VARIANCE HETEROGENEITY IN GENETIC MAPPING

Robert Wallace Corty

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Curriculum of Bioinformatics and Computational Biology

Chapel Hill
2018

Approved by:

Fernando Pardo Manuel de Villena

James Evans

Yun Li

Lisa Tarantino

William Valdar

ABSTRACT

Robert Wallace Corty: Variance Heterogeneity in Genetic Mapping
(Under the direction of William Valdar)

Genetic mapping is a process by which researchers seek to identify genetic factors that influence a trait of interest. Such efforts typically focus on those that either increase or decrease the trait of interest, and assume that the variance of the trait is constant across all individuals. I develop and apply statistical methods that challenge that assumption in two ways. First, I consider the situation where non-genetic factors influence trait variance, which I term "background variance heterogeneity". Though they are not of immediate interest in a genetic mapping study, they can be exploited to align observations' weights with their precisions. Second, I consider the situation where genetic factors influence trait variance, which I term "foreground variance heterogeneity". Such factors are of immediate interest because they represent novel discoveries that could be missed by standard analyses.

I consider both foreground and background variance heterogeneity as they relate to linkage disequilibrium mapping in exchangeable mapping populations. I report three novel genetic factors with strong evidence that they influence medically-important traits in the mouse model system. Finally, I consider the background variance heterogeneity as it relates to association mapping in non-exchangeable populations. I report a mathematical advance that makes possible the fitting of a statistical model that accommodates background variance heterogeneity in non-exchangeable populations.

Happy families are all alike;
every unhappy family is unhappy in its own way.

— Leo Tolstoy, opening line of *Anna Karenina*,
on happiness-dependent variance heterogeneity

I'm convinced the tuxedo was invented by women... "Well, they're all the same; we might as well dress them all the same."

— Jerry Seinfeld,
on ignoring variation

ACKNOWLEDGEMENTS

I thank the Valdar lab for being a supportive and enriching venue in which to work and learn these past five years. I appreciate the efforts Will has made to create this lab environment by recruiting talented students such as Greg Keele, Paul Maurizio, Dan Oreper, Wes Crouse, Yanwei Cai, and Kathie Sun.

I thank the BCB program. They built it and we came. Tim Elston, Jonathan Cornett, and Cara Marlow were supportive in a thousand little ways that helped me stay focused on my academics.

I thank the MD-PhD program. The "big picture" guidance and support I've received from Gene Orringer, Toni Darville, Mohanish Deshmukh, Alison Regan, Carol Herion has kept me moving forward in a productive direction no matter how thick the morass of graduate school felt.

I thank my family. My mom, dad, and brother have been patient with me when I needed it and gave me a little kick sometimes when I needed that too.

I especially want to thank my grandparents. From childhood, my mom's parents shared with me a love of reading and writing. At every step in life, the value of these skills seems to compound. My dad's parents brought in another piece of the puzzle. With visits to the Franklin institute and a home experimenter kit, they kindled my enthusiasm for science. I'm thrilled to be where I am and I am deeply grateful to them for helping me get here.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

DGLM	double generalized linear model
EMMA	efficient mixed model analysis
FPR	false positive rate
FWER	family-wise error rate
GLS	generalized least squares
ISAM	inbred strain association mapping
LD	linkage disequilibrium
LMM	linear mixed model
LRT	likelihood ratio test
ML	maximum likelihood
mQTL	mean-controlling quantitative trait locus
mvQTL	mean or variance controlling quantitative trait locus
SLM	standard linear model
QTL	quantitative trait locus
vQTL	variance-controlling quantitative trait locus

CHAPTER 1
Introduction

1.1 Genetic Mapping

Genetic mapping is a scientific endeavor that has elucidated the genetic underpinnings of hundreds of conditions relevant to human health and disease, both directly in humans (MacArthur et al., 2017) and in model organisms (Grubb et al., 2014), as well as many commercially-important traits in crops and livestock. There are, broadly speaking, three approaches to genetic mapping, which I will discuss below. They are united in their goal and the general process by which they seek to achieve it.

All approaches to genetic mapping involve the collection of phenotype and genotype information on a population of organisms and a statistical analysis to test for associations between the phenotype and each measured, polymorphic locus of the genome. This endeavor is motivated by the belief that the vast majority of the genome, say greater than 99%, does not have any appreciable effect on the phenotype, so a successful genetic mapping experiment allows researchers interested in the phenotype to focus their efforts on the small section of the genome that does have an effect. Thus, a successful genetic mapping effort results in a partition of the genome into a large part with no appreciable effect on the trait of interest, and a small part believed with a high degree of certainty to influence the phenotype.

The three general approaches to genetic mapping are: 1. linkage analysis, 2. linkage disequilibrium mapping, and 3. association mapping. I'll briefly review these approaches and some examples of how each has been productively applied.

Linkage Analysis is most useful for traits where one or a few genetic factors are expected to exert a large effect (Elston and Stewart, 1971; Haseman and Elston, 1972). It is based on a large collection of families with a few individuals per family, where each family must have at least one affected and one unaffected individual and everyone has been genotyped across a sparse panel of markers.

Examples of successful applications of linkage analysis are Mendelian disease phenotypes like Duchenne muscular dystrophy, (Brown et al., 1985; Murray et al., 1982) cystic fibrosis (Tsui et al., 1985; Wainwright et al., 1985; White et al., 1985) and ataxia-telangiectasia (Gatti et al., 1988). In the last decade, as denser marker panels have become available and many of the high-prevalence, near-Mendelian traits have been mapped, linkage analysis has receded in prominence.

Linkage disequilibrium mapping involves an experimental population of organisms where the pattern of descent from a reference population can be inferred. For that reason, it is only possible in model organisms, livestock, and some crops — but never in humans. Strengths of this approach include the tight control of environmental exposures and the opportunity to deeply and invasively measure phenotypes. Two of the most classic designs, the F2 intercross and backcross, mimic the pedigrees of human linkage mapping, but rather than using many families with a few individuals per family, they create a single family with hundreds of siblings (Lynch and Walsh, 1998; Lander and Green, 1987; Lander and Botstein, 1989). Because these designs restrict the total genetic variation to only two parental haplotypes, rather than the vast number of haplotypes represented in a collection of human families, they are able to detect smaller effects and are not restricted to analyze mono- or oligo-genic traits in the way that linkage analysis is.

Modern efforts toward model organism LD mapping have made prominent use of more elaborate breeding designs, most prominently multi-parental outbred populations (Ghazalpour et al., 2012; Svenson et al., 2012) and multi-parental genetic reference populations (The Complex Trait Consortium, 2004; MacKay et al., 2012; King et al., 2012) as well as in commercially-important crops (McMullen et al., 2009; Bandillo et al., 2013). These designs use a larger number of parental haplotypes and therefore explore a larger space of possible genetic effects than classic designs like F2's and backcrosses. In this way, multi-parental populations designed for LD mapping strike a middle ground between classical designs and association analysis, which I discuss below.

Association mapping (GWAS) is based on a large population of individuals with no particular genetic relationship. Because no breeding is required, it can be conducted in human populations. It is most appropriate for traits where many genetic factors are thought to exert an effect, like body mass index (Speliotes et al., 2010; Locke et al., 2015) and height (Allen et al., 2010; Wood et al., 2014) and psychiatric conditions like schizophrenia (Ripke et al., 2014) and depression (of the PGC et al.,

2017). Each genetic locus is tested for association with the phenotype after a correction is made for global genetic similarity between individuals (Lippert et al., 2011; Zhou and Stephens, 2012).

Association mapping in model organisms uses a panel of inbred organisms with no particular genetic relationship to conduct a study similar to a human GWAS. This study design combines the strengths of model organism experiments (the tight control of environmental exposures and the ability to make invasive measurements) with the ability to observe replicates from each genome (Payseur and Place, 2007; Kang et al., 2008; Kirby et al., 2010). One important strength of a study design that allows multiple observations of the same genotype is that it allows for very precise measurement of the average phenotype that results from a given genotype because that genotype can be observed arbitrarily-many times. Additionally, it allows for direct quantification of environmental variance, which is confounded with genetic variance in any population without genetically-identical individuals (Falconer, 1965; Lynch and Walsh, 1998).

Across all these approaches to genetic mapping, the goal remains the same — to identify genetic loci where allelic variation correlates with phenotype variation.

1.2 Variation and Variance

We can say that we have observed "Variation" in some quantity when we have observed at least two different values for that quantity. Without phenotype variation, no analysis of any kind is possible, genetic or otherwise. Imagine a QTL mapping study where a tremendous amount of genotypic variation was measured, but, by chance, all individuals in the mapping population have the same phenotype value to measured precision. Realistically, the problem in such a study is that we did not measure the phenotype to sufficient precision — maybe the scale we used to measure mouse bodyweight was only accurate to the nearest pound, or maybe the phenotype is a molecular phenotype for which the state-of-the-art measurement procedure cannot differentiate between the highest and lowest values in our population. But more theoretically, given a set of observations without any variation, there can be no attempt to correlate it with variation in any other quantity, be they other phenotypes, environmental exposures, or genetic factors.

Analogously, for any genetic locus where all individuals in the mapping population have the same allele, no genetic mapping study can hope to identify an association. This statement is quite

different from a mechanistic assessment that determines the gene products of this locus are irrelevant to the phenotype of interest; no such assessment can be made. Genetic mapping is fundamentally a statistical, rather than mechanistic process, simply testing for correlations between phenotype variation and allelic variation.

The above discussion considered variation as a binary quantity; it's either present or absent. But there are a variety of measures that can be used to quantify variation. Some examples include the range, the interquartile range, the standard deviation, the mean absolute deviation, and the variance. This dissertation deals almost exclusively with the variance because it has the salutary property that the sum of the variance attributable to each individual factor in a regression analysis is equal to the variance of the response (the phenotype in genetic applications). Put simply the variance of a sample of numbers is the sum of the squared differences between each number and the mean. For a large sample of numbers, this quantity accurately estimates the variance of the random process by which the numbers were generated.

At times, this dissertation also considers the standard deviation, which is simply the square root of the variance. The standard deviation has the property that it is on the same scale as the phenotype itself, and is therefore straightforwardly interpretable.

1.3 Sources of Variance

It is important to recognize all potential sources of variance in a population in which a genetic mapping study will be conducted. Understanding genetic parameters such as broad sense and narrow sense heritability, the percentage of variance explained by aggregate additive, dominance, and epistatic effects yields valuable insights into the "genetic architecture" of the trait. Understanding the effect of sex, bodyweight, and nuisance covariates such as housing, diet, and experimenter can help scientists design more efficient experiments (Nettleton, 2006; Datta and Nettleton, 2014). I'll begin by reviewing sources of variance in measurements made on a single organism. As discussed previously, in the absence of any genetic variation, there can be no prospect for genetic insight. I continue with a review of sources of variance in measurements of multiple organisms, keeping in mind that the single-organism sources of variance are still present.

1.3.1 Measurements on a Single Organism

There are surprisingly many sources of variance when multiple measurements are made, even on a single organism.

When multiple measurements of a given trait are made on a single individual at the same time, the only source of variance is technical variance (Rönnegård and Valdar, 2011). An example of this type of measurement is the collection of a single blood sample from a mouse, which is split it into three aliquots and the mRNA content of each aliquot is analyzed independently (Marioni et al., 2008).

When the a phenotype is measured on one individual at multiple times, temporal fluctuation is a potential source of variance. This temporal fluctuation comes in two "flavors". First, the value of the phenotype of the individual may change over time. Second, the measurement device may change over time. Effects of this type are often called "batch effects". An example of this type of measurement is the weighing of each experimental mouse each day of a multi-day experiment (Gray et al., 2015).

When the same organism is observed in multiple different "macro-environments", that variation in macro-environment can contribute variance to the phenotype. The term "macro-environment" is used here to signify that the researcher has intentionally introduced an environmental effect. It is used in contrast to the "micro-environment", which is discussed below. The same individual could be exposed to multiple different macro-environments at different times in its life, in which case temporal variation would potentially be in play, or samples of the organism can be extracted and treated with different environmental factors at a single time point.

When multiple, theoretically-identical structures are measured on a single individual at a single time, "fluctuating asymmetry" is a potential source of variance (Palmer and Strobeck, 1986). An example of this type of experiment is be measurement of the left and right kidney weight of mice (Leamy et al., 2000, 2002). There are valid criticisms to be made about many specific measurements that are said to reflect fluctuating asymmetry. For example, in the case of the left and right kidney in a mouse, some difference in size might be expected due to the right kidney being crowded by the liver during development. But the general concept, that of assessing the extent to which multiple theoretically-identical phenotypes are expressed identically in a given organism, is an important

contribution to understanding the totality of sources of variance in a phenotype, for phenotypes where it is applicable.

A further source of variance in measurements of theoretically-identical structures from the same organism is developmental stochasticity. This term refers to the concept that micro-environmental perturbations during the developmental process can "fix" larger changes later in life, similar to a "butterfly effect" of developmental biology.

1.3.2 Measurements on Multiple Organisms

When multiple organisms are observed, additional layers of variance are possible, depending on the genetics of the organisms (Table 1.2).

The experimental design that most limits the variance amongst multiple organisms is when all the organisms are genetically identical. In the observation of multiple genetically-identical organisms, developmental stochasticity, is a potential source of variation (Fraser and Schadt, 2010). This same source of variance can also be referred to as "micro-environmental variance" (Hill and Mulder, 2010). This type of variance captures all the myriad, subtle exposures that each organism experiences, but which no researcher can hope to standardize. For example, the precise living temperature a mouse experiences depends slightly on where its cage is relative to the air vents, the amount of bedding depends on exactly how much the technician happened to grab when filling the cage, and uncountably many more such small effects could be imagined. Outside of experimental designs that make use of inbred organisms, it is impossible to directly estimate the contribution of micro-environmental variance to phenotype variance.

Consider a population of organisms that is not genetically identical in a global sense, but is genetically identical at one specific locus. A potential source of phenotype variance in such a population is interactions between the locus and factors in which the organisms do vary, such as other genetic loci and micro-environmental exposures. The fact that all the organisms have the same allele at the focal locus precludes any direct contribution from that locus to the phenotype variance. But, the locus may interact with polymorphisms elsewhere in the genome to make a contribution to the phenotype variance through GxG or may interact with micro-environmental factors to make a contribution through GxE (Falconer and Mackay, 1995; Struchalin et al., 2010; Rönnegård and Valdar, 2011).

sources of single organism variance
measurement error
organism fluctuation
device fluctuation
developmental stochasticity
macro-environment effects
fluctuating asymmetry

Table 1.1: Sources of variance that contribute to total phenotype variance in measurements of a single organism. Note that measurement error is present in all measurements. Organismal fluctuation and device fluctuation are both the result of taking measurements at different times, but they can be deconfounded with designs that cross individuals and devices.

	sources of variance			
	single organism variance	developmental stochasticity	locus-by-G and locus-by-E	marginal effects of all polymorphic loci
one organism	•			
genetically-identical organisms	•	•		
organisms with same allele at a locus	•	•	•	
genetically-distinct organisms	•	•	•	•

Table 1.2: Sources of variance that contribute to total phenotype variance in measurements of multiple organisms. The first column represents all the sources of variance that can be present in a measurement of a single organism (Table 1.1). Note the hierarchical nature of the sources of variance as we progress down the table from more-closely related individuals to less-closely related individuals; new sources of variance are added, but never removed.

Consider next a population of organisms where there are multiple alleles present at the focal locus — one could imagine the same population as the above paragraph but simply focus on a different locus. Here, all the same effects described above could be present, and additionally a marginal effect of the locus could contribute to phenotype variance. In fact, this is the reasoning that underlies the vast majority of QTL mapping efforts. Any genetic locus where researchers conclude with high statistical certainty that the proportion of phenotype variance explained by the locus is not zero constitutes a QTL (Broman and Sen, 2009; Broman, 2010).

Having considered thoroughly many possible sources of variance, I turn next to the concept that not all organisms are influenced by them to equal extent.

1.4 Variance Heterogeneity

Conceptually, any of the sources of variance described above could contribute more or less to any one measurement and any factor could determine how much a given source of variance contributes. This situation is termed "variance heterogeneity". The measurements could all end up with the same variance, and it's simply partitioned differently according to sources. Or they could end up with different total variance.

Despite this reality, most genetics studies impose strong assumptions about the nature of phenotype variation. They use statistical models that assume that each measurement has an equal quantity of variance from each source. To put a concrete example to that statement, the statistical analysis most commonly used in LD mapping of an F2 intercross or backcross assumes that the residual variance is constant across all individuals. In this study design, the "residual variance" is the sum of all within-individual sources of variance described above, the genomic variance arising from genetic factors other than the locus currently being tested (which I term the "focal locus"), locus-by-genome interactions, and locus-by-micro-environment interactions. So the assumption is equivalent to the belief that, across all organisms in the mapping population, that sum is equal.

Other study designs that are often analyzed with the assumption of homogeneous variance include pedigree analysis to determine breeding values and heritability, inbred strain association mapping, and human GWAS.

Despite the pervasiveness of these "constant variance" assumptions, there is ample evidence that these sources of variance do not affect each measurement equally. Rather this assumption arises out of analytic convenience. Statistical models that make use of the homogeneous variance assumption have been more straightforward to develop and tend to have faster performance than more complex models that allow for variance heterogeneity. This situation formed the central tension of my dissertation work. **It was my belief that the analysis of many genetic study designs could be improved by using statistical models that recognize, and in some ways even capitalize on, variance heterogeneity.**

1.4.1 Evidence of Variance Heterogeneity

Why might measurements from one organism have more variance than measurements from another organism? If one device is used for one organism and another device is used for another, heterogeneity of measurement error could result in heterogeneity of phenotype variance. Observations of this type have been made in the field of human blood pressure management (Labarthe et al., 1973; Ataman et al., 1996; O'Brien, 2001).

As another example, genetic factors could influence phenotype variance by influencing sensitivity to variation in the micro-environment or influencing the extent of GxG or GxE variance. Family-based designs cannot disentangle these two sources of variance, but they can document their presence. For example, genomic effects on phenotype variance have been documented in cattle (Visscher and Hill, 1992; Mulder et al., 2008; Fasoula, 2012), dairy cow (Clay et al., 1979), pigs (Ibáñez-Escriche et al., 2008), chickens (Rowe et al., 2006), snails (Ros et al., 2004).

Other studies have documented a heterogeneity of phenotype variance across inbred strains, which can only be caused by differences in micro-environmental variation. Theses studies have documented this phenomenon in *Drosophila melanogaster* (Mackay and Lyman, 2005; Mackay, 2014), *Arabidopsis thaliana* (Hall et al., 2007), and crops (Walsh, 2017).

Early theoretical work focused on the notion that organisms with one allele at a locus might all have a similar phenotype, while organisms with the other allele might have very different phenotypes, despite tremendous variation in the rest of the genome in both groups and termed this phenomenon "canalization" (Waddington, 1942, 1959). This work has been extended to include a population genetic theory of how it could come about (Wagner et al., 1997; Gibson and Wagner, 2000; Meiklejohn

and Hartl, 2002). It has been related to the concept of "modularity" (Wagner et al., 2007), of "developmental constraint" (Pavličev and Cheverud, 2015). The original concept is now referred to as "robustness" (Kitano, 2004; Felix and Barkoulas, 2015; Yadav et al., 2015; Fraser and Schadt, 2010) as well as "capacitance" (Pettersson and Carlborg, 2015; Queitsch et al., 2002). Usefully, the concept of robustness is divided into environmental robustness and genomic robustness (Fraser and Schadt, 2010), where the former refers to heterogeneity of locus-by-E variance and the latter to heterogeneity of locus-by-G variance.

1.5 QTL Mapping in the Presence of Variance Heterogeneity

Given the limited focus of the genetics community on variance heterogeneity, despite its seeming ubiquity, I sought to assess the ways in which it could damage QTL mapping efforts, through false positive or false negative results, and whether there were ways in which the presence of variance heterogeneity could actually *strengthen* genetic mapping efforts.

Quantitative trait locus (QTL) mapping in both model organisms and humans has traditionally focused on finding regions of the genome whose allelic variation influences the phenotypic mean. In the past decade, a number of studies and proposed methods have broadened the scope of QTL mapping to consider effects on the phenotypic variance (Paré et al., 2010; Rönnegård and Valdar, 2011; Hulse and Cai, 2013). These studies and their findings have raised interesting questions and possibilities about underlying biology, evolutionary trajectory, and potential utility in agriculture (Wagner et al., 1997; Dworkin, 2005; Mulder et al., 2015). Nonetheless, consideration of variance effects — whether as the target of inference or as a feature of the data to be accommodated — has thus far remained outside of routine genetic analysis. This may be in part because QTL effects on the variance are sometimes considered of esoteric secondary interest, intrinsically controversial in their interpretation due to considerations of data scale (Sun et al., 2013; Shen and Ronnegard, 2013), or a priori too hard to detect (Visscher and Posthuma, 2010). But it is also likely to be in part because familiar software and procedures are currently lacking, and because the advantages of modeling heterogeneous variance, even when targeting QTL effects on the phenotypic mean, remain under-appreciated and largely undemonstrated.

The predominant approach to QTL mapping in model organisms, the focus here, considers each genetic locus in turn, using a standard linear model (SLM) to regress the phenotypes of the mapping population on their genotypes or their inferred genotype probabilities (Lander and Botstein, 1989; Haley and Knott, 1992). This SLM-based approach is primarily able to detect genomic regions containing a subset of genetic factors of interest — those that drive heterogeneity of phenotype mean. Despite this limited scope, however, its use is widespread due to its ease of use, the straightforward interpretation of its detected QTL, its historical importance in the fields of agricultural and livestock genetics, and the fact that many genetic factors truly do influence the expected value of phenotypes. Indeed, SLM-based interval mapping has yielded important insights on commercially- and medically-important traits across many organisms for many years.

The goal of QTL mapping, however, is much broader — to identify genetic factors that influence the phenotype in any way. For example, a genetic factor that influences the sensitivity of the phenotype to micro-environmental variation through a collection of what might be called a locus-by-E interactions is of interest whether the identity of the micro-environmental factors is known or not, but unless it also affects the mean it is undetectable by the SLM. Similarly, a genetic factor that influences the phenotype through many epistatic interactions (a collection of locus-by-G effects), but has an average effect near zero is unlikely to be detected by the SLM. Neither of these important goals, however, was the original motivation for seeking to detect genetic loci that influence phenotype variance. The original motivation was to lower the dimensionality of the search space for large locus-by-locus interactions (Paré et al., 2010), by searching first for variants that influence the variance and then searching for interaction effects between those loci and the rest of the genome. Such QTL that influence phenotype variance are often termed "vQTL". The goal of detecting vQTL and other more exotic types of QTL effects motivated the development and application of statistical tests that can detect genetic effects on other aspects of the phenotypic distribution, most notably the phenotype variance.

A number of statistical models and methods have been developed or adapted to identify associations between genotype and phenotypic variance. These include: Levene's test (Struchalin et al., 2010), the Fligner-Killeen test (Fraser and Schadt, 2010), Bartlett's test (Freund et al., 2013), the double generalized linear model (DGLM) and similar (Rönnegård and Valdar, 2011; Cao et al., 2014), and a host of two-step procedures that involve computing measure of variance for each individual and

testing that quantity for relation to the tested locus (Brown et al., 2014; Ayroles et al., 2015; Forsberg et al., 2015). Tests have also been developed to detect genotype associations with arbitrary functions of the phenotype, for example higher moments. These include a variant of the Komolgorov-Smirnov test (Aschard et al., 2013) and a semi-parametric exponential tilt model (Hong et al., 2016). The additional flexibility of these latter models makes them promising — a genetic factor that influences, e.g., the kurtosis of a phenotype should be of interest — but at present neither can accommodate covariates and the flexibility that affords them the ability to detect higher order effects brings with it a decreased power to detect mean and variance effects.

Efforts to identify vQTL have gathered steam in recent years. A few dozen vQTL have been reported, spanning *Arabidopsis thaliana* (Jimenez-Gomez et al., 2011; Shen et al., 2012; Forsberg et al., 2014), flowers (Lee et al., 2014), dairy cows (Fikse et al., 2012), *Drosophila melanogaster* (Ayroles et al., 2015; Huang et al., 2015), layer chickens (Wolc et al., 2012), maize (Ordas et al., 2008), mouse (Gray et al., 2015), and yeast (Nelson et al., 2013; Ziv et al., 2017; Forsberg et al., 2017). In at least two cases, researchers have identified vQTL and then gone on to identify specific interactions that caused the appearance of that vQTL (Huang et al., 2015; Brown, 2017), one of the original stated goals of vQTL analysis.

The existence of a vQTL, or indeed any factor affecting the variance has implications regarding statistical genetic analyses, both those targeting variance effects and those targeted mean-affecting QTL (hereafter, "mQTL"), and these implications have been relatively unexamined.

In particular, if a genetic (or other) factor influences phenotype variance then it follows that examination and testing of any other QTL effect — for example, that of a QTL elsewhere in the genome — must occur against a backdrop of systematically heterogeneous residual variance. The presence of this "background variance heterogeneity" (BVH) when testing for a (foreground) effect simultaneously presents analytic challenges and opportunities, not only for mapping vQTL but also the validity of studies detecting mQTL.

The impact of BVH on mapping mQTL can be thought of as a disruption of the natural observation weights: The SLM assumes the phenotype of every individual is subject to equal noise variance and therefore equal weight; but if it is known that some individuals' phenotypes are inherently less noisy — due to BVH induced by either a vQTL or other factors such as sex, housing, strain or experimenter — then those data should be upweighted, and this would lead to a more powerful test

for mQTL detection. Conversely, giving equal weight to subgroups of the data that are inherently noisier than average has the potential to leave outliers with overmuch influence on the regression, increasing the potential for false positive mQTL detections A case in point is when an mQTL also has variance effects: here the effects on the variance are a type of proximal BVH, and modeling them explicitly improves ability to detect effects on the mean, as in chapter 3. Knowledge and appropriate modeling of variance heterogeneity therefore has important implications for making mean-controlling QTL studies sensitive, robust and reproducible.

The impact of BVH on detection of foreground vQTL is more subtle. Parametric methods to identify vQTL typically pit heterogeneous variance alternative models against a homoskedastic, normally distributed null. However, under BVH the null model is not homoskedastic — it is a scale mixture — and this risks the null being rejected too readily. BVH could therefore lead to an inflated vQTL false positive rate.

If BVH is disruptive to QTL mapping generally, it is potentially valuable to incorporate it into the QTL mapping model when its source is known, and to use robustifying techniques to protect against it when its source is unknown. Accommodating BVH of known source is most naturally achieved through modeling covariate effects on the variance, something that is straightforward with the DGLM of Rönnegård and Valdar (2011) but not currently with other proposed methods. Protecting BVH when its source is unknown is less obvious, but since the threat manifests through sensitivity to distributional assumptions, natural contenders include side-stepping such assumptions via non-parametric approaches, *e.g.*, permutation testing, or reshaping the distribution prior to analysis through variable transformation. Both have been considered in the vQTL context, with permutation used in Hulse and Cai (2013) and Yang et al. (2012) and transformation in Rönnegård and Valdar (2011), Yang et al. (2012), Sun et al. (2013), and Shen and Carlborg (2013), but not specifically for controlling vQTL false positives in the presence of BVH.

CHAPTER 2

Mean-Variance QTL Mapping on a Background of Variance Heterogeneity [1]

2.1 Introduction

Here we examine the effect of modeled and unmodeled BVH on power and false positive rate when mapping QTL affecting the mean, the variance or both. In doing so we:

1. Develop a robust, straightforward procedure and software based on the DGLM that can be used for routine mQTL and vQTL analysis;
2. Compare alternative proposed methods for mQTL and vQTL analysis;
3. Show how incorporating BVH can improve power for detecting mQTL and vQTL;
4. Show how sensitivity to model assumptions can be rescued by variable transformation and/or permutation.
5. Illustrate the effect of modeling BVH in existing dataset, an F2 cross from Leamy et al, and discover a new QTL for bodyweight.

2.2 Statistical Methods

This section reviews four approaches for modeling the effect of a single QTL on the phenotypic mean and/or variance: the standard linear model, Levene's test, Cao's tests, and our preferred procedure based on the DGLM. For each approach we describe a set of alternative procedures for evaluating significance (*i.e.*, calculating p-values) that provide varying degrees of protection against the impact of BVH and distributional assumptions more generally. The following section, Data and Simulations, then describes a simulation study that assesses the approaches and p-value procedures, and a dataset to which they are applied genomewide.

[1] This chapter has been adapted from a manuscript submitted to G3. The citation will be: Corty, RW. and Valdar, W., 2018 Mean-Variance QTL Mapping on a Background of Variance Heterogeneity. G3: Genes, Genomes, Genetics.

2.2.1 Definitions

We start by defining three partially overlapping classes of QTL:

mQTL: a locus containing a genetic factor that causes heterogeneity of phenotype mean,

vQTL: a locus containing a genetic factor that causes heterogeneity of phenotype variance, and

mvQTL: a locus containing a genetic factor that causes heterogeneity of either phenotype mean, variance, or both — a generalization that includes the other two classes.

In addition, since we restrict our attention to QTL mapping methods that test genetic association with a phenotype one locus at a time, we distinguish two sources of variance effects:

Foreground Variance Heterogeneity (FVH): effects on the variance that arise from the locus under consideration (the focal locus);

Background Variance Heterogeneity (BVH): effects on the variance that arise from outside of the focal locus, *e.g.*, from another locus or an experimental covariate.

2.2.2 Procedures to evaluate the significance of a single test

In comparing different statistical approaches and their sensitivity to BVH, namely the effect of BVH on power and false positive rate (FPR), it is important to acknowledge that various measures could be taken to make significance testing procedures more robust to model misspecification in general and to BVH specifically. The significance testing methods considered here are frequentist, involving the calculation of a test statistic T on the observed data followed by an estimation of statistical significance based on a conception of T's distribution under the null. However, BVH constitutes a departure of distributional assumptions, and in any rigorous applied statistical analysis when departures are expected it would be typical to consider protective measures such as, for example, transforming the response to make asymptotic assumptions more reasonable, or the use of computationally intensive procedures, such as those based on bootstrapping or permutation, to evaluate significance empirically.

Nominal significance (*i.e.*, the p-value for a single hypothesis test) is evaluated using four distinct procedures. The first two rely on asymptotics:

1. Standard: The test statistic T is computed on the observed data and compared with its asymptotic distribution under the null.

2. Rank-based inverse normal transform (RINT): As for standard, except observed phenotypes $\{y_i\}_{i=1}^n$ are first transformed to strict normality using the function $\text{RINT}(y_i) = \Phi^{-1}[(\text{rank}(y_i) - 3/8)/(n + 1/4)]$, where Φ is the normal c.d.f. and $\text{rank}(y_i)$ is gives the rank (from $1, \ldots, n$) (Beasley et al., 2009).

The second two determine significance empirically based on randomization: the test statistic T is recomputed as $T^{(r)}$ under randomizations of the data $r = 1, \ldots, R$, and the resulting set of statistics $\{T^{(r)}\}_{r=1}^R$ is used as the empirical distribution of T under the randomized null. Two alternative randomizations are considered:

3. Residperm: we generate a pseudo-null response $\{y_i^{(r)}\}_{i=1}^n$ based on permuting the residuals of the fitted null model, (Freedman and Lane, 1983; Good, 2013), a process recently applied in the field of QTL mapping by Cao et al. (2014).

4. Locusperm: we leave the response intact, instead permuting the rows of the design matrix (or matrices) that differentiate(s) the null from alternative model.

2.2.3 Procedure to evaluate genomewide significance

In the context of a genome scan, where many hypotheses are tested, we aim to control the genomewide FPR, namely the family-wise error rate (FWER), the probability of making at least one false positive finding across the whole genome. This is done following the general approach of Churchill and Doerge (1994), which is closely related to the locusperm procedure described above, and which we refer to as genomeperm. Briefly, we perform an initial genome scan, recording test statistics $\{T_l\}_{l=1}^L$ for all L loci. Then for each randomization $r = 1, \ldots, R$, and for only the parts of the model that distinguish the null from the alternative model, the genomes are permuted among the individuals; the scan is then repeated to yield simulated null test statistics $\{T_l^{(r)}\}_{l=1}^L$ of which the maximum, $T_{\text{max}}^{(r)}$, is recorded. The collection of $\{T_{\text{max}}^{(r)}\}_{r=1}^R$ from all R such permutations is then used to fit a generalized extreme value distribution (GEV) (Dudbridge and Koeleman, 2004), and the quantiles of this are used to estimate FWER-adjusted p-values for each $\{T_l\}_{l=1}^L$.

2.2.4 Standard linear model (SLM) for detecting mQTL

The standard model of quantitative trait mapping uses a linear regression based on the approximation of Haley and Knott (1992) and Martínez and Curnow (1992) to interval mapping of Lander and Botstein (1989). The effect of a given QTL on quantitative phenotype y_i of individual $i = 1, \ldots, n$ is modeled as

$$y_i \sim \mathrm{N}(m_i, \sigma^2) \tag{2.1}$$

where σ^2 is the residual variance and m_i is a linear predictor for the mean, defined, in what we term the "full model", as

$$\text{Full model:} \quad m_i = \mu + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{q}_i^{\mathrm{T}}\boldsymbol{\alpha}\,, \tag{2.2}$$

where μ is the intercept, $\boldsymbol{x}_i$ is a vector of covariates with effects $\boldsymbol{\beta}$, and $\boldsymbol{q}_i$ is a vector encoding the genetic state at the putative mQTL with corresponding mQTL effects $\boldsymbol{\alpha}$. In the case considered here of biallelic loci arising from a cross of two founders, A and B, the genetic state vector $\boldsymbol{q}_i = (a_i, d_i)^{\mathrm{T}}$ is defined as follows: when genotype is known, for genotypes $(\mathrm{AA}, \mathrm{AB}, \mathrm{BB})$, the additive dosage is $a_i = (0, 1, 2)$ and the dominance predictor is $d_i = (0, 1, 0)$; when genotype is available only as estimated probabilities $p(\mathrm{AA})$, $p(\mathrm{AB})$ and $p(\mathrm{BB})$, following (Haley and Knott, 1992; Martínez and Curnow, 1992), we use the corresponding expectations, $a_i = 2p(\mathrm{AA}) + p(\mathrm{AB})$ and $d_i = p(\mathrm{AB})$.

The test statistic for an mQTL is based on comparing the fit of the full model, acting as an alternative model, with that of a null that omits the locus effect, namely,

$$\text{Null model:} \quad m_i = \mu + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}\,. \tag{2.3}$$

Since the regression in each case provides a maximum likelihood fit, the test statistic used here is likelihood ratio (LR) statistic, $T = 2(\ell_1 - \ell_0)$, where ℓ_1 and ℓ_0 are the log-likelihoods under the alternative and the null respectively. For the biallelic model, the asymptotic test is the likelihood ratio test (LRT) whereby under the null, $T \sim \chi^2_2$. (Note: Alternative evaluation using the F-test is in general more precise but for our purposes provides equivalent results.)

The residperm approach to empirical significance evaluation of T proceeds as follows. We first fit the null model (Equation 2.3) to obtain predicted values $\widehat{m}_i = \boldsymbol{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}$ and estimated residuals $\widehat{\varepsilon}_i$ such that $y_i = \widehat{m}_i + \widehat{\varepsilon}_i$. Then, for each randomization $r = 1, \ldots, R$, we generate pseudo-null phenotypes $\{y_i^{(r)}\}_{i=1}^n$ as

$$y_i^{(r)} = \widehat{m}_i + \widehat{\varepsilon}_{\pi_r(i)}\,,$$

where if $\boldsymbol{\pi}_r$ is a vector containing a random permutation of the indices $i = 1, \ldots, n$, then $\pi_r(i)$ is its ith element, mapping index i to its rth permuted version. The null and alternative models are then fitted to $\{y_i^{(r)}\}_{i=1}^n$ to yield $\ell_1^{(r)}$ and $\ell_0^{(r)}$, and hence $T^{(r)}$.

In the locusperm approach to empirical significance, the response is unchanged but permutations are applied to the locus genotypes. For each randomization r, the full model m_i is

$$\text{Permuted full model:} \quad m_i = \mu + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{q}_{\pi_r(i)}^{\mathrm{T}}\boldsymbol{\alpha} \tag{2.4}$$

where $\pi_r(i)$ is as defined for residperm above. This full model fit yields $\ell_1^{(r)}$, and then $T^{(r)} = 2(\ell_1^{(r)} - \ell_0)$. Note that $\ell_0^{(r)}$ need not be recomputed after randomization because because only the rows of the design matrices that are unique to the alternative model are permuted and thus $\ell_0^{(r)} = \ell_0$. Genomeperm applies locusperm genomewide: specifically, in each randomization $r = 1, \ldots, R$, the same permutation, $\boldsymbol{\pi}_r$, is applied to all L loci.

2.2.5 Levene's Test (LV) for detecting vQTL

Levene's test is a procedure for differences in variance between groups that can be used to detect vQTL. Suppose individuals are in G mutually exclusive groups $g = 1, \ldots, G$. Let $g[i]$ denote the group to which individual i belongs, denote gth group size as $n_g = \sum_{i=1}^n I_{\{g[i]=g\}}$, and gth group mean as $\bar{y}_g = n_g^{-1}\sum_{i=1}^n y_i I_{\{g[i]=g\}}$. Then denote the ith absolute deviation as $z_i = |y_i - \bar{y}_{g[i]}|$, the group mean of these as $\bar{z}_g = n_g^{-1}\sum_{i=1}^n z_i I_{\{g[i]=g\}}$ and overall mean $\bar{z} = n^{-1}\sum_{i=1}^n z_i$. Levene's W statistic is then

$$W = \frac{\sum_{g=1}^G n_g\,(\bar{z}_g - \bar{z})^2}{(G-1)} \left[\frac{\sum_{i=1}^n (z_i - \bar{z}_{g[i]})^2}{(n-G)}\right]^{-1}, \tag{2.5}$$

which under the null model of no variance effect follows the F distribution as $W \sim F(N-G, G-1)$ (Levene, 1960). Note that replacing means of y with medians gives the related Brown-Forsythe

test (Brown and Forsythe, 1973), and replacing all instances of z with y in Equation 2.5 gives the ANOVA F statistic.

Levene's test does not lend itself naturally to the residperm approach because it does not explicitly involve a null model to split the data into hat values and residuals. We therefore use the null model from the SLM (Equation 2.3) to approximate the residperm procedure with Levene's test. To execute the locusperm procedure, for each randomization r, the group labels are permuted among the individuals, which is equivalent to replacing all instances of $g[i]$ above with $g[\pi_r(i)]$, with $\pi_r(i)$ defined as above. A corresponding genomewide procedure, although not performed here, would ensure that each randomization r applies the same permutation $\boldsymbol{\pi}_r$ across all loci.

2.2.6 Cao's Tests

Cao et al. (2014) elaborates the SLM to have a variance parameter that differs by genotype, *i.e.*,

$$y_i \sim \mathrm{N}(m_i,\, \sigma_i^2), \tag{2.6}$$

where m_i is the linear predictor, σ_i^2 is the variance of the ith individual. These are defined in what we term the "full model" as

$$\text{Full model:} \quad \begin{cases} m_i &= \mu + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{q}_i^{\mathrm{T}}\boldsymbol{\alpha} \\ \sigma_i^2 &= \phi_{g[i]} \end{cases}, \tag{2.7}$$

where $g[i]$ indexes the genotype group to which i belongs, and $\{\phi_g\}_{g=1}^{G}$ are the variances of the $g = 1, \ldots, G$ genotype groups. Thus an individual's variance is entirely dictated by its genotype, and that genotype must be categorically known (or otherwise assigned). Cao et al. (2014) fits this model using a two-step, profile likelihood method, which in our applications we observe to be indistinguishable from full maximum likelihood (Figure 2.13).

Cao et al. (2014) describes tests for mQTL, vQTL and mvQTL based on comparing a full model against three different null models; we detail these tests below in our notation, denoting them respectively $\mathrm{Cao}_{\mathrm{M}}$, $\mathrm{Cao}_{\mathrm{V}}$, and $\mathrm{Cao}_{\mathrm{MV}}$.

2.2.6.1 Cao$_{\mathrm{M}}$ test for detection of mQTL

The Cao$_{\mathrm{M}}$ test involves an LRT between Cao's full model and Cao's no-mQTL model:

$$\text{Cao's no-mQTL model:} \quad \begin{cases} m_i &= \mu + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} \\ \sigma_i^2 &= \phi_{g[i]} \end{cases}, \tag{2.8}$$

To execute the residperm procedure for Cao$_{\mathrm{M}}$, pseudo-null phenotypes are generated using $\widehat{m}_i$ and $\widehat{\varepsilon}_i$ from Cao's no-mQTL model (Equation 2.8). The locusperm procedure respecifies the full model (Equation 2.7), leaving the variance model unchanged and specifying the mean predictor as $m_i = \mu + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{q}_{\pi_r(i)}^{\mathrm{T}}\boldsymbol{\alpha}$. The genomeperm procedure similarly applies the locusperm procedure genomewide, ensuring each randomization r applies the same permutation $\boldsymbol{\pi}_r$ to the mean specification across all loci.

2.2.6.2 Cao$_{\mathrm{V}}$ for detection of vQTL

The Cao$_{\mathrm{V}}$ test involves an LRT between Cao's full model and Cao's no-vQTL model:

$$\text{Cao's no-vQTL model:} \quad \begin{cases} m_i &= \mu + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{q}_i^{\mathrm{T}}\boldsymbol{\alpha} \\ \sigma_i^2 &= \sigma^2 \end{cases}, \tag{2.9}$$

where the unsubscripted σ^2 is a single, overall residual variance. This null model is identical to the alternative model in the SLM (Equation 2.2).

To execute the residperm procedure for Cao$_{\mathrm{V}}$, pseudo-null phenotypes are generating using $\widehat{m}_i$ and $\widehat{\varepsilon}_i$ from Cao's no-mQTL model (Equation 2.9). The locusperm procedure respecifies the full model (Equation 2.7), leaving the mean sub-model unchanged and specifying the variance predictor as $\sigma_i^2 = \phi_{g[\pi(i)]}$. The genomeperm procedure applies the locusperm procedure genomewide, ensuring each randomization r applies the same permutation $\boldsymbol{\pi}_r$ to the variance specification across all loci.

2.2.6.3 Cao$_{\text{MV}}$ for detection of generalized mvQTL

The Cao$_{\text{MV}}$ test involves an LRT between Cao's full model and Cao's no-QTL model:

$$\text{Cao's no-QTL model:} \quad \begin{cases} m_i &= \mu + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} \\ \sigma_i^2 &= \sigma^2 \end{cases} . \tag{2.10}$$

This null model is identical to the null model in the SLM (Equation 2.3).

To execute the residperm procedure for Cao$_{\text{MV}}$, pseudo-null phenotypes are generated using $\widehat{m}_i$ and $\widehat{\varepsilon}_i$ from Cao's no-QTL model (Equation 2.10). The locusperm procedure specifies the mean predictor as $m_i = \mu + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{q}_{\pi(i)}$ and the variance predictor as $\sigma^2_{g[i]} = \phi_{[\pi(i)]}$. The genomeperm procedure applies the locusperm procedure genomewide, ensuring each randomization r applies the same permutation $\boldsymbol{\pi}_r$ to the mean and variance specifications across all loci.

2.2.7 Double Generalized Linear Model (DGLM)

The DGLM models the phenotype y_i via two linear predictors as

$$y_i \sim \mathrm{N}(m_i,\, \sigma_i^2)\,, \quad \text{where} \quad \sigma_i^2 = \sigma^2 \times \exp(v_i)$$

where m_i predicts the phenotype mean and v_i predicts the extent to which the baseline residual variance σ^2 is increased in individual i. In what we term the "DGLM full model", these are specified as

$$\text{Full model:} \quad \begin{cases} m_i &= \mu + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{q}_i^{\mathrm{T}}\boldsymbol{\alpha} \\ v_i &= \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\gamma} + \mathbf{q}_i^{\mathrm{T}}\boldsymbol{\theta} \end{cases} , \tag{2.11}$$

where μ is the intercept, $\mathbf{z}_i$ is a vector of covariates (which may be identical to $\mathbf{x}_i$), $\boldsymbol{\gamma}$ is a vector of covariate effects on v_i, and $\boldsymbol{\theta}$ is a vector of locus effects on v_i.

As with Cao's full model, the DGLM full model can be compared, in a likelihood ratio test, with various null models to test for mQTL, vQTL (Rönnegård and Valdar, 2011), or mvQTL. A full maximum likelihood fitting procedure for the DGLM was provided by Smyth (1989).

2.2.7.1 DGLM$_{\mathrm{M}}$ for detecting mQTL:

For detecting mQTL, we use an LRT of the DGLM full model in Equation 3.1 against the no-mQTL model:

$$\text{No-mQTL model:} \quad \begin{cases} m_i &= \mu + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} \\ v_i &= \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\gamma} + \mathbf{q}_i^{\mathrm{T}}\boldsymbol{\theta} \end{cases}, \tag{2.12}$$

where the LR statistic has asymptotic distribution $T \sim \chi^2_2$.

To execute the residperm procedure for DGLM$_{\mathrm{M}}$, pseudo-null phenotypes are generated using $\widehat{m}_i$ and $\widehat{\varepsilon}_i$ from the Equation 2.12. The locusperm procedure respecifies the mean predictor as $m_i = \mu + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{q}_{\pi(i)}^{\mathrm{T}}\boldsymbol{\alpha}$ and does not modify the variance predictor. The genomeperm procedure similarly applies the locusperm procedure genomewide, ensuring each randomization r applies the same permutation $\boldsymbol{\pi}_r$ to the mean specification across all loci.

2.2.7.2 DGLM$_{\mathrm{V}}$ for detecting vQTL:

For detecting vQTL, we use an LRT of the DGLM full model in Equation 3.1 against the no-vQTL model:

$$\text{No-vQTL model:} \quad \begin{cases} m_i &= \mu + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{q}_i^{\mathrm{T}}\boldsymbol{\alpha} \\ v_i &= \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\gamma} \end{cases}, \tag{2.13}$$

where the LR statistic has asymptotic distribution $T \sim \chi^2_2$.

To execute the residperm procedure for DGLM$_{\mathrm{V}}$, pseudo-null phenotypes are generated using $\widehat{m}_i$ and $\widehat{\varepsilon}_i$ from the Equation 2.13. The locusperm procedure does not modify the variance predictor and respecifies the mean predictor as $v_i = \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\gamma} + \mathbf{q}_{\pi(i)}^{\mathrm{T}}\boldsymbol{\theta}$. The genomeperm procedure similarly applies the locusperm procedure genomewide, ensuring each randomization r applies the same permutation $\boldsymbol{\pi}_r$ to the variance specification across all loci.

Category	Test	Description
mQTL	SLM	Conventional test of mean differences; allows neither FVH nor BVH
mQTL	Cao_M	Allows FVH, but not BVH
mQTL	DGLM_M	Allows FVH and BVH
vQTL	Levene's test	Conventional test of variance differences; detects FVH, does not allow BVH
vQTL	Cao_V	Detects FVH, does not allow BVH
vQTL	DGLM_V	Detects FVH, allows BVH
mvQTL	Cao_MV	Detects FVH, does not allow BVH
mvQTL	DGLM_MV	Detects FVH and allows BVH

Table 2.1: The eight tests that were evaluated in the simulation studies. FVH: foreground variance heterogeneity. BVH: background variance heterogeneity.

2.2.7.3 DGLM_MV for detecting mvQTL:

For detecting mvQTL, we use an LRT of the DGLM full model in Equation 3.1 against the no-QTL model:

$$\text{No-QTL model:} \quad \begin{cases} m_i &= \mu + \mathbf{x}_i^\text{T}\boldsymbol{\beta} \\ v_i &= \mathbf{z}_i^\text{T}\boldsymbol{\gamma} \end{cases}, \tag{2.14}$$

where the LR statistic has asymptotic distribution $T \sim \chi^2_4$.

To execute the residperm procedure for DGLM_MV, pseudo-null phenotypes are generated using $\widehat{m}_i$ and $\widehat{\varepsilon}_i$ from the Equation 2.14. The locusperm procedure respecifies the mean predictor as $m_i = \mu + \mathbf{x}_i^\text{T}\boldsymbol{\beta} + \mathbf{q}_{\pi(i)}^\text{T}\boldsymbol{\alpha}$ and the variance predictor as $v_i = \mathbf{z}_i^\text{T}\boldsymbol{\gamma} + \mathbf{q}_{\pi(i)}^\text{T}\boldsymbol{\theta}$. The genomeperm procedure similarly applies the locusperm procedure genomewide, ensuring each randomization r applies the same permutation $\boldsymbol{\pi}_r$ to the mean and variance specifications across all loci.

2.3 Data and Simulations

Simulation was used to assess the ability of the eight tests described above to distinguish each of the three types of QTL — pure mQTL, pure vQTL, and mixed mvQTL — from a null locus in the presence and absence of background variance heterogeneity (BVH). Tests are distinguished by their ability to accommodate and target foreground variance heterogeneity (FVH) and background variance heterogeneity (Table 2.1).

In each simulation, $n = 300$ observations were simulated as

$$y_i \sim \mathrm{N}(q_i^{\mathrm{T}}\alpha, \exp(\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\gamma} + q_i^{\mathrm{T}}\theta))$$

where y_i is the phenotype of individual i, q_i is the genotype, and $\mathbf{z}_i$ is the indicator vector for the factor that drives BVH in scenarios where it is present. In each simulation, each row of $\mathbf{q}$ (each q_i) is drawn randomly from [-1, 0, 1] with probability (0.25, 0.5, 0.25) mimicking an F2 intercross. Across all simulations, $\boldsymbol{Z}$ is fixed to be an indicator matrix mapping the first 60 observations to group 1, the next 60 to group 2, ... and the last 60 to group 5.

Values of α, θ, and $\boldsymbol{\gamma}$ differentiate the eight simulation scenarios as described below.

2.3.1 Scenarios

Simulations varied in two dimensions: locus effect (four options) and BVH (two options) for a total of eight simulation scenarios. Each of the eight scenarios was examined in $S = 10,000$ simulation trials. Locus effect sizes were chosen such that all QTL were detectable with approximately 70% power at a 5% false positive rate for traditional tests in the absence of BVH. Comprehensive details on the simulation setup are described in **Supplementary Materials**.

The four options for a simulated locus effect were as follows:

1. null locus: The locus has no effect on phenotype.

2. pure mQTL: The locus has an additive effect on the phenotype mean that explains 5% of the total variance.

3. pure vQTL: The locus has an additive effect on the log standard deviation that is detectable with approximately 70% power at a 5% false positive rate for traditional tests in the absence of BVH.

4. mixed mvQTL: The locus has both an additive mean effect that explains 3.25% of phenotype variance and an additive variance effect that is approximately equally detectable.

The two options for simulated BVH were as follows:

1. absent: Nothing, except for possibly the locus (if a vQTL or mvQTL), influences the residual variance. ($\boldsymbol{\gamma} = [0, 0, 0, 0, 0]$)

2. present: $\boldsymbol{\gamma} = [-0.4, -0.2, 0, 0.2, 0.4]$, resulting in group-wise standard deviations in the null locus and mQTL scenarios of approximately $[0.67, 0.82, 1, 1.22, 1.49]$. In the vQTL and mvQTL scenarios, these BVH effects combine additively with the locus effects on the log standard deviation scale, yielding 15 distinct standard deviations. These effect sizes generate a spectrum of standard deviations across groups that are consistent with those observed in the real data reanalysis that follows.

2.3.2 Tests and Significance

In each scenario, eleven tests were applied, and four procedures were used to assess the statistical significance of each test, for a total of 32 test-procedures.

The eleven tests comprise four tests for detecting mQTL: LM, Cao_M, and DGLM_M with and without modeling the variance covariate; four for detecting vQTL: Levene's test, Cao_V, DGLM_V with and without modeling the variance covariate; and three for detecting mvQTL: Cao_MV and DGLM_MV with and without modeling the variance covariate, as described in **Methods**.

The eleven tests, however, contain some redundancy, and so in the main text we report results from only eight. Specifically, for a given type of QTL, the DGLM model that omits variance covariates is equivalent to the corresponding Cao's test, and, barring computational errors in fitting, should give equivalent results (as was observed); results from these DGLM models are therefore omitted from the main text, but for completeness are reported in the supplement.

The four procedures for evaluating the statistical significance were: standard, RINT, residperm, and locusperm, as described in the **Methods**.

2.3.3 Evaluation of tests and procedures

Tests and procedures for assessing statistical significance were evaluated based on their receiver operating characteristics (ROC) and their ability to accurately control the FPR to the nominal level. ROC curves display the ability of a test to discriminate between two conditions by plotting FPR against power for all possible cutoffs.

In this case, the ROC curve reflects the ability of a test to discriminate between QTL and null loci. Specifically, for a given method and cutoff c, the FPR was the fraction null simulations in which the nominal p-value p was less than c; the power is the fraction of times this happened in non-null (*i.e.*, QTL) simulations. A test was said to accurately control FPR when, for all c, FPR $= c$; it was said to "dominate" another test when it had higher power across all FPRs.

The ROC curve cannot immediately distinguish between tests that accurately control FPR and those that do not. We added a symbol to each ROC curve at the point where $c = 0.05$. In cases where the point falls on the vertical line at FPR = 0.05, it reflects accurate FPR control. In cases where the point falls to the left or right of the vertical line it reflects a conservative or anti-conservative test, respectively. QQ plots, provided in the supplementary material provide a more holistic view on the FPR control of each test and procedure.

2.3.4 Leamy *et al*. Summary of Original Study

Leamy et al. (2000) backcrossed mice from strain CAST/Ei, a small, lean strain, into mouse strain M16i, a large, obese strain. Nine F1 males were bred with 54 M16i females to produce a total of 421 offspring (208 female, 213 male), which were genotyped at 92 microsatellite markers across the 19 autosomes and phenotyped for body composition and morphometric traits. We retrieved all available data on this cross, which included marker genotypes, covariates, and eight phenotypes (body weight at five ages, liver weight, subcutaneous fat pad thickness, and gonadal fat pad thickness), from the Mouse Phenome Database (Grubb et al., 2014), and estimated genotype probabilities at 2cM intervals across the genome using the hidden Markov model in `R/qtl` (Broman et al., 2003).

This mapping population has been studied for association with several phenotypes: asymmetry of mandible geometry (Leamy et al., 2000), limb bone length (Leamy et al., 2002; Wolf et al., 2006), organ weight (Leamy et al., 2002; Wolf et al., 2006; Yi et al., 2006), fat pad thickness (Yi et al., 2005, 2006, 2007), and body weight (Yi et al., 2006). The most relevant prior study to this reanalysis, Yi et al. (2006), used standard methods to identify QTL for body weight at three weeks on chromosomes 1 and 18. However, we were not able to reproduce this result, despite following their analysis as described.

2.3.5 Availability of Data and Software

Analyses were conducted in the `R` statistical programming language (R Core Team, 2017). The simulation studies used the implementation of the standard linear model from package `stats`, Levene's test from `car`, Cao's tests as published in Cao et al. (2014) and the DGLM tests in package `dglm`. The reanalyzed dataset is available on the Mouse Phenome Database (Grubb et al., 2014) with persistent identifier MPD:206.

The entire project, including data and all analysis scripts, is available as a Zenodo repository at 10.5281/zenodo.1181887. There are six files stored in this repository. Files S1, S2, and S3 contain the R scripts necessary to replicate the simulation studies and their analysis, relying on the `plotROC` package to make ROC plots (Sachs and Others, 2017). File S4 contains the data from Leamy et al. (2000) that was reanalyzed. File S5 contains the attempted replication of the original analysis (Yi et al., 2006) and file S6 contains the new analysis, using package `vqtl`.

2.4 Results

2.4.1 Simulation study on single locus testing

Simulations were performed to examine the ability of the eight tests listed in Table 2.1 to detect nonzero effects belonging to their target QTL types (mQTL, vQTL, mvQTL), and to control the number of false positives when no such QTL effects were present. This was done both in the presence and absence of background variance heterogeneity, and for each test, with p-values calculated by each of the four alternative p-value generation procedures (standard, RINT, residperm, locusperm). The full combination of settings is listed in Table 2.2, which also lists results pertaining to a nominal FPR of 0.05, and described in more detail in **Data and Simulations**.

2.4.1.1 All three mQTL tests have equivalent performance in the absence of BVH.

All three mQTL tests — the standard linear model, the Cao_M test and DGLM_M— accurately controlled FPR under all four significance testing procedures (Figure 2.1, left panel, Figure 2.10, left column, and Table 2.2, column 1, top third). And all twelve test-procedure combinations had indistinguishable power to detect mQTL, in the range [0.692, 0.706] (Table 2.2, column 2 and Figure 2.7).

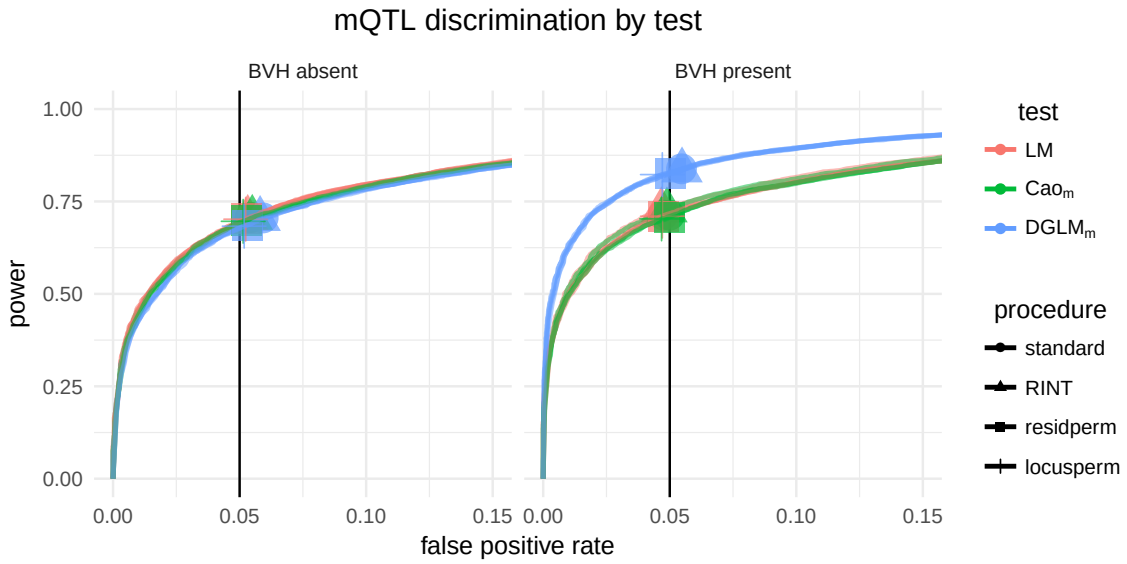


Figure 2.1: ROC curves for detection of mQTL in presence and absence of BVH. Lines are drawn for three different mQTL tests in Table 2.1 and four significance procedures, with a point (circle, square, triangle, tick) corresponding to nominal significance at $p = 0.05$ (more details in **Data and Simulations**). DGLM_M dominates Cao_M and SLM in the presence of BVH, accurately controlling FPR with the locusperm and residperm procedures, but not the standard and RINT procedures, which have FPR of 0.058 and 0.060 (Table 2.2).

These simulation results do not favor any one test over another, but they do favor the standard and RINT assessment procedures over the residperm and locusperm in the sense that the latter two yield no additional improvement in FPR control or power for their additional computational cost.

2.4.1.2 DGLM_M dominates other mQTL tests in the presence of BVH.

The SLM and Cao_M accurately controlled FPR under all four procedures to assess statistical significance (Figure 2.1, right panel, Figure 2.10, right column, and Table 2.2, column 5, top third). In contrast, DGLM_M exhibited modest FPR inflation under the standard and RINT procedures, controlling FPR to the nominal level only under the empirical procedures. Nonetheless, despite requiring an empirical procedure to control FPR, in its power to detect an mQTL under BVH, DGLM_M dominated the other two tests, with power in the range of [0.813, 0.816] compared with the power of SLM and Cao_M in the range of [0.689, 0.718] (Figure 2.1, right panel, Table 2.2).

Based on the results of these simulations, DGLM_M is the preferable mQTL test in the presence of BVH. But, to accurately control FPR, DGLM_M requires an empirical procedure be used to assess statistical significance; both the residperm and locusperm procedures are capable.

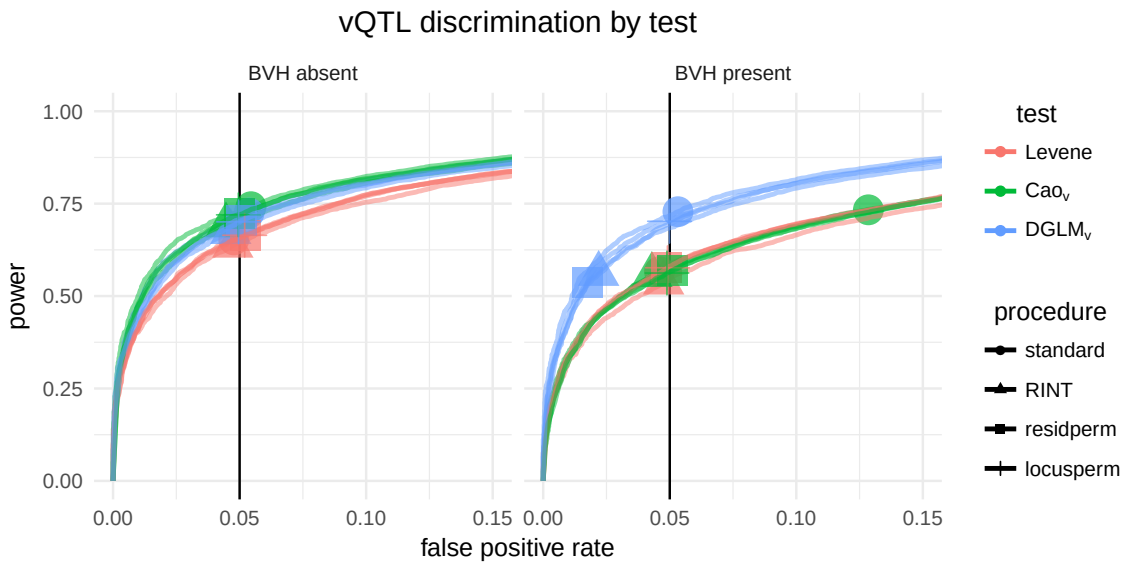


Figure 2.2: ROC curves for detection of vQTL in presence and absence of BVH. Lines are drawn for three different vQTL tests in Table 2.1 and four significance procedures, with a point (circle, square, triangle, tick) corresponding to nominal significance at $p = 0.05$ (more details in **Data and Simulations**). DGLM_V dominates Cao_V and Levene's test in the presence of BVH and accurately controls FPR under the standard and locusperm procedures (Table 2.2). Cao_V suffers a drastic increase in false positives in the presence of BVH under the standard procedure, and DGLM_V would do the same if there were some unmodeled BVH driver, thus DGLM_V under the locusperm procedure is the preferable test.

2.4.1.3 Parametric tests dominate Levene's test for vQTL in the absence of BVH.

In null simulations, Cao_V and DGLM_V exhibited slightly anti-conservative behavior using the standard (*i.e.*, asymptotic) significance testing procedure (FPR = 0.053), modestly conservative behavior under the RINT procedure (FPR = 0.043) and slightly conservative behavior under the residperm and locusperm procedures (FPR in the range [0.046, 0.048], Figure 2.2, left panel, Figure 2.11, left column, and Table 2.2, column 2, middle third). Levene's test, in contrast, was overly conservative using the standard and RINT procedures, but accurately controlled FPR under the empirical procedures (Figure 2.8).

Despite the variation in FPR control among the test-procedure combinations, Cao_V and DGLM_V had more power than Levene's test under all procedures (0.724 vs. 0.667). Thus, the empirical procedures of Cao_V and DGLM_V are the preferred vQTL tests in the absence of BVH, because they have the highest power of the test-procedure combinations that are not anti-conservative. The additional power of Cao_V and DGLM_V relative to Levene's test is consistent with the fact that they make strong parametric assumptions that are exactly true in these simulations and Levene's test does not.

2.4.1.4 DGLM_V dominates other vQTL tests in the presence of BVH.

In the presence of BVH, there were three test-procedure combinations with major departures from accurate FPR control. Cao_V under the standard procedure was drastically anti-conservative, and DGLM_V under both the RINT and residperm procedures was drastically conservative (Figure 2.2, right panel and Figure 2.8, and Figure 2.11, right column). DGLM_V dominated Levene's test and Cao_V, so the standard and locusperm procedure, which accurately control its FPR, seem to be equally preferable and preferable over all other test-procedures.

Nonetheless, there is an important caveat that makes locusperm the strongly preferable significance procedure. In this simulation, there are no BVH driving factors unknown to DGLM_V. If there were such a factor, DGLM_V under the standard procedure would have the same drastic FPR inflation that Cao_V showed under the standard procedure in these simulations (Figure 2.8 (a), third panel). In contrast, the presence of a unknown or unmodeled BVH driving factor does not inflate the FPR of DGLM_V under the locusperm procedure. Due to the practical difficulty of excluding the possibility

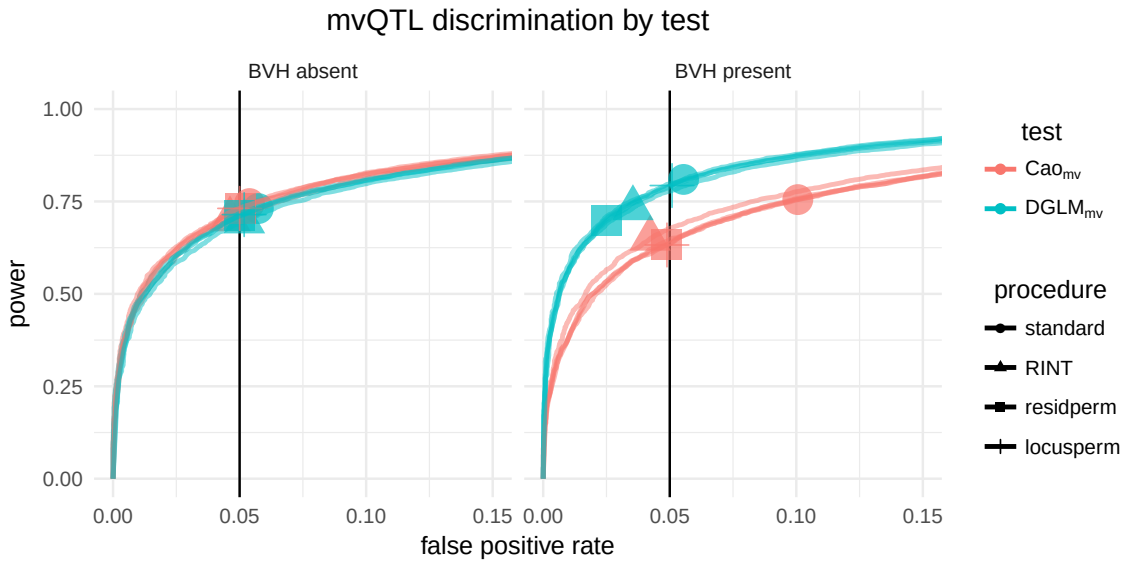


Figure 2.3: ROC curves for detection of mvQTL in presence and absence of BVH. Lines are drawn for two different mvQTL tests in Table 2.1 and four significance procedures, with a point (circle, square, triangle, tick) corresponding to nominal significance at $p = 0.05$ (more details in **Data and Simulations**). mvQTL tests combined the responses of mQTL tests and vQTL tests to BVH, yielding a situation in which DGLM_{MV} dominates Cao_{MV}, but only accurately controls FPR under the locusperm procedure (Table 2.2). The anti-conservative nature of the standard procedure follows from the patterns observed in mQTL tests and the conservative nature of the RINT and residperm procedures follows from the patterns observed in vQTL tests.

of an unknown BVH driver, the most reliable way to guard against covert FPR inflation without giving up the additional power of DGLM_{V} is to use the locusperm procedure.

2.4.1.5 mvQTL mirrors vQTL testing; DGLM_{MV} dominates Cao_{MV} in the presence of BVH.

As with vQTL tests, there was little to distinguish any test or procedure in the absence of BVH except for the modest conservative nature of the RINT procedure and the concomitant decrease in power (Figure 2.3, left panel and Table 2.2, columns 1 and 4, bottom third).

In the presence of BVH, however, DGLM_{MV} dominates Cao_{MV}, with the standard and locusperm procedures accurately controlling FPR (Figure 2.3, right panel and Figure 2.12, right column). As with vQTL testing, due to the difficulty in ruling out BVH from an unknown source and the inflated FPR that results from such BVH under the standard procedure, the DGLM_{MV} under the locusperm procedure is the recommended test for mvQTL.

2.4.1.6 In the presence of BVH, the rank-based inverse normal transformation fails to correct anti-conservative behavior of DGLM_M and over corrects that of DGLM_V and DGLM_MV

A consistent feature of the simulations involving detection of variance effects, whether vQTL or mvQTL, is that FPR control and power is affected, for better or worse, by applying the RINT to the response.

In the presence of BVH, DGLM_M under the standard procedure was anti-conservative (FPR = 0.058 at $\alpha = 0.05$). The RINT procedure had no efficacy in returning this test to accurate FPR control (FPR = 0.060).

In the case of vQTL detection in the presence of BVH, Cao_V under the standard procedure had a drastically inflated FPR (0.123) and the RINT procedure over-corrected it (FPR = 0.044). Similarly, the RINT procedure disrupted DGLM_V, which accurately controlled FPR under the standard procedure, causing overly conservative behavior (FPR = 0.021).

As always, in the presence of BVH, the mvQTL tests exhibited a mixture of the patterns observed in mQTL tests and vQTL tests. Both Cao_MV and DGLM_MV were anti-conservative under the standard procedure, illustrating their relations to Cao_V and DGLM_M respectively. And in both cases, the RINT procedure drove an over-correction into the realm of over conservatism (FPR = 0.046 and 0.038 respectively).

In summary, the RINT procedure is unhelpful in the context of the DGLM_M: it inflates the FPR of a test that is appropriately sized under standard procedures. But, in the context of vQTL testing with BVH from an unknown source, it has one useful and important property: pre-processing the phenotype with the RINT, leads to vQTL tests that are conservative rather than anti-conservative, decreasing the probability of false positives at the expense of false negatives.

2.4.2 Genomewide reanalysis of bodyweight in Leamy et al. backcross

To understand the impact of BVH on mean and variance QTL mapping in real data, we applied both traditional QTL mapping, using SLM, and mean-variance QTL mapping, using Cao's tests and the DGLM, to body weight at three weeks in the mouse backcross dataset of Leamy et al. (2000).

		BVH absent				BVH present			
test	version	null	mQTL	vQTL	mvQTL	null	mQTL	vQTL	mvQTL
LM	standard	0.050	0.706	0.056	0.510	0.051	0.701	0.050	0.508
	RINT	0.050	0.704	0.052	0.495	0.052	0.718	0.049	0.518
	residperm	0.048	0.700	0.054	0.502	0.049	0.694	0.050	0.504
	locusperm	0.049	0.702	0.054	0.499	0.049	0.695	0.049	0.503
Cao_{M}	standard	0.052	0.705	0.051	0.515	0.050	0.700	0.048	0.516
	RINT	0.051	0.705	0.051	0.503	0.052	0.716	0.047	0.523
	residperm	0.049	0.697	0.049	0.503	0.048	0.695	0.047	0.508
	locusperm	0.048	0.691	0.048	0.500	0.048	0.689	0.044	0.503
DGLM_{M}	standard	0.052	0.705	0.051	0.515	0.058	0.832	0.054	0.649
	RINT	0.051	0.705	0.051	0.503	0.060	0.830	0.054	0.644
	residperm	0.049	0.696	0.049	0.504	0.052	0.816	0.046	0.629
	locusperm	0.048	0.692	0.048	0.500	0.050	0.813	0.046	0.624
Levene's test	standard	0.045	0.049	0.660	0.462	0.046	0.046	0.577	0.393
	RINT	0.045	0.043	0.645	0.422	0.047	0.043	0.546	0.344
	residperm	0.049	0.052	0.667	0.474	0.048	0.050	0.585	0.401
	locusperm	0.049	0.052	0.667	0.474	0.048	0.050	0.583	0.400
Cao_{V}	standard	0.053	0.053	0.750	0.543	0.123	0.127	0.744	0.563
	RINT	0.043	0.042	0.700	0.467	0.044	0.048	0.563	0.364
	residperm	0.047	0.051	0.729	0.519	0.045	0.054	0.572	0.388
	locusperm	0.046	0.049	0.726	0.517	0.047	0.051	0.567	0.382
DGLM_{V}	standard	0.053	0.053	0.750	0.543	0.049	0.056	0.732	0.524
	RINT	0.043	0.042	0.700	0.467	0.021	0.027	0.570	0.340
	residperm	0.048	0.051	0.729	0.520	0.015	0.018	0.542	0.329
	locusperm	0.046	0.049	0.724	0.515	0.046	0.050	0.713	0.498
Cao_{MV}	standard	0.050	0.597	0.643	0.741	0.100	0.642	0.649	0.751
	RINT	0.043	0.585	0.574	0.701	0.046	0.600	0.436	0.646
	residperm	0.046	0.587	0.618	0.728	0.050	0.514	0.507	0.632
	locusperm	0.048	0.589	0.617	0.727	0.050	0.516	0.506	0.630
DGLM_{MV}	standard	0.050	0.597	0.643	0.741	0.057	0.741	0.633	0.807
	RINT	0.043	0.585	0.574	0.701	0.038	0.715	0.445	0.726
	residperm	0.046	0.590	0.618	0.729	0.024	0.621	0.469	0.687
	locusperm	0.046	0.590	0.617	0.728	0.051	0.723	0.601	0.788

Table 2.2: Positive rates of all tests in all scenarios based on 10,000 simulations, 1,000 permutations each to estimate empirical null distributions (residperm and locusperm), and a nominal false positive rate (FPR) of $\alpha = 0.05$. Entries in column 1 and 5 through all rows, columns 3 and 7 in the top third, and columns 2 and 6 in the middle third represent FPR. The entries in the rest of the table represent power. The largest standard error for an FPR is 0.001. The largest standard error for a power is 0.0025.

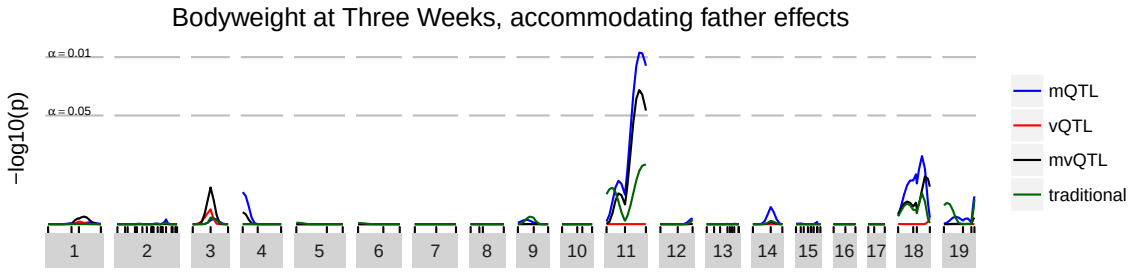


Figure 2.4: FWER-controlling association statistic at each genomic locus for body weight at three weeks. The linear model (green, "traditional") does not detect any statistically-significant associations. The mQTL test takes into account the heterogeneity of both mean and variance due to which F1 male fathered each mouse in the mapping population and detects one mQTL on chromosome 11.

2.4.2.1 Analysis with Traditional QTL Mapping Identifies no QTL.

We first used a traditional, linear modeling-based QTL analysis, with sex and father as additive covariates and genomewide significance based on 1000 genome permutations (Churchill and Doerge, 1994). Although sex was found not to be a statistically significant predictor of body weight ($p = 0.093$ by the likelihood ratio test with 1 degree of freedom), it was included in the mapping model because, based on the known importance of sex in determining body weight, any QTL that could only be identified in the absence of modeling sex effects would be highly questionable. Father was found to be a significant predictor of body weight in the baseline fitting of the SLM ($p = 9.6 \times 10^{-5}$ by the likelihood ratio test with 8 degrees of freedom) and therefore was included in the mapping model.

No associations rose above the threshold that controls family-wise error rate to 5% (Figure 2.4, green line). One region on the distal part of chromosome 11 could be considered "suggestive" with FWER-adjusted $p \approx 0.17$.

To test the sensitivity of the results to the inclusion/exclusion of covariates, the analysis was repeated without sex as a covariate, without father as a covariate, and with no covariates. No QTL were identified in any of these sensitivity analyses. This result is consistent with the fact that none of the publications on this mapping population reported any QTL for body weight at any age.

2.4.2.2 Analysis with Cao's tests Identifies no QTL

The same phenotype was analyzed with Cao's tests, again including sex and father as mean covariates, and using the genome permutation procedures described in **Methods** were used to control FWER. No statistically significant mQTL, vQTL, nor mvQTL were identified (Figure 2.14b).

2.4.2.3 Analysis with DGLM-based tests Identifies an mQTL

The same phenotype was analyzed with the DGLM-based tests. In a baseline fitting of the DGLM, sex was found not to be a statistically significant predictor of mean or residual variance (mean effect $p = 0.18$, variance effect $p = 0.22$, and joint $p = 0.19$ by the LRT with 1, 1, and 2 d.f.). But father was found to be a statistically significant predictor of both mean and variance (mean effect $p = 2.0 \times 10^{-7}$, variance effect $p = 1.8 \times 10^{-11}$, and $p = 4.8 \times 10^{-14}$ by the LRT with 8, 8, and 16 d.f.). Therefore, following the same reasoning as in the mean model described above, both sex and father were included in the mapping model as covariates of both the mean and the variance. As with the other tests, the genome permutation procedures described in **Methods** were used to control FWER.

A genomewide significant mQTL was identified on chromosome 11 (Figure 2.4, blue line). The peak was at 69.6 cM with FWER-adjusted $p = 0.011$, with the closest marker being D11MIT11 at 75.7 cM with FWER-adjusted $p = 0.016$. Nonparametric bootstrap resampling, using 1,000 resamples (after Visscher et al. 1996), established a 90% confidence interval for the QTL from 50 to 75 cM. This region overlaps with the "suggestive" region identified in the traditional analysis.

By the traditional definition of percent variance explained, following from a fitting of the standard linear model, this QTL explains 2.1% of phenotype variance. Though, given the variance heterogeneity inherent in the DGLM that was used to detect this QTL, this quantity is better considered the "average" percent variance explained. The ratio of the QTL variance to the sum of QTL variance, covariate variance, and residual variance ranges from 1% to 6% across the population, based on the heterogeneity of residual variance.

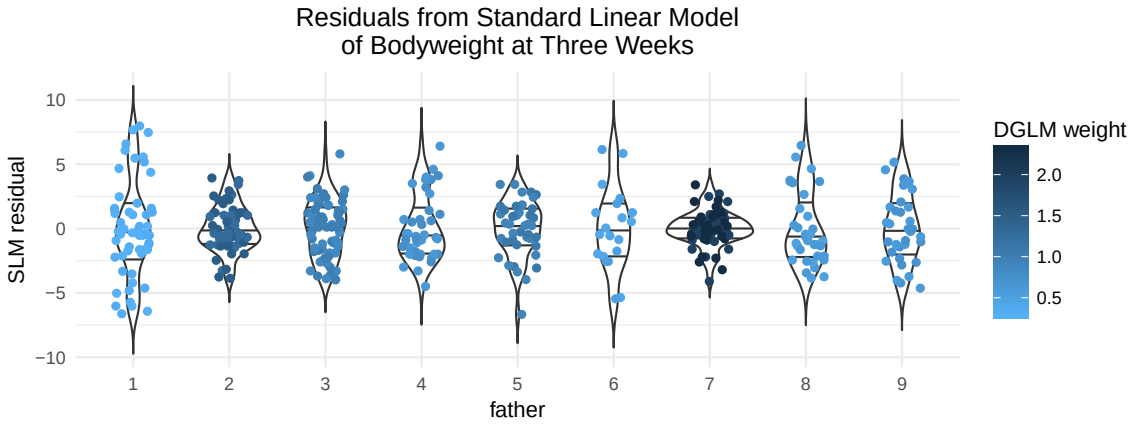


Figure 2.5: Residuals from the standard linear model for body weight at three weeks, with sex and father as covariates, stratified by father. It is evident that fathers differed in the residual variance of the offspring they produced. For example, the residual variance of offspring from father 1 is greater than that of father 2 and 7. Here, points are colored by their predicted residual variance in the fitted DGLM with sex and father as mean and variance covariates.

2.4.2.4 Understanding the Novel QTL

The mQTL on chromosome 11 was identified by the DGLM_M test, but not by by the standard linear model or Cao's mQTL test. The additional power of the DGLM_M test over these other tests relates to its accommodation of background variance heterogeneity (BVH).

Specifically, the DGLM reweighted each observation based on its residual variance, according to the sex and F1 father of the mouse. This BVH is visually apparent when the residuals from the standard linear model are plotted, separated out by father (Figure 2.5).

Some fathers, for example fathers 2 and 7, appear to have offspring with less residual variance than average, whereas others, for example father 1, seem to have offspring with more residual variance than average. The DGLM captured these patterns of variance heterogeneity, and estimated the effect of each father on the log standard deviation of the observations (Figure 2.6). Based on these estimated variance effects, observations were upweighted (e.g. fathers 2 and 7) and downweighted (e.g. father 1). This weighting gave the DGLM-based mapping approach more power to reject the null as compared to the SLM.

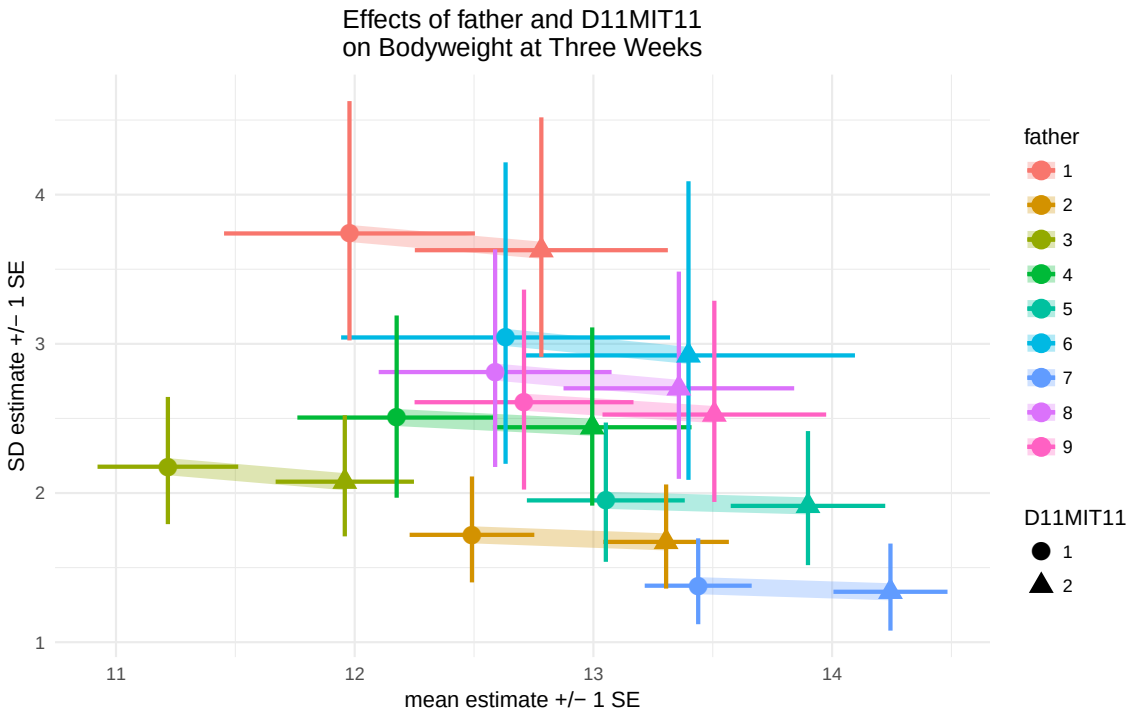


Figure 2.6: The predictive mean and standard deviation of mice in the mapping population based on father and genotype at the top marker, D11MIT11 on chromosome 11. The genotype effect, illustrated by the colored ribbons is almost entirely horizontal, indicating a difference in means across genotype groups but no difference in variance, consistent with the identification of this QTL as a pure mQTL. The father effects, illustrated by the spread of colored crossbars, have both mean and variance components. For example, father 7 (blue) has the highest predictive mean and lowest predictive standard deviation. His offspring were upweighted in the QTL analysis based on their low standard deviation. Father 1 (red) has an average predictive mean and the highest predictive standard deviation. His offspring were downweighted in the QTL analysis based on their high standard deviation. Note: the effect of sex on phenotype mean and variance was modeled, then marginalized out for readability.

2.4.2.5 Other Phenotypes

For brevity, we described in detail only the results of the DGLM-based analysis of body weight at three weeks; but, of the eight phenotypes from this cross available on the Mouse Phenome Database, the mean-variance approach to QTL mapping discovered new QTL in four. Five of the eight phenotypes — body weight at twelve days, three weeks, and six weeks, as well as subcutaneous and gonadal fat pad thickness — exhibited BVH due to father, and for each we performed both traditional QTL mapping using the SLM and mean-variance QTL mapping using the DGLM. For body weight at three weeks and six weeks, we identified one new mQTL and two new vQTL respectively. For subcutaneous fat pad thickness, we "undiscovered" one mean QTL. That is, after reweighting the observations based on the observed variance of each father, two QTL that were detected by the SLM no longer met criteria for statistical significance, as shown in supplementary figures.

2.5 Discussion

The simulation studies revealed that in the presence of background variance heterogeneity (BVH), the DGLM-based tests are uniquely powerful in the detection of mQTL, vQTL, and mvQTL.

Our reanalysis of the Leamy et al. dataset demonstrated that the additional power of DGLM_M in the face of BVH can be used to detect an mQTL that was overlooked by all competitor methods.

2.5.1 Detecting and Modeling BVH

To select the right test and procedure to assess significance, it is important to establish whether there is any BVH present. We advocate fitting the DGLM with all potential BVH drivers as variance covariates, then including any that are statistically significant as variance covariates in the mapping model to improve power to detect QTL.

2.5.2 Guidelines for QTL mapping in the presence of BVH

Given that

1. The DGLM-based tests dominate all other tests in the presence of BVH,

2. the locusperm procedure accurately controls the FPR of the DGLM-based tests in the presence of BVH, whether the source is known or not, and

3. the locusperm procedure can be extended into the genomeperm procedure to control FWER,

we advocate for the analysis of experimental crosses that exhibit BVH with the three DGLM-based tests (DGLM_M, DGLM_V, and DGLM_MV) and, where the individuals in the population are exchangeable (as in an F2 or backcross) or where partial exchangeability can be suitably identified [*e.g.*, see (Churchill and Doerge, 1994; Zou et al., 2006; Churchill and Doerge, 2008)], the use of our described genomeperm procedures, which permute the genome in selective parts of the model, to assess genomewide significance.

Because this procedure involves three families of tests rather than one family as would be typical with an SLM-based analysis, an additional correction may be desired to control experiment-wise error rate. DGLM_M and DGLM_V are orthogonal tests (Smyth, 1989), but DGLM_MV is neither orthogonal nor identical to either, so the effective number of families is between two and three. One reasonable, heuristic approach to control experiment-wise error rate is simply to lower the acceptable FWER, *e.g.* replacing the standard 0.05 with 0.02.

2.5.3 Data reweighting for mQTL detection

The additional power of mean-variance QTL mapping to detect mQTL in general, and of DGLM_M to detect mQTL in the presence of BVH in particular, can be seen as deriving from how data is reweighted. This reweighting is not based on any prior knowledge on the part of the experimenter, but rather based on patterns of residual variance heterogeneity detected by the DGLM.

The impact of reweighting can be illustrated through consideration of the normal likelihood. For $y_i \sim \text{N}(m_i, \sigma^2/w_i)$, with known weights $w_1, \ldots, w_n$ and known baseline variance σ^2, the log-likelihood can be written as $\ell = \text{const} - \text{WRSS}/2\sigma^2$, where the key quantity to be minimized[2],

$$\text{WRSS} = \sum_{i=1}^{n} w_i (y_i - m_i)^2 \,,$$

[2]Note: const $= -0.5(n \log 2\pi - \sum_{i=1}^{n} w_i \log \sigma^2)$ can be ignored.

is the weighted residual sum of squares, that is, the squared discrepancies between the observed phenotype y_i and its predicted value m_i weighted by w_i. The weights therefore affect how much, relatively speaking, each data point contributes to the likelihood: highly imprecise measurements, such as from individuals whose phenotypes are expected to have high variance, have low weight and diminished contribution, whereas as more precise measurements are correspondingly upweighted. In the DGLM, weights are informed by experimental covariates and the QTL genotype itself, as $w_i = e^{-v_i}$. In the SLM, unless weights are specified externally, there is no such mechanism for phenotype precision to be incorporated and so all weights equal 1. The improvement of the DGLM over the SLM and Cao_M, therefore stems entirely from its greater ability to provide this additional information, and thereby give more credence to phenotype values that are expected to be more precise.

This reweighting can be thought of has having two benefits in the QTL mapping endeavor. Geneticists are often rightly concerned about high leverage observations, which can cause to false positives. Less often acknowledged is that high leverage observations may also induce false negatives, disrupting an otherwise good statistical model fit. By bringing the data weights into alignment with their estimated residual variance, the DGLM addresses both of these concerns: by downweighting outliers from systematically noisy subgroups, it reduces the potential for false positives; by upweighting outliers from systematically precise subgroups, it reduces the probability of false negatives.

2.5.4 Covariate correction for vQTL detection

Conceptually, the additional power of the DGLM_V to detect vQTL over Cao_V in the presence of BVH, as demonstrated above, derives from its ability to accommodate a covariate, just as any linear regression analysis benefits from accommodating a covariate. The distinction is that, whereas the response for the in a typical regression analysis is the observed data, in the case of BVH and the DGLM, the response is the squared residuals from the mean sub-model.

As with any regression analysis, when the covariate effect is meaningfully large, its inclusion in the model improves the estimation of the effect of interest. The more precise the estimation of the effect of interest allows a greater model improvement from the null to alternative model and ultimately, a more powerful test.

2.5.5 Percent Variance Explained

Variance heterogeneity complicates the notion of percent variance explained (PVE) by a QTL. Assuming the QTL has the same effect on the expected value of the phenotype of all individuals, it will explain a larger percent of total variance for individuals with lower than average residual variance, and vice versa for individuals with higher than average residual variance. In light of this observation, the percent variance explained can either be reported as "average percent variance explained" or can be calculated for some representative sub-groups. For example, if there is variance heterogeneity across sexes, it would be reasonable to report the PVE of a QTL for both males and females, or if a vQTL is known to be present elsewhere in the genome, report the PVE for each vQTL genotype as in Yang et al. (2012).

2.5.6 Rank inverse normal transformation: pros and cons for vQTL mapping

In the detection of vQTL, foreground variance heterogeneity (FVH) and BVH come into conflict — the goal is to detect FVH and BVH obscures its detection. Both, however, induce excess kurtosis (fatter tails) in the phenotype distribution. Thus, it is logical that the RINT, which reshapes away excess kurtosis without reference to its source, should have both beneficial and harmful properties.

In the case where there is no known driver of BVH, a scenario represented by the simulations examining Cao_V, the RINT procedure acts like an insurance policy: if there truly is no BVH, the test suffers a modest decrease in power; but if there truly is BVH from an unknown source, it averts the drastic FPR inflation under the standard (*i.e.*, non-empirical) p-value procedure.

In the case where the major BVH drivers are known, represented by the DGLM_V simulations, the RINT procedure is unnecessary, costing power with its conservatism in the absence of BVH and paradoxically creating even more conservative behavior in the presence of BVH.

The above disadvantages of RINT assume the phenotype data has an underlying normal distribution, either as given or after a simple (*e.g.*, power) transformation. When this is not so, that is, in cases of highly non-normal data, valid inference would be possible by both the RINT and the locusperm procedure, and perhaps the most robust approach would be to use the two in combination. Nonetheless, where normality approximately holds, whether as given or after a simple transformation, we strongly prefer the locusperm procedure without RINT: across all simulation scenarios it exhibited

at worst slight conservatism when applied to DGLM-based tests and represents a useful step toward FWER control.

2.6 Additional Information

2.6.1 Simulation Details:

In simulation with BVH present, the group-wise effects on the log standard deviation were $\gamma = [-0.4, -0.2, 0, 0.2, 0.4]$. Though $\overline{\gamma} = 0$, the exponential transform connecting these effects to the standard deviation results in a simulated phenotype with slightly more total variance than one without BVH. Therefore, the additive effect of the locus on phenotype mean was adjusted when BVH was introduced, in order to maintain a constant percent variance explained by the mean effect. The following values were used in the simulation.

	no BVH	yes BVH
null	$\alpha = 0, \theta = 0$	$\alpha = 0, \theta = 0$
mQTL	$\alpha = 0.22, \theta = 0$	$\alpha = 0.25, \theta = 0$
vQTL	$\alpha = 0, \theta = 0.17$	$\alpha = 0, \theta = 0.17$
mvQTL	$\alpha = 0.18, \theta = 0.14$	$\alpha = 0.2, \theta = 0.136$

null locus and mQTL in the absence of BVH: All observations have standard deviation 1.

vQTL in the absence of BVH: The genotype-wise standard deviations implied by the additive effect of 0.17 on the log standard deviation are approximately: [0.84, 1.00, 1.19].

mvQTL in the absence of BVH: The genotype-wise standard deviations implied by the additive effect of 0.14 on the log standard deviation are approximately: [0.87, 1.00, 1.15].

null locus and mQTL in the presence of BVH: The covariate-wise standard deviations implied by the effects of [-0.4, -0.2, 0, 0.2, 0.4] on the log standard deviation are approximately: [0.67, 0.82, 1.00, 1.22, 1.49].

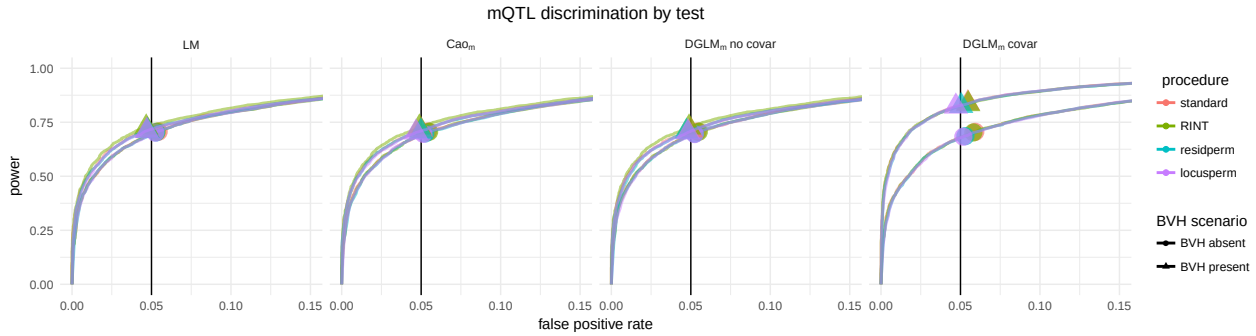
vQTL in the presence of BVH: Locus and covariate effects on the residual variance combine additively on the log standard deviation scale, yielding 15 distinct standard deviations:

mvQTL in the presence of BVH: Locus and covariate effects on the residual variance combine additively on the log standard deviation scale, yielding 15 distinct standard deviations:

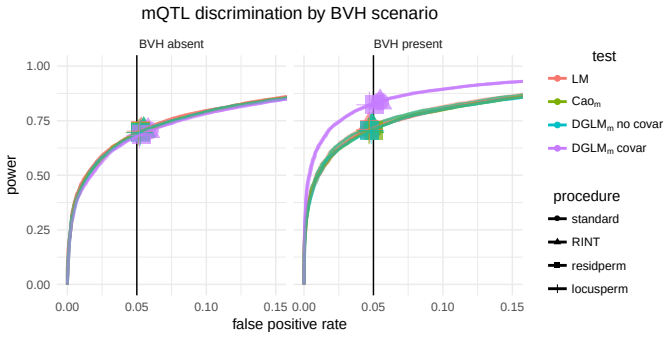
	genotype		
covar	-1	0	1
1	0.57	0.67	0.79
2	0.69	0.82	0.97
3	0.84	1.00	1.19
4	1.03	1.22	1.45
5	1.26	1.49	1.77

	genotype		
covar	-1	0	1
1	0.59	0.67	0.77
2	0.71	0.82	0.94
3	0.87	1.00	1.15
4	1.07	1.22	1.40
5	1.30	1.49	1.71

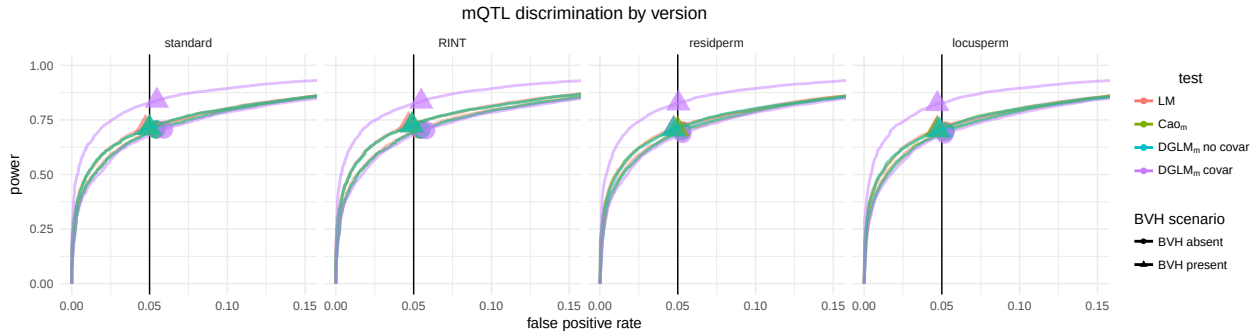
2.6.2 ROC Curves



(a) All test-evaluations accurately control FPR. DGLM$_M$ with BVH of known source is the most powerful test.

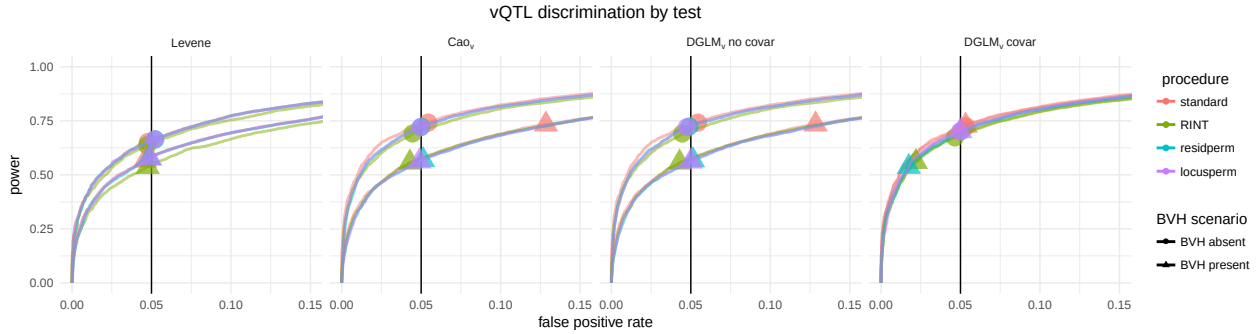


(b) Within BVH scenarios, all mQTL tests perform equivalently.

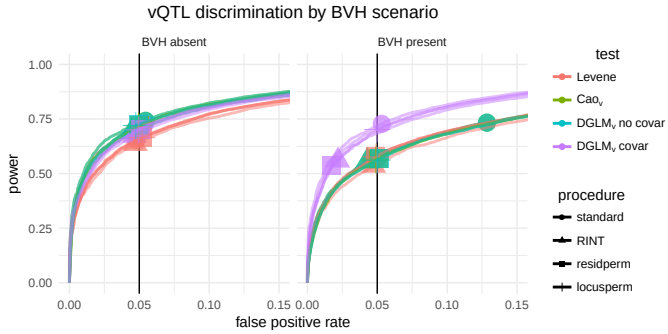


(c) DGLM$_M$ outperforms all other tests across all evaluation methods.

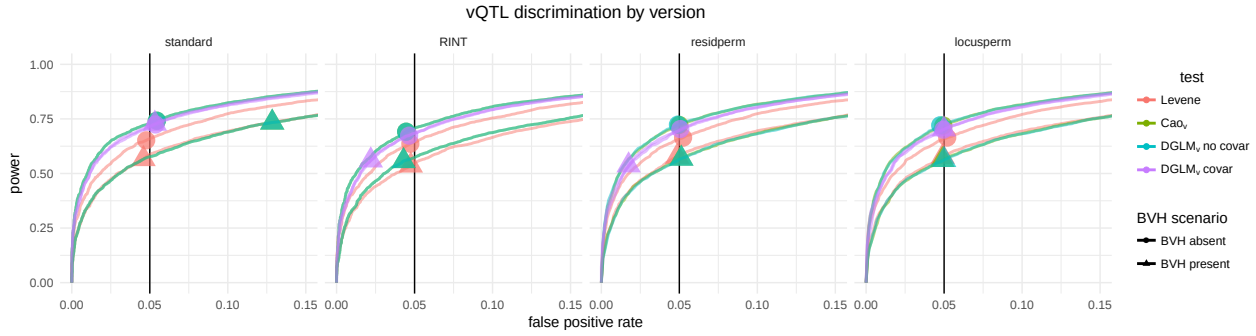
Figure 2.7: ROC Curves for mQTL tests in the detection of mQTL. The same 32 ROC curves are plotted three times, organized by (a) test, (b) BVH scenario, and (c) version to allow for comparisons across all dimensions.



(a) Levene's test accurately controls FPR in all scenarios.Cao$_V$ and DGLM$_V$ have inflated FPR in the presence of BVH of unknown source. DGLM$_V$'s RINT and residperm versions are anti-conservative in the presence of BVH of known source.

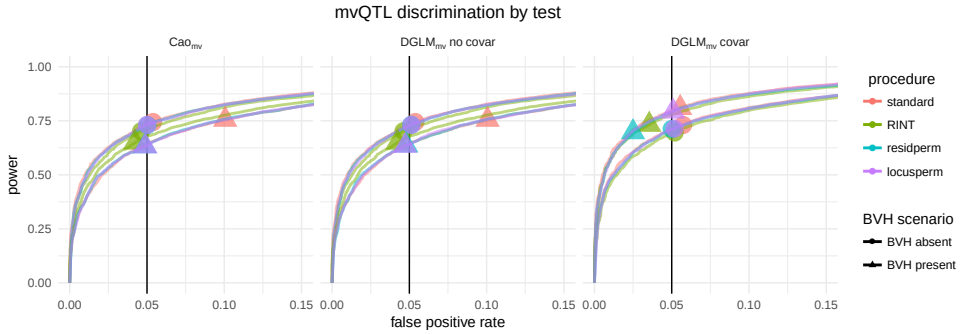


(b) In the absence of BVH, Levene's test is less powerful than Cao$_V$ and DGLM$_V$. In the face of BVH of unknown source, all tests suffer decreased power, except DGLM$_V$'s standard version, which fails to accurately control FPR. In the scenario with BVH of known source, DGLM$_V$ recovers most of the power lost with introduction of BVH, but its RINT and residperm versions are anti-conservative.

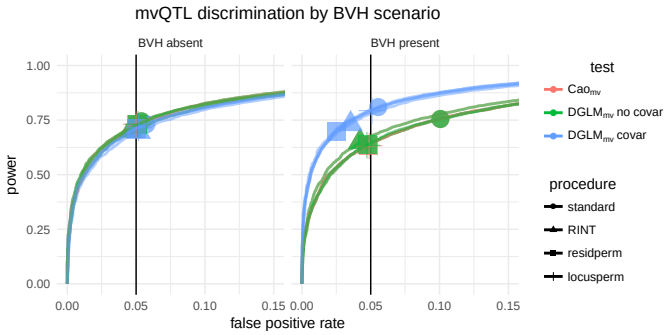


(c) The only version of DGLM$_V$ that accurately controls FPR across all BVH scenarios is locusperm. Its standard version is anti-conservative in the presence of BVH of unknown source and its RINT and residperm versions are conservative in the presence of BVH of known source.

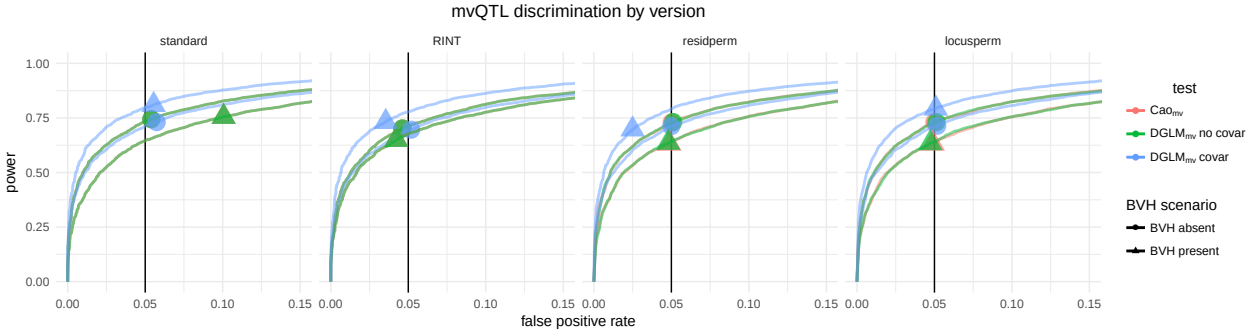
Figure 2.8: ROC Curves for vQTL tests in the detection of vQTL. The same 32 ROC curves are plotted three times, organized by (a) test, (b) BVH scenario, and (c) version to allow for comparisons across all dimensions.



(a) Cao_{MV} and DGLM_{MV} both suffer a decrease in discrimination (down and right shift of ROC curve) in the presence of BVH of unknown (or unmodeled) source. Only DGLM_{MV} can accommodate the source when it is known and therefore can achieve superior discrimination in that case. The standard and locusperm versions of DGLM_{MV} accurately control FPR.



(b) In the absence of BVH, both mvQTL tests accurately control FPR and have similar power. In the presence of BVH of unknown source, the standard version of both mvQTL tests is anti-conservative and the other three versions maintain FPR control but suffer a decrease in power compared to the no-BVH scenario. Only DGLM_{MV} can incorporate information on the BVH-driving covariate. It achieves increased power and accurately controls FPR in its standard and locusperm versions and is conservative in its RINT and locusperm versions.



(c) Only the locusperm version accurately controls FPR in all scenarios. The standard version of both tests are anti-conservative in the presence of BVH of known source and the RINT and residperm versions are conservative in DGLM_{MV} the presence of BVH of known source.

Figure 2.9: ROC Curves for mvQTL tests in the detection of mvQTL. The same 24 ROC curves are plotted three times, organized by (a) test, (b) BVH scenario, and (c) version to allow for comparisons across all dimensions.

2.6.3 QQ Plots

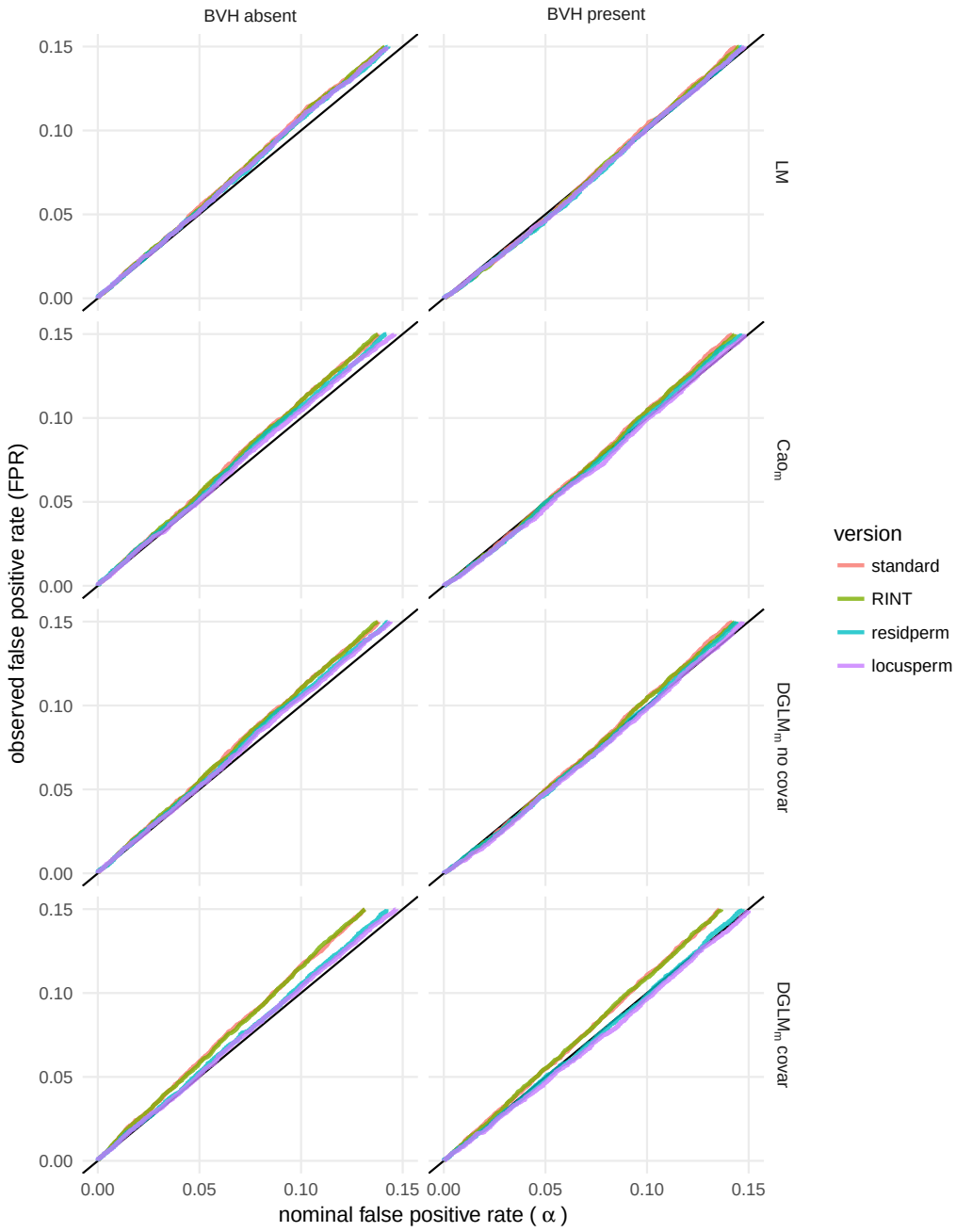


Figure 2.10: The empirical false positive rate of each mQTL test-version for each nominal false positive rate, α, in [0, 0.1]. A test that accurately controls FPR will have empirical FPR $= \alpha$ for all value of α. All mQTL tests accurately control FPR.

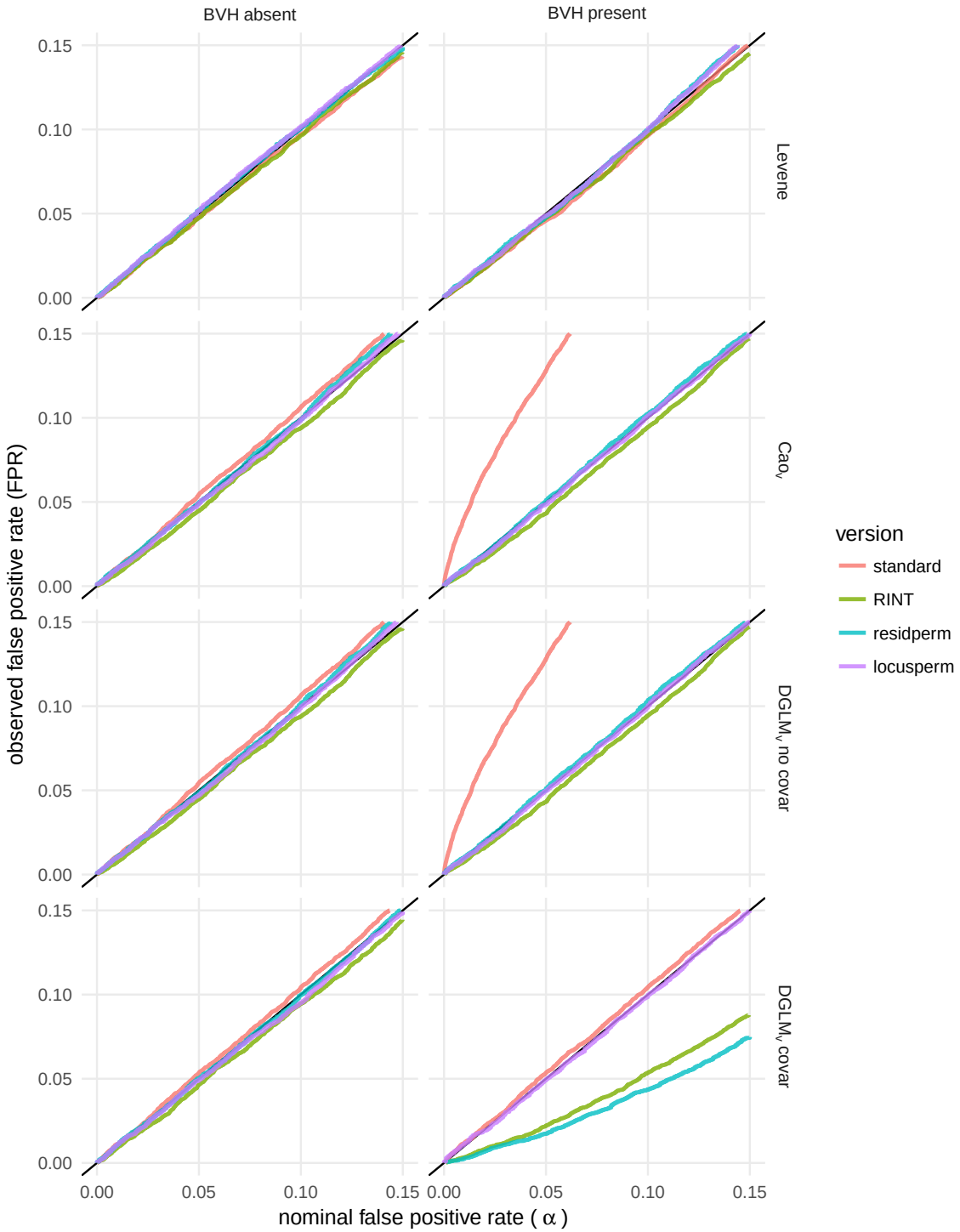


Figure 2.11: The empirical false positive rate of each vQTL test-version for each nominal false positive rate, α, in [0, 0.1]. A test that accurately controls FPR will have empirical FPR $= \alpha$ for all value of α. Amongst vQTL tests, Cao$_V$ has conservative behavior in the presence of BVH when the standard procedure is used, and DGLM$_V$ has anti-conservative behavior when the RINT and residperm procedures are used.

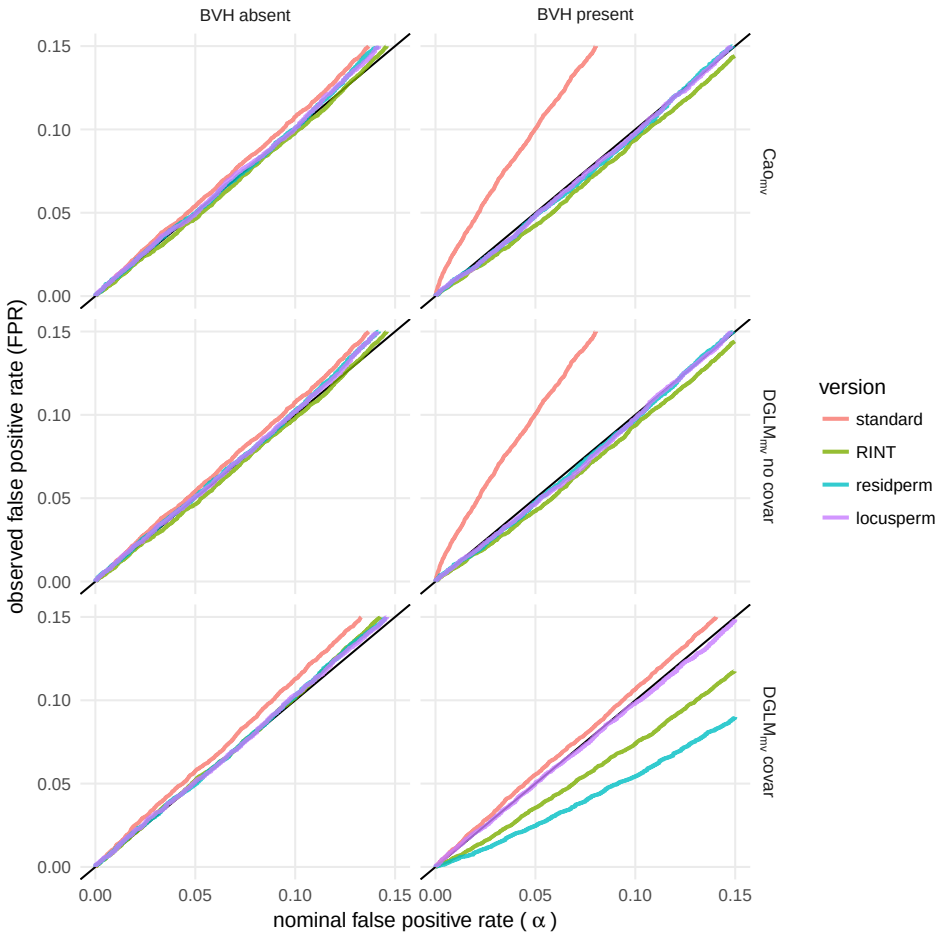


Figure 2.12: The empirical false positive rate of each mvQTL test-version for each nominal false positive rate, α, in [0, 0.1]. A test that accurately controls FPR will have empirical FPR $= \alpha$ for all value of α. mvQTL tests show the same pattern of deviation from accurate FPR control as vQTL tests (Figure 2.11), but to a lesser extent.

2.6.4 False Positive Rates of mQTL tests

		BVH absent				BVH present			
test	version	null	mQTL	vQTL	mvQTL	null	mQTL	vQTL	mvQTL
LM	standard	0.054	0.706	0.055	0.504	0.047	0.713	0.051	0.510
	RINT	0.053	0.706	0.053	0.490	0.047	0.727	0.047	0.522
	residperm	0.052	0.700	0.053	0.495	0.046	0.708	0.050	0.506
	locusperm	0.051	0.698	0.054	0.494	0.046	0.708	0.050	0.506
Cao_M	standard	0.055	0.707	0.049	0.511	0.050	0.713	0.048	0.522
	RINT	0.055	0.706	0.047	0.501	0.049	0.726	0.047	0.534
	residperm	0.051	0.696	0.046	0.504	0.049	0.704	0.046	0.512
	locusperm	0.050	0.693	0.046	0.499	0.046	0.699	0.045	0.510
$DGLM_M$ no covar	standard	0.055	0.707	0.049	0.511	0.050	0.713	0.048	0.522
	RINT	0.055	0.706	0.047	0.501	0.049	0.726	0.047	0.534
	residperm	0.052	0.695	0.046	0.502	0.047	0.703	0.048	0.513
	locusperm	0.051	0.693	0.044	0.500	0.047	0.700	0.044	0.509
$DGLM_M$ with covar	standard	0.059	0.705	0.052	0.512	0.055	0.838	0.057	0.658
	RINT	0.058	0.703	0.050	0.503	0.055	0.835	0.056	0.651
	residperm	0.052	0.683	0.044	0.490	0.049	0.823	0.051	0.634
	locusperm	0.051	0.680	0.045	0.485	0.046	0.821	0.049	0.630

Table 2.3: Positive rates of all four mQTL tests in all scenarios based on 10,000 simulations, 1,000 permutations each to estimate empirical null distributions (residperm and locusperm), and a cutoff of $p = 0.05$. Note that in all cases the DGLM test without the covariate had identical or very nearly identical FPR to the Cao test that tests for the same kind of QTL.

2.6.5 False Positive Rates of vQTL tests

		BVH absent				BVH present			
test	version	null	mQTL	vQTL	mvQTL	null	mQTL	vQTL	mvQTL
Levene's test	standard	0.048	0.045	0.653	0.466	0.046	0.048	0.566	0.387
	RINT	0.048	0.040	0.637	0.422	0.047	0.043	0.536	0.339
	residperm	0.052	0.049	0.661	0.477	0.048	0.051	0.573	0.394
	locusperm	0.051	0.050	0.661	0.475	0.048	0.050	0.573	0.393
Cao_V	standard	0.054	0.051	0.742	0.543	0.128	0.124	0.733	0.571
	RINT	0.045	0.040	0.691	0.468	0.043	0.047	0.557	0.365
	residperm	0.049	0.048	0.720	0.520	0.050	0.050	0.565	0.388
	locusperm	0.048	0.048	0.718	0.517	0.048	0.048	0.559	0.382
DGLM_Vno covar	standard	0.054	0.051	0.742	0.543	0.128	0.124	0.733	0.571
	RINT	0.045	0.040	0.691	0.468	0.043	0.047	0.557	0.365
	residperm	0.048	0.049	0.721	0.522	0.050	0.050	0.565	0.388
	locusperm	0.047	0.048	0.718	0.519	0.049	0.049	0.559	0.381
DGLM_V	standard	0.053	0.053	0.724	0.525	0.054	0.054	0.729	0.531
with covar	RINT	0.046	0.041	0.673	0.453	0.022	0.022	0.560	0.341
	residperm	0.050	0.047	0.699	0.501	0.017	0.015	0.533	0.325
	locusperm	0.049	0.048	0.698	0.501	0.049	0.050	0.700	0.498

Table 2.4: Positive rates of all four vQTL tests in all scenarios based on 10,000 simulations, 1,000 permutations each to estimate empirical null distributions (residperm and locusperm), and a cutoff of $p = 0.05$. Note that in all cases the DGLM test without the covariate had identical or very nearly identical FPR to the Cao test that tests for the same kind of QTL.

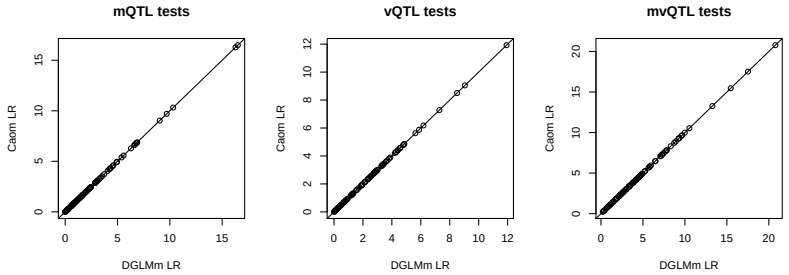
2.6.6 False Positive Rates of mvQTL tests

		BVH absent				BVH present			
test	version	null	mQTL	vQTL	mvQTL	null	mQTL	vQTL	mvQTL
Cao_{MV}	standard	0.054	0.594	0.637	0.745	0.100	0.651	0.644	0.755
	RINT	0.046	0.585	0.570	0.703	0.042	0.608	0.435	0.649
	residperm	0.049	0.587	0.609	0.726	0.048	0.523	0.505	0.628
	locusperm	0.049	0.588	0.611	0.728	0.047	0.523	0.504	0.630
$DGLM_{MV}$	standard	0.054	0.594	0.637	0.745	0.100	0.651	0.644	0.755
no covar	RINT	0.046	0.585	0.570	0.703	0.042	0.608	0.435	0.649
	residperm	0.050	0.588	0.610	0.728	0.047	0.524	0.503	0.632
	locusperm	0.050	0.587	0.612	0.728	0.046	0.522	0.505	0.631
$DGLM_{MV}$	standard	0.058	0.596	0.620	0.732	0.055	0.745	0.621	0.809
with covar	RINT	0.052	0.587	0.554	0.696	0.036	0.720	0.432	0.732
	residperm	0.049	0.576	0.582	0.710	0.025	0.625	0.457	0.694
	locusperm	0.051	0.577	0.582	0.712	0.049	0.731	0.591	0.790

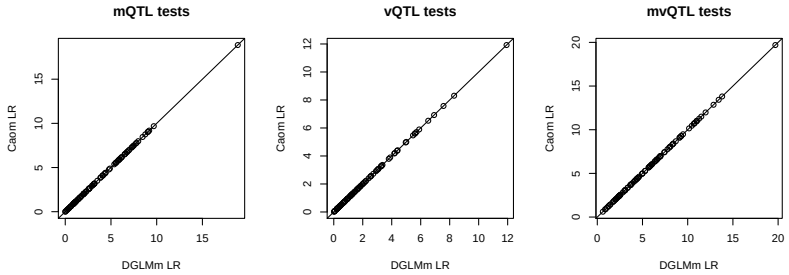
Table 2.5: Positive rates of all three tests in all scenarios based on 10,000 simulations, 1,000 permutations each to estimate empirical null distributions (residperm and locusperm), and a cutoff of $p = 0.05$. Note that in all cases the DGLM test without the covariate had identical or very nearly identical FPR to the Cao test that tests for the same kind of QTL.

2.6.7 Cao's Profile-Likelihood Approximation is Extremely Accurate

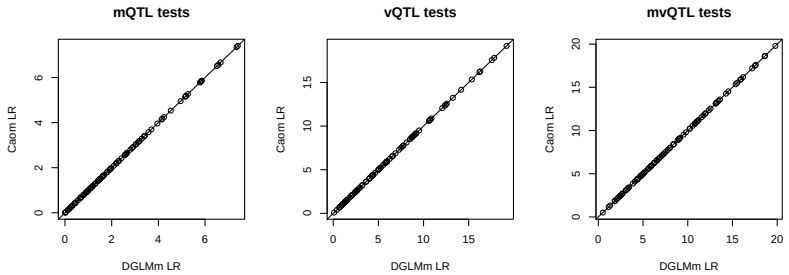
(a) null simulations



(b) mQTL simulations



(c) vQTL simulations



(d) mvQTL simulations

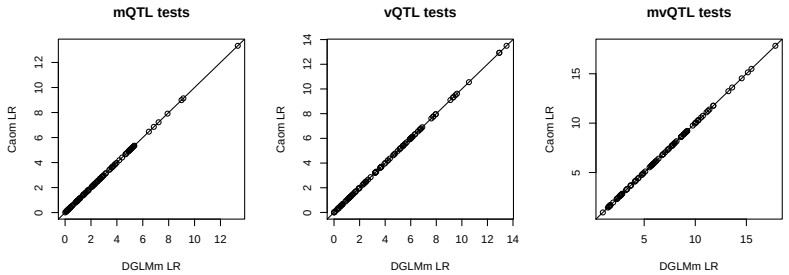
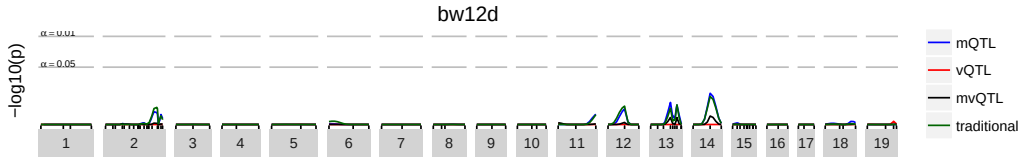
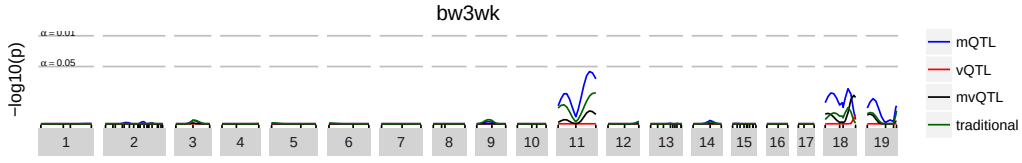


Figure 2.13: On simulated null loci, mQTL, vQTL, and mvQTL, Cao's profile likelihood method had identical likelihood ratio to DGLM when DGLM does not use any variance covariates.

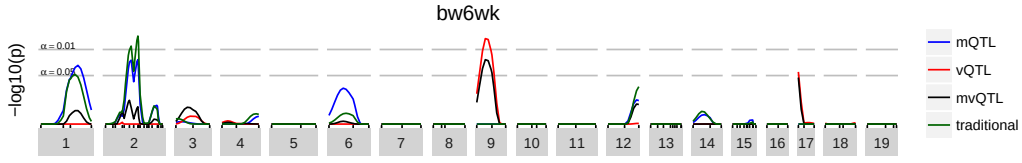
2.6.8 Cao's Tests for All Phenotypes with BVH



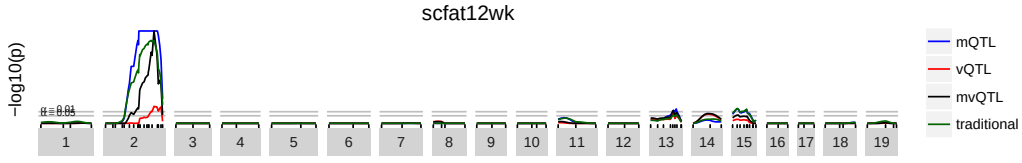
(a) Genome scan for bodyweight at twelve days



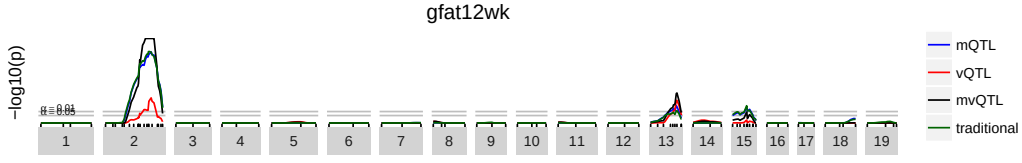
(b) Genome scan for bodyweight at three weeks



(c) Genome scan for bodyweight at six weeks



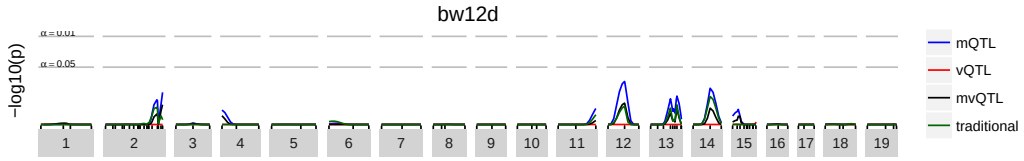
(d) Genome scan for subcutaneous fat pad thickness at twelve weeks



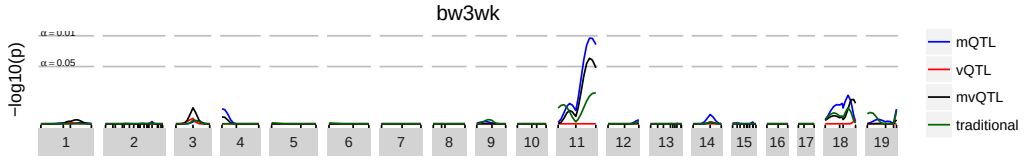
(e) Genome scan for gonadal fat pad thickness at twelve wees

Figure 2.14: Genome scans conducted with the DGLM, without accounting for effects of sex and father on variance, shown by simulation to be identical to Cao's tests (Figure 2.13, Table 2.3, Table 2.4, and Table 2.5).

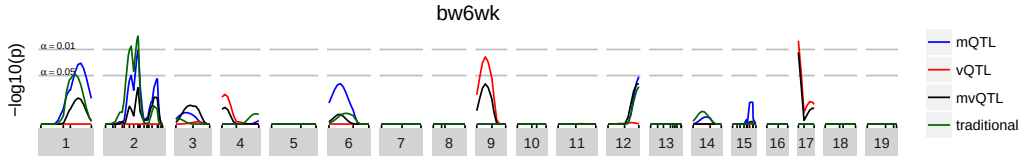
2.6.9 DGLM Tests for All Phenotypes with BVH



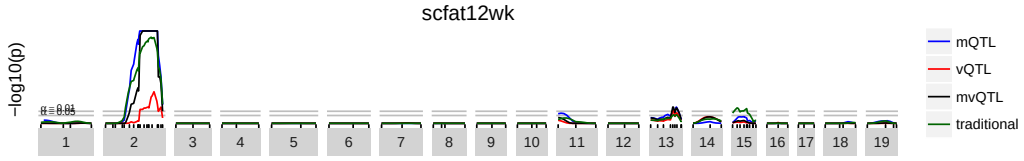
(a) Genome scan for bodyweight at twelve days



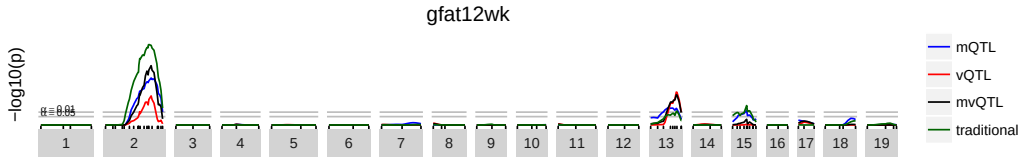
(b) Genome scan for bodyweight at three weeks



(c) Genome scan for bodyweight at six weeks



(d) Genome scan for subcutaneous fat pad thickness at twelve weeks



(e) Genome scan for gonadal fat pad thickness at twelve wees

Figure 2.15: Genome scans conducted with the DGLM, accounting for effects of sex and father on variance.

CHAPTER 3

Mean-Variance QTL Mapping Identifies Novel QTL for Circadian Activity and Exploratory Behavior in Mice [1]

3.1 Introduction

Here we demonstrate, with two real data examples available from the Mouse Phenome Database (Bogue et al., 2015), that QTL mapping using the DGLM, which we term "mean-variance QTL mapping" largely replicates the results of standard QTL mapping and detects additional QTL that the traditional analysis does not.

3.2 Statistical Methods

3.2.1 Traditional QTL mapping based on the standard linear model (SLM)

The traditional approach to mapping a quantitative trait in an experimental cross with no population structure (*e.g.* an F2 intercross or backcross) involves fitting, at each locus in turn, a linear model of the following form. Letting y_i denote the phenotype value of individual i, this phenotype is modeled as

$$y_i \sim \mathrm{N}(m_i,\, \sigma^2)\,,$$

where σ^2 is the residual variance, and the expected phenotype mean, m_i, is predicted by effects of QTL genotype and, optionally, effects of covariates. In the reanalyses performed here, m_i is modeled to include a covariate of sex and additive and dominance effects of QTL genotype, that is,

$$m_i = \mu + \mathrm{sex}_i\beta_{\mathrm{sex}} + a_i\beta_a + d_i\beta_d\,,$$

[1] This chapter has been adapted from a manuscript submitted to G3. The citation will be: Corty, RW., Kumar, V., Tarantino, L., Takahashi, JS., Valdar, W., 2018. Mean-Variance QTL Mapping Identifies Novel QTL for Circadian Activity and Exploratory Behavior in Mice. G3: Genes, Genomes, Genetics.

where μ is the intercept, β_{sex} is the sex effect, with sex_i indicating (0 or 1) the sex of individual i, and β_a and β_d are the additive and dominance effects of a QTL whose genotype is represented by a_i and d_i defined as follows: when QTL genotype is known, a_i is the count (0,1,2) of one parental allele, and d_i indicates heterozygosity (0 or 1); when QTL genotype is inferred based on flanking marker data, as is done here, a_i and d_i are replaced by their corresponding probabilistic expectations (Haley and Knott, 1992; Martínez and Curnow, 1992). The evidence for association at a given putative QTL is based on a comparison of the fit of the model above with that of a null model that is identical except for the QTL effects being omitted. These models and their comparison we henceforth refer to as the standard linear model (SLM) approach.

3.2.2 Mean-variance QTL mapping based on the double generalized linear model (DGLM)

The statistical model underlying mean-variance QTL mapping, the double generalized linear model (DGLM; Smyth 1989 and Rönnegård and Valdar 2011), elaborates the SLM approach by modeling a potentially unique value of σ^2 for each individual, as

$$y_i \sim \mathrm{N}(m_i,\, \sigma_i^2)\,,$$

where m_i has the same meaning as in the SLM, but now σ_i^2 is linked to its own linear predictor v_i as

$$\sigma_i = \exp(v_i),$$

where the exponentiation ensures that σ_i is always positive, though v_i is unconstrained. The linear predictors for m_i and v_i are modeled as

$$\begin{aligned} \text{mean:} \quad & m_i = \mu + \text{sex}_i\beta_{\text{sex}} + a_i\beta_a + d_i\beta_d \\ \text{log(variance):} \quad & v_i = \mu_v + \text{sex}_i\gamma_{\text{sex}} + a_i\gamma_a + d_i\gamma_d \end{aligned} \tag{3.1}$$

where μ, a_i, d_i, sex_i, and the β's are as before, μ_v is an intercept representing the (log of the) "baseline" residual variance, and γ_a, γ_d, and γ_{sex} are the effects of the QTL and covariates on v_i.

The evidence for a QTL association is now defined through three distinct model comparisons, corresponding to testing for an mQTL, a vQTL, or an mvQTL. In each case, the fit of the "full"

model in Equation 3.1 is compared with that of a different fitted null: for the mQTL test, the null model omits the QTL effects on the mean (*i.e.*, $\beta_a = \beta_d = 0$); for the vQTL test, the null model omits the QTL effects on the variance (*i.e.*, $\gamma_a = \gamma_d = 0$); and for the mvQTL test, the null model omits QTL effects on both mean and variance (*i.e.*, $\beta_a = \beta_d = \gamma_a = \gamma_d = 0$). These tests are detailed in chapter 2.

3.2.3 Genomewide significance and FWER-adjusted p-values

The model comparisons described above constitute the SLM test and the three DGLM-based tests and each produces a likelihood ratio (LR) statistic. These LR statistics are converted to p-values that are adjusted for the family-wise error rate (FWER) across loci, *i.e.*, p-values on the scale of genomewide significance. This adjustment is performed separately for each test by calculating an empirical distribution for the LR statistic under permutation, much in the spirit of Churchill and Doerge (1994) but with some modifications, namely that different tests have differently structured permutations. Briefly, let G_i be the full set of genetic information for individual i, that is, the genotypes or genotype probabilities across all loci. For the SLM and mvQTL tests, we define a permutation as randomly shuffling the G_i's across individuals; for the mQTL test, the permutations apply this shuffle only to the genotype information in the full model's mean component; for the vQTL test, the permutations apply the shuffle only to the genotype information in the full model's variance component. For a given test, for each permutation we calculate LR statistics across the genome and record the maximum; the maxima of over all permutations is fitted to a generalized extreme value distribution, and the upper tail probabilities of this fitted distribution are used to calculated the FWER-adjusted p-values for the LR statistics in the unpermuted data [see Dudbridge and Koeleman 2004, and, *e.g.*, Valdar et al. 2006; more details in chapter 2]. An FWER-adjusted p-value can be interpreted straightforwardly: it is the probability of observing an association statistic this large or larger in a genome scan of a phenotype with no true associations.

3.2.4 Data Availability

All data and scripts used to conduct the analyses presented here and plot results are archived in a public, static repository at with DOI: `10.5281/zenodo.1187195`. Specifically, the raw data files are:

- `1_Kumar2014.csv` The phenotype and genotype data from Kumar et al. (2013) that was reanalyzed. This dataset is also available from the Mouse Phenome Database (Bogue et al., 2015) at `https://phenome.jax.org/projects/Kumar1`.
- `4_Bailey2008.csv` The phenotype and genotype data from Bailey et al. (2008) that was reanalyzed. This dataset is also available from the Mouse Phenome Database at `https://phenome.jax.org/projects/Bailey1`.
- `9_actogram_data` The raw data on circadian activity from Kumar et al. (2013) that was used to plot actograms

The analysis and plotting scripts are:

- `2_run_Kumar_scans.R` This script runs genome scans with `R/qtl` and `R/vqtl` on the data from Kumar et al. (2013).
- `3_plot_Kumar_scans.R` This script plots the results of the reanalysis of Kumar et al. (2013).
- `5_run_Bailey_scans.R` This script runs genome scans with `R/qtl` and `R/vqtl` on the data from Bailey et al. (2008).
- `6_plot_Bailey_scans.R` This script plots the results of the reanalysis of Bailey et al. (2008).
- `7_prune_big_files.R` This script strips out redundant information from the results to make the file size smaller to share more easily online.
- `8_power_simulations.R` This script runs the power simulation comparing the DGLM to the SLM at the QTL identified in the Kumar reanalysis.

The results of running the analysis and plotting scripts are:

- `Kumar_scans_1000_perms.RDS` This file contains the results of the reanalysis of Kumar et al. (2013).
- `Bailey_scans_1000_perms.RDS` This file contains the results of the reanalysis of Bailey et al. (2008).

- `Kumar_plots` This directory contains the figures generated by `3_plot_Kumar_scans.R` (Figures 3.1, 3.2, and 3.6).

- `Bailey_plots` This directory contains the figures generated by `6_plot_Bailey_scans.R` (Figures 3.4, 3.5, 3.10, 3.11, 3.12, and 3.13)

3.3 Reanalysis of Kumar et al. Reveals a new mQTL for Circadian Wheel Running Activity

3.3.1 Summary of Original Study

Kumar et al. (2013) intercrossed C57BL/6J and C57BL/6N, two closely-related strains of C57BL6 that diverged in 1951, approximately 330 generations ago. Due to recent coancestry of the parental strains, this cross is termed a "reduced complexity cross", and their limited genetic differences ensure that any identified QTL region can be narrowed to a small set of variants bioinformatically. The intercross resulted in 244 F2 offspring, 113 female and 131 male, which were tested for acute locomotor response to cocaine (20mg/kg) in the open field. One to three weeks following psychostimulant response testing, the mice were tested for circadian wheel running activity.

Analysis of wheel running data was carried out using ClockLab software v6.0.36. For calculation of activity 20 day epoch in DD was used in order to have standard display between actograms. Analysis of other circadian measures such as period (tau) or amplitude were carried out using methods previously described (Shimomura et al., 2001). All animal protocols were approved by the Institutional Animal Care and Use Committee (IACUC) of the University of Texas Southwestern Medical Center

Traditional QTL mapping with the SLM, reported in Kumar et al. (2013), detected a single large-effect QTL for cocaine-response traits on chromosome 11, but no QTL for circadian activity. A later study by another group nonetheless observed that the circadian activity of the two strains showed significant differences (Banks et al., 2015). Though a difference in strain mean is not required for a successful F2 intercross mapping experiment in general, it is a reassuring observation in the context of a reduced complexity cross, where genetic differences are scant.

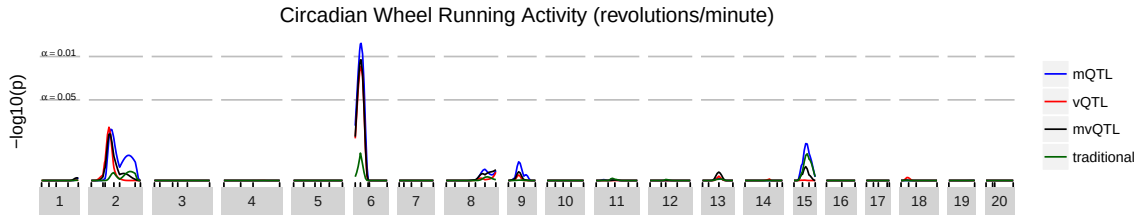


Figure 3.1: Genome scan for Kumar et al. circadian wheel running activity. The horizontal axis shows chromosomal location and the vertical axis shows FWER-controlling p-values for the association between each genomic locus and circadian wheel running activity.

3.3.2 Reanalysis with traditional QTL mapping and mean-variance QTL mapping

For the cocaine response traits, traditional QTL mapping and mean-variance QTL mapping gave results that were nearly identical to the originally-published analysis in Kumar et al. (2013) (Figure 3.6).

For the circadian wheel running activity trait, however, traditional QTL mapping identified no QTL (Figure 3.1 in green) but mean-variance QTL mapping identified one QTL on chromosome 6 (Figure 3.1 in blue, black, and red). In this case, all three tests were statistically significant, but the most significant was the mQTL test (blue), so we discuss it as an mQTL. The most significant genetic marker was rs30314218 on chromosome 6, at 18.83 cM, 40.0 Mb, with a FWER-controlling p-value of 0.0063. The mQTL explains 8.4% of total phenotype variance by the traditional definition of percent variance explained (*e.g.*, Broman and Sen 2009).

3.3.3 Understanding the Novel QTL

Though they test for the same pattern, the mQTL test of mean-variance QTL mapping identified a QTL where the traditional QTL test did not. This discordance may arise when there is variance heterogeneity in the mapping population. In this case, mice homozygous for the C57BL/6N allele at the mQTL have both higher average wheel running activity and lower residual variance in wheel running activity than mice with other genotypes (Figure 3.2a).

The identification of this QTL by mean-variance QTL mapping but not traditional QTL mapping can be understood by contrasting how the DGLM and SLM fit the data at this locus.

For the SLM, a single value of the residual standard deviation σ is estimated for all mice. Approximately 25% of the mice are homozygous for the C57BL/6N allele, so σ is estimated mostly

based on heterozygous mice and homozygous C57BL/6J mice. The SLM estimates $\hat{\sigma} = 7.83$, a slight underestimate for some genotype-sex combinations, and a drastic overestimate for the homozygous C57BL/6N of both sexes (Figure 3.2b). With σ overestimated for the C57BL/6N homozygotes, the addition of a locus effect to the null model results in only a limited increase in the likelihood, one that could reasonably be caused by chance alone. For the DGLM, six different values of σ are estimated, one for each genotype-sex combination (Figure 3.2b). With an better-estimated (lower) $\widehat{\sigma}$ for the C57BL/6N homozygotes, the addition of the locus effect to the null model results in a greater increase in the likelihood, one that is very unlikely due to chance alone.

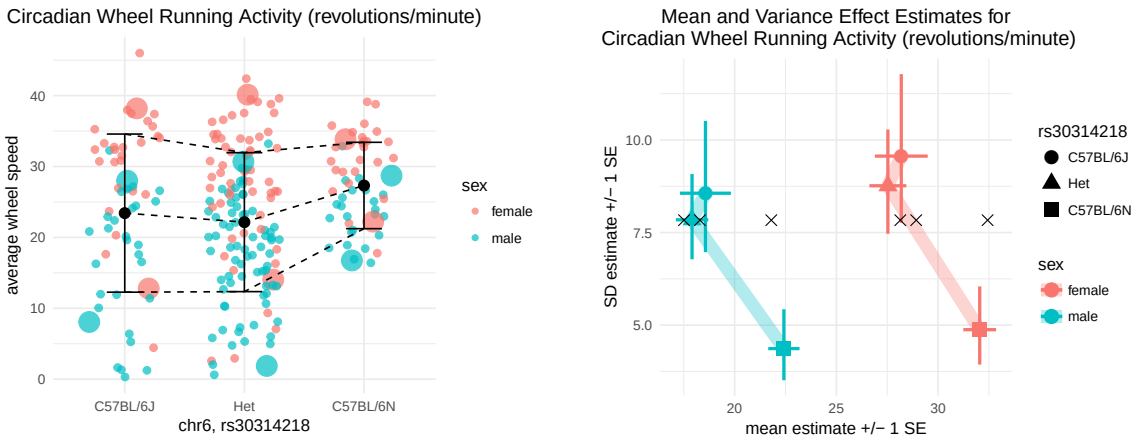
A simulation based on the estimated coefficients shows that at a false positive rate of 5×10^{-4}, relevant for genome-wide significance testing, the SLM has 61% power to reject the null at this locus and the DGLM has 90% power (See file `8_power_simulations.R`).

3.3.4 Variant Prioritization

Reduced complexity crosses allow variant prioritization to proceed quickly because of the number of segregating variants is small. Using 1000 nonparametric bootstrap resamples, the QTL interval was estimated as 13.5-23.5 cM (90% CI), which translates to physical positions of 32.5 - 48.5 Mb using Mouse Map Converter's sex averaged Cox map (Cox et al., 2009). Since this interval contains no genes or previously identified QTL shown to regulate circadian rhythms, we prioritized candidates by identifying variants between C57BL/6J and C57BL/6NJ based on Sanger mouse genome database (Keane et al., 2011; Simon et al., 2013), which yielded 463 SNPs, 124 indels, and 3 structural variants (Table 3.1).

Of these variants, none of the indels or structural variants were nonsynonymous. Two SNPs were predicted to lead to missense changes (T to A at position 6: 39400456 in *Mkrn1*, and A to A/C at 6:48486716 in *Sspo*). The variant in Sspo was a very low confidence call and therefore likely a false positive.

The *Mkrn1* (makorin ring finger protein 1) variant is a mutation in C57BL/6J that changes a highly conserved (Figure 3.8 and Figure 3.9) tyrosine to asparagine. It was determined to be the best candidate variant in the QTL interval. The Mkrn1 protein is a ubiquitin E3 ligase with zinc finger domains with poorly defined function (Kim et al., 2005). It is expressed at low levels widely in the brain according to Allen Brain Atlas and EBI Expression Atlas (Kapushesky et al., 2009; McWilliam



(a) (b)

Figure 3.2: (a) Average wheel speed (revolutions/minute) of all mice. It is visually apparent that female mice had higher circadian wheel running activity than male mice and that mice that homozygous for C57BL/6N had higher circadian wheel running activity and less intra-genotype variation. Large dots indicate the mice whose activity is shown in actogram form (Figure 3.3). (b) Predicted mean and variance of mice according to sex and allele at the QTL. What was visually apparent in (a) is captured by the DGLM. The estimated parameters relating to mice that are homozygous for the C57BL/6N allele imply a higher expected value and a lower residual variance than the other two genotype groups. Black x's indicate the estimates from the SLM, very similar to the DGLM estimates in the horizontal (mean) axis, but homogeneous in the vertical (variance) axis.

et al., 2013; Allen Institute for Brain Science, 2015; McWilliam et al., 2013). Functional analysis will be necessary to experimentally confirm that this variant in Mkrn1 is indeed the causative mutation that led, in a dominant fashion, to the decreased expected value and increased variance of circadian wheel running activity observed in mice with at least one copy of the C57BL/6J haplotype in the QTL region in this study.

3.4 Reanalysis of Bailey et al. Identifies a new vQTL for Rearing Behavior

3.4.1 Summary of Original Study

Bailey et al. (2008) intercrossed C57BL/6J and C58/J mice, two strains known to be phenotypically similar for anxiety-related behaviors, as a control cross for an ethylnitrosourea mutagenesis mapping study. The intercross resulted in 362 F2 offspring, 196 females and 166 males. Six open-field behaviors were measured at approximately 60 days of age in a 43cm by 43cm by 33cm white arena for ten minutes. All phenotypes were transformed with the rank-based inverse normal

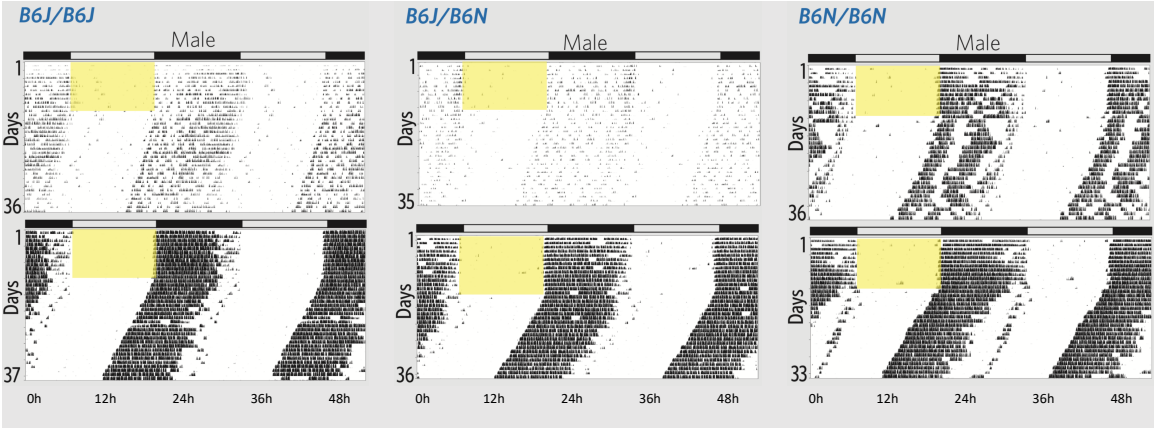


Figure 3.3: Double-plotted actograms illustrate the variation in wheel running activity of male mice based on their genotype at rs30314218. On reading a single actogram: An actogram illustrates the activity of a single mouse over the course of an experiment. Each day of the experiment is represented by a histogram, with bin width of six minutes. Histograms are stacked vertically. Additionally, each day is shown twice (repeated horizontally) so that there is no time of day that is illegible due to the plot edges. Yellow box indicates when lights were on. On reading this six-actogram plot: Recall that the DGLM estimates a unique mean and standard deviation (SD) for each genotype. The mice whose actograms are shown here had an activity level that is one genotype-specific SD greater than (top) or less than (bottom) the genotype-specific mean. The difference between the two is much less in the C57BL/6N homozygotes than in the other genotypes, reflecting the decreased phenotype variance amongst C57BL/6N homozygotes. The animals shown in this figure are marked with large blue circles in Figure 3.2a. A larger figure that also includes female mice as well as the ID's of all plotted mice are in the supplement (Figure 3.7 and Table 3.2).

location	indel	SNP	SV	Total
exon, missense	–	2	–	2
intron, splice region	1	–	–	1
intron, nonsynonymous	57	246	–	303
intron, synonymous	–	1	–	1
3' UTR	–	3	–	3
upstream	6	29	–	35
downstream	7	20	–	27
intergenic	53	161	–	214
unclassified	–	1	3	4

Table 3.1: Genetic Variants in QTL interval for circadian wheel running activity

transform to limit the influence of outliers. The authors reported 7 QTL spread over five of the six measured traits, but none for rearing behavior.

3.4.2 Reanalysis with SLM and DGLM

SLM-based QTL analysis replicated the originally-reported LOD curves. Significance thresholds to control FWER at 0.05 were estimated by 10,000 permutations, using the method described in the original publication, but found to be meaningfully higher than the originally-reported thresholds. Of the 7 originally-reported QTL, 3 exceeded the newly-estimated thresholds (Figure 3.10).

The DGLM-based reanalysis was initially conducted with the rank-based inverse normal transformed phenotypes, to maximize the comparability with the original study. This reanalysis largely replicated the results of the SLM-based analysis and identified a statistically-significant vQTL for rearing behavior on chromosome 2 (Figure 3.4 and Figure 3.10). The top marker under the peak was at 38.6cM and 65.5Mb.

There are well-known and well-founded concerns that inappropriate scaling of phenotypes can produce spurious vQTL (Rönnegård and Valdar, 2012; Sun et al., 2013; Shen and Ronnegard, 2013). Therefore, the rearing phenotype was analyzed under a variety of additional transforms: none, log, square root, and $\frac{1}{4}^{\text{th}}$ power (the transformation recommended by the Box-Cox procedure). Because the trait is a "count" and a positive mean-variance correlation was observed, the trait was further analyzed with a Poisson double generalized linear model with its canonical link function (log). In all cases, the same genomic region on chromosome 2 was identified as a statistically significant vQTL ($p < 0.01$) (Figure 3.11, Figure 3.12, and Figure 3.13). Though all transformations yielded similar results, we highlight the Box-Cox transformed analysis recommended for transformation selection in Rönnegård and Valdar (2011).

3.4.3 Understanding the Novel QTL

In this case, the DGLM-based analysis identified a vQTL, a pattern of variation across genotypes not targeted by traditional, SLM-based, QTL analysis. The phenotype values, when stratified by genotype at the top locus, illustrate clear variance heterogeneity (Figure 3.5a). The effects and their standard errors estimated by the DGLM fitted at the top locus corroborate the impression from simply viewing the data, that the locus is a vQTL but not an mQTL (Figure 3.5b).

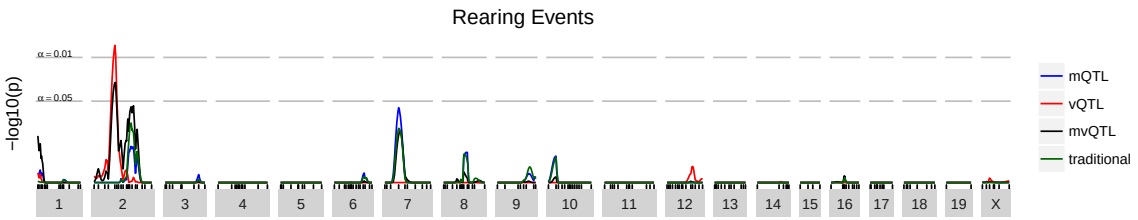
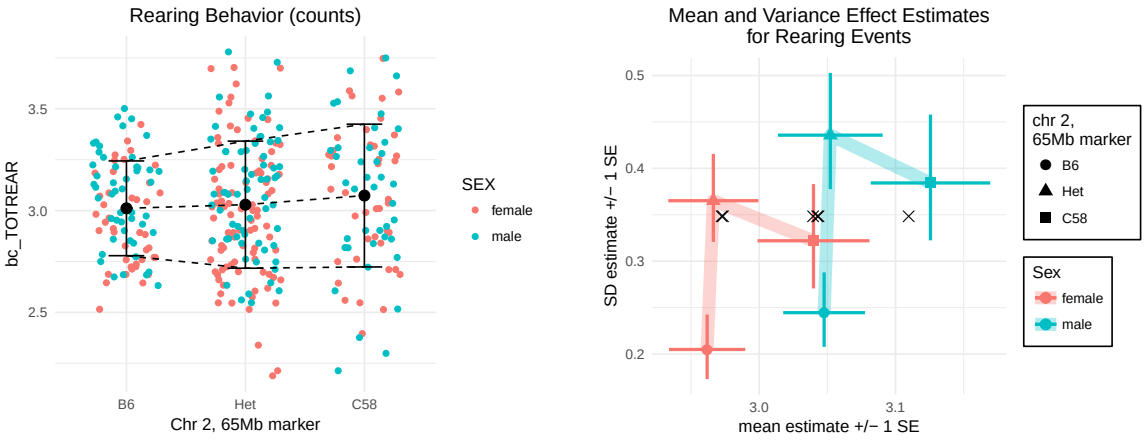


Figure 3.4: Genome scan for Bailey et al. rearing behavior. The x axis shows chromosomal location and the y axis shows FWER-controlling p-values for the association between each genomic locus and the Box-Cox transformed rearing behavior.



(a) (b)

Figure 3.5: (a) "Total Rearing Events", transformed by the Box-Cox procedure, stratified by sex and genotype at the top marker. (b) Predicted mean and variance of mice according to sex and allele at the top marker.

3.5 Discussion

We have demonstrated through two case studies that mean-variance QTL mapping based on the DGLM expands the range of QTL that can be detected, including both mQTL at loci that exhibit variance heterogeneity and vQTL. In an era where ever more complete and complex data on biological systems is becoming available, this modest elaboration of an existing approach represents a step toward the broader goal of characterizing the wide array of patterns of association between genotype, environment, and phenotype.

In the reanalysis of Kumar et al., mean-variance QTL mapping identified the same QTL as traditional, SLM-based QTL mapping for cocaine response traits and one novel mQTL for a circadian behavior trait. Such an mQTL would likely have been detected by a traditional QTL analysis with a larger mapping population: Through simulation, we estimated that the additional power to detect the mQTL was equivalent to the power increase that would have come from increasing the sample size by $\approx$100 mice, from 244 to $\approx$ 350 (See file `8_power_simulations.R`). Given the considerable effort and expense associated with conducting an experimental cross or expanding the size of the mapping population, there seems to be little to be gained by omitting a DGLM-based analysis.

In the reanalysis of Bailey et al., mean-variance QTL mapping identified a novel vQTL for an exploratory behavior. A vQTL such as this would not be detected by the traditional QTL analysis no matter how large the mapping population because the pattern is entirely undetectable by the SLM.

The identification of a vQTL raises important issues related to phenotype transformation and the interpretation of findings, but both are manageable, as we have illustrated here. The criticism that a spurious vQTL can arise as the result of an inappropriate transformation is based on the observation that when genotype means are unequal, there always exists a (potentially exotic) transformation that diminishes the extent of variance heterogeneity (Sun et al., 2013). Thus, any other transformation (including none at all) can be seen as inflationary toward variance heterogeneity. In this context, however, an "inappropriate transformation" leads not to the misclassification of a non-QTL as a QTL, but an mQTL as a vQTL.

To the extent that the goal of QTL mapping is to understand the genetic architecture of a trait, this criticism is valid and should be addressed by considering a wide range of transformations, alternative models, and parameterizations. To the extent that the goal of QTL mapping is to identify

genomic regions that contain genes and regulatory factors that influence a trait, we argue that such a misclassification is largely irrelevant. Whether we pursue bioinformatic follow-up to identify QTN in a region because it was identified as an mQTL or a vQTL need not change our downstream efforts.

In summary, we advocate for the use of mean-variance QTL mapping not as an additional flourish to consider after conducting an SLM-based QTL mapping effort, but rather as a drop-in replacement. This approach should not be too alien — when variance heterogeneity is absent, it simplifies to the well-known SLM-based approach. Full-featured software that implements this approach is described in chapter 4.

Lastly, we note an additional benefit conferred by mean-variance QTL mapping not discussed in depth here. Variance heterogeneity can also derive from factors acting in the "background", that is, arising from experimental or biological variables that are outside the main focus of testing but that nonetheless predict phenotypic variability and thereby inform the relative precision of one individual's phenotype over another. In the case studies presented here, the only background factor considered was sex. But, more generally, any factor that a researcher considers as a potentially important covariate that should be modeled can be included not only as a mean covariate (as with the SLM) but also as a variance covariate. In chapter 2, we describe how accommodating such background factors can deliver additional power to detect mQTL, vQTL, and mvQTL.

3.6 Additional Information

3.6.1 Additional Information on Kumar Reanalysis

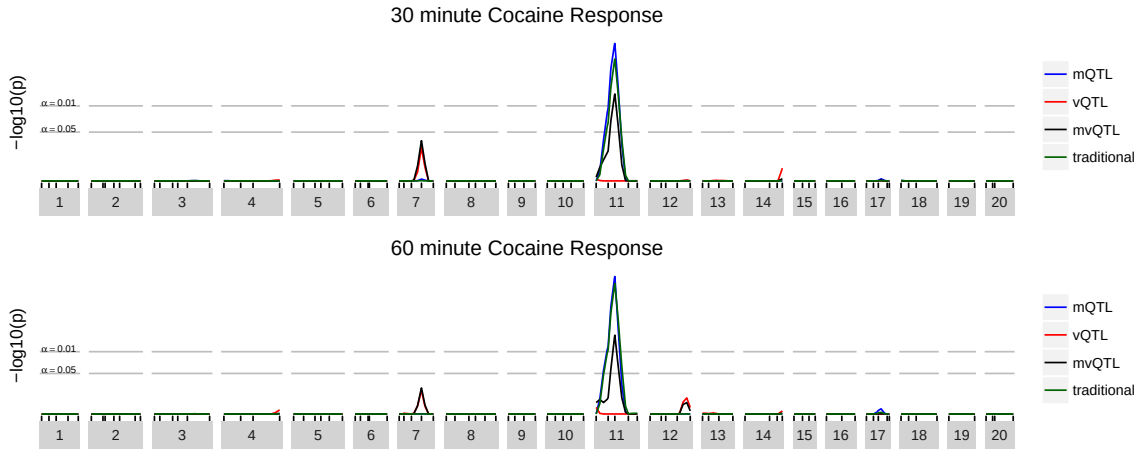


Figure 3.6: Replicated scans from Kumar et al. (2013)

Table 3.2: The characteristics of the mice plotted in Figure 3.3

genotype at rs30314218	sex	activity in the DD (rev/min)
6J	female	12.79
6J	female	38.20
6J	male	8.07
6J	male	27.99
Het	female	14.03
Het	female	40.13
Het	male	1.87
Het	male	30.68
6N	female	22.22
6N	female	33.85
6N	male	16.75
6N	male	28.71

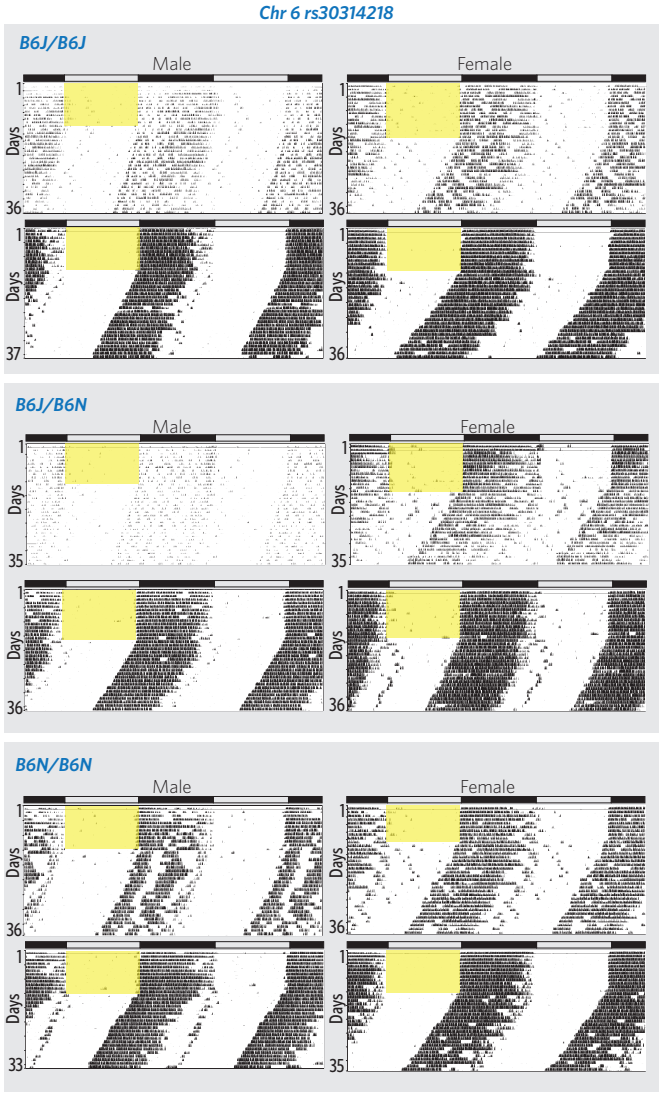


Figure 3.7: Actograms, similar to Figure 3.3, including female mice. The mice depicted here are highlighted with larger circles in Figure 3.2a.

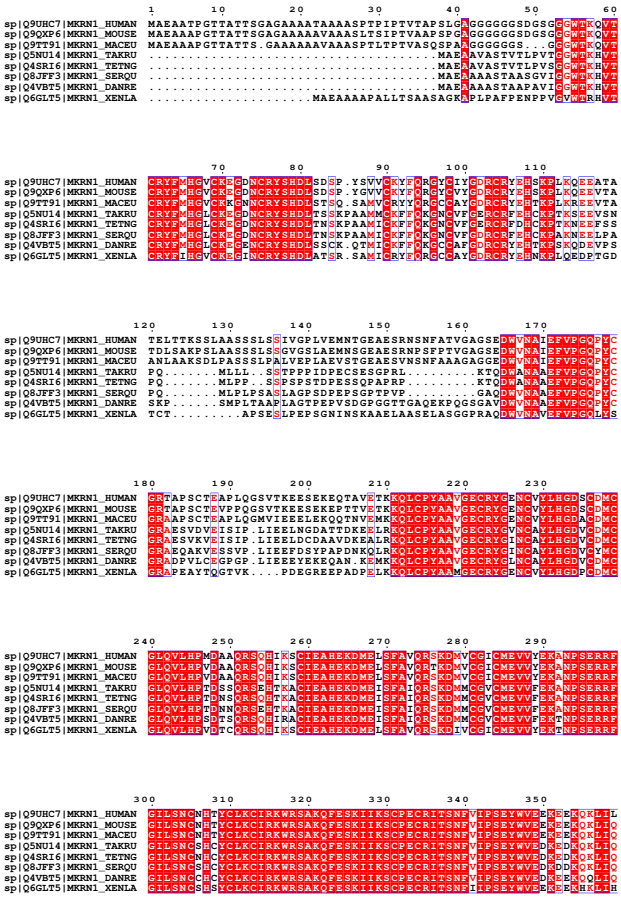


Figure 3.8: Page one of *Mkrn1* alignment. Note that the amino acid at position 346 is conserved across all species. See next page for species labels.

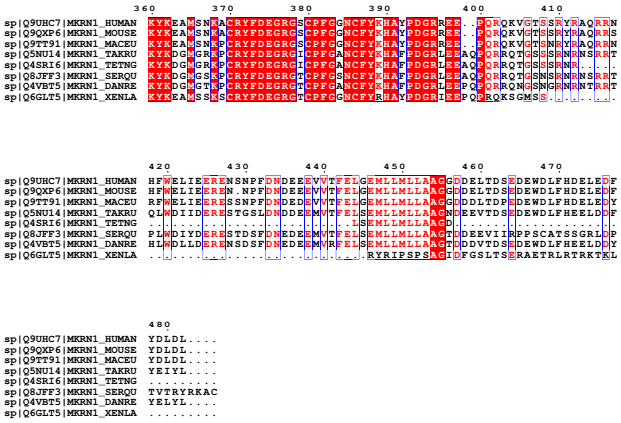


Figure 3.9: Page two of *Mkrn1* alignment.

3.6.2 Additional Information on Bailey Reanalysis

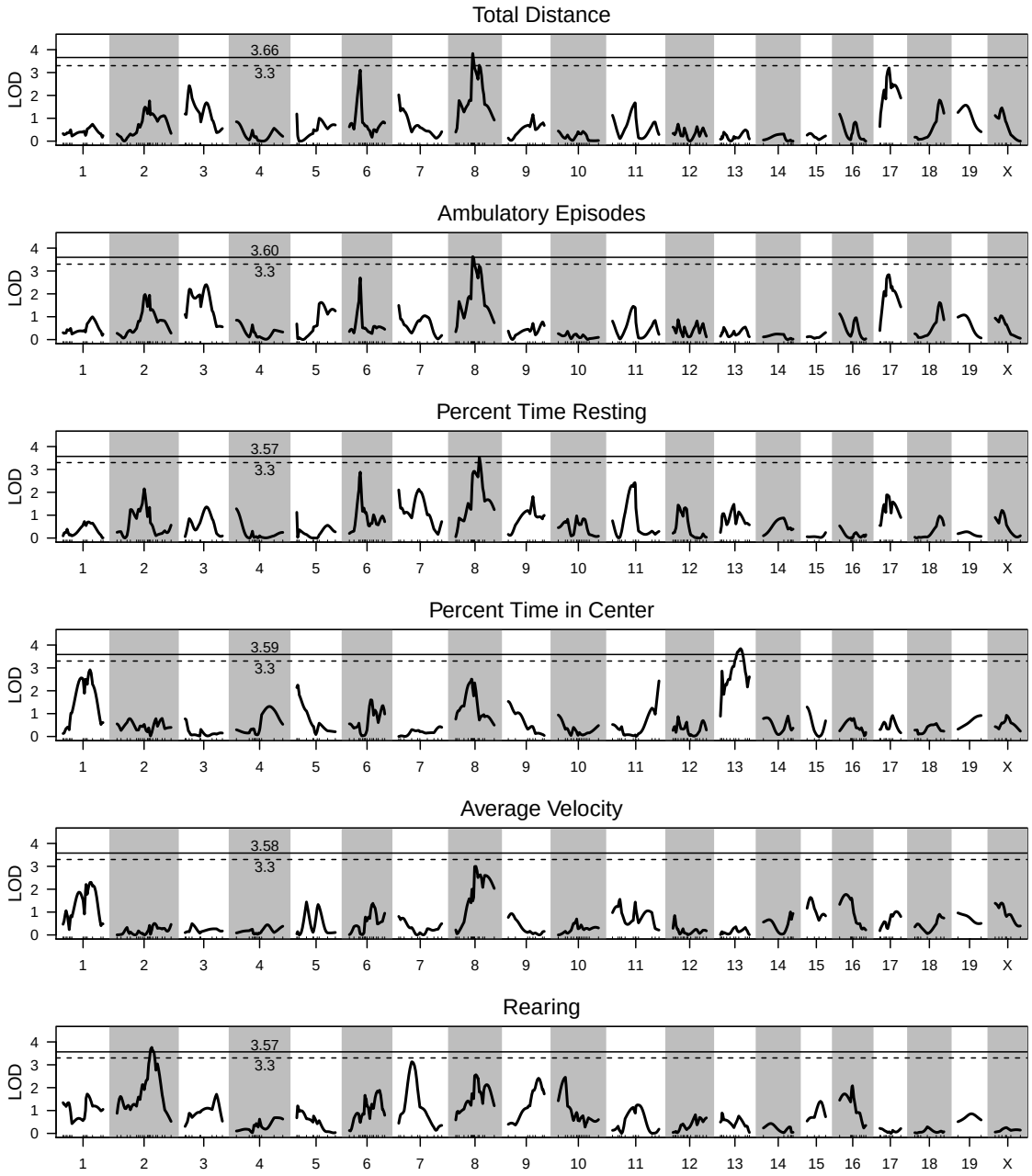


Figure 3.10: Replication of genome scans from original Bailey analysis. LOD curves are visually identical to originally-published LOD curves, but thresholds, estimated based on the described methods, are meaningfully higher.

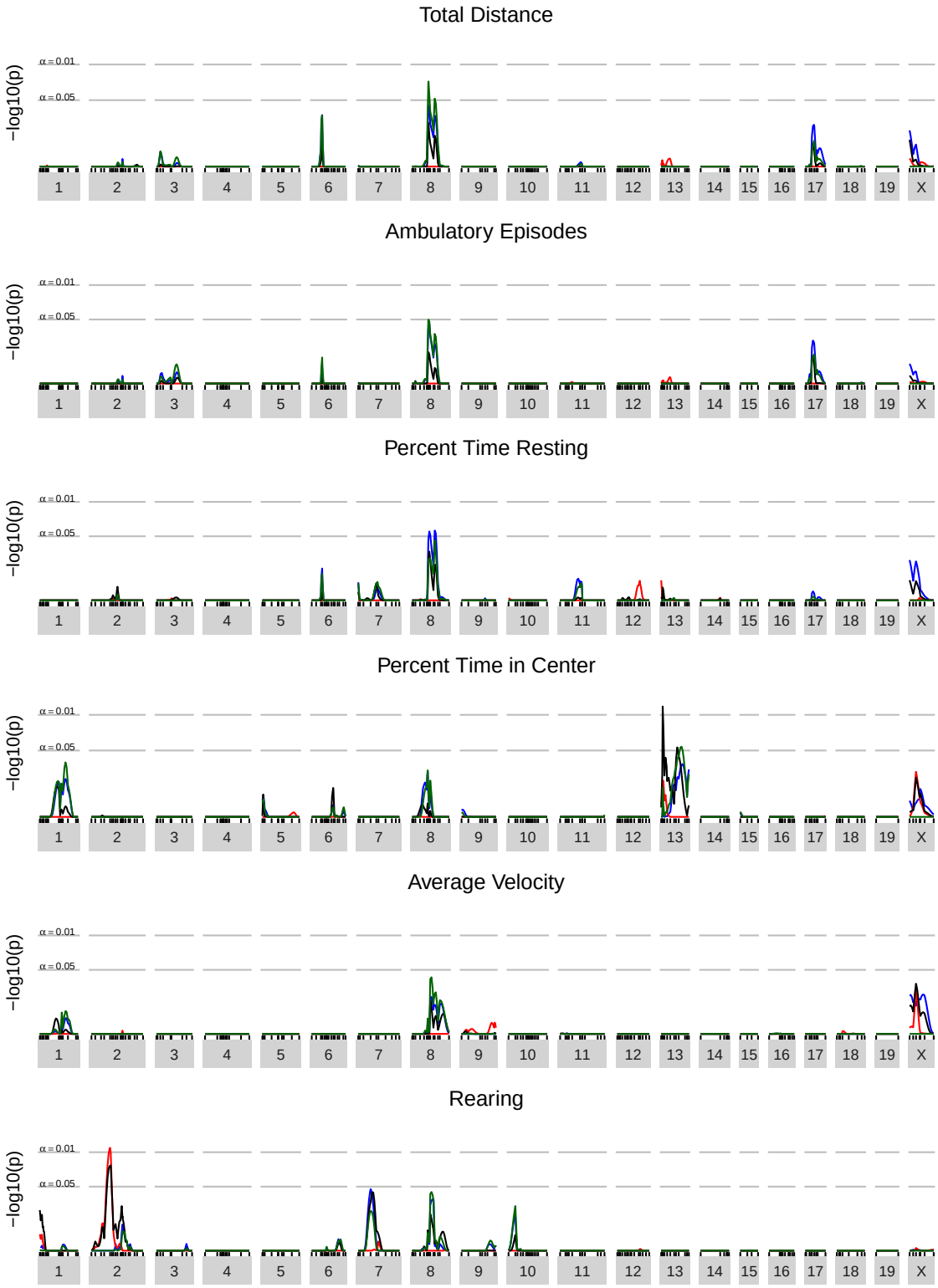


Figure 3.11: DGLM-based reanalysis of all traits measured in Bailey et al., all transformed by the rank-based inverse normal transform.

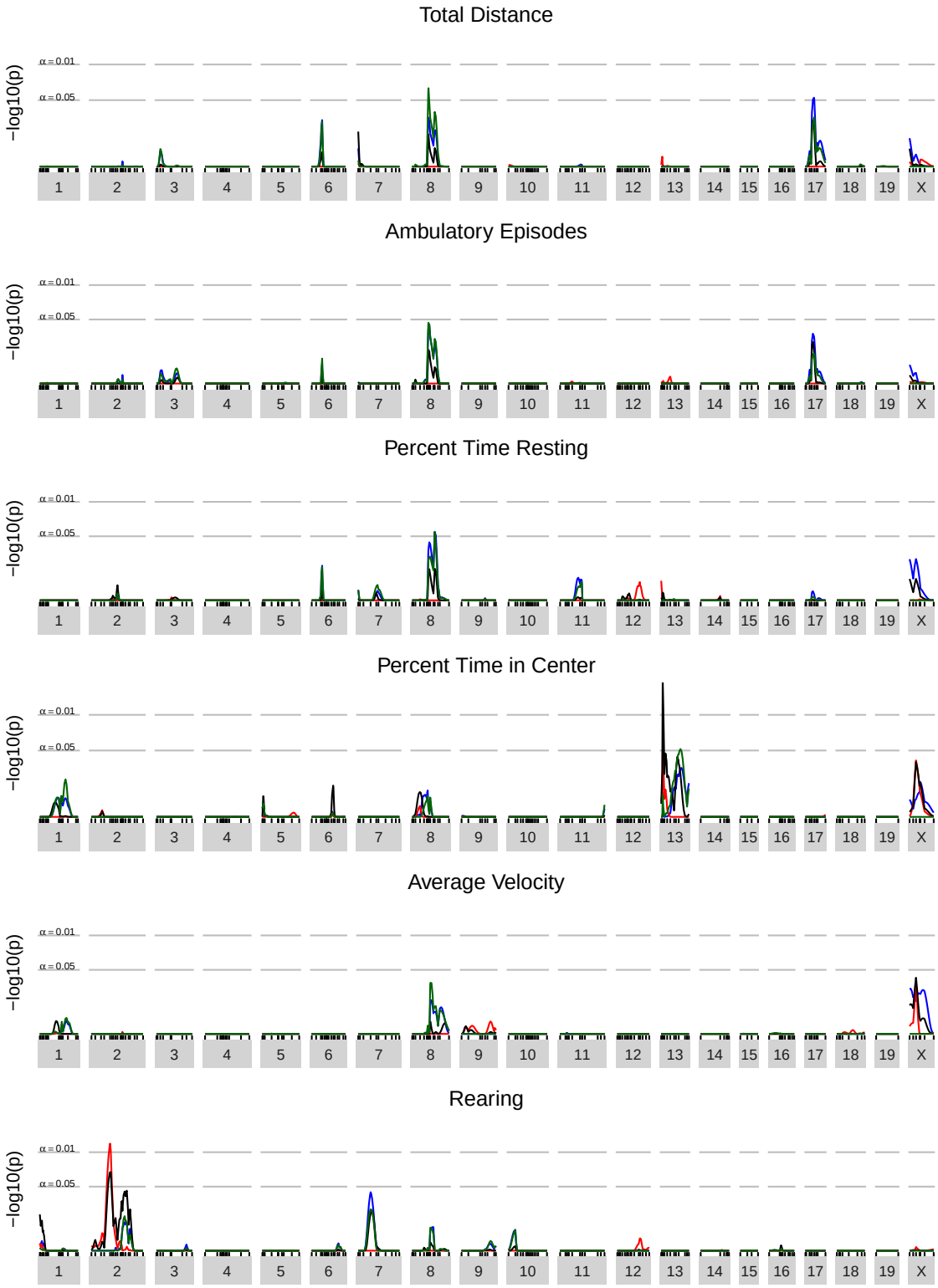
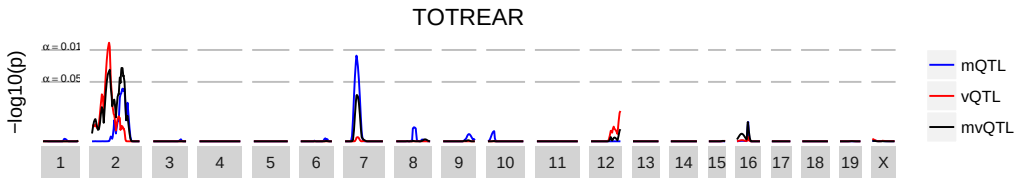
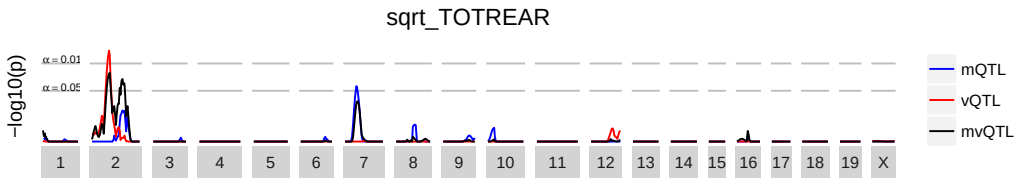


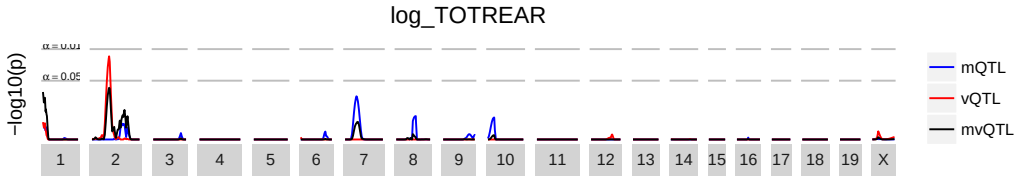
Figure 3.12: DGLM-based reanalysis of all traits measured in Bailey et al., all transformed by the Box-Cox transform. Box-Cox exponents were 1, 1, 0, 0.75, 0, 0.25, respectively.



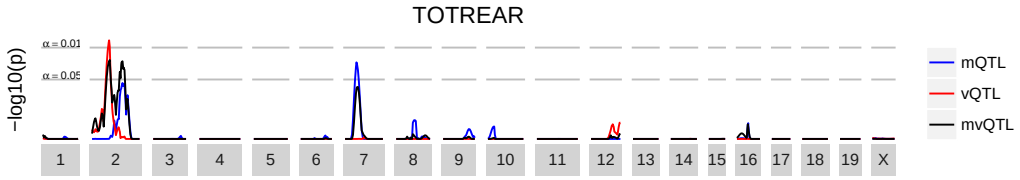
(a) “Rearing” trait analyzed with no transformation.



(b) “Rearing” trait analyzed with square root transformation.



(c) “Rearing” trait analyzed with log transformation.



(d) “Rearing” trait analyzed with Poisson generalized linear model.

Figure 3.13: vQTL for TOTREAR phenotype on chromosome 2 is consistent across various transforms.

CHAPTER 4

`vqtl`: An `R` package for Mean-Variance QTL Mapping [1]

4.1 Introduction

Here, we provide a practical guide to using the `R` package `vqtl`, which implements mean-variance QTL mapping. First, to generate illustrative data, we simulate an F2 intercross and four phenotypes: one phenotype determined entirely by random noise, and one with each of the three kinds of QTL. On each phenotype we then conduct a genome scan using standard approximations to interval mapping (Lander and Botstein, 1989; Martínez and Curnow, 1992), and mean-variance QTL mapping, which includes a test for mQTL, a test for vQTL, and a test for mvQTL. The association statistics of all four tests are then initially plotted in LOD score units, with drawbacks of this plotting unit discussed. Permutation scans are used to determine empirically adjusted p-values, and plotting in these units is shown to to make the results of the four tests more comparable. Last, we describe plots to communicate the effects that led to the detection of a QTL, and use the bootstrap to estimate its confidence interval.

4.2 Example data: Simulated F2 Intercross

To illustrate the use of the `vqtl` package, we first simulated an example F2 intercross using the popular `R/qtl` package (Broman et al., 2003), on which `vqtl` is based. This cross consisted of 200 male and 200 female F2 offspring, with 3 chromosomes of length 100 cM, each tagged by 11 equally-spaced markers and estimated genotype probabilities at 2cM intervals with `R/qtl`'s hidden Markov model. We then generated four phenotypes:

1. `phenotype1` consists only of random noise and will serve as an example of negative results for all tests.

[1] This chapter has been adapted from a manuscript submitted to G3. The citation will be: Corty, RW. and Valdar, W., 2018. vqtl: An R package for Mean-Variance QTL Mapping. G3: Genes, Genomes, Genetics.

2. `phenotype2` has an mQTL that explains 4% of phenotype variance at the center of chromosome one.

3. `phenotype3` has a vQTL at the center of chromosome two. This vQTL acts additively on the log standard deviation scale, and results in residual standard deviation of [0.8, 1, 1.25] for the three genotype groups.

4. `phenotype4` has an mvQTL at the center of chromosome three. This mvQTL has a mean effect that explains 2.7% of phenotype variance and a variance effect that acts additively on the standard deviation scale, resulting in residual standard deviation of [0.85, 1, 1.17] for the three genotype groups.

We additionally consider `phenotype1x` through `phenotype4x`, which have the same type of genetic effects as `phenotype1` through `phenotype4`, but have the additional feature that females have greater residual variance than males. All the same analyses and plots that are shown for `phenotype1` through `phenotype4` are shown for `phenotype1x` through `phenotype4x` in the appendix.

4.3 Scan the Genome

The central function for genetic mapping in package `R/qtl` is `scanone` (Broman et al., 2003). Analogously, the central function for mean-variance QTL mapping in package `vqtl` is `scanonevar`, building on an early version of `scanonevar` in package `qtl`. It takes three required inputs:

1. `cross` is an object that contains the genetic and phenotypic information from an experimental cross, as defined in package `qtl`.

2. `mean.formula` is a two-sided formula, specifying the phenotype to be mapped, the covariates to be corrected for, and the QTL terms to be fitted, with keywords `mean.QTL.add` and `mean.QTL.dom`

3. `var.formula` is a one-sided formula, specifying the variance covariates to be corrected for as well as the QTL terms to be fitted, using keywords `var.QTL.add` and `var.QTL.dom`.

For example, to scan a phenotype named `p1`, we run:

```
scanonevar(
 cross = test_cross,
 mean.formula = p1 ~ sex + mean.QTL.add + mean.QTL.dom,
 var.formula = ~ sex + var.QTL.add + var.QTL.dom
)
```

At each locus in turn, this function tests for the presence of an mQTL, a vQTL, and an mvQTL. The basis of these tests is a comparison between the fit of an alternative model of the form

$$\text{mean} = \text{covariate effects} + \text{locus effects}$$
$$\log(\text{variance}) = \text{covariate effects} + \text{locus effects}$$

with a null model that omits specific terms: for the mQTL test, the null model omits locus effects on phenotype mean; for the vQTL test, the null omits the locus effects on phenotype variance; and for the mvQTL test, the null omits locus effects on both mean and variance. (Note that the mQTL test in mean-variance QTL mapping is different from the traditional test: the traditional test does not have variance predictors of any kind in either null or alternative models.)

4.3.1 LOD scores and nominal p-values

Each type of test (mQTL, vQTL, and mvQTL) yields two association statistics: the LOD score, and the (nominal) p-value. The LOD is a raw measure of association equal to the base 10 logarithm of the likelihood ratio (LR) between the fitted alternative and null models. Higher values indicate greater association when considered across loci for the same type of test; but LOD scores between different types of tests, namely between mvQTL test vs either mQTL or vQTL tests, are not readily comparable. The p-value, which is comparable between different types of tests, transforms the LOD score to take account of the number of parameters being fit: it is calculated from the asymptotic distribution of $2\log_e(\text{LR})$ under the null model, namely the χ^2 distribution with degrees of freedom equal to the difference in the number of parameters between the alternative and null models.

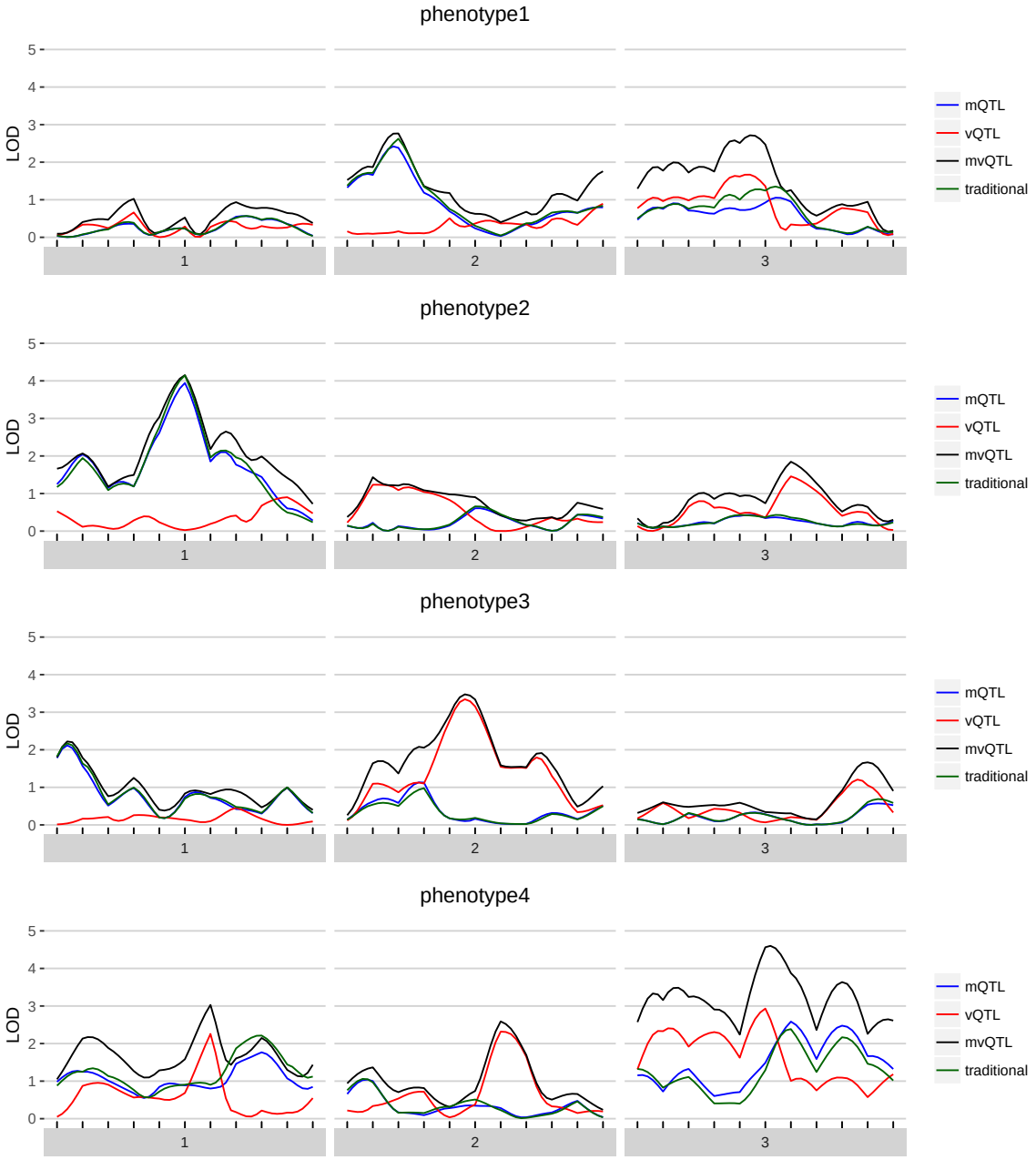


Figure 4.1: For each of the four simulated phenotypes, the genome scan shows the LOD score of each test — mQTL, vQTL, and mvQTL — in blue, red, and black, respectively. The traditional test is in green and globally similar to the mQTL test.

The p-values described above, however, are nominal: they do not take into account multiple testing across the genome. They also rely on asymptotic theory that assumes the underlying phenotype being residually normal; this may not always be the case and when violated will lead to inflated significance.

More robust p-values that are corrected for genomewide significance via control of the family-wise error rate (FWER) can be obtained empirically, through a permutation procedure described below.

4.3.2 Robust, genomewide-adjusted p-values

To calculate the empirical, FWER-controlled p-value of each test at each locus we advocate use of a permutation procedure (chapter 2). Like previous work on permutation-based thresholds for genetic mapping (Churchill and Doerge, 1994; Carlborg and Andersson, 2002), this procedure sidesteps the need to explicitly estimate the effective number of tests.

In brief, this approach involves conducting many genomes scans on pseudo-null data generated through permutation to maintain as much of the character of the data as possible, while breaking the tested phenotype-genotype association. Specifically, the design matrix of the QTL is permuted in the mean portion of the mQTL alternative model, the variance portion of the vQTL alternative model, and in both portions of the mvQTL alternative model.

For each test (mQTL, vQTL, and mvQTL), the highest observed test statistic is extracted from each permutation scan and the collection of statistics that results is used to fit a generalized extreme value (GEV) density (Stephenson, 2002; Dudbridge and Koeleman, 2004; Valdar et al., 2006). The observed LOD scores from the genome scan are then transformed by the cumulative distribution function of the extreme value density to estimate the FWER-controlling p-values. This approach is implemented in the function, `scanonevar.perm`, which requires two inputs:

1. `sov` is the `scanonevar` object, the statistical significance of which will be assessed through permutation.

2. `n.perms` is the number of permutations to conduct.

The object returned by `scanonevar.perm` is a `scanonevar` object with two additional pieces of information: an empirical p-value for each test at each locus and the per-permutation maxima

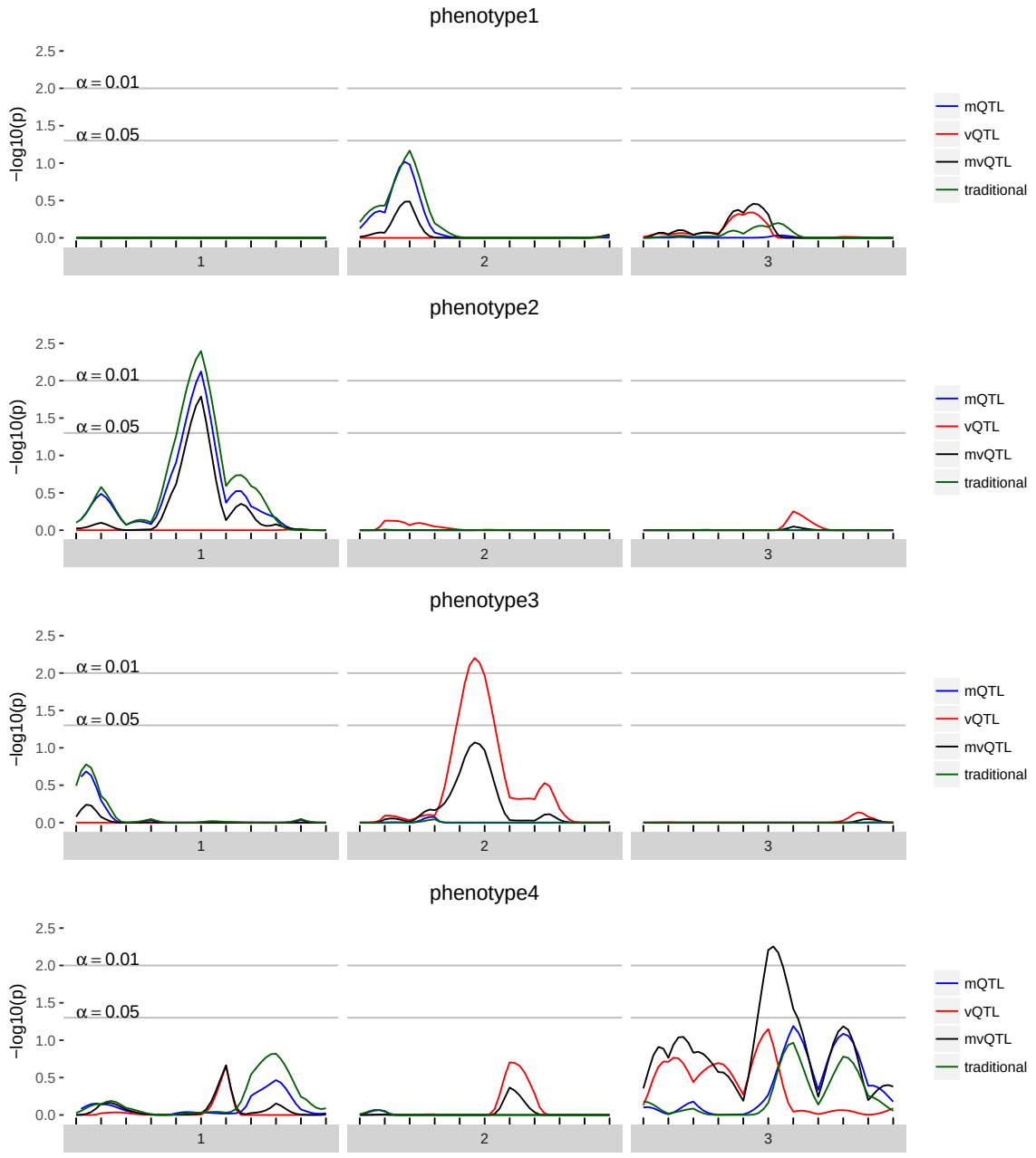


Figure 4.2: For each of the four simulated phenotypes, the genome scan shows the -log10 of the FWER-corrected p-value for each test — mQTL, vQTL, and mvQTL — in blue, red, and black, respectively. The traditional test is in green and globally similar to the mQTL test. A value of 2 implies that the quantity of evidence against the null is such that we expect to see this much or more evidence once per hundred phenotypes no QTL.

that were used to calculate those p-values. These FWER-corrected p-values are straightforwardly interpretable: $p = 0.05$ for a specific test at a specific locus implies that in 5% of similar experiments where there is no true genotype-phenotype association, we would expect to observe some locus with this much or more evidence of association in this test.

Accurate estimation of the FWER-controlled p-values requires many permutation scans: traditionally recommended is 1,000 (*e.g.*, Churchill and Doerge 1994; Carlborg and Andersson 2002), although the efficiency gain of using the GEV rather than raw quantiles means that fewer may be adequate in practice (Valdar et al., 2006). These permutation scans can be run on multiple processors by specifying the optional `n.cores` argument, which defaults to the total number of cores on the computer minus 2. On an Intel Core i5, running 100 permutations on this dataset takes about five minutes. When many phenotypes are studied, or if faster runtimes are needed, these permutation scans can be broken into groups with different values for `random.seed`, run on separate computers, and combined with the `c` function. This function combines the permutations from all the inputted scans, re-estimates the extreme value density, re-evaluates the observed LOD scores in the context of new extreme value density, and returns a new `scanonevar` object with more precisely estimated empirical p-values.

4.3.3 Reporting and plotting genome scans

The results of `scanonevar` can be plotted by calling `plot` on the `scanonevar` output object. This produces a publication-quality figure that shows the association of the phenotype for each location in the genome as different colors for type of test, with y-axis scale being specified by the user, via option `plotting.units` as the LOD (Figure 4.1), nominal p-value, or, provided permutations have been run, empirical, FWER-controlling p-value (Figure 4.2). Of the available y-axis scales, we recommend using the FWER-controlled p-values since this scale puts all tests on a level-footing (unlike the LOD), and allows direct identification of genomewide significance and thereby relevance (unlike the nominal p-value).

Calling `summary` on the output of `scanonevar` produces a summary of how the scan was conducted and what the results were.

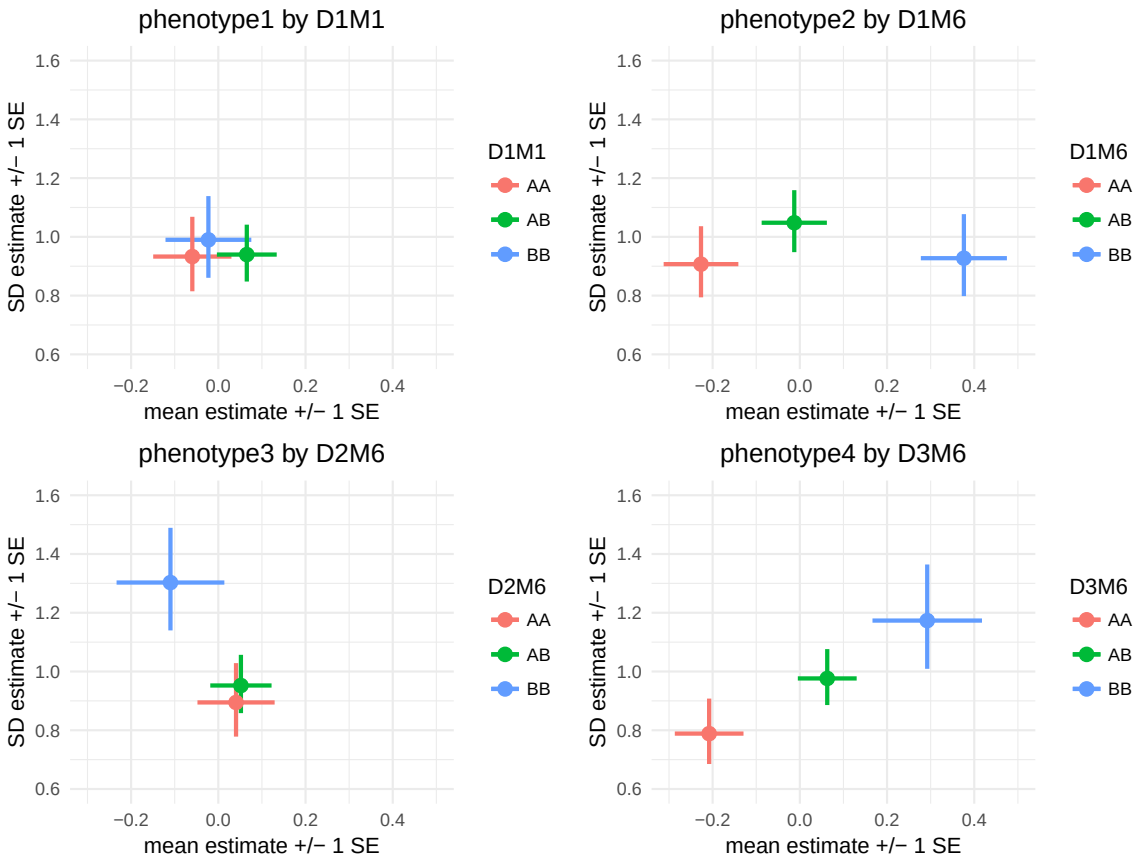


Figure 4.3: `mean_var_plots` show the estimated genotype effects at a locus with mean effects on the horizontal axis and variance effects on the vertical axis. Horizontal lines indicate standard errors for mean effects and vertical lines indicate standard errors for variance effects. For `phenotype1`, the pattern of overlapping estimates and standard errors is consistent with the fact that there are no genetic effects, and the p-value was not statistically significant at any locus. For `phenotype2`, the pattern of horizontal, but not vertical, separation visually illustrates the identified mQTL. For `phenotype3`, the pattern of vertical, but not horizontal, separation visually illustrates the identified vQTL. For `phenotype4`, the pattern of two-dimensional separation illustrates an mvQTL.

4.4 Communicate Significant Findings

Having identified interesting QTL, we want to visualize the their estimated genetic and covariate effects. Because the `vqtl` package models effects for both mean and variance, existing plotting utilities are not able to display the entirety of the modeling results. To understand and communicate the results of a `vqtl` scan at one particular locus, we developed the `mean_var_plot`. This plot illustrates how the mean sub-model and variance sub-model of the DGLM fit the data at a given locus.

In each `mean_var_plot` in Figure 4.3, the location of the dot shows the estimated mean and standard deviation of each genotype group, with the mean indicated by the horizontal position and the standard deviation indicated by the vertical position. The horizontal lines extending to the left and right from each dot show the standard error of the mean estimate, and the vertical lines extending up and down from each dot show the standard error of the standard deviation estimate. There are two types of grouping factors considered by the function `mean_var_plot_model_based`: (1) `focal.groups` are groups that are modeled and the prediction for each group is plotted. For example, a genetic marker is the `focal.group` in each plot in Figure 4.3; `D1M1` in the top left, `D1M6` in the top right, `D2M6` in the bottom left, and `D3M6` in the bottom right. (2) `nuisance.groups` are groups that are modeled, but then averaged over before plotting. When there are many grouping factors thought to play a role in determining the mean and variance of an individual's phenotype, such as sex, treatment, and batch, we recommend putting just one or two in `focal.groups` and the others in `nuisance.groups` for clarity, cycling through which are displayed to gain a thorough understanding of the factors that influence the phenotype.

Additional plotting utilities, `phenotype_plot`, `effects_plot` and `mean_var_plot_model_free` are described in the online documentation, available on CRAN.

4.5 Establish a Confidence Interval for the QTL

Last, it is important to assess the genetic precision of a discovered QTL for bioinformatic follow-up. The function `scanonevar.boot` implements the non-parametric bootstrap (Visscher et al., 1996). This function takes, as arguments, a `scanonevar` object, the type of QTL detected, the name of the chromosome containing the QTL, and `num.resamples`, the number of bootstrap resamplings desired. As with `scanonevar.perm`, the `n.cores` argument can be used to spread the bootstraps over many computational cores and defaults to the number of cores available minus two, and bootstraps can be run on separate computers and combined with `c` to increase the precision of the estimate of the confidence interval.

We recommend 1000 resamples to establish 80% and 90% confidence intervals. With the datasets simulated here, it takes 20 minutes to run 1000 bootstrap resamples on an Intel core i5.

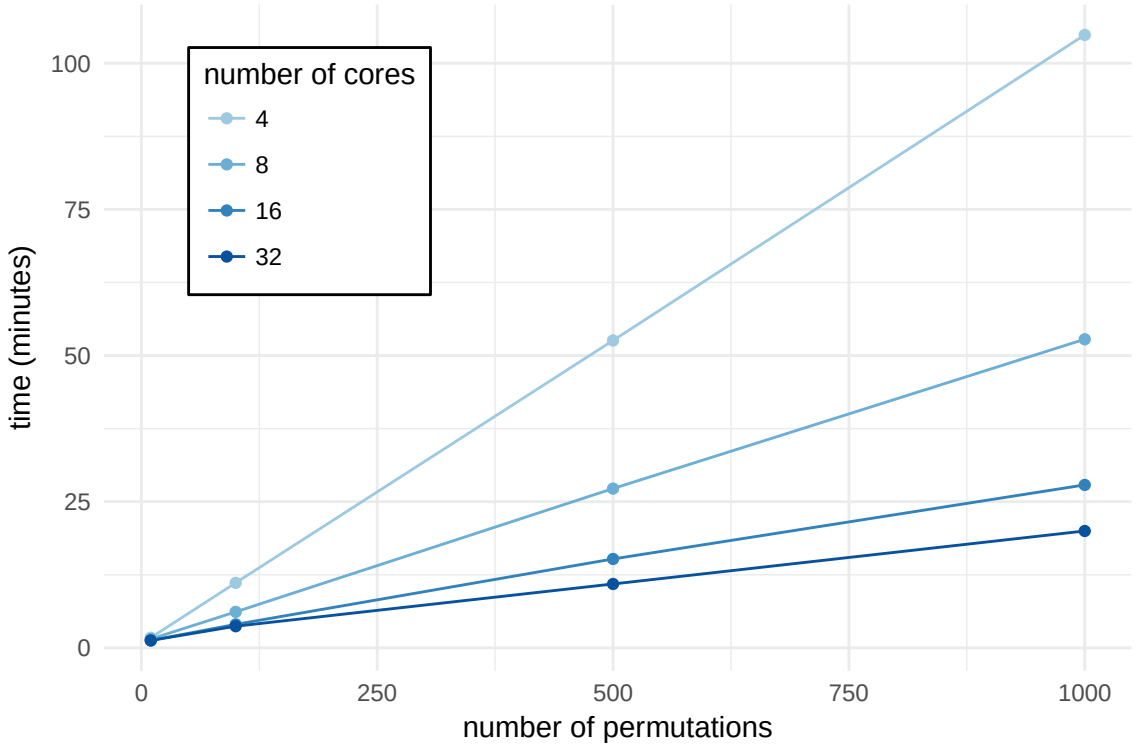


Figure 4.4: Time taken to run `scanonevar.perm` on the data from Kumar et al. (2013) which contains 244 individuals and 582 loci, varying the number of permutations desired and the number of computer cores used. For a given number of cores, there is a linear relationship between number of permutations conducted and time required. The slope the line indicates time required per permutation and is dependent on the number of cores, ranging from $\approx$ 6.3 seconds per permutation with 4 cores to $\approx$ 1.2 second per permutation with 32 cores.

4.6 Performance Benchmarks

By far, the most computationally-intensive step in the mean-variance QTL mapping process is the assessment of genome-wide statistical significance by permutation. The original genome scan is much faster, because it involves only a single scan, and the bootstrap is much faster because it involves only a single chromosome.

For the first benchmark, we ran `scanonevar.perm` on the data from Kumar et al. (2013) and chapter 3, which contains 244 individuals and 582 loci, varying the number of permutations desired and the number of computer cores used. For a given number of cores, there is a linear relationship between number of permutations conducted and time required Figure 4.4. The slope the line indicates time required per permutation and is dependent on the number of cores, ranging from $\approx$ 6.3 seconds per permutation with 4 cores to $\approx$ 1.2 second per permutation with 32 cores.

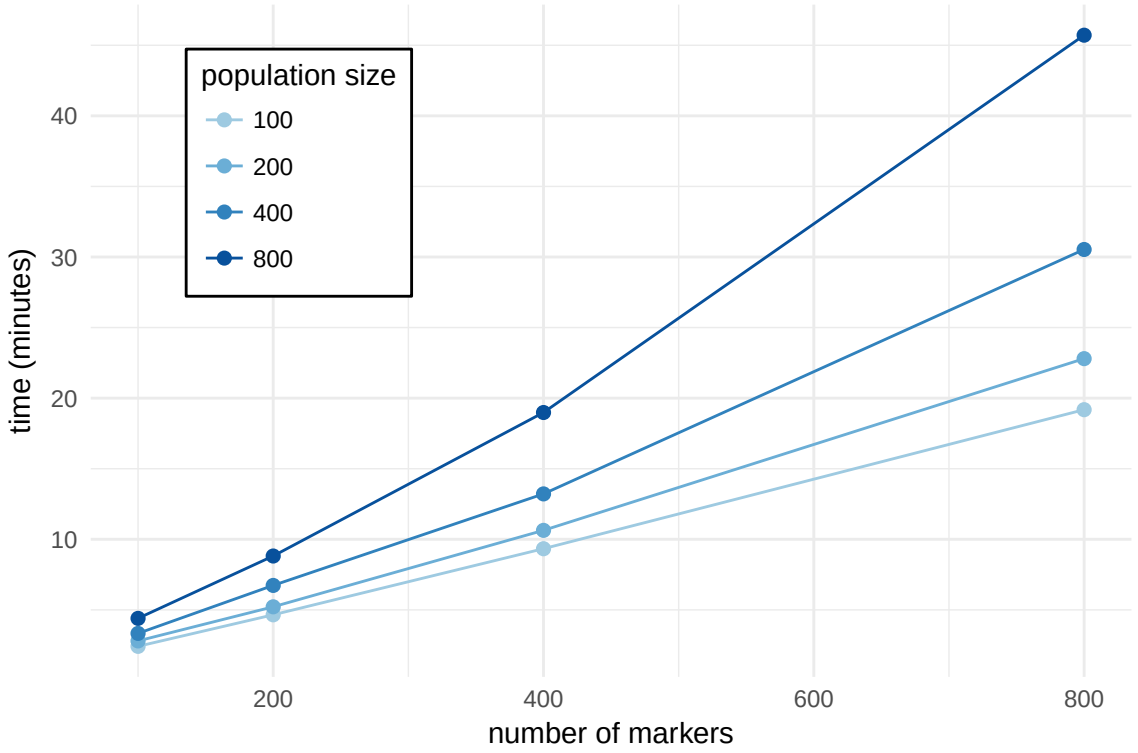


Figure 4.5: Time taken to run 1000 permutation scans on 32 cores on simulated data using `scanonevar.perm`, varying the number of individuals in the mapping population and the number of markers in the genome. For a given population size, there is a slightly supra-linear relationship between number of markers and time required. The average slope of the line indicates the average time required per locus and is dependent on the population size, ranging from $\approx$ 1.4 seconds per locus with a population of size 100 to $\approx$ 3.3 seconds per locus with a population of size 800.

For the second benchmark, we ran `scanonevar.perm` on simulated data, always conducting 1000 permutations and using 32 cores, but varying the number of individuals in the mapping population and the number of markers in the genome. For a given population size, there is a slightly supra-linear relationship between number of markers and time required Figure 4.5, which reflects a linear increase in the time taken to conduct the permuted genome scans, plus an increase in the time taken for "bookkeeping" tasks like organizing and reshaping genetic data. The slope of the line indicates the time required per locus and is dependent on the population size, ranging from $\approx$ 1.4 seconds per locus with a population of size 100 to $\approx$ 3.3 seconds per locus with a population of size 800.

Based on these benchmarks, it is clear that the workflow presented here is practical for QTL mapping F2 intercross and similar populations on modern, multi-core scientific computers. Populations

with many recombinations, where dense genotyping arrays that interrogate $>$ 10,000 loci are relevant, could not be practically analyzed with package `vqtl` in this way. However, both statistical and computational steps could be taken to make such a study feasible. Statistically, techniques could be used that allow for large-scale analysis without permutation testing (Efron, 2004). Computationally, the software could be changed to run on a computer cluster, rather than on a single computer (Jette and Grondona, 2003; Marchand, 2017).

4.7 Conclusion

We have demonstrated typical usage of the `R` package `vqtl` for mean-variance QTL mapping in an F2 intercross. This package is appropriate for crosses and phenotypes where genetic factors or covariates or are known or suspected to influence phenotype variance. In the case of genetic factors, they can be mapped, as illustrated in chapter 3. In the case of covariates, they can be accommodated, which can increase power and improve false positive rate control, as illustrated in chapter 2.

4.8 Resources

The scripts used to simulate genotypes and phenotypes, conduct the genome scans, and plot the results are available as a public, static Zenodo repository at `DOI:10.5281/zenodo.1173799`. The package `vqtl` and its documentation are freely available on `CRAN` at `https://CRAN.R-project.org/package=vqtl`.

4.9 Phenotypes with Background Variance Heterogeneity

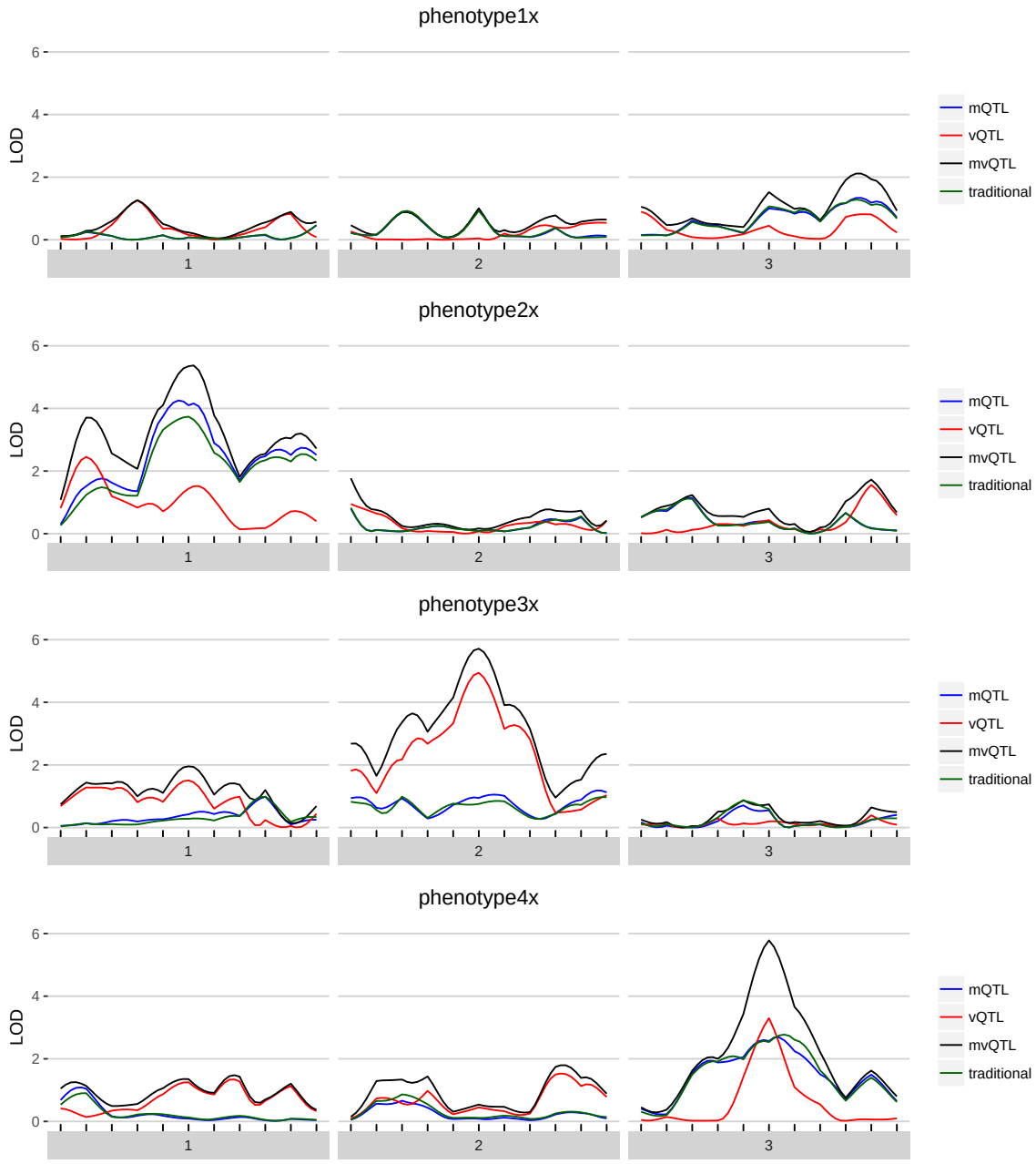


Figure 4.6: For each of the four simulated phenotypes with background variance heterogeneity, the genome scan shows the LOD score of each test – mean, variance, and joint – in blue, red, and black, respectively. The traditional test is in green and globally similar to the mean test.

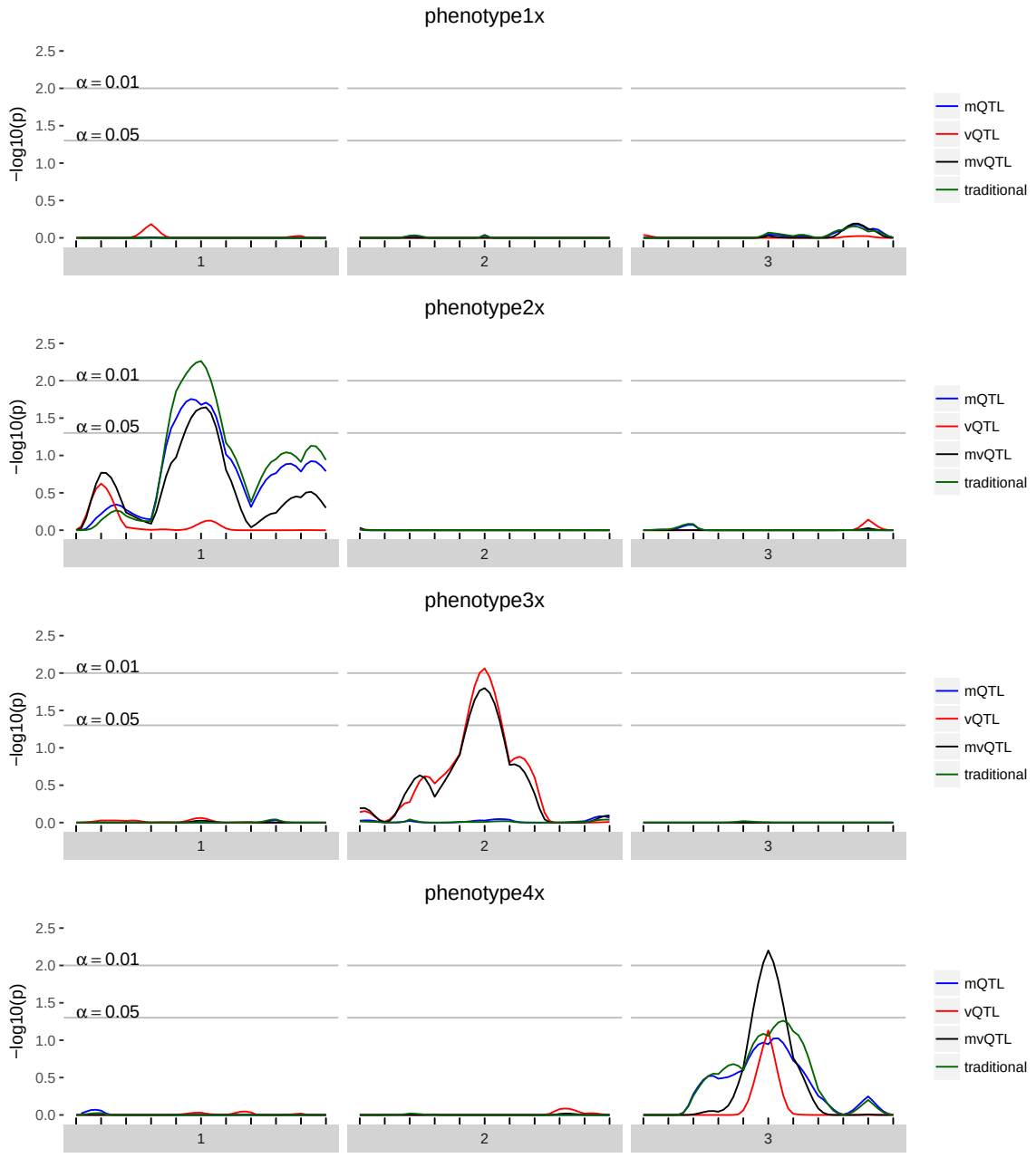


Figure 4.7: For each of the four simulated phenotypes with background variance heterogeneity, the genome scan shows the -log10 of the FWER-corrected p-value of each test – mean, variance, and joint – in blue, red, and black, respectively. Thus, a value of 3 implies that the quantity of evidence against the null is such that we expect to see this much or more evidence once per thousand genome scans when there is no true effect.

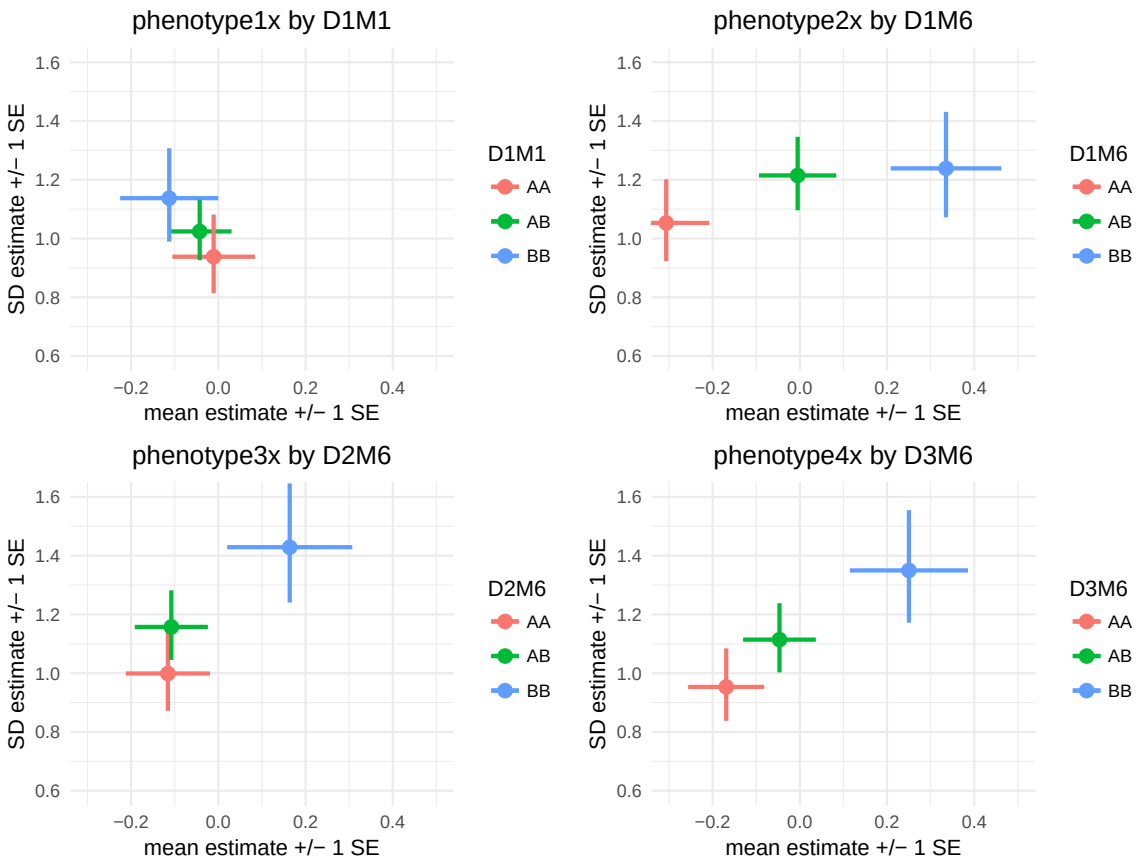


Figure 4.8: `mean_var_plots` show the estimated genotype effects at a locus, with mean effects on the horizontal axis and variance effects on the vertical axis. Horizontal lines indicate standard errors for mean effects and vertical lines indicate standard errors for variance effects. For `phenotype1x`, the pattern of overlapping estimates and standard errors is consistent with the fact that there are no genetic effects, and the p-value was not statistically significant at any locus. For `phenotype2x`, the pattern of horizontal, but not vertical, separation visually illustrates the identified mQTL on a background of variance heterogeneity. For `phenotype3x`, the pattern of vertical, but not horizontal, separation visually illustrates the identified vQTL on a background of variance heterogeneity. For `phenotype4x`, the pattern of two dimensional separation without either total horizontal or vertical separation illustrates an mvQTL with neither mean nor variance effect strong enough to define an mQTL or vQTL on a background of variance heterogeneity.

CHAPTER 5
The Heteroscedastic Linear Mixed Model

This chapter deals with the linear mixed model (LMM). This statistical model can be applied in association mapping and LD mapping in situations where all individuals in the population are not equally-related. In such populations, population structure and cryptic relatedness break the assumption of the standard linear model that all observations are independent, conditional on the effects of the covariates. By estimating so called "random effects" with arbitrary covariance structure, the LMM can accurately accommodate this differential relatedness and thereby maintain the validity of the statistical inference. But, the additional complexity of the LMM relative to the SLM presents a challenge as well — in studies with a large number of organisms and/or a large number of genetic markers, the computational cost associated with using the LMM can be prohibitive.

This chapter proceeds as follows. In section 1, I describe a verbose form of the LMM that emphasizes the meaning of each term in the model and describe how this model can be used to conduct a GWAS. In section 2, I demonstrate a compact, but mathematically equivalent, form of the LMM that will be used for the remainder of the chapter. In section 3, I illustrate how, given the value of one parameter in the LMM (h^2), it can be fit by the simpler generalized least squares (GLS) procedure. In section 4, I illustrate how, given the value of one parameter in the GLS model ($\mathbf{M}$), it can be fit by the simpler ordinary least squares (OLS) procedure. In section 5, I describe how these simplifications can be used to rapidly fit the LMM genome-wide. In section 6, I summarize a previously published result that demonstrates how to rapidly calculate the necessary parameter for the GLS-to-OLS simplification in the situation where the micro-environmental residuals are homoskedastic. In section 7, I demonstrate a novel way to calculate the necessary parameter for the GLS-to-OLS simplification that is valid whether the micro-environmental residuals are homoskedastic or heteroskedastic. In section 8, I show, through simulation, that for phenotypes with heteroskedastic environmental residuals, the novel method of calculating the simplifying parameter leads better false positive rate control and a more powerful test.

5.1 The Linear Mixed Model

The LMM models an observed phenotype, $\boldsymbol{y}$, as,

$$\boldsymbol{y} = \mathbf{1}\mu + \mathbf{X}\boldsymbol{\beta} + \mathbf{G}\boldsymbol{\alpha} + \mathbf{a} + \mathbf{e} \tag{5.1}$$

where $\mathbf{1}$ is a column vector of ones, $\mathbf{X}$ is the design matrix of covariates, $\mathbf{G}$ is the design matrix describing the genetic locus to be tested, and μ, $\boldsymbol{\beta}$, and $\boldsymbol{\alpha}$ are unconstrained parameters that can be referred to as the population mean, the effect(s) of the covariate(s), and the effect(s) of the genetic factor(s) to be tested, respectively. $\mathbf{a}$ and $\mathbf{e}$ are so-called "random effects", estimated in the process of model fitting, but with constraints. Specifically, they are modeled hierarchically as

$$\mathbf{a} \sim \mathrm{N}(0, \mathbf{K}\tau^2), \tag{5.2}$$

$$\mathbf{e} \sim \mathrm{N}(0, \mathbf{D}\sigma^2) \tag{5.3}$$

where $\mathbf{K}$ is a known, positive semi-definite genomic similarity matrix, and $\mathbf{D}$ is a known, diagonal residual variance matrix. The scale parameters, τ^2 and σ^2, are constrained only to be non-negative.

To conduct a GWAS, the LMM is fit to each polymorphic genetic variant, using $\mathbf{G}$ to encode the locus design matrix and testing whether $\boldsymbol{\alpha} = 0$. If $\boldsymbol{\alpha} \neq 0$, the locus is a QTL. Here, we use the likelihood ratio test (LRT) to test how likely the observed difference between $\widehat{\boldsymbol{\alpha}}$ and 0 is, due to chance alone.

In this context, the LRT requires "fitting" both a null and alternative version of the LMM, where the term "fitting" is shorthand for calculating the maximum likelihood value of all parameters and calculating the likelihood of the model at those parameter estimates. The relevant alternative model is written in Equation 5.1 and the relevant null model is identical except it excludes the term $\mathbf{G}\boldsymbol{\alpha}$.

Henderson (1984) described a suite of procedures for fitting the LMM in a variety of situations, but Henderson's methods are of limited use in QTL mapping because they are computationally slow. His focus was on estimation of breeding values ($\mathbf{a}$ in Equation 5.1), which is useful for livestock improvement breeding programs, so the model only needed to be fit once, to one design matrix, and speed was not a primary concern.

5.2 Compact Specification of the LMM

The LMM as specified in Equation 5.1 is equivalent to:

$$\boldsymbol{y} \sim \mathrm{N}(\mathbf{X}_c\boldsymbol{\beta}_c, \mathbf{V}\lambda) \tag{5.4}$$

where fixed effects design matrices are combined into $\mathbf{X}_c$ and the variance-covariance matrices of the random effects are combined into $\mathbf{V}\lambda$. Specifically, the covariate matrices and their effect vectors are compacted as

$$\mathbf{X}_c = [1 \;\; \mathbf{X} \;\; \mathbf{G}] \tag{5.5}$$

$$\boldsymbol{\beta}_c = \left[\mu \;\; \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\alpha}^{\mathrm{T}}\right]^{\mathrm{T}} \tag{5.6}$$

Going forward, only the compact notation is used, so $\mathbf{X}_c$ and $\boldsymbol{\beta}_c$ will be referred to simply as $\mathbf{X}$ and $\boldsymbol{\beta}$. The covariance matrices is compacted by use the re-parametrization,

$$h^2 = \frac{\tau^2}{\tau^2 + \sigma^2} \tag{5.7}$$

$$\lambda = \tau^2 + \sigma^2 \tag{5.8}$$

and "hiding" the h^2 parameter inside the definition of $\mathbf{V}$, which will be mathematically useful. After defining

$$\mathbf{V} = \mathbf{K}h^2 + \mathbf{D}(1 - h^2) \tag{5.9}$$

we have

$$\mathbf{K}\tau^2 + \mathbf{D}\sigma^2 = \left(\mathbf{K}h^2 + \mathbf{D}(1 - h^2)\right)\lambda \tag{5.10}$$

$$= \mathbf{V}\lambda \tag{5.11}$$

This parameterization's usage of the narrow-sense heritability, h^2, has two benefits. First, it is directly interpretable to geneticists. And second, it is bounded to the range $[0, 1]$, which can be useful in a grid-based or gradient-based search for an optimal parameter value.

5.3 Given h^2, the LMM problem reduces to the GLS problem

Given it the value of h^2, the LMM simplifies to the generalized least squares (GLS) model.

$$\boldsymbol{y} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}\lambda) \tag{5.12}$$

The well-known ML estimates for $\boldsymbol{\beta}$ and λ are

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\boldsymbol{y} \tag{5.13}$$

$$\widehat{\lambda} = \left\|(\mathbf{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{y})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{y})\right\|_2 \tag{5.14}$$

which can calculated directly more rapidly than the LMM could be fit directly. However, due to the fact that $\mathbf{X}$ changes with every new genetic locus and the requirement to invert $(\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})$ to solve this problem directly, this solution is still not optimal for GWAS application. The next section describes a further simplification that will make the genome-wide fitting of the GLS (and, by extension, the LMM) computationally tractable.

Note that the likelihood of Equation 5.12 is:

$$\ell(\boldsymbol{\beta}, \lambda; \boldsymbol{y}, \mathbf{X}, \mathbf{V}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{V}\lambda| - \frac{1}{2\lambda}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{V}^{-1}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) \tag{5.15}$$

$$= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{V}| - \frac{n}{2}\log\lambda - \frac{1}{2\lambda}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{V}^{-1}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) \tag{5.16}$$

to which I will refer in the next section.

5.4 Given M, the GLS problem reduces to the OLS problem

Although the simplification of the LMM fitting process to the GLS procedure did not immediately result in sufficient speed-up to make GWAS tractable, the GLS procedure proves is amenable to further simplification. Take as given a matrix, $\mathbf{M}$, that has the property

$$\mathbf{M}^{\mathrm{T}}\mathbf{M} = \mathbf{V}^{-1} \tag{5.17}$$

where $\mathbf{V}$ is the covariance of $\boldsymbol{y}$, as defined in Equation 5.9. I can use $\mathbf{M}$ to define a "rotated" phenotype vector, $\boldsymbol{y}_r = \mathbf{M}\boldsymbol{y}$, which has identity covariance, as can be verified by

$$\begin{aligned}
\text{Var}(\boldsymbol{y}_r) &= \text{Var}(\mathbf{M}\boldsymbol{y}) && (5.18)\\
&= \mathbf{M}\text{Var}(\boldsymbol{y})\mathbf{M}^{\mathrm{T}} && (5.19)\\
&= \mathbf{M}\mathbf{V}\mathbf{M}^{\mathrm{T}} && (5.20)\\
&= \mathbf{I} && (5.21)
\end{aligned}$$

where the last step can be verified by

$$\begin{aligned}
\mathbf{M}\mathbf{V}\mathbf{M}^{\mathrm{T}} &= \mathbf{I} && (5.22)\\
\mathbf{M}^{\mathrm{T}}\mathbf{M}\mathbf{V}\mathbf{M}^{\mathrm{T}}\mathbf{M} &= \mathbf{M}^{\mathrm{T}}\mathbf{M} && (5.23)\\
\mathbf{V}^{-1}\mathbf{V}\mathbf{V}^{-1} &= \mathbf{V}^{-1} && (5.24)\\
\mathbf{V}^{-1} &= \mathbf{V}^{-1} && (5.25)
\end{aligned}$$

Because $\boldsymbol{y}_r$ has identity covariance, it can be modeled with a simple linear model (SLM). In particular, we choose to model it as:

$$\boldsymbol{y}_r \sim \mathrm{N}(\mathbf{X}_r\boldsymbol{\beta}_r,\, \mathbf{I}\lambda_r) \tag{5.26}$$

where $\mathbf{X}_r = \mathbf{M}\mathbf{X}$ is a "rotated" covariate matrix. This linear model can be solved by ordinary least squares (OLS), which is computationally efficient and numerically stable when solved by the QR decomposition. The well-known ML estimates of $\boldsymbol{\beta}_r$ and λ_r are

$$\begin{aligned}
\widehat{\boldsymbol{\beta}_r} &= (\mathbf{X}_r^{\mathrm{T}}\mathbf{X}_r)^{-1}\mathbf{X}_r^{\mathrm{T}}\boldsymbol{y}_r && (5.27)\\
\widehat{\lambda_r} &= \left\|(\mathbf{X}_r\widehat{\boldsymbol{\beta}_r} - \boldsymbol{y}_r)^{\mathrm{T}}(\mathbf{X}_r\widehat{\boldsymbol{\beta}_r} - \boldsymbol{y}_r)\right\|_2 && (5.28)
\end{aligned}$$

For completeness, it should be shown that $\widehat{\boldsymbol{\beta}_r} = \widehat{\boldsymbol{\beta}}$, which can be verified by:

$$\widehat{\boldsymbol{\beta}_r} = (\mathbf{X}_r^{\mathrm{T}}\mathbf{X}_r)^{-1}\mathbf{X}_r^{\mathrm{T}}\boldsymbol{y}_r \tag{5.29}$$

$$= ((\mathbf{M}\mathbf{X})^{\mathrm{T}}(\mathbf{M}\mathbf{X}))^{-1}(\mathbf{M}\mathbf{X})^{\mathrm{T}}\mathbf{M}\boldsymbol{y} \tag{5.30}$$

$$= (\mathbf{X}^{\mathrm{T}}\mathbf{M}^{\mathrm{T}}\mathbf{M}\mathbf{X}))^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{M}^{\mathrm{T}}\mathbf{M}\boldsymbol{y} \tag{5.31}$$

$$= (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\boldsymbol{y} \tag{5.32}$$

$$= \widehat{\boldsymbol{\beta}} \tag{5.33}$$

and that $\widehat{\lambda_r} = \widehat{\lambda}$, which can be verified by:

$$\widehat{\lambda_r} = \left\|(\mathbf{X}_r\widehat{\boldsymbol{\beta}_r} - \boldsymbol{y}_r)^{\mathrm{T}}(\mathbf{X}_r\widehat{\boldsymbol{\beta}_r} - \boldsymbol{y}_r)\right\|_2 \tag{5.34}$$

$$= \left\|(\mathbf{M}\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{M}\boldsymbol{y})^{\mathrm{T}}(\mathbf{M}\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{M}\boldsymbol{y})\right\|_2 \tag{5.35}$$

$$= \left\|(\mathbf{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{y})^{\mathrm{T}}\mathbf{M}^{\mathrm{T}}\mathbf{M}(\mathbf{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{y})\right\|_2 \tag{5.36}$$

$$= \left\|(\mathbf{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{y})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{y})\right\|_2 \tag{5.37}$$

$$= \widehat{\lambda} \tag{5.38}$$

That $\widehat{\boldsymbol{\beta}_r} = \widehat{\boldsymbol{\beta}}$ and $\widehat{\lambda_r} = \widehat{\lambda}$ can also be verified from the log likelihoods. The OLS model described in Equation 5.26 has log likelihood:

$$\ell(\boldsymbol{\beta}_r, \lambda_r; \mathbf{X}_r, \boldsymbol{y}_r) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{I}| - \frac{n}{2}\log\lambda_r - \frac{1}{2\lambda_r}(\boldsymbol{y}_r - \mathbf{X}_r)^{\mathrm{T}}(\boldsymbol{y}_r - \mathbf{X}_r\boldsymbol{\beta}_r) \tag{5.39}$$

$$= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\lambda_r - \frac{1}{2\lambda_r}(\mathbf{M}\boldsymbol{y} - \mathbf{M}\mathbf{X}\boldsymbol{\beta}_r)^{\mathrm{T}}(\mathbf{M}\boldsymbol{y} - \mathbf{M}\mathbf{X}\boldsymbol{\beta}_r) \tag{5.40}$$

$$= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\lambda_r - \frac{1}{2\lambda_r}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}_r)^{\mathrm{T}}\mathbf{M}^{\mathrm{T}}\mathbf{M}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}_r) \tag{5.41}$$

$$= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\lambda_r - \frac{1}{2\lambda_r}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}_r)^{\mathrm{T}}\mathbf{V}^{-1}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}_r) \tag{5.42}$$

which differs from that of the GLS problem (Equation 5.12) by a constant, $\frac{1}{2}\log|\mathbf{V}|$.

$$\ell(\boldsymbol{\beta}_r, \lambda_r; \mathbf{X}_r, \boldsymbol{y}_r) = \ell(\boldsymbol{\beta}, \lambda; \mathbf{X}, \boldsymbol{y}) - \frac{1}{2}\log|\mathbf{V}| \tag{5.43}$$

thus, for fixed h^2 (and thus fixed $\mathbf{V}$), these likelihoods reach their maxima at the same parameter values.

The combination of these two simplifications makes possible a strategy to use the LMM to conduct a GWAS. Specifically, by using Brent's method to optimize over h^2, and therefore using a fixed h^2 at each step, and using $\mathbf{M}$ to fit the model by OLS rather than GLS at each step, the LMM can be rapidly fit to any $\mathbf{X}$. But this procedure requires the ability to rapidly calculate $\mathbf{M}$ to calculate the maximum likelihood parameter values and and $\log|\mathbf{M}|$ to "back correct" the rotated likelihood to the un-rotated frame, which we have thus far not addressed. In the following sections we address these issues.

5.5 $\mathbf{M}$ for the Homoscedastic LMM

Kang et al. (2008) proposed the strategy for GWAS described above and proposed a value of $\mathbf{M}$ that is computationally efficient and is valid when $\mathbf{D} = \mathbf{I}$. This advance was termed "EMMA", an acronym for efficient mixed-model analysis. Their approach used a slightly different, but mathematically equivalent, parameterization of the variance components, but we convert it into the (h^2, λ) parameterization here for consistency with the rest of this chapter. The differences in parameterization do not change the likelihood of the model and do not influence any results.

5.5.1 Kang's $\mathbf{M}$

Kang et al. (2008) proposed

$$\mathbf{M}_{\text{hom}} = \left(h^2\mathbf{\Lambda}_{\mathbf{K}} + (1-h^2)\mathbf{I}\right)^{-\frac{1}{2}} \mathbf{U}_{\mathbf{K}}{}^{\mathrm{T}} \tag{5.44}$$

where $\mathbf{\Lambda}_{\mathbf{K}}$ and $\mathbf{U}_{\mathbf{K}}$ are the eigenvalue matrix and eigenvector matrix of $\mathbf{K}$, respectively. Importantly, $\mathbf{K}$ is fixed for the entire genome scan, so it need only be eigen-decomposed once and its eigen vectors and eigen values can be used to calculate useful locus-specific intermediates as described below.

5.5.2 Validity

It can be verified that $\mathbf{M}_{\text{hom}}^{\mathrm{T}}\mathbf{M}_{\text{hom}} = \mathbf{V}^{-1}$. First, compute a useful form of $\mathbf{V}$ and $\mathbf{V}^{-1}$.

$$\begin{aligned}
\mathbf{V} &= h^2\mathbf{K} + (1-h^2)\mathbf{I} && \text{definition} && (5.45)\\
&= h^2\mathbf{U_K}\mathbf{\Lambda_K}\mathbf{U_K}^{\mathrm{T}} + (1-h^2)\mathbf{I} && \text{eigen decomposition} && (5.46)\\
&= h^2\mathbf{U_K}\mathbf{\Lambda_K}\mathbf{U_K}^{\mathrm{T}} + (1-h^2)\mathbf{U_K}\mathbf{U_K}^{\mathrm{T}} \\
& && \text{eigen vectors of real symmetric are orthonormal} && (5.47)\\
&= \mathbf{U_K}\left(h^2\mathbf{\Lambda_K} + (1-h^2)\mathbf{I}\right)\mathbf{U_K}^{\mathrm{T}} && \text{distributive property} && (5.48)
\end{aligned}$$

This eigen-form can be inverted directly by inverting the eigen values, giving

$$\mathbf{V}^{-1} = \mathbf{U_K}\left(h^2\mathbf{\Lambda_K} + (1-h^2)\mathbf{I}\right)^{-1}\mathbf{U_K}^{\mathrm{T}} \tag{5.49}$$

Now, we can verify the necessary equality

$$\begin{aligned}
\mathbf{M}^{\mathrm{T}}\mathbf{M} &= \left(\left(h^2\mathbf{\Lambda_K} + (1-h^2)\mathbf{I}\right)^{-\frac{1}{2}}\mathbf{U_K}^{\mathrm{T}}\right)^{\mathrm{T}}\left(\left(h^2\mathbf{\Lambda_K} + (1-h^2)\mathbf{I}\right)^{-\frac{1}{2}}\mathbf{U_K}^{\mathrm{T}}\right) \\
& && \text{definition} && (5.50)\\
&= \mathbf{U_K}\left(h^2\mathbf{\Lambda_K} + (1-h^2)\mathbf{I}\right)^{-\frac{1}{2}}\left(h^2\mathbf{\Lambda_K} + (1-h^2)\mathbf{I}\right)^{-\frac{1}{2}}\mathbf{U_K}^{\mathrm{T}} \\
& && \text{transpose of product} && (5.51)\\
&= \mathbf{U_K}\left(h^2\mathbf{\Lambda_K} + (1-h^2)\mathbf{I}\right)^{-1}\mathbf{U_K}^{\mathrm{T}} && \text{product of roots} && (5.52)\\
&= \mathbf{V}^{-1} && \text{from above} && (5.53)
\end{aligned}$$

5.5.3 Calculation of $\log|\mathbf{V}|$

As described in Section 5.4, $\log|\mathbf{V}|$ is necessary to calculate the likelihood of the original model from the likelihood of the rotated model. In the case of $\mathbf{M}_{\text{hom}}$ it is straightforward to calculate. The determinant of any matrix is equal to the product of its eigen values, so the log of the determinant is the sum of its eigen values. We already have the eigen values of the $\mathbf{V}$, using only the eigen

decomposition of $\mathbf{K}$, as indicated in

$$\log|\mathbf{V}| = \sum_{i=1}^{n} h^2 \lambda_K i + (1 - h^2) \tag{5.54}$$

5.6 $\mathbf{M}$ for the Heteroscedastic LMM

The above $\mathbf{M}$ ($\mathbf{M}_{\text{hom}}$) is valid only when the micro-environmental covariance is identity. Note that the second step of the validity proof for $\mathbf{M}_{\text{hom}}$ (Equation 5.46 to 5.47) requires re-expressing that covariance matrix as $\mathbf{U_K U_K}^{\mathrm{T}}$. That equality not generally true — it is true only when that covariance is identity. Said another way, generally $\mathbf{D} \neq \mathbf{U_K U_K}^{\mathrm{T}}$. In the special case where $\mathbf{D} = \mathbf{I}$, though, $\mathbf{D} = \mathbf{I} = \mathbf{U_K U_K}^{\mathrm{T}}$.

In the more general situation, where the phenotype associated with some genotypes is known with more certainty than the phenotype associated with other genotypes and therefore $\mathbf{D} \neq \mathbf{I}$, it would be preferable to use a covariance matrix for the residual variance that reflects this reality. Here, I propose a multiplier matrix that yields the same speed up as $\mathbf{M}_{\text{hom}}$, but remains valid for any diagonal residual covariance matrix.

5.6.1 Proposal

I propose

$$\mathbf{M}_{\text{het}} = (h^2 \mathbf{\Lambda_L} + (1 - h^2)\mathbf{I})^{-\frac{1}{2}} \mathbf{U_L}^{\mathrm{T}} \mathbf{D}^{-\frac{1}{2}} \tag{5.55}$$

where

$$\mathbf{L} = \mathbf{D}^{-\frac{1}{2}} \mathbf{K} \mathbf{D}^{-\frac{1}{2}} \tag{5.56}$$

where $\mathbf{\Lambda_L}$ and $\mathbf{U_L}$ are the eigenvalue matrix and eigenvector matrix of $\mathbf{L}$, respectively. Note that $\mathbf{L}$ has the property that, like $\mathbf{K}$, it is fixed for the entire genome scan, so it need only be eigen-decomposed once, though its eigen vectors and eigen values can be used to calculate useful locus-specific intermediates as described below.

5.6.2 Validity

As before, to be a valid multiplier matrix, $\mathbf{M}$ must have the property:

$$\mathbf{M}^{\mathrm{T}}\mathbf{M} = \mathbf{V}^{-1} \tag{5.57}$$

To verify this equality, we begin by calculating a useful form of $\mathbf{V}$ and $\mathbf{V}^{-1}$.

Proof.

$$\begin{aligned}
\mathbf{V} &= h^2\mathbf{K} + (1-h^2)\mathbf{D} && \text{definition} && (5.58)\\
&= \mathbf{D}^{\frac{1}{2}}\mathbf{D}^{-\frac{1}{2}}(h^2\mathbf{K} + (1-h^2)\mathbf{D}) && \text{pre-multiply by } \mathbf{D}^{\frac{1}{2}}\mathbf{D}^{-\frac{1}{2}} = \mathbf{I} && (5.59)\\
&= \mathbf{D}^{\frac{1}{2}}\mathbf{D}^{-\frac{1}{2}}(h^2\mathbf{K} + (1-h^2)\mathbf{D})\mathbf{D}^{-\frac{1}{2}}\mathbf{D}^{\frac{1}{2}} && \text{post-multiply by } \mathbf{D}^{-\frac{1}{2}}\mathbf{D}^{\frac{1}{2}} = \mathbf{I} && (5.60)\\
&= \mathbf{D}^{\frac{1}{2}}(h^2\mathbf{D}^{-\frac{1}{2}}\mathbf{K}\mathbf{D}^{-\frac{1}{2}} + (1-h^2)\mathbf{D}^{-\frac{1}{2}}\mathbf{D}\mathbf{D}^{-\frac{1}{2}})\mathbf{D}^{\frac{1}{2}} && \text{distribute } \mathbf{D}^{-\frac{1}{2}} \text{ in} && (5.61)\\
&= \mathbf{D}^{\frac{1}{2}}(h^2\mathbf{D}^{-\frac{1}{2}}\mathbf{K}\mathbf{D}^{-\frac{1}{2}} + (1-h^2)\mathbf{I})\mathbf{D}^{\frac{1}{2}} && \text{definition of root inverse} && (5.62)\\
&= \mathbf{D}^{\frac{1}{2}}(h^2\mathbf{L} + (1-h^2)\mathbf{I})\mathbf{D}^{\frac{1}{2}} && \text{define: } \mathbf{L} = \mathbf{D}^{-\frac{1}{2}}\mathbf{K}\mathbf{D}^{-\frac{1}{2}} && (5.63)\\
&= \mathbf{D}^{\frac{1}{2}}(h^2\mathbf{U_L}\mathbf{\Lambda_L}\mathbf{U_L}^{\mathrm{T}} + (1-h^2)\mathbf{I})\mathbf{D}^{\frac{1}{2}} && \text{eigen decomposition of } \mathbf{L} && (5.64)\\
&= \mathbf{D}^{\frac{1}{2}}(h^2\mathbf{U_L}\mathbf{\Lambda_L}\mathbf{U_L}^{\mathrm{T}} + (1-h^2)\mathbf{U_L}\mathbf{U_L}^{\mathrm{T}})\mathbf{D}^{\frac{1}{2}} && \text{property of eigen vectors} && (5.65)\\
&= \mathbf{D}^{\frac{1}{2}}\mathbf{U_L}(h^2\mathbf{\Lambda_L} + (1-h^2)\mathbf{I})\mathbf{U_L}^{\mathrm{T}}\mathbf{D}^{\frac{1}{2}} && \text{distributive property} && (5.66)
\end{aligned}$$

Given that form of $\mathbf{V}$, a useful form of $\mathbf{V}^{-1}$ is near.

$$\begin{aligned}
\mathbf{V}^{-1} &= \left(\mathbf{D}^{\frac{1}{2}}\mathbf{U_L}(h^2\mathbf{\Lambda_L} + (1-h^2)\mathbf{I})\mathbf{U_L}^{\mathrm{T}}\mathbf{D}^{\frac{1}{2}}\right)^{-1} && \text{definition} && (5.67)\\
&= \mathbf{D}^{-\frac{1}{2}}\left(\mathbf{U_L}(h^2\mathbf{\Lambda_L} + (1-h^2)\mathbf{I})\mathbf{U_L}^{\mathrm{T}}\right)^{-1}\mathbf{D}^{-\frac{1}{2}} && \text{inverse of product} && (5.68)\\
&= \mathbf{D}^{-\frac{1}{2}}\mathbf{U_L}(h^2\mathbf{\Lambda_L} + (1-h^2)\mathbf{I})^{-1}\mathbf{U_L}^{\mathrm{T}}\mathbf{D}^{-\frac{1}{2}} && \text{inverse of eigen decomposition} && (5.69)
\end{aligned}$$

It is straightforward to compare $\mathbf{M}^{\mathrm{T}}\mathbf{M}$ with this form of $\mathbf{V}^{-1}$ to verify their equality.

$$\mathbf{M}^{\mathrm{T}}\mathbf{M} = \left((h^2\mathbf{\Lambda_L} + (1-h^2)\mathbf{I})^{-\frac{1}{2}}\mathbf{U_L}^{\mathrm{T}}\mathbf{D}^{-\frac{1}{2}}\right)^{\mathrm{T}}\left((h^2\mathbf{\Lambda_L} + (1-h^2)\mathbf{I})^{-\frac{1}{2}}\mathbf{U_L}^{\mathrm{T}}\mathbf{D}^{-\frac{1}{2}}\right)$$

definition (5.70)

$$= \left(\mathbf{D}^{-\frac{1}{2}}\mathbf{U_L}(h^2\mathbf{\Lambda_L} + (1-h^2)\mathbf{I})^{-\frac{1}{2}}\right)\left((h^2\mathbf{\Lambda_L} + (1-h^2)\mathbf{I})^{-\frac{1}{2}}\mathbf{U_L}^{\mathrm{T}}\mathbf{D}^{-\frac{1}{2}}\right)$$

transpose of product (note only $\mathbf{U_L}$ is non-diagonal) (5.71)

$$= \mathbf{D}^{-\frac{1}{2}}\mathbf{U_L}\left((h^2\mathbf{\Lambda_L} + (1-h^2)\mathbf{I})^{-\frac{1}{2}}(h^2\mathbf{\Lambda_L} + (1-h^2)\mathbf{I})^{-\frac{1}{2}}\right)\mathbf{U_L}^{\mathrm{T}}\mathbf{D}^{-\frac{1}{2}}$$

associative property (5.72)

$$= \mathbf{D}^{-\frac{1}{2}}\mathbf{U_L}(h^2\mathbf{\Lambda_L} + (1-h^2)\mathbf{I})^{-1}\mathbf{U_L}^{\mathrm{T}}\mathbf{D}^{-\frac{1}{2}} \qquad \text{definition of root inverse} \quad (5.73)$$

$$= \mathbf{V}^{-1} \qquad \text{from Equation 5.69} \quad (5.74)$$

□

5.6.3 Calculation of $\log|\mathbf{V}|$

As with $\mathbf{M}_{\text{hom}}$, the calculation of $\log|\mathbf{V}|$ comes almost for free after calculating $\mathbf{M}_{\text{het}}$.

$$\mathbf{V} = \mathbf{D}^{\frac{1}{2}}\left(h^2\mathbf{L} + (1-h^2)\mathbf{I}\right)\mathbf{D}^{\frac{1}{2}} \quad (5.75)$$

$$\log|\mathbf{V}| = \log\left(\left|\mathbf{D}^{\frac{1}{2}}\left(h^2\mathbf{L} + (1-h^2)\mathbf{I}\right)\mathbf{D}^{\frac{1}{2}}\right|\right) \quad (5.76)$$

$$= \log\left(\left|\mathbf{D}^{\frac{1}{2}}\right|\left|\left(h^2\mathbf{L} + (1-h^2)\mathbf{I}\right)\right|\left|\mathbf{D}^{\frac{1}{2}}\right|\right) \quad (5.77)$$

$$= \log\left(|\mathbf{D}|\left|\left(h^2\mathbf{L} + (1-h^2)\mathbf{I}\right)\right|\right) \quad (5.78)$$

$$= \log\left(|\mathbf{D}|\right) + \log\left(\left|\left(h^2\mathbf{L} + (1-h^2)\mathbf{I}\right)\right|\right) \quad (5.79)$$

At this point, the problem is reduced to two terms. The first is log of the determinant of a diagonal matrix, which is simply the sum of its elements. The second is the log of the determinant of a covariance matrix expressed in the same form as was present in the homoskedastic setting, simply with $\mathbf{L}$ in place of $\mathbf{K}$ and can be solved in the same way. Specifically,

$$\log\left(|\mathbf{D}|\right) = \sum_{i=1}^{n} d_i \quad (5.80)$$

and

$$\log\left(\left|\left(h^2\mathbf{L} + (1-h^2)\mathbf{I}\right)\right|\right) = \sum_{i=1}^{n} h^2\lambda_L i + (1-h^2) \tag{5.81}$$

5.7 Simulation Studies

Having laid the mathematical groundwork to rapidly fit the LMM with heteroskedastic environmental residuals, it is important to test the properties of this testing procedure as compared to the existing procedure. It is to be expected that, when the residuals truly are heteroskedastic and their variances are known by oracle, the heteroskedastic LMM should outperform the homoskedastic LMM in two ways.

First, I have observed that, when the homoskedastic LMM is applied to data with heteroskedastic residuals, the false positive rate (FPR) is covertly inflated. Said another way, while the probability of observing a nominal p value less than c under the null should be c, I have observed that with the homoskedastic LMM, in this scenario it is greater than c. Anti-conservative statistical behavior of this type can lead to false positive results and is deeply problematic.

Second, by bringing the model into closer accord with the data generating process, more powerful tests should be possible. That is to say, in situations where the residuals are heteroskedastic, and the locus truly does influence the phenotype, the heteroskedastic LMM is expected to be more likely to be able to reject the null hypothesis than the homoskedastic LMM.

5.7.1 Simulation Setup

I ran 10,000 null simulations, where no SNP directly influences the phenotype and 10,000 alternative simulations, where one genetic factor directly influence the phenotype. I will first describe the null simulations, with parameters that varied across simulations enclosed in curly brackets.

Each simulation consisted of a population of size $\{50, 100, 200\}$. Genomes were simulated by forward simulation, starting with one haploid individual with 100 binary genetic factors. At each time t a randomly chosen individual from the population asexually produced four offspring that were identical to the parent except for a random 15% of the genome was mutated. This process continued until the desired population size was reached.

For each simulation, a phenotype was simulated to have narrow-sense heritability of $\{0.1, 0.3, 0.5, 0.7, 0.9\}$. The genetic contribution was simulated from a multivariate normal with covariance equal to the Manhattan distance between genomes. For homoskedastic simulations, the environmental contribution was simulated from a multivariate normal with identity covariance. For heteroskedastic simulations, the covariance was a diagonal matrix, where one fifth of the values were each of (0.25, 0.5, 1, 2, 4).

A genome scan was conducted on each simulated set of genomes and phenotypes using the linear model (LM), the gold-standard implementation of EMMA (Kang et al., 2008), which uses $\mathbf{M}_{\text{hom}}$, my implementation of the EMMA algorithm, also using $\mathbf{M}_{\text{hom}}$ (ISAM), and my implementation, using $\mathbf{M}_{\text{het}}$ (wISAM).

The alternative simulations were identical to the null simulations except for the fact that individuals who have a 1 at the first genetic factor have 0.25 added to their phenotype value.

5.7.2 False positive rate (FPR) Control

I evaluated the FPR control of each test with quantile-quantile (QQ) plots that plot the sorted p values against the quantiles of the uniform distribution. The ideal behavior of all tests under the null would result in a straight line, starting at the origin and having slope of 1. For an anti-conservative test, the realized p values are "too low", and therefore the line on the QQ plot will be underneath the ideal line. For a conservative test, the realized p values are "too high", and therefore the line on the QQ plot will be underneath the ideal line.

For example, the simulation with 100 organisms and $h^2 = 0.5$ showed that all four tests are anti-conservative and that the weighted ISAM, the only method to use $\mathbf{M}_{\text{het}}$ is the least anti-conservative (Figure 5.1). The QQ plot shown in Figure 5.1 is zoomed to the range [0, 0.01] in both the theoretical (horizontal) and realized (vertical) axis. I show the rest of the QQ plots with increasing zoom from the range [0, 1] to the range [0, 0.001] to show both the global behavior and the behavior in the zone that really matters for large scale analysis, the very small p values. Given the large number of hypotheses that are tested in a GWAS and the multiple hypothesis testing corrections that must be made to account for the large number of hypotheses, the most relevant p-value cutoffs to evaluate are the very small ones.

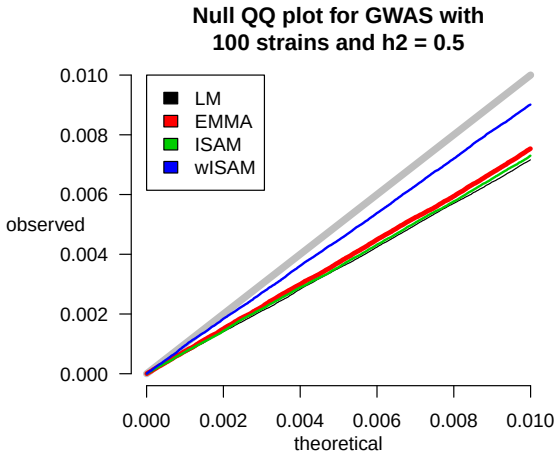


Figure 5.1: This quantile-quantile (QQ) plot of the p values from all four tests shows that they all are anti-conservative and that the weighted ISAM, the only method to use $\mathbf{M}_{\text{het}}$ is the least anti-conservative.

Figure 5.2, Figure 5.3, and Figure 5.4 show the behavior of all four tests across all simulation scenarios. Some consistent patterns emerge.

First, the higher the heritability and the larger the strain panel, the more anti-conservative is the LM. In the most extreme case (Figure 5.4, row 4, column 5), the observed p-values rise less than a hundredth of distance they are expected to over the interval [0, 0.001]. Practically speaking, this observation implies that there is one p value less than 0.001 out of every 10 tests, though should only be one out of every 1,000 tests. This pattern is present to varying degrees in traits with $h^2 \in 0.7, 0.9$ for all panel sizes.

Second, there is never any meaningful difference between EMMA and my unweighted ISAM implementation. This is consistent with a correct implementation, as these tests are theoretically identical.

Third, the EMMA and ISAM tests are overly-conservative when applied to traits with low heritability, especially when few strains are used. The most extreme observation of this behavior can be seen in Figure 5.2, row 4, column 1. The observed p values reach the top of the plot about half way across, indicating that there is one p value less than 0.001 every approximately every 20,000 tests, though there should be one every 10,000 tests. This pattern is present to varying degrees for traits with $h^2 \in 0.1, 0.3$ in a panel with 50 or 100 strains and for a trait with $h^2 = 0.1$ in a panel of 200 organisms.

Finally, the EMMA and ISAM tests are anti-conservative when applied to traits with high heritability when few strains are used. In the most extreme case (Figure 5.2, row 4, column 5), the observed p-values rise about half the distance they are expected to over the interval [0, 0.001]. This pattern indicates that there are two p-values less than 0.001 out of every 1,000 tests, while there should be only one. Said another way, there is one p-value less than 0.001 for every 5,00 tests, while there should be one every 1,000 tests. The weighted ISAM, the only test that uses $\mathbf{M}_{\text{het}}$ rather than $\mathbf{M}_{\text{hom}}$ rises to about three quarters of its expected height, indicating that there are about 1.5 p-values less than 0.001 out of every 1,000 tests, while there should be only one. This behavior is still not ideal, as there should be only 1 p-value less than 0.001 out of every 1,000 tests, but it is closer to ideal than the EMMA and ISAM tests, which use $\mathbf{M}_{\text{hom}}$.

5.7.3 Discrimination between real and spurious signals

Another important way to evaluate the behavior of the tests to compare their ability to discriminate real from spurious signals. Where FPR control under the null used only null simulations, this evaluation combines information from null and alternative simulations. To compare the discrimination of weighted and unweighted LMM-based tests, I compared their receiver operating characteristic (ROC) curves.

This method of evaluation involves collecting test statistics from both null and alternative data and calculating, for each possible cutoff, the true positive rate and the false positive rate. For a cutoff, c, the true positive rate is the fraction of alternative that have a test statistic greater than the cutoff. Similarly, the false positive rate is the fraction of null simulations that have a test statistic greater than the cutoff.

Two points are guaranteed to be in every ROC plot. If $\inf$ is used as the cutoff, no tests will be called positive, so both the true positive rate and the false positive rate will be 0 and thus (0, 0) is on every ROC curve. If $-\inf$ is used as the cutoff all tests will be called positive, so both the true positive rate and the false positive rate will be 1 and thus (1, 1) is on every ROC curve. For a test with no ability to discriminate between null and alternative simulations, the ROC curve progresses from (0, 0) to (1, 1) along a straight line with slope 1. For a test that perfectly discriminates between null and alternative simulations, the ROC curve progresses directly up from (0, 0) to (0, 1) and then directly to the right to (1, 1). Most tests have intermediate discrimination ability, between random

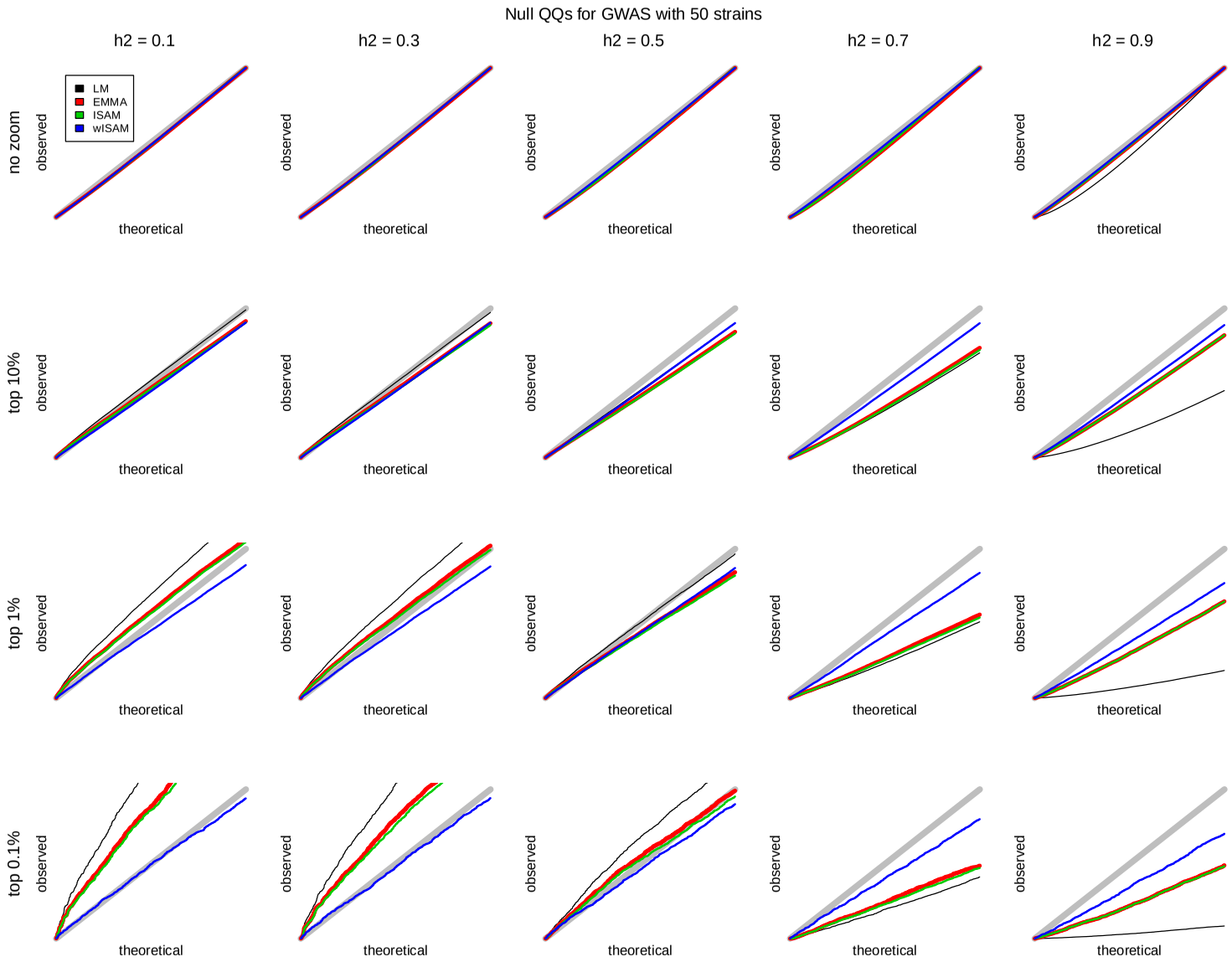


Figure 5.2: QQ plots for simulations with 50 organisms (or strains) in the mapping panel. Heritability increases from left to right and zoom increases from top to bottom. For traits with low and very low heritability (0.1 and 0.3), wISAM is uniquely resistant to overly conservative behavior. For traits with high and very high heritability (0.7 and 0.9), wISAM is uniquely resistant to anti-conservative behavior. For traits with heritability of 0.5, all tests have nearly-accurate FPR control.

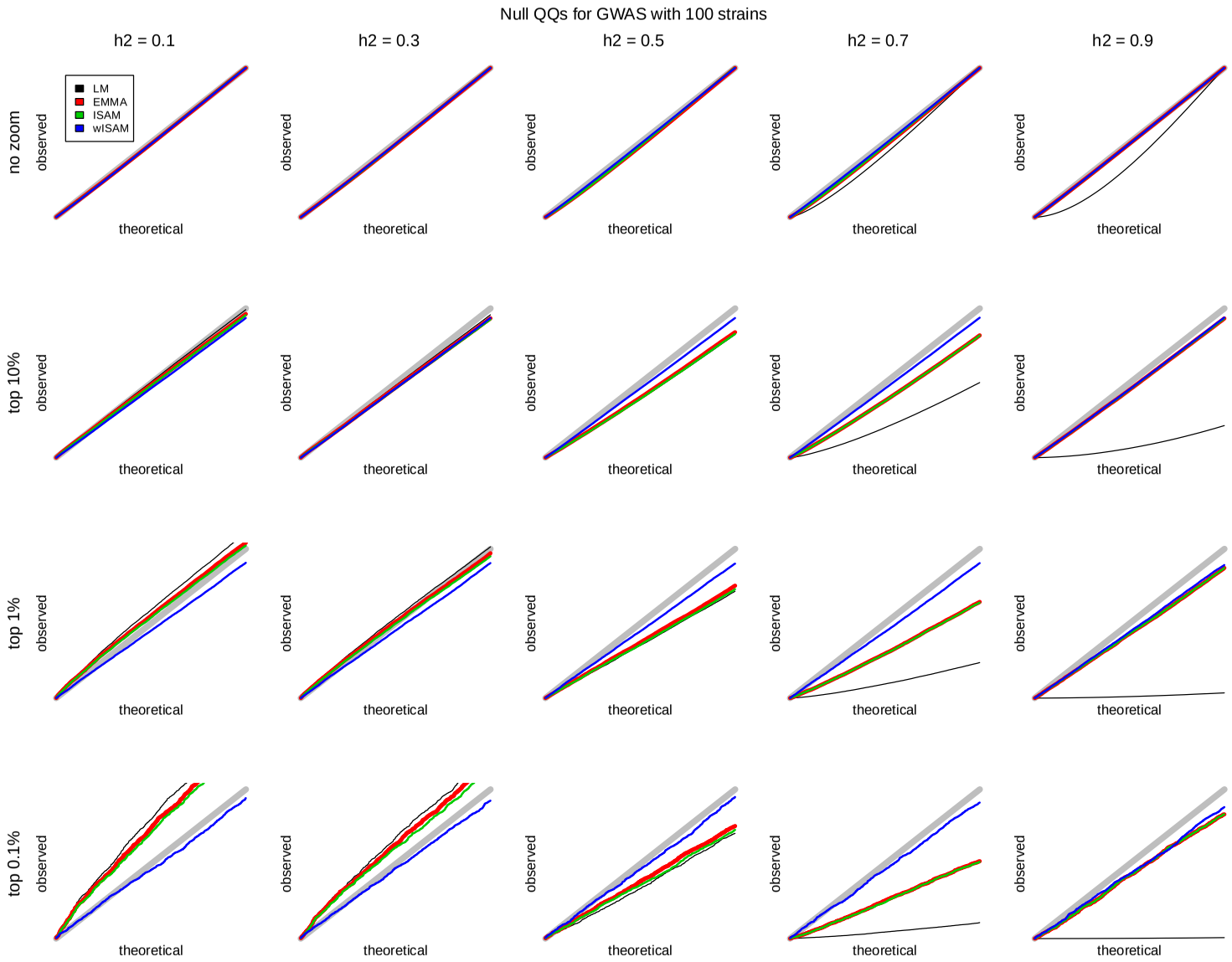


Figure 5.3: QQ plots for simulations with 100 organisms (or strains) in the mapping panel. Heritability increases from left to right and zoom increases from top to bottom. For traits with low and very low heritability (0.1 and 0.3), wISAM is uniquely resistant to overly conservative behavior. For traits with moderate and high heritability (0.5 and 0.7), wISAM is uniquely resistant to anti-conservative behavior. For traits with heritability of 0.9, all LMM-based tests have nearly-accurate FPR control.

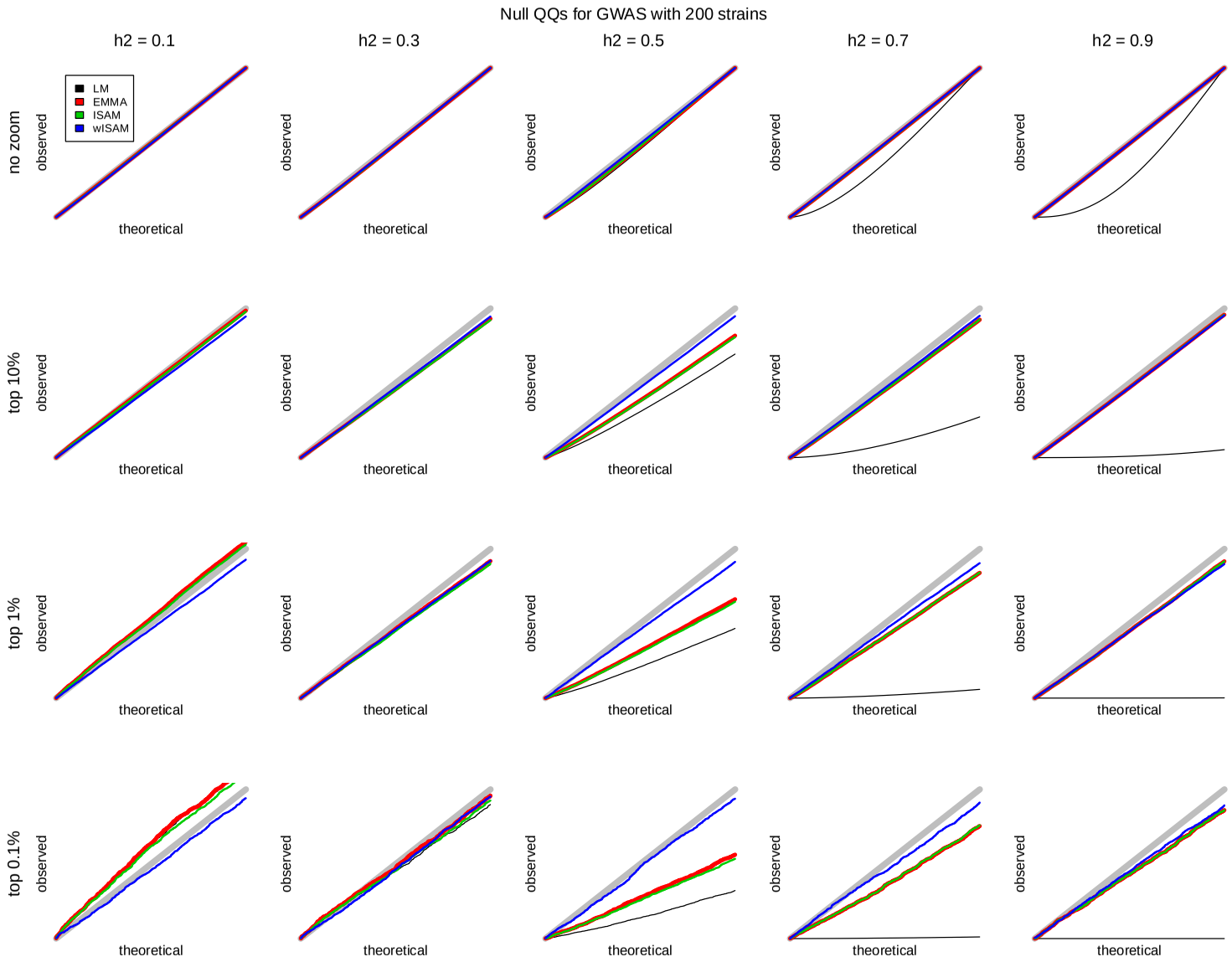


Figure 5.4: QQ plots for simulations with 200 organisms (or strains) in the mapping panel. Heritability increases from left to right and zoom increases from top to bottom. For traits with heritability of 0.1, wISAM is uniquely resistant to overly conservative behavior. For traits with heritability of 0.3, all tests have nearly-accurate FPR control. For traits with moderate to high heritability (0.5 to 0.7), wISAM is uniquely resistant to anti-conservative behavior. For traits with heritability of 0.9, all LLM-based tests have nearly-accurate FPR control.

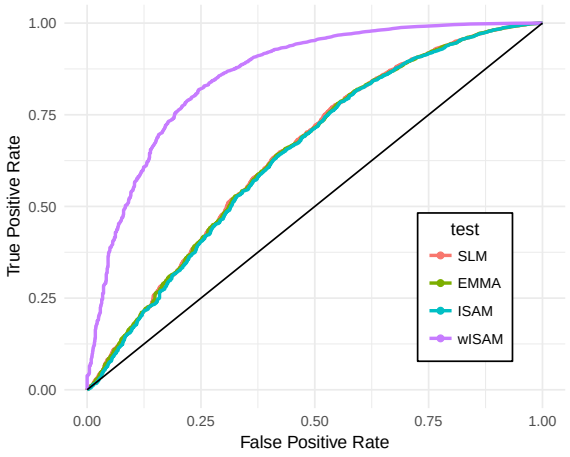


Figure 5.5: Receiver operating characteristics (ROC) curve for a GWAS with 100 organisms on a trait with $h^2 = 0.05$. The ROC curve for the weighted ISAM is superior to that of the other LMM-based methods and the LM.

guessing and perfect, so their ROC curve progresses through the upper left portion of the plot. The closer the ROC curve is to that of a perfect test, the better the test.

ROC analysis exploring the entire parameter space describe for the FPR control under the null has not yet been completed. Here, I show the ROC curve that resulted for a study with $h^2 = 0.5$ and 100 organisms, the middle-of-the-road parameter set Figure 5.5.

5.8 Software

I implemented a linear mixed model fitter that can handle heteroskedastic residual variance using the method described in this chapter in `R` package `wISAM`. This package is tailored to conduct GWAS using the heteroskedastic LMM. It is freely available on github at `https://github.com/rcorty/wISAM` and on CRAN at `https://CRAN.R-project.org/package=wISAM`.

CHAPTER 6
Conclusion and Future Directions

6.1 Summary

In section 1.5, I explored the potential of the double generalized linear model (DGLM) to detect novel QTL in linkage disequilibrium (LD) mapping experiments. I described how the DGLM, unlike the standard linear model (SLM), can detect mean QTL, variance QTL, and joint mean-variance QTL. This work was based on long-public, but little used statistical methods (Smyth, 1989) applied to genetics experiments in a way that is relatively novel (Paré et al., 2010; Rönnegård and Valdar, 2011). I extended previous work by developing a permutation approach that accurately controls the false positive rate (FPR) of an individual test to the desired level and can be applied naturally in a genome-wide context to accurately control the family-wise error rate (FWER). Additionally, I developed novel plots for visualizing and interpreting the results of a genome scan based on these tests and have distributed software that implements this framework in `R` package `vqtl`, which is available on `CRAN` and is interoperable with the popular `R/qtl`.

I retrieved data from the Mouse Phenome Database to test my framework for QTL mapping with the DGLM and found three novel QTL. In reanalyzing the data from Bailey et al. (2008), I discovered a novel vQTL for rearing behavior, which was not detected previously because the standard analysis framework does not detect vQTL. In reanalyzing the data from Kumar et al. (2013), I discovered a novel mQTL for circadian behavior. The additional power of the DGLM-based test to detect this QTL came from accommodating the variance heterogeneity across alleles at the QTL. In reanalyzing the data from Leamy et al. (2000), I discovered a novel mQTL for bodyweight at three weeks of age. The additional power of the DGLM-based test to detect this QTL came from accommodating the variance heterogeneity across levels of a nuisance covariate — which father sired the mouse on which the bodyweight was measured. Gaining power to detect a QTL in this manner was novel and therefore I investigate the statistical power of DGLM-based tests in cases of what

I've termed "background variance heterogeneity". Simulations confirmed that when the source of variance heterogeneity is known, the DGLM-based tests for mQTL, vQTL, and mvQTL are uniquely powerful while maintaining accurate FPR control.

In Figure 4.9, I described the utility of the linear mixed model (LMM) for genetic mapping in non-exchangeable populations. In such populations, both the SLM and the DGLM are inappropriate because they cannot account for the differential relatedness of individuals in the mapping population. I described the established procedure for fitting the LMM that is fast enough to make genome-wide analysis tractable, and noted that one of its limitations is its requirement that the environmental variance be identical across all measurements. I went on to report a novel mathematical method for rapidly fitting the LMM that allows for heteroskedastic residual variance, removing a limitation of the previous standard method. I tested this method on simulated data and found it to have beneficial effects on FPR control and power to detect mQTL. I have developed an `R` package that implements GWAS analysis using this method and it is available on `CRAN`. The next step on this project is clear — to apply the software to existing datasets to attempt to characterize the extent to which it changes the results of completed GWAS analyses and to potentially detect new QTL.

6.2 Outstanding Specific Aims

This dissertation to this point describes the results of aims 1a, 1b, and 2a. Briefly, aim 1a was to accommodate variance heterogeneity in an LD mapping panel. This aim was accomplished and its results were reported in the second reanalysis in chapter 3 and chapter 2. Aim 1b was to accommodate variance heterogeneity in an association mapping panel, and the results of that work are reported in chapter 5. Aim 2a was to identify variance heterogeneity in an LD mapping panel, which was reported in the first reanalysis in chapter 3. Here, I will discuss aims 2b and 3.

Aim 2b was to develop a statistical approach to detect vQTL in an inbred strain panel. I began addressing this aim with a two-stage approach. The first step was to use a Bayesian statistical modeling language like JAGS or STAN (Plummer, 2003; Carpenter et al., 2017) to estimate the mean and variance of each strain. The second intended step was to use a standard GWAS approach to detect genetic associations with those strain parameters. Though I did not make meaningful progress on this work to date, I consider it a promising avenue of future research.

Aim 3 was to develop principled methods for combining evidence from LD mapping and association mapping panels. With the benefit of three more years of study, I understand better the challenges that this aim faces. Foremost among them is the question of how to reconcile the fact that a given genetic variant is likely to have drastically different effects depending on the genetic background in which it is expressed with the stated goal to share information across study designs. Based on this challenge, I consider this aim to be a less-promising avenue for future research.

6.3 Human Studies

In the last two years, I worked on an exciting project to apply DGLM-based genetic mapping approaches to cardiovascular disease risk traits in humans. I worked with Ethan and Leslie Lange and Laura Raffield, a postdoctoral researcher in their lab, to attempt to identify genetic loci that influence lipid levels and blood pressure traits. This work did not achieve any meaningful results yet, but I believe it has tremendous potential to identify GxG and GxE interactions through the detection of vQTL.

The problem of differential relatedness was managed by including the first ten principle components of the genome as covariates in both the mean and variance sub-models. This is a heuristic correction, and is less mathematically-sound than the application the linear mixed model described in chapter 5. Given the extremely large sample size involved in human GWAS studies, it is likely that even the efficient implementation of the LMM described here and elsewhere (Kang et al., 2008) will not be tenable for the near future. When I turned my focus away from this project and toward those that I completed, we were dealing with issues related to correction for p-value inflation with genomic control and how that might be different in the mean sub-model as compared to the variance sub-model.

BIBLIOGRAPHY

Allen, H. L., Estrada, K., Lettre, G., Berndt, S. I., Weedon, M. N., Rivadeneira, F., Willer, C. J., Jackson, A. U., Vedantam, S., Raychaudhuri, S., Ferreira, T., Wood, A. R., Weyant, R. J., Segrè, A. V., Speliotes, E. K., Wheeler, E., Soranzo, N., Park, J. H., Yang, J., Gudbjartsson, D., Heard-Costa, N. L., Randall, J. C., Qi, L., Smith, A. V., Mägi, R., Pastinen, T., Liang, L., Heid, I. M., Luan, J., Thorleifsson, G., Winkler, T. W., Goddard, M. E., Lo, K. S., Palmer, C., Workalemahu, T., Aulchenko, Y. S., Johansson, Å., Zillikens, M. C., Feitosa, M. F., Esko, T., Johnson, T., Ketkar, S., Kraft, P., Mangino, M., Prokopenko, I., Absher, D., Albrecht, E., Ernst, F., Glazer, N. L., Hayward, C., Hottenga, J. J., Jacobs, K. B., Knowles, J. W., Kutalik, Z., Monda, K. L., Polasek, O., Preuss, M., Rayner, N. W., Robertson, N. R., Steinthorsdottir, V., Tyrer, J. P., Voight, B. F., Wiklund, F., Xu, J., Zhao, J. H., Nyholt, D. R., Pellikka, N., Perola, M., Perry, J. B., Surakka, I., Tammesoo, M. L., Altmaier, E. L., Amin, N., Aspelund, T., Bhangale, T., Boucher, G., Chasman, D. I., Chen, C., Coin, L., Cooper, M. N., Dixon, A. L., Gibson, Q., Grundberg, E., Hao, K., Junttila, M. J., Kaplan, L. M., Kettunen, J., König, I. R., Kwan, T., Lawrence, R. W., Levinson, D. F., Lorentzon, M., McKnight, B., Morris, A. P., Müller, M., Ngwa, J. S., Purcell, S., Rafelt, S., Salem, R. M., Salvi, E., Sanna, S., Shi, J., Sovio, U., Thompson, J. R., Turchin, M. C., Vandenput, L., Verlaan, D. J., Vitart, V., White, C. C., Ziegler, A., Almgren, P., Balmforth, A. J., Campbell, H., Citterio, L., De Grandi, A., Dominiczak, A., Duan, J., Elliott, P., Elosua, R., Eriksson, J. G., Freimer, N. B., Geus, E. J., Glorioso, N., Haiqing, S., Hartikainen, A. L., Havulinna, A. S., Hicks, A. A., Hui, J., Igl, W., Illig, T., Jula, A., Kajantie, E., Kilpeläinen, T. O., Koiranen, M., Kolcic, I., Koskinen, S., Kovacs, P., Laitinen, J., Liu, J., Lokki, M. L., Marusic, A., Maschio, A., Meitinger, T., Mulas, A., Paré, G., Parker, A. N., Peden, J. F., Petersmann, A., Pichler, I., Pietiläinen, K. H., Pouta, A., Ridderstråle, M., Rotter, J. I., Sambrook, J. G., Sanders, A. R., Schmidt, C. O., Sinisalo, J., Smit, J. H., Stringham, H. M., Walters, G. B., Widen, E., Wild, S. H., Willemsen, G., Zagato, L., Zgaga, L., Zitting, P., Alavere, H., Farrall, M., McArdle, W. L., Nelis, M., Peters, M. J., Ripatti, S., Van Meurs, J. B., Aben, K. K., Ardlie, K. G., Beckmann, J. S., Beilby, J. P., Bergman, R. N., Bergmann, S., Collins, F. S., Cusi, D., Den Heijer, M., Eiriksdottir, G., Gejman, P. V., Hall, A. S., Hamsten, A., Huikuri, H. V., Iribarren, C., Kähönen, M., Kaprio, J., Kathiresan, S., Kiemeney, L., Kocher, T., Launer, L. J., Lehtimäki, T., Melander, O., Mosley, T. H., Musk, A. W., Nieminen, M. S., O'Donnell, C. J., Ohlsson, C., Oostra, B., Palmer, L. J., Raitakari, O., Ridker, P. M., Rioux, J. D., Rissanen, A., Rivolta, C., Schunkert, H., Shuldiner, A. R., Siscovick, D. S., Stumvoll, M., Tönjes, A., Tuomilehto, J., Van Ommen, G. J., Viikari, J., Heath, A. C., Martin, N. G., Montgomery, G. W., Province, M. A., Kayser, M., Arnold, A. M., D.Atwood, L., Boerwinkle, E., Chanock, S. J., Deloukas, P., Gieger, C., Grönberg, H., Hall, P., Hattersley, A. T., Hengstenberg, C., Hoffman, W., Lathrop, G. M., Salomaa, V., Schreiber, S., Uda, M., Waterworth, D., Wright, A. F., Assimes, T. L., Barroso, I., Hofman, A., Mohlke, K. L., Boomsma, D. I., Caulfield, M. J., Cupples, L. A., Erdmann, J., Fox, C. S., Gudnason, V., Gyllensten, U., Harris, T. B., Hayes, R. B., Jarvelin, M. R., Mooser, V., Munroe, P. B., Ouwehand, W. H., Penninx, B. W., Pramstaller, P. P., Quertermous, T., Rudan, I., Samani, N. J., Spector, T. D., Völzke, H., Watkins, H., Wilson, J. F., Groop, L. C., Haritunians, T., Hu, F. B., Kaplan, R. C., Metspalu, A., North, K. E., Schlessinger, D., Wareham, N. J., Hunter, D. J., O'Connell, J. R., Strachan, D. P., Wichmann, H. E., Borecki, I. B., Van Duijn, C. M., Schadt, E. E., Thorsteinsdottir, U., Peltonen, L., Uitterlinden, A. G., Visscher, P. M., Chatterjee, N., Loos, R. J., Boehnke, M., McCarthy, M. I., Ingelsson, E., Lindgren, C. M., Abecasis, G. R.,

Stefansson, K., Frayling, T. M., and Hirschhorn, J. N. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 467(7317):832–838.

Allen Institute for Brain Science (2015). Allen Mouse Brain Atlas. *Allen Mouse Brain Atlas*, 2(November).

Aschard, H., Zaitlen, N., Tamimi, R. M., Lindström, S., and Kraft, P. (2013). A Nonparametric Test to Detect Quantitative Trait Loci Where the Phenotypic Distribution Differs by Genotypes. *Genet. Epidemiol.*, 37(4):323–333.

Ataman, S. L., Cooper, R., Rotimi, C., McGee, D., Osotimehin, B., Kadiri, S., Kingue, S., Muna, W., Fraser, H., Forrester, T., and Wilks, R. (1996). Standardization of blood pressure measurement in an international comparative study. *J. Clin. Epidemiol.*, 49(8):869–77.

Ayroles, J. F., Buchanan, S. M., O'Leary, C., Skutt-Kakaria, K., Grenier, J. K., Clark, A. G., Hartl, D. L., de Bivort, B. L., and Kusters, J. (2015). Behavioral idiosyncrasy reveals genetic control of phenotypic variability. *Proc. Natl. Acad. Sci.*, 1(21).

Bailey, J. S., Grabowski-Boase, L., Steffy, B. M., Wiltshire, T., Churchill, G. A., and Tarantino, L. M. (2008). Identification of quantitative trait loci for locomotor activation and anxiety using closely related inbred strains. *Genes. Brain. Behav.*, 7(7):761–9.

Bandillo, N., Raghavan, C., Muyco, P. A., Sevilla, M. A. L., Lobina, I. T., Dilla-Ermita, C. J., Tung, C. W., McCouch, S., Thomson, M., Mauleon, R., Singh, R. K., Gregorio, G., Redoña, E., and Leung, H. (2013). Multi-parent advanced generation inter-cross (MAGIC) populations in rice: Progress and potential for genetics research and breeding. *Rice*, 6(1):1–15.

Banks, G., Heise, I., Starbuck, B., Osborne, T., Wisby, L., Potter, P., Jackson, I. J., Foster, R. G., Peirson, S. N., and Nolan, P. M. (2015). Genetic background influences age-related decline in visual and nonvisual retinal responses, circadian rhythms, and sleep. *Neurobiol. Aging*, 36(1):380–393.

Beasley, T. M., Erickson, S., Public, R., Building, H., and Allison, D. B. (2009). Rank-Based Inverse Normal Transformations are Increasingly Used, but are they Merited? *Behav. Genet.*, 39(5):580–595.

Bogue, M. a., Churchill, G. a., and Chesler, E. J. (2015). Collaborative Cross and Diversity Outbred data resources in the Mouse Phenome Database. *Mamm. Genome*, 1(Cc).

Broman, K. W. (2010). *A Guide to QTL Mapping with R*, volume 32.

Broman, K. W. and Sen, Ś. (2009). Single QTL Analysis. In *A Guid. to QTL Mapp. with R/qtl*, pages 75–133.

Broman, K. W., Wu, H., Sen, Ś., and Churchill, G. A. (2003). R/qtl: QTL mapping in experimental crosses. *Bioinformatics*, 19(7):889–890.

Brown, A. A. (2017). veqtl-mapper: variance association mapping for molecular phenotypes. *Bioinformatics*, 33(17):2772–2773.

Brown, A. A., Buil, A., Viñuela, A., Lappalainen, T., Zheng, H.-F., Richards, J. B., Small, K. S., Spector, T. D., Dermitzakis, E. T., and Durbin, R. (2014). Genetic interactions affecting human gene expression identified by variance association mapping. *Elife*, 3:e01381.

Brown, C. S., Thomas, N. S., Sarfarazi, M., Davies, K. E., Kunkel, L., Pearson, P. L., Kingston, H. M., Shaw, D. J., and Harper, P. S. (1985). Genetic linkage relationships of seven DNA probes with Duchenne and Becker muscular dystrophy.

Brown, M. B. and Forsythe, A. B. (1973). Robust Tests for the Equality of Variances. *J. Am. Stat. Assoc.*, 69(346):364–367.

Cao, Y., Wei, P., Bailey, M., Kauwe, J. S. K., and Maxwell, T. J. (2014). A versatile omnibus test for detecting mean and variance heterogeneity. *Genet. Epidemiol.*, 38(1):51–59.

Carlborg, O. and Andersson, L. (2002). Use of randomization testing to detect multiple epistatic QTLs. *Genet. Res.*, 79(2):175–184.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *J. Stat. Softw.*, 76(1):1–32.

Churchill, G. A. and Doerge, R. W. (1994). Empirical Threshold Values for Quantitative Trait Mapping. *Genetics*, 138(3):963–971.

Churchill, G. A. and Doerge, R. W. (2008). Naive application of permutation testing leads to inflated type I error rates. *Genetics*, 178(1):609–610.

Clay, J. S., Vinson, W. E., and White, J. M. (1979). Heterogeneity of Daughter Variances of Sires for Milk Yield1. *J. Dairy Sci.*, 62(6):985–989.

Cox, A., Ackert-Bicknell, C. L., Dumont, B. L., Ding, Y., Bell, J. T., Brockmann, G. A., Wergedal, J. E., Bult, C., Paigen, B., Flint, J., Tsaih, S.-W., Churchill, G. A., and Broman, K. W. (2009). A New Standard Genetic Map for the Laboratory Mouse. *Genetics*, 182(4):1335–1344.

Datta, S. and Nettleton, D. (2014). *Statistical Analysis of Next Generation Sequencing Data*.

Dudbridge, F. and Koeleman, B. P. C. (2004). Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies. *Am. J. Hum. Genet.*, 75(3):424–35.

Dworkin, I. (2005). Canalization, cryptic variation, and developmental buffering: A critical examination and analytical perspective. *Variation*, pages 131–158.

Efron, B. (2004). Large-Scale Simultaneous Hypothesis Testing. *J. Am. Stat. Assoc.*, 99(465):96–104.

Elston, R. C. and Stewart, J. (1971). A general model for the genetic analysis of pedigree data. *Hum. Hered.*, 21(6):523–542.

Falconer, D. S. (1965). The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Ann. Hum. Genet.*, 29:51–76.

Falconer, D. S. and Mackay, T. (1995). lntroduction to quantitative genetics.

Fasoula, D. A. (2012). Nonstop selection for high and stable crop yield by two prognostic equations to reduce yield losses. *Collect. FAO Agric.*, 2(3):211–227.

Felix, M.-A. and Barkoulas, M. (2015). Pervasive robustness in biological systems. *Nat. Rev. Genet.*, 16(8):483–496.

Fikse, W. F., Rönnegård, L., Mulder, H. A., and others (2012). Genome-wide assocation study for genetic heterogeneity for milk yield and somatic cell score. *Proceedings of the*.

Forsberg, S., Andreatta, M. E., Huang, X. Y., Danku, J., and others (2015). multi-allelic genetic architecture of a variance-heterogeneity locus for molybdenum concentration in leaves acts as a source of unexplained additive genetic . *Genetics*.

Forsberg, S., Pettersson, M. E., Sheng, Z., and others (2014). Natural CMT2 variation is associated with genome-wide methylation changes and temperature seasonality. *PLoS*.

Forsberg, S. K. G., Bloom, J. S., Sadhu, M. J., Kruglyak, L., and Carlborg, Ö. (2017). Accounting for genetic interactions improves modeling of individual quantitative trait phenotypes in yeast. *Nat. Genet.*, 49(4):497–503.

Fraser, H. B. and Schadt, E. E. (2010). The quantitative genetics of phenotypic robustness. *PLoS One*, 5(1).

Freedman, D. and Lane, D. (1983). A nonstochastic interpretation of reported significance levels. *J. Bus. Econ. Stat.*, 1(4):292–298.

Freund, J., Brandmaier, A. M., Lewejohann, L., Kirste, I., Kritzler, M., Krüger, A., Sachser, N., Lindenberger, U., Kempermann, G., Online, S., and Vrc, A. (2013). Emergence of individuality in genetically identical mice. *Science*, 340(6133):756–9.

Gatti, R. A., Berkel, I., Boder, E., Braedt, G., Charmley, P., Concannon, P., Ersoy, F., Foroud, T., Jaspers, N. G. J., Lange, K., Lathrop, G. M., Leppert, M., Nakamura, Y., O'Connell, P., Paterson, M., Salser, W., Sanal, O., Silver, J., Sparkes, R. S., Susi, E., Weeks, D. E., Wei, S., White, R., and Yoder, F. (1988). Localization of an ataxia-telangiectasia gene to chromosome 11q2223. *Nature*, 336(6199):577–580.

Ghazalpour, A., Rau, C. D., Farber, C. R., Bennett, B. J., Orozco, L. D., Van Nas, A., Pan, C., Allayee, H., Beaven, S. W., Civelek, M., Davis, R. C., Drake, T. A., Friedman, R. A., Furlotte, N., Hui, S. T., Jentsch, J. D., Kostem, E., Kang, H. M., Kang, E. Y., Joo, J. W., Korshunov, V. A., Laughlin, R. E., Martin, L. J., Ohmen, J. D., Parks, B. W., Pellegrini, M., Reue, K., Smith, D. J., Tetradis, S., Wang, J., Wang, Y., Weiss, J. N., Kirchgessner, T., Gargalovic, P. S., Eskin, E., Lusis, A. J., and LeBoeuf, R. C. (2012). Hybrid mouse diversity panel: A panel of inbred mouse strains suitable for analysis of complex genetic traits.

Gibson, G. and Wagner, G. (2000). Canalization in evolutionary genetics: A stabilizing theory? *BioEssays*, 22(4):372–380.

Good, P. (2013). *Permutation tests: a practical guide to resampling methods for testing hypotheses*. Springer Science & Business Media.

Gray, M. M., Parmenter, M. D., Hogan, C. A., Ford, I., Cuthber, R. J., Ryan, P. G., Broman, K. W., and Payseur, B. A. (2015). Genetics of rapid and extreme size evolution in Island mice. *Genetics*, 201(1):213–228.

Grubb, S. C., Bult, C. J., and Bogue, M. A. (2014). Mouse Phenome Database. *Nucleic Acids Res.*, 42(D1).

Haley, C. S. and Knott, S. a. (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity (Edinb).*, 69(4):315–24.

Hall, M. C., Dworkin, I., Ungerer, M. C., and Purugganan, M. (2007). Genetics of microenvironmental canalization in Arabidopsis thaliana. *Pnas*, 104(34):13717–13722.

Haseman, J. K. and Elston, R. C. (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behav. Genet.*, 2(1):3–19.

Henderson, C. R. (1984). Applications of Linear Models in Animal Breeding Models. *Univ. Guelph*, page 384.

Hill, W. G. and Mulder, H. A. (2010). Genetic analysis of environmental variation. *Genet. Res. (Camb).*, 92(5-6):381–395.

Hong, C., Ning, Y., Wei, P., Cao, Y., and Chen, Y. (2016). A semiparametric model for vQTL mapping.

Huang, W., Carbone, M. A., Magwire, M. M., Peiffer, J. A., Lyman, R. F., Stone, E. A., Anholt, R. R. H., and Mackay, T. F. C. (2015). Genetic basis of transcriptome diversity in drosophila melanogaster. *Proc. Natl. Acad. Sci. U. S. A.*, 112(44):E6010–9.

Hulse, A. M. and Cai, J. J. (2013). Genetic variants contribute to gene expression variability in humans. *Genetics*, 193(1):95–108.

Ibáñez-Escriche, N., Varona, L., Sorensen, D., and Noguera, J. L. (2008). A study of heterogeneity of environmental variance for slaughter weight in pigs. *Animal*, 2(1):19–26.

Jette, M. and Grondona, M. (2003). SLURM: Simple Linux Utility for Resource Management. In *Clust. Conf. Expo CWCE*, volume 2682, pages 44–60.

Jimenez-Gomez, J. M., Corwin, J. a., Joseph, B., Maloof, J. N., and Kliebenstein, D. J. (2011). Genomic analysis of QTLs and genes altering natural variation in stochastic noise. *PLoS Genet.*, 7(9):e1002295.

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):1709–23.

Kapushesky, M., Emam, I., Holloway, E., Kurnosov, P., Zorin, A., Malone, J., Rustici, G., Williams, E., Parkinson, H., and Brazma, A. (2009). Gene expression Atlas at the European Bioinformatics Institute. *Nucleic Acids Res.*, 38(SUPPL.1).

Keane, T. M., Goodstadt, L., Danecek, P., White, M. A., Wong, K., Yalcin, B., Heger, A., Agam, A., Slater, G., Goodson, M., Furlotte, N. A., Eskin, E., Nellåker, C., Whitley, H., Cleak, J., Janowitz, D., Hernandez-Pliego, P., Edwards, A., Belgard, T. G., Oliver, P. L., McIntyre, R. E., Bhomra, A., Nicod, J., Gan, X., Yuan, W., Van Der Weyden, L., Steward, C. A., Bala, S., Stalker, J., Mott, R., Durbin, R., Jackson, I. J., Czechanski, A., Guerra-Assunçao, J., Donahue, L. R., Reinholdt, L. G., Payseur, B. A., Ponting, C. P., Birney, E., Flint, J., and Adams, D. J. (2011). Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature*, 477(7364):289–294.

Kim, J. H., Park, S. M., Kang, M. R., Oh, S. Y., Lee, T. H., Muller, M. T., and Chung, I. K. (2005). Ubiquitin ligase MKRN1 modulates telomere length homeostasis through a proteolysis of hTERT. *Genes Dev.*, 19(7):776–781.

King, E. G., Macdonald, S. J., and Long, A. D. (2012). Properties and power of the Drosophila synthetic population resource for the routine dissection of complex traits. *Genetics*, 191(3):935–949.

Kirby, A., Kang, H. M., Wade, C. M., Cotsapas, C., Kostem, E., Han, B., Furlotte, N., Kang, E. Y., Rivas, M., Bogue, M. A., Frazer, K. A., Johnson, F. M., Beilharz, E. J., Cox, D. R., Eskin, E., and Daly, M. J. (2010). Fine mapping in 94 inbred mouse strains using a high-density haplotype resource. *Genetics*, 185(3):1081–1095.

Kitano, H. (2004). Biological robustness. *Nat. Rev. Genet.*, 5(11):826–837.

Kumar, V., Kim, K., Joseph, C., Kourrich, S., Yoo, S.-H., Huang, H. C., Vitaterna, M. H., Pardo-Manuel de Villena, F., Churchill, G., Bonci, A., and Takahashi, J. S. (2013). C57BL/6N Mutation in Cytoplasmic FMRP interacting protein 2 Regulates Cocaine Response. *Science (80-.).*, 342(6165):1508–1512.

Labarthe, D. R., Hawkins, C. M., and Remington, R. D. (1973). Evaluation of performance of selected devices for measuring blood pressure. *Am. J. Cardiol.*, 32(4):546–553.

Lander, E. S. and Botstein, S. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, 121(1):185.

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

Leamy, L. J., Pomp, D., Eisen, E. J., and Cheverud, J. M. (2000). Quantitative trait loci for directional but not fluctuating asymmetry of mandible characters in mice. *Genet. Res.*, 76(1):27–40.

Leamy, L. J., Pomp, D., Eisen, E. J., and Cheverud, J. M. (2002). Pleiotropy of quantitative trait loci for organ weights and limb bone lengths in mice. *Physiol. Genomics*, 10(1):21–9.

Lee, C. R., Anderson, J. T., and Mitchell-Olds, T. (2014). Unifying genetic canalization, genetic constraint, and genotype-by-environment interaction: QTL by genomic background by environment interaction of . *PLoS Genet.*

Levene, H. (1960). Robust tests for equality of variances. In Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling. *Stanford Univ. Press*, pages 278–292.

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):833–5.

Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., Day, F. R., Powell, C., Vedantam, S., Buchkovich, M. L., Yang, J., Croteau-Chonka, D. C., Esko, T., Fall, T., Ferreira, T., Gustafsson, S., Kutalik, Z., Luan, J., Mägi, R., Randall, J. C., Winkler, T. W., Wood, A. R., Workalemahu, T., Faul, J. D., Smith, J. A., Zhao, J. H., Zhao, W., Chen, J., Fehrmann, R., Hedman, Å. K., Karjalainen, J., Schmidt, E. M., Absher, D., Amin, N., Anderson, D., Beekman, M., Bolton, J. L., Bragg-Gresham, J. L., Buyske, S., Demirkan, A., Deng, G., Ehret, G. B., Feenstra, B., Feitosa, M. F., Fischer, K., Goel, A., Gong, J., Jackson, A. U., Kanoni, S., Kleber, M. E., Kristiansson, K., Lim, U., Lotay, V., Mangino, M., Leach, I. M., Medina-Gomez, C., Medland, S. E., Nalls, M. A., Palmer, C. D., Pasko, D., Pechlivanis, S., Peters, M. J., Prokopenko, I., Shungin, D., Stančáková, A., Strawbridge, R. J., Sung, Y. J., Tanaka, T., Teumer, A., Trompet, S., Van Der Laan, S. W., Van Setten, J., Van Vliet-Ostaptchouk, J. V., Wang, Z., Yengo, L.,

Zhang, W., Isaacs, A., Albrecht, E., Ärnlöv, J., Arscott, G. M., Attwood, A. P., Bandinelli, S., Barrett, A., Bas, I. N., Bellis, C., Bennett, A. J., Berne, C., Blagieva, R., Blüher, M., Böhringer, S., Bonnycastle, L. L., Böttcher, Y., Boyd, H. A., Bruinenberg, M., Caspersen, I. H., Chen, Y. D. I., Clarke, R., Daw, E. W., De Craen, A. J., Delgado, G., Dimitriou, M., Doney, A. S., Eklund, N., Estrada, K., Eury, E., Folkersen, L., Fraser, R. M., Garcia, M. E., Geller, F., Giedraitis, V., Gigante, B., Go, A. S., Golay, A., Goodall, A. H., Gordon, S. D., Gorski, M., Grabe, H. J., Grallert, H., Grammer, T. B., Gräßler, J., Grönberg, H., Groves, C. J., Gusto, G., Haessler, J., Hall, P., Haller, T., Hallmans, G., Hartman, C. A., Hassinen, M., Hayward, C., Heard-Costa, N. L., Helmer, Q., Hengstenberg, C., Holmen, O., Hottenga, J. J., James, A. L., Jeff, J. M., Johansson, Å., Jolley, J., Juliusdottir, T., Kinnunen, L., Koenig, W., Koskenvuo, M., Kratzer, W., Laitinen, J., Lamina, C., Leander, K., Lee, N. R., Lichtner, P., Lind, L., Lindström, J., Lo, K. S., Lobbens, S., Lorbeer, R., Lu, Y., Mach, F., Magnusson, P. K., Mahajan, A., McArdle, W. L., McLachlan, S., Menni, C., Merger, S., Mihailov, E., Milani, L., Moayyeri, A., Monda, K. L., Morken, M. A., Mulas, A., Müller, G., Müller-Nurasyid, M., Musk, A. W., Nagaraja, R., Nöthen, M. M., Nolte, I. M., Pilz, S., Rayner, N. W., Renstrom, F., Rettig, R., Ried, J. S., Ripke, S., Robertson, N. R., Rose, L. M., Sanna, S., Scharnagl, H., Scholtens, S., Schumacher, F. R., Scott, W. R., Seufferlein, T., Shi, J., Smith, A. V., Smolonska, J., Stanton, A. V., Steinthorsdottir, V., Stirrups, K., Stringham, H. M., Sundström, J., Swertz, M. A., Swift, A. J., Syvänen, A. C., Tan, S. T., Tayo, B. O., Thorand, B., Thorleifsson, G., Tyrer, J. P., Uh, H. W., Vandenput, L., Verhulst, F. C., Vermeulen, S. H., Verweij, N., Vonk, J. M., Waite, L. L., Warren, H. R., Waterworth, D., Weedon, M. N., Wilkens, L. R., Willenborg, C., Wilsgaard, T., Wojczynski, M. K., Wong, A., Wright, A. F., Zhang, Q., Brennan, E. P., Choi, M., Dastani, Z., Drong, A. W., Eriksson, P., Franco-Cereceda, A., Gådin, J. R., Gharavi, A. G., Goddard, M. E., Handsaker, R. E., Huang, J., Karpe, F., Kathiresan, S., Keildson, S., Kiryluk, K., Kubo, M., Lee, J. Y., Liang, L., Lifton, R. P., Ma, B., McCarroll, S. A., McKnight, A. J., Min, J. L., Moffatt, M. F., Montgomery, G. W., Murabito, J. M., Nicholson, G., Nyholt, D. R., Okada, Y., Perry, J. R., Dorajoo, R., Reinmaa, E., Salem, R. M., Sandholm, N., Scott, R. A., Stolk, L., Takahashi, A., Tanaka, T., Van T'Hooft, F. M., Vinkhuyzen, A. A., Westra, H. J., Zheng, W., Zondervan, K. T., Heath, A. C., Arveiler, D., Bakker, S. J., Beilby, J., Bergman, R. N., Blangero, J., Bovet, P., Campbell, H., Caulfield, M. J., Cesana, G., Chakravarti, A., Chasman, D. I., Chines, P. S., Collins, F. S., Crawford, D. C., Cupples, L. A., Cusi, D., Danesh, J., De Faire, U., Den Ruijter, H. M., Dominiczak, A. F., Erbel, R., Erdmann, J., Eriksson, J. G., Farrall, M., Felix, S. B., Ferrannini, E., Ferrières, J., Ford, I., Forouhi, N. G., Forrester, T., Franco, O. H., Gansevoort, R. T., Gejman, P. V., Gieger, C., Gottesman, O., Gudnason, V., Gyllensten, U., Hall, A. S., Harris, T. B., Hattersley, A. T., Hicks, A. A., Hindorff, L. A., Hingorani, A. D., Hofman, A., Homuth, G., Hovingh, G. K., Humphries, S. E., Hunt, S. C., Hyppönen, E., Illig, T., Jacobs, K. B., Jarvelin, M. R., Jöckel, K. H., Johansen, B., Jousilahti, P., Jukema, J. W., Jula, A. M., Kaprio, J., Kastelein, J. J., Keinanen-Kiukaanniemi, S. M., Kiemeney, L. A., Knekt, P., Kooner, J. S., Kooperberg, C., Kovacs, P., Kraja, A. T., Kumari, M., Kuusisto, J., Lakka, T. A., Langenberg, C., Le Marchand, L., Lehtimäki, T., Lyssenko, V., Männistö, S., Marette, A., Matise, T. C., McKenzie, C. A., McKnight, B., Moll, F. L., Morris, A. D., Morris, A. P., Murray, J. C., Nelis, M., Ohlsson, C., Oldehinkel, A. J., Ong, K. K., Madden, P. A., Pasterkamp, G., Peden, J. F., Peters, A., Postma, D. S., Pramstaller, P. P., Price, J. F., Qi, L., Raitakari, O. T., Rankinen, T., Rao, D. C., Rice, T. K., Ridker, P. M., Rioux, J. D., Ritchie, M. D., Rudan, I., Salomaa, V., Samani, N. J., Saramies, J., Sarzynski, M. A., Schunkert, H., Schwarz, P. E., Sever, P., Shuldiner, A. R., Sinisalo, J., Stolk, R. P., Strauch, K., Tönjes, A., Trégouët, D. A., Tremblay, A., Tremoli, E., Virtamo, J., Vohl, M. C., Völker, U., Waeber, G., Willemsen, G., Witteman, J. C., Zillikens, M. C., Adair, L. S., Amouyel, P., Asselbergs, F. W., Assimes, T. L., Bochud, M., Boehm, B. O., Boerwinkle,

E., Bornstein, S. R., Bottinger, E. P., Bouchard, C., Cauchi, S., Chambers, J. C., Chanock, S. J., Cooper, R. S., De Bakker, P. I., Dedoussis, G., Ferrucci, L., Franks, P. W., Froguel, P., Groop, L. C., Haiman, C. A., Hamsten, A., Hui, J., Hunter, D. J., Hveem, K., Kaplan, R. C., Kivimaki, M., Kuh, D., Laakso, M., Liu, Y., Martin, N. G., März, W., Melbye, M., Metspalu, A., Moebus, S., Munroe, P. B., Njølstad, I., Oostra, B. A., Palmer, C. N., Pedersen, N. L., Perola, M., Pérusse, L., Peters, U., Power, C., Quertermous, T., Rauramaa, R., Rivadeneira, F., Saaristo, T. E., Saleheen, D., Sattar, N., Schadt, E. E., Schlessinger, D., Slagboom, P. E., Snieder, H., Spector, T. D., Thorsteinsdottir, U., Stumvoll, M., Tuomilehto, J., Uitterlinden, A. G., Uusitupa, M., Van Der Harst, P., Walker, M., Wallaschofski, H., Wareham, N. J., Watkins, H., Weir, D. R., Wichmann, H. E., Wilson, J. F., Zanen, P., Borecki, I. B., Deloukas, P., Fox, C. S., Heid, I. M., O'Connell, J. R., Strachan, D. P., Stefansson, K., Van Duijn, C. M., Abecasis, G. R., Franke, L., Frayling, T. M., McCarthy, M. I., Visscher, P. M., Scherag, A., Willer, C. J., Boehnke, M., Mohlke, K. L., Lindgren, C. M., Beckmann, J. S., Barroso, I., North, K. E., Ingelsson, E., Hirschhorn, J. N., Loos, R. J., and Speliotes, E. K. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518(7538):197–206.

Lynch, M. and Walsh, B. (1998). *Genetics and analysis of quantitative traits*.

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., MayPendlington, Z., Welter, D., Burdett, T., Hindorff, L., Flicek, P., Cunningham, F., and Parkinson, H. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.*, 45(D1):D896–D901.

Mackay, T. F. and Lyman, R. F. (2005). Drosophila bristles and the nature of quantitative genetic variation. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, 360(1459):1513–1527.

MacKay, T. F., Richards, S., Stone, E. A., Barbadilla, A., Ayroles, J. F., Zhu, D., Casillas, S., Han, Y., Magwire, M. M., Cridland, J. M., Richardson, M. F., Anholt, R. R., Barrón, M., Bess, C., Blankenburg, K. P., Carbone, M. A., Castellano, D., Chaboub, L., Duncan, L., Harris, Z., Javaid, M., Jayaseelan, J. C., Jhangiani, S. N., Jordan, K. W., Lara, F., Lawrence, F., Lee, S. L., Librado, P., Linheiro, R. S., Lyman, R. F., MacKey, A. J., Munidasa, M., Muzny, D. M., Nazareth, L., Newsham, I., Perales, L., Pu, L. L., Qu, C., Ràmia, M., Reid, J. G., Rollmann, S. M., Rozas, J., Saada, N., Turlapati, L., Worley, K. C., Wu, Y. Q., Yamamoto, A., Zhu, Y., Bergman, C. M., Thornton, K. R., Mittelman, D., and Gibbs, R. A. (2012). The Drosophila melanogaster Genetic Reference Panel. *Nature*, 482(7384):173–178.

Mackay, T. F. C. (2014). Epistasis and quantitative traits: using model organisms to study gene-gene interactions. *Nat. Rev. Genet.*, 15(1):22–33.

Marchand, P. (2017). *rslurm: Submit R Calculations to a Slurm Cluster*.

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., and Gilad, Y. (2008). RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, 18(9):1509–1517.

Martínez, O. and Curnow, R. N. (1992). Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. *Theor. Appl. Genet.*, 85(4):480–488.

McMullen, M. D., Kresovich, S., Villeda, H. S., Bradbury, P., Li, H., Sun, Q., Flint-Garcia, S., Thornsberry, J., Acharya, C., Bottoms, C., Brown, P., Browne, C., Eller, M., Guill, K., Harjes, C., Kroon, D., Lepak, N., Mitchell, S. E., Peterson, B., Pressoir, G., Romero, S., Rosas, M. O.,

Salvo, S., Yates, H., Hanson, M., Jones, E., Smith, S., Glaubitz, J. C., Goodman, M., Ware, D., Holland, J. B., and Buckler, E. S. (2009). Genetic properties of the maize nested association mapping population. *Science (80-.).*, 325(5941):737–740.

McWilliam, H., Li, W., Uludag, M., Squizzato, S., Park, Y. M., Buso, N., Cowley, A. P., and Lopez, R. (2013). Analysis Tool Web Services from the EMBL-EBI. *Nucleic Acids Res.*, 41(Web Server issue).

Meiklejohn, C. D. and Hartl, D. L. (2002). A single mode of canalization. *Trends Ecol. Evol.*, 17(10):468–473.

Mulder, H. A., Bijma, P., and Hill, W. G. (2008). Selection for uniformity in livestock by exploiting genetic heterogeneity of residual variance. *Genet. Sel. Evol.*, 40(1):37–59.

Mulder, H. a., Hill, W. G., and Knol, E. F. (2015). Heritable environmental variance causes nonlinear relationships between traits: application to birth weight and stillbirth of pigs. *Genetics*, 199(4):1255–69.

Murray, J. M., Davies, K. E., Harper, P. S., Meredith, L., Mueller, C. R., and Williamson, R. (1982). Linkage relationship of a cloned DNA sequence on the short arm of the X chromosome to Duchenne muscular dystrophy. *Nature*, 300(5887):69–71.

Nelson, R. M., Pettersson, M. E., Li, X., and Carlborg, Ö. (2013). Variance heterogeneity in saccharomyces cerevisiae expression data: trans-regulation and epistasis. *PLoS One*, 8(11):e79507.

Nettleton, D. (2006). A discussion of statistical methods for design and analysis of microarray experiments for plant scientists. *Plant Cell*, 18(September):2112–2121.

O'Brien, E. (2001). State of the market for devices for blood pressure measurement. *Blood Press. Monit.*, 6(6):281–6.

of the PGC, M. D. D. W. G., Wray, N. R., and Sullivan, P. F. (2017). Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *bioRxiv*, page 167577.

Ordas, B., Malvar, R. a., and Hill, W. G. (2008). Genetic variation and quantitative trait loci associated with developmental stability and the environmental correlation between traits in maize. *Genet. Res. (Camb).*, 90(5):385–395.

Palmer, A. R. and Strobeck, C. (1986). Fluctuating Asymmetry: Measurement, Analysis, Patterns. *Annu. Rev. Ecol. Syst.*, 17(1):391–421.

Paré, G., Cook, N. R., Ridker, P. M., and Chasman, D. I. (2010). On the use of variance per genotype as a tool to identify quantitative trait interaction effects: A report from the women's genome health study. *PLoS Genet.*, 6(6):1–10.

Pavličev, M. and Cheverud, J. M. (2015). Constraints evolve: Context dependency of gene effects allows evolution of pleiotropy.

Payseur, B. A. and Place, M. (2007). Prospects for association mapping in classical inbred mouse strains. *Genetics*, 175(4):1999–2008.

Pettersson, M. E. and Carlborg, Ö. (2015). Capacitating Epistasis—Detection and role in the genetic architecture of complex traits. In Moore, J. H. and Williams, S. M., editors, *Epistasis: Methods and Protocols*, pages 185–196. Springer New York, New York, NY.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proc. 3rd Int. Work. Distrib. Stat. Comput. (DSC 2003)*, pages 20–22.

Queitsch, C., Sangster, T. A., and Lindquist, S. (2002). Hsp90 as a capacitor of phenotypic variation. *Nature*, 417(6889):618–624.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ripke, S., Neale, B. M., Corvin, A., Walters, J. T., Farh, K. H., Holmans, P. A., Lee, P., Bulik-Sullivan, B., Collier, D. A., Huang, H., Pers, T. H., Agartz, I., Agerbo, E., Albus, M., Alexander, M., Amin, F., Bacanu, S. A., Begemann, M., Belliveau, R. A., Bene, J., Bergen, S. E., Bevilacqua, E., Bigdeli, T. B., Black, D. W., Bruggeman, R., Buccola, N. G., Buckner, R. L., Byerley, W., Cahn, W., Cai, G., Campion, D., Cantor, R. M., Carr, V. J., Carrera, N., Catts, S. V., Chambert, K. D., Chan, R. C., Chen, R. Y., Chen, E. Y., Cheng, W., Cheung, E. F., Chong, S. A., Cloninger, C. R., Cohen, D., Cohen, N., Cormican, P., Craddock, N., Crowley, J. J., Curtis, D., Davidson, M., Davis, K. L., Degenhardt, F., Del Favero, J., Demontis, D., Dikeos, D., Dinan, T., Djurovic, S., Donohoe, G., Drapeau, E., Duan, J., Dudbridge, F., Durmishi, N., Eichhammer, P., Eriksson, J., Escott-Price, V., Essioux, L., Fanous, A. H., Farrell, M. S., Frank, J., Franke, L., Freedman, R., Freimer, N. B., Friedl, M., Friedman, J. I., Fromer, M., Genovese, G., Georgieva, L., Giegling, I., Giusti-Rodríguez, P., Godard, S., Goldstein, J. I., Golimbet, V., Gopal, S., Gratten, J., De Haan, L., Hammer, C., Hamshere, M. L., Hansen, M., Hansen, T., Haroutunian, V., Hartmann, A. M., Henskens, F. A., Herms, S., Hirschhorn, J. N., Hoffmann, P., Hofman, A., Hollegaard, M. V., Hougaard, D. M., Ikeda, M., Joa, I., Julià, A., Kahn, R. S., Kalaydjieva, L., Karachanak-Yankova, S., Karjalainen, J., Kavanagh, D., Keller, M. C., Kennedy, J. L., Khrunin, A., Kim, Y., Klovins, J., Knowles, J. A., Konte, B., Kucinskas, V., Kucinskiene, Z. A., Kuzelova-Ptackova, H., Kähler, A. K., Laurent, C., Keong, J. L. C., Lee, S. H., Legge, S. E., Lerer, B., Li, M., Li, T., Liang, K. Y., Lieberman, J., Limborska, S., Loughland, C. M., Lubinski, J., Lönnqvist, J., Macek, M., Magnusson, P. K., Maher, B. S., Maier, W., Mallet, J., Marsal, S., Mattheisen, M., Mattingsdal, M., McCarley, R. W., McDonald, C., McIntosh, A. M., Meier, S., Meijer, C. J., Melegh, B., Melle, I., Mesholam-Gately, R. I., Metspalu, A., Michie, P. T., Milani, L., Milanova, V., Mokrab, Y., Morris, D. W., Mors, O., Murphy, K. C., Murray, R. M., Myin-Germeys, I., Müller-Myhsok, B., Nelis, M., Nenadic, I., Nertney, D. A., Nestadt, G., Nicodemus, K. K., Nikitina-Zake, L., Nisenbaum, L., Nordin, A., O'Callaghan, E., O'Dushlaine, C., O'Neill, F. A., Oh, S. Y., Olincy, A., Olsen, L., Van Os, J., Pantelis, C., Papadimitriou, G. N., Papiol, S., Parkhomenko, E., Pato, M. T., Paunio, T., Pejovic-Milovancevic, M., Perkins, D. O., Pietiläinen, O., Pimm, J., Pocklington, A. J., Powell, J., Price, A., Pulver, A. E., Purcell, S. M., Quested, D., Rasmussen, H. B., Reichenberg, A., Reimers, M. A., Richards, A. L., Roffman, J. L., Roussos, P., Ruderfer, D. M., Salomaa, V., Sanders, A. R., Schall, U., Schubert, C. R., Schulze, T. G., Schwab, S. G., Scolnick, E. M., Scott, R. J., Seidman, L. J., Shi, J., Sigurdsson, E., Silagadze, T., Silverman, J. M., Sim, K., Slominsky, P., Smoller, J. W., So, H. C., Spencer, C. C., Stahl, E. A., Stefansson, H., Steinberg, S., Stogmann, E., Straub, R. E., Strengman, E., Strohmaier, J., Stroup, T. S., Subramaniam, M., Suvisaari, J., Svrakic, D. M., Szatkiewicz, J. P., Söderman, E., Thirumalai, S., Toncheva, D., Tosato, S., Veijola, J., Waddington, J., Walsh, D., Wang, D., Wang, Q., Webb, B. T., Weiser,

M., Wildenauer, D. B., Williams, N. M., Williams, S., Witt, S. H., Wolen, A. R., Wong, E. H., Wormley, B. K., Xi, H. S., Zai, C. C., Zheng, X., Zimprich, F., Wray, N. R., Stefansson, K., Visscher, P. M., Adolfsson, R., Andreassen, O. A., Blackwood, D. H., Bramon, E., Buxbaum, J. D., Børglum, A. D., Cichon, S., Darvasi, A., Domenici, E., Ehrenreich, H., Esko, T., Gejman, P. V., Gill, M., Gurling, H., Hultman, C. M., Iwata, N., Jablensky, A. V., Jönsson, E. G., Kendler, K. S., Kirov, G., Knight, J., Lencz, T., Levinson, D. F., Li, Q. S., Liu, J., Malhotra, A. K., McCarroll, S. A., McQuillin, A., Moran, J. L., Mortensen, P. B., Mowry, B. J., Nöthen, M. M., Ophoff, R. A., Owen, M. J., Palotie, A., Pato, C. N., Petryshen, T. L., Posthuma, D., Rietschel, M., Riley, B. P., Rujescu, D., Sham, P. C., Sklar, P., St Clair, D., Weinberger, D. R., Wendland, J. R., Werge, T., Daly, M. J., Sullivan, P. F., and O'Donovan, M. C. (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511(7510):421–427.

Rönnegård, L. and Valdar, W. (2011). Detecting major genetic loci controlling phenotypic variability in experimental crosses. *Genetics*, 188(2):435–447.

Rönnegård, L. and Valdar, W. (2012). Recent developments in statistical methods for detecting genetic loci affecting phenotypic variability. *BMC Genet.*, 13:63.

Ros, M., Sorensen, D., Waagepetersen, R., Dupont-Nivet, M., SanCristobal, M., Bonnet, J. C., and Mallard, J. (2004). Evidence for genetic control of adult weight plasticity in the snail Helix aspersa. *Genetics*, 168(4):2089–2097.

Rowe, S. J., White, I. M., Avendaño, S., and Hill, W. G. (2006). Genetic heterogeneity of residual variance in broiler chickens. *Genet. Sel. Evol.*, 38(6):617–635.

Sachs, M. C. and Others (2017). plotROC: A Tool for Plotting ROC Curves. *J. Stat. Softw.*, 79(c02).

Shen, X. and Carlborg, ö. (2013). Beware of risk for increased false positive rates in genome-wide association studies for phenotypic variability. *Front. Genet.*, 4(MAY).

Shen, X., Pettersson, M., Ronnegard, L., and Carlborg, O. (2012). Inheritance beyond plain heritability: variance-controlling genes in Arabidopsis thaliana. *PLoS Genet.*, 8(8):e1002839.

Shen, X. and Ronnegard, L. (2013). Issues with data transformation in genome-wide association studies for phenotypic variability. *F1000Research*, 2:200.

Shimomura, K., Low-Zeddies, S. S., King, D. P., Steeves, T. D., Whiteley, A., Kushla, J., Zemenides, P. D., Lin, A., Vitaterna, M. H., Churchill, G. A., and Takahashi, J. S. (2001). Genome-wide epistatic interaction analysis reveals complex genetic determinants of circadian behavior in mice. *Genome Res.*, 11(6):959–80.

Simon, M. M., Greenaway, S., White, J. K., Fuchs, H., Gailus-Durner, V., Wells, S., Sorg, T., Wong, K., Bedu, E., Cartwright, E. J., Dacquin, R., Djebali, S., Estabel, J., Graw, J., Ingham, N. J., Jackson, I. J., Lengeling, A., Mandillo, S., Marve, J., Meziane, H., Preitner, F., Puk, O., Roux, M., Adams, D. J., Atkins, S., Ayadi, A., Becker, L., Blake, A., Brooker, D., Cater, H., Champy, M. F., Combe, R., Danecek, P., Di Fenza, A., Gates, H., Gerdin, A. K., Golini, E., Hancock, J. M., Hans, W., Hölter, S. M., Hough, T., Jurdic, P., Keane, T. M., Morgan, H., Müller, W., Neff, F., Nicholson, G., Pasche, B., Roberson, L. A., Rozman, J., Sanderson, M., Santos, L., Selloum, M., Shannon, C., Southwel, A., Tocchini-Valentini, G. P., Vancollie, V. E., Westerberg, H., Wurst, W., Zi, M., Yalcin, B., Ramirez-Solis, R., Steel, K. P., Mallon, A. M., De Angelis, M. H., Herault, Y., and Brown, S. D. (2013). A comparative phenotypic and genomic analysis of C57BL/6J and C57BL/6N mouse strains. *Genome Biol.*, 14(7).

Smyth, G. K. (1989). Generalized linear models with varying dispersion. *J. R. Stat. Soc. Ser. B Methodol.*, 51(1):47–60.

Speliotes, E. K., Willer, C. J., Berndt, S. I., Monda, K. L., Thorleifsson, G., Jackson, A. U., Allen, H. L., Lindgren, C. M., Luan, J., Mägi, R., Randall, J. C., Vedantam, S., Winkler, T. W., Qi, L., Workalemahu, T., Heid, I., Steinthorsdottir, V., Stringham, H., Weedon, M. N., Wheeler, E., Wood, A. R., Ferreira, T., Weyant, R. J., Segrè, A. V., Eestrada, K., Liang, L., Nemesh, J., Park, J. H., Gustafsson, S., Kilpeläinen, T. O., Yang, J., Bouatia-Naji, N., Eesko, T., Feitosa, M. F., Kutalik, Z., Mangino, M., Raychaudhuri, S., Scherag, A., Smith, A. V., Welch, R., Zhao, J. H., Aben, K. K., Absher, D. M., Amin, N., Dixon, A. L., Fisher, E., Glazer, N., Goddard, M. E., Heard-Costa, N., Hoesel, V., Hottenga, J. J., Johansson, Å., Johnson, T., Ketkar, S., Lamina, C., Li, S., Moffatt, M. F., Myers, R. H., Narisu, N., Perry, J. R., Peters, M. J., Preuss, M., Ripatti, S., Rivadeneira, F., Sandholt, C., Scott, L. J., Timpson, N. J., Tyrer, J. P., Van Wingerden, S., Watanabe, R., White, C. C., Wiklund, F., Barlassina, C., Chasman, D. I., Cooper, M. N., Jansson, J. O., Lawrence, R. W., Pellikka, N., Prokopenko, I., Shi, J., Thiering, E., Alavere, H., Alibrandi, M. T., Almgren, P., Arnold, A., Aspelund, T., Atwood, L. D., Balkau, B., Balmforth, A. J., Bennett, A. J., Ben-Shlomo, Y., Bergman, R., Bergmann, S., Biebermann, H., Blakemore, A. I., Boes, T., Bonnycastle, L., Bornstein, S. R., Brown, M. J., Buchanan, T. A., Busonero, F., Campbell, H., Cappuccio, F. P., Cavalcanti-Proença, C., Ida Chen, Y. D., Chen, C. M., Chines, P., Clarke, R., Coin, L., Connell, J., Day, I., Den Heijer, M., Duan, J., Eebrahim, S., Eelliott, P., Eelosua, R., Eeiriksdottir, G., Eerdos, M. R., Eeriksson, J. G., Facheris, M. F., Felix, S. B., Fischer-Posovszky, P., Folsom, A. R., Friedrich, N., Freimer, N. B., Fu, M., Gaget, S., Gejman, P. V., Geus, E. J., Gieger, C., Gjesing, A. P., Goel, A., Goyette, P., Grallert, H., Gräßler, J., Greenawalt, D., Groves, C. J., Gudnason, V., Guiducci, C., Hartikainen, A. L., Hassanali, N., Hall, A., Havulinna, A., Hayward, C., Heath, A., Hengstenberg, C., Hicks, A. A., Hinney, A., Hofman, A., Homuth, G., Hui, J., Igl, W., Iribarren, C., Isomaa, B., Jacobs, K. B., Jarick, I., Jewell, E., John, U., Jørgensen, T., Jousilahti, P., Jula, A., Kaakinen, M., Kajantie, E., Kaplan, L., Kathiresan, S., Kettunen, J., Kinnunen, L., Knowles, J., Kolcic, I., König, I. R., Koskinen, S., Kovacs, P., Kusisto, J., Kraft, P., Kvaløy, K., Laitinen, J., Lantieri, O., Lanzani, C., Launer, L. J., Lecoeur, C., Lehtimäki, T., Lettre, G., Liu, J., Lokki, M. L., Lorentzon, M., Luben, R., Ludwig, B., Magic, Manunta, P., Marek, D., Marre, M., Martin, N. G., McArdle, W., McCarthy, A., McKnight, B., Meitinger, T., Melander, O., Meyre, D., Midthjell, K., Montgomery, G., Morken, M. A., Morris, A. P., Mulic, R., Ngwa, J., Nelis, M., Neville, M. J., Nyholt, D. R., O'Ddonnell, C. J., O'Rahilly, S., Ong, K., Ostra, B., Paré, G., Parker, A., Perola, M., Pichler, I., Pietiläinen, K. H., Platou, C. P., Polasek, O., Pouta, A., Rafelt, S., Raitakari, O., Rayner, N., Ridderstråle, M., Rief, W., Ruokonen, A., Robertson, N. R., Rzehak, P., Salomaa, V., Sanders, A. R., Sandhu, M., Sanna, S., Saramies, J., Savolainen, M. J., Scherag, S., Schipf, S., Schreiber, S., Schunkert, H., Silander, K., Sinisalo, J., Siscovick, D. S., Smit, J. H., Soranzo, N., Sovio, U., Stephens, J., Surakka, I., Swift, A. J., Tammesoo, M. L., Tardif, J. C., Teder-Laving, M., Teslovich, T., Thompson, J. R., Thomson, B., Tönjes, A., Tuomi, T., Van Meurs, J. B., Van OMen, G. J., Vatin, V., Viikari, J., Visvikis-Siest, S., Vitart, V., Vogel, C. I., Voight, B. F., Waite, L., Wallaschofski, H., Walters, B., Widen, E., Wiegand, S., Wild, S. H., Willemsen, G., Witte, D. R., Witteman, J., Xu, J., Zhang, Q., Zgaga, L., Ziegler, A., Zitting, P., Beilby, J. P., FarOqi, I. S., Hebebrand, J., Huikuri, H. V., James, A., Kähönen, M., Levinson, D. F., MacCiardi, F., Nieminen, M. S., Ohlsson, C., Palmer, L. J., Ridker, P., Stumvoll, M., Beckmann, J., Boeing, H., Boerwinkle, E., BOmsma, D. I., Caulfield, M. J., Chanock, S. J., Collins, F., Cupples, L. A., Smith, G. D., Eerdmann, J., Frogue, P., Grönberg, H., Gyllensten, U., Hall, P., Hansen, T., Harris, T. B., Hattersley, A. T., Hayes, R. B., Heinrich, J., Hu, F. B., Hveem, K., Illig, T.,

Jarvelin, M. R., Kaprio, J., Karpe, F., Khaw, K. T., Kiemeney, L. A., Krude, H., Laakso, M., Lawlor, D. A., Metspalu, A., Munroe, P. B., Ouwehand, W. H., Pedersen, O., Penninx, B. W., Peters, A., Pramstaller, P. P., Quertermous, T., Reinehr, T., Rissanen, A., Rudan, I., Samani, N. J., Schwarz, P. E., Shuldiner, A. R., Spector, T. D., Tuomilehto, J., Uda, M., Uitterlinden, A., Valle, T., Wabitsch, M., Waeber, G., Wareham, N. J., Watkins, H., Wilson, J. F., Wright, A. F., Zillikens, M. C., ChatterjE, N., McCarroll, S. A., Purcell, S., Schadt, E., Visscher, P., Assimes, T. L., Borecki, I. B., Deloukas, P., Fox, C. S., Groop, L. C., Haritunians, T., Hunter, D. J., Kaplan, R., Mohlke, K., O'ConneL, J. R., Peltonen, L., SchleSinger, D., P Strachan, D. P., Van Duijn, C., Wichmann, H. E., Frayling, T. M., Thorsteinsdottir, U., Abecasis, G. R., Barroso, I., Boehnke, M., StefanSon, K., North, K. E., McArthy, M. I., Hirschhorn, J. N., IngelSon, E., and Loos, R. J. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.*, 42(11):937–948.

Stephenson, A. G. (2002). evd: Extreme Value Distributions. *R News*, 2(2):0.

Struchalin, M. V., Dehghan, A., Witteman, J. C., van Duijn, C., and Aulchenko, Y. S. (2010). Variance heterogeneity analysis for detection of potentially interacting genetic loci: method and its limitations. *BMC Genet.*, 11:92.

Sun, X., Elston, R., Morris, N., and Zhu, X. (2013). What is the significance of difference in phenotypic variability across SNP genotypes? *Am. J. Hum. Genet.*, 93(2):390–397.

Svenson, K. L., Gatti, D. M., Valdar, W., Welsh, C. E., Cheng, R., Chesler, E. J., Palmer, A. A., McMillan, L., and Churchill, G. A. (2012). High-resolution genetic mapping using the mouse Diversity Outbred population. *Genetics*, 190(2):437–447.

The Complex Trait Consortium (2004). The Collaborative Cross, a community resource for the genetic analysis of complex traits. *Nat. Genet.*, 36:1133–1137.

Tsui, L., Buchwald, M., Barker, D., Braman, J., Knowlton, R., Schumm, J., Eiberg, H., Mohr, J., Kennedy, D., Plavsic, N., and al. Et (1985). Cystic fibrosis locus defined by a genetically linked polymorphic DNA marker. *Science (80-.).*, 230(4729):1054–1057.

Valdar, W., Flint, J., and Mott, R. (2006). Simulating the Collaborative Cross: power of quantitative trait loci detection and mapping resolution in large sets of recombinant inbred strains of mice. *Genetics*, 172(3):1783–1797.

Visscher, P. M. and Hill, W. G. (1992). Heterogeneity of variance and dairy cattle breeding. *Anim. Sci.*, 55(3):321–329.

Visscher, P. M. and Posthuma, D. (2010). Statistical power to detect genetic loci affecting environmental sensitivity. *Behav. Genet.*, 40(5):728–733.

Visscher, P. M., Thompson, R., and Haley, C. S. (1996). Confidence intervals in QTL mapping by bootstrapping.

Waddington, C. H. (1942). Canalization of Development and the Inheritance of Acquired Characters. *Nature*, 150(3811):563–565.

Waddington, C. H. (1959). Canalization of development and genetic assimilation of acquired characters. *Nature*, 183:1654–1655.

Wagner, G. P., Booth, G., and Bagheri-Chaichian, H. (1997). A Population Genetic Theory of Canalization. *Evolution (N. Y).*, 51(2):329–347.

Wagner, G. P., Pavlicev, M., and Cheverud, J. M. (2007). The road to modularity. *Nat. Rev. Genet.*, 8(12):921–931.

Wainwright, B. J., Scambler, P. J., Schmidtke, J., Watson, E. A., Law, H. Y., Farrall, M., Cooke, H. J., Eiberg, H., and Williamson, R. (1985). Localization of cystic fibrosis locus to human chromosome 7cen-q22. *Nature*, 318(6044):384–385.

Walsh, B. (2017). Crops can be strong and sensitive. *Nat Plants*, 3(9):694–695.

White, R., Woodward, S., Leppert, M., and O'Connell, P. (1985). A closely linked genetic marker for cystic brosis. *Nature*, 318(December 2015):382–384.

Wolc, A., Arango, J., Settar, P., Fulton, J. E., and others (2012). Genomewide association analysis and genetic architecture of egg weight and egg uniformity in layer chickens. *Genetics*.

Wolf, J. B., Pomp, D., Eisen, E. J., Cheverud, J. M., and Leamy, L. J. (2006). The contribution of epistatic pleiotropy to the genetic architecture of covariation among polygenic traits in mice. *Evol. Dev.*, 8(5):468–476.

Wood, A. R., Esko, T., Yang, J., Vedantam, S., Pers, T. H., Gustafsson, S., Chu, A. Y., Estrada, K., Luan, J., Kutalik, Z., Amin, N., Buchkovich, M. L., Croteau-Chonka, D. C., Day, F. R., Duan, Y., Fall, T., Fehrmann, R., Ferreira, T., Jackson, A. U., Karjalainen, J., Lo, K. S., Locke, A. E., Mägi, R., Mihailov, E., Porcu, E., Randall, J. C., Scherag, A., Vinkhuyzen, A. A. E., Westra, H.-J., Winkler, T. W., Workalemahu, T., Zhao, J. H., Absher, D., Albrecht, E., Anderson, D., Baron, J., Beekman, M., Demirkan, A., Ehret, G. B., Feenstra, B., Feitosa, M. F., Fischer, K., Fraser, R. M., Goel, A., Gong, J., Justice, A. E., Kanoni, S., Kleber, M. E., Kristiansson, K., Lim, U., Lotay, V., Lui, J. C., Mangino, M., Mateo Leach, I., Medina-Gomez, C., Nalls, M. A., Nyholt, D. R., Palmer, C. D., Pasko, D., Pechlivanis, S., Prokopenko, I., Ried, J. S., Ripke, S., Shungin, D., Stancáková, A., Strawbridge, R. J., Sung, Y. J., Tanaka, T., Teumer, A., Trompet, S., van der Laan, S. W., van Setten, J., Van Vliet-Ostaptchouk, J. V., Wang, Z., Yengo, L., Zhang, W., Afzal, U., Arnlöv, J., Arscott, G. M., Bandinelli, S., Barrett, A., Bellis, C., Bennett, A. J., Berne, C., Blüher, M., Bolton, J. L., Böttcher, Y., Boyd, H. A., Bruinenberg, M., Buckley, B. M., Buyske, S., Caspersen, I. H., Chines, P. S., Clarke, R., Claudi-Boehm, S., Cooper, M., Daw, E. W., De Jong, P. A., Deelen, J., Delgado, G., Denny, J. C., Dhonukshe-Rutten, R., Dimitriou, M., Doney, A. S. F., Dörr, M., Eklund, N., Eury, E., Folkersen, L., Garcia, M. E., Geller, F., Giedraitis, V., Go, A. S., Grallert, H., Grammer, T. B., Gräßler, J., Grönberg, H., de Groot, L. C. P. G. M., Groves, C. J., Haessler, J., Hall, P., Haller, T., Hallmans, G., Hannemann, A., Hartman, C. A., Hassinen, M., Hayward, C., Heard-Costa, N. L., Helmer, Q., Hemani, G., Henders, A. K., Hillege, H. L., Hlatky, M. A., Hoffmann, W., Hoffmann, P., Holmen, O., Houwing-Duistermaat, J. J., Illig, T., Isaacs, A., James, A. L., Jeff, J., Johansen, B., Johansson, Å., Jolley, J., Juliusdottir, T., Junttila, J., Kho, A. N., Kinnunen, L., Klopp, N., Kocher, T., Kratzer, W., Lichtner, P., Lind, L., Lindström, J., Lobbens, S., Lorentzon, M., Lu, Y., Lyssenko, V., Magnusson, P. K. E., Mahajan, A., Maillard, M., McArdle, W. L., McKenzie, C. A., McLachlan, S., McLaren, P. J., Menni, C., Merger, S., Milani, L., Moayyeri, A., Monda, K. L., Morken, M. A., Müller, G., Müller-Nurasyid, M., Musk, A. W., Narisu, N., Nauck, M., Nolte, I. M., Nöthen, M. M., Oozageer, L., Pilz, S., Rayner, N. W., Renstrom, F., Robertson, N. R., Rose, L. M., Roussel, R., Sanna, S., Scharnagl, H., Scholtens, S., Schumacher, F. R., Schunkert, H., Scott, R. A., Sehmi,

J., Seufferlein, T., Shi, J., Silventoinen, K., Smit, J. H., Smith, A. V., Smolonska, J., Stanton, A. V., Stirrups, K., Stott, D. J., Stringham, H. M., Sundström, J., Swertz, M. A., Syvänen, A.-C., Tayo, B. O., Thorleifsson, G., Tyrer, J. P., van Dijk, S., van Schoor, N. M., van der Velde, N., van Heemst, D., van Oort, F. V. A., Vermeulen, S. H., Verweij, N., Vonk, J. M., Waite, L. L., Waldenberger, M., Wennauer, R., Wilkens, L. R., Willenborg, C., Wilsgaard, T., Wojczynski, M. K., Wong, A., Wright, A. F., Zhang, Q., Arveiler, D., Bakker, S. J. L., Beilby, J., Bergman, R. N., Bergmann, S., Biffar, R., Blangero, J., Boomsma, D. I., Bornstein, S. R., Bovet, P., Brambilla, P., Brown, M. J., Campbell, H., Caulfield, M. J., Chakravarti, A., Collins, R., Collins, F. S., Crawford, D. C., Cupples, L. A., Danesh, J., de Faire, U., den Ruijter, H. M., Erbel, R., Erdmann, J., Eriksson, J. G., Farrall, M., Ferrannini, E., Ferrières, J., Ford, I., Forouhi, N. G., Forrester, T., Gansevoort, R. T., Gejman, P. V., Gieger, C., Golay, A., Gottesman, O., Gudnason, V., Gyllensten, U., Haas, D. W., Hall, A. S., Harris, T. B., Hattersley, A. T., Heath, A. C., Hengstenberg, C., Hicks, A. A., Hindorff, L. A., Hingorani, A. D., Hofman, A., Hovingh, G. K., Humphries, S. E., Hunt, S. C., Hypponen, E., Jacobs, K. B., Jarvelin, M.-R., Jousilahti, P., Jula, A. M., Kaprio, J., Kastelein, J. J. P., Kayser, M., Kee, F., Keinanen-Kiukaanniemi, S. M., Kiemeney, L. A., Kooner, J. S., Kooperberg, C., Koskinen, S., Kovacs, P., Kraja, A. T., Kumari, M., Kuusisto, J., Lakka, T. A., Langenberg, C., Le Marchand, L., Lehtimäki, T., Lupoli, S., Madden, P. A. F., Männistö, S., Manunta, P., Marette, A., Matise, T. C., McKnight, B., Meitinger, T., Moll, F. L., Montgomery, G. W., Morris, A. D., Morris, A. P., Murray, J. C., Nelis, M., Ohlsson, C., Oldehinkel, A. J., Ong, K. K., Ouwehand, W. H., Pasterkamp, G., Peters, A., Pramstaller, P. P., Price, J. F., Qi, L., Raitakari, O. T., Rankinen, T., Rao, D. C., Rice, T. K., Ritchie, M., Rudan, I., Salomaa, V., Samani, N. J., Saramies, J., Sarzynski, M. A., Schwarz, P. E. H., Sebert, S., Sever, P., Shuldiner, A. R., Sinisalo, J., Steinthorsdottir, V., Stolk, R. P., Tardif, J.-C., Tönjes, A., Tremblay, A., Tremoli, E., Virtamo, J., Vohl, M.-C., Electronic Medical Records and Genomics (eMEMERGEGE) Consortium, MIGen Consortium, PAGEGE Consortium, LifeLines Cohort Study, Amouyel, P., Asselbergs, F. W., Assimes, T. L., Bochud, M., Boehm, B. O., Boerwinkle, E., Bottinger, E. P., Bouchard, C., Cauchi, S., Chambers, J. C., Chanock, S. J., Cooper, R. S., de Bakker, P. I. W., Dedoussis, G., Ferrucci, L., Franks, P. W., Froguel, P., Groop, L. C., Haiman, C. A., Hamsten, A., Hayes, M. G., Hui, J., Hunter, D. J., Hveem, K., Jukema, J. W., Kaplan, R. C., Kivimaki, M., Kuh, D., Laakso, M., Liu, Y., Martin, N. G., März, W., Melbye, M., Moebus, S., Munroe, P. B., Njølstad, I., Oostra, B. A., Palmer, C. N. A., Pedersen, N. L., Perola, M., Pérusse, L., Peters, U., Powell, J. E., Power, C., Quertermous, T., Rauramaa, R., Reinmaa, E., Ridker, P. M., Rivadeneira, F., Rotter, J. I., Saaristo, T. E., Saleheen, D., Schlessinger, D., Slagboom, P. E., Snieder, H., Spector, T. D., Strauch, K., Stumvoll, M., Tuomilehto, J., Uusitupa, M., van der Harst, P., Völzke, H., Walker, M., Wareham, N. J., Watkins, H., Wichmann, H.-E., Wilson, J. F., Zanen, P., Deloukas, P., Heid, I. M., Lindgren, C. M., Mohlke, K. L., Speliotes, E. K., Thorsteinsdottir, U., Barroso, I., Fox, C. S., North, K. E., Strachan, D. P., Beckmann, J. S., Berndt, S. I., Boehnke, M., Borecki, I. B., McCarthy, M. I., Metspalu, A., Stefansson, K., Uitterlinden, A. G., van Duijn, C. M., Franke, L., Willer, C. J., Price, A. L., Lettre, G., Loos, R. J. F., Weedon, M. N., Ingelsson, E., O'Connell, J. R., Abecasis, G. R., Chasman, D. I., Goddard, M. E., Visscher, P. M., Hirschhorn, J. N., and Frayling, T. M. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.*, 46(11):1173–1186.

Yadav, A., Dhole, K., and Sinha, H. (2015). Phenotypic robustness determines genetic regulation of complex traits. *bioRxiv*.

Yang, J., Loos, R., Goddard, M., and Visscher, P. M. (2012). FTO genotype is associated with phenotypic variability of body mass index. *Nature*, 490(7419):267–272.

Yi, N., Shriner, D., Banerjee, S., Mehta, T., Pomp, D., and Yandell, B. S. (2007). An efficient Bayesian model selection approach for interacting quantitative trait loci models with many effects. *Genetics*, 176(3):1865–1877.

Yi, N., Yandell, B. S., Churchill, G. A., Allison, D. B., Eisen, E. J., and Pomp, D. (2005). Bayesian model selection for genome-wide epistatic quantitative trait loci analysis. *Genetics*, 170(3):1333–1344.

Yi, N., Zinniel, D. K., Kim, K., Eisen, E. J., Bartolucci, A., Allison, D. B., and Pomp, D. (2006). Bayesian analyses of multiple epistatic QTL models for body weight and body composition in mice. *Genet. Res.*, 87(1):45–60.

Zhou, X. and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.*, 44(7):821–4.

Ziv, N., Shuster, B. M., Siegal, M. L., and Gresham, D. (2017). Resolving the complex genetic basis of phenotypic variation and variability of cellular growth. *Genetics*, 206(3):1645–1657.

Zou, F., Xu, Z., and Vision, T. (2006). Assessing the significance of quantitative trait loci in replicable mapping populations. *Genetics*, 174(2):1063–1068.