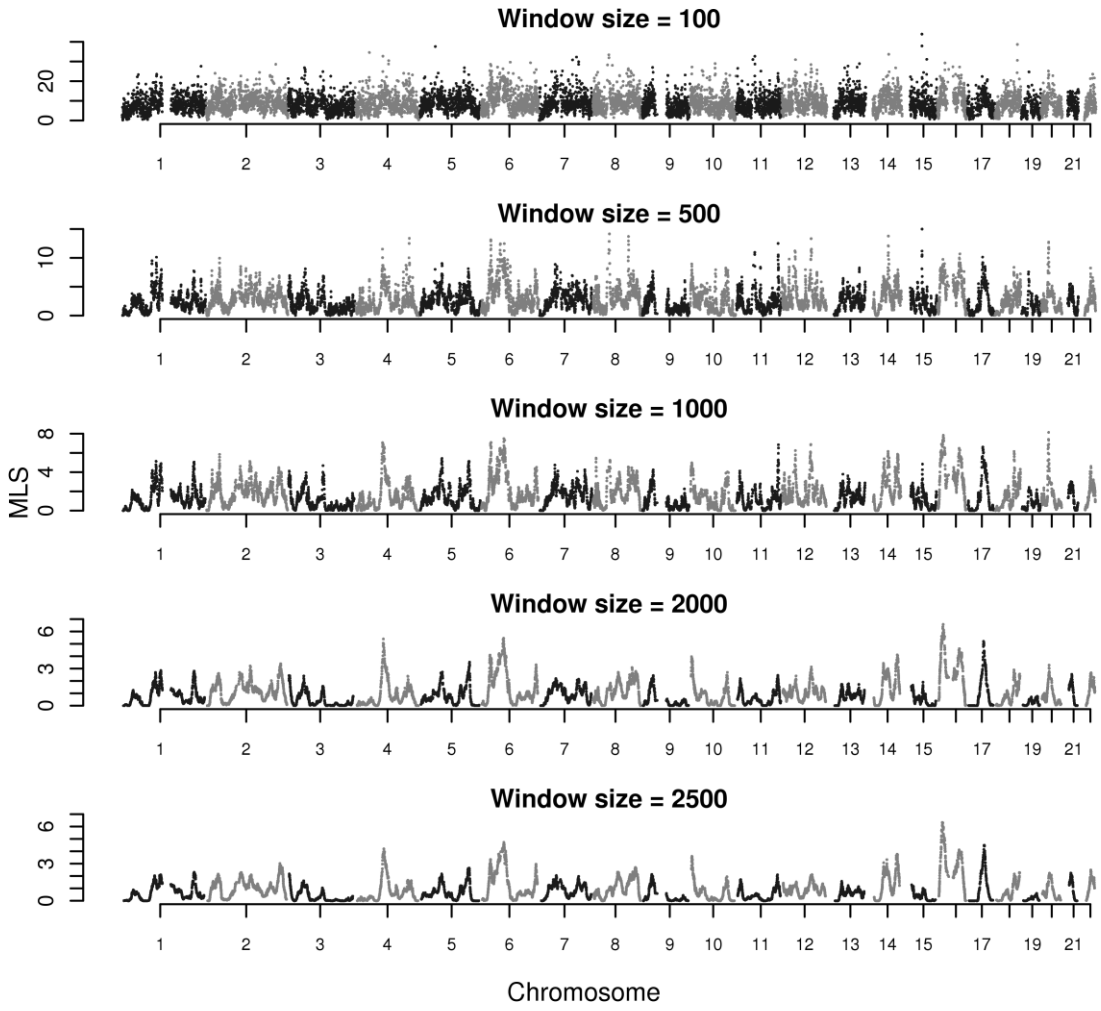

Figure 6.16 Manhattan plots for full complexity, real VL data set using RIA with IBD estimation from PLINK with different window sizes.

6.5. Discussion

This chapter demonstrates the use of RIA, a new non-parametric linkage analysis method based entirely on genetically estimated IBD, in both small and large family data sets. It is particularly advantageous in a data set with larger families, where an exact method cannot be used. It also seems to give better results than simulation-based linkage analysis, and should in theory have more computational advantage when pedigree becomes larger, as, unlike traditional linkage analysis methods, it is not affected by pedigree complexity. An additional advantage is that it does not require prior knowledge of the family structure, and can be used even when the pedigree information is absent or incorrect.

For efficiency reasons, only family-based linkage analysis was performed here. However, in theory, RIA should be applicable for population-based linkage analysis (that is, analysis using individuals who are not known to be related) as well, provided that a suitable IBD estimation method is used.

For non-nuclear family data such as the VL data set, it is possible that some of the affected relative pairs are not actually biologically related. An example of this can be seen in family 245 (Figure 6.9, page 160): affected individual A is related to the two other affected and genotyped individuals C and D only through a series of marriages, yet will be analysed as affected relative to them. Strictly speaking, linkage analysis using the pairs A-C and A-D is in fact population-based. However, RIA is capable of handling this; and such pairs were included in the analysis presented here as this would resemble the realistic analysis of this type of data sets.

The current implementation of KING only allows the allele frequencies to be estimated from the data set in use. This poses a particular theoretical problem when applied to affected relative pairs methods like RIA, as the genotype data could potentially be available only from these individuals, which could affect the allele frequency estimation and consequently the accuracy of the IBD and ultimately the MLS calculations. With a more inclusive data set like those used here, it is possible to let KING estimate the IBD based on the whole data set, so that more accurate allele frequency estimation can be achieved, and then select only the relevant affected individual pairs for the MLS calculation. However, doing so is inefficient and more complicated practically. The method actually used here allowed KING to estimate the IBD based only on the affected individuals, which is computationally more efficient and could resemble more extreme data sets (perhaps those collected with the aim of using only affected relative pairs analysis or case-only analysis), at the cost of potentially less accurate IBD estimation. The results from this approach are in fact reassuringly similar to PLINK, perhaps because the aim was not to accurately estimate the IBD in itself, but rather to compare downstream analysis from two sets of estimated IBD, in which case any bias from inaccurate allele frequencies may have cancelled each other out.

Another issue affecting the accuracy of IBD estimation is the size of SNP windows used. As has been demonstrated here, smaller window size results in more noisy results, probably because the IBD estimates are less stable and subject more to local fluctuations. This, however, needs to be balanced against using too many SNPs in each window, which could average out any local effect thus defeat the purpose of the test itself. There is no established rule as to how large the windows should be, and this is likely to depend on at least a few factors such as the density of the chip and the complexity of the pedigrees. In practice, it may be useful to perform a few trial runs with varying window sizes on a chromosome to optimise the size of the window to be used in the full analysis. Empirically, based on the data sets used here, a window size of 500 SNPs on the 100K chip, or 2,000 SNPs on the 600K chip, both corresponding approximately to the median span of 10 cM /10 million base pairs, seems to have given the best results.

On this note, it may be worth comparing these window sizes to another study which used a similar concept of global and local (or 'regional' in their terminology) relatedness estimated from windows of SNPs, but using different analytic framework. Nagamine *et al.* (2012) demonstrated the use of a mixed model method for quantitative trait analysis that incorporated genetically estimated global relatedness (in a manner similar to that in the previous chapters, but using the 'full' instead of 'pruned' set of SNPs) and 'regional' relatedness—representing the genetic block under test, and calculated from a short window of up to 100 adjacent SNPs—as two separate random effect components. Since the total number of SNPs in their data set was 275,564, the maximum window size of 100 SNPs means that their SNP windows would span approximately a tenth of the windows used here (or about 1 cM / 1 Mbps). The smaller window size would presumably be appropriate for their purpose of finding 'association' effect (although the method also shares a common characteristic of being able to integrate over multiple allelic effects with linkage analysis; the association in this case operates at the window level). However, this window size seems to lead to very 'noisy' signals in the data sets used in this thesis; for example, in Figure 6.16, the 100-SNP windows used by RIA on the full VL data set, which would be approximately equivalent to 50-SNP windows in Nagamine *et al.* (2012), gave many presumably false signals, and it would require much larger number of SNPs in a window to produce linkage signals analogous to those achieved by traditional linkage analysis software. One factor which may contribute to this phenomenon is that the method by Nagamine *et al.* uses a single summary measure for relatedness, whereas RIA uses all three IBD states, and would therefore be able to detect more variation. (As a very simplified hypothetical example: suppose the global IBD states between all affected pairs in the data set are (¼, ½, ¼) for 0, 1, and 2 alleles shared IBD, respectively (as is the case in 'ideal' full-sibs), but the local IBD states at a particular window in all pairs are (0, 1, 0), then RIA would detect these as very different; on the contrary, the global and regional summary kinship measures in the method by Nagamine *et al.* would both be equal at 0.25, and would imply no additional contribution from the regional window.)

The window size also affects the magnitude of the test statistics. This could be because the larger windows are more likely to include non-effect area, which dilutes the effect of the true SNPs. This dependence of the magnitude of the test statistics on the window size can complicate the calculation of the theoretical distribution for the test statistics, and therefore the calculation of the p-values.

A related problem with the current implementation of RIA, which is actually specific to most affected relative pairs methods, is that the calculated MLS could be anticonservative as the relative pairs are not jointly independent (Meunier *et al.*, 1997; Greenwood and Bull, 1999; Cordell *et al.*, 2000). To address this, Cordell *et al.* (2000)

proposed that the calculated MLS be used as pseudolikelihood instead, and calculate the significance level by using simulation. This issue is beyond the scope of this thesis. Nevertheless, the current usage of RIA depends on analysing the visual patterns for linkage signals rather than the numerical MLS or p-values (which are not currently calculated), and so is not affected by this issue.

Although RIA using genetically estimated IBD appeared to have detected positive signals as identified by Merlin, the baseline seemed rather noisy. Nevertheless, it seems that the strongest signals could still be separated from the background noise by visual inspection. This could potentially mean that RIA would have less power than exact methods, as many weaker signals would be obscured by the background noise. However, the aim here is to develop a method that is practical for complex pedigrees rather than to develop a superior method for small pedigree analysis; as such, RIA should be useful in certain circumstances.

Although attempted, it was not possible to properly compare the performance of RIA to the simulation-based method (lm_ibdtests) using the current VL data set, as it appeared that it may not actually contain a linkage signal, but at least the 'baselines' seemed quite similar among the methods. A different problem arose in the VUR analyses, as the true effect loci are not really known, so the relative merit of each method could not be judged. The next chapter will attempt to address these issues by means of phenotype and genotype simulations.

Chapter 7. Application of Genomic IBD Estimates in Non-parametric Linkage Analyses of Simulated Visceral Leishmaniasis Data Sets

In this chapter, the performance of RIA will be compared with a traditional non-parametric linkage analysis method using simulated data sets. In the same spirit as the LMM GWAS comparisons using simulated data sets in Chapter 5, this ensures that there is an effect locus, and its location is known, so that the relative merit of each method can be assessed. Because of the complexity of the pedigrees in this data set, Merlin cannot be used, and the simulation-based method lm_ibdtests is once again used for comparison.

Additionally, the results were also compared to those from FaST-LMM, as the ultimate goal of this chapter is to study the performance of RIA when there is linkage but no association signal in the data. The simulation settings were chosen to try to ensure a situation where there was linkage but no visible association signal in the region. Unfortunately, as will be seen later, I was not entirely successful at achieving this goal.

7.1. Comparison with Simulation-based Non-parametric Linkage Analysis of a SNP-based Qualitative Trait

The first simulation is a simple SNP-based qualitative trait simulation as described in Section 2.5.1. To briefly recap, this is an 'association'-type simulation with logistic additive effect based on the minor allele of SNP rs9271252 on chromosome 6, using unmodified genotype data from the VL data set. The window size for local IBD estimation used in this section as well as the next is 2,000 SNPs, advancing by 50 SNPs at a time, similar to that used in the VL data set analysis in the previous chapter.

As can be seen from Figures 7.1 and 7.2, results from RIA are roughly similar to those from lm_ibdtests, although RIA seems to have better discriminatory power (notice, in particular, the height of the 'signal' peak in chromosome 6 relative to the other 'false' peaks). However, what seems to be the top signal in both linkage methods (RIA and lm_ibdtests) is actually 'false', in a sense that it occurred about 20 cM before the simulated locus, while FaST-LMM detected this signal correctly (Figure 7.2). Nevertheless, RIA—and to a lesser extent, lm_ibdtests—detected the simulated locus as a weaker, secondary signal. This concordance between lm_ibdtests and RIA seems to suggest that the main signal detected by the linkage methods may, after all, be the true

linkage signal; and was perhaps caused by the simulation process combined with the genetic linkage structure in that area.

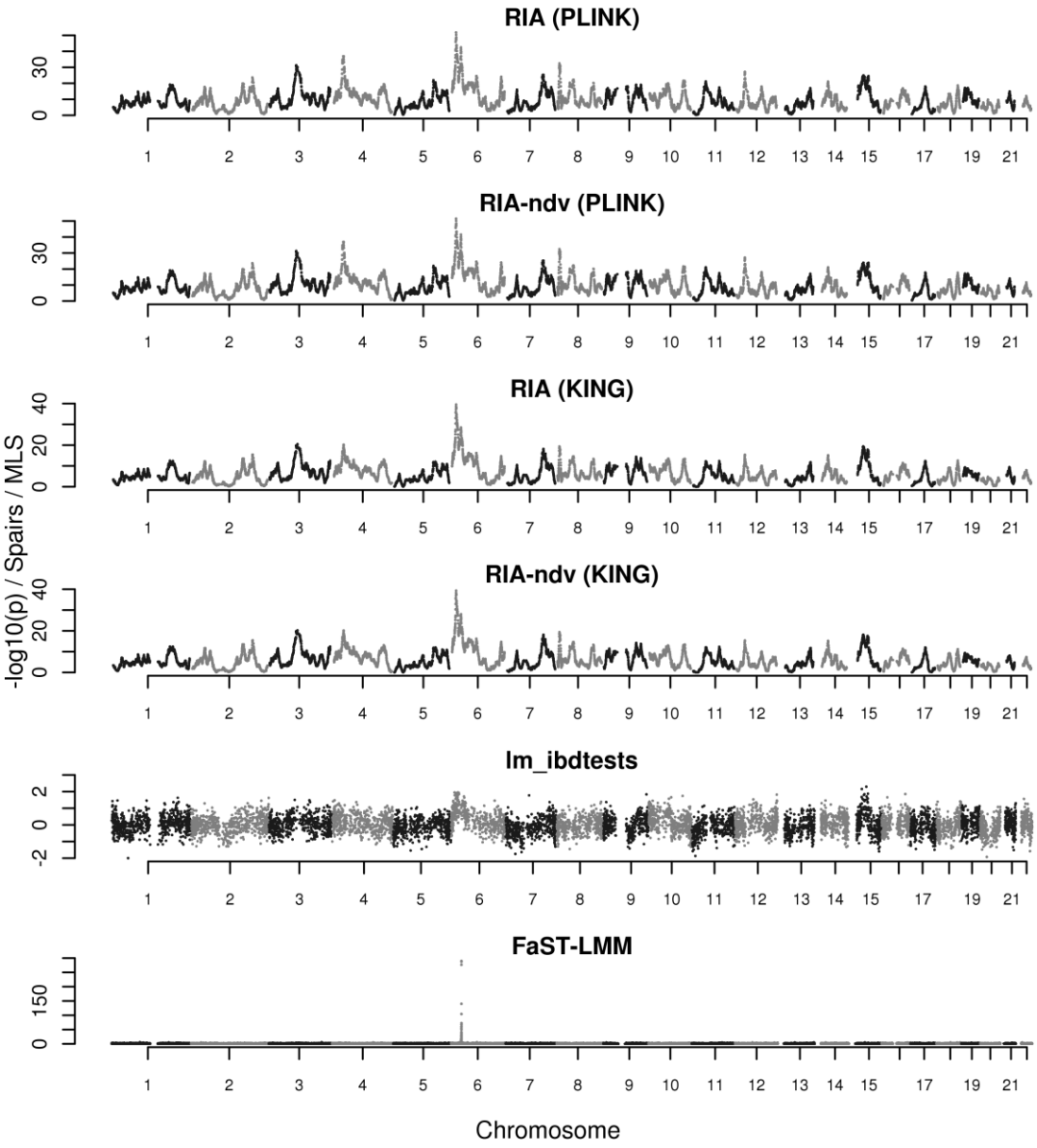

Figure 7.1 Manhattan plots for VL data set, SNP-based simulated qualitative phenotype, using various non-parametric linkage analysis and association methods. RIA = Regional IBD Analysis, RIA-ndv = RIA with dominance variance set to 0, (PLINK) = using IBD estimated by PLINK, (KING) = using IBD estimated by KING under homogeneous population assumption.

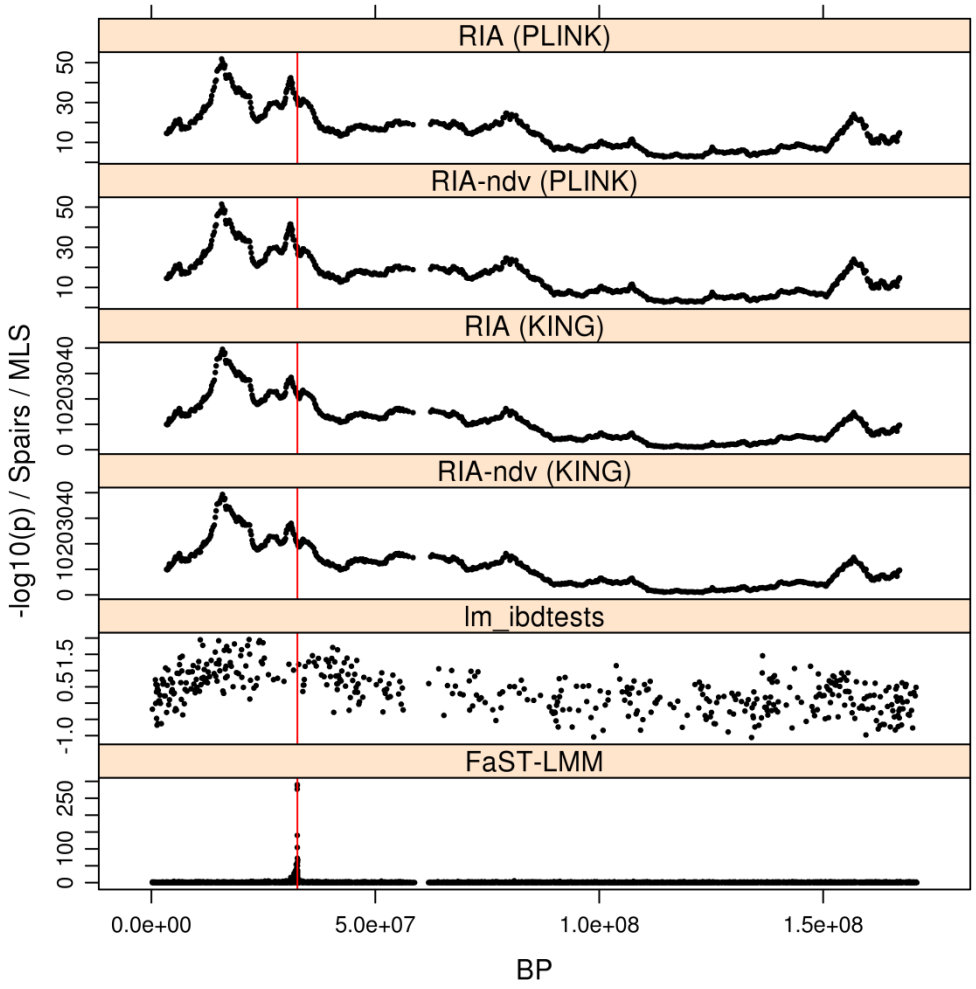

Figure 7.2 Comparisons of test statics for chromosome 6 of the VL data set, SNP-based simulated qualitative phenotype, using various non-parametric linkage analysis and association methods. RIA = Regional IBD Analysis, RIA-ndv = RIA with dominance variance set to 0, (PLINK) = using IBD estimated by PLINK, (KING) = using IBD estimated by KING under homogeneous population assumption.

Because of the uncertainty surrounding the accuracy of the MLS statistic to represent the true likelihood, and its sensitivity to the window size used, as described in the previous chapter (Section 6.5), it is not possible to assign the significance level to the test, and therefore not possible to perform a formal power and type I error evaluation for RIA. However, some intuition regarding RIA's discriminatory power can be gained from the following experiments:

Using the same model as the original simulation, 1,000 phenotype replicates were generated (using a very similar procedure to the generation of replicates for the GWAS data set, as described in Section 2.4.4). These simulated qualitative phenotypes were then used in very restricted RIA analyses involving just two loci: the first was the five adjacent local IBD windows surrounding the simulated effect SNP ('effect locus'); the other, another five adjacent local IBD windows in chromosome 7 ('null locus': this was

chosen at random). Additionally, for each replicate of the simulated phenotypes, a set of permutated phenotypes was created. These comprise 'null' phenotypes, which have no genetic contribution, and the 1,000 replicates of these were used in restricted RIA analysis involving only the 'effect locus'. The maximum MLS among the five local IBD windows was then selected from each analysis. The histograms of these maximum MLS are shown in Figure 7.3, which clearly demonstrate the discriminatory power of RIA: the distribution of RIA MLS test statistics under the alternative (when there is a genuine simulated effect) is seen to be well separated from their distribution under the null. As a side note, the MLS from the simulated phenotype / null locus replicates tend to be slightly higher than those from the null phenotype / effect locus replicates. This may be because in the null phenotype set, the phenotypes were completely uncoupled from the genotypes, whereas in the simulated phenotype set, some effects may still be seen in other loci due to random correlation of the genotypes.

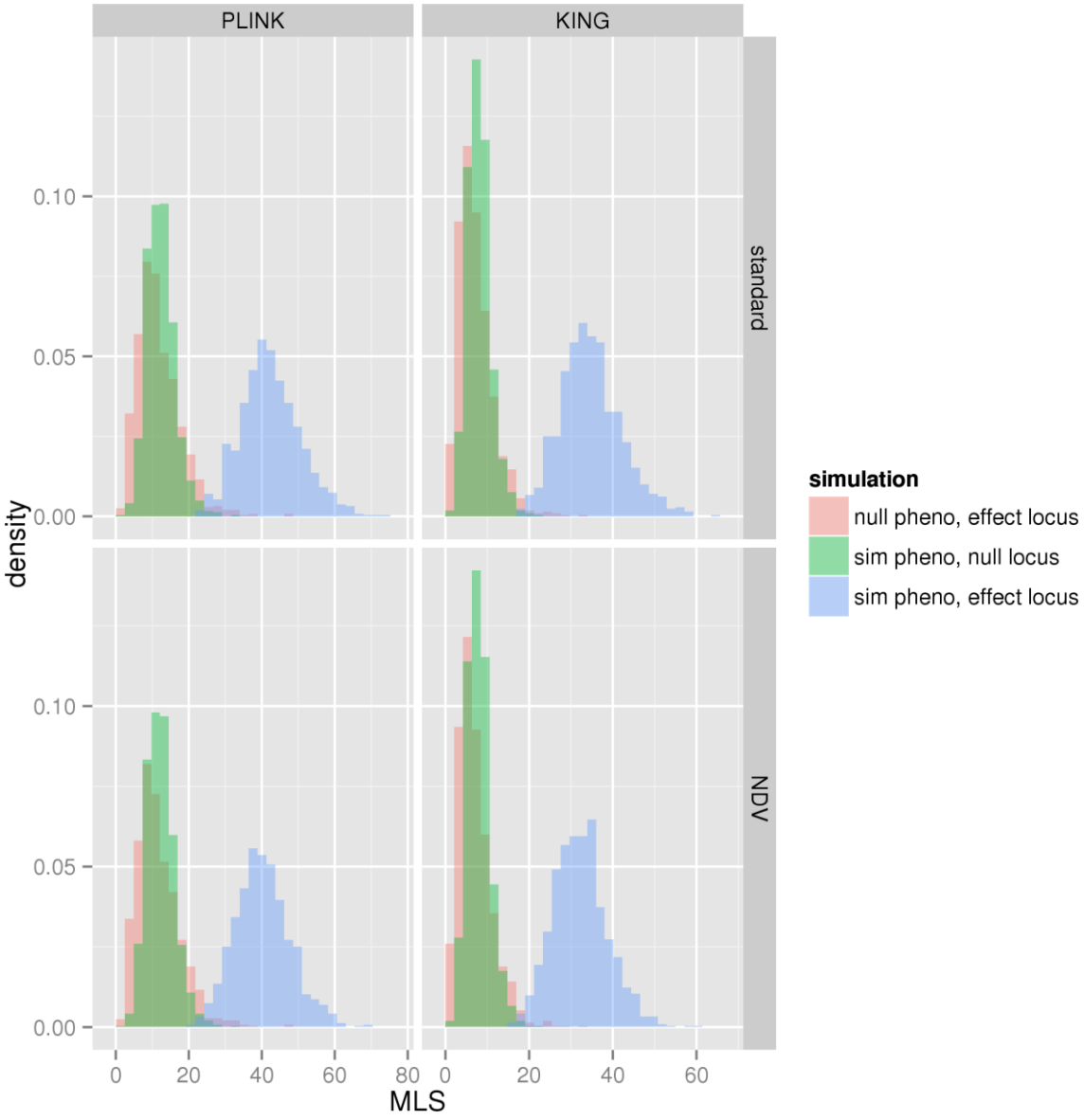

Figure 7.3 Histograms of MLS statistics from various RIA analyses of VL data set with SNP-based simulated qualitative phenotypes. Each panel shows histograms of the MLS statistics calculated using respective RIA methods (standard or no-dominance-variance (NDV)) and IBD estimation methods (PLINK or KING), with either simulated 'association' or 'null' phenotypes, at either the null or the simulated effect locus; each with 1,000 simulation replicates.

The analysis time for this data set was also measured on a standard memory worknode of the older HPC cluster (with similar caveats to Section 5.5). The results show a clear advantage of RIA, especially when using KING for IBD estimation (Table 7.1).

Method	Analysis time (hours)
lm_ibdtests	66
RIA (PLINK)	43
RIA (KING)	2

Table 7.1 Analysis time for various non-parametric linkage analysis methods used in this section. Based on a computer with a single 2.67 MHz CPU running a single process.

In attempt to create a situation where a linkage signal is present without an association signal, a slightly different simulation strategy was tried: the families in the data set were randomly split into two groups; in one group the phenotype was simulated based on the minor allele count similar to the above, but in the other group, the phenotype was based on the major allele count. The reason for this was that doing so should result in the association pattern in the two groups cancelling each other out, while having little impact on the linkage signal. However, the results actually showed marked reduction in the power of the linkage methods, whereas the association method (FaST-LMM) was relatively unaffected. With hindsight, this actually makes sense. Because the major allele is by definition more common than the minor allele (usually by a large margin), and combined with the way the simulation model works, most cases in the simulation were caused by the major rather than the minor allele. However, because the major allele is less likely to have been inherited IBD, this means that the linkage effect was severely diluted. On the contrary, because LMM GWAS methods make adjustment for genetic relatedness, it seems that FaST-LMM has successfully captured the family effect imposed in this simulation and adjusted for it. This simulation was therefore abandoned in favour of the more sophisticated simulations described in the next section.

7.2. Comparison with Simulation-based Non-parametric Linkage Analysis of a Haplotype-based Qualitative Trait

In further attempt to create a situation where the linkage signal is strong but the association signal is weak or absent, a more 'proper' linkage model was used. This has been described in detail in Section 2.5.2. Very briefly, this involves determining the haplotypes of a SNP cluster surrounding rs9271252, and inserting a disease SNP into this cluster. In certain haplotypes, the disease SNP allele was set to an 'affected' allele (effectively assigning those haplotypes as 'affected'), while in the remaining haplotypes, it was set to an 'unaffected' allele (thus assigning those haplotypes as 'unaffected'). With affected and unaffected haplotypes defined, the genotype data in the 10 cM range on chromosome 6 (47 cM to 57 cM, which contains the simulated SNP cluster) were generated by gene dropdown process. The simulated genotype data were then used for

phenotype simulation, based on the disease SNP, using exactly the same model as in the previous section, before removing the disease SNP from the data set.

What remained to be determined at this stage was which haplotypes were to be assigned as affected or unaffected. Several strategies of assignment (with increasing degree of sophistication) were tried, including, but not limited to:

1. Setting haplotypes 2 and 7 (see Table 2.3 on page 34 for details) as affected haplotypes. These were chosen for the low correlations among their alleles.
2. Setting haplotypes 16 and 17 as affected haplotypes. These are two of the rarer haplotypes, with estimated frequency of 0.0007 each, which were least correlated. The rationale was that their rarity would ensure they were inherited IBD, and so should the SNPs surrounding them, thus creating a strong linkage signal around them.
3. Setting haplotypes 6, 16 and 17 as affected haplotypes. The addition here of another rare haplotype (haplotype 6, which was chosen to be as much different to the other two haplotypes as possible) was an attempt to further dilute the association signal from each individual SNP.
4. Randomly divide families into three groups, then *remove* one of the above affected haplotypes from each group (so, for example, group 1 would only have haplotypes 16 and 17 as their affected haplotypes). This was also an attempt to better balance the allelic association in each SNP.
5. Assigning each affected haplotype from strategy 3 as the affected haplotype for each group of families from strategy 4. This has similar rationale to strategy 4.
6. Randomly assign one of the 9 rare haplotypes (see Table 2.3) as the affected haplotype for each family. The randomisation process was in fact stratified, so that each haplotype was assigned roughly equally among the larger and the smaller families.

Despite the attempts, none of these strategies successfully produced a data set with a strong linkage effect but little or no association effect. Neither could this be achieved through adjustment of model parameters: it appeared that the linkage methods lost their power faster than FaST-LMM when the model parameters were changed.

As an example, results from the simulation using strategy 2 are shown here (Figures 7.4 and 7.5). All methods were able to detect a signal at the simulated locus, with RIA again appearing to have slightly better power than lm_ibdtests, but not as good as FaST-LMM. Results from simulations using other strategies follow a similar pattern and are not shown.

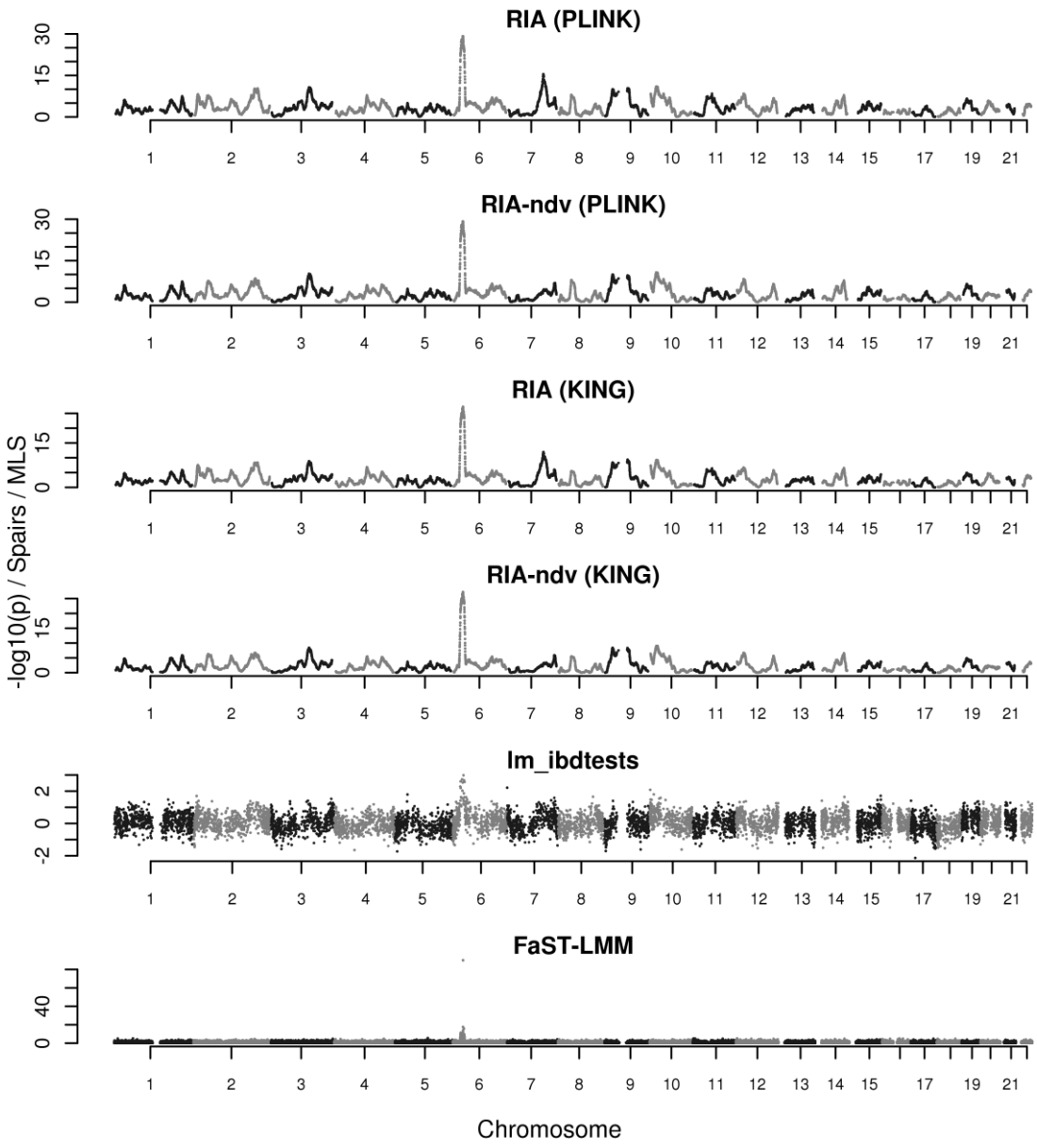

Figure 7.4 Manhattan plots for VL data set, haplotype-based simulated qualitative phenotype with gene dropdown, using various non-parametric linkage analysis and association methods. RIA = Regional IBD Analysis, RIA-ndv = RIA with dominance variance set to 0, (PLINK) = using IBD estimated by PLINK, (KING) = using IBD estimated by KING under homogeneous population assumption. The lone, extreme dot ($-\log_{10}(p) = 89.94$) in chromosome 6 of the FaST-LMM plot represents rs9271252, which is close to the simulated locus, and is not a plotting artefact.

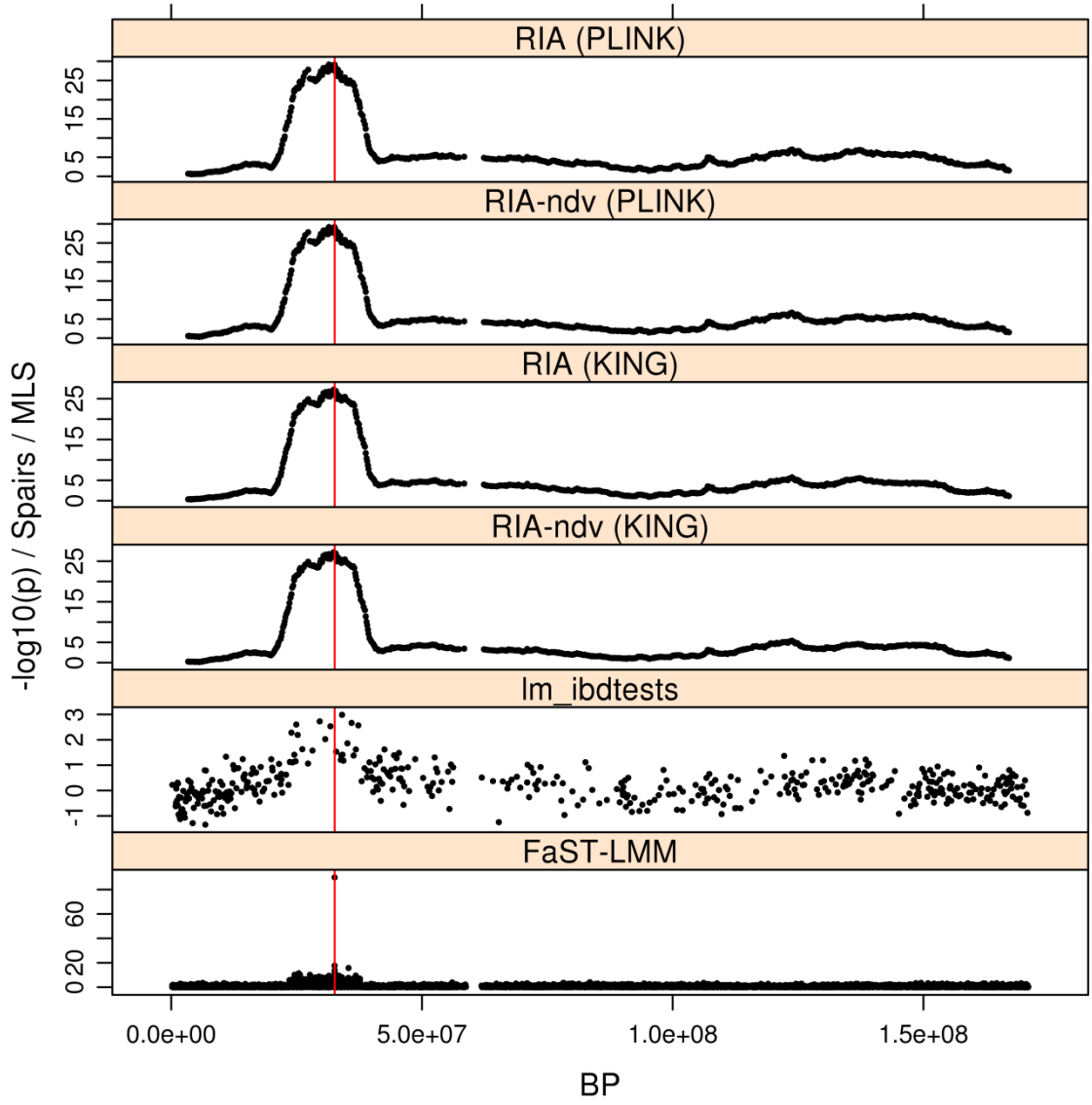

Figure 7.5 Comparisons of test statics for chromosome 6 of the VL data set, haplotype-based simulated qualitative phenotype with gene dropdown, using various non-parametric linkage analysis and association methods. RIA = Regional IBD Analysis, RIA-ndv = RIA with dominance variance set to 0, (PLINK) = using IBD estimated by PLINK, (KING) = using IBD estimated by KING under homogeneous population assumption. The lone, extreme dot ($-\log_{10}(p) = 89.94$) in chromosome 6 of the FaST-LMM plot represents rs9271252, which is close to the simulated locus, and is not a plotting artefact.

An analogous procedure to the previous section was used to assess RIA's overall power on this simulation. However, the gene dropdown had to be performed in each replicate before its phenotype can be simulated, with similar constraints to the original simulation, namely, the gene dropdown and phenotype simulation was repeated in each family until there were at least two affected individuals in that family. The null phenotype replicates were also created on a family-based basis: for each replicate, the phenotypes were permutated only within each family. This was done to preserve the total number of affected relative pairs that can be analysed, but a consequence is that

the null phenotypes are slightly correlated with the respective simulated phenotypes as there are limited ways to permutate the phenotypes in smaller families. An extreme case is a family with only two individuals (who both need to be affected due to the simulation constraint) in which permutation is not possible.

The null loci in this simulation were also selected differently. The second SNP window from each chromosome apart from chromosome 6 (which contains the effect locus) were analysed in each replicate, and all of their MLS statistics were used.

The histograms of these are shown in Figure 7.6. Again, the discriminatory power of RIA can be clearly seen. One difference between this and Figure 7.3 in the previous section is that, in this figure, the MLS from the simulated phenotype / null locus tend to be lower than those from the null phenotype / effect locus replicates. This is because of the restriction in the phenotype permutation which means there is still some effect remaining in the null phenotypes. On the other hand, because the null loci now came from 21 independent locations, the correlation with the effect locus is much less than that seen in the previous section where only a single null locus was used.

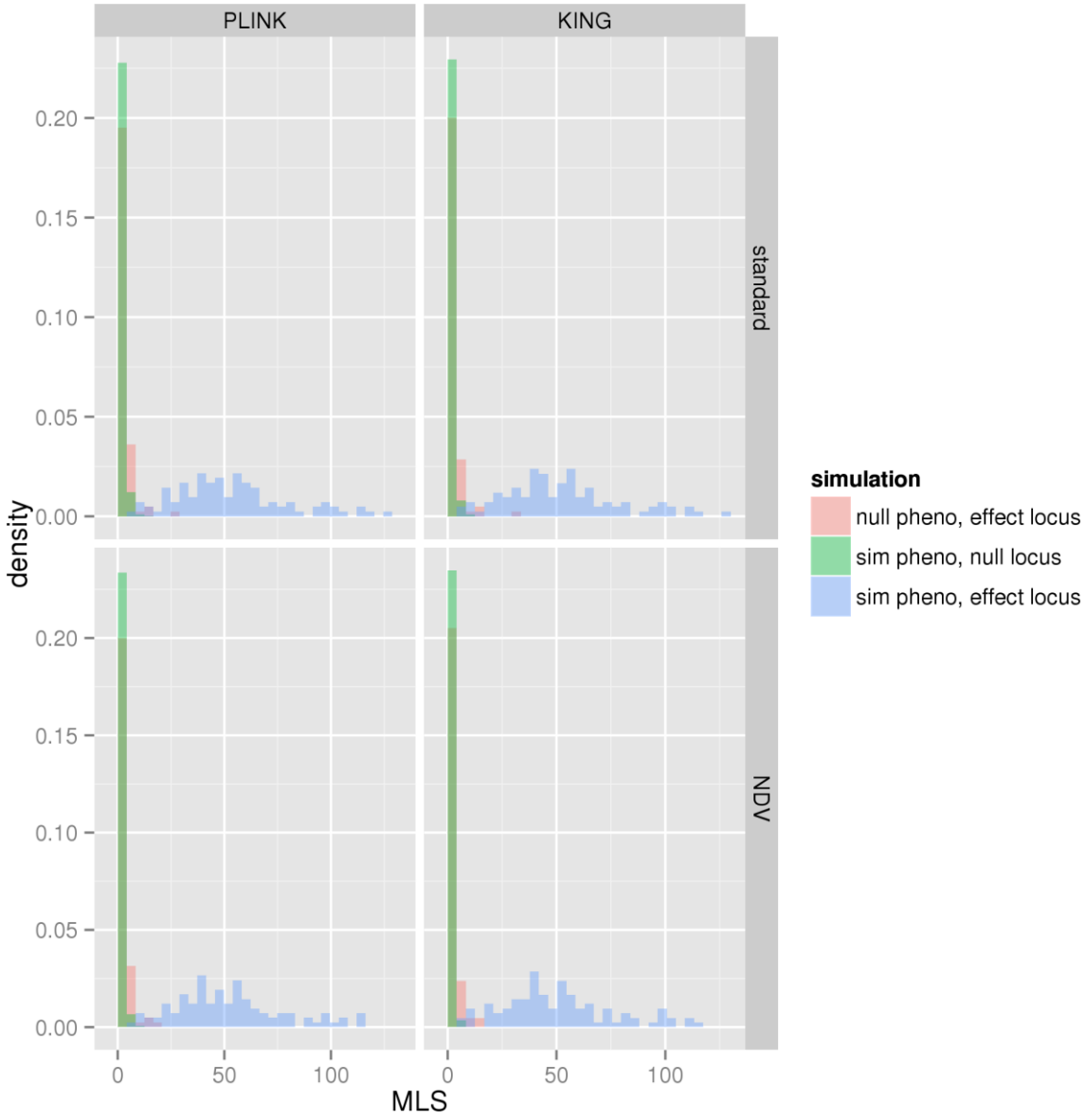

Figure 7.6 Histograms of MLS statistics from various RIA analyses of VL data set with haplotype-based simulated qualitative phenotypes. Each panel shows histograms of the MLS statistics calculated using respective RIA methods (standard or no-dominance-variance (NDV)) and IBD estimation methods (PLINK or KING), with either simulated 'disease' or 'null' phenotypes, at either the null or the simulated effect locus; each with 100 simulation replicates.

7.3. Discussion

This chapter demonstrates the success of RIA in detecting linkage signals from different type of simulations. However, there are two rather striking and unintended observations to be made here.

Firstly, it was not possible, by using the various strategies employed here, to recreate a situation similar to that observed in the original VUR data set, namely, the presence of a linkage signal without an association signal. To achieve this may require a more drastic measure such as artificially constructing haplotypes so that the affected

haplotypes are totally uncorrelated. However, this was decided against in this thesis, as the aim was also to use 'realistic' genotypes, so that RIA's performance can be studied in 'real life' data.

Secondly, regardless of the simulation settings, FaST-LMM (and, by extension, LMM GWAS programs) always outperformed both RIA and lm_ibdtests. This was of course because the association signal was still present. In fact, this observation and the fact that the linkage analysis methods lost power faster than FaST-LMM when model parameters were adjusted may be seen as analogous to the analysis made by Risch and Merikangas (1996) that association analysis tended to be more powerful than linkage analysis (however, their analysis was based on an affected sib pairs method vs transmission disequilibrium and with a moderate effect locus, so may not be fully applicable here). Nevertheless, if there is a data set in which there is linkage but not association, RIA is expected to perform well and probably better than lm_ibdtests, with an additional advantage of being more computationally efficient.

An additional advantage of RIA is that, at least in theory, it does not require any knowledge about the pedigrees. This could be useful, perhaps, in a situation where accurate determination of pedigree relatedness is difficult. In fact, this also seems to be the motivation for a method proposed by Day-Williams *et al.* (2011a), which also uses genetically estimated global and local kinships in variance component analysis of a quantitative trait (in an approach quite similar to that of Nagamine *et al.* (2012), although the motivation and the methods used for kinship calculation differ)

At present, RIA is still a work in progress. The idea seems promising, but there are still many aspects to explore.

One issue with RIA that may need to be explored is the assessment of significance, although in practice this may not be entirely necessary, as previously discussed (Section 6.5). Another is the IBD estimation. Although both PLINK and KING seemed to work reasonably well here, there are both theoretical and practical issues with them. The theoretical issue is that the current methods used for IBD estimation assume that the markers are independent—which is likely to be wrong when estimating local IBD using dense genome-wide data—and also requires accurate allele frequency estimation (Purcell *et al.*, 2007; Manichaikul *et al.*, 2010; Browning and Browning, 2011; Han and Abney, 2011). The practical issues, which are specific to PLINK (abrupt termination upon encountering a pair of individuals without any non-missing SNP in common) and KING (inability to use externally-estimated allele frequencies for population-based IBD estimations), have been described in the previous chapter (Section 6.1.3). To solve these would require modification of the programs. An alternative is perhaps to try other IBD estimation methods that allow linkage disequilibrium between markers such as

those proposed by Browning and Browning (2011) or Han and Abney (2011). However, since these methods use hidden Markov models (HMMs) and may take longer to run, a good compromise may be to use them only for local IBD estimation where each set of markers will be smaller but are more likely to be in linkage disequilibrium, while the global IBD can be calculated using simpler methods based on a pruned set of SNPs which will be in linkage equilibrium (an approach conceptually similar to that of Day-Williams *et al.* (2011a) where different procedures were used for the estimation of global and local kinship coefficients).

Another area that could be explored is the application of genetically estimated IBD to other non-parametric linkage analysis methods, particularly the 'score' methods as these only require a single IBD measure (instead of three IBD states as in the MLS methods), and would allow more IBD estimation methods to be used.

Chapter 8. Discussion and Conclusion

This chapter will focus on discussing issues regarding the use of theoretical and empirical IBD estimates that are relevant to more than one chapter. Discussion specific to any analysis method is in the relevant chapter.

8.1. Discussion

In this thesis, the utility of empirical (genetically estimated) and theoretical IBD—either as probabilities of IBD states or as kinship coefficients—in two principal types of genetic data analysis method has been demonstrated. It would appear that empirical relatedness estimates have outperformed pedigree relatedness, clearly for the association studies, and also in some of the linkage analyses. This is perhaps due to the slight difference between *theoretical* (pedigree) relatedness and *realised* relatedness (which empirical relatedness directly measures) caused by Mendelian segregation (Guo, 1996; Weir *et al.*, 2006; Hayes *et al.*, 2009). As discussed in Chapter 4, the reason why this may cause an issue could be that—unlike when attempting to categorise individuals' relationships into a pedigree where empirical relatedness is only an approximation of pedigree relatedness— the concern in genetic data analysis is with the modelling of the genetic relatedness itself. Pedigree relatedness is then only an approximation and may not necessarily be correct (Nordborg, 2001), which may then lead to inaccuracies in downstream analyses.

Studies from the field of animal breeding have indeed shown that realised relatedness predicts trait values better than pedigree relatedness (Nejati-Javaremi *et al.*, 1997; Hayes *et al.*, 2009). Perhaps the situation here is similar. The best IBD estimators for use in genetic data analysis should, then, be ones that most precisely estimate the IBD based on the observed IBS data rather than ones that correlate most to pedigree relatedness (although the latter would of course have their own utility in some other ways), and relative merits of the methods should be judged accordingly.

Although the aim of this thesis is not to compare the merits of methods of kinship estimation (and therefore only some convenient selections of them were included), it can be seen that all kinship estimation methods used for the LMM GWAS analysis in Chapters 3-5 (including the LMM software's own methods) performed quite well in that context—perhaps with the exception of PLINK, which resulted in a higher inflation of test statistics than was obtained using other empirical methods or theoretical kinships (Section 4.4). This was quite puzzling, as the algorithm used in PLINK is more elaborate and should give better estimation of IBD than those used in other programs

except, perhaps, for KING; and since, according to Kang *et al.* (2008), the kinship estimates based directly on IBS (such as those from the various LMM GWAS software) tend to capture distant relatedness better, while those estimated based on IBD (such as those from PLINK) tend to capture recent relatedness better. In this data set, which contains strong family structure, LMM using PLINK's IBD estimates should have performed at least as well as other methods, even in presence of additional population stratification. Nevertheless, as discussed briefly in Chapter 4, the reason for this could be that PLINK's IBD estimates deviate most from the *realised* genetic correlations among pairs of individuals. More detailed discussion now follows.

The differences between PLINK and other methods of IBD estimation considered here (KING and native methods in each LMM GWAS software) are that PLINK estimates the probabilities of the three IBD sharing states first and constrains them so that their values are between 0 to 1, then uses these for the calculation of the proportion of alleles shared IBD (which under a probabilistic viewpoint is equivalent to the coefficient of relationship and equals twice the kinship coefficient (C. C. Li and Sacks, 1954; Ritland, 1996; Blouin, 2003)), which is again constrained to biologically plausible values (Purcell *et al.*, 2007) whereas other methods including KING derived their 'kinship coefficients' directly from the genotype data without constraint (Aulchenko *et al.*, 2007b; Kang *et al.*, 2010; Manichaikul *et al.*, 2010; Lippert *et al.*, 2011; Zhou and Stephens, 2012; Pirinen *et al.*, 2013).

Although it was first proposed under a probabilistic viewpoint, kinship coefficients can also be viewed as reflecting genetic correlation between two individuals (Ritland, 1996). It is in fact under this latter viewpoint that it is used in LMM modelling—to model the polygenic effect, and therefore the genetic correlation between a pair of individuals. Since a pair of individuals intuitively can never be less related than unrelated (this is not strictly true: individuals from different populations can be less related than unrelated individuals from the same population; the assumption of homogeneity is implicit in this statement), their *theoretical* genetic correlation can never be less than zero. However, this is not the case for the *realised* genetic correlation: due to population stratification as well as the stochastic nature of Mendelian segregation, a pair of individuals may have negative genetic correlation, which can be interpreted as their sharing fewer alleles than can be expected in unrelated individuals (Astle and Balding, 2009). Constraining the estimated probabilities of empirical IBD states also introduces similar types of errors. Hence, when PLINK attempts to reconcile its empirical kinship coefficients to the theoretical ones, the adjusted values may no longer accurately reflect the realised genetic correlations; and when used in LMM—which requires genetic correlations—some degrees of error can be expected in the results. The discrepancies between theoretical and realised genetic correlation could also be the

explanation of the inferior performance of LMM using theoretical kinship to that using PLINK's empirical kinship estimates: although constrained to realistic values, PLINK's IBD estimates would still reflect the underlying genetic correlation better than the theoretical kinship estimates, which could be more affected by population stratification as well as cryptic relatedness.

Purcell *et al.* (2007) in fact suggested that *unconstrained* IBD sharing estimates from PLINK be used for diagnosing sample and genotyping error or for detection of misspecification of family relationship. These estimates may also be more suitable for use in LMM GWAS analysis and may give the best results from this data set; however, it is not clear from PLINK's documentation how these estimates could be obtained. Nevertheless, perhaps the conclusion that can be drawn from this is that empirical kinship coefficients estimated from the genotype data (whether IBD or IBS) without constraint are the best to use in LMM GWAS analysis, and methods that attempt to recreate pedigree relatedness should be avoided in this context.

A different problem arises when the empirical IBD estimates are used in RIA's MLS calculation. This time, the IBD sharing estimates are used under the probabilistic paradigm: Onelocarp expects probabilities, and feeding negative values to Onelocarp gives undefined results. But because these IBD estimates are intrinsically calculated as genetic correlations, certain transformations and constraints are needed. For PLINK, this was done automatically through its own algorithm; For KING, this was done through a user-defined algorithm. Both gave similar results, and resulted in similar MLS scores among the RIA methods. However, this process could give rise to small inaccuracies, which could result in a slight disadvantage when compared with the exact methods that can calculate the IBD probabilities directly, but may be comparable with the MCMC methods which are also at a disadvantage due to their stochastic nature.

In the cross-sectional VL data set (Chapters 4 and 5), LMM GWAS programs performed quite similarly to each other when provided with the same set of SNPs for kinship coefficient estimation, but the performance of each program was significantly affected by the choice of SNPs used or by the use of theoretical kinship estimates (Section 4.2.3). As discussed in Chapter 4, this could be because the thinned set of SNPs did not have enough information to accurately model complete relationships within or between the pedigrees, whereas the theoretical kinships were affected by additional relatedness or population structure. Intuitively, the equivalence between the analyses using the full and the pruned sets of SNPs suggests that pruning only remove redundant markers from the full set while still retaining most of the information (in other words, the effective number of independent markers (Yang *et al.*, 2014) remains the same). However, when more SNPs were removed from the pruned set of SNPs, some

information would also be removed, which eventually leads to underperformance in the thinned SNP set.

On this note, it appears that the choice of SNPs to be used in kinship estimation may be more important than the estimation method itself. The SNPs used in kinship estimation should contain adequate information to capture the genetic relatedness within the data set. For practical reasons, I believe a pruned set of SNPs is the set that should be used in LMM GWAS if possible as it contains similar amount of information to the full data set, while requiring less computational time. This choice is of course also governed by the availability of each set of SNPs, given the similar performance between the full and pruned sets.

This, however, seems to be somewhat contradictory to the various strategies proposed by the developers of FaST-LMM to reduce the amount of SNPs that are required for kinship estimation (Lippert *et al.*, 2011; Listgarten *et al.*, 2012; Lippert *et al.*, 2013). Nevertheless, results from Zhou and Stephens (2012), using a strategy similar to the earlier version of FaST-LMM-Select (selecting top SNPs that are associated with the phenotype in unadjusted analysis), also showed inadequate control of the inflation of test statistics; and a detailed analysis of the effect of population stratification by Yang *et al.* (2014) showed that, with subtle population stratification, the SNPs provided by either the randomly selected, reduced SNPs (equivalent to the thinned set in this thesis) or the equivalent of the earlier version of FaST-LMM-Select (selecting top n SNPs according to unadjusted association) were inadequate to correct for the population stratification, whereas the later version of FaST-LMM-Select (based on out-of-sample prediction accuracy) tends to select a set of SNPs that maximises power rather than providing effective population structure correction. Yang *et al.* (2014) therefore recommended the use of all available pruned SNPs in analyses that are concerned about population stratification. It seems that the lower number of SNPs produced by these strategies did not adequately correct for the high degree of relatedness seen in the VL data set used in this thesis either. In fact, this could be expected, given that relatedness has higher dimensionality than population stratification (Hoffman, 2013). Given that the computational advantage of using the thinned set of SNPs is not noticeable, using the pruned set would be a more prudent choice.

When the empirical IBD estimates are used in non-parametric linkage analyses, there are two further issues that are worthy of consideration: linkage disequilibrium and uncertainty surrounding IBD estimation.

Although all the *global* IBD estimates (including those from the LMM chapters) were theoretically valid, the same could not be said for the *local* IBD. Apart from the issues discussed in Chapter 6, both PLINK and KING, being method of moments estimators,

require that the SNPs are in linkage equilibrium (Browning and Browning, 2011). Additionally, they also require a sufficient number of SNPs for the estimation to be stable. This is generally not a problem in global IBD estimation, where genome-wide SNPs are available and can be pruned down so that they are independent, but in local IBD estimation there are competing demands between using only SNPs that are in linkage equilibrium and having an estimation window that contain sufficient number of SNPs while not spanning too great distance. All these demands may not necessarily be satisfied at the same time. An interesting option to resolve this could be to use methods that can handle (or even utilise) linkage disequilibrium (e.g. Albers *et al.*, 2008; Browning and Browning, 2011; Han and Abney, 2011).

Since RIA needs to estimate both the global and local IBD probabilities from the genotype data, there is more uncertainty in its input than in a traditional linkage analysis method where prior IBD estimates can be obtained exactly from the pedigree, provided that the pedigree is accurate. This may put RIA at a slight disadvantage. On the other hand, if the pedigree is misspecified, or if there is cryptic relatedness among the founders, then RIA may perform better than the traditional methods.

8.2. Conclusions

This thesis has investigated the use of theoretical and empirical IBD estimates in LMM GWAS analyses and in a new non-parametric linkage analysis method. In LMM GWAS analyses, the IBD estimates are used in the form of kinship matrix to model genetic relatedness between individuals. This is an area where the empirical kinship estimates performed much better than the theoretical estimates. Under standard conditions, all LMM GWAS programs investigated appeared to work well, especially when given empirical kinship estimates. However, when encountering model misspecification through the use of simulated longitudinal data, differences among the programs began to show; even so, most were still successful at controlling type I error.

In RIA, a new non-parametric linkage analysis method, both the globally and locally estimated IBD probabilities are used for the calculation of MLS statistics in affected relative pairs. This has the advantage of being able to completely bypass even the most complex pedigree structure, resulting in a substantial improvement in speed, with an additional advantage of not requiring pedigree information and not affected by pedigree misspecification. Compared with exact methods of non-parametric linkage analysis, RIA seemed to have less power; but in large, complex pedigrees where exact method can no longer be used, RIA performed well and seemed slightly more powerful than an MCMC-based method which would otherwise be the only class of programs that can operate in that condition.

In all, empirical IBD estimates have been shown to be useful in genetic data analysis, and in most cases resulted in better performance than theoretical IBD estimates. Considering the continuous improvement in methodology as well as the rapid advance in computational capability, it is foreseeable that they will become a powerful tool in genetic data analysis in the near future.

References

Abecasis, G. R., Cherny, S. S., Cookson, W. O. and Cardon, L. R. (2002) 'Merlin--rapid analysis of dense genetic maps using sparse gene flow trees', *Nat Genet*, 30(1), pp. 97-101.

Abecasis, G. R. and Wigginton, J. E. (2005) 'Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers', *Am J Hum Genet*, 77(5), pp. 754-67.

Abney, M., Ober, C. and McPeek, M. S. (2002) 'Quantitative-trait homozygosity and association mapping and empirical genomewide significance in large, complex pedigrees: fasting serum-insulin level in the Hutterites', *Am J Hum Genet*, 70(4), pp. 920-34.

Albers, C. A., Stankovich, J., Thomson, R., Bahlo, M. and Kappen, H. J. (2008) 'Multipoint approximations of identity-by-descent probabilities for accurate linkage analysis of distantly related individuals', *Am J Hum Genet*, 82(3), pp. 607-22.

Alexander, D. H., Novembre, J. and Lange, K. (2009) 'Fast model-based estimation of ancestry in unrelated individuals', *Genome Res*, 19(9), pp. 1655-64.

Almasy, L., Dyer, T., Peralta, J., Jun, G., Wood, A., Fuchsberger, C., Almeida, M., Kent, J., Fowler, S., Blackwell, T., Puppala, S., Kumar, S., Curran, J., Lehman, D., Abecasis, G., Duggirala, R., Blangero, J. and The, T. D. G. C. (2014) 'Data for Genetic Analysis Workshop 18: human whole genome sequence, blood pressure, and simulated phenotypes in extended pedigrees', *BMC Proceedings*, 8(Suppl 1), p. S2.

Amin, N., van Duijn, C. M. and Aulchenko, Y. S. (2007) 'A genomic background based method for association analysis in related individuals', *PLoS One*, 2(12), p. e1274.

Anderson, A. D. and Weir, B. S. (2007) 'A maximum-likelihood method for the estimation of pairwise relatedness in structured populations', *Genetics*, 176(1), pp. 421-40.

Anderson, C. A., Pettersson, F. H., Clarke, G. M., Cardon, L. R., Morris, A. P. and Zondervan, K. T. (2010) 'Data quality control in genetic case-control association studies', *Nat Protoc*, 5(9), pp. 1564-73.

Anderson, D., Cordell, H. J., Fakiola, M., Francis, R. W., Syn, G., Scaman, E. S., Davis, E., Miles, S. J., McLeay, T., Jamieson, S. E. and Blackwell, J. M. (2015) 'First genome-wide association study in an Australian aboriginal population provides insights into genetic risk factors for body mass index and type 2 diabetes', *PLoS One*, 10(3), p. e0119333.

Astle, W. and Balding, D. J. (2009) 'Population Structure and Cryptic Relatedness in Genetic Association Studies', *Statistical Science*, 24(4), pp. 451-471.

Aulchenko, Y. S., de Koning, D. J. and Haley, C. (2007a) 'Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis', *Genetics*, 177(1), pp. 577-85.

Aulchenko, Y. S., Ripke, S., Isaacs, A. and van Duijn, C. M. (2007b) 'GenABEL: an R library for genome-wide association analysis', *Bioinformatics*, 23(10), pp. 1294-6.

Bacanu, S. A., Devlin, B. and Roeder, K. (2000) 'The power of genomic control', *Am J Hum Genet*, 66(6), pp. 1933-44.

Balding, D. J. (2006) 'A tutorial on statistical methods for population association studies', *Nat Rev Genet*, 7(10), pp. 781-91.

Balding, D. J. and Nichols, R. A. (1995) 'A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity', *Genetica*, 96(1-2), pp. 3-12.

Basu, S., Di, Y. and Thompson, E. A. (2008) 'Exact trait-model-free tests for linkage detection in pedigrees', *Ann Hum Genet*, 72(Pt 5), pp. 676-82.

Basu, S., Stephens, M., Pankow, J. S. and Thompson, E. A. (2010) 'A likelihood-based trait-model-free approach for linkage detection of binary trait', *Biometrics*, 66(1), pp. 205-13.

Bates, D., Maechler, M., Bolker, B. and Walker, S. (2014) *lme4: Linear mixed-effects models using Eigen and S4* (Version 1.1-7) [Computer program]. Available at: http://CRAN.R-project.org/package=lme4.

Bellenguez, C., Ober, C. and Bourgain, C. (2009) 'A multiple splitting approach to linkage analysis in large pedigrees identifies a linkage to asthma on chromosome 12', *Genet Epidemiol*, 33(3), pp. 207-16.

Bennett, B. J., Farber, C. R., Orozco, L., Kang, H. M., Ghazalpour, A., Siemers, N., Neubauer, M., Neuhaus, I., Yordanova, R., Guan, B., Truong, A., Yang, W. P., He, A., Kayne, P., Gargalovic, P., Kirchgessner, T., Pan, C., Castellani, L. W., Kostem, E., Furlotte, N., Drake, T. A., Eskin, E. and Lusis, A. J. (2010) 'A high-resolution association mapping panel for the dissection of complex traits in mice', *Genome Res*, 20(2), pp. 281-90.

Blouin, M. S. (2003) 'DNA-based methods for pedigree reconstruction and kinship analysis in natural populations', *Trends in Ecology & Evolution*, 18(10), pp. 503-511.

Boehnke, M. and Cox, N. J. (1997) 'Accurate inference of relationships in sib-pair linkage studies', *Am J Hum Genet*, 61(2), pp. 423-9.

Bourgain, C. and Zhang, Q. (2009) *KinInbcoef: Calculation of Kinship and Inbreeding Coefficients Based on Pedigree Information* [Computer program]. Available at: http://www.stat.uchicago.edu/~mcpeek/software/KinInbcoef/index.html.

Broman, K. W. and Weber, J. L. (1999) 'Long homozygous chromosomal segments in reference families from the centre d'Etude du polymorphisme humain', *American journal of human genetics*, 65(6), pp. 1493-500.

Browning, B. L. and Browning, S. R. (2011) 'A fast, powerful method for detecting identity by descent', *Am J Hum Genet*, 88(2), pp. 173-82.

Cardon, L. R. and Palmer, L. J. (2003) 'Population stratification and spurious allelic association', *Lancet*, 361(9357), pp. 598-604.

Chen, W. M. and Abecasis, G. R. (2007) 'Family-based association tests for genomewide association scans', *Am J Hum Genet*, 81(5), pp. 913-26.

Clayton, D. G., Walker, N. M., Smyth, D. J., Pask, R., Cooper, J. D., Maier, L. M., Smink, L. J., Lam, A. C., Ovington, N. R., Stevens, H. E., Nutland, S., Howson, J. M., Faham, M., Moorhead, M., Jones, H. B., Falkowski, M., Hardenbol, P., Willis, T. D. and Todd, J. A. (2005) 'Population structure, differential bias and genomic control in a large-scale, case-control association study', *Nat Genet*, 37(11), pp. 1243-6.

Cordell, H. J., Darlay, R., Charoen, P., Stewart, A., Gullett, A. M., Lambert, H. J., Malcolm, S., Feather, S. A., Goodship, T. H., Woolf, A. S., Kenda, R. B., Goodship, J. A. and Group, U. V. S. (2010) 'Whole-genome linkage and association scan in primary, nonsyndromic vesicoureteric reflux', *J Am Soc Nephrol*, 21(1), pp. 113-23.

Cordell, H. J., Topf, A., Mamasoula, C., Postma, A. V., Bentham, J., Zelenika, D., Heath, S., Blue, G., Cosgrove, C., Granados Riveron, J., Darlay, R., Soemedi, R., Wilson, I. J., Ayers, K. L., Rahman, T. J., Hall, D., Mulder, B. J., Zwinderman, A. H., van Engelen, K., Brook, J. D., Setchfield, K., Bu'Lock, F. A., Thornborough, C., O'Sullivan, J., Stuart, A. G., Parsons, J., Bhattacharya, S., Winlaw, D., Mital, S., Gewillig, M., Breckpot, J., Devriendt, K., Moorman, A. F., Rauch, A., Lathrop, G. M., Keavney, B. D. and Goodship, J. A. (2013) 'Genome-wide association study identifies loci on 12q24 and 13q32 associated with tetralogy of Fallot', *Hum Mol Genet*, 22(7), pp. 1473-81.

Cordell, H. J., Wedig, G. C., Jacobs, K. B. and Elston, R. C. (2000) 'Multilocus linkage tests based on affected relative pairs', *Am J Hum Genet*, 66(4), pp. 1273-86.

Cotterman, C. W. (1940) *A calculus for statistico-genetics*. Ohio State University [Online]. Available at: http://rave.ohiolink.edu/etdc/view?acc_num=osu1298297334.

Cudworth, A. G. and Woodrow, J. C. (1975) 'Evidence for HL-A-linked genes in "juvenile" diabetes mellitus', *Br Med J*, 3(5976), pp. 133-5.

Darlow, J. M., Dobson, M. G., Darlay, R., Molony, C. M., Hunziker, M., Green, A. J., Cordell, H. J., Puri, P. and Barton, D. E. (2014) 'A new genome scan for primary nonsyndromic vesicoureteric reflux emphasizes high genetic heterogeneity and shows linkage and association with various genes already implicated in urinary tract development', *Mol Genet Genomic Med*, 2(1), pp. 7-29.

Day-Williams, A. G., Blangero, J., Dyer, T. D., Lange, K. and Sobel, E. M. (2011a) 'Linkage analysis without defined pedigrees', *Genet Epidemiol*, 35(5), pp. 360-70.

Day-Williams, A. G., Blangero, J., Dyer, T. D., Lange, K. and Sobel, E. M. (2011b) 'Unifying ideas for non-parametric linkage analysis', *Hum Hered*, 71(4), pp. 267-80.

Devlin, B. and Roeder, K. (1999) 'Genomic control for association studies', *Biometrics*, 55(4), pp. 997-1004.

Dudbridge, F. (2008) 'Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data', *Hum Hered*, 66(2), pp. 87-98.

Dudbridge, F., Holmans, P. A. and Wilson, S. G. (2011) 'A flexible model for association analysis in sibships with missing genotype data', *Ann Hum Genet*, 75(3), pp. 428-38.

Ehm, M. and Wagner, M. (1998) 'A test statistic to detect errors in sib-pair relationships', *Am J Hum Genet*, 62(1), pp. 181-8.

Ellinger, T. (1920) 'On the numerical expression of the degree of inbreeding and relationship in a pedigree', *Am Nat*, 54(635), pp. 540-5.

Elston, R. C. (1998) 'Methods of linkage analysis--and the assumptions underlying them', *Am J Hum Genet*, 63(4), pp. 931-4.

Epstein, M. P., Duren, W. L. and Boehnke, M. (2000) 'Improved inference of relationship for pairs of individuals', *Am J Hum Genet*, 67(5), pp. 1219-31.

Eu-ahsunthornwattana, J., Howey, R. A. and Cordell, H. J. (2014a) 'Accounting for relatedness in family-based association studies: application to Genetic Analysis Workshop 18 data', *BMC Proc*, 8(Suppl 1 Genetic Analysis Workshop 18), p. S79.

Eu-ahsunthornwattana, J., Miller, E. N., Fakiola, M., Wellcome Trust Case Control, C., Jeronimo, S. M., Blackwell, J. M. and Cordell, H. J. (2014b) 'Comparison of methods to account for relatedness in genome-wide association studies with family-based data', *PLoS Genet*, 10(7), p. e1004445.

Ewens, W. J. and Spielman, R. S. (1995) 'The transmission/disequilibrium test: history, subdivision, and admixture', *Am J Hum Genet*, 57(2), pp. 455-64.

Fabregat-Traver, D., Sharapov, S. Z., Hayward, C., Rudan, I., Campbell, H., Aulchenko, Y. and Bientinesi, P. (2014) 'High-Performance Mixed Models Based Genome-Wide Association Analysis with omicABEL software', *F1000Research*, 3.

Fakiola, M., Strange, A., Cordell, H. J., Miller, E. N., Pirinen, M., Su, Z., Mishra, A., Mehrotra, S., Monteiro, G. R., Band, G., Bellenguez, C., Dronov, S., Edkins, S., Freeman, C., Giannoulatou, E., Gray, E., Hunt, S. E., Lacerda, H. G., Langford, C., Pearson, R., Pontes, N. N., Rai, M., Singh, S. P., Smith, L., Sousa, O., Vukcevic, D., Bramon, E., Brown, M. A., Casas, J. P., Corvin, A., Duncanson, A., Jankowski, J., Markus, H. S., Mathew, C. G., Palmer, C. N., Plomin, R., Rautanen, A., Sawcer, S. J., Trembath, R. C., Viswanathan, A. C., Wood, N. W., Wilson, M. E., Deloukas, P., Peltonen, L., Christiansen, F., Witt, C., Jeronimo, S. M., Sundar, S., Spencer, C. C., Blackwell, J. M. and Donnelly, P. (2013) 'Common variants in the HLA-DRB1-HLA-DQA1 HLA class II region are associated with susceptibility to visceral leishmaniasis', *Nat Genet*, 45(2), pp. 208-13.

Falk, C. T. and Rubinstein, P. (1987) 'Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations', *Annals of Human Genetics*, 51(Pt 3), pp. 227-33.

Freedman, M. L., Reich, D., Penney, K. L., McDonald, G. J., Mignault, A. A., Patterson, N., Gabriel, S. B., Topol, E. J., Smoller, J. W., Pato, C. N., Pato, M. T., Petryshen, T. L., Kolonel, L. N., Lander, E. S., Sklar, P., Henderson, B., Hirschhorn, J. N. and Altshuler, D. (2004) 'Assessing the impact of population stratification on genetic association studies', *Nat Genet*, 36(4), pp. 388-93.

Furlotte, N. A., Eskin, E. and Eyheramendy, S. (2012) 'Genome-wide association mapping with longitudinal data', *Genet Epidemiol*, 36(5), pp. 463-71.

George, A. W., Visscher, P. M. and Haley, C. S. (2000) 'Mapping quantitative trait loci in complex pedigrees: a two-step variance component approach', *Genetics*, 156(4), pp. 2081-92.

Goddard, M. E., Wray, N. R., Verbyla, K. and Visscher, P. M. (2009) 'Estimating Effects and Making Predictions from Genome-Wide Marker Data', *Statistical Science*, 24(4), pp. 517-529.

Greenwood, C. M. and Bull, S. B. (1999) 'Down-weighting of multiple affected sib pairs leads to biased likelihood-ratio tests, under the assumption of no linkage', *Am J Hum Genet*, 64(4), pp. 1248-52.

Guo, S. W. (1996) 'Variation in genetic identity among relatives', *Hum Hered*, 46(2), pp. 61-70.

Han, L. and Abney, M. (2011) 'Identity by descent estimation with dense genome-wide genotype data', *Genet Epidemiol*, 35(6), pp. 557-67.

Hayes, B. J., Visscher, P. M. and Goddard, M. E. (2009) 'Increased accuracy of artificial selection by using the realized relationship matrix', *Genet Res (Camb)*, 91(1), pp. 47-60.

Henderson, C. R., Kempthorne, O., Searle, S. R. and Krosigk, C. M. v. (1959) 'The Estimation of Environmental and Genetic Trends from Records Subject to Culling', *Biometrics*, 15(2), pp. 192-218.

Hirschhorn, J. N. and Daly, M. J. (2005) 'Genome-wide association studies for common diseases and complex traits', *Nat Rev Genet*, 6(2), pp. 95-108.

Hoffman, G. E. (2013) 'Correcting for Population Structure and Kinship Using the Linear Mixed Model: Theory and Extensions', *PLoS ONE*, 8(10), p. e75707.

Horvath, S., Xu, X. and Laird, N. M. (2001) 'The family based association test method: strategies for studying general genotype--phenotype associations', *Eur J Hum Genet*, 9(4), pp. 301-6.

Howey, R. and Cordell, H. J. (2011) *MapThin* (Version 1.02) [Computer program]. Available at: http://www.staff.ncl.ac.uk/richard.howey/mapthin/.

Jacquard, A. (1972) 'Genetic information given by a relative', *Biometrics*, 28(4), pp. 1101-14.

Jakobsdottir, J. and McPeek, M. S. (2013) 'MASTOR: mixed-model association mapping of quantitative traits in samples with related individuals', *Am J Hum Genet*, 92(5), pp. 652-66.

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S. Y., Freimer, N. B., Sabatti, C. and Eskin, E. (2010) 'Variance component model to account for sample structure in genome-wide association studies', *Nat Genet*, 42(4), pp. 348-54.

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J. and Eskin, E. (2008) 'Efficient control of population structure in model organism association mapping', *Genetics*, 178(3), pp. 1709-23.

Kennedy, B. W., Quinton, M. and van Arendonk, J. A. (1992) 'Estimation of effects of single genes on quantitative traits', *J Anim Sci*, 70(7), pp. 2000-12.

Knowler, W. C., Williams, R. C., Pettitt, D. J. and Steinberg, A. G. (1988) 'Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture', *Am J Hum Genet*, 43(4), pp. 520-6.

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

Kruglyak, L., Daly, M. J., Reeve-Daly, M. P. and Lander, E. S. (1996) 'Parametric and nonparametric linkage analysis: a unified multipoint approach', *Am J Hum Genet*, 58(6), pp. 1347-63.

Laird, N. M., Horvath, S. and Xu, X. (2000) 'Implementing a unified approach to family-based tests of association', *Genet Epidemiol*, 19 Suppl 1, pp. S36-42.

Lake, S. L., Blacker, D. and Laird, N. M. (2000) 'Family-based tests of association in the presence of linkage', *Am J Hum Genet*, 67(6), pp. 1515-25.

Lander, E. S. and Green, P. (1987) 'Construction of multilocus genetic linkage maps in humans', *Proc Natl Acad Sci U S A*, 84(8), pp. 2363-7.

Lander, E. S. and Schork, N. J. (1994) 'Genetic dissection of complex traits', *Science*, 265(5181), pp. 2037-48.

Lange, C., DeMeo, D., Silverman, E. K., Weiss, S. T. and Laird, N. M. (2004) 'PBAT: tools for family-based association studies', *Am J Hum Genet*, 74(2), pp. 367-9.

Lange, K., Papp, J. C., Sinsheimer, J. S., Sripracha, R., Zhou, H. and Sobel, E. M. (2013) 'Mendel: the Swiss army knife of genetic analysis programs', *Bioinformatics*, 29(12), pp. 1568-70.

Li, C. C. and Sacks, L. (1954) 'The Derivation of Joint Distribution and Correlation between Relatives by the Use of Stochastic Matrices', *Biometrics*, 10(3), pp. 347-360.

Li, Q. and Yu, K. (2008) 'Improved correction for population stratification in genome-wide association studies by identifying hidden population structures', *Genet Epidemiol*, 32(3), pp. 215-26.

Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I. and Heckerman, D. (2011) 'FaST linear mixed models for genome-wide association studies', *Nat Methods*, 8(10), pp. 833-5.

Lippert, C., Quon, G., Kang, E. Y., Kadie, C. M., Listgarten, J. and Heckerman, D. (2013) 'The benefits of selecting phenotype-specific variants for applications of mixed models in genomics', *Sci Rep*, 3, p. 1815.

Listgarten, J., Lippert, C., Kadie, C. M., Davidson, R. I., Eskin, E. and Heckerman, D. (2012) 'Improved linear mixed models for genome-wide association studies', *Nature Methods*, 9(6), pp. 525-526.

Liu, N., Zhao, H., Patki, A., Limdi, N. A. and Allison, D. B. (2011) 'Controlling Population Structure in Human Genetic Association Studies with Samples of Unrelated Individuals', *Stat Interface*, 4(3), pp. 317-326.

Lynch, M. and Ritland, K. (1999) 'Estimation of pairwise relatedness with molecular markers', *Genetics*, 152(4), pp. 1753-1766.

Malécot, G. (1969) *The mathematics of heredity*. revised English edn. Translated by Yermanos, D. M. San Francisco: W. H. Freeman.

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M. and Chen, W. M. (2010) 'Robust relationship inference in genome-wide association studies', *Bioinformatics*, 26(22), pp. 2867-73.

Marchini, J., Cardon, L. R., Phillips, M. S. and Donnelly, P. (2004) 'The effects of human population structure on large genetic association studies', *Nat Genet*, 36(5), pp. 512-7.

Marchini, J., Howie, B., Myers, S., McVean, G. and Donnelly, P. (2007) 'A new multipoint method for genome-wide association studies by imputation of genotypes', *Nat Genet*, 39(7), pp. 906-13.

Martin, E. R., Monks, S. A., Warren, L. L. and Kaplan, N. L. (2000) 'A test for linkage and association in general pedigrees: the pedigree disequilibrium test', *Am J Hum Genet*, 67(1), pp. 146-54.

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. and Hirschhorn, J. N. (2008) 'Genome-wide association studies for complex traits: consensus, uncertainty and challenges', *Nat Rev Genet*, 9(5), pp. 356-69.

McPeek, M. S. (1999) 'Optimal allele-sharing statistics for genetic mapping using affected relatives', *Genet Epidemiol*, 16(3), pp. 225-49.

McPeek, M. S. and Sun, L. (2000) 'Statistical tests for detection of misspecified relationships by use of genome-screen data', *Am J Hum Genet*, 66(3), pp. 1076-94.

Menozzi, P., Piazza, A. and Cavallisforza, L. (1978) 'Synthetic Maps of Human Gene-Frequencies in Europeans', *Science*, 201(4358), pp. 786-792.

Meunier, F., Philippi, A., Martinez, M. and Demenais, F. (1997) 'Affected sib-pair tests for linkage: type I errors with dependent sib-pairs', *Genet Epidemiol*, 14(6), pp. 1107-11.

Milligan, B. G. (2003) 'Maximum-likelihood estimation of relatedness', *Genetics*, 163(3), pp. 1153-67.

Mitchell, B. D., Kammerer, C. M., Blangero, J., Mahaney, M. C., Rainwater, D. L., Dyke, B., Hixson, J. E., Henkel, R. D., Sharp, R. M., Comuzzie, A. G., VandeBerg, J. L., Stern, M. P. and MacCluer, J. W. (1996) 'Genetic and environmental contributions to cardiovascular risk factors in Mexican Americans. The San Antonio Family Heart Study', *Circulation*, 94(9), pp. 2159-70.

Morton, N. E. (1969) 'Preface to Probabilités et Hérédité (reprinted from French edition, 1966)', in *The mathematics of heredity*. revised English edn. San Francisco: W. H. Freeman.

Nagamine, Y., Pong-Wong, R., Navarro, P., Vitart, V., Hayward, C., Rudan, I., Campbell, H., Wilson, J., Wild, S., Hicks, A. A., Pramstaller, P. P., Hastie, N., Wright, A. F. and Haley, C. S. (2012) 'Localising loci underlying complex trait variation using Regional Genomic Relationship Mapping', *PLoS One*, 7(10), p. e46501.

Nejati-Javaremi, A., Smith, C. and Gibson, J. P. (1997) 'Effect of total allelic relationship on accuracy of evaluation and response to selection', *Journal of Animal Science*, 75(7), pp. 1738-1745.

Newman, D. L., Abney, M., McPeek, M. S., Ober, C. and Cox, N. J. (2001) 'The importance of genealogy in determining genetic associations with complex traits', *American journal of human genetics*, 69(5), pp. 1146-8.

Nordborg, M. (2001) 'Coalescent theory', in Balding, D. J., Bishop, M. and Cannings, C. (eds.) *Handbook of statistical genetics*. Chichester: John Wiley & Son, pp. 179-212.

Nyholt, D. R. (2008) 'Principles of linkage analysis', in Neale, B. M., Ferreira, M. A., Medland, S. E. and Posthuma, D. (eds.) *Statistical genetics : gene mapping through linkage and association*. New York: Taylor & Francis Group, pp. 113-134.

Oliehock, P. A., Windig, J. J., van Arendonk, J. A. M. and Bijma, P. (2006) 'Estimating relatedness between individuals in general populations with a focus on their use in conservation programs', *Genetics*, 173(1), pp. 483-496.

Ott, J. (1999) *Analysis of human genetic linkage*. 3rd edn. Baltimore: Johns Hopkins University Press.

Ott, J., Kamatani, Y. and Lathrop, M. (2011) 'Family-based designs for genome-wide association studies', *Nat Rev Genet*, 12(7), pp. 465-474.

Pardo, L. M., MacKay, I., Oostra, B., van Duijn, C. M. and Aulchenko, Y. S. (2005) 'The effect of genetic drift in a young genetically isolated population', *Ann Hum Genet*, 69(Pt 3), pp. 288-95.

Patterson, N., Price, A. L. and Reich, D. (2006) 'Population structure and eigenanalysis', *PLoS Genet*, 2(12), p. e190.

Pearl, R. (1914) 'Studies on inbreeding. V. Inbreeding and relationship coefficients', *Am Nat*, 48(573), pp. 513-23.

Peloso, G. M., Dupuis, J. and Lunetta, K. L. (2011) 'Evaluation of methods accounting for population structure with pedigree data and continuous outcomes', *Genet Epidemiol*, 35(6), pp. 427-36.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. and R Core Team (2013) *nlme: Linear and Nonlinear Mixed Effects Models* (Version 3.1-108) [Computer program]. Available at: http://CRAN.R-project.org/package=nlme.

Pirinen, M., Donnelly, P. and Spencer, C. C. A. (2013) 'Efficient Computation with a Linear Mixed Model on Large-Scale Data Sets with Applications to Genetic Studies', *Annals of Applied Statistics*, 7(1), pp. 369-390.

Powell, J. E., Visscher, P. M. and Goddard, M. E. (2010) 'Reconciling the analysis of IBD and IBS in complex trait studies', *Nature Reviews Genetics*, 11(11), pp. 800-805.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A. and Reich, D. (2006) 'Principal components analysis corrects for stratification in genome-wide association studies', *Nat Genet*, 38(8), pp. 904-9.

Price, A. L., Zaitlen, N. A., Reich, D. and Patterson, N. (2010) 'New approaches to population stratification in genome-wide association studies', *Nat Rev Genet*, 11(7), pp. 459-63.

Pritchard, J. K. and Donnelly, P. (2001) 'Case-control studies of association in structured or admixed populations', *Theor Popul Biol*, 60(3), pp. 227-37.

Pritchard, J. K. and Rosenberg, N. A. (1999) 'Use of unlinked genetic markers to detect population stratification in association studies', *Am J Hum Genet*, 65(1), pp. 220-8.

Pritchard, J. K., Stephens, M. and Donnelly, P. (2000a) 'Inference of population structure using multilocus genotype data', *Genetics*, 155(2), pp. 945-59.

Pritchard, J. K., Stephens, M., Rosenberg, N. A. and Donnelly, P. (2000b) 'Association mapping in structured populations', *Am J Hum Genet*, 67(1), pp. 170-81.

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

Queller, D. C. and Goodnight, K. F. (1989) 'Estimating Relatedness Using Genetic Markers', *Evolution*, 43(2), pp. 258-275.

Rabinowitz, D. and Laird, N. (2000) 'A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information', *Human Heredity*, 50(4), pp. 211-23.

Rakovski, C. S. and Stram, D. O. (2009) 'A Kinship-Based Modification of the Armitage Trend Test to Address Hidden Population Structure and Small Differential Genotyping Errors', *PLoS One*, 4(6).

Reich, D. E. and Goldstein, D. B. (2001) 'Detecting association in a case-control study while correcting for population stratification', *Genet Epidemiol*, 20(1), pp. 4-16.

Risch, N. (1990) 'Linkage strategies for genetically complex traits. II. The power of affected relative pairs', *Am J Hum Genet*, 46(2), pp. 229-41.

Risch, N. and Merikangas, K. (1996) 'The future of genetic studies of complex human diseases', *Science*, 273(5281), pp. 1516-7.

Ritland, K. (1996) 'Estimators for pairwise relatedness and individual inbreeding coefficients', *Genet Res (Camb)*, 67, pp. 175-185.

Rosenberg, N. A. and Nordborg, M. (2006) 'A general population-genetic model for the production by population structure of spurious genotype-phenotype associations in discrete, admixed or spatially distributed populations', *Genetics*, 173(3), pp. 1665-78.

Sawcer, S., Hellenthal, G., Pirinen, M., Spencer, C. C., Patsopoulos, N. A., Moutsianas, L., Dilthey, A., Su, Z., Freeman, C., Hunt, S. E., Edkins, S., Gray, E., Booth, D. R., Potter, S. C., Goris, A., Band, G., Oturai, A. B., Strange, A., Saarela, J., Bellenguez, C., Fontaine, B., Gillman, M., Hemmer, B., Gwilliam, R., Zipp, F., Jayakumar, A., Martin, R., Leslie, S., Hawkins, S., Giannoulatou, E., D'Alfonso, S., Blackburn, H., Martinelli Boneschi, F., Liddle, J., Harbo, H. F., Perez, M. L., Spurkland, A., Waller, M. J., Mycko, M. P., Ricketts, M., Comabella, M., Hammond, N., Kockum, I., McCann, O. T., Ban, M., Whittaker, P., Kemppinen, A., Weston, P., Hawkins, C., Widaa, S., Zajicek, J., Dronov, S., Robertson, N., Bumpstead, S. J., Barcellos, L. F., Ravindrarajah, R., Abraham, R., Alfredsson, L., Ardlie, K., Aubin, C., Baker, A., Baker, K., Baranzini, S. E., Bergamaschi, L., Bergamaschi, R., Bernstein, A., Berthele, A., Boggild, M., Bradfield, J. P., Brassat, D., Broadley, S. A., Buck, D., Butzkueven, H., Capra, R., Carroll, W. M., Cavalla, P., Celius, E. G., Cepok, S., Chiavacci, R., Clerget-Darpoux, F., Clysters, K., Comi, G., Cossburn, M., Cournu-Rebeix, I., Cox, M. B., Cozen, W., Cree, B. A., Cross, A. H., Cusi, D., Daly, M. J., Davis, E., de Bakker, P. I., Debouverie, M., D'Hooghe M, B., Dixon, K., Dobosi, R., Dubois, B., Ellinghaus, D., Elovaara, I., Esposito, F., et al. (2011) 'Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis', *Nature*, 476(7359), pp. 214-9.

Sham, P. (1997) *Statistics in human genetics*. London: Arnold.

Shih, M.-C. and Whittemore, A. S. (2001) 'Allele-sharing among affected relatives: non-parametric methods for identifying genes', *Statistical Methods in Medical Research*, 10(1), pp. 27-55.

Sieh, W., Basu, S., Fu, A. Q., Rothstein, J. H., Scheet, P. A., Stewart, W. C., Sung, Y. J., Thompson, E. A. and Wijsman, E. M. (2005) 'Comparison of marker types and map assumptions using Markov chain Monte Carlo-based linkage analysis of COGA data', *BMC Genet*, 6 Suppl 1, p. S11.

Spielman, R. S., McGinnis, R. E. and Ewens, W. J. (1993) 'Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM)', *American journal of human genetics*, 52(3), pp. 506-16.

Suarez, B. K., Rice, J. and Reich, T. (1978) 'The generalized sib pair IBD distribution: its use in the detection of linkage', *Ann Hum Genet*, 42(1), pp. 87-94.

Sun, L., Wilder, K. and McPeek, M. S. (2002) 'Enhanced pedigree error detection', *Hum Hered*, 54(2), pp. 99-110.

Svishcheva, G. R., Axenovich, T. I., Belonogova, N. M., van Duijn, C. M. and Aulchenko, Y. S. (2012) 'Rapid variance components-based method for whole-genome association analysis', *Nat Genet*.

Terwilliger, J. D. and Ott, J. (1992) 'A haplotype-based 'haplotype relative risk' approach to detecting allelic associations', *Human Heredity*, 42(6), pp. 337-46.

The Wellcome Trust Case Control Consortium (2007) 'Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls', *Nature*, 447(7145), pp. 661-78.

Therneau, T. (2012) *coxme: Mixed Effects Cox Models* (Version 2.2-3) [Computer program]. Available at: http://CRAN.R-project.org/package=coxme.

Thompson, E. A. (1974) 'Gene identities and multiple relationships', *Biometrics*, 30(4), pp. 667-80.

Thompson, E. A. (1975) 'The estimation of pairwise relationships', *Ann Hum Genet*, 39(2), pp. 173-88.

Thornton, T. and McPeek, M. S. (2007) 'Case-control association testing with related individuals: a more powerful quasi-likelihood score test', *Am J Hum Genet*, 81(2), pp. 321-37.

Thornton, T. and McPeek, M. S. (2010) 'ROADTRIPS: case-control association testing with partially or completely unknown population and pedigree structure', *Am J Hum Genet*, 86(2), pp. 172-84.

Tian, C., Gregersen, P. K. and Seldin, M. F. (2008) 'Accounting for ancestry: population substructure and genome-wide association studies', *Hum Mol Genet*, 17(R2), pp. R143-50.

Trager, E. H., Khanna, R., Marrs, A., Siden, L., Branham, K. E., Swaroop, A. and Richards, J. E. (2007) 'Madeline 2.0 PDE: a new program for local and web-based pedigree drawing', *Bioinformatics*, 23(14), pp. 1854-6.

Visscher, Peter M., Brown, Matthew A., McCarthy, Mark I. and Yang, J. (2012) 'Five Years of GWAS Discovery', *The American Journal of Human Genetics*, 90(1), pp. 7-24.

Voight, B. F. and Pritchard, J. K. (2005) 'Confounding from cryptic relatedness in case-control association studies', *PLoS Genet*, 1(3), p. e32.

Wang, J. (2002) 'An estimator for pairwise relatedness using molecular markers', *Genetics*, 160(3), pp. 1203-15.

Weeks, D. E. and Lange, K. (1988) 'The affected-pedigree-member method of linkage analysis', *Am J Hum Genet*, 42(2), pp. 315-26.

Weir, B. S., Anderson, A. D. and Hepler, A. B. (2006) 'Genetic relatedness analysis: modern data and new challenges', *Nat Rev Genet*, 7(10), pp. 771-80.

Whittemore, A. S. and Halpern, J. (1994) 'A Class of Tests for Linkage Using Affected Pedigree Members', *Biometrics*, 50(1), pp. 118-127.

Wright, S. (1921) 'Systems of Mating. I. the Biometric Relations between Parent and Offspring', *Genetics*, 6(2), pp. 111-23.

Wright, S. (1922) 'Coefficients of inbreeding and relationship', *Am Nat*, 56, pp. 330-8.

Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. and Price, A. L. (2014) 'Advantages and pitfalls in the application of mixed-model association methods', *Nat Genet*, 46(2), pp. 100-106.

Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., Kresovich, S. and Buckler, E. S. (2006) 'A unified mixed-model method for association mapping that accounts for multiple levels of relatedness', *Nat Genet*, 38(2), pp. 203-8.

Zhang, S., Zhu, X. and Zhao, H. (2003) 'On a semiparametric test to detect associations between quantitative traits and candidate genes using unrelated individuals', *Genet Epidemiol*, 24(1), pp. 44-56.

Zhang, Z., Ersoz, E., Lai, C. Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., Bradbury, P. J., Yu, J., Arnett, D. K., Ordovas, J. M. and Buckler, E. S. (2010) 'Mixed linear model approach adapted for genome-wide association studies', *Nat Genet*, 42(4), pp. 355-60.

Zheng, G., Li, Z., Gail, M. H. and Gastwirth, J. L. (2010) 'Impact of population substructure on trend tests for genetic case-control association studies', *Biometrics*, 66(1), pp. 196-204.

Zhou, X. and Stephens, M. (2012) 'Genome-wide efficient mixed-model analysis for association studies', *Nat Genet*, 44(7), pp. 821-4.

Appendix

Published articles from this project:

A. Eu-ahsunthornwattana, J., Howey, R. A. and Cordell, H. J. (2014) 'Accounting for relatedness in family-based association studies: application to Genetic Analysis Workshop 18 data', BMC Proc, 8(Suppl 1 Genetic Analysis Workshop 18), p. S79.
B. Eu-ahsunthornwattana, J., Miller, E. N., Fakiola, M., Wellcome Trust Case Control, C., Jeronimo, S. M., Blackwell, J. M. and Cordell, H. J. (2014) 'Comparison of methods to account for relatedness in genome-wide association studies with family-based data', PLoS Genet, 10(7), p. e1004445.

Eu-ahsunthornwattana et al. BMC Proceedings 2014, 8(Suppl 1):S79
http://www.biomedcentral.com/1753-6561/8/S1/S79

BMC Proceedings

PROCEEDINGS Open Access

Accounting for relatedness in family-based association studies: application to Genetic Analysis Workshop 18 data

Jakris Eu-ahsunthornwattana[1,2], Richard AJ Howey[1], Heather J Cordell[1*]

From Genetic Analysis Workshop 18
Stevenson, WA, USA. 13-17 October 2012

Abstract

In the last few years, a bewildering variety of methods/software packages that use linear mixed models to account for sample relatedness on the basis of genome-wide genomic information have been proposed. We compared these approaches as implemented in the programs EMMAX, FaST-LMM, Gemma, and GenABEL (FASTA/GRAMMAR-Gamma) on the Genetic Analysis Workshop 18 data. All methods performed quite similarly and were successful in reducing the genomic control inflation factor to reasonable levels, particularly when the mean values of the observations were used, although more variation was observed when data from each time point were used individually. From a practical point of view, we conclude that it makes little difference to the results which method/software package is used, and the user can make the choice of package on the basis of personal taste or computational speed/convenience.

Background

A number of different methods/software packages have been proposed in the last few years that implement linear mixed-model approaches to account for population structure and relatedness among samples in genome-wide association studies (GWAS), but no detailed comparisons among them have been made before our effort. Indeed, when a new method/package is developed, it is often quite unclear whether or how it differs substantially from those already available. To address this question, we explored the performance of various implementations of such methods in the longitudinal Genetic Analysis Workshop 18 (GAW18) data set.

Methods

We analyzed the GAW18 GWAS data [1] using the real phenotypes and the first set of simulated phenotypes. This analysis was performed without knowledge of the underlying simulating model. The genotype data were cleaned using standard procedures [2]. This resulted in 4 individuals being excluded because of their total lack of genotype data, and another individual being excluded because of outlying ethnicity (Chinese [CHB] or Japanese [JPT]), leaving 954 individuals whose genotype data were used. We removed 43,987 monomorphic or low-frequency (minor allele frequency [MAF] <1%) single-nucleotide polymorphisms (SNPs), 109 SNPs with missing rate above 10% (this criterion took into account the apparently high missing rate in some SNPs likely to be caused by the differences in genotyping technology used in the samples), and 1 SNP that failed Hardy-Weinberg equilibrium testing in the control founder population. A total of 427,952 SNPs were retained for analysis.

We conducted linear regression of the real and simulated systolic blood pressure and simulated diastolic blood pressure at each time point regressed on age, medication, and smoking status. For the real diastolic blood pressure–which, as could be physiologically expected, seemed to have a nonlinear relationship with age–we used a quadratic regression, including age and age squared as predictors. The phenotype data from all individuals were used for these regressions. Residuals from

* Correspondence: heather.cordell@ncl.ac.uk
[1]Institute of Genetic Medicine, Newcastle University, Central Parkway, Newcastle upon Tyne, NE1 3BZ, UK
Full list of author information is available at the end of the article

BioMed Central

© 2014 Eu-ahsunthornwattana et al.; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

these regressions in subjects who also had genotype data were then used for the genome-wide analyses.

Genome-wide association analyses, adjusting for familial relatedness using genomic data, were performed using a variety of linear mixed model approaches. All approaches attempt to fit the model $\boldsymbol{Y}=\beta+Q+\boldsymbol{\varepsilon}$, where $\boldsymbol{Y}=(y_1, ..., y_n)^T$ is a vector of responses on n subjects; $X=(x_{ik})$ is the $n \times K$ matrix of predictor values for variables to be modeled as fixed effects (including covariates and genotypes at any SNPs currently under test); $\beta=(\beta_1, ... \beta_K)^T$ are regression coefficients (to be estimated) representing the linear effects of the predictors on the response; Q are random effects, $Q \sim N(0,2\sigma_g^2\Phi)$, and ε are random errors, $\varepsilon \sim N(0,\sigma_e^2 I)$, where σ_g^2 and σ_e^2 are parameters (to be estimated) representing the genetic and environmental components of variance respectively; Φ is the $n \times n$ matrix of

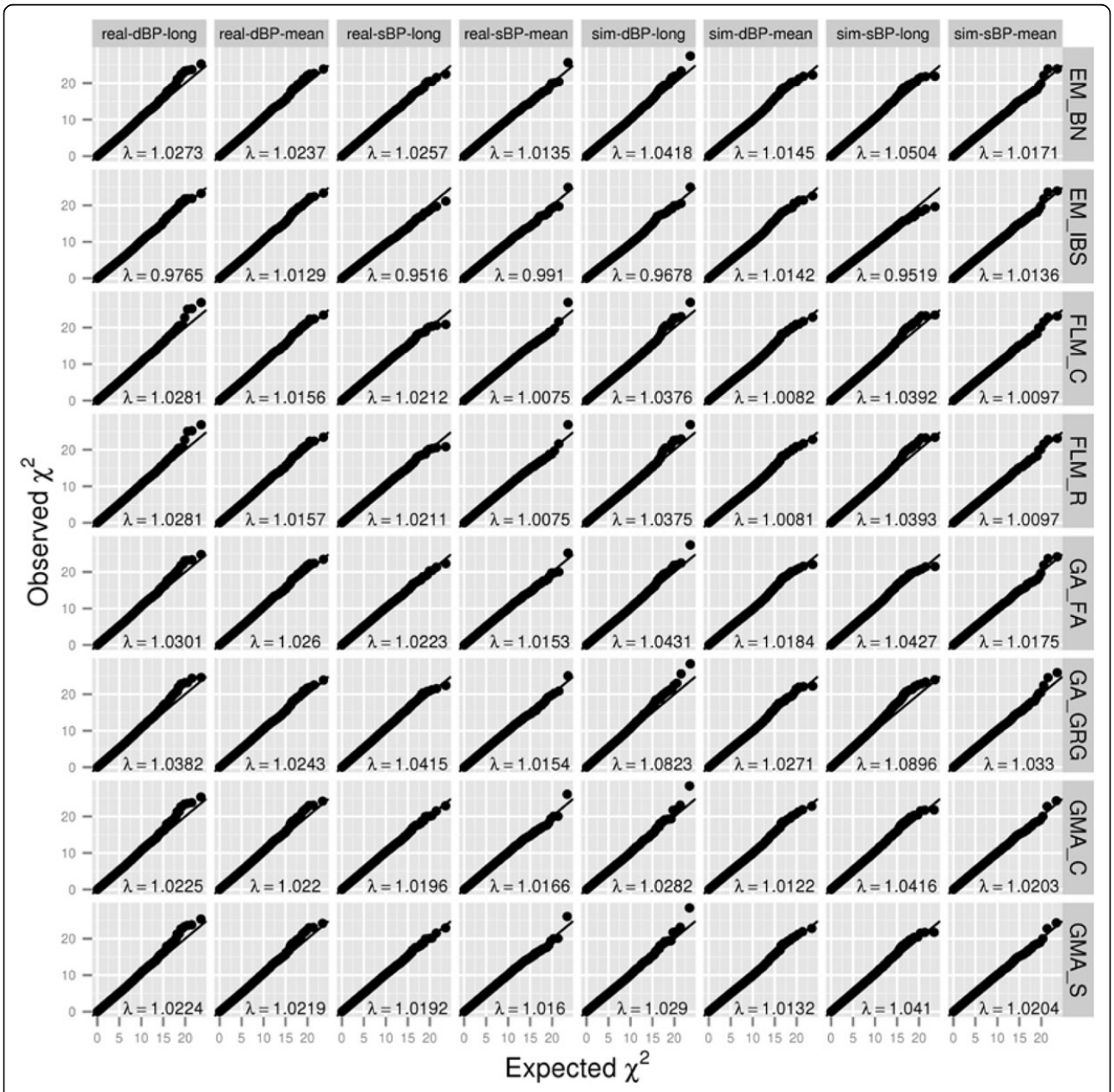

Figure 1 Q-Q plots and genomic inflation factors for different methods. These were calculated for each phenotype (real diastolic blood pressure [DBP], real systolic blood pressure [SBP], simulated DBP, and simulated SBP), using either longitudinal ("long") or average ("mean") residuals. EM_BN, EMMAX using Balding-Nichols matrix; EM_IBS, EMMAX using IBS matrix; FLM_C, FaST-LMM using standard covariance matrix; FLM_R, FaST-LMM using realized relationship matrix; GA_FA, GenABEL/FASTA; GA_GRG, GenABEL/GRAMMAR-Gamma; GMA_C, Gemma using centralized covariance matrix; GMA_S, Gemma using standardized covariance matrix. The diagonal line represents the identity line in each panel.

pairwise kinship coefficients; and I is the $n \times n$ identity matrix. The approaches vary with respect to precise details of the calculation of kinship or "relatedness" and with respect to whether an exact method or a fast approximation is used (for more details, see descriptions in references [3-9]). In each case we used a subset of 21,153 SNPs to perform the relatedness calculations, namely SNPs with MAF >0.4, <5% missing data, and "pruned" to be in approximate linkage equilibrium via the PLINK command "-indep 50 5 2". In analyses of other data sets we have found little difference between results when using such a pruned set of SNPs for calculating relatedness and when using the full set of SNPs (data not shown).

The methods considered were: (a) EMMAX [3], which implements 2 methods for relatedness calculations: one

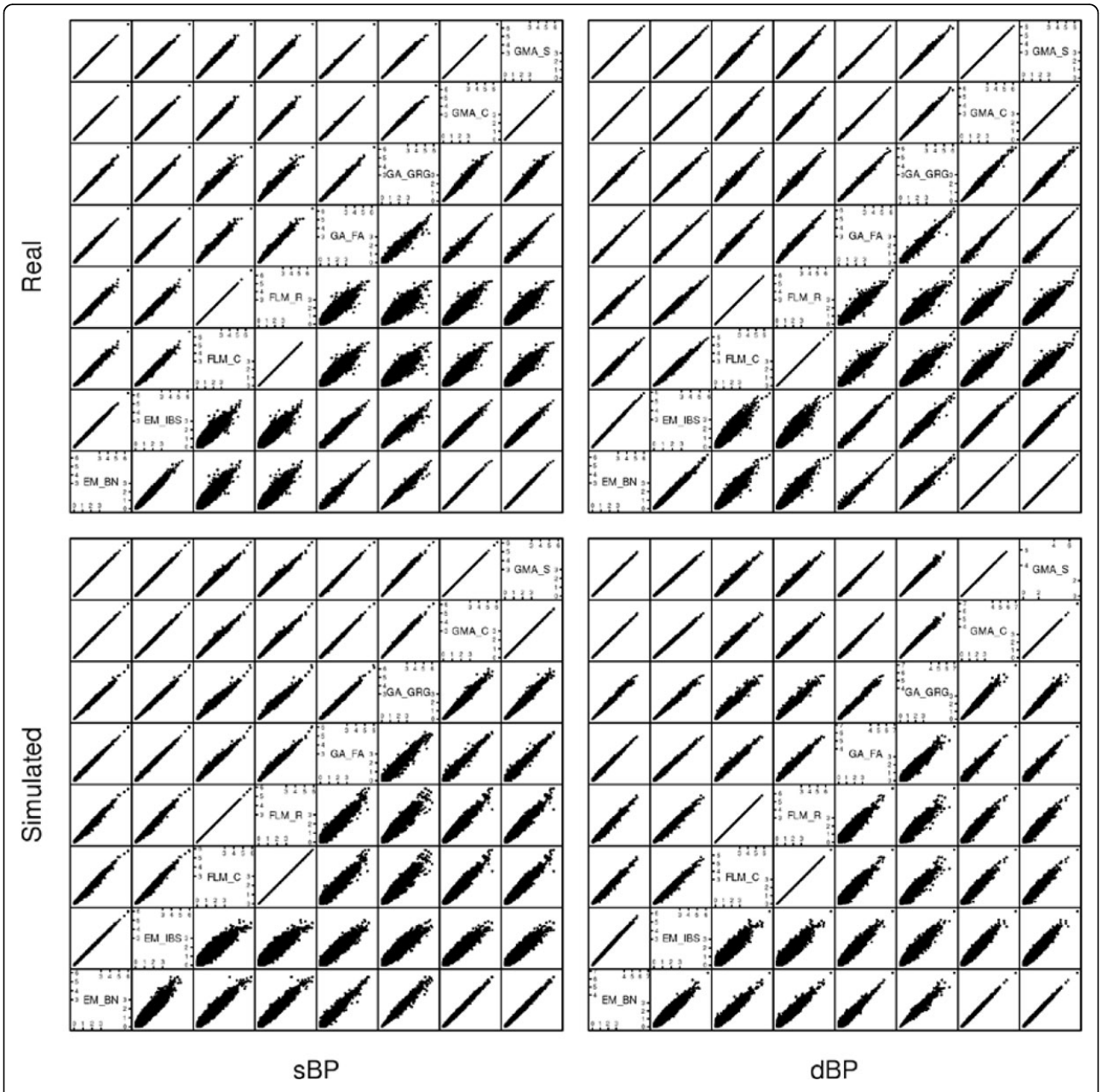

Figure 2 Comparison of $-\log_{10}$ *p* values at each SNP calculated using different methods. The upper triangles show the values based on mean residuals, while the lower triangles show the values calculated using longitudinal data. DBP, diastolic blood pressure; EM_BN, EMMAX using Balding-Nichols matrix; EM_IBS, EMMAX using IBS matrix; FLM_C, FaST-LMM using standard covariance matrix; FLM_R, FaST-LMM using realized relationship matrix; GA_FA, GenABEL/FASTA; GA_GRG, GenABEL/GRAMMAR-Gamma; GMA_C, Gemma using centralized covariance matrix; GMA_S, Gemma using standardized covariance matrix; SBP, systolic blood pressure.

based on identity-by-state (IBS) sharing and one based on the Balding-Nichols method [4]; (b) FaST-LMM [5], which also implements 2 methods to adjust for relatedness: one using a standard covariance matrix and one using the realized relationship matrix; (c) the polygenic/mmscore functions in GenABEL [6], which implement the FASTA method [7]; (d) the polygenic/grammar functions in GenABEL, which implement the GRAMMAR-Gamma approximation [8]; and (e) Gemma [9], which uses an efficient exact method. Simple linear regression without any relatedness adjustment was also performed in FaST-LMM. All analyses were performed using both the residual from each individual observation (modeled without regard to its true longitudinal nature, or *longitudinal*) and the mean of the residuals for each subject, or *mean*. Genomic inflation factors (λ) were calculated as proposed by Devlin and Roeder [10]. We also assessed the genomic inflation factors for unadjusted χ^2 and Cochran-Armitage trend tests of hypertension status at each time point as calculated using PLINK [11].

Results and discussion

Figure 1 shows the Q-Q plots and genomic inflation factors for different methods. It is well known that population substructure and relatedness will cause an inflated distribution of genome-wide association test statistics ($\lambda > 1.00$) if not appropriately modeled. All methods performed reasonably well for the mean residuals, controlling the λ to 0.99 to 1.03. For longitudinal data, most methods also performed well, with λ in the range of 0.95 to 1.05, except perhaps for GRAMMAR-Gamma, which achieved λs of approximately 1.08 to 1.09 for the simulated phenotypes. However, even these values were much less inflated compared to the λ values of 1.22 to 1.68 (mean) and 2.04 to 3.41 (longitudinal) seen in the unadjusted analyses. The higher inflation in longitudinal analyses (even when adjusting for relatedness) could be expected from the fact that additional (nongenetic) within-subject correlation was not allowed for in these analyses; indeed, one could argue that this behavior is statistically the "correct" behavior, with GRAMMAR-Gamma (which gave the highest

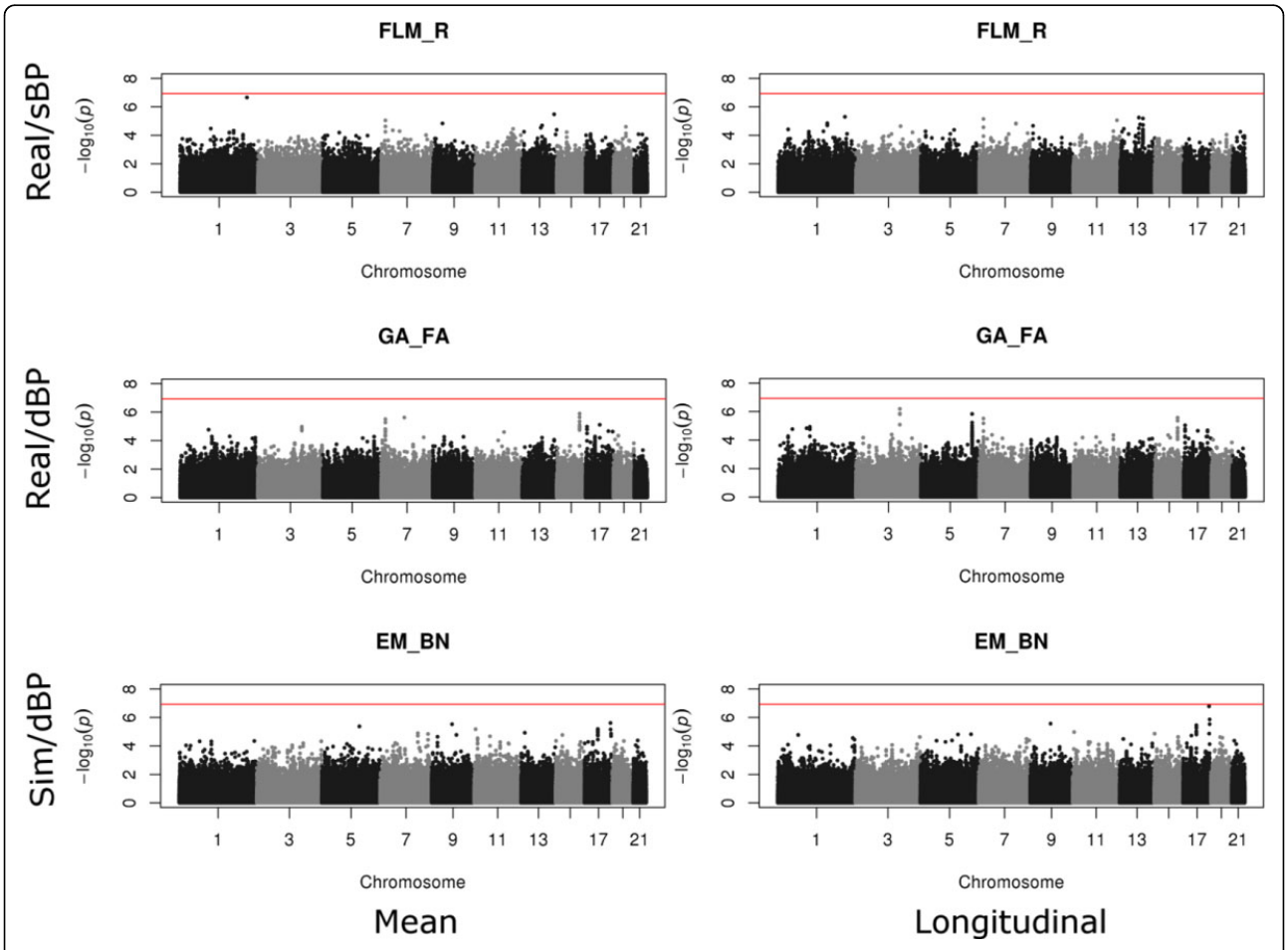

Figure 3 A selection of Manhattan plots showing *p* values calculated using various methods. DBP, diastolic blood pressure; EM_BN, EMMAX using Balding-Nichols matrix; FLM_R, FaST-LMM using realized relationship matrix; GA_FA, GenABEL/FASTA; SBP, systolic blood pressure.

inflation) showing the "most correct" behavior. Interestingly, EMMAX using the IBS matrix seemed to have the opposite behavior, for reasons we are currently unable to determine.

For the analyses using hypertension status, the unadjusted genomic inflations were between 1.21 and 1.55 for the Cochran-Armitage trend test and between 1.01 and 1.27 for the χ^2 test.

Figure 2 compares the individual $-\log_{10} p$ values from different methods. Most methods gave highly concordant results, particularly EMMAX (BN) and Gemma, whereas the 2 GenABEL methods were similar but less concordant. This is analogous to findings on single-observation data by Zhou and Stephens [9]. FaST-LMM tended to perform slightly differently from the other methods at SNPs with lower significance, although the results overall were still quite similar.

Figure 3 shows a selection of Manhattan plots. For each phenotype, the results from all methods were quite similar, although the longitudinal data tended to show stronger signals. No clearly significant SNP was found in any phenotype, which is not surprising given the relatively small size of the GAW18 data set, which is underpowered for detecting (at genome-wide levels of significance) anything other than strong genetic effects. The high concordance in significance levels (at any given SNP) achieved by the different software packages (see Figure 2) indicates that no package is substantially more powerful than another, as expected from the fact that all packages implement slightly different versions of essentially the same statistical model.

Although the results from all packages considered here were similar, the implementations did vary in speed. All packages performed the analysis in reasonable time (less than 1 day) on our system. Precise timings will depend on the computer resources and architecture available, but as a rule of thumb we found FaST-LMM and GRAMMAR-Gamma to be the fastest (taking just a few hours), followed by EMMAX and Gemma, which took 12 to 16 hours, and GenABEL/FASTA, which took 18 to 20 hours.

Conclusions

All methods performed well and results were similar, particularly at the most significant SNPs. We conclude that (at least for nonlongitudinal traits) it makes little difference to the results which method/software package is used, and the user can make the choice of package on the basis of personal taste, speed, or computational convenience. For longitudinal traits (modeled without regard to their longitudinal nature) the slight differences seen between the methods would be an interesting topic for further investigation, but it is beyond the scope of the current article.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
JE conducted the statistical analyses and drafted the manuscript. RAJH prepared the data and conducted statistical analyses. HJC conceived the overall study and critically revised the manuscript. All authors read and approved the final manuscript.

Acknowledgements
This work was supported by the Wellcome Trust (grant reference 087436). JE receives scholarship and funding from Faculty of Medicine, Ramathibodi Hospital, Mahidol University, Bangkok, Thailand. The GAW18 whole genome sequence data were provided by the T2D-GENES Consortium, which is supported by NIH grants U01 DK085524, U01 DK085584, U01 DK085501, U01 DK085526, and U01 DK085545. The other genetic and phenotypic data for GAW18 were provided by the San Antonio Family Heart Study and San Antonio Family Diabetes/Gallbladder Study, which are supported by NIH grants P01 HL045222, R01 DK047482, and R01 DK053889. The Genetic Analysis Workshop is supported by NIH grant R01 GM031575.
This article has been published as part of *BMC Proceedings* Volume 8 Supplement 1, 2014: Genetic Analysis Workshop 18. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcproc/supplements/8/S1. Publication charges for this supplement were funded by the Texas Biomedical Research Institute.

Authors' details
[1]Institute of Genetic Medicine, Newcastle University, Central Parkway, Newcastle upon Tyne, NE1 3BZ, UK. [2]Division of Medical Genetics, Department of Internal Medicine, Faculty of Medicine Ramathibodi Hospital, Mahidol University, Rama VI Rd, Ratchathevi, Bangkok 10400, Thailand.

Published: 17 June 2014

References

1. Almasy L, Dyer T, Peralta J, Jun G, Fuchsberger C, Almeida M, Kent JW Jr, Fowler S, Duggirala R, Blangero J: **Data for Genetic Analysis Workshop 18: human whole genome sequence, blood pressure, and simulated phenotypes in extended pedigrees.** *BMC Proc* 2014, **8(suppl 2)**:S2.
2. Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT: **Data quality control in genetic case-control association studies.** *Nat Protoc* 2010, **5**:1564-1573.
3. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, Sabatti C, Eskin E: **Variance component model to account for sample structure in genome-wide association studies.** *Nat Genet* 2010, **42**:348-354.
4. Balding DJ, Nichols RA: **A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity.** *Genetica* 1995, **96**:3-12.
5. Listgarten J, Lippert C, Kadie CM, Davidson RI, Eskin E, Heckerman D: **Improved linear mixed models for genome-wide association studies.** *Nat Methods* 2012, **9**:525-526.
6. Aulchenko YS, Ripke S, Isaacs A, van Duijn CM: **GenABEL: An R library for genome-wide association analysis.** *Bioinformatics* 2007, **23**:1294-1296.
7. Chen WM, Abecasis GR: **Family-based association tests for genomewide association scans.** *Am J Hum Genet* 2007, **81**:913-926.
8. Svishcheva GR, Axenovich TI, Belonogova NM, van Duijn CM, Aulchenko YS: **Rapid variance components-based method for whole-genome association analysis.** *Nat Genet* 2012, **44**:1166-1170.
9. Zhou X, Stephens M: **Genome-wide efficient mixed-model analysis for association studies.** *Nat Genet* 2012, **44**:821-824.
10. Devlin B, Roeder K: **Genomic control for association studies.** *Biometrics* 1999, **55**:997-1004.
11. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, *et al*: **PLINK: a tool set for whole-genome association and population-based linkage analyses.** *Am J Hum Genet* 2007, **81**:559-575.

doi:10.1186/1753-6561-8-S1-S79
Cite this article as: Eu-ahsunthornwattana *et al.*: Accounting for relatedness in family-based association studies: application to Genetic Analysis Workshop 18 data. *BMC Proceedings* 2014 **8**(Suppl 1):S79.

OPEN ACCESS Freely available online

PLOS | GENETICS

Comparison of Methods to Account for Relatedness in Genome-Wide Association Studies with Family-Based Data

Jakris Eu-ahsunthornwattana[1,2], E. Nancy Miller[3†], Michaela Fakiola[3], Wellcome Trust Case Control Consortium 2[¶], Selma M. B. Jeronimo[4], Jenefer M. Blackwell[3,5], Heather J. Cordell[1]*

1 Institute of Genetic Medicine, Newcastle University, International Centre for Life, Newcastle upon Tyne, United Kingdom, **2** Division of Medical Genetics, Department of Internal Medicine, Faculty of Medicine Ramathibodi Hospital, Mahidol University, Ratchathevi, Bangkok, Thailand, **3** Cambridge Institute for Medical Research, University of Cambridge School of Clinical Medicine, Addenbrooke's Hospital, Cambridge, United Kingdom, **4** Department of Biochemistry, Center for Biosciences, Universidade Federal do Rio Grande do Norte, Natal, Brazil, **5** Telethon Institute for Child Health Research, Centre for Child Health Research, The University of Western Australia, Subiaco, Western Australia, Australia

Abstract

Approaches based on linear mixed models (LMMs) have recently gained popularity for modelling population substructure and relatedness in genome-wide association studies. In the last few years, a bewildering variety of different LMM methods/software packages have been developed, but it is not always clear how (or indeed whether) any newly-proposed method differs from previously-proposed implementations. Here we compare the performance of several LMM approaches (and software implementations, including EMMAX, GenABEL, FaST-LMM, Mendel, GEMMA and MMM) via their application to a genome-wide association study of visceral leishmaniasis in 348 Brazilian families comprising 3626 individuals (1972 genotyped). The implementations differ in precise details of methodology implemented and through various user-chosen options such as the method and number of SNPs used to estimate the kinship (relatedness) matrix. We investigate sensitivity to these choices and the success (or otherwise) of the approaches in controlling the overall genome-wide error-rate for both real and simulated phenotypes. We compare the LMM results to those obtained using traditional family-based association tests (based on transmission of alleles within pedigrees) and to alternative approaches implemented in the software packages MQLS, ROADTRIPS and MASTOR. We find strong concordance between the results from different LMM approaches, and all are successful in controlling the genome-wide error rate (except for some approaches when applied naively to longitudinal data with many repeated measures). We also find high correlation between LMMs and alternative approaches (apart from transmission-based approaches when applied to SNPs with small or non-existent effects). We conclude that LMM approaches perform well in comparison to competing approaches. Given their strong concordance, in most applications, the choice of precise LMM implementation cannot be based on power/type I error considerations but must instead be based on considerations such as speed and ease-of-use.

Citation: Eu-ahsunthornwattana J, Miller EN, Fakiola M, Wellcome Trust Case Control Consortium 2, Jeronimo SMB, et al. (2014) Comparison of Methods to Account for Relatedness in Genome-Wide Association Studies with Family-Based Data. PLoS Genet 10(7): e1004445. doi:10.1371/journal.pgen.1004445

Editor: Gonçalo R. Abecasis, University of Michigan, United States of America

Received September 20, 2013; **Accepted** May 2, 2014; **Published** July 17, 2014

Copyright: © 2014 Eu-ahsunthornwattana et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by the Wellcome Trust (Grant Reference 087436). This study makes use of data generated by the Wellcome Trust funded WTCCC2 project (Grant Reference 085475). JEa receives scholarship and funding from Faculty of Medicine Ramathibodi Hospital, Mahidol University, Bangkok, Thailand. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: heather.cordell@newcastle.ac.uk

¶ Membership of the Wellcome Trust Case Control Consortium 2 is listed in Text S1.

† Deceased.

Introduction

Recently, linear mixed models based approaches have been proposed as appealing alternatives to principal component based approaches when adjusting for population substructure in genome-wide association studies of apparently unrelated individuals [1–4]. These methods build upon work originally described in the animal breeding literature, and subsequently developed in the human genetics literature, in which a genetic effect of interest (e.g. the number of copies of a particular allele at a particular test SNP) is included as a fixed effect in a regression model, with an additional random effect also included to model genetic correlation between individuals. The covariance structure for the random effect is generally assumed to correspond to that implied by a polygenic model, incorporating the genetic relationship (kinship) between each pair of individuals. Although use of this linear mixed model (LMM) was originally proposed for pedigrees with known relationships [5–10], this approach has recently gained popularity for use with samples of unknown or uncertain relationship [1–3,11–13], including apparently unrelated samples who may nevertheless display distant levels of common ancestry. For this purpose, the kinship coefficients between all pairs of individuals modelling either close or distant relatedness are estimated (prior to fitting the linear mixed model) on the basis of genome-wide genotype data, rather than being fixed at their known theoretical values.

Fitting a full linear mixed model for each SNP in turn across the genome is computationally challenging. These computational considerations have led to the development of several faster

Author Summary

Recently, statistical approaches known as linear mixed models (LMMs) have become popular for analysing data from genome-wide association studies. In the last few years, a bewildering variety of different LMM methods/software packages have been developed, but it has not always been clear how (or indeed whether) any newly-proposed method differs from previously-proposed implementations. Here we compare the performance of several different LMM approaches (and software implementations) via their application to a genome-wide association study of visceral leishmaniasis in 348 Brazilian families comprising 3626 individuals. We also compare the LMM results to those obtained using alternative analysis methods. Overall, we find strong concordance between the results from the different LMM approaches and high correlation between the results from LMMs and most alternative approaches. We conclude that LMM approaches perform well in comparison to competing approaches and, in most applications, the precise LMM implementation will not be too important, and can be chosen on the basis of speed or convenience.

approximations for constructing tests of the fixed SNP effects of interest in the linear mixed model [1,2,9,10,14]. These approximate tests have been implemented in various software packages including MERLIN, GenABEL, EMMAX, TASSEL, FaST-LMM, Mendel and MMM. The MMM [15] and FaST-LMM [4] packages, in common with the package GEMMA [16], also provide fast implementations of an exact (rather than an approximate) model, which in principle can lead to a small increase in power [15,16], depending on the true underlying level of relatedness.

A limited comparison of several LMM implementations, via application to real and simulated data from Genetic Analysis Workshop 18 (GAW18) [17], was performed by Eu-ahsunthornwattana et al. [18]. In the GAW18 data, which comprised 959 Mexican-American individuals from 20 families, the LMM implementations investigated performed rather similarly to one another in terms of the association test statistics and p-values achieved; however, no formal quantification of power or type 1 error was performed. Eu-ahsunthornwattana et al. [18] also investigated the performance of the various LMM implementations when applied naively to longitudinal traits (repeated measures) available in GAW18, simply by treating each measurement as if it came from a separate person and expanding out the genetic data set accordingly (resulting in an expanded data set containing many apparent twins, triplets, quadruplets etc., depending on how many measurements are available for each person). Although this approach is not strictly 'correct' (as it does not distinguish between correlations in trait values due to genetic factors and correlations due to non-genetic within-individual factors), Eu-ahsunthornwattana et al. found this procedure generated only minimal inflation in the resulting distribution of genome-wide test statistics.

Here we expand the investigation of Eu-ahsunthornwattana et al. [18] to perform a more comprehensive comparison of LMM approaches (involving a larger number of software implementations) and to conduct a formal investigation of power and type 1 error. We also compare the LMM approaches to traditional family-based approaches ('within-family association tests' based on the transmission of high-risk alleles within pedigrees [19–23]), and to alternative previously-proposed approaches based on extending standard case/control tests (such as the Armitage trend test) to allow for either known [24,25] or known and unknown [26] relatedness. The programs compared (see Table 1) differ in the precise details of the methodology implemented (such as whether an LMM approach is used, and, if so, whether an exact method or an approximation is used) and through various user-chosen options such as the specific method and number of SNPs used to estimate the kinship matrix. We investigate the sensitivity to these choices and the success (or otherwise) of the approaches in controlling the overall genome-wide error-rate in both real and simulated data (into which artificial simulated disease loci have been inserted).

The approaches are compared via application to real and simulated data derived from a genome-wide association study of visceral leishmaniasis (VL) in 348 Brazilian families comprising 3636 individuals (1970 with both genotype and phenotype data). This Brazilian family data set was used (together with a larger Indian case/control data set) by Fakiola et al. [13] to identify, at genome-wide levels of significance, a replicable association between variants in the HLA region on chromosome 6 and visceral leishmaniasis. Although in [13] the HLA locus (analysed using the LMM package MMM [15]) did not achieve genome-wide levels of significance in the Brazilian data set alone (p-value $=2\times10^{5}$), this locus was the only one to show strong evidence of association in both Brazilian and Indian data sets, and achieved convincing replication in a separate Indian cohort.

Results

Estimation of kinship coefficients using genome-wide SNP data

Before embarking on a detailed comparison of different methods, we explored the use of different SNP sets (containing different numbers of SNPs) for estimating pairwise kinship measures, in order to identify a robust set of SNPs that could be used for subsequent comparisons. We considered using either the full genome-wide set of SNPs (545,433 SNPs), a 'pruned' set of 50,129 SNPs selected to have minor allele frequencies >0.4 and chosen to be in approximate linkage equilibrium via the `--indep 50 5 2` command in PLINK [27]), or a 'thinned' set of 1900 evenly-spaced SNPs that were selected from the 'pruned' SNPs based purely on physical position using the software package MapThin (http://www.staff.ncl.ac.uk/richard.howey/mapthin/). In addition to exploring the kinship estimates provided by various LMM software packages, we also investigated those provided by the software packages PLINK [27] and KING [28]. KING implements two different kinship estimation methods: KING-homo (KING_H), which assumes population homogeneity, and KING-robust (KING_R), which provides robust relationship inference in the presence of population substructure.

A comparison of the kinship estimates output by different software packages based on the pruned set of SNPs is shown in Figure 1 (similar results were seen for the full and thinned SNP sets, data not shown). Although the scale on which the kinship estimates are measured differs between different packages, the measures themselves are highly correlated, particularly those from EMMAX-BN, FaST-LMM, GenABEL, GEMMA and MMM. Kinship measures from EMMAX-IBS and PLINK were also quite well correlated, although they tended to differ slightly from those in the previous group. Kinship measures are used within the LMM framework to structure the variance/covariance matrix of the genetic random effect (see Methods). Thus, the scale of measurement (i.e. whether the kinship measure actually reflects an estimate of the kinship per se, or a rescaled measure such as twice the

Table 1. Summary of methods/software packages investigated.

Package/method and version	Approach	Kinship estimation method	Reference(s)
EMMAX emmax-intel-20120210.tar.gz	LMM (approximate)	Kinship matrix estimated internally using user-supplied set of SNPs, or set to theoretical/estimated values calculated externally	[1]
FaST-LMM v2.04	LMM (approximate or exact)	Kinship matrix estimated internally using user- supplied set of SNPs, using SNPs selected through FaST-LMM-Select procedure, or set to theoretical/estimated values calculated externally	[4] [30] [31]
GEMMA v0.91	LMM (exact)	Kinship matrix estimated internally using user-supplied set of SNPs, or set to theoretical/estimated values calculated externally	[16]
GenABEL v1.7-6 (FASTA)	LMM (approximate)	Kinship matrix estimated internally using user-supplied set of SNPs, or set to theoretical/estimated values calculated externally	[9] [39]
GenABEL v1.7-6 (Grammar-Gamma)	LMM (approximate)	Kinship matrix estimated internally using user-supplied set of SNPs, or set to theoretical/estimated values calculated externally	[14] [39]
GTAM (implemented in MASTOR v0.3)	LMM (approximate)	Kinship matrix calculated externally (assumed to reflect 'known' (theoretical) pedigree relationships)	[8]
Mendel v13.2	LMM (approximate or exact)	Kinship matrix estimated internally using theoretical pedigree relationships, estimated within estimated pedigree clusters (using all SNPs), or fully estimated (using all SNPs)	[35]
MMM v1.01	LMM (approximate or exact)	Kinship matrix estimated internally using user-supplied set of SNPs, or set to theoretical/estimated values calculated externally	[15]
FBAT v2.0.4	Transmission of alleles within pedigrees	Method by definition uses 'known' (theoretical) pedigree relationships	[21] [23]
MASTOR v0.3	Retrospective quantitative trait version of MQLS	Kinship matrix calculated externally (assumed to reflect 'known' (theoretical) pedigree relationships)	[25]
MQLS v1.5	Adjusted version of retrospective case/control test	Kinship matrix calculated externally (assumed to reflect 'known' (theoretical) pedigree relationships)	[24]
ROADTRIPS v1.2 (RM test)	Adjusted version of retrospective case/control test	Kinship matrix calculated externally (assumed to reflect 'known' (theoretical) pedigree relationships). Further correction based on genome-wide set of SNPs applied internally.	[26]

doi:10.1371/journal.pgen.1004445.t001

kinship) should not be too important, as any rescaling will be compensated for by a similar rescaling of the estimated genetic variance parameter σ_g^2 (see Methods). Kinship estimates from both KING methods tended to differ most from the other methods, with the frequent output of negative kinship estimates (compared to most other methods for which the kinship estimates are bounded at 0) among the less related individuals. This was more pronounced for KING_R than for KING_H. We consider later the possible implications of these (rather small) differences in estimated kinships for subsequent association testing.

Within any given method, we found the kinship measures (for each pair of individuals) and p-values obtained (in the real data set) based on the full SNP set to be very similar to those based on the pruned set, whereas those calculated based on the thinned set were less similar (see Figure S1). The performance of the different SNP sets in terms of controlling the genome-wide type 1 error rate (i.e. controlling the genomic inflation factor λ [29] to the desired level of $\lambda = 1$) in the real data set is shown in Figure 2 (see Figure S2 for full QQ plots). All packages performed well when using the full or pruned set of SNPs ($\lambda = 0.99$–1.00), but performance deteriorated when the thinned set was used (λ mostly about 1.08–1.10). This was most pronounced for GenABEL (GRAMMAR-Gamma), for which λ was 1.16. Our intuition is that, although 1900 SNPs may be sufficient to accurately model close relationships (such as full sib or parent-offspring), many more SNPs will be required to accurately model distant relationships within pedigrees (such as cousins, second cousins, third cousins etc.) or even more distant relationships between pedigrees. Results obtained using theoretical kinships were inflated for all methods ($\lambda \approx 1.11$), suggesting the presence of additional relatedness/population structure that is not well accounted for by known family relationships. Regardless of the method or SNP set used, adjustment always resulted in substantially lower inflation than was seen ($\lambda = 1.23$) in unadjusted analysis.

Listgarten et al. [30] proposed an automated method, FaST-LMM-Select, to select the most appropriate set of SNPs to use for kinship estimation when testing for association in a LMM framework. The method proceeds by ordering SNPs according to their linear regression p-values and then constructing kinship matrices with an increasing number of ordered SNPs, until the first minimum genomic control factor λ is obtained. We investigated this strategy within the FaST-LMM package using either the full or pruned set of SNPs as a starting point (see Figure S3). We found that the first minimum genomic control factor (achieved using 3–10 ordered SNPs) was generally higher than the desired value of $\lambda = 1$, the genomic control factor subsequently decreased to considerably less than 1, and then increased back to 1 once all (pruned or full) SNPs had been included.

The automated version of FaST-LMM-Select available as an option within the current version of the FaST-LMM package uses a slightly different strategy involving k-fold cross-validation [31], with the ordering of SNPs and calculation of genomic control factors as varying numbers of SNPs are included in the kinship calculation carried out within the training data (and then used to predict the test data) within each cross-validation fold. The final number of SNPs to be used in the kinship calculation for the entire data set is that which minimizes the mean-squared error summed over all folds. (See FaST-LMM documentation and [31] for more

details). Lippert et al. [31] found this procedure to show some advantage over using all SNPs (including a large number of presumably irrelevant SNPs) in simulations that included population stratification (but not familial relatedness) of quantitative phenotypes in randomly ascertained individuals. Application of this automated procedure to the real disease phenotype in our highly ascertained set of Brazilian pedigrees resulted in no SNPs selected for calculation of kinships when applied to the full SNP set, or two SNPs selected when applied to the pruned SNP set, resulting in a genomic control value of $\lambda = 1.17$ when these two SNPs were used to adjust for relatedness in the subsequent association analysis. We conclude that, at least for our data set, there is no particular advantage in using the FaST-LMM-Select procedure, indeed this procedure seems to work less well than simply using all pruned or full SNPs for estimating pairwise kinships. For the remainder of the manuscript we therefore focus on results obtained using the pruned set of SNPs to estimate kinships (apart for genome-wide analysis in the program Mendel, which by default always uses the entire set of SNPs that has been read in).

Comparison of LMM and alternative analysis approaches

We compared the performance of the different LMM and alternative approaches listed in Table 1 through their application to real and simulated data derived from the Brazilian family data set of Fakiola et al. [13]. The simulation scenarios (see Methods) included a binary disease trait influenced by either two strong (sim-D1) or two weak (sim-D2) genetic effects or a quantitative trait (sim-Q) influenced by two strong genetic effects. In all cases the genetic effects were governed by two SNPs (rs9271252 and rs233722) located on chromosomes 6 and 12 respectively. In addition to the effects at rs9271252 and rs233722, we also allowed for 22 weaker 'polygenic' effects caused by genotype at the 100th SNP on each autosomal chromosome. Where applicable, we used either the default analysis options within each program, or else explored the use of different options as indicated below. The program FaST-LMM uses either maximum likelihood (ML) or restricted maximum likelihood (REML). (In early versions of FaST-LMM the default was ML but in later versions the default became REML). After some experimentation, we deemed the ML option to be the most reliable in the presence of strong genetic effects, and have therefore used ML for all results presented here.

The success of the various approaches in controlling the overall genome-wide type 1 error rate (i.e. controlling the genomic inflation factor [29] λ to the desired level of $\lambda = 1$) is shown in Table 2. All methods that made use of estimated kinships performed well, apart from Mendel when estimation was restricted only to estimated pedigree clusters (which gave $\lambda = 1.10$) and MQLS, for which use of estimated kinships (in the 1972 genotyped individuals) appeared to result in slightly deflated genomic inflation factors. For all other methods, use of estimated kinships reduced the genomic inflation factor to around 1, compared to a value of $\lambda = 1.23$ in the real data (and up to 1.43 in the simulated data) when performing an unadjusted analysis. Methods that used only theoretical kinships based on 'known' pedigree information performed well in the simulated data sets, but were less successful at controlling inflation for the real data set, suggesting that our real data contains additional, more complicated, relatedness or population substructure that is not accounted for by known family relationships.

The Brazilian populations studied here are believed to be long-term (>200 years) admixtures of Caucasian, Negroid and Native Indian ethnic backgrounds, as confirmed in recent analysis of a subset of our families [32]. The discrepancy between the genomic inflation factors seen in our real and simulated data results suggests that our (relatively simplistic) simulation scenarios have not been able to fully mimic the underlying population structure existant in the real data; although our simulation strategy (see Methods) was designed to generate trait correlations that reflect close familial relationships, we did not specifically endeavour to generate correlations due to population stratification or more distant/cryptic relationships. To investigate the relative contributions of phenomena such as admixture/population stratification/cryptic relationships to the inflation observed in our real data when using theoretical (pedigree-based) kinships, we applied the ADMIXTURE program [33] to our pruned set of SNPs to estimate ancestry proportions (assuming 3 ancestral populations) in each individual. Although the variation in ancestry proportion estimated within each individual was quite large (standard deviation $\approx 0.08-0.15$ depending on ancestral population) there was no evidence ($P > 0.14$) for a relationship between estimated ancestry proportion and disease status, suggesting that the inflation in test statistics observed when using theoretical kinships is more likely to be due to unmeasured cryptic relationships and/or subtle population substructure, than to population substructure or admixture directly related to the Caucasian, Negroid and Native Indian ethnicities. This conclusion was supported by the fact that logistic regression analysis allowing for the ancestry proportions as covariates resulted in a genomic control inflation factor of 1.17, only slightly reduced from the unadjusted genomic control inflation factor of 1.23.

We also used as covariates in a logistic regression analysis the first nine coordinates obtained from a multidimensional scaling (MDS) analysis of the pruned SNPs in PLINK (having considered between one and ten coordinates, nine was the number that minimised the genomic control inflation factor). The resulting genomic control inflation factor was 1.08, considerably smaller than the unadjusted inflation factor of 1.23, but still not perfectly controlled. Inclusion of MDS coordinates as covariates, similar to including principal components scores, might be expected to account for more subtle levels of population substructure than are accounted for by the use of the ADMIXTURE program (and may possibly also indirectly account for relatedness), which perhaps explains the greater success of this procedure. However the fact that LMM approaches based on estimated kinships still do better (with respect to controlling λ) than does the MDS approach suggests there may still be levels of known or cryptic relatedness that are not well-captured by these first nine coordinates.

An intuitive overview of the expected power provided by the different (real and simulated) data sets can be obtained from Figure S4, which shows Manhattan plots from a FaST-LMM analysis of a single replicate of real or simulated data. The real phenotype data shows a noticeable signal in the HLA region on chromosome 6, consistent with the main finding in [13], while for all simulated traits the primary associated regions are correctly identified without any obvious false signals. A formal comparison of power and type 1 error for the different analysis methods using 1000 simulation replicates is shown in Figure 3. All methods apart from an unadjusted analysis show acceptable levels of type 1 error (although note that the type 1 error rate for FBAT appears to be slightly conservative). In terms of power, all LMM approaches (including GTAM and Mendel) and MASTOR show similar performance, apart from MMM which shows slightly higher power than other methods for detection of loci involved in the (strong) simulated quantitative trait. ROADTRIPS and MQLS show slightly lower power than the LMM approaches, while the approaches implemented in FBAT appear to be considerably less powerful than those implemented in the LMM and other packages

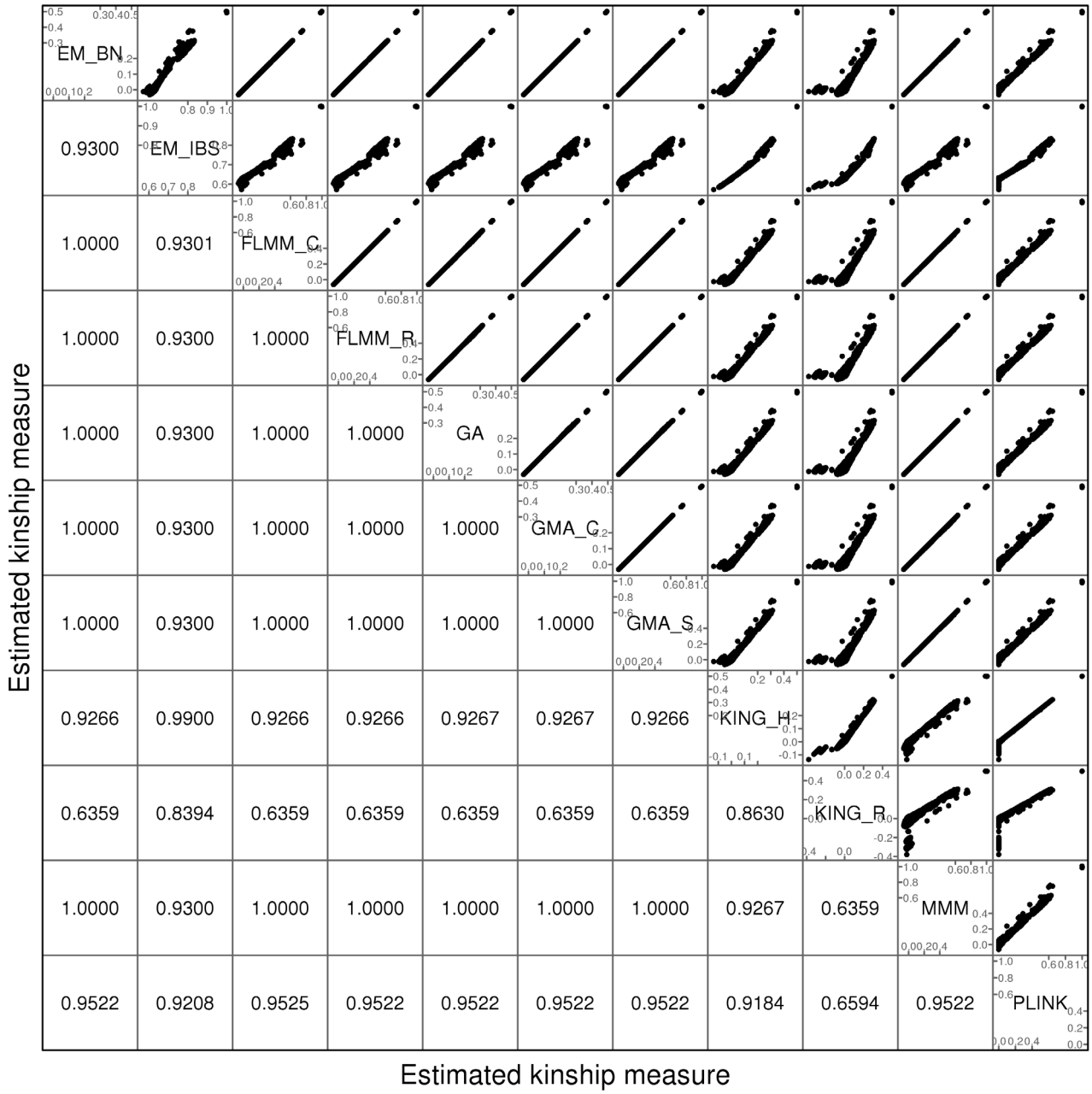

Figure 1. Comparison of kinship estimates (pruned SNPs) using different software packages. Plots above the diagonal show a comparison of kinship measures, with correlations between the kinship measures indicated below the diagonal. EM_BN = EMMAX (Balding-Nichols), EM_IBS = EMMAX (IBS method), FLMM_C = FaST-LMM using covariance matrix, FLMM_R = FaST-LMM using realised relationship matrix, GA = GenABEL, GMA_C = GEMMA using centred genotypes, GMA_S = GEMMA using standardised genotypes, KING_H = KING with homogeneous population assumption, KING_R = KING with robust estimation.
doi:10.1371/journal.pgen.1004445.g001

(even allowing for FBAT's slightly conservative levels of type 1 error). The lower power of FBAT is likely to be caused by the smaller effective sample size (357 cases compared to 357 'pseudo' controls in FBAT, versus 357 cases compared to 1613 genuine controls in the LMM and other alternative approaches), due to the way the FBAT test statistics are constructed. These results are consistent with a visual examination of the Manhattan plots obtained from the different methods using either the real data or a single replicate of the simulated data (Figure 4, Supplementary Figures S5–S6), with FBAT achieving much lower levels of significance around the true or simulated phenotype-associated SNPs than do the other methods. (The results from all LMM methods not displayed in Figure 4 and Supplementary Figures S5–S6 were indistinguishable from FLMM_E, data not shown).

Although the LMM (and several alternative) approaches show similar overall levels of power, an interesting separate question is the degree of concordance between the different methods with respect to the association signals detected. In the real data set we found the p-values obtained at each SNP from the different LMM methods to be highly concordant (Figure S7), while the concordance between the LMM methods and alternative approaches (Figure S8) is high for all methods other than FBAT

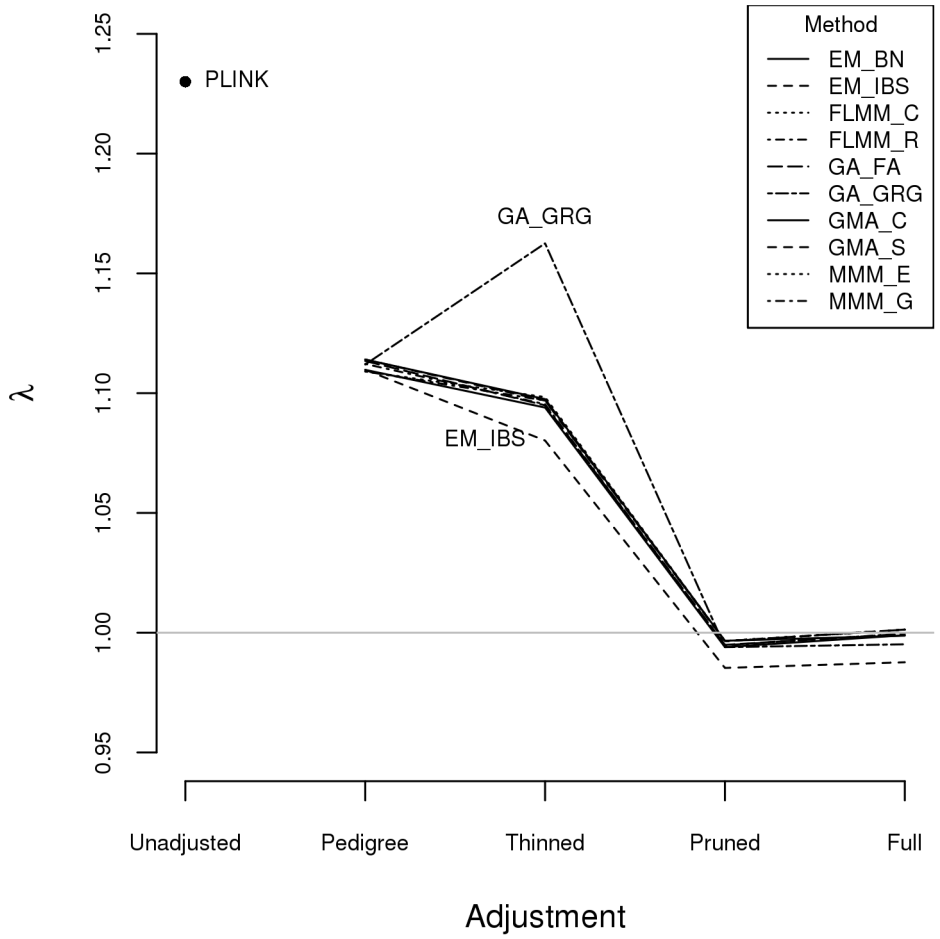

Figure 2. Genomic control factors obtained using different software packages and different strategies for modelling kinships. PLINK = analysis in PLINK with no adjustment made for relatedness. Other methods/software packages are listed in Table 1 (see Table 2 for abbreviated names of methods). Pedigree = theoretical kinships based on known pedigree relationships used to adjust for relatedness. Thinned = kinships based on 1900 'thinned' SNPs used to adjust for relatedness. Pruned = kinships based on 50,129 'pruned' SNPs used to adjust for relatedness. Full = kinships based on 545,433 SNPs used to adjust for relatedness.
doi:10.1371/journal.pgen.1004445.g002

(although lower than is observed among methods within the LMM class). The test implemented in FBAT is statistically uncorrelated with that implemented in the LMM and other alternative approaches, therefore it is not surprising that little concordance is seen between the test statistics achieved at the vast majority of (presumably null) SNPs. Figure S8 also shows that methods that use phenotype information from non-genotyped family members (MQLS3626 and RT3626, which use all 3626 individuals regardless of whether or not they have genotype data) are most similar to each other and less similar to methods that use information only from the genotyped individuals.

The high concordance between the different LMM methods (and, to a slightly lesser extent, between LMM methods and all methods other than FBAT) is also seen for the simulated (weak disease) trait (Figure S9); similar results were found for the other simulated traits and other LMM methods (data not shown). A formal comparison of the concordance between 'top hits' identified by the different methods in the simulated data (1000 simulation replicates, comparison restricted to true and null simulated regions) is shown in Table 3. Using EM_BN as reference, the concordance between the top SNPs identified is seen to be extremely high for all other methods except FBAT, suggesting again that all methods except FBAT provide essentially the same inference.

Feeding externally estimated kinship coefficients into LMMs

Most LMM packages (although not Mendel) allow a separation between the 'estimation of kinships' step and the 'association testing' step. This is convenient as it allows the user to read in theoretical or estimated kinships as desired, and to consider using an alternative package for estimating kinships to the one used for the actual association testing. We investigated performing an analysis in FaST-LMM (exact calculation), but with the kinships estimated from various different software packages (see Figure S10 and Table S1). Use of the 'wrong' kinship estimates (chosen to be inversely related to the theoretical kinship value) resulted in very similar results to unadjusted analyses ($\lambda = 1.23$ in the real trait, 1.12 in the simulated strong disease trait, and 1.43 in the simulated quantitative trait). Results based on kinship estimates from KING_R and KING_H were very similar to those obtained using FaST-LMM's own realised relationship matrix (FLMM-R) for all traits, and provided good control of the genome-wide error rate ($\lambda \approx 1$) in spite of the unusual pattern in KING's estimated kinships that had been noted in Figure 1. Estimation of kinships using PLINK was less satisfactory, leading to inflated genomic control factors in both real and simulated data sets. This is consistent with previous results [28] suggesting that PLINK

performs less well than KING for relationship estimation. Interestingly, although KING_R has been shown to have an advantage over KING_H in non-homogeneous populations when the goal is relationship estimation for its own sake [28], this advantage is not apparent here, where the goal is instead to adjust for potentially different levels of relatedness, from close family relationships to more distant relationships (perhaps mimicking population membership), while performing association testing.

Computational efficiency and ease-of-use

Given that many of the software implementations we investigated (and in particular all the various LMM implementations) showed similar levels of power and type 1 error, and gave rather similar inference in terms of localisation of signals and $-\log_{10}$ p-values achieved, an important practical consideration when deciding what implementation to use is the ease-of-use and computational efficiency. Ease-of-use is necessarily somewhat subjective as it depends on a user's prior experience and software/operating system preferences. Computational efficiency can, in theory, be examined more objectively, however, in practice, the total time required to perform an analysis is dependent on the computer architecture available (in particular the ability of the system and of any given program to allow multi-threading), demands of competing users and the availability of (and ability of any given program to make use of) facilities for parallel processing e.g. a multi-node compute cluster. These considerations make it hard to perform a genuine 'head-to-head' comparison between different packages. In Table S2 we present an approximate comparison (carried out on the same machine, without use of parallel processing) together with some comments concerning ease-of-use. Since many groups (including ourselves) use PLINK [27] to perform initial quality control of genome-wide association data, we considered programs that could use PLINK files directly (or with just a few easily-implemented transformation steps) to be the easiest to use, while those programs that required more extensive data transformation, creation of additional input files and/or external estimation of kinships were considered harder.

With respect to computational speed, as a rule of thumb we found Mendel (theoretical kinships), FaST-LMM (approximate) and GenABEL (GRAMMAR-Gamma) to be the fastest LMM implementations, taking between 3 minutes and a quarter of an hour on our system to analyse 545,433 SNPs in 1972 genotyped individuals. These were closely followed by EMMAX and MMM (approximate) which took around half an hour, GenABEL (FASTA), GEMMA, FaST-LMM (exact) and MMM (exact) which typically took 1–2 hours, Mendel (estimated kinships) which took around 2.5 hours, and GTAM which took around 4 hours. Of the non-LMM methods, FBAT, MQLS and MASTOR were the fastest, taking a few hours to perform the analysis, while ROADTRIPS was the slowest, taking several days. Inputting estimated (rather than theoretical) kinships into MQLS increased the time taken to around 4 days (and appeared to over-correct the genomic inflation, see Table 2), while an analysis inputting estimated (rather than theoretical) kinships into ROADTRIPS was still running (with analysis completed for only 38,926 of the desired 545,433 SNPs) after more than 2 months. Neither MQLS nor ROADTRIPS were designed for analysis of unrelated individuals and so are most likely optimised for reading in and working with relatively sparse kinship matrices (in which individuals from different pedigrees are assumed to have kinships equal to 0); to force the programs to consider estimated kinships between all individuals we had to recode the pedigree names to pretend that everyone comes from the same pedigree, which most likely considerably increases processing and memory requirements.

Analysis of longitudinal phenotypes

Eu-ahsunthornwattana et al. [18] investigated a strategy for analysing longitudinal traits (repeated measures) in a linear mixed model framework simply by treating each measurement as if it came from a different individual, and expanding out the genetic data set accordingly (resulting in an expanded data set containing many apparent twins, triplets, quadruplets etc., depending on how many measurements are available for each person). We investigated this strategy in the current data set using a single replicate of data (498 individuals) simulated under either a longitudinal (sim-L20) or longitudinal polygenic (sim-P20) model (see Methods). Results (Table 4) showed that EMMAX, FaST-LMM and GEMMA were successful in maintaining the genomic inflation factor to about 1, whereas GenABEL (FASTA) and MMM showed some inflation, particularly in the polygenic longitudinal simulation, and GenABEL (GRAMMAR-Gamma) showed strong *deflation*. Comparison of the concordance in $-\log_{10}$ p-values achieved by the different methods (data not shown) indicated that, although the results from different methods were highly correlated (in terms of the top SNPs identified), the actual p-values achieved were very different, consistent with the differences seen in overall distribution of test statistics.

Analysing each repeated measure as if it comes from a different individual treats our data set as a larger 'pseudo data set' containing many apparent twins/triplets/quadruplets (actually, in this case, 20-tuplets). Although less satisfactory than a proper longitudinal analysis that takes into account correlations due to both relatedness between individuals and repeated measures within individuals [34], our intuition was that the LMM framework would absorb the effect of repeated measures within individuals into the genetic component of variance estimated, resulting in an overall correct distribution of test statistics. For EMMAX, FaST-LMM and GEMMA, this intuition appears to have been correct. Although for GenABEL (FASTA) and MMM the resulting distribution of test statistics is inflated, the linear relationship between the observed and desired test statistics means that test statistics following the desired distribution could be obtained simply by dividing the observed χ^2 test statistics by the observed genomic control inflation factor, in an approach akin to standard genomic control [29].

We also investigated a 'proper' longitudinal analysis implemented within the R software package longGWAS [34]. QQ plots from longGWAS (data not shown) indicated acceptable genomic control inflation factors ($\lambda = 1.00$ and 0.97 for sim-L20 and sim-P20 respectively). A comparison of longGWAS with our (improper) approach using FaST-LMM (data not shown) indicated that the results (in terms of the $-\log_{10}$ p-values obtained at each SNP) from longGWAS and FaST-LMM were highly correlated for both sim-L20 and sim-P20. Although the 'proper' analysis implemented in longGWAS might be considered theoretically most appealing, we note that longGWAS was considerably slower than FaST-LMM, taking approximately 19 hours (in comparison to 5.5 minutes for FaST-LMM), when run in parallel for each of 22 chromosomes. If run as a single process (all chromosomes), this translates to about 9.5 days for longGWAS versus 7.6 hours for FaST-LMM. Thus, given the satisfactory performance of FaST-LMM, and the high correlation between the results obtained from FaST-LMM and those from longGWAS, from a practical point of view, FaST-LMM (or possibly EMMAX or GEMMA) would seem the more attractive option.

Table 2. Genomic control inflation factors achieved in real data or in a single replicate of the simulated data sets.

			Trait analysed			
Method	**Description**	**Kinships used**	**Real disease (VL)**	**Simulated strong (sim-D1)**	**Simulated weak (sim-D2)**	**Simulated quantitative (sim-Q)**
Unadjusted	Standard linear or logistic regression	None	1.23	1.12	1.04	1.43
EM_BN	EMMAX (Balding-Nichols kinships)	Estimated	0.99	0.99	1.00	0.99
EM_IBS	EMMAX (IBS kinships)	Estimated	0.99	0.99	1.00	1.00
FLMM_A	FaST-LMM (approximate calculation)	Estimated	0.99	0.99	1.00	1.00
FLMM_E	FaST-LMM (exact calculation)	Estimated	1.00	0.99	1.01	1.00
GA_FA	GenABEL (FASTA)	Estimated	0.99	0.99	1.00	0.99
GA_GRG	GenABEL (GRAMMAR-Gamma)	Estimated	0.99	0.99	1.00	1.00
GMA_C	GEMMA using centred genotypes	Estimated	1.00	0.99	1.01	1.00
GMA_S	GEMMA using standardised genotypes	Estimated	1.00	0.99	1.01	1.00
GTAM	GTAM (implemented in MASTOR)	Pedigree	1.20	1.00	0.99	0.99
Mendel_T	Mendel with theoretical kinships	Pedigree	1.11	1.00	0.99	0.99
Mendel_P	Mendel with kinships estimated within estimated pedigree clusters	Estimated	1.10	1.00	0.99	0.99
Mendel	Mendel with fully estimated kinships	Estimated	1.03	0.99	1.00	1.00
MMM_E	MMM (exact calculation)	Estimated	1.00	0.99	1.01	1.00
MMM_G	MMM (GLS approximation)	Estimated	0.99	0.99	1.00	0.99
FBATaff[a]	FBAT (transmissions to affecteds only)	Pedigree	1.02	1.01	1.00	–
FBATboth	FBAT (transmissions to all individuals)	Pedigree	1.01	1.00	1.01	1.00
MASTOR	MASTOR (implemented in MASTOR)	Pedigree	1.15	1.00	0.99	0.99
MQLS1972[a]	MQLS (using 1972 genotyped individuals)	Pedigree	1.15	1.01	0.99	–
MQLS3626[a,b]	[a,b]MQLS (using all 3626 individuals with or without genotype data)	Pedigree	1.16	–	–	–
MQLS1972_E	MQLS using 1972 genotyped individuals and estimated kinships	Estimated	0.94	0.90	0.91	–
RT1972[a]	ROADTRIPS (using 1972 genotyped individuals)	Pedigree & estimated	1.00	1.00	0.99	–
RT3626[a,b]	ROADTRIPS (using all 3626 individuals with or without genotype data)	Pedigree & estimated	1.00	–	–	–

[a]FBATaff, MQLS and ROADTRIPS are only applicable to binary traits and so do not have results in the 'Simulated quantitative' column.
[b]In the simulated data sets, MQLS and RT could only be based on the 1972 individuals with simulated phenotypes, and so no simulated trait results are displayed in the MQLS3626 and RT3626 rows.
doi:10.1371/journal.pgen.1004445.t002

Another program that can, in theory, implement a 'proper' longitudinal analysis is the lmekin function within the R package coxme. We found this function to be computationally infeasible for analysis of genome-wide data, but application to a selected set of 2423 SNPs (of different effect sizes) in the sim-L20 data suggested that the results were very similar to those obtained from GenABEL (FASTA), EMMAX, FaST-LMM, GEMMA and MMM. However, we were unable to get lmekin to give meaningful results (most results were "NA") when applied to the sim-P20 data. We also speculated that a 'proper' longitudinal analysis should, in theory, be implementable in the package Mendel [35], through making use of Mendel's ability to include household effects. (Effectively one would trick Mendel into fitting the correct model by designating all 'individuals' (with each timepoint considered as a separate individual) to be members of a single pedigree, with the individuals corresponding to separate timepoints within a single real individual designated as belonging to the same household). We attempted to fit this model in Mendel for our sim-L20 and sim-P20 data sets, but were unable to obtain reliable

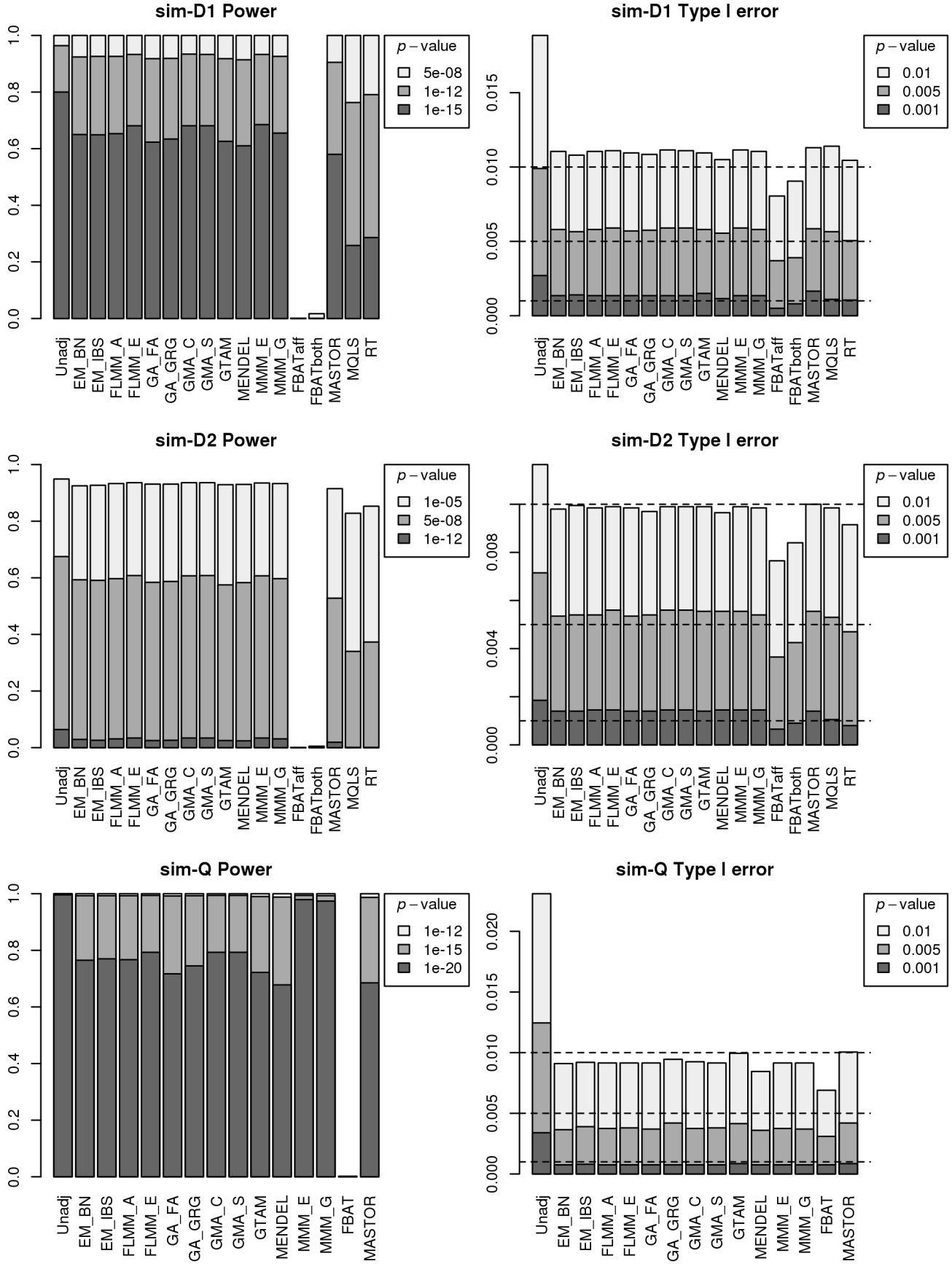
sim-D1 Power
p − value
5e-08
1e-12
1e-15
sim-D1 Type I error
p − value
0.01
0.005
0.001
sim-D2 Power
p − value
1e-05
5e-08
1e-12
sim-D2 Type I error
p − value
0.01
0.005
0.001
sim-Q Power
p − value
1e-12
1e-15
1e-20
sim-Q Type I error
p − value
0.01
0.005
0.001
Unadj
EM_BN
EM_IBS
FLMM_A
FLMM_E
GA_FA
GA_GRG
GMA_C
GMA_S
GTAM
MENDEL
MMM_E
MMM_G
FBATaff
FBATboth
MASTOR
MQLS
RT
FBAT

Figure 3. Power and type 1 error of different methods. Powers (left hand plots) are defined as the proportion of replicates (out of 1000) in which both simulated disease loci are detected, with 'detection' corresponding to any SNP within 40 kb of the simulated disease locus reaching the specified *p*-value threshold. Type 1 errors (right hand plots) are defined as the proportion of null SNPs (out of 20,000 = 20 null SNPs times 1000 simulation replicates) that reach the specified *p*-value threshold. Horizontal dashed lines indicate the target *p*-value thresholds (i.e. the expected type 1 error rates).
doi:10.1371/journal.pgen.1004445.g003

results. (If included, household effects were continually estimated at 0, and, regardless of whether or not household effects were included, the SNP association tests showed highly inflated significance values, with no correct localisation of true sim-L20 signals as had been seen for FaST-LMM (Figure S4) and little correlation between $-\log_{10}$ p-values from Mendel and those from these other packages). We speculate that the algorithm used by Mendel may be adversely affected by the presence of many highly-related individuals (e.g. repeated measures that in actuality pertain to a single individual), causing the test statistics generated to be unreliable.

Discussion

Here we have demonstrated, through simulations and application to real data, that linear mixed model approaches such as those implemented in the packages GenABEL, EMMAX, FAST-LMM,

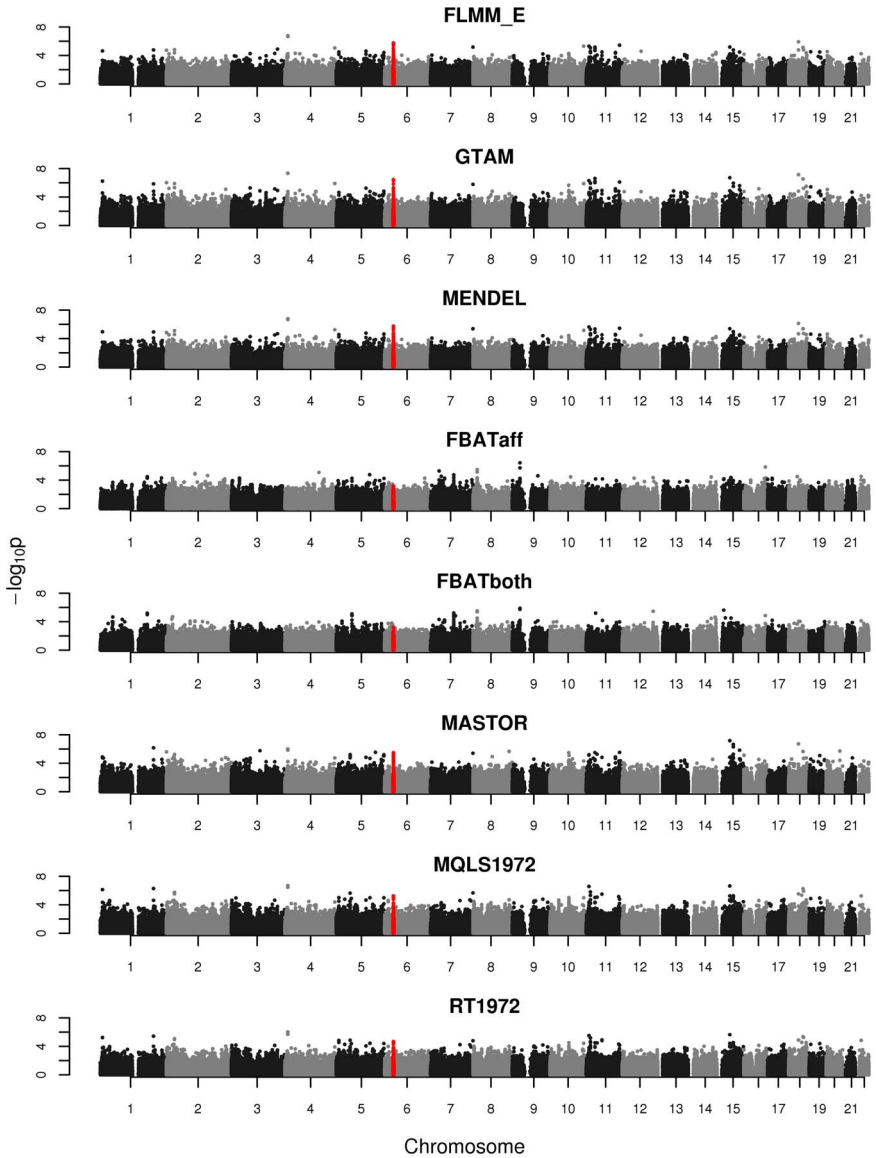

Figure 4. Manhattan plots for the real phenotype using FaST-LMM exact and alternative software packages. The points marked in red denote the confirmed significant region from Fakiola et al. (2013). FLMM_E = FaST-LMM using exact calculation, MQLS1972 = MQLS using 1972 genotyped individuals, RT1972 = ROADTRIPS using 1972 genotyped individuals, FBATaff = FBAT using transmissions to affecteds only, FBATboth = FBAT using transmissions to both affecteds and unaffecteds. Results from all other LMM methods were indistinguishable from FLMM_E and so are not shown.
doi:10.1371/journal.pgen.1004445.g004

Table 3. Concordance between top SNPs identified by different methods.

		Mean (standard deviation) in 1000 replicates of proportion of top *t* SNPs within null and true regions that overlap with top *t* SNPs from EM_BN				
Trait	**Method**[a]	***t*=5**	***t*=10**	***t*=15**	***t*=20**	***t*=25**
sim-D1	Unadjusted	0.991 (0.042)	0.990 (0.030)	0.981 (0.033)	0.975 (0.032)	0.973 (0.027)
	EM_IBS	0.999 (0.017)	0.999 (0.009)	0.997 (0.015)	0.997 (0.013)	0.996 (0.012)
	FLMM_A	1.000 (0.009)	1.000 (0.003)	1.000 (0.007)	1.000 (0.004)	1.000 (0.003)
	FLMM_E	0.998 (0.021)	1.000 (0.005)	0.999 (0.008)	0.999 (0.005)	1.000 (0.004)
	GA_FA	0.998 (0.018)	1.000 (0.005)	0.999 (0.011)	0.999 (0.008)	0.998 (0.008)
	GA_GRG	0.998 (0.021)	0.999 (0.011)	0.996 (0.017)	0.998 (0.010)	0.998 (0.008)
	GMA_C	0.998 (0.021)	1.000 (0.004)	0.999 (0.009)	0.999 (0.005)	1.000 (0.004)
	GMA_S	0.998 (0.021)	1.000 (0.005)	0.999 (0.008)	0.999 (0.005)	1.000 (0.004)
	GTAM	0.998 (0.022)	0.995 (0.022)	0.990 (0.025)	0.988 (0.022)	0.987 (0.020)
	Mendel	0.997 (0.025)	0.996 (0.019)	0.991 (0.024)	0.989 (0.021)	0.989 (0.018)
	MMM_E	0.991 (0.041)	1.000 (0.004)	0.999 (0.009)	0.999 (0.005)	1.000 (0.004)
	MMM_G	0.993 (0.036)	1.000 (0.003)	1.000 (0.007)	1.000 (0.005)	0.999 (0.005)
	FBATaff	0.684 (0.253)	0.790 (0.115)	0.773 (0.090)	0.771 (0.080)	0.760 (0.072)
	FBATboth	0.859 (0.130)	0.844 (0.084)	0.811 (0.078)	0.795 (0.075)	0.777 (0.071)
	MASTOR	0.993 (0.038)	0.994 (0.024)	0.989 (0.027)	0.985 (0.024)	0.985 (0.022)
	MQLS	0.978 (0.062)	0.981 (0.040)	0.960 (0.043)	0.951 (0.041)	0.941 (0.038)
	RT	0.981 (0.059)	0.984 (0.037)	0.962 (0.042)	0.952 (0.041)	0.942 (0.038)
sim-D2	Unadjusted	0.982 (0.060)	0.984 (0.041)	0.979 (0.039)	0.974 (0.040)	0.973 (0.036)
	EM_IBS	0.997 (0.029)	0.997 (0.024)	0.995 (0.025)	0.994 (0.028)	0.994 (0.024)
	FLMM_A	0.998 (0.027)	0.998 (0.024)	0.997 (0.025)	0.997 (0.029)	0.997 (0.026)
	FLMM_E	0.995 (0.035)	0.997 (0.025)	0.997 (0.025)	0.996 (0.030)	0.997 (0.026)
	GA_FA	0.992 (0.044)	0.998 (0.024)	0.997 (0.026)	0.996 (0.030)	0.996 (0.026)
	GA_GRG	0.994 (0.038)	0.997 (0.026)	0.996 (0.027)	0.995 (0.030)	0.996 (0.026)
	GMA_C	0.995 (0.035)	0.997 (0.025)	0.997 (0.025)	0.996 (0.030)	0.997 (0.026)
	GMA_S	0.995 (0.035)	0.997 (0.025)	0.997 (0.025)	0.996 (0.030)	0.997 (0.026)
	GTAM	0.988 (0.050)	0.990 (0.036)	0.983 (0.037)	0.982 (0.036)	0.982 (0.032)
	Mendel	0.988 (0.051)	0.992 (0.033)	0.986 (0.035)	0.984 (0.036)	0.987 (0.031)
	MMM_E	0.995 (0.037)	0.997 (0.025)	0.997 (0.025)	0.996 (0.030)	0.997 (0.026)
	MMM_G	0.998 (0.028)	0.998 (0.024)	0.997 (0.025)	0.997 (0.029)	0.997 (0.026)
	FBATaff	0.413 (0.255)	0.571 (0.201)	0.614 (0.157)	0.639 (0.128)	0.651 (0.102)
	FBATboth	0.664 (0.246)	0.718 (0.146)	0.699 (0.111)	0.691 (0.099)	0.686 (0.088)
	MASTOR	0.971 (0.075)	0.988 (0.038)	0.981 (0.038)	0.978 (0.039)	0.979 (0.033)
	MQLS	0.934 (0.107)	0.962 (0.056)	0.942 (0.053)	0.928 (0.051)	0.917 (0.047)
	RT	0.943 (0.099)	0.965 (0.055)	0.943 (0.053)	0.930 (0.052)	0.919 (0.047)
sim-Q	Unadjusted	0.987 (0.049)	0.983 (0.038)	0.962 (0.040)	0.963 (0.034)	0.954 (0.033)
	EM_IBS	0.998 (0.020)	0.998 (0.016)	0.993 (0.020)	0.994 (0.017)	0.993 (0.015)
	FLMM_A	1.000 (0.000)	1.000 (0.000)	1.000 (0.004)	1.000 (0.005)	1.000 (0.004)
	FLMM_E	1.000 (0.009)	0.999 (0.008)	1.000 (0.005)	1.000 (0.005)	0.999 (0.005)
	GA_FA	1.000 (0.006)	0.999 (0.010)	0.998 (0.010)	0.998 (0.010)	0.996 (0.012)
	GA_GRG	0.994 (0.034)	0.999 (0.010)	0.995 (0.018)	0.996 (0.014)	0.996 (0.012)
	GMA_C	1.000 (0.009)	1.000 (0.007)	1.000 (0.004)	1.000 (0.004)	1.000 (0.004)
	GMA_S	1.000 (0.009)	0.999 (0.008)	1.000 (0.005)	1.000 (0.005)	0.999 (0.005)
	GTAM	0.995 (0.032)	0.991 (0.028)	0.984 (0.030)	0.985 (0.024)	0.984 (0.022)
	Mendel	0.998 (0.021)	0.996 (0.020)	0.987 (0.027)	0.988 (0.022)	0.988 (0.019)
	MMM_E	0.899 (0.100)	0.999 (0.008)	1.000 (0.004)	1.000 (0.004)	1.000 (0.004)
	MMM_G	0.903 (0.100)	1.000 (0.003)	1.000 (0.003)	1.000 (0.004)	1.000 (0.003)
	FBAT	0.906 (0.101)	0.896 (0.067)	0.869 (0.059)	0.844 (0.067)	0.814 (0.066)
	MASTOR	0.998 (0.020)	0.992 (0.027)	0.984 (0.030)	0.984 (0.025)	0.983 (0.023)

[a]See Table 2 for description of methods.

doi:10.1371/journal.pgen.1004445.t003

Table 4. Genomic control factors achieved in naive analysis of a single replicate of the simulated longitudinal data sets.

	Trait analysed	
Method[a]	**Longitudinal (sim-L20)**	**Longitudinal polygenic (sim-P20)**
Unadjusted	20.82	21.53
EM_BN	1.01	1.01
EM_IBS	0.99	0.97
FLMM_A	1.01	1.01
FLMM_E	1.01	1.01
GA_FA	1.06	2.39
GA_GRG	0.66	0.47
GMA_C	1.01	1.01
GMA_S	1.01	1.01
MMM_E	1.01	3.52
MMM_G	1.01	3.52

[a]See Table 2 for description of methods.

doi:10.1371/journal.pgen.1004445.t004

GEMMA and MMM offer a convenient and robust approach for family-based GWAS of quantitative or binary traits, are successful in controlling the overall genomic inflation factor to an appropriate level, and offer higher power than traditional family-based association analysis approaches such as those implemented in FBAT. Similar inference is also provided by related and alternative approaches implemented in the software packages Mendel, ROADTRIPS, MQLS and MASTOR, although our results from analysis of the real data suggest that, for Mendel, MQLS and MASTOR, care may need to be taken to use estimated kinships based on SNP data rather than known pedigree relationships, if one is to avoid any inflation in the test statistics.

Our current study focused mostly on family data in which genuine close relationships between many individuals exist. Nevertheless we found similar results with respect to the LMM methods investigated (adequate control of type 1 error and extremely similar performance in terms of power and concordance between top findings) when applied to a subset of 462 founder individuals from our pedigrees, selected to be approximately unrelated to one another (see Figure S11 and Table S3). Therefore, we believe that our results highlighting the concordance between different LMM methods are equally relevant to researchers carrying out genome-wide association studies of apparently unrelated individuals as to researchers carrying out family-based studies.

Traditional methods for family-based association analysis make use of pedigree relationships either (e.g. FBAT) through direct use of known pedigree structure or else (e.g. MQLS, ROADTRIPS and all LMM methods) through use of a covariance matrix that involves the known kinship between each pair of individuals (the probability that a randomly chosen allele at a locus in each individual is identical by descent i.e. is a copy of a common ancestral allele, under the assumption that the pedigrees are correctly specified and all founders in a pedigree are completely unrelated i.e. share no alleles identical by descent). The assumption that all founders in a pedigree share no alleles identical by descent is clearly a fiction, given human population history, while the assumption that all pedigrees are correctly specified and unrelated to one another is also likely to be violated in most real studies. The use of estimated kinships based on SNP data rather than theoretical kinships based on known pedigree relationships removes the reliance on these untenable assumptions, and allows essentially the same analysis approaches to be applied to apparently unrelated individuals (who may nevertheless display distant levels of shared ancestry). The question then arises as to what exactly these estimated kinships (or related measures) are actually measuring? We consider a detailed discussion of this issue to be beyond the scope of the current manuscript, but we refer the reader to the more detailed expositions given in [36] and [37] which discuss some differences between different kinship measures as well as pointing out the difficulty of directly modelling identity by descent in the absence of an explicit pedigree. A key point when using estimated kinships to structure the covariance matrix in an association analysis (as here) is that our goal is not relationship estimation (close or distant) in its own right, but rather to adjust our analysis for phenotypic correlations between individuals due to genetic factors (usually assumed to be polygenic effects) that would otherwise result in inflated association test statistics. Therefore, one could argue that the extent to which the estimated kinship measures do or do not reflect genuine relationships between individuals (and how one should interpret such relationships) is largely irrelevant; the important issue is whether or not use of such kinships succeeds with respect to adequately modelling phenotypic correlations between individuals. On that note, in the analyses performed here we did not find large differences between the results obtained using different kinship measures, although use of the kinship measures output by PLINK (as well as use of completely incorrect kinship measures) did perform worse than the other kinship measures investigated.

The recent popularity of LMM approaches for the analysis of apparently unrelated individuals [1–4] has been partly motivated by a desire to correct for more complicated models of population structure including population stratification, rather than (or in addition to) correcting for relatedness between individuals. Population stratification can be thought of as a type of relatedness in that members of the same sub-population are effectively more closely related to one another than to individuals in other sub-populations, although it has been noted [36] that this sub-population or 'island model' underlying the traditional view of population stratification may be unduly simplistic. The observation that LMM approaches have sometimes worked better than traditional principal component approaches at correcting for apparent population structure [1] may reflect the fact that the inflation seen in genome-wide test statistics (in the absence of any correction) results not from population stratification under an 'island model' per se, but rather from more complicated

population structure (involving distant ancestral relationships between individuals). A recent paper by Wang et al. [38] showed that, in the presence of cryptic relatedness between study subjects (but no population stratification), both principal component and LMM methods are valid (in the sense of generating test statistics with the desired distribution under the null hypothesis), but LMM approaches are more powerful for detecting association. In contrast, in the presence of population stratification, neither principal component nor LMM methods are strictly valid, but LMM methods seem to display better overall performance.

An interesting finding of our current study was the fact that longitudinal traits (repeated measures) could be successfully analysed in an LMM framework simply by treating each measurement as if it came from a separate person and expanding out the genetic data set accordingly (resulting in an expanded data set containing many apparent twins, triplets, quadruplets etc.). From a practical point of view this is useful as analysis of an expanded data set in standard LMM software is computationally convenient; we found a 'proper' analysis using software such as longGWAS [34] to be prohibitively slow when applied to our data set.

A caveat to all the results presented here is that they relate to genotypes derived from a single data set, our Brazilian family study of visceral leishmaniasis [13]. (Although the results in terms of the performance and power of different methods were comparable across both real and simulated data sets, even in the simulated data all genotypes were held fixed and only phenotypes were re-simulated). However, we have good reason to believe that the high concordance between different LMM implementations seen here (as well as their performance from when applied naively to longitudinal data) will hold more generally for genetic studies of diverse phenotypes carried out in diverse human populations. We observed essentially the same pattern of results described here when we applied a more limited set of LMM implementations to GWAS data from Genetic Analysis Workshop 18 (959 Mexican-American individuals from 20 families, with real and simulated phenotypes) [18] as well as when we applied these approaches to GWAS data from 402 Aboriginal Australian individuals that cluster loosely into 4 large nominal pedigrees (unpublished data). Therefore, although it is possible that highly structured populations (such as those encountered in plant or animal breeding experiments) may uncover subtle differences between the various LMM approaches, for researchers carrying out complex genetic disease studies in human populations, we anticipate there will be little difference between the results seen from one approach over another, and the choice of which method/software package to use will be largely dictated by personal taste or convenience.

On this note, we point out that each package has its own particular advantages (and disadvantages). These include the fact that EMMAX, GEMMA and MMM allow the input of dosages derived from imputed (in addition to real) genotypes; MMM has the advantage of allowing the output of regression coefficients and standard errors for the SNP effects on the (log) odds ratio scale, making it convenient to compare or combine the results with results from traditional case/control studies analysed via logistic regression; GenABEL (GRAMMAR-Gamma) has the advantage of scaling linearly with sample size, which makes it attractive for the analysis of very large data sets; FaST-LMM has the advantage, along with EMMAX and Mendel, of internally imputing missing data at any (genetic or non-genetic) covariates, which can make it convenient for implementing stepwise conditional analyses; and, unlike most LMM implementations, ROADTRIPS, MQLS and MASTOR have the advantage of using all phenotype information, including that for individuals that have not been genotyped, which can in theory generate a small increase in power.

One of the main differences between the different software implementations we investigated was the time taken to perform the analysis (not including the time required to re-format data into an appropriate format for a given package). We were unable to do a strict head-to-head comparison as the precise timings depend on a number of factors including the computer architecture available (in particular the ability of the system and of any given program to allow multi-threading and/or parallel processing), however our rough comparison (Table S2), assuming that kinships are to be estimated on the basis of SNP data, implicated FaST-LMM (approximate calculation), GenABEL (GRAMMAR-Gamma) and EMMAX as generally the fastest implementations.

In conclusion, we recommend linear mixed model approaches as a convenient and powerful approach for family-based GWAS of quantitative or binary traits. We find these approaches to be successful in controlling the overall genome-wide error rate and to perform well in comparison to competing approaches.

Materials and Methods

Ethics statement

Ethical approval for the Belem Family Study was obtained originally from the local ethics committee at the Instituto Evandro Chagas, Belém, Para, Brazil. Approval for continued use of the Belem Family Study samples, and for collection and use of the samples from Natal, has been granted from the local Institutional Review Board at the Universidade Federal do Rio Grande do Norte (CEP-UFRN 94–2004), nationally from the Comissão Nacional de Ética em Pesquisa (CONEP: 11019), and from the Ministerios Cencia e Tecnologia for approval to ship samples out of Brazil (portaria 617; 28 September 2005). Informed written consent for sample collection was obtained from adults, and from parents of children <18 years old.

Subjects and genotyping

Sample collection and genotyping of the Brazilian subjects used here is described in detail in [13]. In brief, we ascertained 348 families comprising 65 families collected from sites around Belém and 283 families collected from sites around Natal in north east Brazil. All families were ascertained on the basis of containing multiple individuals that had been diagnosed with clinical visceral leishmaniasis. DNA from 2159 family members was genotyped at the Wellcome Trust Sanger Institute using the Illumina Human660-Quad chip. Extensive quality control checks were employed to retain only high quality samples [13], and to exclude samples whose apparent relatedness (as assessed based on estimated genome-wide average identity by descent, calculated using a subset of 11,177 high-quality autosomal SNPs via the `--genome` command in PLINK [27]) was incompatible with their known pedigree relationships (and for whom such discrepancies could not be resolved on further investigation). SNP quality control checks were used to retain only a subset of the genome-wide SNPs that could be expected to be of high quality. For the current investigation, we used slightly more stringent SNP exclusion thresholds than had been used in [13], namely SNPs were excluded if their minor allele frequency was <0.01, if the Fisher information for the allele frequency <0.98, if call rate <0.99, or if the p-value for a test of Hardy Weinburg Equilibrium $<10^{-6}$. These quality control checks resulted in the retention of 1972 genotyped individuals (357 cases, 1613 controls and two individuals of unknown phenotype) from 308 families (244 from Natal, 64 from Belém), each genotyped at 545,433 autosomal SNPs.

For the majority of analyses considered here, we used either the 1972 genotyped individuals or else the entire set of 3626

individuals (with or without genotype data) that are required to define the 'known' (theoretical) pedigree relationships. For power comparisons between LMM methods, we also investigated use of a subset of 462 'founder' individuals, chosen on the basis of theoretical relationships and estimated kinships to be approximately unrelated to one another.

Generation of simulated phenotypes

We generated simulated phenotypes for the 1972 individuals that had genome-wide SNP data available. We used two different models for generating binary (disease) traits, one corresponding to 'strong' genetic effects (sim-D1) and one corresponding to 'weak' genetic effects (sim-D2), with the trait in both cases governed by two SNPs (rs9271252 and rs233722) located on chromosomes 6 and 12 respectively. In addition to modelling genetic effects at rs9271252 and rs233722, we allowed for 22 weaker 'polygenic' effects caused by genotype at the 100th SNP on each autosomal chromosome. Each effect contributed multiplicatively to the probability of developing disease. Thus, the mathematical model for generating the simulated phenotype was

$$\text{Penetrance} = \alpha \prod_{j=1}^{24} \beta_j^{x_j}$$

where x_j was a variable coded (0, 1, 2) according to the number of copies of the risk allele possessed at causal SNP j (with $j=1$ corresponding to rs9271252 and $j=2$ corresponding to rs233722), the baseline penetrance α was set to equal 0.017 for the 'strong' scenario and 0.022 for the 'weak' scenario, β_1 was set to equal 2 for the 'strong' scenario and 1.6 for the 'weak' scenario, β_2 was set to equal 1.8 for the 'strong' scenario and 1.55 for the 'weak' scenario, and β_j $(j=3,\ldots 24)$ was set to equal 1.1 under both scenarios. Resulting penetrances greater than 1.0 were assigned to equal 1.0.

We also simulated a model (sim-Q) for quantitative traits, again governed by rs9271252 and rs233722 on chromosomes 6 and 12. The traits were generated as a linear combination of the effect from each of the strong and polygenic effect SNPs, with a normally distributed error component, thus:

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \sum_{j=3}^{24} \beta_j x_{ij} + \epsilon_i$$

where x_{ij} was a genotype variable for person i at SNP j coded as above, α represents the baseline trait and was set to 100, β_1 was set to 3, β_2 to 2, β_j $(j=3,\ldots 24)$ which correspond to polygenic contributions for SNP i were set to 1, and ϵ_i was a randomly generated variable following a normal distribution with mean 0 and standard deviation 5.

We simulated a model (sim-L20) for longitudinal quantitative traits (with $k=20$ repeated measures for each individual) in a rather similar manner, with individuals' non-genetic variation accounted for by another error term δ_i:

$$y_{i_k} = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \sum_{j=3}^{24} \beta_j x_{ij} + \delta_i + \epsilon_{i_k}$$

The baseline trait α remained 100, β_1 was set to 5, β_2 to 4, β_j $(j=3,\ldots 24)$ were set to 1.5, δ_i was a random variable following a normal distribution with mean 0 and standard deviation 4, generated once for each individual. The residual error term ϵ_{i_k} was a randomly generated variable following a normal distribution with mean 0 and standard deviation 2.

To make the analyses feasible whilst still maintaining the overall degree of relatedness, the longitudinal data set was constructed based on a subset of 498 individuals selected through stratified sampling from the original data set, with number of individuals randomly selected from each extended family approximately proportional to their family size while also ensuring that every family is represented by at least one individual. Phenotypes for these 498 individuals were then generated 20 times to create the final longitudinal data set.

In addition we simulated a purely polygenic longitudinal model (sim-P20) in which the strong effects β_1 and β_2 did not exist, and the 22 polygenic effects β_j $(j=3,\ldots 24)$ were replaced by 402 polygenic effects β_j $(j=3,\ldots 404)$ which were set to 0.75. In this model, α was set to 20, δ_i followed a normal distribution with mean 0 and standard deviation 16, and ϵ_{i_k} followed a normal distribution with mean 0 and standard deviation 1.

We generated 1000 replicates of each simulated data set, apart from the longitudinal and polygenic longitudinal data sets for which we only simulated a single replicate. For visualisation of results from a whole genome scan, we analysed only a single replicate (replicate 1). For investigation of power, type 1 error and concordance, to reduce computation time we analysed all 1000 replicates but only generated test statistics at 40 SNPs that lay within 40 kb of the simulated disease loci (for evaluation of power) and 20 SNPs that lay well outside the region of any simulated disease loci (for evaluation of type 1 error). By default, the programs Mendel and ROADTRIPS require all SNPs that are being used to estimate genome-wide relatedness to also be read in and tested for association; to perform the analysis of all 1000 replicates in reasonable time we therefore included the 50,129 'pruned' SNPs rather than the full genome-wide set of SNPs that would normally be used by these programs.

Linear mixed models methods and software

All the LMM implementations evaluated here attempt to fit either an exact or an approximate version of the standard linear mixed model:

$$\mathbf{y} = X\beta + Q + \epsilon$$

where $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ is a vector of responses (either quantitative traits or binary traits coded 1/0 for case/control status) on n subjects, $X = (x_{ij})$ is the $n \times J$ matrix of predictor variables to be modelled as fixed effects, including variables representing genetic and/or non-genetic covariates as well as a vector of variables $\mathbf{x}_1$ representing the genotypes at a particular SNP currently being tested (generally coded as (0,1,2) according to the number of copies of a particular allele possessed), $\beta = (\beta_1, \beta_2, \ldots, \beta_J)$ are regression coefficients (to be estimated) representing the linear effects of predictors on response, and Q and ϵ are random effects assumed to follow the distributions $Q \sim N(0, 2\Phi\sigma_g^2)$ and $\epsilon \sim N(0, \sigma_e^2 I)$ respectively (where σ_g^2 and σ_e^2 are parameters to be estimated representing genetic and environmental components of variance, I is the $n \times n$ identity matrix and Φ is an $n \times n$ matrix of pairwise kinship coefficients).

GenABEL (FASTA). The `mmscore` and `polygenic` functions of the GenABEL package [39] together allow implementation of the FAmily based Score Test Approximation (FASTA) method proposed by Chen and Abecasis [9]. The FASTA method is also implemented in the `--fast-Assoc` option of the MERLIN [40]

package, however MERLIN calculates the kinship matrix Φ internally on the basis of known (theoretical) kinships constructed from known pedigree relationships, rather than allowing the pairwise kinship coefficients to be estimated using genome-wide SNP genotype data [12]. We therefore preferred to use GenABEL, which can read in a user-specified matrix Φ constructed on the basis of either theoretical or estimated kinship coefficients.

Rather than fitting the full linear mixed model $\mathbf{y}=X\beta+Q+\epsilon$ and estimating β, σ_g^2 and σ_e^2 by maximum likelihood for each SNP across the genome, FASTA implements an 'approximate' two-stage approach. At the first stage a reduced model is fitted, where the regression coefficient β_1 (corresponding to the effect at the SNP currently under test) is assumed to equal 0. At the second stage, a score statistic for testing the null hypothesis that β_1 does indeed equal 0 is constructed as:

$$T_{\mathrm{FA}}=\frac{([\mathbf{x}_1-E(\mathbf{x}_1)]^T\Omega^{-1}[\mathbf{y}-E(\mathbf{y})])^2}{[\mathbf{x}_1-E(\mathbf{x}_1)]^T\Omega^{-1}[\mathbf{x}_1-E(\mathbf{x}_1)]}$$

where $E(\mathbf{y})$ refers to an n-dimensional vector of fitted values of the response from the reduced model, $E(\mathbf{x}_1)$ refers to an n-dimensional vector of unconditional expectations of genotype scores at the test SNP (each element of which equals twice the allele frequency of the particular allele being counted), and Ω refers to the estimated variance/covariance matrix, $\Omega=2\Phi\sigma_g^2+\sigma_e^2 I$, with σ_g and σ_e taking their maximum likelihood estimates as calculated under the reduced model. The score statistic is calculated repeatedly using the appropriate n-dimensional vector $\mathbf{x}_1$ for each test SNP (typically between 500,000 and several million SNPs) across the genome, but the time-consuming maximum likelihood step for estimating σ_g^2, σ_e^2 and $(\beta_2,...,\beta_J)$ need only be performed once, at the start.

GenABEL (Grammar-Gamma). The `grammar` function of the GenABEL package [39] implements the GRAMMAR-Gamma method proposed by Svishcheva et al. [14]. This method can be considered as an extension of the original GRAMMAR method [10,12] to produce a test that is essentially a fast approximation to FASTA.

In GRAMMAR [10], similarly to FASTA, the first step is to fit a reduced version of the full linear mixed model in which β_1 is set to 0. Phenotype residuals $\tilde{\mathbf{y}}=(\tilde{y}_1,\tilde{y}_2,...,\tilde{y}_n)^T$ may be constructed as $\tilde{y}_i=y_i-E(y_i)$ where $E(y_i)$ refers to the fitted value of the response for person i from the reduced model. These residuals are then used as the independent trait in a simple linear regression model:

$$\tilde{y}_i=\mu+\tilde{\beta}_1 x_{i1}+e_i$$

where the error term e_i is assumed to be independently normally distributed. Estimation of $\tilde{\beta}_1$ and testing of the null hypothesis that $\tilde{\beta}_1=0$ can be accomplished through maximum likelihood or least squares approaches. Alternatively, a rapid test of $\tilde{\beta}_1=0$ can be achieved [12,14] through construction of a score statistic:

$$T_{\mathrm{GR}}=\frac{n([\mathbf{x}_1-E(\mathbf{x}_1)]^T[\tilde{\mathbf{y}}^*])^2}{[\mathbf{x}_1-E(\mathbf{x}_1)]^T[\mathbf{x}_1-E(\mathbf{x}_1)][\tilde{\mathbf{y}}^*]^T[\tilde{\mathbf{y}}^*]}$$

where $\tilde{\mathbf{y}}^*=(\tilde{y}_1^*,\tilde{y}_2^*,...\tilde{y}_n^*)$ are transformed version of the residuals $\tilde{\mathbf{y}}^*=\sigma_e^2\Omega^{-1}\tilde{\mathbf{y}}$. Again, the time-consuming maximum likelihood step for estimating σ_g^2, σ_e^2 and $(\beta_2,...,\beta_J)$ (and thus for calculating the transformed residuals $\tilde{\mathbf{y}}^*$) need only be performed once.

In the original GRAMMAR publication [10], the assumption was that pedigree relationships between individuals would be known and so Φ would be constructed on the basis of theoretical kinship coefficients. Subsequently it was suggested [12] that the use of estimated kinship coefficients (estimated on the basis of genome-wide SNP data) could perform as well or better. Regardless of which kinship coefficients are used, GRAMMAR was found to be conservative and to result in biased regression coefficients representing the SNP effects of interest [12], and so it was suggested that the final χ^2 test statistics should be 're-inflated' by multiplying by an appropriate estimated correction factor (in a procedure analogous to the 'deflation' of χ^2 test statistics via genomic control [29]) to result in a final test statistic with the appropriate null distribution. This 'genomic control corrected' version of GRAMMAR was denoted GRAMMAR-GC by [12].

The GRAMMAR-Gamma method [14] is similar to GRAMMAR but, unlike GRAMMAR, produces unbiased SNP effect estimates and test statistics that do not require any deflation. The method involves calculating a GRAMMAR-Gamma correction factor γ (see [14] for details) that is used to adjust a new statistic

$$T_{\mathrm{new}}=\frac{([\mathbf{x}_1-E(\mathbf{x}_1)]^T\Omega^{-1}[\mathbf{y}-E(\mathbf{y})])^2}{[\mathbf{x}_1-E(\mathbf{x}_1)]^T[\mathbf{x}_1-E(\mathbf{x}_1)]}$$

which can be calculated from a standard linear regression analysis of $\Omega^{-1}[\mathbf{y}-E(\mathbf{y})]$ on $[\mathbf{x}_1-E(\mathbf{x}_1)]$. This results in a final GRAMMAR-Gamma statistic $T_{\mathrm{GRG}}=T_{\mathrm{new}}/\gamma$ that can be shown to be approximately equivalent to the FASTA statistic T_{FA}. Svishcheva et al. [14] argue that their GRAMMAR-Gamma method has similar computational complexity to alternative methods such as FASTA, EMMAX and FaST-LMM at stage 1, while achieving computational savings over these methods at stage 2 (achieving a stage 2 computational complexity of $O(sn)$, where n is the sample size and s the number of SNPs to be tested).

EMMAX. Kang et al. [1] proposed a method that appears to be essentially equivalent to the FASTA method proposed by Chen and Abecasis [9], except for the following caveats:

1. In the approach of Kang et al. [1], there is no expectation that the individuals will be closely related, indeed the method is motivated as an alternative to principal component based approaches when adjusting for population substructure in genome-wide association studies of unrelated individuals. Thus, the kinship coefficients used to construct Φ are not based on any 'known' pedigree relationships but are estimated based on genome-wide SNP data (using either a simple estimate based on the proportion of alleles identical-by-state (IBS) measure, or else an estimate that Kang et al. [1] describe as a Balding-Nichols (BN) estimate), resulting in a procedure essentially identical to that proposed by Amin et al. [12].
2. In the approach of Kang et al. [1], rather than applying the method solely to quantitative traits as had been done previously [9,10,12], the method is also proposed to apply to case/control data (with the response coded as 0 or 1, but analysed as if it were, in fact, a quantitative trait, i.e. assuming a normally distributed random environmental/error term ϵ). Kang et al. argue that this is computationally more convenient than fitting a generalized linear mixed model with a logit or probit link function (which would be the usual way to analyse binary response data) and should not result in increased type 1 error for testing the null hypothesis.
3. Although not entirely clear from the description in Kang et al. [1], it appears that, at the second stage, in contrast to [9], any

covariates other than the SNP currently under test are re-estimated i.e. the entire vector of fixed effect predictors $\beta=(\beta_1,\beta_2,...,\beta_J)$ is estimated, rather than fixing $(\beta_2,...,\beta_J)$ at their estimated values from the first stage.

The method of Kang et al. [1] has been implemented in the software package EMMAX. As pointed out by Lippert et al. [4], EMMAX, along with its predecessor EMMA [41], achieves additional computational efficiency (over and above that achieved by simply estimating parameters σ_g^2 and σ_e^2 only once) by reparameterising the likelihood in terms of a parameter $\delta=\sigma_e^2/\sigma_g^2$ (which is estimated only once) and by making clever use of spectral decompositions. This results in a computational complexity of $O(n^3+rn)$ at stage 1 (where r the number of iterations i.e. the number of evaluations of the likelihood required) together with a computational complexity of $O(sn^2)$ at stage 2, resulting in a total computational complexity of $O(n^3+sn^2+rn)$.

A similar approach to [1] and [9] was proposed by Zhang et al. [2] and implemented in a software package TASSEL. The main focus of the paper by Zhang et al [2] was to describe a clustering algorithm that results in an approximation to the kinship matrix with lower effective dimensionality, which can be used in place of the full known or estimated kinship matrix. Similarly to EMMAX, in TASSEL the values of σ_g^2 and σ_e^2 (as well as a cluster membership variable C) are estimated under the null hypothesis that $\beta_1=0$ (at stage 1) and are then held fixed while estimating $\beta=(\beta_1,\beta_2,...,\beta_J)$ (at stage 2). The motivation for the clustering approximation is to reduce computation time. However, existing software packages (e.g. EMMAX and the `mmscore` and `polygenic` functions in GenABEL) that address the problem without making such an approximation are not computationally prohibitively time consuming. Therefore it is unclear why use of this approximation should be preferred. For this reason, given the extreme similarity between the methods implemented in EMMAX and TASSEL when no clustering is performed, we have not included TASSEL in our comparisons.

FaST-LMM. Lippert et al. [4] developed a fast 'exact' LMM implementation that, in common with EMMAX, reparameterises the likelihood in terms of a parameter $\delta=\sigma_e^2/\sigma_g^2$, and also requires only a single spectral decomposition at the first stage of the algorithm, resulting in a total time complexity of $O(n^3+sn^2+rsn)$. This exact method is the default in the current (2.04) version of FaST-LMM. (In previous versions the default was to use an approximate method in which δ is fixed to its value from fitting a null model containing no fixed SNP effects, as is done in EMMAX, TASSEL and FASTA, resulting in a reduced complexity of $O(n^3+sn^2+rn)$. This approximate method is now available in FaST-LMM as an optional alternative to the exact method). A further speed-up can be achieved in FaST-LMM by restricting the number of SNPs used to construct the kinship matrix Φ to a number less than the number of individuals.

FaST-LMM uses either maximum likelihood (ML) or restricted maximum likelihood (REML). In early versions of FaST-LMM the default was ML but in later versions the default became REML. After some experimentation, we deemed ML to be the most reliable and have used that for all results presented here.

GEMMA. Zhou and Stephens [16] implemented an exact approach extremely similar to that of FaST-LMM in their package GEMMA. Indeed, Zhou and Stephens themselves point out that GEMMA should give essentially identical inference to FaST-LMM in the same time complexity $O(n^3+sn^2+rsn)$, but note that the number of iterations (r) required to reach convergence in GEMMA is expected to be slightly smaller than in FaST-LMM, owing to the use of a more efficient optimization method. GEMMA also has an attractive practical advantage of allowing the input of imputed [42] genotype data, rather than real measured genotype data, if desired.

MMM. Pirinen et al. [15] have implemented approximate and exact approaches similar to the approximate and exact approaches of FaST-LMM (and the exact approach of GEMMA) in their package MMM. An advantage of MMM in comparison to the other packages is that it allows the output of regression coefficients and standard errors for the SNP effects on the (log) odds ratio scale, making it convenient to compare or combine the results with results from traditional case/control studies analysed via logistic regression. In addition, MMM allows the input of imputed genotype data rather than real measured genotype data, if desired. MMM was used in the original analysis of the Brazilian VL family data described in [13]. For more details on the methodology implemented in MMM, see [15].

Mendel. An approximate (score test) LMM implementation, suitable for analysis of GWAS data, has also been implemented in the software package Mendel [35] (versions 13.0 and higher). A slower (exact) LMM implementation is also available, but we only considered the approximate test here. Mendel can a. calculate kinship coefficients on the basis of known pedigree relationships, b. use the full set of genome-wide SNP data to cluster people into apparent pedigrees and then estimate kinship coefficients within those pedigree clusters, or c. use kinship coefficients estimated for all pairs of genotyped individuals on the basis of their full set of genome-wide SNPs. The resulting tests should be conceptually extremely similar to the LMM tests implemented in other software packages such as EMMAX and FaST-LMM.

Alternative methods and software

FBAT. Traditional approaches for family-based association analysis focus on the transmission of high-risk alleles through pedigrees, in an approach that is closely related to traditional linkage analysis. Indeed, the well-known transmission disequilibrium test (TDT) [19], which tests whether a particular allele is transmitted preferentially from heterozygous parents to affected offspring, was originally developed as a test of linkage in the presence of association, rather than as a test of association per se. In this context, by 'linkage' we mean the transmission from parent to offspring of alleles in coupling at a test (marker) locus and an unobserved causal locus, i.e. the phenomenon whereby alleles that are in coupling (on the same haplotype) in the parent tend to be transmitted together to the offspring, whereas by 'association' we mean population-level correlation between alleles at the two loci (usually referred to as linkage disequilibrium (LD)), i.e. the tendency for alleles at the two loci to occur in coupling in the founders of a pedigree.

The TDT was originally designed for the analysis of case/parent trios (i.e. units consisting of an affected child together with their parents) but has been extended to allow analysis of nuclear families and larger pedigrees [20,21,23,43–46]. Here we focus on the family-based association test (FBAT) [21,23], as implemented in the FBAT software package. FBAT can be thought of as a general class of test statistics of the form

$$\frac{S-E(S)}{\sqrt{\mathrm{Var}(S)}}$$

where $S=\sum_{ij}T_{ij}X_{ij}$ and X_{ij} is some genotype variable and T_{ij} some trait variable for offspring i in nuclear family j. The exact form of FBAT thus depends on the genotype and trait coding

used. Genotype is generally coded in allelic fashion with a variable coded (0, 1, 2) according to the number of copies of the high-risk allele possessed. The trait variable is constructed as $T_{ij} = Y_{ij} - \mu_{ij}$ where Y_{ij} is coded 0/1 (for binary traits such as disease status) and μ_{ij} is an offset that can be chosen to consider transmissions to affected offspring only (the default), or else to contrast transmissions to affected offspring with transmissions to unaffected offspring, either weighted equally ($\mu_{ij} = 0.5$) or with μ_{ij} chosen to minimize the variance of test statistic. For quantitative traits, Y_{ij} would generally correspond to the measured trait for offspring i in nuclear family j, with μ_{ij} set to equal the mean trait value or else chosen to minimize the variance of test statistic.

Although, for binary traits, contrasting transmissions to affecteds with transmissions to unaffecteds seems an attractive idea, in practice this results in comparing the probability of transmission of high-risk alleles to affected individuals (which is expected, under the alternative hypothesis, to exceed 0.5) with an *estimate* of the probability of transmission of high-risk alleles to unaffected individuals (which is expected, under both null and alternative hypotheses, to approximately equal 0.5, unless the effect of the risk allele is large), rather than comparing the transmission probability to affecteds with an assumed fixed value of 0.5. For complex diseases, where the effects of risk alleles are likely to be modest (allelic odds ratios in the order 1.2–1.5), this means that greater power would be expected from the default offset that considers transmissions to affected offspring only, without paying a penalty for (imperfect) estimation of the expected 0.5 transmission probability (along with a measure of uncertainty in the estimate) from the data at hand.

By default, FBAT divides larger pedigrees into nuclear families and constructs a test that corresponds to testing 'linkage in the presence of association' [23]. The '-e' option in FBAT allows the alternative construction of a test for 'association in the presence of linkage' [22], through use of an empirical variance/covariance estimator that adjusts for the correlation among sibling genotypes and for different nuclear families within a single pedigree. Use of the '-e' option is expected to give smaller test statistics (larger p-values) than the default analysis, since it accounts for the fact that the effective sample size is smaller when considering FBAT as a test of association than as a test of linkage. Since, for complex diseases, we are interested in maximizing the power for detection of an effect, rather than in ensuring that the detection is genuinely driven by association (rather than linkage) between alleles at our test locus and the underlying unobserved causal locus, we use the default option in all analyses presented here. From a practical point of view, this means that any signal we detect may in fact be marking a true effect that lies some distance away, rather than necessarily being located in the immediate vicinity of the detected signal.

ROADTRIPS and MQLS. Thornton and McPeek [26] implemented a '**RO**bust **A**ssociation- **D**etection **T**est for **R**elated **I**ndividuals with **P**opulation **S**ubstructure' in a package called ROADTRIPS. ROADTRIPS can be thought of as an extension of their previously-proposed Maximum Quasi-Likelihood Statistic (MQLS) [24]. Both MQLS and ROADTRIPS construct adjusted versions of standard case/control χ^2 (or Armitage Trend) tests, adjusting for the known relatedness between individuals (that would ordinarily cause an inflation in standard case/control tests) through a kinship matrix that models the known pedigree relationships. ROADTRIPS (but not MQLS) additionally makes use of a covariance matrix based on estimated kinships (as estimated from genome-wide SNP data) to further correct for additional unknown relatedness and population stratification.

The ROADTRIPS test statistic takes the form:

$$\frac{(\mathbf{V}^T\mathbf{Y})^2}{\hat{\sigma}^2\mathbf{V}^T\hat{\mathbf{\Psi}}\mathbf{V}} \sim \chi^2_1$$

Thornton and McPeek note that many commonly-used case/control statistics can be coerced into this form. Here $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ is genotype vector at a test SNP for n individuals (coded using an allelic coding), $\mathbf{V}$ is a vector of length n coding for phenotype information (disease status) and known (or externally estimated) relationships (see [26] for details of its construction), $\hat{\sigma}^2\hat{\mathbf{\Psi}}$ is an estimate of the null variance/covariance matrix of $\mathbf{Y}$ (so that $\hat{\sigma}^2\mathbf{V}^T\hat{\mathbf{\Psi}}\mathbf{V}$ is an estimate of null variance/covariance of $(\mathbf{V}^T\mathbf{Y})^2$), $\hat{\sigma}^2$ is an estimate of Var($\mathbf{Y}$) in an outbred population and $\hat{\mathbf{\Psi}}$ is an internally estimated matrix used to simultaneously adjust for unknown relatedness/pedigree relationship errors and population stratification.

MASTOR and GTAM. Recently, Jakobsdottir and McPeek [25] proposed a retrospective approach (MASTOR) for analysis of quantitative traits that can be considered essentially as a quantitative trait version of MQLS. In common with MQLS, kinships are assumed to be estimated on the basis of known pedigree relationships, but in principle kinships estimated from genome-wide SNP data could be read in instead. Jakobsdottir and McPeek compared MASTOR to a previously-proposed LMM method, GTAM [8], and found MASTOR to have some advantages. The main advantage of MASTOR over GTAM (and many other approaches) is that, in common with MQLS and ROADTRIPS, MASTOR allows information to be gained from individuals who are phenotyped but not genotyped. Both MASTOR and GTAM are implemented within the MASTOR software package. Although designed for analysis of quantitative (rather than binary) traits, given that the spirit of recent LMM approaches has been to apply approaches originally designed for quantitative traits to binary traits (coded as 0 and 1), we investigated the performance of MASTOR and GTAM when applied to both binary and quantitative traits.

Calculation of kinship coefficients

The LMM approaches considered here, as well as methods such as MQLS, ROADTRIPS, MASTOR and GTAM, all involve modelling the relatedness between individuals through one or more kinship matrices, constructed either on the basis of known (hypothesized) pedigree relationships between individuals, or through estimating kinships on the basis of genome-wide SNP data (or from a subset of available genome-wide SNPs). The precise algorithms used to estimate kinships on the basis of genome-wide SNP data vary [36,37,47], although we have found the kinship matrices from the different packages we considered to be largely comparable (see Results). Most packages allow a separation between the estimation of the kinship matrix step and the analysis (incorporating the desired kinship matrix) step. This is convenient as it allows a potentially different set of SNPs to be used for estimating the kinship matrix as is used for genome-wide association testing. It also means that kinships estimated using one package can potentially be read in to another package at the analysis stage, if desired. For the majority of analyses performed here, we used the same software package (or a recommended accompanying software package) to calculate the kinship matrix as we used for subsequent association testing, and to estimate the kinship matrix we used a subset of 50,129 'pruned' SNPs with minor allele frequencies >0.4 and 'pruned' to be in approximate

linkage equilibrium via the `--indep 50 5 2` command in PLINK [27]). (We found little difference between the results obtained when using such a pruned set of SNPs and using the full genome-wide set of SNPs, see Results).

We also explored the use of a smaller set of 1900 'thinned' SNPs to estimate kinships. This number was chosen to capitalise on the speed-up that can be achieved in FaST-LMM by restricting the number of SNPs used to construct the kinship matrix Φ to a number less than the number of individuals. The 'thinned' SNPs comprised an evenly-spaced subset of the 'pruned' SNPs selected based purely on physical position using the software package MapThin (http://www.staff.ncl.ac.uk/richard.howey/mapthin/). In addition we explored the use of the FaST-LMM-Select procedure [30], implemented within the FaST-LMM package, that uses an iterative procedure to select SNPs for inclusion in the construction of the kinship matrix on the basis of their nominal association with phenotype (as evaluated through a fixed effects linear regression analysis). However, we did not find this procedure to be superior to using either the pruned or the full set of SNPs (see Results).

Several alternative packages exist for estimating genetic relationships from genome-wide SNP data, either for subsequent use in LMM type analyses [48] or in order to infer pedigree relationships as an end in itself [28]. We investigated use of the kinship estimates output by the packages PLINK [27] and KING [28], in comparison to those calculated internally by the various LMM packages we had used. Another popular package is GCTA [48]; we note that the realised relationship matrix (RRM) kinship estimation approach used internally by FaST-LMM is theoretically equivalent to that used by GCTA.

Supporting Information

Figure S1 Comparison of estimated kinship measures and −log10(p-values) obtained based on full, pruned and thinned SNPs. (A) Estimated kinship measures (B) $-\log_{10}$ p-values obtained. F = full set, P = pruned set, T = thinned set. EM_BN = EMMAX (Balding-Nichols), EM_IBS = EMMAX (IBS method), FLMM_C = FaST-LMM using covariance matrix, FLMM_R = FaST-LMM using realised relationship matrix, GA = GenABEL, GA_FA = GenABEL (FASTA), GA_GRG = GenABEL (GRAMMAR-Gamma), GMA_C = GEMMA using centred genotypes, GMA_S = GEMMA using standardised genotypes, KING_H = KING with homogeneous population assumption, KING_R = KING with robust estimation, MMM_E = MMM using full mixed model (exact) calculation, MMM_G = MMM using GLS approximation.
(TIF)

Figure S2 QQ plots of real VL phenotype GWAS results, using different LMM software packages and different SNP sets for kinship estimation. The black diagonal lines represent the line of equality. The "theoretical" set used pedigree structure to derive theoretical kinship coefficients. EM_BN = EMMAX (Balding-Nichols), EM_IBS = EMMAX (IBS method), FLMM_C = FaST-LMM using covariance matrix, FLMM_R = FaST-LMM using realised relationship matrix, GA_FA = GenABEL (FASTA), GA_GRG = GenABEL (GRAMMAR-Gamma), GMA_C = GEMMA using centred genotypes, GMA_S = GEMMA using standardised genotypes, MMM_E = MMM using full mixed model (exact) calculation, MMM_G = MMM using GLS approximation, Unadj = unadjusted analysis. For methods with two ways to estimate the kinships, the same "theoretical" results were plotted twice. Unadjusted analysis results were plotted once in each column only for comparison, and did not use the kinship estimates for adjustment.
(TIF)

Figure S3 Performance of FaST-LMM-Select. Genomic control factor (λ_{GC}) achieved in analysis of the real disease phenotype as different numbers of ordered SNPs are added in when calculating the kinship matrix (= realised relationship matrix, RRM). Method implemented manually in FaST-LMM v2.0.
(TIF)

Figure S4 Manhattan plots for real and simulated data sets using FaST-LMM. The points marked in red denote either the confirmed significant region from Fakiola et al. (2013) (real phenotype), or the regions close to the simulated strong/weak effect SNPs (simulated phenotypes). real = real VL phenotype, sim-D1 = simulated strong binary (disease) trait, sim-D2 = simulated weak binary (disease) trait, sim-Q = simulated quantitative trait, sim-L20 = simulated longitudinal quantitative trait with 20 observations, sim-P20 = simulated polygenic longitudinal quantitative trait with 20 observations.
(TIF)

Figure S5 Manhattan plots for the simulated weak binary (disease) phenotype using FaST-LMM exact and alternative software packages. The points marked in red denote the regions close to the simulated weak effect SNPs. FLMM_E = FaST-LMM using exact calculation, RT = ROADTRIPS, FBATaff = FBAT using transmissions to affecteds only, FBATboth = FBAT using transmissions to both affecteds and unaffecteds. Results from all other LMM methods were indistinguishable from FLMM_E and so are not shown. MQLS and RT gave identical results with either 1972 or 3626 individuals, as phenotypes could only be simulated for the 1972 genotyped individuals.
(TIF)

Figure S6 Manhattan plots for the simulated strong binary (disease) phenotype using FaST-LMM exact and alternative software packages. The points marked in red denote the regions close to the simulated weak effect SNPs. FLMM_E = FaST-LMM using exact calculation, RT = ROADTRIPS, FBATaff = FBAT using transmissions to affecteds only, FBATboth = FBAT using transmissions to both affecteds and unaffecteds. Results from all other LMM methods were indistinguishable from FLMM_E and so are not shown. MQLS and RT gave identical results with either 1972 or 3626 individuals, as phenotypes could only be simulated for the 1972 genotyped individuals.
(TIF)

Figure S7 Comparison of −log10(p-values) using different LMM software packages, real disease phenotypes. Plots above the diagonal show a comparison of −log10(p-values), with correlations between the -log10(p-values) indicated below the diagonal. The grey solid lines represents the line of equality; the black dashed lines the linear regression line of the variable on the y axis on the variable on the x axis. EM_BN = EMMAX (Balding-Nichols), EM_IBS = EMMAX (IBS method), FLMM_A = FaST-LMM using approximate calculation, FLMM_E = FaST-LMM using exact calculation, GA_FA = GenABEL (FASTA), GA_GRG = GenABEL (GRAMMAR-Gamma), GMA_C = GEMMA using centred genotypes, GMA_S = GEMMA using standardised genotypes, MMM_E = MMM using full mixed model (exact) calculation, MMM_G = MMM using GLS approximation, Unadj = unadjusted analysis.
(TIF)

Figure S8 Comparison of −log(p-values) using LMM and alternative software packages, real disease phenotypes. Plots above the diagonal show a comparison of −log10(p-values), with correlations between the −log10(p-values) indicated below the diagonal. The grey solid lines represent the line of equality; the

black dashed lines the linear regression line of the variable on the y axis on the variable on the x axis. FLMM_E = FaST-LMM using exact calculation, MQLS1972 = MQLS using 1972 genotyped individuals, MQLS3626 = MQLS using all 3626 individuals with or without genotype data, RT1972 = ROADTRIPS using 1972 genotyped individuals, RT3626 = ROADTRIPS using all 3626 individuals with or without genotype data, FBATaff = FBAT using transmissions to affecteds only, FBATboth = FBAT using transmissions to both affecteds and unaffecteds, MQLS_E = MQLS using estimated (rather than theoretical) kinships.
(TIF)

Figure S9 Comparison of −log(p-values) using LMM and alternative software packages, simulated weak binary (disease) phenotype. Plots above the diagonal show a comparison of –log10(p-values), with correlations between the –log10(p-values) indicated below the diagonal. The grey solid lines represent the line of equality; the black dashed lines the linear regression line of the variable on the y axis on the variable on the x axis. The colours denote: red = the two weak effect SNPs, magenta = SNPs within 500 kb of the weak effect SNPs, blue = 22 polygenic SNPs, green = SNPs within 500 kb of the polygenic SNPs, black = all other SNPs. Because the black/green/blue SNPs were plotted before the magenta/red SNPs, they may be obscured by the latter. FLMM_E = FaST-LMM using exact calculation, MQLS = MQLS using 1972 or 3626 individuals, RT = ROADTRIPS using 1972 or 3626 individuals, FBATaff = FBAT using transmissions to affecteds only, FBATboth = FBAT using transmissions to both affecteds and unaffecteds. MQLS and RT gave identical results with either 1972 or 3626 individuals, as phenotypes could only be simulated for the 1972 genotyped individuals.
(TIF)

Figure S10 Comparison of −log10(p-values) obtained from FaST-LMM using alternative kinship estimates, real disease phenotypes. Plots above the diagonal show a comparison of –log10(p-values), with correlations between the –log10(p-values) indicated below the diagonal. The grey solid lines represents the line of equality; the black dashed lines the linear regression line of the variable on the y axis on the variable on the x axis. KING_H = KING homogeneous method, KING_R = KING robust method, Ped = theoretical kinship estimates based on pedigree information, FLMM_R = FaST-LMM's own realised relationship matrix, Unadj = unadjusted, Wrong = misspecified kinships, chosen to be inversely related to the true kinship value.
(TIF)

Figure S11 Power and type 1 error of different LMM methods applied to 462 Brazilian founders. Powers (left hand plots) are defined as the proportion of replicates (out of 1000) in which both simulated disease loci are detected, with 'detection' corresponding to any SNP within 40 kb of the simulated disease locus reaching the specified p-value threshold. Type 1 errors (right hand plots) are defined as the proportion of null SNPs (out of 20,000 = 20 null SNPs times 1000 simulation replicates) that reach the specified p-value threshold. Horizontal dashed lines indicate the target p-value thresholds (i.e. the expected type 1 error rates).
(TIF)

Table S1 Genomic control factors achieved in analysis of the real data, or a single replicate of the simulated data, when feeding externally estimated kinships into FaST-LMM.
(PDF)

Table S2 Computational speed and ease of use of various packages.
(PDF)

Table S3 Concordance between top SNPs identified by different LMM methods when using 462 founder individuals.
(PDF)

Text S1 Membership of Wellcome Trust Case Control Consortium 2.
(DOC)

Author Contributions

Conceived and designed the experiments: JMB HJC. Performed the experiments: JEa ENM MF HJC. Analyzed the data: JEa ENM MF HJC. Contributed reagents/materials/analysis tools: SMBJ JMB. Wrote the paper: JEa MF JMB HJC.

References

1. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, et al. (2010) Variance component model to account for sample structure in genome-wide association studies. Nat Genet 42: 348–354.
2. Zhang Z, Ersoz E, Lai CQ, Todhunter J R, Tiwari HK, et al. (2010) Mixed linear model approach adapted for genome-wide association studies. Nat Genet 42: 355–360.
3. Sawcer S, Hellenthal G, Pirinen M, Spencer CC, Patsopoulos NA, et al. (2011) Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature 476: 214–219.
4. Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, et al. (2011) FaST linear mixed models for genome-wide association studies. Nature Methods 8: 833–835.
5. Fisher R (1918) The correlation between relatives on the supposition of Mendelian inheritance. Trans R Soc Edin 52: 399–433.
6. Henderson CR (1953) Estimation of variance and covariance components. Biometrics 9: 226–252.
7. Boerwinkle E, Chakraborty R, Sing CF (1986) The use of measured genotype information in the analysis of quantitative phenotypes in man. I. Models and analytical methods. Ann Hum Genet 50: 181–94.
8. Abney M, Ober C, McPeek MS (2002) Quantitative-trait homozygosity and association mapping and empirical genomewide significance in large, complex pedigrees: fasting serum-insulin level in the Hutterites. Am J Hum Genet 70: 920–934.
9. Chen WM, Abecasis GR (2007) Family-based association tests for genomewide association scans. Am J Hum Genet 81: 913–926.
10. Aulchenko YS, de Koning DJ, Haley C (2007) Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. Genetics 177: 577–585.
11. Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, et al. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38: 203–208.
12. Amin N, van Duijn CM, Aulchenko YS (2007) A genomic background based method for association analysis in related individuals. PLoS One 2: e1274.
13. Fakiola M, Strange A, Cordell HJ, Miller EN, Pirinen M, et al. (2013) Common variants in the HLA-DRB1-HLA-DQA1 HLA class II region are associated with susceptibility to visceral leishmaniasis. Nat Genet 45: 208–213.
14. Svishcheva GR, Axenovich TI, Belonogova NM, van Duijn CM, Aulchenko YS (2012) Rapid variance components-based method for whole-genome association analysis. Nat Genet 44: 1166–1170.
15. Pirinen M, Donnelly P, Spencer C (2013) Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. Annals of Applied Statistics 7: 369–390.
16. Zhou X, Stephens M (2012) Genome-wide efficient mixed-model analysis for association studies. Nat Genet 44: 821–824.
17. Almasy L, Dyer TD, Peralta JM, Jun G, Wood AR, et al. (2014) Data for Genetic Analysis Workshop 18: Human whole genome sequence, blood pressure, and simulated phenotypes in extended pedigrees. Genet Epidemiol in press.
18. Eu-ahsunthornwattana J, Howey RAJ, Cordell HJ (2014) Accounting for relatedness in family-based association studies: application to GAW18 data. BMC Proceedings 8(Suppl 1):S79.
19. Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: The insulin gene region and insulin–dependent diabetes mellitus. Am J Hum Genet 52: 506–516.
20. Rabinowitz D, Laird NM (2000) A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. Hum Hered 50: 211–223.
21. Laird NM, Horvath S, Xu X (2000) Implementing a unified approach to family based tests of association. Genet Epidemiol Suppl 19: S36–S42.

22. Lake SL, Blacker DB, Laird NM (2000) Family-based tests of association in the presence of linkage. Am J Hum Genet 67: 1515–1525.
23. Horvath S, Xu X, Laird NM (2001) The family based association test method: strategies for studying general genotype–phenotype associations. Eur J Hum Genet 9: 301–306.
24. Thornton T, McPeek MS (2007) Case-control association testing with related individuals: a more powerful quasi-likelihood score test. Am J Hum Genet 81: 321–337.
25. Jakobsdottir J, McPeek MS (2013) MASTOR: Mixed-Model Association Mapping of Quantitative Traits in Samples with Related Individuals. Am J Hum Genet 92: 652–666.
26. Thornton T, McPeek MS (2010) ROADTRIPS: case-control association testing with partially or completely unknown population and pedigree structure. Am J Hum Genet 86: 172–184.
27. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559–575.
28. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, et al. (2010) Robust relationship inference in genome-wide association studies. Bioinformatics 26: 2867–2873.
29. Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55: 997–1004.
30. Listgarten J, Lippert C, Kadie CM, Davidson RI, Eskin E, et al. (2012) Improved linear mixed models for genome-wide association studies. Nature Methods 9: 525–526.
31. Lippert C, Quon G, Kang EY, Kadie CM, Listgarten J, et al. (2013) The benefits of selecting phenotype-specific variants for applications of mixed models in genomics. Sci Rep 3: 1815.
32. Ettinger NA, Duggal P, Braz RF, Nascimento ET, Beaty TH, et al. (2009) Genetic admixture in Brazilians exposed to infection with Leishmania chagasi. Ann Hum Genet 73: 304–313.
33. Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome Research 19: 1655–1664.
34. Furlotte NA, Eskin E, Eyheramendy S (2012) Genome-wide association mapping with longitudinal data. Genet Epidemiol 36: 463–471.
35. Lange K, Papp JC, Sinsheimer JS, Sripracha R, Zhou H, et al. (2013) Mendel: the Swiss army knife of genetic analysis programs. Bioinformatics 29: 1568–1570.
36. Astle W, Balding DJ (2009) Population structure and cryptic relatedness in genetic association studies. Statistical Science 24: 451–471.
37. Speed D, Hemani G, Johnson MR, J BD (2012) Improved heritability estimation from genome-wide SNPs. Am J Hum Genet 91: 1011–1021.
38. Wang K, Hu X, Peng Y (2013) An analytical comparison of the principal component method and the mixed effects model for association studies in the presence of cryptic relatedness and population stratification. Hum Hered 76: 1–9.
39. Aulchenko YS, Ripke S, Isaacs A, van Duijn CM (2007) GenABEL: an R library for genome-wide association analysis. Bioinformatics 23: 1294–1296.
40. Abecasis GR, Cherney SS, Cookson WO, Cardon LR (2002) Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30: 97–101.
41. Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, et al. (2008) Efficient control of population structure in model organism association mapping. Genetics 178: 1709–1723.
42. Marchini J, Howie B, Myers S, McVean G, Donnelly P (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 39: 906–913.
43. Martin ER, Monks SA, Warren LL, Kaplan NL (2000) A test for linkage and association in general pedigrees: the pedigree disequilibrium test. Am J Hum Genet 67: 147–154.
44. Lange C, DeMeo D, Silverman EK, Weiss ST, Laird NM (2004) PBAT: tools for family-based association studies. Am J Hum Genet 74: 367–369.
45. Dudbridge F (2008) Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data. Hum Hered 66: 87–98.
46. Dudbridge F, Holmans PA, Wilson SG (2011) A flexible model for association analysis in sibships with missing genotype data. Ann Hum Genet 75: 428–438.
47. Powell JE, Visscher P, Goddard ME (2010) Reconciling the analysis of IBD and IBS in complex trait studies. Nat Rev Genet 11: 800–805.
48. Yang J, Lee SH, Goddard ME, Visscher PM (2011) GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet 88: 76–82.