Bayesian Inference of the Allelic Series
in Multiparental Populations with Applications

Wesley L. Crouse

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
2019

Approved by:

Terry S. Furey

Samir N.P. Kelada

Yun Li

Michael I. Love

Fernando Pardo-Manuel de Villena

William Valdar

ABSTRACT

WESLEY L. CROUSE: Bayesian Inference of the Allelic Series
in Multiparental Populations with Applications
(Under the direction of Samir N.P. Kelada and William Valdar)

Multiparental populations (MPPs) are experimental populations in which the genome of every individual is a random mosaic of known founder haplotypes. These populations provide advantages for detecting quantitative trait loci (QTL) because tests of association between phenotypes and genetic variation can leverage inferred founder haplotype descent. It is difficult, however, to determine how haplotypes at a locus group into distinct functional alleles, termed the allelic series. The allelic series is important because it provides information about the number of casual variants at a QTL and their combined effects.

We begin by analyzing QTL mapping power in a particular MPP, the Collaborative Cross (CC). We find that QTL mapping power depends on the allelic series and whether it is balanced or imbalanced with respect to the founder haplotypes. More generally, this study serves as a much-needed resource for designing CC experiments that are well-powered to detect QTL using haplotype-based approaches.

Next, we introduce a fully-Bayesian framework for inferring the allelic series. This framework accounts for sources of uncertainty found in typical MPPs, including individual haplotype states at the QTL, the size of the allele effects, and most importantly, the number and composition of functional alleles. Our prior distribution for the allelic series is based on the Chinese restaurant process, and we leverage its connection to the coalescent to introduce additional prior information about haplotype relatedness via a phylogenetic tree. This is the primary innovation of our research.

We evaluate our approach via simulation and find that posterior inference of the allelic series is uncertain even when power is high. Despite this uncertainty, allele-based inference still improves effect estimation when the true number of functional alleles is small. Phylogenetic information improves posterior certainty of the allelic series, effect estimation, and statistical signal. We find

only marginal improvements in QTL mapping power using the allele-based approach without tree information, and although the tree-informed approach may perform better, implementing it in practice is challenging.

We also apply our method to real data from the CC and the Drosophila Synthetic Population Resource, highlighting new insights facilitated by our allele-based association approach.

For all my teachers...

ACKNOWLEDGEMENTS

I would like to thank my advisors, Samir Kelada and Will Valdar, for investing their time, resources, and expertise into my training. I am especially appreciative of their trust to pursue this research in the directions that I find most interesting. They are supportive mentors, and working with them has been a pleasure.

I also want to acknowledge all the current and former members of the Kelada and Valdar labs, who I have enjoyed learning from and working with every day. In particular, I have worked closely with Greg Keele, who shares many of my interests and always provides valuable insight. I feel fortunate to be surrounded by such capable colleagues and friends.

The same can be said of my classmates Nur Shahir, Jessime Kirk, Sherif Farag, and all my friends in the Biological and Biomedical Sciences Program. I have shared this experience with many great people.

This research was facilitated by the guidance of my committee, and more broadly, by institutional support from the research community at UNC. This community is made possible by the hard work of many faculty and staff, and by continued public investment in research and education.

Finally, I am thankful for the love and support of my friends and family, especially my parents Daryl and Debbie, and my sister Erin, who were my first teachers. Most of all, I am thankful for my partner Lauren, who is always positive and encouraging, and who is a model of hard work and excellence for both me and our child, Sam.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ARG	Ancestral recombination graph
BF	Bayes factor
CC	Collaborative Cross
CeMEE	Caenorhabditis elegans Multiparental Experimental Evolution
CRP	Chinese restaurant process
DGRP	Drosophila Genetic Reference Panel
DO	Diversity Outbred
DSPR	Drosophila Synthetic Population Resource
DP	Dirichlet process
eQTL	Expression quantitative trait locus
FE	Fixed effects
FPR	False positive rate
GEV	Generalized extreme value distribution
GWER	Genome-wide type I error rate
HK	Haley-Knott
HMM	Hidden Markov model
HPD	Highest posterior density
JAX	Jackson Laboratory
LD	Linkage disequilibrium
lnBF	Natural log of the Bayes factor
MAGIC	Multigenerational advanced intercross
MAP	Maximum a posteriori
MCV	Mean red blood cell volume
MPP	Multiparental population
MSE	Mean squared error
NAM	Nested association mapping
PreCC	Incipient lines of the Collaborative Cross
QTL	Quantitative trait locus

RE   Random effects

RI   Recombinant inbred

ROP   Regression on probabilities

SNP   Single nucleotide polymorphism

SPARCC   Simulated Power Analysis of the Realized Collaborative Cross

TIMBR   Tree-Based Inference of Multiallelism via Bayesian Regression

CHAPTER 1

Introduction

Achieving precision medicine requires an improved understanding of how genetic variation affects disease state, therapeutic responsiveness, and, more broadly, underlying disease biology (Collins and Varmus, 2015). Many genetic risk factors have been identified for disease susceptibility (MacArthur et al., 2017), drug efficacy (Madian et al., 2012), and adverse drug response (Wilke et al., 2007), and early results from pharmacological interventions that account for these genetic risk factors are encouraging (Relling and Evans, 2015). Genetic association is an effective and popular approach for identifying genetic risk factors, and for interrogating the relationship between genetic variation and quantitative traits more generally (Visscher et al., 2017). The most common genetic association approach, the genome-wide association study, involves testing genetic variants throughout the genome for an association with a quantitative trait in order to characterize the effects of those variants on that trait. Typically, the impact of each variant is assessed individually, with adjustment for global population structure as necessary, and without considering the possibility of interactions with other variants in the local genetic environment. For example, a linear model for the effect of a single biallelic variant (with alternatives A and a, and assuming homozygosity for simplicity) is given by

$$\mathbf{y} = \mu \mathbf{1} + \mathbf{x}_a \beta_a + \boldsymbol{\epsilon}$$

where $\mathbf{y}$ is vector quantitative trait observations for N individuals, μ is the intercept, $\mathbf{x}_a$ is a vector indicating the presence of alternative a, β_a is the effect of a, and $\boldsymbol{\epsilon}$ is vector of normally distributed individual error. This approach only considers a single variant, with parameters μ and β_a accounting for the two levels of the data.

A variant may act on a trait in combination with other nearby variants in a region, however, a process known as local epistasis (Wei et al., 2014). For example, a linear model for epistasis between

1

two biallelic variants (with alternatives B and b) is given by

$$\mathbf{y} = \mu\mathbf{1} + \mathbf{x}_a\beta_a + \mathbf{x}_b\beta_b + \mathbf{x}_{ab}\beta_{ab} + \boldsymbol{\epsilon},$$

where $\mathbf{x}_b$ is a vector indicating b, β_b is the effect of b, $\mathbf{x}_{ab}$ is a vector indicating both a and b, and β_{ab} is the effect of this interaction. This model considers both variants simultaneously, with the parameters μ, β_a, β_b, and β_{ab} accounting for the four levels of the data. Epistasis occurs when the combined effect of the variants cannot be reduced to the additive sum of their individual effects, i.e. $\beta_{ab} \neq 0$. Epistasis can mask the effects of the individual variants, making the single-variant approach inappropriate for detecting such a genetic interaction. Given that nearby variants tend to be inherited together as haplotypes, local epistasis is a potentially large source (Eichler et al., 2010; Zuk et al., 2012) of the "missing" genetic contribution to complex traits (Manolio et al., 2009) that is not fully interrogated by single-variant methods.

An alternative to the single-variant association approach that accounts for local epistasis is haplotype-based association. Returning to the previous example, the two biallelic variants comprise $J = 4$ possible haplotypes (*AB*, *Ab*, *aB*, *ab*), and the linear model is reformulated with respect to these haplotypes and their effects:

$$\mathbf{y} = \mathbf{x}_{AB}\beta_{AB} + \mathbf{x}_{Ab}\beta_{Ab} + \mathbf{x}_{aB}\beta_{aB} + \mathbf{x}_{ab}\beta_{ab} + \boldsymbol{\epsilon},$$

where $\mathbf{x}_j$ is a vector indicating haplotype j, and $\boldsymbol{\beta}_j$ is the effect of haplotype j. More generally, the haplotype-based association approach is given by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\mathbf{X}$ is a $N \times J$ matrix of indicators denoting J haplotypes, and $\boldsymbol{\beta}$ is a vector of J haplotype effects. Haplotypes-based association approaches can reveal complex genetic interactions that improve our understanding of the genetic architecture of complex traits (Hamazaki and Iwata, 2019; Yano et al., 2016; Zhang et al., 2012a; Hamblin and Jannink, 2011; Druet and Georges, 2010; Shim et al., 2009; McClurg et al., 2007).

Implementing a haplotype-based association approach can be challenging because, in many situations, the underlying haplotypes are unknown. In this case, haplotypes must be defined empirically using combinations of adjacent variants (Meuwissen et al., 2014), or otherwise inferred as a reduced number of ancestral haplotypes (Davies et al., 2016; Pook et al., 2019). Haplotype-based association is straightforward, though, using model systems such as the multiparental population (MPP) (Churchill et al., 2004; Beyer et al., 2008; King et al., 2012). MPPs are experimental populations generated by breeding a small but genetically diverse set of inbred parents to produce individual offspring whose genomes are mosaics of the original founder haplotypes. These populations are ideal for haplotype-based association because the underlying founder haplotypes are known by construction and are identical-by-descent in the population (Valdar et al., 2006b; Aylor et al., 2011; Collaborative Cross Consortium et al., 2012). Typically, haplotypes are inferred for each individual from genotype information (termed haplotype reconstruction) (Broman et al., 2019; Zheng et al., 2015; Mott et al., 2000), and haplotype-based association proceeds using each unique combination of haplotypes observed in the population. As we just described, haplotypes provide richer information (i.e. more levels of the data, equal to the number of founder haplotypes J) than single variants (two levels when biallelic) (Haley and Knott, 1992; Martínez and Curnow, 1992). This approach assumes that all haplotypes at a given locus are functionally distinct with respect to the phenotype, testing the combined and potentially epistatic effect of all variants within the genomic interval, including variants that are unobserved or undiscovered. This haplotype-based association approach facilitates the detection of complex genetic signals that may not be revealed by single-variant approaches, an advantage that has contributed to the widespread development of MPPs across a variety of biomedically (Churchill et al., 2004; Collaborative Cross Consortium et al., 2012; Macdonald and Long, 2007; King et al., 2014; Kover et al., 2009) and agriculturally (Huang et al., 2015) important model organisms and species.

Although haplotype-based association approaches are useful for detecting complex genetic signals, it can be difficult to translate their results into knowledge about causal variants. Haplotype-based approaches typically assume that all haplotypes are functionally distinct and that the effects of the haplotypes are independent. This assumption, however, is biologically unrealistic, since it is reasonable to expect that there are only a few causal variants at a locus, and that combinations of these variants may be shared across haplotypes. More specifically, we expect that sets of shared

causal variants partition the haplotypes into a potentially smaller number of functionally distinct alleles, with the assignment of haplotypes to functional alleles termed the allelic series. For example, a linear model in the case of K functional alleles is given by

$$\mathbf{y} = \mathbf{XM}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where the allelic series $\mathbf{M}$ is a $J \times K$ indicator matrix which collapses J haplotypes into K functional alleles (i.e. the allelic series), and $\boldsymbol{\beta}$ is a vector of effects for those K alleles (Jannink and Wu, 2003). Knowledge of the allelic series, and in particular, whether it is biallelic ($K = 2$) or multiallelic ($K > 2$), is critical for inference about the number of causal variants at a locus. This allelic perspective also suggests that the haplotype-based association approach is inefficient, in that it involves estimating redundant parameters when some haplotypes may be functionally equivalent ($K \leq J$). Thus, an allele-based association approach would provide valuable insights into the number of causal variants, while potentially improving effect estimation and the power to detect quantitative trait loci (QTL).

Here, we introduce an allele-based association approach that explicitly models the allelic series, treating it as an unknown quantity that must be inferred from the data. In the context of the previous linear model, this means estimating the indicator matrix $\mathbf{M}$ while K is also unknown. This is a challenging problem because the number of possible allelic configurations is large even when the number of haplotypes is small. Currently, there are no established methods for inferring the allelic series, with association methods focused instead on, for example, accommodating uncertainty due to haplotype reconstruction (Kover et al., 2009; Mott et al., 2000; Haley and Knott, 1992), random effects (Wei and Xu, 2016), population structure (Zhou and Stephens, 2012; Lippert et al., 2011; Yuan et al., 2011; Kang et al., 2010; Valdar et al., 2009; Eskin et al., 2008), or several of these, with non-additive effects (Zhang et al., 2014; Durrant and Mott, 2010). In practice, inference of the allelic series is often subjective, combining patterns in the haplotype effect estimates with some intuition about the number of functional alleles, as in Aylor et al. (2011) and Kelada et al. (2012). Yalcin et al. (2005) developed a method to compare biallelic contrasts of "merged" haplotypes with the full haplotype model. In the linear model framework, this assumes that $\mathbf{M}$ is known with $K = 2$. This is essentially a single-variant approach, and it fails to consider multiallelic series. King et al. (2014)

generalized merge analysis to interrogate multiallelic contrasts. This approach implies a uniform prior distribution over the allelic series, $\mathrm{p}(\mathbf{M}) \propto 1$. However, their ad hoc procedure was not embedded within a broader statistical framework that could account for prior information. Our approach most closely resembles a model developed by Jannink and Wu (2003) to infer the allelic series in doubled haploid (homozygous) lines. Their method places either a uniform or Poisson distribution on K, with the conditional allelic series then distributed uniformly, $\mathrm{p}(\mathbf{M}|K) \propto 1$. This study found that the allele-based model improved haplotype effect estimation but that inference of the allelic series itself was generally uncertain. Notably, their approach did not incorporate prior information about the relatedness of the strains, which they identified as a key limitation of their approach. It is reasonable to expect that closely related haplotypes are more likely to be functionally identical than distantly related haplotypes (Morgan and Welsh, 2015) and that including this information would improve allelic series inference. Accounting for haplotype relatedness in an allele-based association framework is the primary innovation of our research.

In the approach presented here, inference of the allelic series is framed as a Bayesian model selection problem. As suggested above, this requires specifying a prior distribution over the space of allelic configurations, $\mathrm{p}(\mathbf{M})$. The space of allelic configurations is often much larger than the number of observations. In this low-information environment, the prior distribution is critical, as it provides the basis for setting expectations about the number of functional alleles and their haplotype composition. Our approach is based on the Chinese restaurant process (CRP), which is the distribution over partitions that underlies the popular Dirichlet process mixture model (Escobar and West, 1995; Müller et al., 2015). In this framework, the haplotypes are partitioned into a potentially smaller set of functional alleles, with the alleles having independent effects. The CRP allows for control over the prior number of alleles via its concentration parameter, but it implicitly assumes equal relatedness between individual haplotypes. Our approach generalizes the CRP to allow for unequal relatedness between the haplotypes by leveraging a particular property of the CRP, namely, that it can be described as the distribution of partitions induced by functional mutations on random coalescent trees, a representation known as Ewens's sampling formula (Ewens, 1972; Kingman, 2006). Examples are given in **Figure 1.1**. Ewens's sampling formula provides an intuitive mechanism for introducing prior information about haplotype relatedness: assuming that the phylogenetic tree of the haplotypes is known rather than random. This defines a prior distribution over the allelic
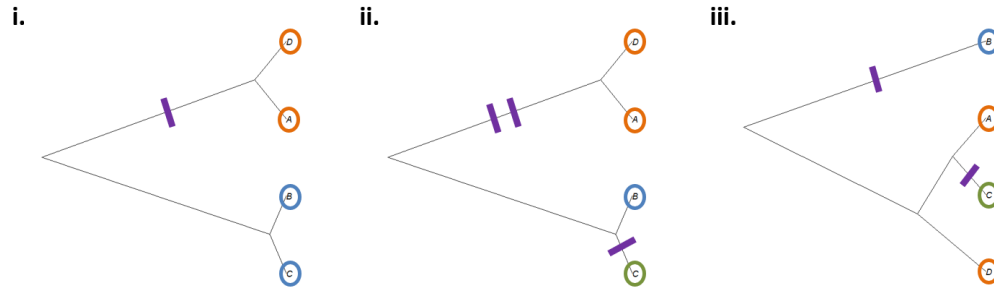
Figure 1.1: Allelic series induced by functional mutations on coalescent trees of haplotypes. **i.** One functional mutation on a tree partitions four haplotypes into two functional alleles: (A,D)|(B,C). **ii.** Additional mutations on the same tree partition the haplotypes into three functional alleles: (A,D)|(B)|(C). The second mutation does not affect (A,D). **iii.** Two functional mutations on a different tree partition the haplotypes into the same allelic series: (A,D)|(B)|(C). Note that the allelic series from the first example, (A,D)|(B,C), is impossible given this tree.

series that is informed by the tree, $p(\mathbf{M}|T)$. In this way, our approach is similar to partition models which include phylogenetic information, for example, by modeling distributional "changepoints" on a tree (Azim Ansari and Didelot, 2016) or by using phylogenetic distance as an input for a distance-dependent CRP (Cybis et al., 2018), among others (Zhang et al., 2012b; Thompson and Kubatko, 2013). In particular, the changepoint model of Azim Ansari and Didelot (2016) specifies a prior distribution over the allelic series by defining the prior probability that each branch of a tree is functionally mutated with respect to a phenotype (in their case, a categorical trait). This is also how we define $p(\mathbf{M}|T)$, and we highlight that it is embedded within the broader population genetics framework of Ewens's sampling formula (Ewens, 1972; Kingman, 2006), with its connections to both the CRP and the coalescent (Berestycki, 2009).

In this document, we introduce an allelic series model and evaluate properties of both the haplotype-based and allele-based association approaches. In **Chapter 2**, we focus on a specific property of the haplotype-based approach, the power to map QTL, in a particular MPP, the Collaborative Cross (CC) (Churchill et al., 2004; Collaborative Cross Consortium et al., 2012). This published study (Keele et al., 2019) stands on its own, serving as a much-needed resource for researchers interested in designing CC experiments that are well-powered to detect QTL using haplotype-based approaches. This chapter identifies a key finding: The power to detect QTL using a haplotype-based approach depends on assumptions about the underlying allelic series. In particular, imbalanced allelic series, where most of the haplotypes are assigned to a single allele, are poorly detected using existing

methods. This motivates the development of an allele-based association approach that directly models the allelic series.

In **Chapter 3**, we introduce a fully-Bayesian framework for inferring the allelic series and additive allele effects in MPPs. As described previously, this approach represents the allelic series as a latent partition of the haplotypes, leveraging Ewens's sampling formula in order to define an informative prior distribution that generalizes the CRP to account for unequal haplotype relatedness. This model places the allelic series on a continuum that encompasses both the single-variant and haplotype-based approaches at the limits. Our approach accounts for multiple sources of uncertainty found in typical MPPs, including uncertainty due to haplotype reconstruction, the number of functional alleles (Escobar and West, 1995; Müller et al., 2015), and the magnitude of their effects (Gelman, 2006). We outline a strategy for posterior inference of this model using a partially-collapsed Gibbs sampler (Neal, 2000; van Dyk and Park, 2008; Park and Van Dyk, 2009) and show how to use posterior samples and Rao-Blackwellization to calculate the marginal likelihood (Blackwell, 2007; Chib, 1995), which is useful for calculating Bayes Factors (BFs) and comparing competing model assumptions (Kass and Raftery, 1995).

In **Chapter 4**, we evaluate various properties of the allelic series model via simulation. We consider the accuracy and posterior certainty of allelic series inference for a given experiment size and effect size, using several alternative prior distributions for the allelic series. We find that accuracy is generally low, but we observe that including phylogenetic information can improve these measures. We also evaluate the error of haplotype effect estimates using the allelic-based approach and the haplotype-based approach. On this measure, we find that the allele-based approach outperforms the haplotype-based approach when the true number of functional alleles is small, even in the absence of additional phylogenetic information. These results are concordant with are most similar study (Jannink and Wu, 2003). We also evaluate the relative statistical signal, as measured by BFs, between the haplotype-based and allele-based approaches in order to quantify the potential improvement from using an allele-based method, both with and without additional phylogenetic information.

In the **Chapter 5**, we present a series of illustrative real-data examples that showcase the inferences that are facilitated by our allele-based approach. The first example, an analysis of a QTL for a red blood cell phenotype in incipient lines of the CC (PreCC) (Kelada et al., 2012), highlights how local phylogenetic information can improve posterior inference of the allelic series. The second

example, an analysis of cis-eQTL in the PreCC(Kelada et al., 2014), demonstrates that our method can detect QTL which are highly multiallelic. The final two examples are analyses of cis-eQTL in the Drosophila Synthetic Population Resource (DSPR) (King et al., 2014), a MPP with fifteen founder haplotypes rather than the typical eight; these indicate that our approach (without tree information) can be applied even in the challenging case of many founder haplotypes.

In **Chapter 6**, we return to the topic of QTL mapping power in the CC, but reconsidered in the context of our allele-based association approach. We use the simulation framework described in **Chapter 2**, this time comparing a fully-Bayesian implementation of the haplotype-based association approach with our allele-based approach, along with two other frequentist approaches. This analysis does not include phylogenetic relatedness at a QTL, as this requires the strong assumption that local phylogeny for the CC founders is known at every locus. Instead, we compare the allele-based but tree-naive CRP approach with the haplotype-based approach, finding a modest improvement in power when the number of functional alleles is small, and especially if the allelic series is imbalanced. Notably, this improvement comes with a considerable computational burden.

In **Chapter 7**, we summarize the previous chapters and comment on future directions for allele-based association approaches in MPPs. We consider phylogeny-informed QTL mapping using our approach, and more generally, we discuss the challenge of inferring local phylogeny with recombination. We also discuss how our allelic series model may be useful for identifying candidate causal variants. Finally, we comment on some of the limitations of our allele-based association approach and how these could be addressed.

Determinants of QTL Mapping Power in the Realized Collaborative Cross

2.1 Overview

This chapter was previously published by Keele et al. (2019) and is reproduced here in its entirety, with only minor changes due to formatting. In this chapter, we focus on the power to map quantitative trait loci (QTL) in the Collaborative Cross (CC) using a haplotype-based association approach. This facilitates the design of future CC experiments that are well-powered to detect QTL. With respect to the allelic series, we find that QTL mapping power is low when the underlying allelic series is imbalanced. This motivates the development of an allele-based association approach that directly models the allelic series, which is the primary focus of future chapters.

The specific contributions of Wesley L. Crouse to this chapter include designing the phenotype simulations (**Section 2.3.2**), implementing the simulation procedures at scale, collating and analyzing the results, and writing sections of the manuscript.

2.2 Introduction

The Collaborative Cross (CC) is a multiparental population (MPP) recombinant inbred (RI) strain panel of laboratory mice derived from eight inbred founder strains (letter abbreviation in parentheses): A/J (A), C57BL/6J (B), 129S1/SvImJ (C), NOD/ShiLtJ (D), NZO/H1LtJ (E), CAST/EiJ (F), PWK/PhJ (G), and WSB/EiJ (H) (Threadgill et al., 2002; Churchill et al., 2004; Chesler et al., 2008; Threadgill and Churchill, 2012). This set of founder strains represents three subspecies of the house mouse *Mus musculus* (Yang et al., 2011) and, in large part due to the inclusion of three wild-derived founders (F-H), imbues the CC panel with far greater genetic variation than previous RI panels derived solely from pairs of classical inbred strains. As an RI panel, the CC thus provides a diverse

set of reproducible genomes and represents a powerful tool for genetic analysis (Collaborative Cross Consortium et al., 2012; Srivastava et al., 2017). Indeed, although the CC RI panel has only become available in the last six years (Welsh et al., 2012), it has already yielded new insights into human disease and basic mouse biology (Shusterman et al., 2013; Rogala et al., 2014; Rasmussen et al., 2014a; Lorè et al., 2015; Levy et al., 2015; Gralinski et al., 2015; Venkatratnam et al., 2017; Orgel et al., 2019; Molenhuis et al., 2018).

As originally envisaged, a key use of the CC was as a resource for QTL mapping (Threadgill et al., 2002; Churchill et al., 2004). In theory, its broad genetic diversity makes it ideal for this purpose, and its replicability permits the mapping of phenotypes such as drug-response that are otherwise hard to measure in outbreds (Mosedale et al., 2017). Its utility for QTL mapping in practice was also predicted by studies in the incipient CC lines (PreCC) (Aylor et al., 2011; Durrant et al., 2011; Philip et al., 2011; Mathes et al., 2011; Kelada et al., 2012; Ferris et al., 2013; Ram et al., 2014; Rutledge et al., 2014; Kelada, 2016; Donoghue et al., 2017; Phillippi et al., 2014)

Nonetheless, QTL mapping power depends in part on the number of strains available, and the number strains available in the CC is, and will remain, far less than the 1,000 proposed in Churchill et al. (2004): At the time of this work, mice were available for 59 CC strains from the UNC Systems Genetics Core, with a subset from these 59 and an additional 11 expected to be offered through the Jackson Laboratory (JAX), a total of 70 CC strains potentially.

A reduction in strain numbers as a function of allelic incompatibilities between subspecies (Shorter et al., 2017) was expected, and winnowed the number of resulting CC strains down to 50-70. Although smaller than originally intended, this population size reflects the biological and financial realities of maintaining a sustainable mammalian genome reference population. [Whereas cost grows proportional to the the number of strains, demand does not, and a much larger number of strains would threaten the economic viability of the operation (F. Pardo-Manuel de Villena, *pers. comm.*).] Nonetheless, subsets of the available CC strains have already been used to map QTL , as evidenced by a growing list of studies (Vered et al., 2014; Mosedale et al., 2017; Graham et al., 2017). Beyond these successes, however, it is unclear how much the reduction has affected the ability to map QTL in the CC in general.

The initially proposed figure of 1,000 CC strains in Churchill et al. (2004) was more formally justified in Valdar et al. (2006a) as being necessary to provide enough power both to map single QTL

and for robust, genome-wide detection of epistasis. That estimate was based on simulations involving larger numbers (500-1,000) of hypothetical CC genomes. Those simulations, performed before any CC strains existed and with the goal of guiding the CC's design, had a broad scope, exploring the effect of varying strain numbers, alternative mapping approaches [association of single nucleotide polymorphisms (SNPs) vs association of inferred haplotypes], and alternative breeding strategies. As such, the power estimates that were reported do not reflect the number of CC strains now available, nor their actual, realized founder mosaic genomes. An updated, more focused power analysis that both exploits and works within the constraints of the realized genomes is therefore timely.

Power analyses have been performed previously for a number of RI panels. For biparental RIs, they have been performed analytically in plants (*e.g.*, Kaeppler 1997), animals [*e.g.*, the BXD lines in mice (Belknap et al., 1996; Peirce et al., 2004)], and in general (Cowen, 1988; Soller and Beckmann, 1990; Knapp and Bridges, 1990), as well as through simulation (Falke and Frisch, 2011; Takuno et al., 2012). For MPP RIs, they have most often been reported as those resources are introduced to the community. This includes, in plants: *Arabidopsis* (Kover et al., 2009; Klasen et al., 2012), nested association mapping (NAM) populations (Li et al., 2011) in maize (Yu et al., 2008) and sorghum (Bouchet et al., 2017), and multigenerational advanced intercross (MAGIC) populations of rice (Yamamoto et al., 2014) and maize (Dell'Acqua et al., 2015). In animals, other than aforementioned prospective study of Valdar et al. (2006a): (Noble et al., 2017) assessed mapping power of SNP association while introducing a 507-strain nematode resource, the *Caenorhabditis elegans* Multiparental Experimental Evolution (CeMEE) panel; and King et al. (2012) estimated haplotype-based association power while introducing the *Drosophila* Synthetic Population Resource (DSPR), a fly panel with more than 1,600 lines. In a follow-up DSPR power analysis, King and Long (2017) compared the DSPR with the related *Drosophila* Genetic Reference Panel (DGRP) (Mackay et al., 2012). They illustrated how QTL effect size differs between a population whose allele frequencies are balanced (DSPR) vs one whose allele frequencies are less balanced (DGRP) and explored implications for cross-population validation; they also compared mapping power for bi-allelic QTL, based on single SNPs, and multi-allelic QTL constructed from actual adjacent SNPs within genes.

Here we examine related topics on QTL mapping power in the realized CC, including: 1) how power is affected by the number of strains and replicates; 2) how it is affected by the number of

functional alleles and their distributions among the founders; and 3) how the QTL effect size is specific to a particular population or sample and how that influences a power estimate's interpretation.

To allow researchers to repeat our analyses, but tailored to their own specific requirements or with updated CC genome lists, we provide an R package SPARCC (Simulated Power Analysis of the Realized Collaborative Cross), a tool that evaluates the power to map QTL by performing efficient haplotype regression-based association analysis of simulated QTL using the currently available CC genomes. SPARCC is highly flexible, allowing QTL to be simulated with any possible allele-to-founder pattern and scaled with respect to different reference populations. As a re-usable resource, researchers could estimate power calculations based on the CC strains available to them and potentially incorporate prior knowledge about the genetic architecture of the likely QTL or the phenotype as whole.

2.3 Methods

Our power calculations are based on three main processes:

1. Simulation of CC data, including selection of CC strains from a fixed set of realized CC genomes, and QTL location, and simulation of phenotypes.

2. QTL mapping, including determination of significance thresholds.

3. Evaluation of QTL detection accuracy, power and false positive rate (FPR).

These are described in detail below, after a description of the genomic data that serves as the basis for the simulations.

2.3.1 Data on realized CC genomes

2.3.1.1 CC strains.

Genome data was obtained for a set of 72 CC strains (listed in **Appendix C**) available at the time of writing from `http://csbio.unc.edu/CCstatus/index.py?run=FounderProbs`. Genome data was in the form of founder haplotype mosaics (see below) for each strain, this based on genotype data from the MegaMUGA genotyping platform (Morgan et al., 2016) applied to

composites of multiple mice per strain. Since genotyping, some of the 72 strains have become extinct, and more may do so in the future (Darla Miller *pers. comm.*), although it is also possible that more may be added. At the time of writing, however, these were all genomes that had been observed by workers at UNC.

Of the 72 CC strains used in the simulations, it is planned that 54 will be maintained and distributed by The UNC Systems Genetics Core, along with another 5 whose genome data were not available in time for this study (see **Discussion**) to give a UNC total of 59 strains (listed in **Appendix C**). A subset of the UNC 59 will also eventually be maintained by The Jackson Laboratory, which will also potentially maintain 11 of the 72 not among the UNC 59.

The 72 strains used in the simulations included two that were more closely related than others: CC051 and CC059. These strains, which are among the UNC 59, were derived from the same breeding funnel; the number of independent strains available from UNC is thus arguably 58. This relatedness, though not explicitly modeled in the simulations, is nonetheless marked in the figures, which include an indicator denoting 58 as a currently realistic maximum for strain number in CC studies.

2.3.1.2 Reduced dataset of haplotype mosaics.

The genomes of the CC, as with other MPPs, can be represented by inferred mosaics of the original founder haplotypes (Mott et al., 2000). Founder haplotype mosaics were inferred previously by the UNC Systems Genetics Core (`http://csbio.unc.edu/CCstatus/index.py?run=FounderProbs`) using the hidden Markov model (HMM) of Fu et al. (2012) applied to genotype calls from MegaMUGA, a genotyping platform providing data on 77,800 SNP markers (Morgan et al., 2016). The HMM inference provides a vector of 36 diplotype probabilities for each CC strain for each of 77,551 loci (each defined as the interval between adjacent, usable SNPs) across the genome. Rather than using all of the available data for our simulations, we used a reduced version: since adjacent loci often have almost identical descent, mapping using all loci is both computationally expensive and—at least for the purposes of the power analysis—largely redundant. Thus, prior to analysis the original dataset was reduced by averaging adjacent genomic intervals whose diplotype probabilities were highly similar. Specifically, adjacent genomic intervals were averaged if the maximum L2 norm between the probability vectors of all individuals is less than 10%

of the maximum possible L2 norm ($\sqrt{2}$); this reduced the file storage from 610 MB to 288 MB, and the genome from 77,551 to 17,900 intervals (76.9% reduction in positions to be evaluated in a scan).

2.3.2 Phenotype simulation

Phenotypes for CC strains were simulated based on effects from a single QTL, plus effects of polygenic background ("strain effects"), and noise. Within our simulation framework, we specified: 1) the QTL location, which randomly was sampled from the genome; 2) the sample size in terms of both strains and replicates; 3) how the eight possible haplotypes at that location are grouped into eight or fewer functional alleles (the "allelic series"; see below); and 4) how those alleles, along with strain information, are used to generate phenotype values (see below).

2.3.2.1 Underlying phenotype model.

Simulated phenotypes were generated according to the following linear mixed model. For given QTL with $m \leq 8$ functional alleles, phenotype values $\mathbf{y} = \{y_i\}_{i=1}^{N}$ for N individuals in $n \leq N$ strains were generated so that

$$\mathbf{y} = \mathbf{1}\mu + \underbrace{\mathbf{ZX}\boldsymbol{\beta}}_{\text{QTL effect}} + \underbrace{\mathbf{Zu}}_{\text{Strain effect}} + \underbrace{\boldsymbol{\varepsilon}}_{\text{Noise}}, \tag{2.1}$$

where $\mathbf{1}$ is an N-vector of 1's, μ is an intercept, $\mathbf{Z}$ is an $N \times n$ incidence matrix mapping individuals to strains, $\mathbf{X}$ is an $n \times m$ allele dosage matrix mapping strains to their estimated dosage of each of the m alleles, $\boldsymbol{\beta}$ is an m-vector of allele effects, $\mathbf{u}$ is an n-vector of strain effects (representing polygenic background variation), and $\boldsymbol{\varepsilon}$ is an N-vector of unstructured, residual error. The parameter vectors $\boldsymbol{\beta}$, $\mathbf{u}$ and $\boldsymbol{\varepsilon}$ were each generated as being equivalent to independent normal variates rescaled to have specific variances: the strain effects $\mathbf{u}$ and residual $\boldsymbol{\varepsilon}$ were rescaled to have population (rather than sample) variances h_{strain}^2 and σ^2 respectively; the allele effects $\boldsymbol{\beta}$ were rescaled so that the QTL contributes a variance h_{QTL}^2, with this latter rescaling performed in one of three distinct ways (described later).

The relative contributions of the QTL, polygenic background, and noise were thus controlled through three parameters: the QTL effect size, h_{QTL}^2, the strain effect size, h_{strain}^2, and the residual variance σ^2. By convention, these were specified as fractions summing to exactly 1.

14

The allele dosage matrix $\mathbf{X}$ was generated by collapsing functionally equivalent haplotypes according to a specified allelic series. Let $\mathbf{D}$ be an $n \times 36$ incidence matrix describing the haplotype pair (diplotype) state of of each CC strain at the designated QTL, with columns corresponding to AA,..., HH, AB, ..., GH, such that, for example, $\{D\}_{3,1} = 1$ implies CC strain 3 has diplotype AA. Then

$$\mathbf{X} = \mathbf{DAM}, \tag{2.2}$$

where $\mathbf{A}$ is an 36×8 additive model matrix that maps diplotype state to haplotype dosage (*e.g.*, diplotype AA equals 2 doses of A), and $\mathbf{M}$ is an $8 \times m$ "merge matrix" [after Yalcin et al. (2005)] that encodes the allelic series, mapping the 8 haplotypes to m alleles, such that if haplotypes A and B were both in the functional group "allele 1", then diplotype AB in $\mathbf{D}$ would correspond to 2 doses of allele 1 in $\mathbf{X}$ (see examples in **Appendix D**).

2.3.2.2 QTL allelic series.

The specification of an allelic series, rather than assuming all haplotype effects are distinct, acknowledges that for many QTL we would expect the same functional allele to be carried by multiple founder haplotypes. For our main set of simulations, the allelic series was randomly sampled from all possible configurations (examples in **Figure 2.1**); in a smaller, more focused investigation of the effects of allele frequency imbalance, we sampled from all possible configurations of bi-alleles.

2.3.2.3 Alternative definitions of QTL effect size: B and DAMB.

The QTL effect size (h_{QTL}^2) is a critical determinant of mapping power; yet its precise definition and its corresponding interpretation often varies between studies and according to what question is being asked. We used two alternative definitions, "B" and "DAMB", described below. These alternatives acknowledge that the proportion of variance explained by a particular QTL, and thus the power to detect that QTL, is not determined solely by h_{QTL}^2, but rather depends on several additional factors, namely: the variance of the finite sample of allele effects $\boldsymbol{\beta}$; the allelic series configuration $\mathbf{M}$; and the particular set of CC strains and their locus diplotypes $\mathbf{D}$.

Definition B scales the allele effects so that $h_{\mathrm{QTL}}^2 = \mathrm{V}(2\boldsymbol{\beta})$, where $\mathrm{V}()$ denotes the population variance (rather than the sample variance). The QTL effect size is interpretable as the variance
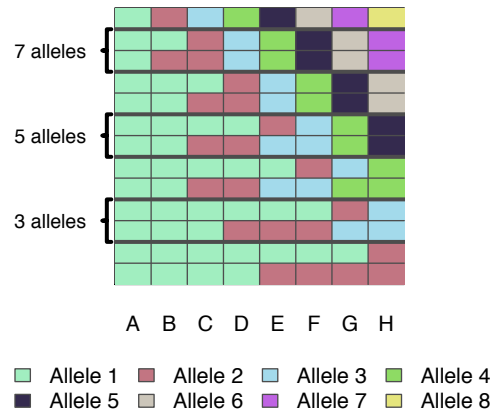
Figure 2.1: Example allelic series with differing numbers of functional alleles. Each row is an allelic series, each column of the grid is a CC founder, and colors correspond to functional allele. Two examples of allelic series are provided for each number of functional alleles: a balanced series and an imbalanced series. The entire space of allelic series are not shown here; however, the full space of series with two alleles is shown in **Figure 2.9A**.

that would be explained by the QTL in a theoretical population that is balanced with respect to the functional alleles. As such, the proportion of variance explained by the QTL in the mapping population will deviate from h^2_{QTL} due to imbalance in both **M** and **D**. Conversely, for a given h^2_{QTL}, the allelic values at a QTL will be constant across populations. (Note: the 2 multiplier ensures proper scaling since **X** from Eq 2.2 includes dosages of founder haplotypes at the QTL, ranging from 0 to 2.)

Definition DAMB scales the QTL effect so that $h^2_{\text{QTL}} = \text{V}(\mathbf{DAM\beta})$. The QTL effect size is exactly the variance explained by the QTL in the mapping population, essentially the R^2. As such, it depends on both **M** and **D**. Correspondingly, for a given h^2_{QTL}, the allelic values will adjust depending on which population they are in. [In the **Supplement**, for completeness, we also describe a further, intermediate option, Definition MB, where $h^2_{\text{QTL}} = \text{V}(2\mathbf{M\beta})$, corresponding to balanced founder contributions.]

The earlier power study of Valdar et al. (2006a), which considered only bi-allelic QTL, defined effect size in a manner comparable to Definition B.

2.3.2.4 Averaging over strains and causal loci.

The previous subsections described simulation of a single phenotype conditional on a set of strains and a causal genomic locus. For each of S simulations, $s = 1, \ldots, S$, we averaged over these variables by uniformly sampling 1) the set of strains included in the experiment (for a specified number of strains), 2) the causal locus underlying the QTL, and 3) the allelic series (for a specified number of functional alleles). This was intended to produce power estimates that take into account many sources of uncertainty and are thus broadly applicable.

2.3.3 QTL detection and power estimation

2.3.3.1 QTL mapping model.

QTL mapping of the simulated data was performed using a variant of Haley-Knott (HK) regression (Haley and Knott, 1992; Martínez and Curnow, 1992) that is commonly used in MPP studies (Mott et al., 2000; Liu et al., 2010; Fu et al., 2012; Gatti et al., 2014; Zheng et al., 2015) whereby association is tested between the phenotype and the local haplotype state, the latter having been inferred probabilistically from genotype (or sequence data) and represented as a set of diplotype probabilities or, in the case of an additive model, a set of haplotype dosages then used as predictors in a linear regression. Specifically, we used HK regression on the strain means (Valdar et al., 2006a; Zou et al., 2006) via the linear model

$$\bar{\mathbf{y}}^{(s)} = \mathbf{1}\mu + \mathbf{PA}\boldsymbol{\beta} + \boldsymbol{\epsilon} \,, \tag{2.3}$$

where $\bar{\mathbf{y}}^{(s)}$ is the s$^{\text{th}}$ simulated n-vector of strain means, $\mathbf{P}$ is an $n \times 36$ matrix of inferred diplotype probabilities for the sampled CC genomes at the QTL [*i.e.*, $\mathbf{P} = p(\mathbf{D}|\text{genotype data})$; see Zhang et al. (2014)], and $\boldsymbol{\epsilon}$ is the n-vector of residual error on the means, distributed as $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}(h_{\text{strain}}^2 + \sigma^2/r))$. The above implies an eight-allele model (cf Eq 2.1 with $\mathbf{M} = \mathbf{I}$). Although this could lead to reduced power when there are fewer functional alleles, particularly at loci in which the functional alleles are not well represented, it is most common in practice, in accordance with the fact that the allelic series of an unmapped QTL would typically be unknown in advance [*e.g.*, (Mott et al., 2000; Valdar et al., 2006a,b; Svenson et al., 2012; Gatti et al., 2014)]. Additional factors that might contribute to

variation in an experiment, such as covariates or batch effects, are neither simulated nor modeled; it is assumed that such factors would be adequately accounted for by, for instance, addition of suitable covariates, pre-processing (*e.g.*, residualizing) of phenotype values or similar, and ultimately lead to a more-or-less equivalent analysis to that described here. The fit of Eq 2.3 was compared with that of an intercept-only null model via an F-test, and produced a p-value, reported as its negative base 10 logarithm, the logP. This procedure was performed for all loci across the genome, resulting in a genome scan for $\mathbf{y}^{(s)}$.

2.3.3.2 Genome-wide significance thresholds and QTL detection.

Genome-wide significance thresholds were determined empirically by permutation. The CC panel is a balanced population with respect to founder genomic contributions and, by design, has minimal population structure. These features support the assumption of exchangeability among strain genomes: that under a null model in which the genetic contribution to the phenotype is entirely driven by infinitesimal (polygenic) effects, all permutations of the strain labels (or equivalently, of the strain means vector $\mathbf{y}^{(s)}$) are equally likely to produce a given configuration of $\mathbf{y}^{(s)}$. Permutation of the strain means, $\mathbf{y}^{(s)}$, was therefore used to find the logP critical value controlling genome-wide type I error rate (GWER) (Doerge and Churchill, 1996). Briefly, we sampled 100 permutations and perform genome scans for each; this was done efficiently using a standard matrix decomposition approach (**Appendix A**). The maximum logPs per genome scan and simulation s were then recorded, and these are fitted to a generalized extreme value distribution (GEV) (Dudbridge and Koeleman, 2004; Valdar et al., 2006a) using R package evir (Pfaff and McNeil, 2018). The upper $\alpha = 0.05$ quantile of this fitted GEV was then taken as the α-level significance threshold, $T_\alpha^{(s)}$. If the maximum observed logP for $\mathbf{y}^{(s)}$ in the region of the simulated QTL exceeded $T_\alpha^{(s)}$, then the corresponding locus was considered to be a (positively) detected QTL (see immediately below).

2.3.3.3 Performance evaluation.

For a given simulation, we declared a true positive if the detected QTL was within ± 5Mb of the true (simulated) QTL. The 5Mb window size was used to approximate a QTL support interval, which is partly a function of linkage disequilibrium (LD) in the CC. (LD has been characterized in the CC previously but not summarized with a single point estimate (Collaborative Cross Consortium et al.,

2012); our choice of 5Mb is therefore an approximation, but we find that it only marginally increased mapping power relative to using smaller window widths.) A false positive was declared if one or more QTL were detected on chromosomes other than the chromosome harboring the simulated QTL. Simulations in which a QTL was detected on the correct chromosome but outside the 5Mb window were disregarded; although this was potentially wasteful of data and biased FPR slightly downward due to loss of false positives on the chromosome with the simulated QTL, it avoided the need for arbitrary rules to handle edge cases in which it was ambiguous whether the simulated signal had been detected or not. Power for a given simulation setting was then defined as the proportion of true positives among all simulations at that setting, and the FPR was defined as the proportion of false positives.

As a measurement of mapping resolution, for true positive detection, we recorded the mean and the 95% quantile of the genomic distance from the true QTL. Given our criterion for calling true positives, the maximum distance was necessarily 5Mb, and experimental settings that correspond to low power would be expected to have fewer data points, yielding estimates that are unstable. In order to obtain more stable estimates, we used a regularization procedure, estimating the mean distance and 95% quantiles as weighted averages of the observed values and prior pseudo-observations. Specifically, for an arbitrarily small but detected true positive QTL, it is reasonable to expect the peak signal to be distributed uniformly within the $\pm$ 5Mb window. This implies a mean location error of 2.5Mb and a 95% quantile of 4.75Mb. Thus, when calculating the regularized mean location error we assumed 10 prior pseudo-observations of 2.5Mb, and when calculating the regularized 95% quantile we assume 10 prior pseudo-observations of 4.75Mb. This number of pseudo-observations represents 1% of the maximum number of possible data points.

2.3.4 Overview of the simulations

2.3.4.1 Simulation settings.

Simulations for all combinations of the following parameter settings:

- Number of strains: [(10-70 by 5), 72]

- QTL effect size (%): [1, (5-95 by 5)]

- Number of functional alleles: [2, 3, 8]

The number of observations per strain were fixed at $r = 1$ and the background strain effect size was fixed at $h_{\text{strain}}^2 = 0\%$ with the understanding that results from these simulations provide information on other numbers of replicates and strain effect sizes implicitly. Specifically, a simulated mapping experiment on strain means that assumes r replicates, strain effect h_{strain}^2, and QTL effect size h_{QTL}^2 is equivalent to a single-observation mapping experiment with no strain effect and QTL effect size $\bar{h}_{\text{QTL}}^2$, where

$$\bar{h}_{\text{QTL}}^2 = \frac{h_{\text{QTL}}^2}{h_{\text{QTL}}^2 + h_{\text{strain}}^2 + \sigma^2/r} \tag{2.4}$$

[Valdar et al. (2006a), after Soller and Beckmann (1990); Knapp and Bridges (1990); Belknap (1998)].

For example, a mapping experiment on strain means with QTL effect size $h_{\text{QTL}}^2 = 0.3$, $h_{\text{strain}}^2 = 0.4$, $\sigma^2 = 0.3$, and $r = 10$, is equivalent to our simulation of a single-observation with no strain effect but QTL effect size $\bar{h}_{\text{QTL}}^2 \simeq 0.41$ (**Supplement**).

We conducted $s = 1,000$ simulation trials per setting. CC strains and the position of the QTL were sampled for each simulation, providing estimates of power that are effectively averaged over the CC population. We ran these settings for QTL effect sizes specified with respect to the observed mapping population (Definition DAMB) and a theoretical population that is balanced in terms of the functional alleles (Definition B). Confidence intervals for power were calculated based on Jeffreys interval (Brown et al., 2001) for a binomial proportion. A description of the computing environment and run-times are provided in **Appendix B**.

2.3.5 Examining FPR when accounting for non-exchangeability of CC strain genomes

In the simulations and mapping procedures described above, strain effects are modeled under the assumption that all CC strains are (at least approximately) equally related. That is, the effects $\mathbf{u} = u_1, \ldots, u_{72}$ in Eq 2.1 are simulated as $\mathbf{u} \sim \text{N}(\mathbf{0}, \mathbf{I}h_{\text{strain}}^2)$ such that any permutation of the values is equally likely (the effects are exchangeable), and this same assumption is made in both the mapping model of Eq 2.3 and the permutation-based estimation of significance thresholds.

An assumption of equal relatedness among CC strains is commonplace: it is suggested by the exchangeable random funnel design used in the CC, is supported by the results of (Valdar et al., 2006a), and has been made in every CC or pre-CC mapping analysis to our knowledge. Making this assumption simplifies QTL mapping analysis by obviating the need for an explicit modeling of genomic similarity [as in, *e.g.*, Eskin et al. (2008)], since, when those similarities are approximately equal and the analysis is performed on strain means, the strain effects are absorbed into the residual error.

Nonetheless, CC strains are equally related only in expectation. Much like the "equal" relatedness of siblings, realized relatedness will depart from expectation due to chance at the point of mixing, and, in the case of the CC, due to selection [*e.g.*, arising from male sterility (Shorter et al., 2017)] and genetic drift during inbreeding [as reflected in unequal founder contributions by Srivastava et al. (2017)]. This combination of stochastic forces can produce unequal relatedness, correlated effects among strains, and population structure, at least at some level.

To quantify population structure in the realized CC, we compared the eigenvalues of the realized genetic relationship matrix $\mathbf{K}$, calculated from the founder mosaic probabilities [after (Gatti et al., 2014)], with those from an idealized $\mathbf{K}$ that reflects equal relatedness of the CC strains, whose off-diagonal elements were set to the mean value observed for the off-diagonal elements in the realized $\mathbf{K}$. We observed that slightly fewer principal components are required to explain 95% of the variation in the realized $\mathbf{K}$ than are required for the balanced $\mathbf{K}$ (64 vs 68 components, respectively; **Figure 2.14A**). This reduction was attenuated with the omission of CC059, one of the two cousin strains, but not completely (64 vs 67 components; **Figure 2.14B**). This suggested that the realized CC strains have mild population structure.

To evaluate to what degree the population structure in the realized CC genomes could inflate FPR when mapping using an analytic model and threshold procedure that ignores it (*i.e.*, that assumes exchangeability), we performed an additional set of null simulations in which strain effects were generated according to additive infinitesimal model (Lynch and Walsh, 1998) based on the actual genomic similarities. Specifically, we set $h_{\mathrm{QTL}}^2 = 0$ and $\mathbf{u} \sim \mathrm{N}(\mathbf{0}, \mathbf{K}h_{\mathrm{strain}}^2)$ but left our mapping protocol unchanged. We conducted 10,000 such null simulations with r=1 for each setting of strain effect size (%): [0-100 by 20]. These simulations were performed using either all 72 founder strains or 71 strains with the omission of CC059, one of the two highly-related cousin strains. A

false positive was declared if any QTL were detected based on the permutation-based significance threshold.

2.3.6 Measuring the Beavis effect

The "Beavis effect" (Beavis, 1994) refers to an upward bias in estimated effect sizes for detected QTL. This phenomenon, also known as the "winner's curse" (Zollner and Pritchard, 2007), arises because the data used for effect estimation has already been substantially selected during QTL discovery; the resulting (post-selection) estimates are thus inflated due to ascertainment bias. The Beavis effect was evaluated theoretically in (Xu, 2003) and found to be most pronounced in studies of smaller sample size ($n < 100$), suggesting that it could be a significant feature of CC mapping studies.

To assess the extent of the Beavis effect in CC mapping experiments, we performed simulations ($s = 1,000$) mapping a bi-allelic QTL, with one replicate ($r = 1$) and zero background strain effect ($h^2_{\text{strain}} = 0$) for all combinations of simulated QTL effect size under Definition DAMB $h^2_{\text{QTL}} \in \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7\}$ and numbers of strains $n \in \{40, 50, 60, 72\}$. If an association was detected within the 10Mb window (using permutation-based thresholds as above), then we recorded the QTL effect size as the R^2 of the model fit at the peak locus (which may or may not be the locus at which the QTL was simulated).

2.3.7 Availability of data and software

2.3.7.1 R package.

All analyses were conducted in the statistical programming language R (R Core Team, 2018). SPARCC is available as an R package on GitHub at `https://github.com/gkeele/sparcc`. Specific arguments that control the phenotype simulations, the strains used, genomic position of simulated QTL, and allelic series, are listed in the **Supplement**. A static version of SPARCC is also provided there (File S2).

Also included within the SPARCC R package are several results datasets. These include data tables of power summaries from our simulations, as well as table summaries from simulations of a bi-allelic QTL that is balanced in the founders, maximally unbalanced in the founders, and the distance between detected and simulated QTL. Further details are provided in File S1 of the **Supplement**, an

account of all the supplemental files. These files are available at figshare, including data, and scripts to run the analysis and produce the figures. File S3 contains the founder haplotype mosaics required for the SPARCC package. Files S4, S5, and S6 can be used to perform the large-scale power analysis. File S7 describes options in the SPARCC package, and also provides two simple tutorials. File S8 produces the figures in this paper and **Supplement**. File S9 is the supplemental tables and figures.

2.3.7.2 CC strains.

The 72 CC strains with available data that were included in the simulations are described in **Appendix C**. Founder diplotype probabilities for each CC strain are available on the CC resource website (`http://csbio.unc.edu/CCstatus/index.py?run=FounderProbs`). We used probabilities corresponding to build 37 (mm9) of the mouse genome, though build 38 (mm10) is also available at the same website.

We store the founder haplotype data in a directory structure that SPARCC is designed to use, and was initially established by the HAPPY software package (Mott et al., 2000). The reduced data are available on GitHub at `https://github.com/gkeele/sparcc_cache`.

2.4   Results

Power simulations were performed for varying numbers of strains, replicates and functional alleles, and for a ladder of QTL effect sizes. QTL effect size was defined in two ways: as the variance explained in a hypothetical populations that is balanced with respect to the alleles (Definition B; see **Methods**), or as the variance explained in the realized population (Definition DAMB). In this section we focus on results using the first of these, Definition B, owing to its more consistent theoretical interpretation. Under that definition, plots of power against numbers of strains are shown in **Figure 2.3**, and power across a representative selection of conditions is shown in **Table 2.1**. For comparison, these numbers are also provided for simulations under Definition DAMB in **Table 2.2**. Throughout these simulations the false positive rate was controlled at the target 0.05 level (**Figure 2.11**).
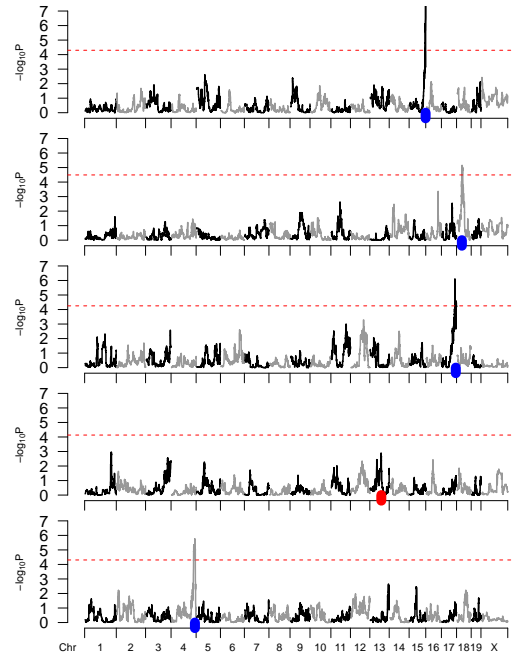
Figure 2.2: Simulated CC data and resulting genome scans. Five simulated genome scans are generated by the code provided in a simple example using our package SPARCC. Red dashed lines represent 95% significance thresholds based on 100 permutation scans. A blue tick represents the simulated position for a QTL that was successfully detected, whereas a red tick marks a QTL that was missed. These simulations were based on a specified set of 65 CC strains, five replicates of each strain, two functional alleles, 10% QTL effect size, and no background strain effect. The QTL is not mapped in the fourth simulation, ranked top to bottom, resulting in a power of 80%. Actual power calculations are based on a greater number of simulations.

2.4.1 Large effect QTL usually detected by 50 or more strains

As a baseline for describing mapping power in the CC, an experiment using one replicate ($r = 1$) of all 72 strains is well-powered to detect QTL explaining $>40\%$ of phenotypic variance but moderately or low powered for QTL explaining 30% or less (**Table 2.1**). Specifically, assuming eight functional alleles, there is 96.4% power to detect a 50% QTL, 79.2% for a 40% QTL, 44.1% for a 30% QTL, and 12.4% for a 20% QTL.

More broadly, simulations across different allele effect types and numbers of strains showed that studies without replicates and with large numbers of strains (>50) were found to be well-powered to detect large effect QTL ($>40\%$) (**Figure 2.3 [top]**).

Identifying smaller effect QTL is feasible, however, using replicates. Replicates improve power by reducing the individual noise variance; as such the extent of the power improvement diminishes as

more variance is attributable to background strain effects than noise. Assuming no background strain effect, and using 50 strains, the power to detect a 20% effect-size QTL with a single replicate is near zero; with 5 replicates it approaches 80%; detecting QTL with effect sizes $\leq 10\%$ is challenging. For example, achieving 80% power to detect an effect size of 10% when all 72 CC strains were used required more than 5 replicates per strain (**Figure 2.3 [middle right]**). Assuming a background strain effect, as would be expected with a complex trait, can reduce the QTL mapping power of small effect QTL substantially (**Figure 2.3 [bottom]**).

2.4.2 Additional strains improve power more than additional replicates

We investigated the relationship between power and the total number of mice, evaluating whether power gains were greater with additional CC strains or additional replicate observations. Power was interpolated over a grid of values for number of replicates and total number of mice from simulations based on a single observation per strain (**Figure 2.5**). This showed that additional CC strains improved mapping power more than additional replicates; this is indicated by higher power values for lower numbers of replicates while holding number of mice constant (see **Figure 2.5**, bordered vertical section at 250 mice).

2.4.3 Location error of detected QTL

To obtain an approximation of mapping resolution, for all true positive detections we recorded the location error, or the genomic distance between simulated and detected QTL. The mean and the 95% quantile of the location error are reported as stabilized estimates for different numbers of strains and QTL effect sizes, but averaged over all other conditions, in **Figure 2.4**. (The stabilization procedure is described in **Methods**; raw, unstabilized estimates provided **Figure 2.12**.) The location error statistics require careful interpretation: for a detection to be classed as a true positive it had to be within 5Mb of the simulated QTL; therefore, location error was artificially capped at 5Mb. Mediocre performance thus corresponds to when that location seems uniformly (and therefore arbitrarily) distributed over the $\pm$5Mb interval, that is, having a mean of 2.5Mb and a 95% quantile of 4.8Mb.

Location error was improved (reduced) by increasing the number of strains, increasing the QTL effect size, or both. In particular, as with power, location error was improved by increasing the number of strains even when while holding the total number of mice constant (**Figure 2.13**),
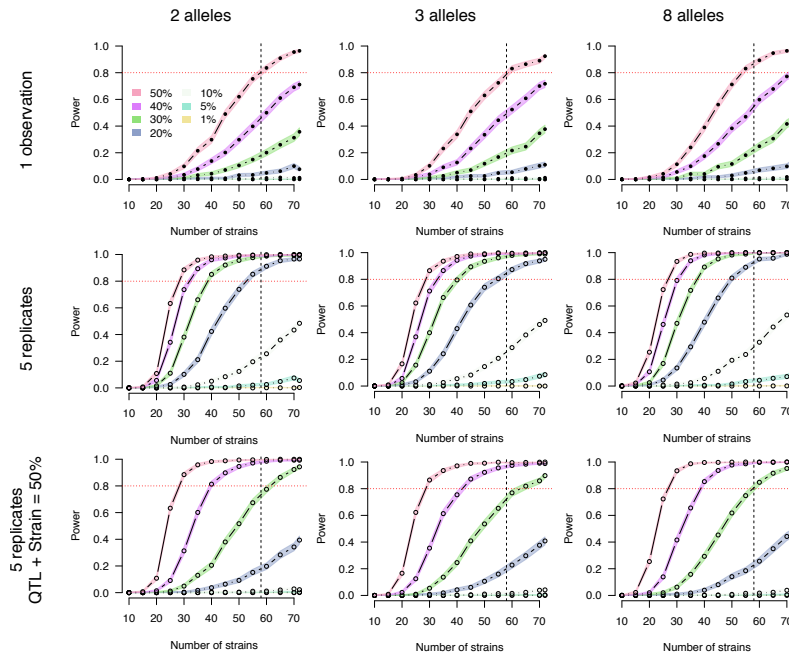
Figure 2.3: Power curves by number of CC strains. Results are stratified by a number of replicates, background strain effect size, and the number of functional alleles. The **[top]** row is based on a single observation per strain and no background strain effect. The **[middle]** row corresponds to five replicates per strain and no background strain effect. For the **[bottom row]**, five replicates are observed and the QTL effect size and background strain effect size sum to 50%, thus penalizing smaller QTL more harshly. The horizontal red dotted line marks 80% power. The vertical black dashed line marks 58 strains, which is currently the number of unrelated strains available from UNC. The columns, left to right, correspond to two, three, and eight functional alleles. Closed circles represent power estimates that were directly assessed, whereas open circles were interpolated. Simulations are based on Definition B.

consistent with mapping resolution being improved by an increased number of recombination events in the QTL region. Distributions of raw location error, stratified by levels of the number of strains, the number of functional alleles, and the QTL effect size can be found in **Figure 2.15**.

2.4.4 False positive rate

The FPR for the QTL power simulations was estimated as the percentage of scans (per setting) that produced a statistically significant signal on a chromosome without a QTL, shown in **Figure 2.11**. As expected, FPR was not elevated from 5% when the strain effects were simulated independently, as the effects were exchangeable by construction. The FPR did not vary with the number of strains or the number of alleles.

| 1 obs | QTL | | Power | | | | | | | | |
| | 3 rep | 5 rep | 30 strains | | | 50 strains | | | 72 strains | | |
			2 alleles	3 alleles	8 alleles	2 alleles	3 alleles	8 alleles	2 alleles	3 alleles	8 alleles
0.01	0.003	0.002	0.001	0.000	0.000	0.000	0.001	0.001	0.001	0.000	0.000
0.05	0.017	0.010	0.001	0.001	0.002	0.004	0.000	0.001	0.007	0.000	0.003
0.1	0.036	0.022	0.001	0.001	0.001	0.006	0.003	0.004	0.013	0.013	0.014
0.15	0.056	0.034	0.001	0.003	0.002	0.009	0.011	0.014	0.035	0.054	0.041
0.2	0.077	0.048	0.006	0.009	0.003	0.032	0.026	0.030	0.077	0.110	0.124
0.25	0.100	0.062	0.002	0.011	0.015	0.076	0.061	0.066	0.207	0.231	0.252
0.3	0.125	0.079	0.011	0.014	0.010	0.105	0.118	0.116	0.357	0.377	0.441
0.35	0.152	0.097	0.018	0.024	0.034	0.194	0.207	0.261	0.553	0.564	0.633
0.4	0.182	0.118	0.035	0.038	0.056	0.298	0.335	0.383	0.711	0.717	0.792
0.45	0.214	0.141	0.048	0.063	0.078	0.456	0.467	0.539	0.858	0.857	0.905
0.5	0.250	0.167	0.098	0.102	0.114	0.620	0.630	0.712	0.964	0.924	0.964
0.55	0.289	0.196	0.156	0.180	0.208	0.789	0.784	0.860	0.977	0.961	0.993
0.6	0.333	0.231	0.272	0.251	0.304	0.914	0.896	0.935	0.990	0.984	0.998
0.65	0.382	0.271	0.387	0.412	0.486	0.953	0.934	0.985	0.993	0.992	0.999
0.7	0.438	0.318	0.603	0.582	0.635	0.983	0.965	0.994	0.998	0.993	1.000
0.75	0.500	0.375	0.780	0.746	0.818	0.990	0.986	0.999	0.998	0.999	1.000
0.8	0.571	0.444	0.890	0.851	0.923	0.995	0.991	1.000	0.999	1.000	1.000
0.85	0.654	0.531	0.932	0.927	0.983	0.997	0.995	0.999	1.000	1.000	1.000
0.9	0.750	0.643	0.970	0.955	0.994	0.999	0.999	1.000	1.000	0.999	1.000
0.95	0.864	0.792	0.976	0.966	1.000	0.999	0.998	1.000	1.000	1.000	1.000

Table 2.1: QTL mapping power in the Collaborative Cross based on QTL effect sizes in a balanced population (Definition B). Replicates assume on no background strain effect. Convert QTL effect sizes from experiments with replicates to mean scale with Eq 2.4.

In additional null simulations that where strain effects were correlated due to realized genomic similarity, QTL scans assuming independent strain effects (and thus, exchangeability) had elevated FPR (**Figure 2.6** and **Table 2.3**). Using all 72 CC strains, the FPR varied from a maximum of 14.5% when strain effects explain all variability to the well controlled FPR of 5.5% when the strain effects were relatively small. Omitting CC059, one of the highly-related cousin strains (CC053 and CC059), because of its obvious violation of equal relatedness, reduced the FPR, although it was still elevated (12.9% for maximum strain effect). This demonstrates that, when strain effects are large relative to individual error (i.e. highly heritable trait, or the use of many replicates), failure to account for population structure due to realized imbalance in founder contributions can increase the risk of false positives.

2.4.5 Beavis effect

It is an expected feature of QTL mapping studies that estimates of QTL effect size, when calculated only for detected QTL, will be biased upwards. This phenomenon, known as the Beavis effect, is a form of selection bias and as such is expected to be most extreme under low power conditions, *e.g.*, when detection rates are low and/or estimates have high variance.

We explored the Beavis effect in our simulations. Assuming a one-replicate ($r = 1$) experiment, we found that, for example, the estimated effect size of a simulated 20% QTL was inflated by 3-fold when mapping in 40 CC strains, and by 2-fold when mapped in 72 CC strains. More generally, and as expected, the Beavis effect was reduced with larger numbers of strains and larger QTL effect sizes (**Figure 2.7**).

These results also imply that the Beavis effect is reduced by replication, at least to the extent that replication boosts effective QTL effect size. For example, consider again the mapping of a 20% QTL effect in 40 strains, which with $r = 1$ replicates implies 3-fold effect size inflation. Although this inflation could be reduced to 2-fold by increasing the number of strains to 72, the same reduction could be achieved by replication: assuming no background strain effect, increasing replicates to a theoretical $r = 1.8$ (so as to give a total sample size of $N = 40 \times 1.8 = 72$) would boost the QTL effect size to an effective $\approx 31\%$ (according to Eq 2.4) and, as shown in **Figure 2.7**, have approximately the same result. The ability of replicates to reduce the Beavis effect, however, will

28

diminish to the extent that there is a significant background strain effect, following the general relationship of replicates and QTL effect size described in Eq 2.4.

2.4.6 Allele frequency imbalance reduces power

For a fixed set of QTL allele effects, it is expected that power will always be greatest when allele frequencies are balanced. Accordingly, when QTL effect size was defined in terms of the variance that would be explained in a theoretical population with balanced allele frequencies (Definition B), deviations from balance in the mapping population—either from imbalance in functional alleles among the founders or imbalance of the founders among the CC strains—inevitably reduce power (**Figure 2.8A**). This reduction in power under Definition B is most evident for bi-allelic QTL (pink), in which the potential imbalance in allelic series is most extreme, namely when a single founder carries one functional allele and the other seven possess the alternative allele (7v1).

Conversely, when the QTL effect size is defined in terms of variance explained in the mapping population (Definition DAMB, which is similar to an R^2 measure), power remains constant across different allelic series and degrees of balance. Although note that this definition carries with it the (possibly unrealistic) implication that allele effects vary depending what population they are in.

When averaged over many allelic series, QTL mapping power based on Definition B is reduced relative to Definition DAMB, with the greatest reduction occurring for bi-allelic QTL (**Figure 2.8 B**). Though this modest reduction in power may seem to suggest that simulating with respect to a balanced population (Definition B) versus the mapping population (Definition DAMB) is unimportant in terms of designing a robust mapping experiment in the CC, we reiterate the value of using Definition B. Specifically, simulating with respect to Definiton DAMB is overly optimistic regarding mapping power for QTL with imbalanced allelic series.

We performed additional simulations to evaluate bi-allelic QTL in more detail, these being more prone to drastic imbalance under Definition B. All 127 possible bi-allelic series are visualized as a grid in **Figure 2.9A**, ordered from balance and high power to imbalance and low power. The corresponding power estimates are shown in **Figure 2.9B**. Power was maximized when the bi-allelic series is balanced (4v4; 35/127 possible allelic series) and minimized when imbalanced (7v1; 8/127 possible allelic series). Uniform sampling of bi-allelic series, the approach in the more general simulations described earlier, slightly reduced power relative to balanced 4v4 allelic series due to

29
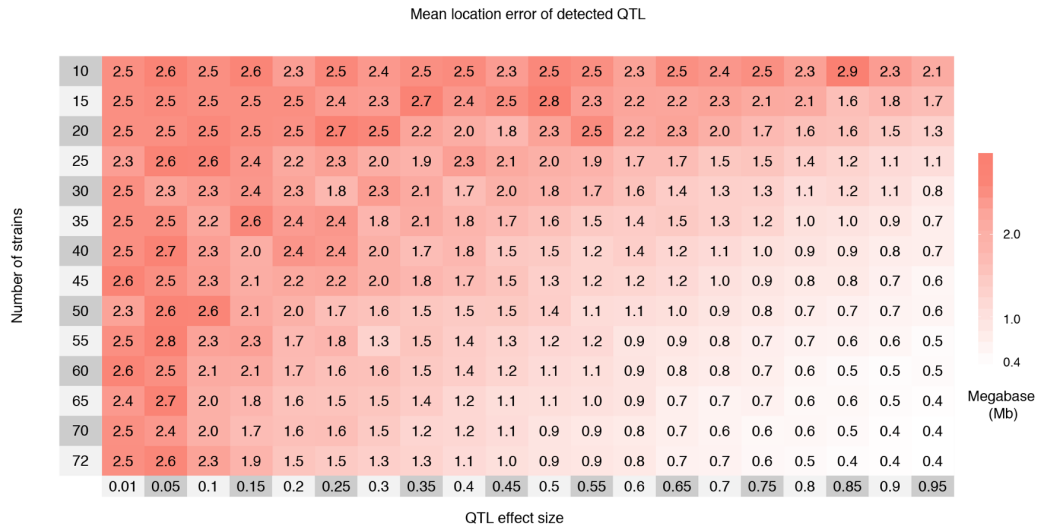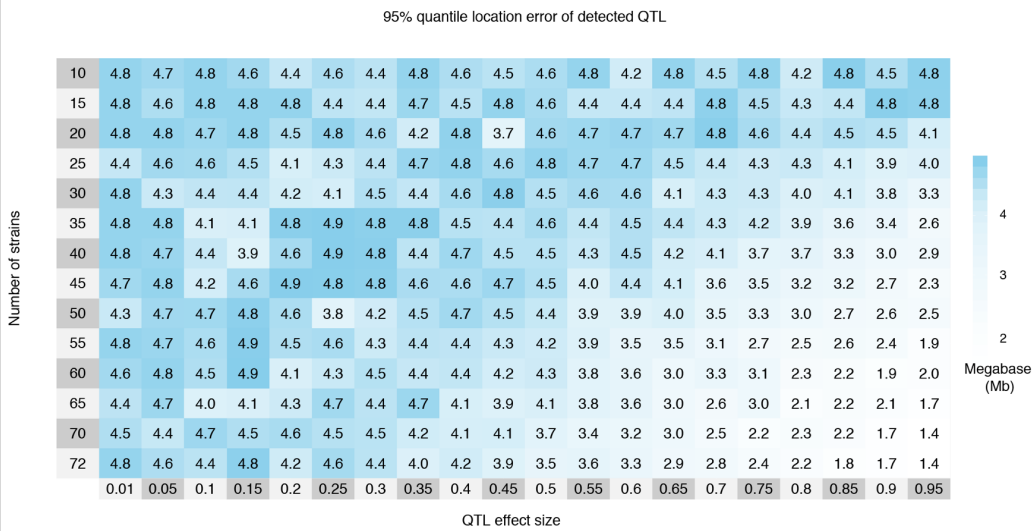
Figure 2.4: The mean (A) and 95% quantile (B) of location error, the distance in Mb between the detected and simulated QTL, by effect size and number of strains for 1,000 simulations of each setting. The simulations are based on Definition B with an eight allele QTL, and only a single observation per strain. Cells are colored red to white with decreasing mean and blue to white with decreasing 95% quantile. Regularization of the means and 95% quantile was accomplished through averaging the observed results with pseudo-counts; see **Figure 2.12** for the raw measurements. Increasing the number of strains reduces the location error, both in terms of the mean and 95% quantile, more so than QTL effect size, also shown in **Figure 2.15**. The maximum possible location error was 5Mb due to the 10Mb window centered around the true QTL position used for detecting QTL.
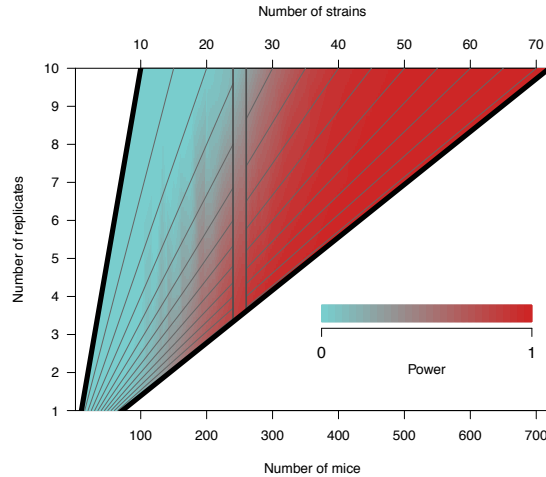
Figure 2.5: Heatmap of QTL mapping power by number of replicates and total number of mice in the experiment. Power is based on a QTL effect size of 20%, no background strain effect, and two functional alleles, though varying these parameters does not affect the dynamic between number of strains and replicates. The gray diagonal lines represent fixed values of the number of CC strains, ranging from 10 to 70 in intervals of five. Holding the total number of mice fixed, power is reduced as the percentage of the sample that are replicates is increased. This is illustrated with a cutout band centered on 250 mice, where power is lower at the top of the band when replicate mice are a relatively higher proportion of the total number of mice.
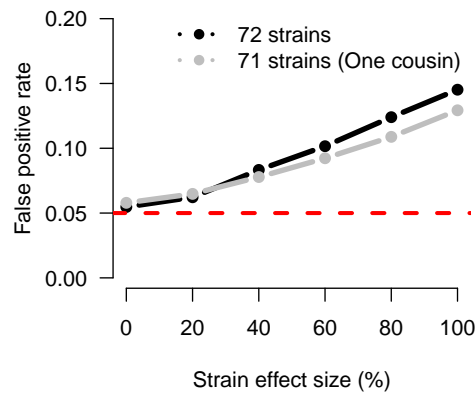
Figure 2.6: The FPR increases due to population structure among the realized genomes of the CC strains in the presence of a background strain effect and no QTL. Curves are based on 10,000 simulations for each setting of strain effect and strain sample, based on a single observation per strain. The inflation in FPR is greater for all 72 CC strains, which includes two closely related cousin strains (CC051 and CC059). Removing CC059 reduces the inflation in FPR (gray line). The dashed red line marks the specified type I error rate of 0.05, which is approximately met as expected when no strain effect is simulated, as in **Figure 2.11**. **Table 2.3** reports the specific FPR values.
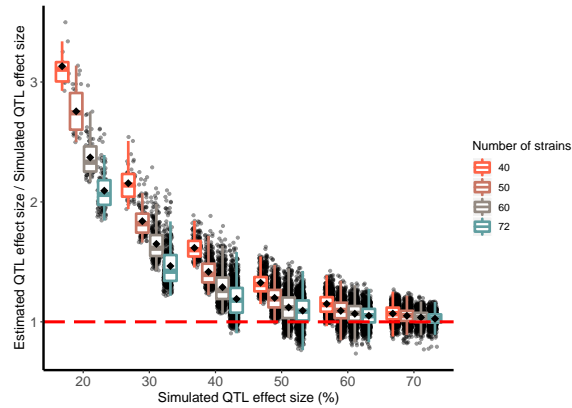
Figure 2.7: The Beavis effect (inflation of QTL effect size estimates) is more pronounced with smaller simulated QTL effect sizes and reduced numbers of strains. For different settings of numbers of strains (40, 50, 60, 72) and simulated QTL effect sizes (20%, 30%, 40%, 50%, 60%, 70%), black dots plot the ratio of the estimated effect size at a detected QTL peak to the effect size that was simulated at the true QTL locus. Out of 1,000 simulations under each setting, only successful detections are shown. Black diamonds represents the mean ratio for a category; horizontal red dashed line marks a ratio of 1, when QTL effect size estimates are unbiased (*i.e.*, no Beavis effect).
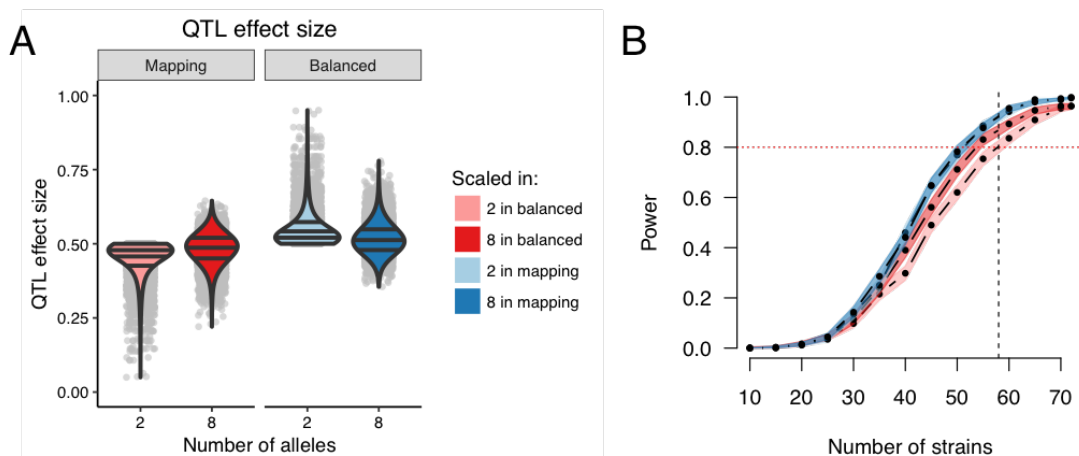
Figure 2.8: QTL effect sizes are in reference to a population, though effect size in the specific mapping population will determine the mapping power. Consider two populations as examples: the mapping population (definition DAMB) and a population balanced in the functional alleles (definition B). (A) QTL effect size distributions based on 10,000 simulations of the QTL for 72 strains. Using definition B, the effect sizes for the mapping population for two alleles is pink and eight alleles is red. Using definition DAMB, the effect sizes in the balanced population for two alleles is light blue and eight alleles is dark blue. Horizontal lines within the violin plots represent the 25th, 50th, and 75th quantiles from the estimated densities. Gray dots represent actual data points. (B) Power curves corresponding to the previously described settings of alleles and QTL effect size definitions. Power curves are estimated from 1,000 simulations per number of strains for a 50% QTL, no background strain effect, and a single observation per strain. The horizontal red dotted line marks 80% power. The vertical black dashed line marks 58 strains, which is currently the number of unrelated strains available from UNC.
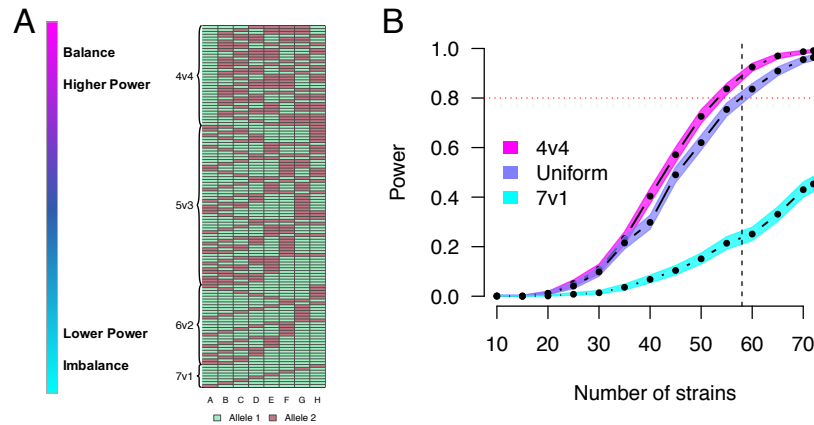
Figure 2.9: The balance of the allelic series for QTL with two functional alleles, and its effect on QTL mapping power. (A) The 127 possible allelic series for a bi-allelic QTL, categorized by the balance in the distribution of alleles among the CC founder strains, and ordered with balanced allelic series at the top and imbalanced at the bottom. (B) Power curves comparing three different sampling approaches for the allelic series with two functional alleles, for populations simulated to have a QTL effect size of 50% in a balanced theoretical population, with a single observation per CC strain. The horizontal red dotted line marks 80% power. The vertical black dashed line marks 58 strains, which is currently the number of unrelated strains available from UNC.

averaging over many cases of balance and some cases of extreme imbalance. These latter, more focused simulations highlight the extent that the reduction in QTL effect size, and thus mapping power, when simulating based on Definition B, is highly dependent on the allelic series. This could be of particular importance when considering QTL that result from a causal variant inherited from a wild-derived founder, such as CAST, which will present as both imbalanced and bi-allelic.

2.5 Discussion

Now that the CC strains have been largely finalized, it is possible to investigate more deeply how, in potential mapping experiments, power is affected by factors such as the number of strains, the number of replicates, and the allelic series at the QTL. We find that the CC can powerfully map large effect QTL ($\geq 50\%$) with single observations of >60 strains. Through the use of replicates, the power to map QTL can be greatly improved, potentially mapping QTL $\geq 20\%$ in 60 strains with 5 replicates per strain with no background strain effect. To guide the design of new CC experiments, we provide broad power curves and tables in **Figure 2.3** and **Tables 2.1** and **2.2**.

The power calculations described here take advantage of realized CC genomes, allowing the power estimates to be highly specific to the available strains but also necessarily restricting the number that can be used. This differs from the simulations of (Valdar et al., 2006a), which primarily focused on comparing potential breeding designs with numbers of strains that far exceed (500-1,000) the realized population (50-70). As such, directly comparing these studies is challenging. The closest comparison case is for a 5% QTL with 45% background strain effect with 100 simulated strains with 10 replicates, for which (Valdar et al., 2006a) estimates 4% power. Matching those settings with the exception of 72 strains instead of 100, and using the DAMB definition of QTL effect size, we find 0.4% power. The relatively lower power with the realized data likely reflects both reduction in the number of strains by 28% (72 to 100) and the deviations from an ideally-randomized population, such as the observed reduction in contributions from the CAST and PWK founders (Srivastava et al., 2017). This emphasizes the challenge in projecting the results from (Valdar et al., 2006a) into the realized population for the purpose of designing an experiment.

We did not attempt power simulations with epistatic QTL or phenotypes with large background strain effect. From the results of (Valdar et al., 2006a), it was clear that mapping studies in the realized CC, even with replicates, would not be well-powered in those contexts. Nonetheless, despite the reduced number of strains of realized population, we found that successful mapping experiments can be designed in the realized CC, particularly by harnessing the ability of genetic replicates to reduce random noise, as well as within the context of molecular phenotypes such as gene expression for which the genetic architecture is relatively simple.

2.5.1 Interpreting QTL effect sizes

Our simulations suggest that QTL mapping experiments in the CC are well-powered for large-effect QTL, in the neighborhood of 20-40%, depending on the number of strains and replicates, and the presence of a background strain effect. As such, it is useful to provide some context for what traits might plausibly yield QTL of this size. That said, we note that comparisons of reported estimates of QTL effect size should be interpreted with caution since they vary across different traits and model systems, are calculated under different experimental protocols that may imply different levels of noise, such as different numbers of strains or replicates, and may be estimated by different analysis

conditions (statistical methods, data transformations, etc.). And ultimately, these estimates are subject to overestimation due to both the aforementioned Beavis effect and reporting bias.

Multiple studies in the pre-CC, which had more strains than the realized CC population, have reported QTL effect sizes for a variety of traits. (Philip et al., 2011) report effect sizes for 17 QTL for 102 morphological and behavioral traits in 235 incipient CC strains, ranging from 5.3% (tail-clip latency) to 26% (red cell distribution width). (Durrant et al., 2011) mapped seven QTL for susceptibility to *Aspergillus fumigatus* infection in 371 mice from 66 strains, with effects ranging from 12.2-16.2%. (Gralinski et al., 2015) identified four SARS susceptibility QTL in 140 strains with effect sizes between 21-26% (vascular cuffing, 21% and 26%; viral titer, 22%; eosinophilia, 26%).

More closely mirroring the number of strains considered here, (Levy et al., 2015) detected six strong QTL for traits related to trabecular bone microstructure using 160 mice from 31 strains, which ranged from 61-86%. In an ongoing project involving the mapping of expression QTL (eQTL) from RNA-seq data collected from three tissues of single individuals from 47 strains, 478-739 eQTL were detected at genome-wide significance, ranging in effect size from 60-90%. These results reiterate that QTL mapping studies in the CC are best suited for detection of large effect QTL, as are more common in molecular traits.

In considering the above, it is useful to understand how this relates to effect sizes seen in humans, for which the CC is often used as a model system (Flint and Mackay, 2009). In particular, human GWASs, which often use much larger sample sizes, routinely report QTL with estimated effect sizes far smaller than is detectable in the CC. Nonetheless, there are reasons to expect effect sizes in the CC to be larger than in humans. Human GWASs are observational, and as such include many additional sources of noise, reducing QTL effect sizes relative to what would be possible in more tightly-controlled experimental designs. Experimental populations will also have larger QTL effect sizes because: 1) they typically have more balanced allele frequencies; 2) in the case of panels of RILs such as the CC, because they are homozygous across the genome, which increases the contrast in additive allele effects and thus boosts additive QTL effect size; and 3), again for RILs, because they furnish biological replicates, which, as illustrated in Eq 2.4, can increase effect size by reducing individual error.

2.5.2 Strains versus replicates

When holding the total number of mice fixed, we found that adding more strains improves power and reduces location error to a greater degree than does adding more replicates. Moreover, this inference was made in the absence of a background strain effect—given that replicates reduce individual-level variance but not strain-level variance, the presence of background effects would reduce the relative value of replicates yet further. These observations are consistent with the results of Valdar et al. (2006a) and established theoretical arguments (Soller and Beckmann, 1990; Knapp and Bridges, 1990).

Nonetheless, for many CC mapping experiments we predict that adding replicates will provide considerable value. First, for all but the most highly polygenic traits, mapping on the means of replicates, a strategy originally termed "replicated progeny" (Cowen, 1988) or "progeny testing" (Lander and Botstein, 1989), will always provide additional power. Indeed, with a limited number of strains available, and the possibility that all available strains are used, replication may sometimes be the only way power can be further increased (Belknap, 1998).

Second, replicates provide not only an insurance policy against phenotyping errors, but also a way to average over batches and similar nuisance parameters (Cowen, 1988), thus protecting against the negative consequences of gene by environment interactions while also providing the opportunity for such interactions to be detected [*e.g.*, Kafkafi et al. (2005, 2018)].

Third, replicates enable deeper phenotypic characterization and in particular measurement of strain-level phenotypes that are necessarily a function of multiple individuals. For example, treatment response phenotypes (*e.g.*, response to drug) are ideally defined in terms of counterfactual-like observations of drug-treated and vehicle-treated strain replicates [*e.g.*, Festing (2010); Crowley et al. (2014)] and recombinant inbred lines such as the CC are uniquely able to combine such definitions with QTL mapping [*e.g.*, Mosedale et al. (2017) and also, in flies, Kislukhin et al. (2013); Najarro et al. (2015)]. Similarly, strain-specific phenotypic variance ideally requires replicates (Rönnegård and Valdar, 2011; Ayroles et al., 2015). We did not consider such elaborations here, but we expect the trade-off between number of strains vs replicates will be more nuanced in such cases.

2.5.3 Population structure in the CC

Our simulations indicate that deviations from equal relatedness in the realized CC strains have introduced a degree of population structure that potentially increases the risk of false positives if not addressed, albeit to a far lesser extent than has been observed in traditional inbred strain association (Eskin et al., 2008). In particular, null simulations that assumed correlated strain effects due to genetic relatedness increased FPR for our mapping approach when the strain effect was large relative to individual error, as would be the case for a highly heritable polygenic trait or when using many replicates. This elevated FPR supports the use of QTL mapping approaches that account for the effect of genetic similarity on phenotypes, such as a mixed effect model (Eskin et al., 2008; Kang et al., 2010; Lippert et al., 2011; Zhou and Stephens, 2012), especially in the context of marginally significant QTL, which may not remain significant given a higher threshold that controls FPR more appropriately. Software packages that can fit the LMM specifically with CC data include our miQTL package (available on GitHub at `https://github.com/gkeele/miqtl`) and R/qtl2 (Broman et al., 2019).

For the analyses reported here, a mixed effect model approach was not feasible owing to its increased computational burden (and in particular, its incompatibility with the computational shortcut in **Appendix A**). Instead, we simulated independent strain effects and employed a fixed effect mapping procedure due to its computational efficiency, especially when computing permutation-based significance thresholds. Nonetheless, the conclusions drawn in this study should be largely consistent with the use of a mixed effect model that correctly controls for correlated strain effects due to genetic relatedness.

2.5.4 Allelic series, and use of an eight allele mapping model

We found that the allelic series can strongly affect power through its influence on observed allele frequencies. Specifically, imbalanced bi-allelic QTL have significantly reduced mapping power whereas highly multi-allelic QTL do not because the potential for imbalance is reduced.

Regardless of the true allelic series at a QTL, which is unknown in practice, our statistical procedure assumed an eight allele model. For QTL with fewer functional alleles than founder strains, this assumption could reduce power due to the estimation of redundant allele effect parameters.

Indeed, QTL consistent with a bi-allelic series have been more powerfully detected in some MPP studies using SNP association (Baud et al. 2013; Keele et al. 2018).

Nonetheless, multi-allelic QTL (with more than two alleles) do occur. This has been seen, for example, in cis-regulation of gene expression that largely corresponds to the three subspecies lineages of *Mus musculus*, present in the CC (Crowley et al., 2015). Moreover, multi-allelic QTL will not be as powerfully detected through SNP association, as seen, for example, in Aylor et al. (2011). SNP (or more generally, variant) association also poses additional challenges, such as how to handle regions of the genome (and variants) that are difficult to genotype, as well as the requirement of extensive quality control filtering to remove markers with low minor allele frequencies. These challenges are implicitly reduced in haplotype analysis.

An ideal statistical procedure would formally model the unknown allelic series and their corresponding uncertainty. Though challenging, the development of alternative mapping strategies that specifically account for the allelic series is clearly an imperative methodological advance that would greatly benefit QTL analyses in MPPs with diverse founder alleles. That said, allelic series-aware approaches would likely be computationally expensive and poorly suited to simulation-based power analyses. Meanwhile, in the absence of more sophisticated approaches, the eight allele model, though potentially redundant, has several advantages over SNP association that suggest it will remain a useful (and maybe the default) tool for CC mapping, namely: it encompasses all possible simpler allelic series, implicitly models local epistasis, and, in reflecting the LD decay around detected QTL, more clearly delineates the limits of mapping resolution.

2.5.5 Inclusion of extinct CC strains in simulations

Our simulations included genomes from CC strains that are now extinct, and also did not include all the CC strains that are currently available. This discrepancy reflects the inherent challenge of maintaining a stable genetic population resource. RI panels, such as the CC, are an approximation to an ideal: they attempt to provide reproducible genomes that can be observed multiple times as well as across multiple studies; yet, as a biological population, the genomes are mutable, and through time will accumulate mutations and drift, and even potentially go extinct.

Although the inclusion of genomes of extinct strains, or those that have drifted since the strains were genotyped, result in power calculations that do not perfectly correspond to the current CC

population, they are preferable to simulated genomes, since they represent genomes that were viable at some point. We view the use of extinct genomes as realistic observations of possible genomes that reflect both the potential that more strains will become extinct or be gained from other breeding sites with time, and thus can be reasonably extended to the realized population, now and into the future.

2.5.6 Future use and directions

Any analysis of power is subject to the assumptions underlying that analysis. One of the advantages of simulation is the ability to evaluate the impact of many of these assumptions, as well as the consideration of new scenarios by re-running the simulation under different settings, or by elaborating the simulation itself. We have attempted to make re-running the simulations under different settings straightforward for other researchers by developing a software package for this purpose. This package could be used to investigate highly-specialized questions, such as the power for specific combinations of CC strains or assessing how the power to detect QTL varies depending on genomic position. In future work, the simulation code itself could be expanded to investigate additional topics of interest, such as how variance heterogeneity or model mis-specification influence power.

2.5.7 Conclusion

We used a focused simulation approach that incorporates realized CC genomes to provide more accurate estimates of QTL mapping power than were previously possible. As such, the results of our simulations provide tailored power calculations to aide the design of future QTL mapping experiments using the CC. Additionally, we evaluate how the balance of alleles at the QTL can strongly influence power to map QTL in the CC. We make available the R package SPARCC that we developed for running these simulations and analyses. It leverages an efficient model fitting approach in order to explore power in a level of detail that has previously been impractical, it is replicable, and it can be extended to user-specified questions of interest.

2.6 Acknowledgments

2.7 Appendix A: QR decomposition for fast regression

To maximize power to detect QTL while controlling the FPR, permutations to determine significance thresholds are needed, which is computationally expensive and thus the underlying regression functionality must be highly optimized. We accomplish this through the QR matrix decomposition, which we will describe briefly (Venables and Ripley, 2002).

Let $\mathbf{X} = \mathbf{PA}$ be the $n \times m$ design matrix included in Eq 2.3, with $m = 8$. The solution for $\boldsymbol{\beta}$ from the least squares normal equations is $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}\mathbf{X}^\mathrm{T}\mathbf{y}$. Through the QR decomposition, $\mathbf{X} = \mathbf{QR}$, for which $\mathbf{Q}$ is an $n \times p$ orthonormal matrix ($\mathbf{Q}^\mathrm{T}\mathbf{Q} = \mathbf{I}$) and $\mathbf{R}$ is a $m \times m$ upper triangular matrix. With matrix algebra, it is fairly straightforward to show that $\widehat{\boldsymbol{\beta}} = \mathbf{R}^{-1}\mathbf{Q}^\mathrm{T}\mathbf{y}$, which is also more numerically stable than calculating $\widehat{\boldsymbol{\beta}}$ through $(\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}$. After solving for $\widehat{\boldsymbol{\beta}}$, the residual sums of squares, and ultimately logP, can be rapidly calculated. Because our simulation approach involves regressing many permuted outcomes $\mathbf{U}_p\mathbf{y}^{(s)}$, where $\mathbf{U}_p$ is a permutation matrix that re-orders $\mathbf{y}^{(s)}$ randomly, on the same design matrices, computational efficiency can be vastly increased by pre-computing and saving the QR decompositions for all $\mathbf{X}$.

Once the QR decomposition has been stored for a design matrix $\mathbf{X}_j$, j indexing locus, it is highly computationally efficient to conduct additional tests for any $\mathbf{y}$, thus encompassing all permuted outcomes $\mathbf{U}_p\mathbf{y}$. If $\mathbf{X}_j$ is the same across S simulations, the boost in computation can extend beyond permutations to samples of $\mathbf{y}^{(s)}$, as is the case when the set of CC strains is fixed. In effect, two cases result for our R package SPARCC: when the set of CC strains is fixed, and when the set varies.

- Fixed set of CC strains

    1. Store QR decompositions of $\mathbf{X}_j$ for $j = 1, 2, \ldots, J$

    2. Run genome scans for $\mathbf{y}^{(s)}$ and $\mathbf{U}_p\mathbf{y}^{(s)}$ for $s = 1, 2, \ldots, S \times p = 1, 2, \ldots, P$

- Varied set of CC strains

1. Store QR decompositions of $\mathbf{X}_{js}$ for $j = 1, 2, \ldots, J$

2. Run genome scans for $\mathbf{y}^{(s)}$ and $\mathbf{U}_p\mathbf{y}^{(s)}$ for $p = 1, 2, \ldots, P$

3. Repeat steps 1 and 2 for $s = 1, 2, \ldots, S$

Varying the sets of CC strains increases computation time linearly with respect to S. If the investigators do not have a predefined set of strains, it is appropriate that this source of variability be incorporated into the power calculation.

2.8   Appendix B: Computing environment and performance

We performed 1,000 simulations (in batches of 100) for each combination of the parameters, resulting in 8,400 individual jobs. These jobs were submitted in parallel to a distributed computing cluster (`http://its.unc.edu/rc-services/killdevil-cluster/`). Runtime varied depending on parameter settings and the hardware used, with the longest jobs taking approximately seven hours to complete.

2.9   Appendix C: CC strains

This study used haplotype mosiac data available from `http://csbio.unc.edu/CCstatus/` `index.py?run=FounderProbs` for the following 72 CC strains: CC001, CC002, CC003, CC004, CC005, CC006, CC007, CC008, CC009, CC010, CC011, CC012, CC013, CC014, CC015, CC016, CC017, CC018, CC019, CC020, CC021, CC022, CC023, CC024, CC025, CC026, CC027, CC028, CC029, CC030, CC031, CC032, CC033, CC034, CC035, CC036, CC037, CC038, CC039, CC040, CC041, CC042, CC043, CC044, CC045, CC046, CC047, CC048, CC049, CC050, CC051, CC052, CC053, CC054, CC055, CC056, CC057, CC058, CC059, CC060, CC061, CC062, CC063, CC065, CC068, CC070, CC071, CC072, CC073, CC074, CC075, CC076. This includes two strains CC051 and CC059 that are derived from the same breeding funnel and thus more closely related than typical pairs of CC strains.

Of the the 72 CC strains used here, 54 are among a larger set of 59 that are currently maintained and distributed by UNC (personal correspondence with Darla Miller, UNC). These 54/59 strains are CC001, CC002, CC003, CC004, CC005, CC006, CC007, CC008, CC009, CC010, CC011, CC012,

CC013, CC015, CC016, CC017, CC019, CC021, CC023, CC024, CC025, CC026, CC027, CC029, CC030, CC031, CC032, CC033, CC035, CC036, CC037, CC038, CC039, CC040, CC041, CC042, CC043, CC044, CC045, CC046, CC049, CC051, CC053, CC055, CC057, CC058, CC059, CC060, CC061, CC062, CC065, CC068, CC071, CC072. The remaining 5/59 strains (CC078, CC079, CC080, CC081, CC083) lacked haplotype mosaic data at the time of simulation and so were not included (although note that their mosaics have since been added to the website).

2.10   Appendix D: Additive model and allelic series matrices

2.10.1   Additive matrix

We can use matrices to specify simplifying linear combinations of the 36 diplotypes. The additive model matrix $\mathbf{A}$ is commonly used, and we use it here. Post-multiplication of the diplotype design matrix $\mathbf{D}$ with the $\mathbf{A}$ rotates the diplotypes at the locus to dosages of the founder haplotypes. If there is no uncertainty on the diplotype identities, $\mathbf{DA}$ will be the matrix of founder haplotype counts at the locus.

$$
\mathbf{A} = \begin{array}{c} \\ \text{AA} \\ \text{BB} \\ \text{CC} \\ \text{DD} \\ \text{EE} \\ \text{FF} \\ \text{GG} \\ \text{HH} \\ \text{AB} \\ \text{AC} \\ \text{AD} \\ \text{AE} \\ \text{AF} \\ \text{AG} \\ \text{AH} \\ \text{BC} \\ \text{BD} \\ \text{BE} \\ \vdots \end{array}
\begin{array}{cccccccc}
\text{A} & \text{B} & \text{C} & \text{D} & \text{E} & \text{F} & \text{G} & \text{H} \\
\left[\begin{array}{cccccccc}
2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots
\end{array}\right]
\end{array}
$$

$$
\mathbf{A} =
\begin{array}{c}
\\
\vdots \\
\text{BF} \\
\text{BG} \\
\text{BH} \\
\text{CD} \\
\text{CE} \\
\text{CF} \\
\text{CG} \\
\text{CH} \\
\text{DE} \\
\text{DF} \\
\text{DG} \\
\text{DH} \\
\text{EF} \\
\text{EG} \\
\text{EH} \\
\text{FG} \\
\text{FH} \\
\text{GH}
\end{array}
\begin{array}{cccccccc}
A & B & C & D & E & F & G & H \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1
\end{array}
$$

2.10.2 Allelic series matrices

We explore the influence of the allelic series on QTL mapping power through the simulation procedure. The QTL mapping procedure estimates separate parameters for each founder, though in reality, there are likely fewer functional alleles. We denote the qth functional allele as k_q. The allelic series can be sampled and encoded in the M.ID argument within the sim.CC.data() function of SPARCC. Below are examples of balanced (4v4) and unbalanced (7v1) bi-allelic series, as well as tri-allelic series.

2.10.2.1 Allelic series with eight alleles (maximum)

$$\texttt{M.ID = ``0,1,2,3,4,5,6,7''}$$

$$
\mathbf{M} = \mathbf{I} =
\begin{array}{c}
\\
A \\ B \\ C \\ D \\ E \\ F \\ G \\ H
\end{array}
\begin{array}{cccccccc}
k_0 & k_1 & k_2 & k_3 & k_4 & k_5 & k_6 & k_7 \\
\left[\begin{array}{cccccccc}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
\end{array}\right]
\end{array}
$$

2.10.2.2 Example balanced (4v4) bi-allelic series

$$\texttt{M.ID = ``0,1,0,0,1,0,1,1''}$$

$$
\mathbf{M} =
\begin{array}{c}
\\
A \\ B \\ C \\ D \\ E \\ F \\ G \\ H
\end{array}
\begin{array}{cc}
k_0 & k_1 \\
\left[\begin{array}{cc}
1 & 0 \\
0 & 1 \\
1 & 0 \\
1 & 0 \\
0 & 1 \\
1 & 0 \\
0 & 1 \\
0 & 1 \\
\end{array}\right]
\end{array}
$$

$$\texttt{M.ID = ``0,1,1,1,0,0,1,0''}$$

$$\mathbf{M} = \begin{array}{c} \\ \text{A} \\ \text{B} \\ \text{C} \\ \text{D} \\ \text{E} \\ \text{F} \\ \text{G} \\ \text{H} \end{array} \begin{array}{cc} k_0 & k_1 \\ \left[\begin{array}{cc} 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{array} \right] \end{array}$$

2.10.2.3 Example unbalanced (7v1) bi-allelic series

$$\text{M.ID} = \text{``}0,0,0,0,0,1,0,0\text{''}$$

$$\mathbf{M} = \begin{array}{c} \\ \text{A} \\ \text{B} \\ \text{C} \\ \text{D} \\ \text{E} \\ \text{F} \\ \text{G} \\ \text{H} \end{array} \begin{array}{cc} k_0 & k_1 \\ \left[\begin{array}{cc} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \end{array} \right] \end{array}$$

$$\text{M.ID} = \text{``}0,1,0,0,0,0,0,0\text{''}$$

$$\mathbf{M} = \begin{array}{c} \\ \text{A} \\ \text{B} \\ \text{C} \\ \text{D} \\ \text{E} \\ \text{F} \\ \text{G} \\ \text{H} \end{array} \begin{array}{cc} k_0 & k_1 \\ \left[\begin{array}{cc} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{array} \right] \end{array}$$

2.10.2.4 Example tri-allelic series

$$\texttt{M.ID = ``0,0,1,2,2,0,2,0''}$$

$$\mathbf{M} = \begin{array}{c} \\ \text{A} \\ \text{B} \\ \text{C} \\ \text{D} \\ \text{E} \\ \text{F} \\ \text{G} \\ \text{H} \end{array} \begin{array}{ccc} k_0 & k_1 & k_2 \\ \left[\begin{array}{ccc} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{array} \right] \end{array}$$

$$\texttt{M.ID = ``0,1,0,0,0,0,2,2''}$$

$$\mathbf{M} = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \\ G \\ H \end{array} \overset{\displaystyle k_0 \quad k_1 \quad k_2}{\left[\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{array} \right]}$$

2.11   Data and Supplement details

Data and analysis files in the **Supplement** include:

- File_S1: data_supplement_details.pdf - Overview of data and **Supplement**.

- File_S2: sparcc_1.1.1.tar.gz - SPARCC R package used for all analyses.

- File_S3: sparcc_cache.zip - CC haplotype mosaics formatted for SPARCC.

- File_S4: sparcc_powersim.R - Example R script to perform large-scale power analysis.

- File_S5: sparcc_powersim.sh - Example shell script to coordinate separate calls to
  sparcc_powersim.R.

- File_S6: collapse_results.R - Example R script to aggregate results from multiple calls to
  sparcc_powersim.R.

- File_S7: sparcc_options_tutorial.pdf - Provides description of options in SPARCC, and provides
  simple tutorials for using the package.

- File_S8: generate_figures.R - R script to generate figures in manuscript and **Supplement**.

- File_S9: supplement_tables_figures.pdf - The supplementary tables and figures.

2.11.1   SPARCC Package

The static version of the SPARCC R package (1.1.1) used in this paper is provided in the **Supplement** as File_S2. R and shell scripts are provided to generate the same results as File_S4, File_S5, File_S6, and File_S8. Static versions of these files have been placed on figshare. The current version of the SPARCC package is available here: `https://github.com/gkeele/sparcc`. SPARCC can be installed using the command 'R CMD INSTALL' at the terminal. The current version can be conveniently installed using the devtools R package and the following command within R: 'install_github("gkeele/sparcc")'.

2.11.2 Data objects included in SPARCC package

- r1.dat: Data frame of power and FPR results for Definition B from 1,000 simulations per combination of

  - Number of strains: [(10-70 by 5), 72]

  - Single observation per strain

  - Number of alleles: [2, 3, 8]

  - $h^2_{\text{strain}} = 0$

  - h^2_{QTL}: [0.01, (0.05-0.95 by 0.05)]

- r1.damb.dat: Data frame of power and FPR results for Definition DAMB from 1,000 simulations per combination of

  - Number of strains: [(10-70 by 5), 72]

  - Single observation per strain

  - Number of alleles: [2, 3, 8]

  - $h^2_{\text{strain}} = 0$

  - h^2_{QTL}: [0.01, (0.05-0.95 by 0.05)]

- r1.4v4.dat: Data frame of power and FPR results for Definition B with bi-allelic series forced to be balanced (4 founders per functional allele) from 1,000 simulations per combination of

  - Number of strains: [(10-70 by 5), 72]

  - Single observation per strain

  - Number of alleles: 2

  - $h^2_{\text{strain}} = 0$

  - $h^2_{\text{QTL}} = 0.5$

- r1.7v1.dat: Data frame of power and FPR results for Definition B with bi-allelic series forced to be imbalanced (singler founder with one allele) from 1,000 simulations per combination of

  - Number of strains: [(10-70 by 5), 72]

- Single observation per strain

- Number of alleles: 2

- $h^2_{\text{strain}} = 0$

- $h^2_{\text{QTL}} = 0.5$

- r1.dist.dat: Data frame of mean location error results for Definition B from 1,000 simulations per combination of

  - Number of strains: [(10-70 by 5), 72]

  - Single observation per strain

  - Number of alleles: [2, 3, 8]

  - $h^2_{\text{strain}} = 0$

  - h^2_{QTL}: [0.01, (0.05-0.95 by 0.05)]

- r1.exchange.dat: Data frame of FPR results for Definition B from 10,000 null simulations with correlated strain effects based on **K** per combination of

  - Number of strains: [71 (excluding CC059), 72]

  - Single observation per strain

  - Number of alleles: 0

  - h^2_{strain}: [0-1 by 0.2]

  - $h^2_{\text{QTL}} = 0$

- r1.beavis.dat: Data frame of QTL effect size estimations for Definition DAMB from 1,000 simulations per combination of

  - Number of strains: [(40-60 by 10), 72]

  - Single observation per strain

  - Number of alleles: 2

  - $h^2_{\text{strain}} = 0$

  - h^2_{QTL}: [(20-70 by 10)]

51

- K: Matrix of dimension 72×72 representing the realized genetic relationship matrix of the 72 CC strains, calculated from the founder mosaics.

2.11.3 File types

- *.R - These are R scripts used for the analyses and figures.

- *.sh - These are bash scripts used to run large-scale analysis.

- *.RData - These files are contained in the happy_formated sparcc_cache directory that can be loaded in R and are necessary for SPARCC to run.

2.12 Supplemental Tables and Figures
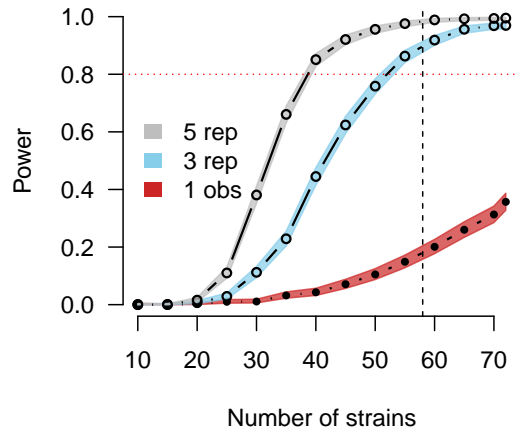
Figure 2.10: Power estimates for experiments with three and five replicates interpolated from estimates from only a single observation per CC strain. Power curves correspond to a QTL with effect size of 30% and two functional alleles. QTL effect sizes for experiments with replicates are adjusted based on Eq 2.4, allowing for results from single observation simulations to be projected into experiments with replicates. Pre-computed power estimates for single observation simulations are stored in SPARCC and can conveniently be extrapolated into other settings, as is demonstrated here. The horizontal red dotted line marks 80% power. The vertical black dashed line marks 58 strains, which is currently the number of unrelated strains available from UNC. Closed circles represent power estimates that were directly evaluated. Open circles represent power estimates that were interpolated from single observation results.

| | QTL | | 30 strains | | | 50 strains | | | 72 strains | | |
| | | | | | | Power | | | | | |
1 obs	3 rep	5 rep	2 alleles	3 alleles	8 alleles	2 alleles	3 alleles	8 alleles	2 alleles	3 alleles	8 alleles
0.01	0.003	0.002	0.000	0.002	0.001	0.000	0.000	0.000	0.000	0.003	0.000
0.05	0.017	0.010	0.000	0.001	0.001	0.000	0.002	0.000	0.005	0.003	0.004
0.1	0.036	0.022	0.004	0.002	0.000	0.009	0.007	0.003	0.015	0.016	0.018
0.15	0.056	0.034	0.002	0.006	0.001	0.016	0.017	0.018	0.056	0.051	0.046
0.2	0.077	0.048	0.008	0.007	0.003	0.032	0.038	0.032	0.119	0.141	0.135
0.25	0.100	0.062	0.009	0.005	0.008	0.071	0.066	0.088	0.264	0.264	0.281
0.3	0.125	0.079	0.015	0.013	0.014	0.141	0.120	0.134	0.460	0.466	0.492
0.35	0.152	0.097	0.028	0.029	0.030	0.234	0.229	0.262	0.695	0.664	0.684
0.4	0.182	0.118	0.045	0.038	0.040	0.415	0.376	0.413	0.854	0.848	0.854
0.45	0.214	0.141	0.082	0.074	0.078	0.603	0.594	0.620	0.958	0.964	0.974
0.5	0.250	0.167	0.136	0.134	0.143	0.769	0.783	0.783	0.996	0.998	0.999
0.55	0.289	0.196	0.198	0.204	0.248	0.911	0.922	0.924	1.000	0.999	1.000
0.6	0.333	0.231	0.334	0.331	0.328	0.985	0.980	0.994	0.999	1.000	1.000
0.65	0.382	0.271	0.519	0.489	0.534	0.998	0.995	0.999	0.999	1.000	1.000
0.7	0.438	0.318	0.707	0.703	0.756	0.998	0.999	1.000	0.999	0.999	1.000
0.75	0.500	0.375	0.866	0.864	0.914	0.998	0.998	1.000	0.999	1.000	1.000
0.8	0.571	0.444	0.940	0.954	0.979	0.996	0.998	1.000	1.000	1.000	1.000
0.85	0.654	0.531	0.962	0.967	0.995	0.998	0.999	1.000	1.000	0.999	1.000
0.9	0.750	0.643	0.978	0.981	0.998	0.999	0.998	1.000	1.000	1.000	1.000
0.95	0.864	0.792	0.970	0.988	0.999	0.999	0.999	1.000	0.999	1.000	1.000

Table 2.2: QTL mapping power in the Collaborative Cross based on QTL effect sizes in a balanced population (Definition DAMB). Replicates assume on no background strain effect. Convert QTL effect sizes from experiments with replicates to mean scale with Eq 2.4.

54

| | False positive rate | |
Background Strain	71 strains	72 strains
0	0.058	0.055
0.2	0.0648	0.062
0.4	0.078	0.083
0.6	0.092	0.102
0.8	0.109	0.124
1	0.129	0.145

Table 2.3: False positive rate in the Collaborative Cross with no simulated QTL and the presence of population structure (also in Figure 2.6). Background strain effects were assumed to be correlated based on the realized genomic similarity of the 72 strains. We also excluded CC059, cousin strain of CC051, for a total of 71 strains.
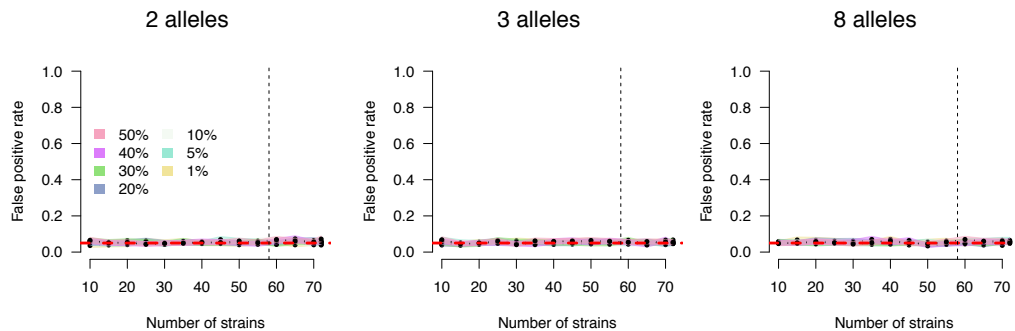
Figure 2.11: False positive rate (FPR) based on 1,000 simulations per setting with respect to number of CC strains, stratified by the number of functional alleles. The horizontal red dashed line marks the 5% type I error (false positive) rate. CC strains and loci were varied in simulations, resulting in false positive rates that average over loci and strain combinations. Confidence intervals were calculated based on Jeffreys interval (Brown et al., 2001) for a binomial proportion. Plots, left to right, correspond to two, three, and eight functional alleles. The FPR represents the probability that any QTL is detected on chromosomes other than the chromosome on which the simulated QTL is located. The significance thresholds maintain the desired type I error rate of 0.05. As expected, the allelic series does not appear to influence FPR. The vertical black dashed line marks 58 strains, which is currently the number of unrelated strains available from UNC.

A

Raw mean location error of detected QTL

Number of strains	0.01	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95
10	■	4.0	■	3.6	0.6	2.8	1.9	■	2.7	1.6	2.8	■	1.1	■	1.8	■	1.6	3.5	2.2	2.0
15	■	2.7	■	■	■	1.1	0.8	4.5	1.5	■	3.7	0.7	1.4	1.9	2.0	1.9	2.0	1.5	1.7	1.7
20	■	■	2.6	■	2.3	3.4	3.0	1.4	1.7	1.2	2.1	2.5	2.1	2.2	2.0	1.7	1.6	1.5	1.4	1.3
25	0.6	3.3	3.5	1.9	1.1	2.0	1.5	1.5	2.2	2.1	1.9	1.8	1.6	1.6	1.4	1.5	1.3	1.2	1.1	1.1
30	■	1.2	0.8	2.0	1.4	1.3	2.2	2.0	1.5	1.9	1.8	1.7	1.5	1.4	1.3	1.3	1.1	1.2	1.0	0.8
35	■	■	1.5	2.6	2.2	2.4	1.6	2.0	1.8	1.7	1.6	1.4	1.4	1.4	1.3	1.2	1.0	0.9	0.9	0.7
40	■	4.7	0.5	1.3	2.4	2.3	1.9	1.6	1.7	1.4	1.4	1.2	1.3	1.2	1.1	1.0	0.9	0.9	0.7	0.7
45	3.8	■	2.0	1.7	2.0	2.2	1.9	1.7	1.7	1.5	1.3	1.2	1.2	1.1	1.0	0.9	0.8	0.8	0.7	0.6
50	0.0	4.0	2.9	1.9	1.8	1.5	1.5	1.4	1.4	1.4	1.3	1.1	1.1	1.0	0.9	0.8	0.7	0.7	0.6	0.6
55	■	3.7	2.0	2.2	1.6	1.7	1.2	1.5	1.4	1.2	1.2	1.1	0.9	0.9	0.8	0.7	0.6	0.6	0.6	0.5
60	3.6	■	1.7	1.9	1.6	1.5	1.6	1.4	1.3	1.2	1.1	1.0	0.9	0.8	0.8	0.7	0.6	0.5	0.4	0.5
65	1.2	3.2	1.5	1.6	1.4	1.5	1.5	1.4	1.2	1.1	1.0	1.0	0.9	0.7	0.6	0.7	0.5	0.5	0.5	0.4
70	2.4	2.2	1.7	1.5	1.5	1.6	1.4	1.2	1.2	1.1	0.9	0.9	0.8	0.7	0.6	0.5	0.5	0.5	0.4	0.3
72	■	2.7	2.1	1.7	1.4	1.4	1.3	1.2	1.1	1.0	0.9	0.9	0.8	0.7	0.7	0.5	0.5	0.4	0.4	0.3

QTL effect size — Megabase (Mb)

B

Raw 95% quantile location error of detected QTL

Number of strains	0.01	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95
10	■	4.0	■	3.6	0.6	2.8	2.9	■	2.7	3.0	2.8	■	1.4	■	1.8	■	2.5	4.9	4.3	4.8
15	■	2.7	■	■	■	1.1	0.8	4.5	1.5	■	4.1	0.7	3.1	4.0	4.8	4.4	4.2	4.3	4.8	4.8
20	■	■	4.5	■	2.3	4.8	3.0	2.3	4.8	2.8	4.6	4.7	4.7	4.7	4.8	4.6	4.4	4.5	4.5	4.1
25	0.6	4.1	3.5	1.9	1.9	3.6	4.1	4.7	4.8	4.6	4.8	4.7	4.7	4.5	4.3	4.3	4.3	4.1	3.9	4.0
30	■	2.1	0.8	2.6	2.3	3.7	4.3	4.3	4.6	4.8	4.5	4.6	4.6	4.1	4.3	4.3	4.0	4.1	3.7	3.3
35	■	■	2.4	3.4	4.8	4.9	4.8	4.8	4.4	4.4	4.6	4.4	4.5	4.4	4.3	4.2	3.9	3.6	3.4	2.6
40	■	4.7	0.5	2.4	4.6	4.9	4.8	4.4	4.7	4.5	4.5	4.3	4.5	4.2	4.1	3.6	3.7	3.2	3.0	2.9
45	3.8	■	2.7	4.5	4.9	4.9	4.8	4.6	4.6	4.7	4.5	4.0	4.4	4.1	3.6	3.5	3.2	3.2	2.7	2.3
50	0.0	4.0	4.5	4.8	4.5	3.6	4.2	4.5	4.7	4.5	4.4	3.9	3.9	4.0	3.5	3.3	3.0	2.7	2.6	2.5
55	■	4.7	4.4	4.9	4.5	4.6	4.3	4.4	4.4	4.3	4.2	3.9	3.5	3.5	3.1	2.7	2.5	2.5	2.4	1.9
60	3.6	■	4.4	5.0	4.0	4.3	4.5	4.4	4.4	4.2	4.2	3.8	3.5	3.0	3.3	3.0	2.3	2.2	1.9	1.9
65	1.2	4.7	3.2	3.9	4.2	4.7	4.4	4.7	4.1	3.9	4.1	3.7	3.6	2.9	2.6	3.0	2.1	2.2	2.0	1.7
70	2.4	3.6	4.7	4.5	4.6	4.5	4.4	4.2	4.1	4.1	3.7	3.3	3.2	3.0	2.5	2.1	2.3	2.2	1.7	1.3
72	■	3.9	4.2	4.8	4.2	4.6	4.4	4.0	4.2	3.9	3.5	3.6	3.3	2.9	2.8	2.3	2.1	1.7	1.6	1.4

QTL effect size — Megabase (Mb)

Figure 2.12: The raw mean (A) and 95% quantile (B) of the location error, the distance in Mb between the detected and simulated QTL, by effect size and number of strains for 1,000 simulations of each setting. These simulations are based on Definition B with an eight allele QTL, and only a single observation per strain. Cells are colored red to white with decreasing mean and blue to white with decreasing 95% quantile. Black cells represent the case in which no simulated QTL were detected. Estimates from poorly-powered settings are more likely to be unobserved or unstable from low detection. Regularized measurements are provided in **Figure 2.4**. Increasing the number of strains reduces both the mean and 95% quantile location error more so than QTL effect size, also shown in **Figure 2.15**. The maximum possible location error was 5Mb due to the 10Mb window centered around the true QTL position used for detecting QTL.

56

Total sample size: 250 mice
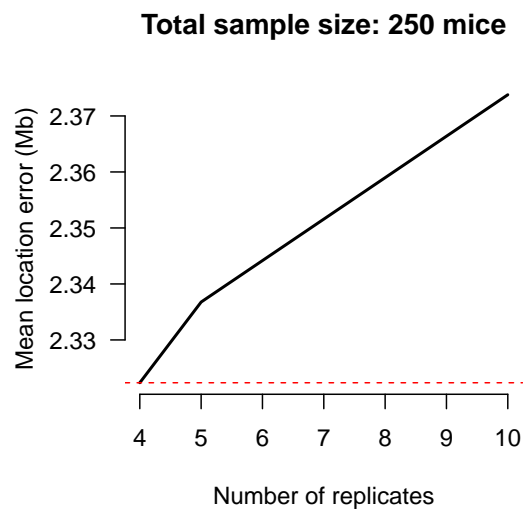
Figure 2.13: Mean location error of detected QTL increases with the number of replicates while keeping total sample size fixed. Estimates are based on linear interpolation from dense simulations using Definition B with single observations per strains. The total number of mice and the QTL effect size are fixed at 250 and 50%, respectively. The red dotted line highlights that the lowest mean location error occurs at 4, the lowest number of replicates possible for a sample of 250 mice, given the 72 strains used in the simulations.
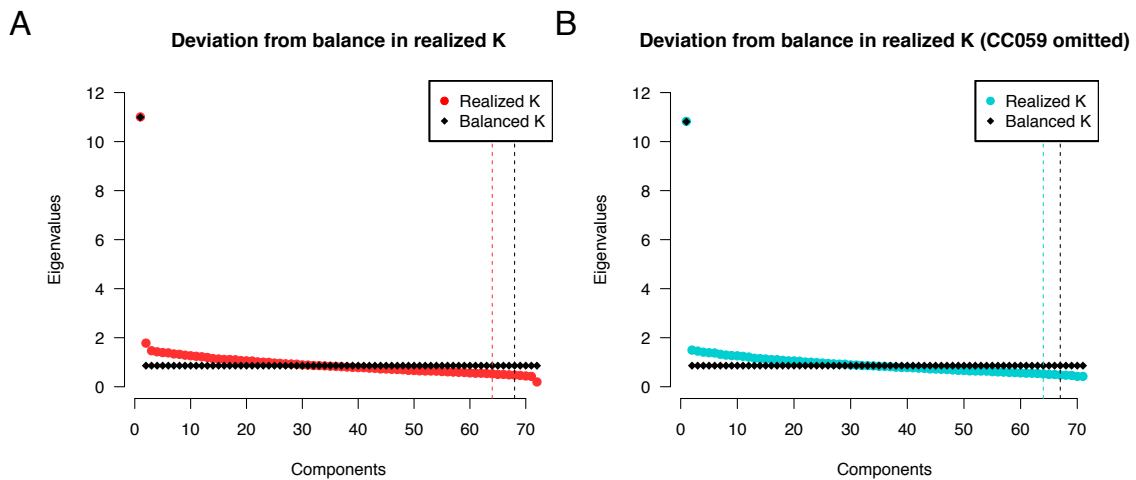
Figure 2.14: The realized genetic relationship matrix **K** deviates from a perfectly balanced population. Red and blue circles represent the eigenvalues of the eigendecomposition of the realized **K**, when including both cousin strains (A) and excluding one (B). Black diamonds represent the eigenvalues of a balanced **K**, with the relationship fixed at the mean relationship observed in the realized **K**. Vertical dashed lines represent the number of components necessary to explain 95% of the variation for the different **K**. The first eigenvalue represents the variation accounted for by the overall mean of **K**. In the balanced **K**, after removing the effect of the mean, all components contribute equally to the variance. The eigenvalue of the second component for the 72 strains is slightly inflated, representing the cousin strains, a notable deviation from equal relatedness. This inflation disappears when one of the cousin strains is removed, however population structure still persists.
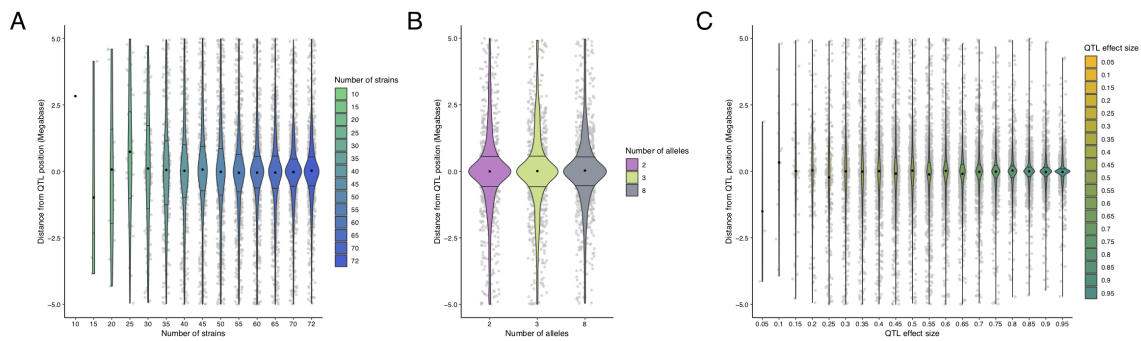
Figure 2.15: Distributions of the un-regularized location error, by number of strains (A), number of alleles (B), and QTL effect size (C). Observed distances are between -5 and 5Mb due to the 10Mb window centered around the simulated QTL that was used for QTL detection in the large scale results. Gray dots represent the distances for a single simulations. The colored violin plots represent the distribution of distances across the simulations. The black dot marks the mean location error for each category. Horizontal lines represent the 25th and 75th quantiles. (A) With QTL effect size fixed at 50% and the number of alleles at 8, as the number of CC strains increases, the distribution of location error becomes more concentrated around zero, meaning the mapping resolution improves with increasing numbers of strains. (B) With the QTL effect size again fixed at 50% and the number of strains fixed at 72, the distribution of distances does not appear to differ based on the number of functional alleles. (C) With the number of strains fixed at 72 and the number of alleles fixed at 8, as the QTL effect size increase, the distribution of distances becomes more concentrated around zero. These simulations are based on Definition B and single observations per strain. See **Figures 2.4** and **2.12** for specific estimates of location error over different settings of QTL effect size and numbers of strains.

CHAPTER 3

Methods for Allele-Based Approach

3.1 Overview

In this chapter, we introduce a fully-Bayesian framework for inferring the allelic series in multi-parental populations (MPPs). This framework encompasses both the single-variant and haplotype-based association approaches, and it permits optional prior information about haplotype relatedness, as encoded by a phylogenetic tree. The allele-based approach described in this chapter is the basis for the remaining chapters, which evaluate its properties and demonstrate its applications.

3.2 Introduction

At a QTL, a quantitative trait $\mathbf{y} = (y_1, ..., y_N)^T$ measured in N individuals $i = 1, ..., N$ is associated with genetic variation at a particular location in the genome. In a multiparental population with $J \geq 2$ founder strains $j = 1, ..., J$, this genetic variation is encoded by the pair of founder haplotypes, or the diplotype, present at the locus, denoted for each individual by the indicator vector $\mathbf{d}_i$. We are interested in understanding the genetic architecture of the phenotype, or the form of the relationship between $\mathbf{y}_i$ and $\mathbf{d}_i$.

Assume that the diplotype state of each individual is known. Then $\mathbf{d}_i$ is an indicator vector with length $J + \binom{J}{2}$ corresponding to the number of possible pairs of founder haplotypes. The diplotype states of all individuals are given by the matrix $\mathbf{D} = (\mathbf{d}_1, ..., \mathbf{d}_N)^T$ with dimension $N \times J + \binom{J}{2}$. For example, in a population with $J = 3$ founder haplotypes labeled as $\{A, B, C\}$, a possible diplotype

state matrix is given by

$$
\mathbf{D} = \begin{array}{c}
\begin{array}{cccccc} AA & BB & CC & AB & AC & BC \end{array} \\
\begin{array}{c} 1 \\ 2 \\ \vdots \\ N \end{array}
\left[\begin{array}{cccccc}
0 & 0 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 1 & 0 & 0 & 0 & 0
\end{array} \right]
\end{array}.
$$

The first individual has the AC diplotype, or one copy of both the A and C haplotypes, while the second individual has the AA diplotype, or two copies of the A haplotype.

We assume an additive genetic model, which defines the phenotypic effect of each diplotype as the sum of two haplotype effects. In our example, the AC diplotype state of the first individual has effect $0.5\beta_A + 0.5\beta_C$ on the phenotype, while the AA diplotype state of the second individual has effect $0.5\beta_A + 0.5\beta_A$. More generally, the diplotype effects are given by the matrix product

$$
\boldsymbol{\beta}_{\text{dip}} = \mathbf{A}\boldsymbol{\beta}_{\text{hap}},
$$

where $\boldsymbol{\beta}_{\text{dip}}$ is a length $J + \binom{J}{2}$ vector of diplotype state effects, $\boldsymbol{\beta}_{\text{hap}}$ is a length J vector of haplotype effects, and $\mathbf{A}$ is a $J + \binom{J}{2} \times J$ matrix that maps diplotype states to (half) counts of haplotypes, termed the additive design matrix. The diplotype effects for our example, decomposed into the

product of the additive design matrix and the haplotype effects, are given by

$$
\mathbf{A}\boldsymbol{\beta}_{\mathrm{hap}} =
\begin{array}{c}
\\ AA \\ BB \\ CC \\ AB \\ AC \\ BC
\end{array}
\begin{array}{ccc}
A & B & C \\
\left[\begin{array}{ccc}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1 \\
0.5 & 0.5 & 0 \\
0.5 & 0 & 0.5 \\
0 & 0.5 & 0.5
\end{array}\right]
\end{array}
\times
\left[\begin{array}{c}
\beta_A \\
\beta_B \\
\beta_C
\end{array}\right],
$$

$$
\boldsymbol{\beta}_{\mathrm{dip}} =
\begin{array}{c}
AA \\ BB \\ CC \\ AB \\ AC \\ BC
\end{array}
\left[\begin{array}{c}
\beta_A \\
\beta_B \\
\beta_C \\
0.5\beta_A + 0.5\beta_B \\
0.5\beta_A + 0.5\beta_C \\
0.5\beta_B + 0.5\beta_C
\end{array}\right],
$$

and the diplotype effects of all individuals are then the matrix product $\mathbf{D}\boldsymbol{\beta}_{\mathrm{dip}} = \mathbf{D}\mathbf{A}\boldsymbol{\beta}_{\mathrm{hap}}$.

Assume that the phenotype is completely explained by the diplotype effects and normally-distributed individual error, i.e. there are no other covariates, replicates observations, or population structure. Then the relationship between $\mathbf{y}$ and $\mathbf{D}$ is given by the likelihood function

$$
\mathbf{y}|\mu, \mathbf{D}, \boldsymbol{\beta}_{\mathrm{hap}}, \sigma \sim \mathrm{N}(\mu\mathbf{1} + \mathbf{D}\mathbf{A}\boldsymbol{\beta}_{\mathrm{hap}}, \sigma^2\mathbf{I}_N), \tag{3.1}
$$

where μ is the intercept and σ^2 scales the error variance of the phenotype. This model is overparameterized, with μ and $\boldsymbol{\beta}_{\mathrm{hap}}$ fitting $J + 1$ variables to the J levels of the data, but we ignore this for now and address it later. A more familiar form substitutes $\mathbf{X} = \mathbf{D}\mathbf{A}$, where $\mathbf{X}$ is a $N \times J$ design matrix of haplotype counts; however, we use the previous notation because it emphasizes the additive relationship between $\mathbf{y}$ and $\mathbf{D}$.

A standard Bayesian regression approach assumes that the haplotype effects are *a priori* distributed according to

$$\boldsymbol{\beta}_{\text{hap}}|\sigma, \phi \sim \text{N}(\mathbf{0}, \sigma^2\phi^2\mathbf{I}_J),$$

where ϕ controls the size of the haplotype effects relative to individual error (Servin and Stephens, 2007). This fits an independent haplotype effect for each of the J founder haplotypes, implicitly assuming that, with respect to the phenotype, each founder haplotype is functionally distinct. This assumption is unrealistic, however, if the causal variant underlying the QTL is a biallelic SNP. Then, the founder haplotypes fall into two functionally distinct groups based on the presence or absence of the SNP, and additional distinctions among the haplotypes are redundant, as in Yalcin et al. (2005). For example, if haplotypes A and C both have functional allele k, then their effects are equal to the effect of that allele: $\beta_A = \beta_C = \beta_k$. More complicated cases, such as local epistasis between multiple nearby variants, could lead to some number of K functional alleles $k = 1, ..., K$ where $2 \le K \le J$, as in King et al. (2014). Thus, the haplotype effects are dependent: they depend on the number of functional alleles and how those alleles are distributed among the haplotypes. This mapping of haplotypes to functional alleles is the allelic series.

We propose an approach that extends the additive genetic model to explicitly account for the allelic series, as in Jannink and Wu (2003). Our approach decomposes the haplotype effects into the product of the allelic series matrix and a vector of allele effects:

$$\boldsymbol{\beta}_{\text{hap}} = \mathbf{M}\boldsymbol{\beta}_{\text{alle}}, \tag{3.2}$$

where $\boldsymbol{\beta}_{\text{alle}} = (\beta_1, ..., \beta_K)^T$ is a length K vector of allele effects, $\mathbf{M} = (\mathbf{m}_1, ..., \mathbf{m}_J)^T$ is a $J \times K$ matrix denoting the allelic series, and $\mathbf{m}_j$ is a length K indicator vector denoting the allele assignment of strain j. In our example, haplotypes A and C share one of two functional alleles, and

the corresponding haplotype effects $\boldsymbol{\beta}_{\mathrm{hap}}$ are given by the matrix product

$$
\mathbf{M}_{(0,1,0)}\boldsymbol{\beta}_{\mathrm{alle}} =
\begin{array}{c}
\\
A \\
B \\
C
\end{array}
\overset{\displaystyle k=1 \quad k=2}{\left[
\begin{array}{cc}
1 & 0 \\
0 & 1 \\
1 & 0
\end{array}
\right]}
\times
\left[
\begin{array}{c}
\beta_1 \\
\beta_2
\end{array}
\right],
$$

$$
\boldsymbol{\beta}_{\mathrm{hap}} =
\begin{array}{c}
A \\
B \\
C
\end{array}
\left[
\begin{array}{c}
\beta_1 \\
\beta_2 \\
\beta_1
\end{array}
\right],
$$

where the subscript of $\mathbf{M}$ is a unique identifier of the allelic series configuration it encodes. The haplotype effects $\boldsymbol{\beta}_{\mathrm{hap}}$ are no longer independent and functionally distinct, but instead are comprised of repeated values of a smaller set of allele effects $\boldsymbol{\beta}_{\mathrm{alle}}$. More generally, the allelic series matrix $\mathbf{M}$ partitions the J haplotypes into K functional alleles, which also determines the number of allele effects in $\boldsymbol{\beta}_{\mathrm{alle}}$. If the allelic series is known and $K < J$, this approach will estimate $\boldsymbol{\beta}_{\mathrm{hap}}$ more efficiently than the standard haplotype-based approach because it fits only K allele effects, rather than J redundant haplotype effects.

The allelic series is rarely known *a priori*, but it may be inferred from the data. From a Bayesian perspective, we are interested in the allelic series' posterior distribution,

$$
\mathrm{p}(\mathbf{M}|\mathbf{y}) = \frac{\mathrm{p}(\mathbf{y}|\mathbf{M})\mathrm{p}(\mathbf{M})}{\mathrm{p}(\mathbf{y})},
$$

which requires specifying a prior distribution over the space of possible partitions encoded by $\mathbf{M}$. In our example with $J = 3$ haplotypes, the space of allelic series is easily enumerated and includes a null configuration where all haplotypes are functionally identical,

$$
\mathbf{M}_{(0,0,0)} =
\begin{array}{c}
A \\
B \\
C
\end{array}
\overset{\displaystyle k=1}{\left[
\begin{array}{c}
1 \\
1 \\
1
\end{array}
\right]},
$$

64

all the partitions of three haplotypes into two functional alleles,

$$
\mathbf{M}_{(0,0,1)} = \begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} k{=}1 \quad k{=}2 \\ \left[\begin{array}{cc} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{array} \right] \end{array},
$$

$$
\mathbf{M}_{(0,1,0)} = \begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} k{=}1 \quad k{=}2 \\ \left[\begin{array}{cc} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{array} \right] \end{array},
$$

$$
\mathbf{M}_{(0,1,1)} = \begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} k{=}1 \quad k{=}2 \\ \left[\begin{array}{cc} 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{array} \right] \end{array},
$$

and a configuration where all the haplotypes are functionally distinct,

$$
\mathbf{M}_{(0,1,2)} = \begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} k{=}1 \quad k{=}2 \quad k{=}3 \\ \left[\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{array} \right] \end{array}.
$$

Specifying a prior distribution over this space involves simultaneously defining expectations about the number of functional alleles and which combinations of haplotypes are more or less likely to be functionally distinct. This is particularly challenging when there are more founder haplotypes, as the space of allelic series partitions becomes exceedingly large.

Partition problems are common in Bayesian nonparametric statistics, and our approach is closely related to the popular Dirichlet process (DP) with a normal base distribution, i.e.,

$$
\boldsymbol{\beta}_{\mathrm{hap}} | \sigma, \phi, \alpha \sim \mathrm{DP}(\alpha, \mathrm{N}(0, \sigma^2 \phi^2)),
$$

where α is the concentration parameter (Escobar and West, 1995). Under the DP, the corresponding prior distributions on the decomposed elements of $\boldsymbol{\beta}_{\mathrm{hap}}$ are

$$\mathbf{M}|\alpha \sim \mathrm{CRP}(\alpha),$$

$$\boldsymbol{\beta}_{\mathrm{alle}}|\sigma, \phi \sim \mathrm{N}(\mathbf{0}, \sigma^2\phi^2\mathbf{I}_K),$$

where CRP denotes the Chinese restaurant process. In the CRP, the concentration parameter α controls the prior distribution of the number of functional alleles, and the distribution over particular allelic configurations is implied by the process itself. Specifically, the CRP assigns a haplotype to an allele conditionally, in proportion to the number of haplotypes already assigned to that allele, without considering which particular haplotypes comprise the allele. In this way, the CRP is uninformative with respect to the relationship between individual haplotypes. We use the CRP as a starting point for directly modeling the allelic series, eventually modifying it in order to introduce additional prior information about haplotype relatedness.

In the remainder of the methods, we first describe the likelihood function in more detail, relaxing some assumptions and specifying convenient conjugate prior distributions for several variables. Then, we discuss the CRP and its properties in more detail, including using Ewens's sampling formula to show how it can be interpreted as a distribution over random coalescent trees with the haplotypes at the leaves. Next, we use this connection to the coalescent to define an informative prior distribution for the allelic series which reflects information about haplotype relatedness, as encoded by a phylogenetic tree. We then specify prior distributions for the remaining model parameters and discuss elicitation of prior hyperparameters. Last, we describe posterior inference via a partially-collapsed Gibbs sampler and show how the output of this sampler can be used to calculate the approximate marginal likelihood and Bayes factors (BFs).

3.3   Likelihood Function

The likelihood function defines the relationship between the phenotype and the diplotype states. We introduced a simplified version of the likelihood in **Equation 3.1** that describes the phenotype as a linear combination of additive haplotype effects and normally-distributed individual error. We also

decomposed the haplotype effects into the product of an allelic series configuration and allele effects in **Equation 3.2**. Substituting this decomposition into the likelihood, and allowing for the possibility of both other covariates and replicate observations, the likelihood function is given by

$$\mathbf{y}|\mu, \mathbf{D}, \mathbf{M}, \boldsymbol{\beta}_{\text{alle}}, \boldsymbol{\delta}, \sigma \sim$$

$$\mathrm{N}(\mu\mathbf{1} + \mathbf{DAM}\boldsymbol{\beta}_{\text{alle}} + \mathbf{Z}\boldsymbol{\delta}, \sigma^2\mathbf{W}^{-1}),$$

where $\mathbf{Z}$ is a $N \times P$ matrix of optional covariates, $\boldsymbol{\delta}$ is a length P vector of covariate effects, and $\mathbf{W}$ is a $N \times N$ diagonal matrix with the number of replicates for each observation on the diagonal.

Our approach requires evaluating the likelihood over many settings of $\mathbf{M}$ with varying dimension. This motivates the use of conjugate priors for the allele effects, which allow us to simplify the likelihood by integrating, or collapsing, the allele effects out of the expression (Servin and Stephens, 2007). Thus, we use the conjugate normal-gamma prior distribution for the precision, intercept, allele effects, and covariate effects:

$$\sigma^{-2} \sim \mathrm{Ga}(0.5\kappa, 0.5\lambda),$$

$$\mu|\sigma \sim \mathrm{N}(0, \sigma^2\tau_\mu^2),$$

$$\boldsymbol{\beta}_{\text{alle}}|\sigma, \phi \sim \mathrm{N}(\mathbf{0}, \sigma^2\phi^2\mathbf{CC}^T),$$

$$\boldsymbol{\delta}|\sigma \sim \mathrm{N}(\mathbf{0}, \sigma^2\tau_\delta^2\mathbf{I}_P),$$

where κ, λ are shape and rate hyperparameters that control prior precision, and τ_μ, τ_δ are hyperparameters related to the prior intercept and covariate effect sizes.

We made note in the overview that the likelihood function is overparameterized, with μ and (now) $\boldsymbol{\beta}_{\text{alle}}$ fitting $K + 1$ parameters to K allelic levels of the data. For this reason, the prior covariance of $\boldsymbol{\beta}_{\text{alle}}$ includes a sum-to-zero constraint encoded by $\mathbf{C}$, a $K \times (K - 1)$ contrast matrix (Crowley

et al., 2014). For example, if $K = 3$, then the scaled prior covariance of the allele effects is given by

$$
\mathbf{C}\mathbf{C}^T = \begin{array}{c} \\ k=1 \\ k=2 \\ k=3 \end{array} \begin{array}{c} k=1 \quad k=2 \quad k=3 \\ \left[\begin{array}{ccc} 1 & -0.5 & -0.5 \\ -0.5 & 1 & -0.5 \\ -0.5 & -0.5 & 1 \end{array} \right] \end{array}.
$$

This matrix product is singular, reducing the dimension of the allele effect vector from K to $K - 1$, and the likelihood function is no longer overparameterized. Thus, the allele effects can be expressed as

$$
\boldsymbol{\beta}_{\text{alle}} = \mathbf{C}\boldsymbol{\beta},
$$

which is the product of the contrast matrix C and a $K - 1$ vector of independent effects β with prior distribution

$$
\boldsymbol{\beta} | \sigma, \phi \sim \text{N}(\mathbf{0}, \sigma^2 \phi^2 \mathbf{I}_{K-1}).
$$

Note that under this constraint, the intercept μ is interpreted as the mean phenotype, adjusted for covariates, in a population that is balanced with respect to the functional alleles, and the allele effects are defined relative to this point. Thus, comparing haplotype effects across allelic series requires making inference on the full prediction $\mu\mathbf{1} + \mathbf{M}\mathbf{C}\boldsymbol{\beta}$, which accounts for this shifting point of reference.

The intercept, independent allele effects and covariate effects, given together by the vector $\boldsymbol{\theta}^T = \begin{bmatrix} \mu & \boldsymbol{\beta}^T & \boldsymbol{\delta}^T \end{bmatrix}$, are jointly normally distributed according to

$$
\boldsymbol{\theta} | \sigma, \phi \sim \text{N}(\mathbf{0}, \sigma^2 \mathbf{V}),
$$

where $\mathbf{V}$ is a $(K + P) \times (K + P)$ diagonal matrix encoding the scaled prior covariance

$$
\mathbf{V} = \begin{array}{c} \\ \mu \\ \boldsymbol{\beta} \\ \boldsymbol{\delta} \end{array} \begin{array}{c} \mu \qquad \boldsymbol{\beta} \qquad \boldsymbol{\delta} \\ \left[\begin{array}{ccc} \tau_\mu & 0 & 0 \\ 0 & \phi^2 \mathbf{I}_{K-1} & 0 \\ 0 & 0 & \tau_\delta \mathbf{I}_P \end{array} \right] \end{array}.
$$

The likelihood can be rewritten as

$$\mathbf{y}|\mu, \mathbf{D}, \mathbf{M}, \boldsymbol{\beta}_{\mathrm{alle}}, \boldsymbol{\delta}, \sigma \sim \mathrm{N}(\mathbf{X}\boldsymbol{\theta}, \sigma^2 \mathbf{W}^{-1}),$$

where $\mathbf{X}$ is the $N \times (K + P)$ concatenated design matrix

$$\mathbf{X} = \begin{array}{ccc} {\mu} & {\boldsymbol{\beta}} & {\boldsymbol{\delta}} \\ \left[\begin{array}{ccc} \mathbf{1} & \mathbf{DAMC} & \mathbf{Z} \end{array} \right]. \end{array}$$

Conjugacy yields a closed form for a simplified, t-distributed likelihood function:

$$\mathbf{y}|\mathbf{D}, \mathbf{M}, \phi \sim \mathrm{t}_\kappa(\mathbf{0}, \lambda[\mathbf{W}^{-1} + \mathbf{X}\mathbf{V}\mathbf{X}^T]).$$

This simplified likelihood depends only on the diplotype states, allelic series configuration, and relative variance of the allele effects—this proves useful during posterior inference. We will describe prior distributions for $\mathbf{D}$ and ϕ later, focusing next on prior distributions for $\mathbf{M}$, the space of possible allelic configurations.

3.4 Prior Distribution of the Allelic Series

3.4.1 Chinese Restaurant Process

Specifying a prior distribution over the allelic series involves defining expectations about the number of functional alleles and likely allelic configurations. This is challenging because the space of possible allelic series is quite large even when the number of haplotypes is small. For example, the CC has $J = 8$ founder haplotypes and 4,140 possible allelic series; the DSPR has $J = 15$ and over 1.3 billion possibilities (Rota, 1964). Encoding specific prior intuitions about such a space is difficult. It is tempting to consider a uniform prior over the allelic series $[\mathrm{p}(\mathbf{M}) \propto 1]$ which allows the likelihood to drive posterior inference about the allelic series. This is the implicit prior for frequentist approaches such as King et al. (2014), or Yalcin et al. (2005) (conditional on $K = 2$). However, in most cases the number of observations will be much smaller than the number of possible allelic configurations, and this low-data scenario is precisely when prior information is most important. Instead of posterior

inference being dominated by the likelihood, it will be sensitive to the properties of the uniform distribution, which include a strong prior belief in an intermediate number of functional alleles and a lack of flexibility to calibrate this belief.

Partition problems occur frequently in Bayesian nonparametric statistics, and a common and more flexible prior distribution is the CRP,

$$\mathbf{M}|\alpha \sim \mathrm{CRP}(\alpha),$$

with probability density function

$$\mathrm{p}(\mathbf{M}|\alpha) = \Gamma(\alpha)\Gamma(\alpha+J)^{-1}\alpha^K \prod_{k=1}^{K} \Gamma(J_k),$$

where α is a concentration parameter which controls the expected number of functional alleles, and J_k is the number of haplotypes assigned to allele k (Escobar and West, 1995). The CRP is widely used in partition problems because it is exchangeable, making it amenable to posterior sampling. Exchangeability means that the density function of the CRP can be factored into conditional distributions that describe the allele assignment of a particular haplotype given the allelic configuration of all the other haplotypes. It also means that this conditional density can be applied iteratively (and in any order), beginning with all haplotypes unassigned, to construct the unconditional density of $\mathbf{M}|\alpha$ (Welling, 2006).

The conditional probability density function of the CRP is given by

$$\mathrm{p}(\mathbf{M}_j|\alpha, \mathbf{M}_{-j}) = \begin{cases} J_k(J-1+\alpha)^{-1} & k \leq K \\ \alpha(J-1+\alpha)^{-1} & k = K+1, \end{cases}$$

where $\mathbf{M}_j$ is the allele assignment of haplotype j, and $\mathbf{M}_{-j}$ is the allelic configuration of the other $J-1$ haplotypes. The probability that haplotype j is assigned to allele k is proportional to the number of haplotypes already assigned to that allele, and the probability that haplotype j is assigned to a new allele is proportional to the concentration parameter α. This proportionality induces a "rich-get-richer" property that favors imbalanced allelic configurations (e.g. for $J = 8$, a biallelic contrast of 7 haplotypes vs 1 haplotype for $J = 8$, "7v1") over balanced configurations (e.g. an

even biallelic contrast, "4v4") (Wallach et al., 2008). Note that the conditional probability that a haplotype is assigned to an existing functional allele does not depend on which particular haplotypes have already been assigned to that allele, only the number that have been assigned. In this way, the CRP is uninformative with respect to the relationship between individual haplotypes.

However, the CRP does allow for control over the prior number of functional alleles via the concentration parameter. When $\alpha \to \infty$, all of the haplotypes will be assigned to a unique functional allele ($\mathbf{M} = \mathbf{I}$), which is identical to the standard haplotype approach which assumes that all J haplotypes are functionally distinct. When $\alpha \to 0$, all of the haplotypes will be assigned to a single functional allele ($\mathbf{M} = \mathbf{1}$), which is equivalent to a null model with no genetic effect.

To allow for additional flexibility, we place a prior distribution over the concentration parameter:

$$\alpha \sim \mathrm{Ga}(a_\alpha, b_\alpha), \tag{3.3}$$

where a_α and b_α are hyperparameters that control the shape and rate of the concentration parameter. We discuss prior elicitation for these hyperparameters in a later subsection.

3.4.2 Ewens's Sampling Formula and the CRP

The CRP is equivalently represented via Ewens's sampling formula as the distribution over partitions induced by functional mutations on random coalescent trees with the founder haplotypes at the leaves (Ewens, 1972; Kingman, 2006). The intuition for this interpretation is as follows. At a QTL, there is a tree that describes the relatedness of the founder haplotypes. At various points during the evolution of this locus, functional mutations that altered the phenotype occurred at a constant rate on the branches of the tree. These functional mutations were transmitted to the founder haplotypes at the leaves of the tree, partitioning the haplotypes into groups which carry the same set of functional mutations. This partition is the allelic series. Examples of allelic series induced by functional mutations on coalescent trees of haplotypes are given in **Figure 1.1**. If we assume that the tree relating the founder haplotypes is unknown, but that it is distributed according to the coalescent process, then the resulting distribution over partitions is the CRP (Berestycki, 2009).

More formally, Ewens's sampling formula describes the allelic series as a function of a tree and which branches of that tree are functionally mutated:

$$\mathbf{M} = f(T, \mathbf{b}),$$

where T denotes a tree with J leaves and $2J - 2$ branches, and $\mathbf{b} = (b_1, ..., b_{2J-2})^T$ is a length $2J - 2$ vector of indicators that denote whether or not a branch is mutated. The conditional probability density function of the allelic series given the tree and branch mutations is

$$\mathrm{p}(\mathbf{M}|T, \mathbf{b}) = \mathbf{1}_{\{T, \mathbf{b} \Rightarrow M\}},$$

where $\mathbf{1}_{\{T, \mathbf{b} \Rightarrow M\}}$ is an indicator function that takes value 1 when T and $\mathbf{b}$ imply $\mathbf{M}$ and 0 otherwise.

The tree T is an unknown random graph that is distributed according to the coalescent process with J leaves:

$$T \sim \mathrm{Coalescent}(J).$$

Coalescent trees are defined by sequential coalescent events that join lineages of the tree in random order, beginning with the leaves, as well as the times at which these coalescent events occur, which are exponentially distributed and depend on the number of lineages remaining prior to each coalescence (Kingman, 1982). For our purposes, it is sufficient to note that there is a probability distribution over trees, $\mathrm{p}(T)$, and that this distribution assumes equal relatedness of the haplotypes via the random order of coalescent events. We also note that each branch of the tree has a corresponding length, which is contained in the length $2J - 2$ vector $\boldsymbol{\ell}$ and described in coalescent units.

The mutation status of the branches $\mathbf{b}$ is an unknown vector of indicators. Assume that functional mutations occur on the branches of the tree as a Poisson process with constant rate 0.5α. Then the number of mutations on each branch is Poisson distributed with rate proportional to branch length, and the probability density function for $\mathbf{b}$, which indicates whether or not each branch is mutated, is

$$\mathrm{p}(\mathbf{b}|T, \alpha) = \prod_{m=1}^{2J-2} (e^{-0.5\alpha\ell_m})^{1-b_m} (1 - e^{-0.5\alpha\ell_m})^{b_m}.$$

This is similar to Azim Ansari and Didelot (2016), but with branch lengths scaled by 0.5 and in coalescent units. The concentration parameter α controls the functional mutation rate per half-unit of coalescent branch length. Note that moving forward, we will refer to α interchangeably as the concentration parameter (of the CRP) or the functional mutation rate (on the tree), depending on context. When $\alpha \to \infty$, the probability that each branch is mutated approaches 1, the tree is saturated with functional mutations, and all of the founder haplotypes are functionally distinct ($\mathbf{M} = \mathbf{I}$). When $\alpha \to 0$, the probability that each branch is mutated approaches 0, there are no functional mutations on the tree, and all of the founders are functionally identical ($\mathbf{M} = \mathbf{1}$).

The probability density function for the allelic series is thus

$$\mathrm{p}(\mathbf{M}|\alpha) = \int_T \left\{ \sum_{\mathbf{b}} \mathrm{p}(\mathbf{M}|T, \mathbf{b})\mathrm{p}(\mathbf{b}|T, \alpha) \right\} \mathrm{p}(T) \, \mathrm{d}T,$$

which involves identifying the allelic series implied by each combination of mutated branches on a tree, weighing by the probability of that combination, summing over all possible combinations, and then integrating over all possible coalescent trees. Remarkably, this is identical to the probability density function of the CRP described previously (Berestycki, 2009).

3.4.3 Tree-Informed CRP

In the previous subsection, we described how the CRP can be interpreted as a process of functional mutation on random coalescent trees, with these mutations partitioning the leaves (haplotypes) into the allelic series. This process is uninformative with respect to the relationship between individual haplotypes because it integrates over all possible coalescent trees, which are generated by randomly selecting lineages to coalesce, implying equal relatedness of the haplotypes. While the CRP does not encode information about the relatedness of the haplotypes, it does, however, allow control over the prior number of functional alleles, via the rate of functional mutation on the branches of the trees.

Within this framework, it is straightforward to define a distribution over the allelic series that encodes specific prior information about the relatedness of the haplotypes in the form of a tree. Conditional on a tree, the distribution over the allelic series reflects the relationships defined by the structure of the tree and the lengths of its branches. The tree topology reduces the space of possible partitions because many settings of $\mathbf{M}$ violate the relationships defined by T, making this information

highly informative. The branch lengths of T also provide information about the allelic series, as long branches are more likely to be functionally mutated than short branches. Consequently, haplotypes separated by longer branches are more likely to be functionally distinct than haplotypes separated by shorter branches. The functional mutation rate still controls the prior number of functional alleles, now in combination with the tree structure and branch lengths.

If the tree is known, the conditional probability density function of the allelic series is given by

$$p(\mathbf{M}|T, \alpha) = \sum_{\mathbf{b}} p(\mathbf{M}|T, \mathbf{b}) p(\mathbf{b}|T, \alpha),$$

which does not involve integrating over coalescent trees, but does involve (weighted) summation over all 2^{2J-2} possible configurations of $\mathbf{b}$. This approach is computationally intractable when the number of haplotypes is large, but provided J is small (e.g. $J = 8$, the case for many MPPs, but not $J = 15$, the case for the DSPR), it is possible to compute $p(\mathbf{M}|T, \alpha)$ directly. We focus on this approach and consider alternatives in the discussion.

Recall that the functional mutation rate (concentration parameter) α is an unknown variable with a prior distribution. To avoid the computational burden of calculating $p(\mathbf{M}|T, \alpha)$ for many settings of α during posterior inference, we instead marginalize over this variable and compute $p(\mathbf{M}|T)$ directly. The conditional probability density function is given by

$$p(\mathbf{M}|T) = \sum_{\mathbf{b}} \left\{ p(\mathbf{M}|T, \mathbf{b}) \int_{\alpha} p(\mathbf{b}|T, \alpha) p(\alpha) \, d\alpha \right\}.$$

In Appendix A, we show that the integral over α can be computed exactly when α has a gamma prior distribution.

Finally, to this point, we have assumed that the tree is known, but it may be unknown and inferred with uncertainty from a sequence alignment (Drummond et al., 2012). In this case, we are interested in the allelic series prior distribution conditional on the sequence alignment S,

$$p(\mathbf{M}|S) = \int_{T} p(\mathbf{M}|T) p(T|S) dT$$

This can be approximated by averaging $p(\mathbf{M}|T)$ over a sample of trees from $p(T|S)$.

3.5   Prior Distribution of Diplotype States

The diplotype state of each individual is an unobserved latent variable that is probabilistically inferred via haplotype reconstruction. To account for this uncertainty, the diplotype state of each individual is given a categorical prior distribution

$$\mathbf{d}_i \sim \mathrm{Cat}(\mathbf{p}_i),$$

where $\mathbf{p}_i$ is a $J + \binom{J}{2}$ length vector of prior diplotype probabilities for each individual.

3.6   Prior Distribution of Allele Effect Size

The variable ϕ controls the size of the allele effects relative to individual error. Importantly, it also controls the degree to which model complexity is penalized in Bayesian regression. An uninformative prior on the allele effects ($\phi \to \infty$) has the undesirable property of universally favoring a null model with no genetic effect. It is possible to specify a single value for ϕ, but this represents exact prior knowledge about the relative size of the allele effects, which is inappropriate if there is uncertainty about QTL effect size. Instead, it is preferable to specify a prior distribution over ϕ in order to accommodate this uncertainty.

We place a half-t prior distribution on the scaled standard deviation of the allele effects:

$$\phi \sim \mathrm{Half\text{-}t}_{2b_\phi},$$

where $2b_\phi$ is degrees of freedom of the half-t distribution. The half-t is a preferred prior choice for variance components in normally-distributed models due to its behavior at the boundary ($\phi = 0$) and its convenient representation as the product of two conditionally-conjugate latent variables (Gelman, 2006).

3.7 Prior Elicitation

The model is now fully elaborated, but we have not specified values for the prior hyperparameters. In this subsection, we guide the selection of these hyperparameters, discussing relevant considerations that may influence these selections.

3.7.1 Individual Error, Mean and Covariate Effects

We specify uninformative prior distributions for the size of the individual error ($\kappa, \phi \to 0$) and the intercept of the data ($\tau_\mu \to \infty$). If optional covariates are included, we also specify an uninformative prior distribution for the corresponding covariate effects ($\tau_\delta \to \infty$). These prior distributions are improper, but posterior inference is proper when these quantities are informed by the data.

3.7.2 Concentration Parameter / Functional Mutation Rate

The shape and rate parameters a_α and b_α control the prior distribution of the concentration parameter (functional mutation rate), which in turn controls the prior distribution over the number of functional alleles. An uninformative prior distribution for the concentration parameter is given by $a_\alpha, b_\alpha \to 0$. However, posterior learning about the concentration parameter depends only on the number of founder haplotypes J and the number of functional alleles K. For this reason, even if $\mathbf{M}$ is known, the concentration parameter is poorly informed when J is small. This necessitates a prior distribution that reflects reasonable prior expectations about the number of functional alleles.

We consider two alternatives for a_α and b_α. The first alternative is an exponential distribution which places 50% of the prior probability on the null model, given by $a_\alpha = 1$ and $b_\alpha \approx 2.33$ when $J = 8$ (termed "Exponential" in the next chapter). This prior distribution favors small numbers of functional alleles with low variance. The second alternative is a weakly-informative distribution which places moderate probability at the tails of the distribution [$\mathrm{p}(K = 1) = 0.05$ and $\mathrm{p}(K = J) = 0.01$, respectively], given by $a_\alpha \approx 2.30$ and $b_\alpha \approx 0.75$ when $J = 8$ (termed "Gamma" in the next chapter). This prior distribution expects an intermediate number of alleles with higher variance. We note that it is possible to calibrate other expectations similarly, for example regarding the frequency of biallelic contrasts, or the expected number of functional mutations on a tree, as in Azim Ansari and Didelot (2016). We emphasize that the reasonableness of these expectations is

specific to the number of founder haplotypes, the nature of the analysis (pre-/post-QTL detection), and other population- or trait-specific prior beliefs.

3.7.3 Coalescent Tree

Specifying a prior tree for the haplotypes is highly informative with respect to the allelic series. Our framework assumes that the phylogenetic tree is coalescent (with branches in coalescent units), satisfying assumptions of no recombination, selection or population structure. In the context of QTL mapping, the exact location of the causal sequence is often uncertain, making it difficult to satisfy the assumption of no recombination in particular. We discuss inferring trees in recombinant organisms in more detail in **Chapter 7**. We evaluate the consequences of tree misspecification in **Chapter 4**.

3.7.4 Diplotype States

We assume that the prior probability of the diplotype state of each individual $\mathbf{p}_i$ is inferred using established methods for haplotype reconstruction (Mott et al., 2000).

3.7.5 Relative Allele Effect Size

The half-t prior distribution on ϕ, the scaled standard deviation of allele effect size, is controlled by degrees of freedom $2b_\phi$. We set $b_\phi = 1$, which is the minimum value of b_ϕ that yields a monotonically decreasing prior distribution for the proportion of variance explained by the QTL, defined as

$$h^2_{QTL} = \frac{\phi^2\sigma^2}{\phi^2\sigma^2 + \sigma^2},$$
$$= \frac{\phi^2}{\phi^2 + 1}.$$

Larger values of b_ϕ reflect a stronger prior belief in "small" effect sizes and increase the degree of shrinkage in the allele effects.

3.8 Posterior Inference

We perform posterior inference via a partially-collapsed Gibbs sampler (van Dyk and Park, 2008; Park and Van Dyk, 2009). Conceptually, this involves four steps: *1)* updating the allelic series with

the effects and scale of the error integrated, or collapsed, from the model; *2)* jointly sampling the effects and scale of the error; *3)* updating the relative size of the allelic effects; and *4)* jointly updating the diplotype states. The effects and scale of the error are integrated from the model during the first step in order to avoid mismatching the dimension of β and the dimension of $\mathbf{M}$ when updating the allelic series. After updating the allelic series, the effects and scale of the error are reintroduced into the model in order to take advantage of a convenient latent variable sampling scheme for the relative size of the allele effects and in order to facilitate a joint update of the diplotype states. We will now briefly discuss each of these steps in more detail.

3.8.1 Updating the Allelic Series

We update the allele assignment of each haplotype individually, conditional on the allele assignment of the other haplotypes. In the case of the CRP, we must also update the concentration parameter. The conditional posteriors of the allelic series and the concentration parameter under the CRP are given by

$$\mathrm{p}(\mathbf{M}_j|\mathbf{y}, \mathbf{D}, \mathbf{M}_{-j}, \phi, \alpha) \propto \mathrm{p}(\mathbf{y}|\mathbf{D}, \mathbf{M}, \phi)\mathrm{p}(\mathbf{M}_j|\mathbf{M}_{-j}, \alpha),$$

$$\mathrm{p}(\alpha|\mathbf{y}, \mathbf{D}, \mathbf{M}, \phi) \propto \mathrm{p}(\mathbf{M}|\alpha)\mathrm{p}(\alpha).$$

The first equation is the product of the t-distributed likelihood and the categorical, exchangeable, conditional prior distribution of the CRP. The conditional posterior is calculated directly by evaluating the likelihood at all possible (conditional) settings of the allelic series (Neal, 2000). The conditional posterior of the concentration parameter depends only on the (number of alleles in the) allelic series, and there is a convenient, well-established latent variable approach for sampling from this posterior distribution (Escobar and West, 1995; Müller et al., 2015).

In the case of the tree-informed prior distribution, the concentration parameter has already been integrated from the allelic series prior. Thus, the conditional posterior of the allelic series under the tree-informed prior is given by

$$\mathrm{p}(\mathbf{M}_j|\mathbf{y}, \mathbf{D}, \mathbf{M}_{-j}, \phi, T) \propto \mathrm{p}(\mathbf{y}|\mathbf{D}, \mathbf{M}, \phi)\mathrm{p}(\mathbf{M}_j|\mathbf{M}_{-j}, T),$$

which is nearly identical to the previous equation. However, the conditional prior distribution of the allelic series given the tree, $p(\mathbf{M}_j|\mathbf{M}_{-j}, T)$, is not exchangeable and was not directly evaluated. Thus, we assume that the conditional prior distribution is proportional to the marginal prior distribution

$$p(\mathbf{M}_j|\mathbf{M}_{-j}, T) \propto p(\mathbf{M}|T),$$

treating this distribution as if it were exchangeable. At each iteration of the sampler, we randomize the order in which the haplotypes assignments are updated. This is to avoid bias introduced by ordered updates of nonexchangeable variables, as observed in (Wallach et al., 2008). We have not observed issues with mixing using this approach, suggesting that this violation of exchangeability is mild.

3.8.2 Sampling the Effects and Error

The conditional posterior of the intercept, allele effects, covariate effects and error scale is given by

$$p(\boldsymbol{\theta}, \sigma|\mathbf{y}, \mathbf{D}, \mathbf{M}, \phi, [\alpha \vee T]) \propto p(\mathbf{y}|\mathbf{D}, \mathbf{M}, \boldsymbol{\theta}, \sigma)p(\boldsymbol{\theta}|\sigma, \phi)p(\sigma),$$

which is the product of the normally-distributed likelihood and a conjugate normal-gamma prior distribution, yielding a normal-gamma conditional posterior distribution, as in Servin and Stephens (2007). We use the notation $[\alpha \vee T]$ to indicate that the distribution is conditional on either α or T, depending on which allelic series prior distribution is used.

3.8.3 Updating the Relative Size of the Allele Effects

The conditional posterior of the relative size of the allele effects is give by

$$p(\phi|\mathbf{y}, \mathbf{D}, \mathbf{M}, \boldsymbol{\theta}, \sigma, [\alpha \vee T]) \propto p(\boldsymbol{\beta}|\sigma, \phi)p(\phi),$$

which is the product of a normal distribution and a half-t prior distribution for the standard deviation, which is not conjugate. However, the half-t prior distribution can be re-expressed as the product of two latent variables: one the square root of an inverse-gamma-distributed variable and the other the absolute value of a normally-distributed variable. Respectively, these variables are conditionally

conjugate to the prior distribution of β and the likelihood function, allowing for straightforward updating of these latent variables (Gelman, 2006).

3.8.4 Updating the Diplotype States

The conditional posterior of the diplotype states is given by

$$p(\mathbf{D}|\mathbf{y}, \mathbf{M}, \boldsymbol{\theta}, \sigma, \phi, [\alpha \vee T]) \propto p(\mathbf{y}|\mathbf{D}, \mathbf{M}, \boldsymbol{\theta}, \sigma)p(\mathbf{D}),$$

which is the product of the normally-distributed likelihood and a categorical prior distribution. The diplotype states are conditionally independent, and the joint conditional posterior is computed directly by evaluating the likelihood of each individual observation over all possible possible diplotype states.

3.9 Marginal Likelihood and Hypothesis Testing

The marginal likelihood is useful for comparing alternative hypotheses about the data. For example, two competing hypothesis can be evaluated by computing a BF, which is the ratio of the marginal likelihoods under the assumptions of each hypothesis (Kass and Raftery, 1995). Computing the marginal likelihood directly via integration is difficult because it involves exploring the space of all combinations of prior values for the model variables. Instead, we use the output of the Gibbs sampler to construct an estimate of the marginal likelihood (Chib, 1995).

From Bayes theorem, the natural logarithm of the marginal likelihood can be expressed as

$$\ln p(\mathbf{y}) = \ln p(\mathbf{y}|\mathbf{D}, \mathbf{M}, \boldsymbol{\theta}, \sigma) + \ln p(\mathbf{D}, \mathbf{M}, \boldsymbol{\theta}, \sigma) -$$
$$\ln p(\mathbf{D}, \mathbf{M}, \boldsymbol{\theta}, \sigma|\mathbf{y}),$$

which is true at every point $(\mathbf{D}^*, \mathbf{M}^*, \boldsymbol{\theta}^*, \sigma^*)$. Obtaining an estimate of the marginal likelihood involves factoring the joint posterior (and prior) into quantities that can either be calculated directly or can be well-approximated from the output of the Gibbs sampler at the specified point.

The estimate of the joint posterior is given by

$$\ln\hat{p}(\mathbf{D}, \mathbf{M}, \boldsymbol{\theta}, \sigma | \mathbf{y}) = \ln\hat{p}(\mathbf{M}|\mathbf{y}) + \ln\hat{p}(\boldsymbol{\theta}, \sigma | \mathbf{y}, \mathbf{M}) +$$
$$\ln p(\mathbf{D}|\mathbf{y}, \mathbf{M}, \boldsymbol{\theta}, \sigma).$$

The first term, $\ln\hat{p}(\mathbf{M}|\mathbf{y})$, is the marginal posterior density of the allelic series. Samples from the marginal posterior are given by the Gibbs sampler, and the density estimate at $\mathbf{M}^*$ is the proportion of posterior samples equal to this value. This estimate is most accurate at the *maximum a posteriori* (MAP) of the allelic series, which is used to select $\mathbf{M}^*$. The second term of our estimate, $\ln\hat{p}(\boldsymbol{\theta}, \sigma | \mathbf{y}, \mathbf{M})$, is the joint posterior of the coefficients and the scale of the error variance, which is is not given directly by the Gibbs output. Thus, after selecting $\mathbf{M}^*$, we fix the allelic series in the Gibbs sampler and resume iterating, obtaining samples from this conditional posterior. The conditional posterior yields sufficient statistics which are then used to obtain an accurate Rao-Blackwellized estimate of the joint conditional posterior density.(Blackwell, 2007) These sufficient statistics are also used to select high-probability values of $\boldsymbol{\theta}^*$ and σ^* at which to evaluate the estimate. Finally, the third term of our estimate, $\ln p(\mathbf{D}|\mathbf{y}, \mathbf{M}, \boldsymbol{\theta}, \sigma)$, is the full conditional posterior of the diplotype states. This is calculated directly, and $\mathbf{D}^*$ is selected as the marginal MAP of the diplotype states.

Factoring the joint prior is straightforward, as many of the priors are independent by construction. The joint prior distribution is given by

$$\ln p(\mathbf{D}, \mathbf{M}, \boldsymbol{\theta}, \sigma) = \ln p(\mathbf{D}) + \ln p(\mathbf{M}) +$$
$$\ln p(\boldsymbol{\theta}|\sigma) + \ln p(\sigma).$$

Note that calculating $\ln p(\mathbf{M})$, or $\ln p(\mathbf{M}|T)$ when using a tree, involves integrating over α. This has already been calculated directly when using the tree, but it must be approximated when using the CRP. Calculating $\ln p(\boldsymbol{\theta}|\sigma)$ involves integrating over ϕ, which has a closed form involving a confluent hypergeometric function (Abramowitz and Stegun, 1972). Finally, when using an improper prior distribution for μ, $\boldsymbol{\delta}$ and σ, the final two terms in the prior are equal to zero, and as such are only are evaluated up to a constant. Comparisons of these marginal likelihood via BFs are valid provided $\boldsymbol{\delta}$ is

unchanged, as these constants cancel when evaluating the ratio of likelihoods (Servin and Stephens, 2007).

3.10   Availability of Software

These methods are implemented in the Tree-Based Inference of Multiallelism via Bayesian Regression ('TIMBR') R package, available on GitHub at `https://github.com/wcrouse/TIMBR`.

3.11   Appendix A - Marginalizing the Tree-Informed Prior

In this appendix, we show that the prior density of the mutation status of the branches $\mathbf{b}$ can be marginalized over the concentration parameter α if the concentration parameter has a gamma prior distribution. This is useful for computing the tree-informed allelic series prior distribution. Our approach includes considerable bookkeeping of signs and coefficients, so we demonstrate using a minimal example, $\mathbf{b} = (0, 0, 1, 1, 1, 0)$ for $J = 4$.

We begin by expanding $\mathrm{p}(\mathbf{b}|T, \alpha)$:

$$
\begin{aligned}
\mathrm{p}(\mathbf{b}|T, \alpha) =& \prod_{m=1}^{2J-2} (e^{-0.5\alpha\ell_m})^{1-b_m}(1 - e^{-0.5\alpha\ell_m})^{b_m}, \\
=& e^{-0.5\alpha\ell_1}e^{-0.5\alpha\ell_2}(1 - e^{-0.5\alpha\ell_3})(1 - e^{-0.5\alpha\ell_4})(1 - e^{-0.5\alpha\ell_5})e^{-0.5\alpha\ell_6}, \\
=& e^{-0.5\alpha(\ell_1+\ell_2+\ell_6)}(1 - e^{-0.5\alpha\ell_3})(1 - e^{-0.5\alpha\ell_4})(1 - e^{-0.5\alpha\ell_5}), \\
=& e^{-0.5\alpha(\ell_1+\ell_2+\ell_6)}\big(1 - e^{-0.5\alpha\ell_5} - e^{-0.5\alpha\ell_4} + e^{-0.5\alpha(\ell_4+\ell_5)} - e^{-0.5\alpha\ell_3} \\
& + e^{-0.5\alpha(\ell_3+\ell_5)} + e^{-0.5\alpha(\ell_3+\ell_4)} - e^{-0.5\alpha(\ell_3+\ell_4+\ell_5)}\big), \\
=& e^{-0.5\alpha(\ell_1+\ell_2+\ell_6)} - e^{-0.5\alpha(\ell_1+\ell_2+\ell_5+\ell_6)} - e^{-0.5\alpha(\ell_1+\ell_2+\ell_4+\ell_6)} \\
& + e^{-0.5\alpha(\ell_1+\ell_2+\ell_4+\ell_5+\ell_6)} - e^{-0.5\alpha(\ell_1+\ell_2+\ell_3+\ell_6)} \\
& + e^{-0.5\alpha(\ell_1+\ell_2+\ell_3+\ell_5+\ell_6)} + e^{-0.5\alpha(\ell_1+\ell_2+\ell_3+\ell_4+\ell_6)} - e^{-0.5\alpha(\ell_1+\ell_2+\ell_3+\ell_4+\ell_5+\ell_6)}.
\end{aligned}
$$

Each term is an exponentiated sum of branch lengths. All terms includes the branch lengths of the branches that are not mutated ($\{1, 2, 6\}$). The eight terms correspond to the eight possible subsets of the set of mutated branches ($\{2, 3, 4\}$), whose lengths are either included or excluded from the sum.

If the sum includes an even number of mutated branches, the sign of the term is positive, and if the sum includes an odd number, the sign is negative.

We are interested in

$$p(\mathbf{b}|T) = \int_\alpha p(\mathbf{b}|T, \alpha) p(\alpha) \, d\alpha.$$

From the previous expansion, we know that this is an integral of a sum, which allows us to evaluate the integral separately for each term:

$$p(\mathbf{b}|T) = \int_\alpha e^{-0.5\alpha(\ell_1+\ell_2+\ell_6)} * p(\alpha) \, d\alpha - \int_\alpha e^{-0.5\alpha(\ell_1+\ell_2+\ell_5+\ell_6)} * p(\alpha) \, d\alpha - \ldots.$$

The prior distribution for the concentration parameter is

$$\alpha \sim \mathrm{Ga}(a_\alpha, b_\alpha),$$

which has probability density function (omitting subscripts for clarity)

$$p(\alpha) = b^a \Gamma(a)^{-1} \alpha^{a-1} e^{-b\alpha}.$$

Focusing only on the first term, we have

$$
\begin{aligned}
p(\mathbf{b}|T) &= \int_\alpha e^{-0.5\alpha(\ell_1+\ell_2+\ell_6)} * p(\alpha) \, d\alpha - \ldots \\
&= \int_\alpha e^{-0.5\alpha(\ell_1+\ell_2+\ell_6)} b^a \Gamma(a)^{-1} \alpha^{a-1} e^{-b\alpha} \, d\alpha - \ldots \\
&= b^a \Gamma(a)^{-1} \int_\alpha \alpha^{a-1} e^{-(b+0.5[\ell_1+\ell_2+\ell_6])\alpha} \, d\alpha - \ldots
\end{aligned}
$$

The quantity within the integral is the kernel of a gamma distribution with shape a and rate $b + 0.5(\ell_1 + \ell_2 + \ell_6)$, and the integral is equal to the inverse of its normalizing constant:

$$\mathrm{p}(\mathbf{b}|T) = b^a \Gamma(a)^{-1} \Gamma(a)(b + 0.5[\ell_1 + \ell_2 + \ell_6])^{-a} - \ldots$$

$$= b^a(b + 0.5[\ell_1 + \ell_2 + \ell_6])^{-a} - \ldots$$

$$= \left(\frac{b}{b + 0.5(\ell_1 + \ell_2 + \ell_6)} \right)^a - \ldots$$

The other terms in the expression are solved similarly. This provides a closed form expression for $\mathrm{p}(\mathbf{b}|T)$, which in turn is used to compute $\mathrm{p}(\mathbf{M}|T)$.

CHAPTER 4

Performance in Single-Locus Simulations

4.1 Overview

In this chapter, we investigate various properties of the haplotype- and allele-based association approaches via simulation. These single-locus simulations evaluate the performance of various alternative prior distributions for the allelic series at simulated QTL with varying numbers of functional alleles. The alternative priors are evaluated with respect to their accuracy in identifying the underlying allelic series, their error in estimating haplotype effects, and their relative statistical signal. These three criteria inform the utility of the allele-based association approach in inferring genetic architecture, predicting phenotypes, and identifying QTL. This chapter focuses on two main topics: prior selection for the allele-based model in the absence of additional phylogenetic information, and the utility of including that additional prior information, with varying levels of accuracy, as a coalescent tree.

In the first set of simulations, we consider three alternative prior distributions for the allelic series:

- a uniform prior which assumes that all configurations of the allelic-series are *a priori* equally likely (termed the "Uniform" model):

$$p(\mathbf{M}) \propto 1.$$

This is implemented as a uniform process prior (Wallach et al., 2008).

- the CRP-based model with a weakly informative gamma prior distribution on the concentration parameter (the "Gamma" model):

$$\mathbf{M}|\alpha \sim \mathrm{CRP}(\alpha)$$

$$\alpha \sim \mathrm{Ga}(a_\alpha \approx 2.30, b_\alpha \approx 0.75),$$

In the common case of eight possible haplotypes (the case considered in these simulations), this prior distribution corresponds to a 5% probability of the null model with one functional allele and a 1% probability of the full model with all eight haplotypes functionally distinct,

- and the CRP-based model with a more informative and less permissive exponential prior distribution on the concentration parameter (the "Exponential" model), parameterized in terms of the gamma for consistency:

$$\mathbf{M}|\alpha \sim \mathrm{CRP}(\alpha)$$

$$\alpha \sim \mathrm{Ga}(a_\alpha = 1, b_\alpha \approx 2.33),$$

With eight possible haplotypes, this prior distribution corresponds to a 50% probability of one functional allele, and monotonically favors smaller numbers of functional alleles *a priori*.

These alternative prior distributions for the allelic series were introduced in the preceding chapter, and their implied distributions over functional alleles are depicted in **Figure 4.1**. This first set of simulations provides guidance on prior elicitation regarding the allelic series and the number of functional alleles.

Given this guidance, we then focus on the Exponential model while investigating the second main topic: the utility of introducing additional prior information on the phylogenetic relatedness of the founder haplotypes. Specifically, we consider four alternatives:

Prior Distribution of Alleles
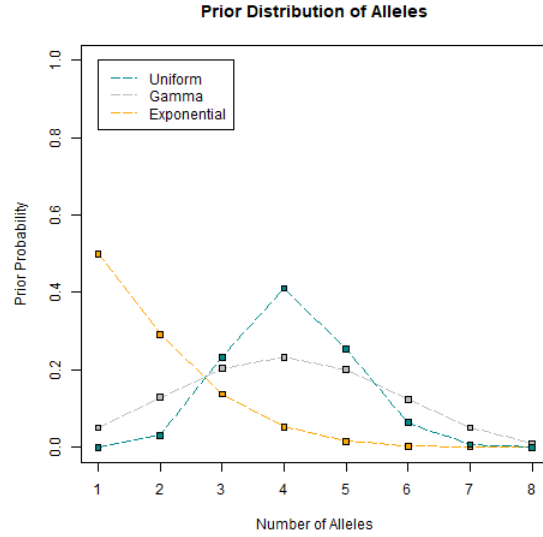
Figure 4.1: Prior distribution of number of functional alleles for alternative prior distributions. The Uniform model places high prior weight on an intermediate number of functional alleles, as these permit many allelic configurations. The Gamma model is less informative with respect to the number of alleles, with fatter tails than the Uniform. The Exponential model favors smaller numbers of functional alleles relative to the other models.

.

- the CRP-based "Exponential" model described above, which implicitly assumes that the phylogenetic tree of the haplotypes is unknown and distributed according to coalescent theory:

$$\mathrm{p}(\mathbf{M}|\alpha) = \int_T \mathrm{p}(\mathbf{M}|T, \alpha)\mathrm{p}(T)\, \mathrm{d}T$$

$$T \sim \mathrm{Coalescent}(J = 8)$$

$$\alpha \sim \mathrm{Ga}(a_\alpha = 1, b_\alpha \approx 2.33),$$

- the Exponential model with the true tree known *a priori* (the "Tree" model), in which latent combinations of functional mutations on branches of the tree provide additional information about the prior distribution of the allelic series:

$$\mathrm{p}(\mathbf{M}|T, \alpha) = \sum_{\mathbf{b}} \mathrm{p}(\mathbf{M}|T, \mathbf{b})\mathrm{p}(\mathbf{b}|T, \alpha),$$

87

- the Tree model with the tree information partially misspecified as T^* (the "Miss" model):

$$p(\mathbf{M}|T^*, \alpha) = \sum_{\mathbf{b}} p(\mathbf{M}|T^*, \mathbf{b})p(\mathbf{b}|T^*, \alpha).$$

This assumes imperfect but correlated information about the phylogenetic tree of the founder haplotypes. This is described in more detail in the next section,

- and the Tree model with the tree information completely misspecified as T' (the "Wrong" model):

$$p(\mathbf{M}|T', \alpha) = \sum_{\mathbf{b}} p(\mathbf{M}|T', \mathbf{b})p(\mathbf{b}|T', \alpha).$$

This assumes that a tree is supplied but that it is independent of the true tree, thus providing incorrect phylogenetic information. This is also described in more detail in the next section.

Comparing these alternatives highlight the relative benefits of including prior tree information, and the sensitivity of those benefits to varying levels of incorrectness. Finally, in some cases, we also evaluate these scenarios against two additional alternatives:

- the standard haplotype-based approach where all haplotypes are functionally distinct (the "Full" model):

$$\mathbf{M} = \mathbf{I},$$

- an oracle approach where the true allelic series is known *a priori* (the "Known" model).

These comparisons provide insight into the performance of the allele-based association approach relative to the standard haplotype-based approach and provide an upper bound on its potential.

This chapter continues with a description of the simulation procedure, followed by definitions of the evaluation metrics. We then present results from the two sets of simulations as described above. Finally, we discuss the implications of these results and compare them with similar studies (Jannink and Wu, 2003; Azim Ansari and Didelot, 2016; King et al., 2014).

4.2 Simulation Procedure

We iteratively simulate QTL for an eight-haplotype MPP with known but varied coalescent phylogeny of the haplotypes at the locus. Throughout the simulations, we assume a fixed experiment size of $N = 400$ individuals, balanced with respect to haplotypes and with known homozygous diplotype states, $\mathbf{D}$. Rather than vary experiment size, we instead vary QTL effect size, as measured by the proportion of total phenotype variance explained by the QTL, h_{QTL}^2, as previously described in **Chapter 2**. We anticipate that the properties explored during these simulations depend primarily on power (a function of experiment size, haplotype balance, and QTL effect size), rather than experiment size *per se*, and thus hold both experiment size and haplotype balance fixed for simplicity. We also assume that the population does not have structure in genetic background. Thus, with non-replicate observations, any variance attributable to strain effects, h_{strain}^2, is indistinguishable from individual-level error and can be ignored. Finally, we consider only additive QTL effects because the diplotypes are assumed to be homozygous, and thus dominance effects are not revealed.

Subject to these assumptions, we performed the following simulation procedure:

- Sample a coalescent tree T to describe the local phylogenetic relationship of the eight founder haplotypes:

$$T \sim \mathrm{Coalescent}(J = 8),$$

- For a given functional mutation rate α, calculate the distribution of allelic series implied by the tree:

$$\mathrm{p}(\mathbf{M}|T, \alpha) = \sum_{\mathbf{b}} \mathrm{p}(\mathbf{M}|T, \mathbf{b})\mathrm{p}(\mathbf{b}|T, \alpha),$$

- For a given number of functional alleles K, sample an allelic series conditional on T that satisfies K:

$$\mathrm{p}(\mathbf{M}|T, \alpha, K) \propto \mathrm{p}(K|T, \alpha, \mathbf{M})\mathrm{p}(\mathbf{M}|T, \alpha),$$

- For a given QTL effect size h_{QTL}^2, zero-center and scale K equally-spaced allele effects $\boldsymbol{\beta}_{\mathrm{alle}}$ to satisfy $\mathrm{mean}(\boldsymbol{\beta}_{\mathrm{alle}}) = 0$ and $\mathrm{var}(\boldsymbol{\beta}_{\mathrm{alle}}) = h_{QTL}^2$

- Sample a vector of N individual errors from

$$\boldsymbol{\epsilon} \sim \mathrm{N}(0, 1),$$

and zero-center and scale to satisfy $\mathrm{mean}(\boldsymbol{\epsilon}) = 0$ and $\mathrm{var}(\boldsymbol{\epsilon}) = 1 - h_{QTL}^2$,

- Calculate the simulated phenotypes:

$$\mathbf{y} = \mathbf{DAM}\boldsymbol{\beta}_{\mathrm{alle}} + \boldsymbol{\epsilon}.$$

Rather than sample normally-distributed allele effects, we elected to use uniformly-spaced allele effects, similar to (King et al., 2014). As the number of functional alleles increases, when the allele effects are normally-distributed, the minimum distance between any two allele effects becomes increasingly small, making it harder to detect these effects. Our uniformly-spaced approach eliminates the possibility of arbitrarily small and undetectable differences between allele effects, with the consequence that our results potentially overstate the power to detect larger numbers of functional alleles if allele effects are in fact normally-distributed.

To explore the impact of incorrect tree information in our simulations, we also performed two additional steps:

- Sample a completely misspecified tree T' from the coalescent:

$$T' \sim \mathrm{Coalescent}(J = 8).$$

In this case, T' does not contain any information about the true tree T. This is used in the "Wrong" tree scenario.

- Sample a partially misspecified tree T^* using the procedure outlined in Azim Ansari and Didelot (2016). This involves resampling the patristic distances between each pair of haplotypes from

$$d \sim \mathrm{Unif}(0.1 * d, 0.9 * d)$$

and generating a new tree T^* based on the unweighted pair group method with arithmetic mean. This procedure results in trees with generally smaller total branch lengths, so we took the additional step of resampling the total branch length of T^* from the coalescent. This procedure generates a misspecified tree T^* that is based on information from the true tree T, but which potentially has rearrangements of the tree structure and different branch lengths. This is used in the "Miss" tree scenario.

We elected to use the patristic distance resampling method to generate T^* because of its use in Azim Ansari and Didelot (2016), which is similar to our approach. There are alternative methods for "scrambling" trees, such as the subtree-prune-regraft procedure, in which part of the tree is removed and reattached in another position. However, applying the this procedure is nontrivial in the context of trees with meaningful branch lengths (Song, 2006), as it requires difficult decisions about selecting where to prune and how to handle height restrictions during regraft.

We performed 1000 simulations for each combination of the following parameter settings:

- Number of functional alleles K: [1-8]

- QTL effect size h^2_{QTL}: [10%, 50%]

- Concentration parameter α: [0.1, 1, 10]

In all cases, we show results stratified by the true number of functional alleles K. In the first section of simulation results, we do not consider tree information, and thus the results are independent of the true α conditional on K. In the second section of simulation results, we do consider tree information, and the results are no longer independent of α conditional on K. However, we find that our results are relatively insensitive to α, and for this reason, we only present results from the $\alpha = 1$ case. Note that the relative benefit of tree information (conditional on K) is somewhat higher when $\alpha = 10$ and lower when $\alpha = 0.1$.

4.3 Evaluation Metrics

We evaluate the performance of our alternative prior distributions across three general categories: their accuracy in identifying the underlying allelic series, their error in estimating haplotype effects,

and their relative statistical signal. Within these categories, we define several specific metrics by which the alternative prior distributions are evaluated. With respect to accuracy in identifying the allelic series, these include:

- whether or not the *maximum a posteriori* (MAP) allelic series is the correct allelic series ("0-1 Accuracy")

- the posterior mass on the correct allelic series, $p(\mathbf{M}|\mathbf{y})$ ("Posterior Certainty")

- the posterior expectation on the number of functional alleles, $E(K|\mathbf{y})$ ("Expected Number of Alleles")

- the posterior probability of over- or underestimating the correct number of alleles.

For error in estimating haplotype effects, we evaluate using

- the mean squared error (MSE) of the posterior haplotype effects, $p(\boldsymbol{\beta}_{\mathrm{hap}}|\mathbf{y})$, relative to the true effects. This is averaged over posterior samples of the allelic series and allele effects, which together determine the haplotype effects, $\boldsymbol{\beta}_{\mathrm{hap}} = \mathbf{M}\boldsymbol{\beta}_{\mathrm{alle}}$.

Finally, for relative statistical signal, we consider

- the natural log Bayes Factor (lnBF), which uses the ratio of marginal likelihoods to compare competing hypothesis. When considering the allele-based Uniform, Gamma and Exponential alternatives, we compare against the Full haplotype-based model or a null model of no effect. When considering Tree information, we compare relative to the Exponential approach, specifically highlighting the addition of tree information within the allele-based framework.

When discussing lnBFs, use refer to interpretation thresholds given by (Kass and Raftery, 1995). In the positive direction, these ranges correspond to "Positive" from 1-3, "Strong" from 3-5, and "Very Strong" above 5. The interpretation is similar in the negative direction, but with the directionality reversed. All these metrics are presented with confidence intervals. For "0-1 Accuracy", we use Jeffreys interval (Brown et al., 2001) for a binomial proportion. We use a t-distribution for all other metrics.

4.4 Results - Alternative Prior Distributions

In this section, we evaluate the Uniform, Gamma, and Exponential alternatives for encoding prior information about the allelic series, across the three categories described in the previous section, using their corresponding metrics.

4.4.1 Accuracy of Allelic Series

For all alternative prior distributions, the accuracy of allelic series inference depends on the true number of functional alleles. **Figure 4.2** shows the 0-1 accuracy of the MAP allelic series under the Uniform, Gamma and Exponential alternatives, for a varying number of true functional alleles. In the relatively lower power scenario (10% variance explained by QTL), the ability to detect multiallelic series is low, but the two CRP-based models, Exponential and Gamma, have high accuracy when the QTL has one functional allele (the null model) or when it is biallelic. In the high power scenario (50% variance explained by the QTL), the CRP-based models have reasonably high accuracy (approximately 80%) up to four functional alleles, with the Exponential outperforming the Gamma through this range. The more-diffuse Gamma prior maintains some limited accuracy through eight alleles, outperforming the Exponential, although accuracy for highly-multiallelic series is generally low even when power is high. In both scenarios, the Uniform approach is worse than the Exponential and Gamma, except when the number of alleles is intermediate.

Figure 4.3 is similar to the previous figure but shows the posterior certainty of the correct allelic series rather than the accuracy of the MAP allelic series. The Gamma and Uniform priors have relatively low certainty across both power scenarios and for all true numbers of functional alleles. In contrast, relative to Gamma, the Exponential prior is decisive when the true number of alleles is low, but at the expensive of accuracy when the true number of alleles is high. Notably, in the high power scenario, the Gamma prior has reduced certainty when the QTL is null relative to biallelic, suggesting it has a tendency to overestimate under the null.

Figure 4.4 is also similar to the previous figures but shows the posterior expectation of number of alleles. In the low power scenario, the expectations for both Gamma and Uniform are insensitive to the true number of alleles, consistently reporting an intermediate number of functional alleles, while the Exponential is accurate when the true number of alleles is one or two, but then reports
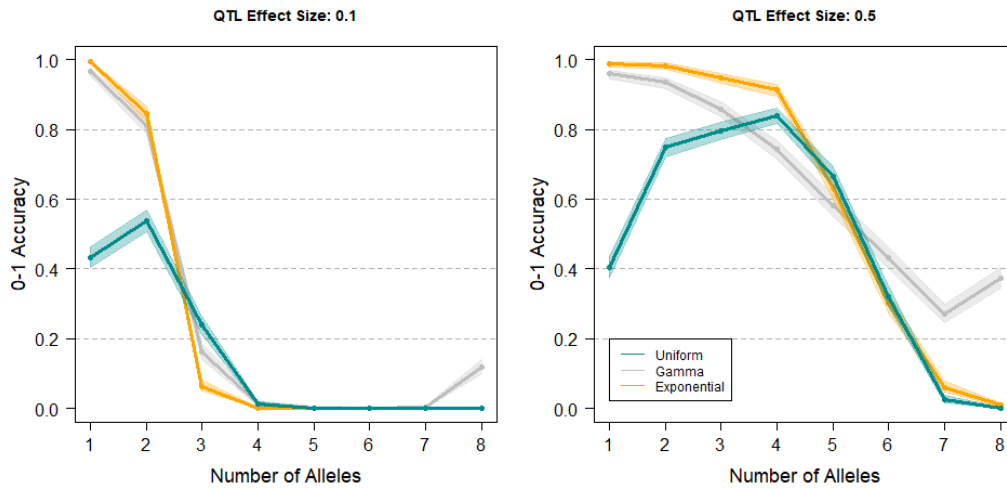
Figure 4.2: 0-1 accuracy of posterior allelic series inference for alternative prior distributions, in low and high power scenarios (respectively, $h^2_{QTL} = [0.1, 0.5]$), for varying numbers of true functional alleles.
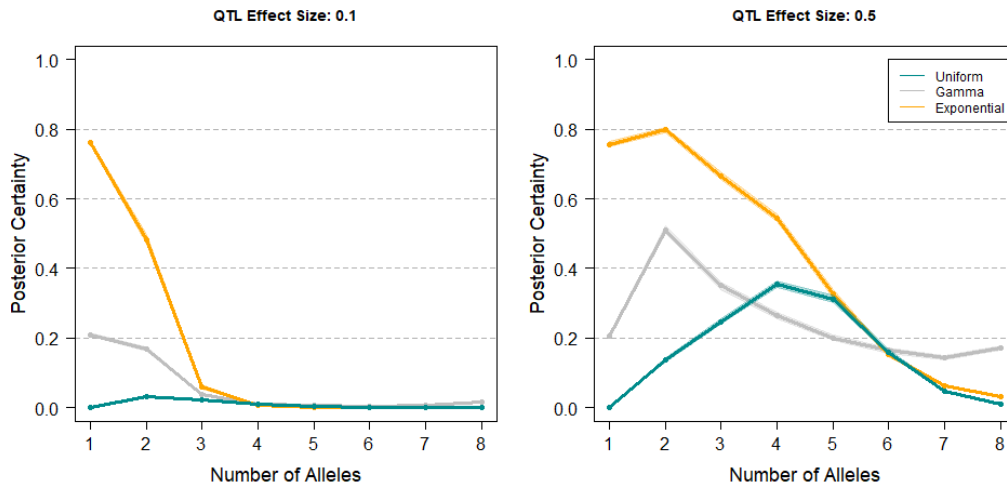
Figure 4.3: Posterior certainty of allelic series inference for alternative prior distributions, in low and high power scenarios (respectively, $h^2_{QTL} = [0.1, 0.5]$), for varying numbers of true functional alleles.
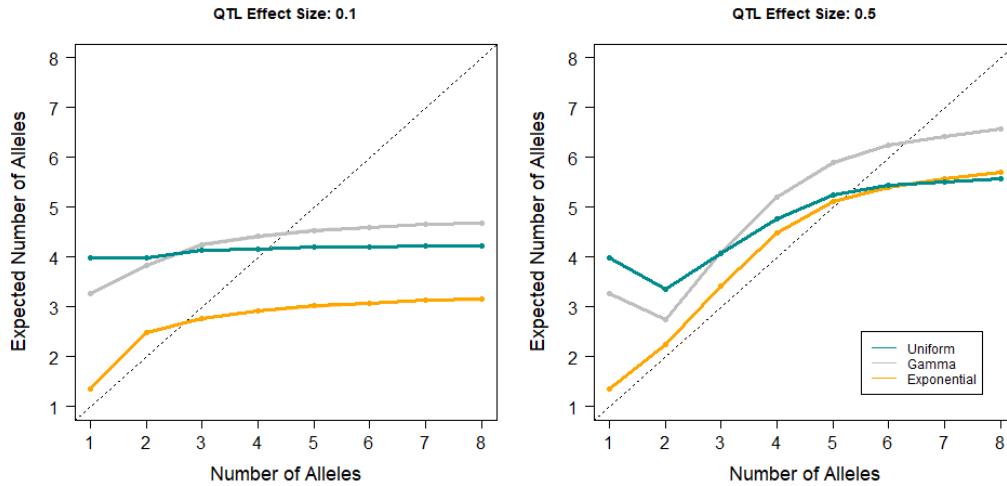
Figure 4.4: Posterior expectation of number of alleles for alternative prior distributions, in low and high power scenarios (respectively, $h^2_{QTL} = [0.1, 0.5]$), for varying numbers of true functional alleles.

approximately three alleles when the true number of alleles is higher higher. In the high power scenario, all the alternatives are more sensitive to the true number of alleles, with the exponential in particular approaching the correct expectation through the five alleles range. Notably, both the Exponential and Uniform show an exaggerated tendency to overestimate when the QTL is null. We hypothesize that this is due to their relatively fat tails with respect to prior number of functional alleles, and that this acts in combination with a prior for QTL effect size that can accommodate small effects. When the true number of alleles is one, the QTL effect is necessarily zero, it then becomes "easier" for these permissive allelic series priors to estimate many effects, each of very small size.

Figure 4.5 shows the posterior probability of under- and overestimating the true number of alleles. These results show that all the alternative prior distributions tend to underestimate the number of functional alleles when the true number is large, even in the high power scenario. In the low power scenario, the Exponential is more prone to underestimation than the other two alternatives. However, the Exponential is also considerably less likely to overestimate the true number of alleles across both power scenarios. Thus, the Exponential is more conservative than the Gamma or Uniform in reporting that a QTL is multiallelic.
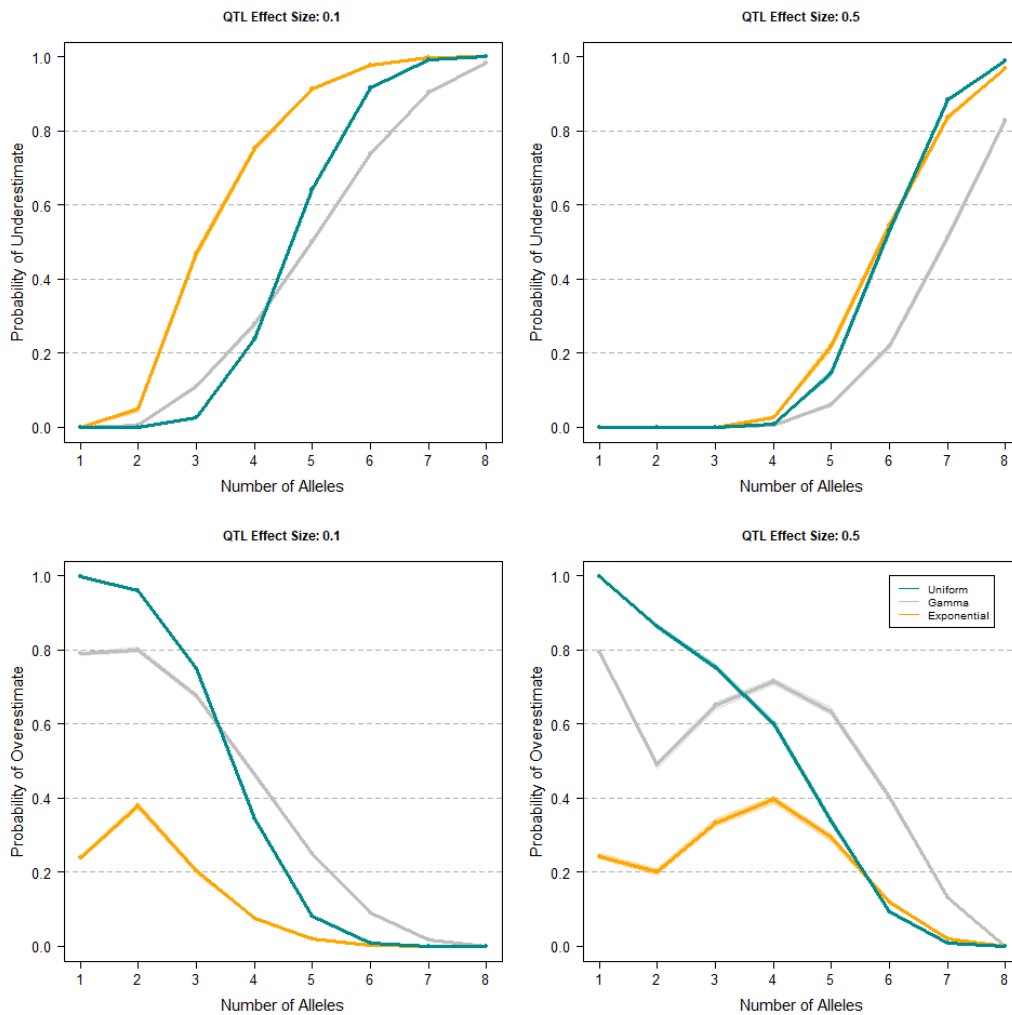
Figure 4.5: Posterior probability of under- and overestimating the number of alleles for alternative prior distributions, in low and high power scenarios (respectively, $h^2_{QTL} = [0.1, 0.5]$), for varying numbers of true functional alleles.
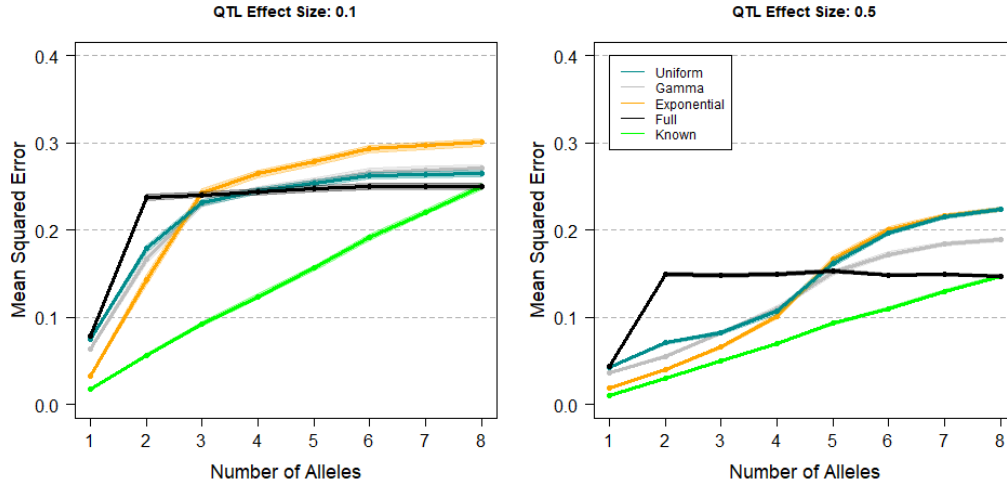
Figure 4.6: MSE of haplotype effect estimates for alternative prior distributions, in low and high power scenarios (respectively, $h^2_{QTL} = [0.1, 0.5]$), for varying numbers of true functional alleles.

4.4.2 Error of Haplotype Effect Estimation

Figure 4.6 shows the MSE of haplotype effect estimates for the alternative allelic series prior distributions, now additionally compared with the Full and Known models. As expected, for all alternatives, MSE is lower in the high power scenario than in the low power scenario. The Full model has near-constant MSE when the true number of alleles is greater than one, but when there is only one functional allele (the null model), the Full model has lower MSE, shrinking near-zero effects. In both the high and low power scenarios, the allelic series models outperform the Full model when the true number of alleles is low. This trend continues into intermediate numbers of alleles when power is high. However, when the true number of alleles is high, the Full model is better than the allelic series models. As with accuracy, the Exponential is relatively better than Gamma and Uniform when the true number of alleles is low, and it is relatively worse when the true number of is high. The Known model establishes a lower-bound on MSE in the case where the allelic series is known *a priori*. The allelic series models approach this lower bound in the high power scenario, when the true number of alleles is low, which is consistent with their high accuracy in this case.

4.4.3 Statistical Signal

Figure 4.7 shows the lnBF for the alternative allelic series prior distributions relative to the Full model, provided that the true number of functional alleles is greater than one. In the low power scenario, when allelic series inference is highly uncertain, the lnBFs are near zero for all alternative priors, indicating that there is little additional statistical signal provided by the allelic series models. In the high power scenario, the lnBFs strongly support the allelic series models relative to the Full model when there are two or three true functional alleles. However, as the true number of functional alleles increases, the allelic series models fare worse than the full model, with negative evidence for Gamma, very strong negative evidence for Exponential, and negative to strongly negative evidence for Exponential.

Figure 4.8 shows the lnBF for the alternative allelic series prior distributions and the Full haplotype-based model, relative to a null model of no genetic effect, when the QTL is in fact null (one functional allele). These results were pooled across power scenarios because QTL effect size is irrelevant in this case. As expected, we find that there is generally negative evidence for all of these alternative prior distributions relative to the null. The Full lnBFs are the most negative, but with a long right tail that extends into a range of weakly positive support in its 99% quantile. The Exponential lnBFs are more tightly distributed near zero and have a more restrained right tail. The Gamma and Uniform lnBFs are intermediate to these two alternatives, but with right tails that are more similar to the Full and longer than the Exponential.

Finally, to establish an upper bound on the statistical signal present in the allelic series, **Figure 4.9** shows the lnBF for the Known model relative to the Full model, provided that the true number of functional alleles is greater than one and less than eight. There is increased signal from the allelic series in all cases, and this increase is greatest when the true number of functional alleles is small. In the low power scenario, there is strong to positive support in favor of the Known model, and in the high power scenario this support is often decisive. These results indicate that there is substantial statistical signal in the allelic series and that this signal increases with power.
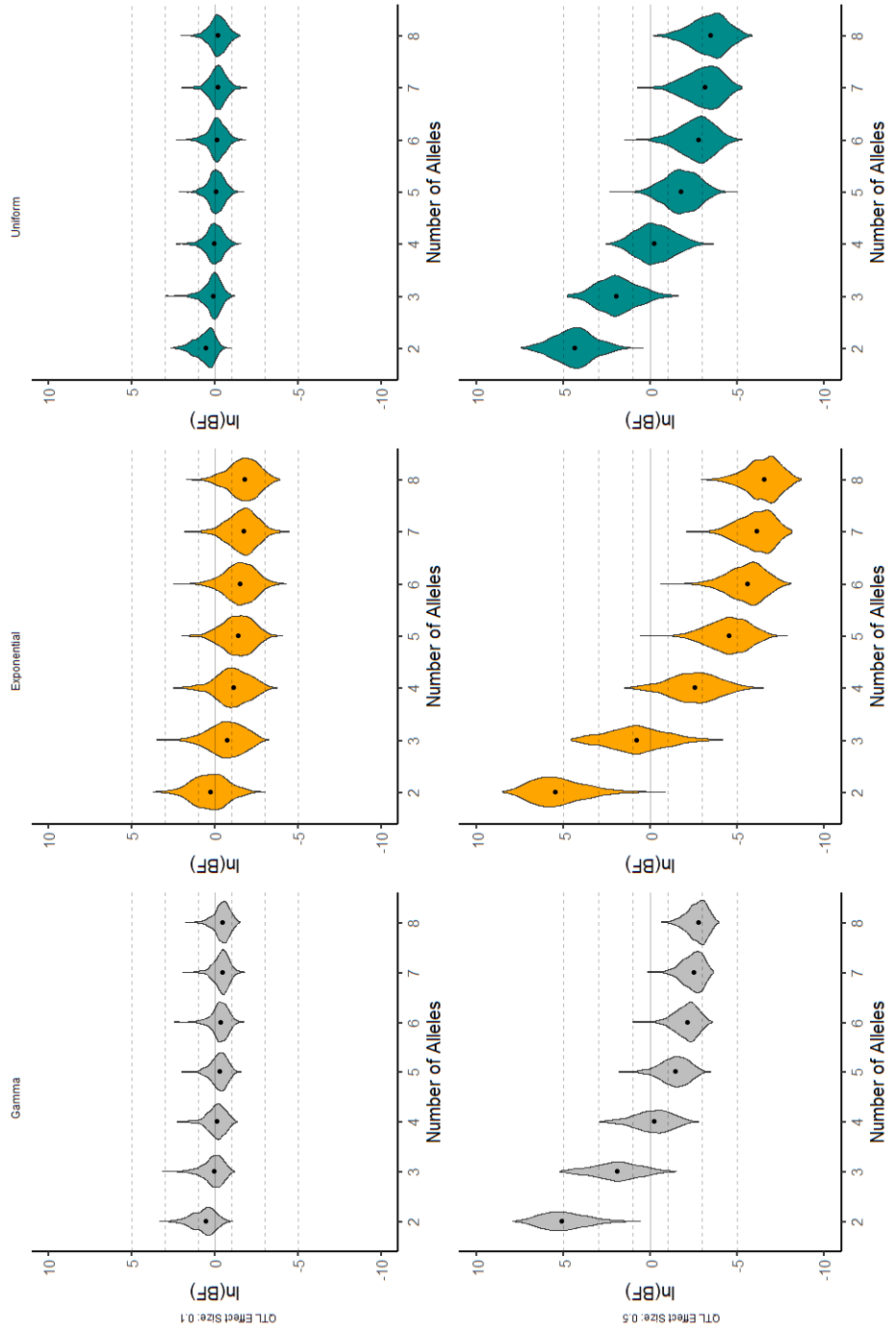
Figure 4.7: lnBFs for alternative prior distributions relative to Full model, on natural log scale, in low and high power scenarios (respectively, $h^2_{QTL} = [0.1, 0.5]$), for varying numbers of true functional alleles. Horizontal lines correspond to interpretation thresholds. In the positive direction, these ranges correspond to "Positive" from 1-3, "Strong" from 3-5, and "Very Strong" above 5. The interpretation is similar in the negative direction, but with the directionality reversed.
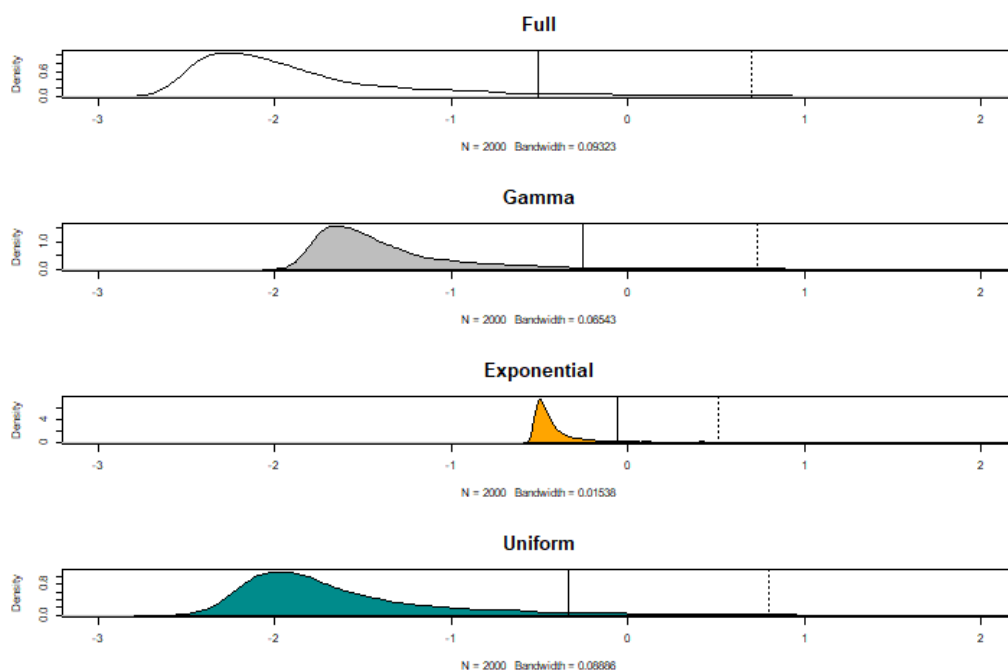
Figure 4.8: lnBFs for alternative prior distributions relative to null model, on natural log scale, pooled across high and low power scenarios (respectively, $h^2_{QTL} = [0.1, 0.5]$), when the true allelic series is null. The solid and dotted vertical lines correspond to the 95% and 99% observed quantiles.
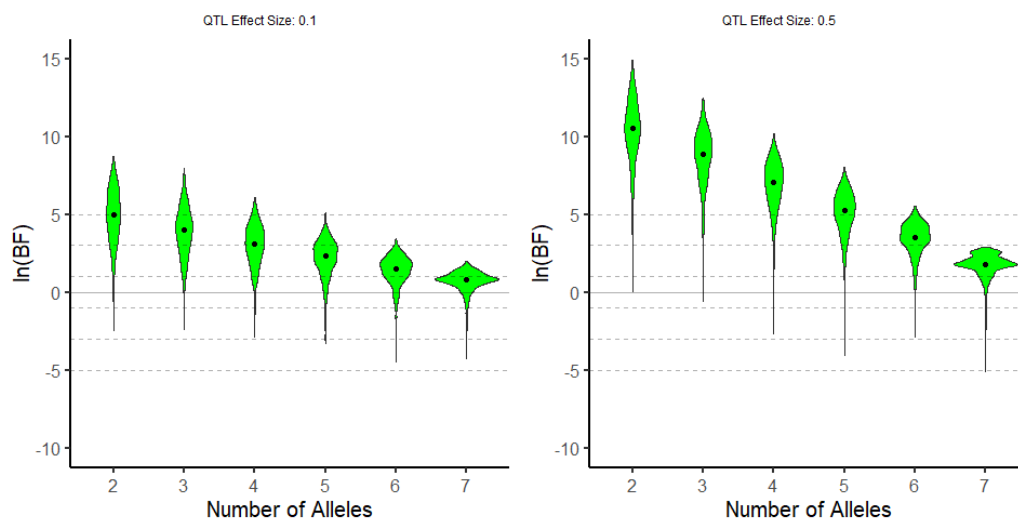
Figure 4.9: lnBFs for the Known model relative to Full model, on natural log scale, in low and high power scenarios (respectively, $h^2_{QTL} = [0.1, 0.5]$), for varying numbers of true functional alleles. Horizontal lines correspond to interpretation thresholds. In the positive direction, these ranges correspond to "Positive" from 1-3, "Strong" from 3-5, and "Very Strong" above 5. The interpretation is similar in the negative direction, but with the directionality reversed.
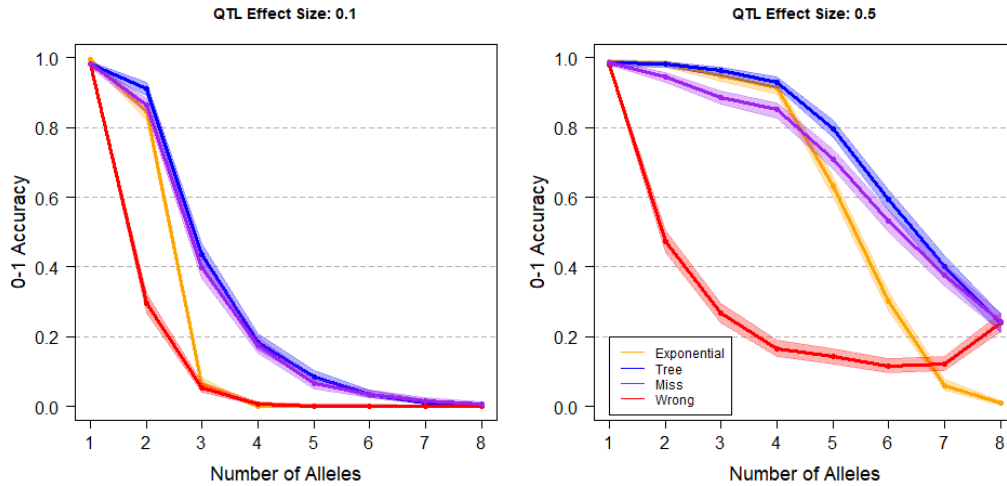
Figure 4.10: 0-1 accuracy of posterior allelic series inference for alternative prior distributions, in low and high power scenarios (respectively, $h^2_{QTL} = [0.1, 0.5]$), for varying numbers of true functional alleles.

4.5   Results - Tree Information

In this section, we investigate the utility of including additional prior information on the phylogenetic relatedness of the founder haplotypes, comparing the Exponential and Tree alternatives. We also consider the possibility that this tree information is either partially misspecified (Miss) or completely misspecified (Wrong).

4.5.1   Accuracy of Allelic Series

Figure 4.10 shows the 0-1 accuracy of the MAP allelic series under the Exponential, Tree, Miss and Wrong alternatives, for varying numbers of true functional alleles. In both power scenarios, including phylogenetic tree information provides a modest increase in accuracy relative to not including tree information, even when that tree information is partially misspecified. However, including Wrong tree information substantially decreases accuracy relative to not including tree information.

Figure 4.11 shows the posterior certainty of the allelic series under the same alternatives. As before, tree information increases accuracy even when partially misspecified. Notably, posterior certainty when the tree is completely misspecified is still similar to posterior certainty when no tree information is provided, despite reducing 0-1 accuracy. This suggests that Wrong tree information tends to increase certainty on incorrect allelic series, rather than reducing certainty on the correct al-
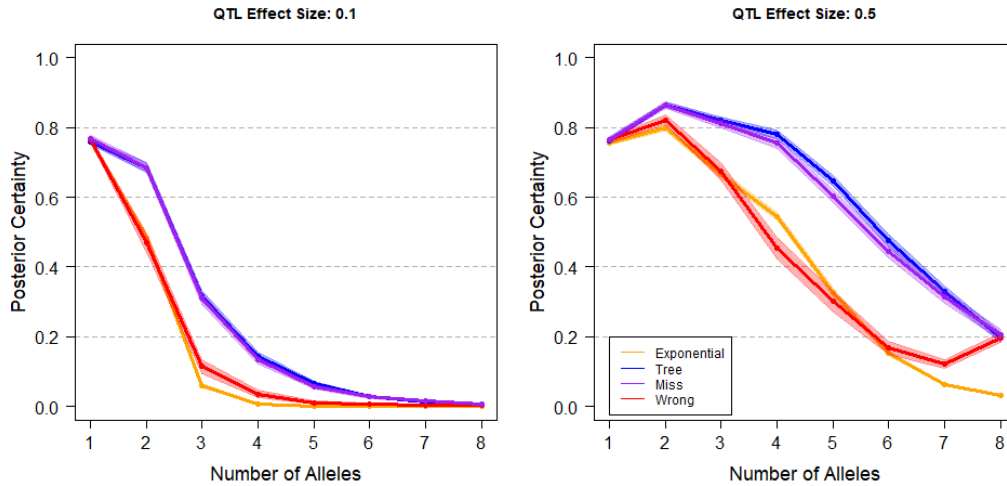
Figure 4.11: Posterior certainty of allelic series inference for alternative prior distributions, in low and high power scenarios (respectively, $h^2_{QTL} = [0.1, 0.5]$), for varying numbers of true functional alleles.

lelic series. This hypothesis could be assessed more directly using the full distribution of simulations, rather than just the mean that we show here.

Figure 4.12 is similar to the previous figures but shows the posterior expectation of number of alleles. Again, these results show incremental improvement when using tree information, subject to the same general trends that we observed in the previous set of simulation results. These results also indicate a relative tendency towards overestimation when Wrong tree information is provided.

4.5.2   Error of Haplotype Effect Estimation

Figure 4.13 shows the MSE of haplotype effect estimates for the Exponential, Tree, Miss, and Wrong alternatives, additionally compared with the Full and Known models. In the low power scenario, both the Tree and Miss alternatives improve on Exponential for low to intermediate numbers of alleles, and all numbers of alleles are improved in the high power scenario. Wrong tree information generally increases MSE relative to Exponential. In the low power scenario, this is sufficient to make the Wrong worse than the Full for most numbers of functional alleles. However, in the high power scenario, even when the tree is completely misspecified, the allele-based Wrong alternative still improves on the haplotype-based Full model when the true number of functional alleles is small. This suggests that, although Wrong tends to overestimate the true number of functional alleles, when
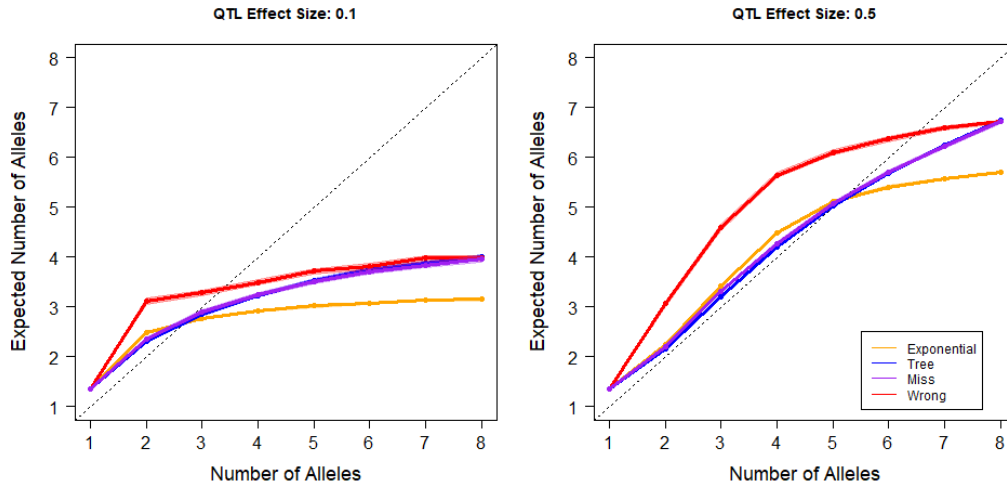
Figure 4.12: Posterior expectation of number of alleles for alternative prior distributions, in low and high power scenarios (respectively, $h^2_{QTL} = [0.1, 0.5]$), for varying numbers of true functional alleles.

the true number of functional alleles is small, it is still possible to find more-parsimonious allelic series that provide better haplotype effect estimates than the Full model.

4.5.3 Statistical Signal

Figure 4.14 shows the lnBF for the Tree, Miss, and Wrong alternatives relative to the Exponential. In the low power scenario, there is positive evidence in favor of Tree information when the number of alleles is between two and four, and in the high power scenario, there is positive to strong evidence in favor of Tree information when the number of alleles is between two and six. The evidence for Tree information is negative for higher numbers of alleles across both power scenarios. When there is only one functional allele (the null model), the Tree lnBFs are centered near zero, suggesting that Tree and Exponential behave similarly under the null. In all cases, the Miss alternative is similar to but slightly worse than the Tree alternative, reflecting that it is partially misspecified but still contains some correct information. When the true number of alleles is greater than zero and power is low, there is negative evidence for the Wrong tree information relative to the Exponential, which is reflects that it contains incorrect phylogenetic information. When power is high, the evidence for the Wrong alternative is strongly negative. In general, the distributions for all lnBFs are diffuse, indicating a high degree of variability around the central tendencies we have described.

103
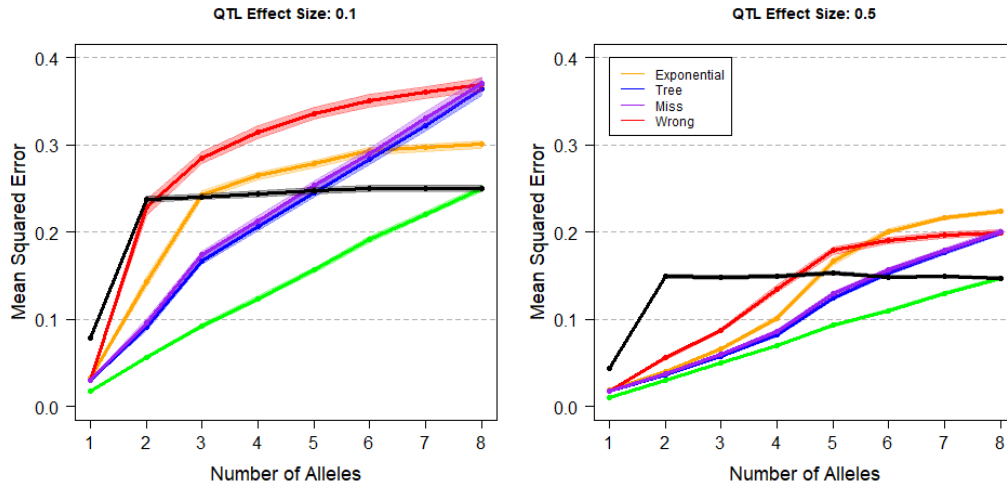
Figure 4.13: MSE of haplotype effect estimates for alternative prior distributions, in low and high power scenarios (respectively, $h^2_{QTL} = [0.1, 0.5]$), for varying numbers of true functional alleles.

4.6 Discussion

In this chapter, we used simulations to evaluate the haplotype- and allele-based association approaches with respect to their accuracy, error in estimating haplotype effects, and relative statistical signal. In the first set of results, we focused on allele-based models in the absence of additional phylogenetic information. We found that, for the CRP-based Gamma and Exponential approaches, the accuracy to identify the correct allelic series is high when the QTL is biallelic, with the Exponential being decisive in this case. However, when the QTL is multiallelic, accuracy deteriorates for all alternatives as the true number of alleles increases, even when power is high. We found that the Uniform prior on the allelic series, which places considerable weight on an intermediate number of alleles, is only competitive in that case. All the allele-based methods tend to underestimate the number of alleles when the true number of alleles was high. This is consistent with simulations results in both Azim Ansari and Didelot (2016) and King et al. (2014). Respectively, these studies used approaches most similar to our Tree and Uniform approaches, and both showed a tendency to underestimate either distributional "changepoints" or number of alleles. With respect to overestimation, we found that when the true number of alleles is low, the Exponential alternative is less prone to overestimation than the other alternatives.

104
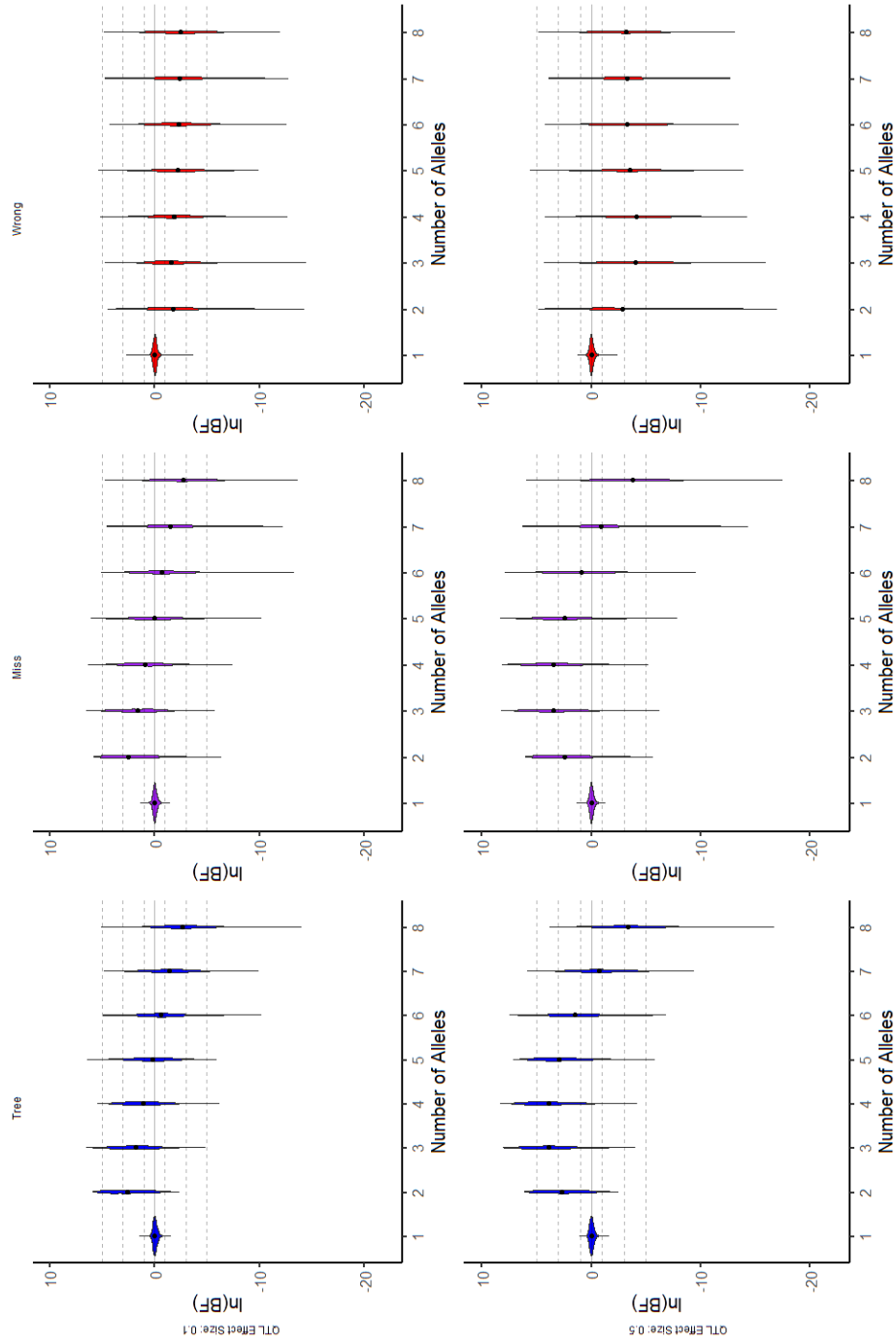
Figure 4.14: lnBFs for alternative prior distributions relative to Exponential model, on natural log scale, in low and high power scenarios (respectively, $h^2_{QTL} = [0.1, 0.5]$), for varying numbers of true functional alleles. Horizontal lines correspond to interpretation thresholds. In the positive direction, these ranges correspond to "Positive" from 1-3, "Strong" from 3-5, and "Very Strong" above 5. The interpretation is similar in the negative direction, but with the directionality reversed

Despite a general uncertainty about the posterior allelic series, when the true number of alleles is small, allele-based methods reduce error in haplotype effects relative to the Full haplotype-based approach, even when power is low. When power is high, this improvement is also apparent with an intermediate number of alleles, and it approaches the lower-bound established by Known allele effects. However, when the true number of alleles is large, the allele-based methods are worse than the Full haplotype-based method. As implied by their prior distributions, the Exponential is better than the Gamma when the true number of alleles is small and worse when the true number of alleles is large.

These observations point to a somewhat paradoxical conclusion: Although there is often insufficient information to identify the allelic series with certainty, accounting for the allelic series can still improve effect estimation, provided that the true number of alleles is small. This is consistent with Jannink and Wu (2003), which also used an allelic series model and performed many similar simulations. The allele-based model improves effect estimation because it allows the data to be represented using fewer parameters, and this reduction in parameters can still be beneficial even when the allelic series is only partially known.

To provide perspective on identifying QTL, we also evaluated the statistical signal of the allele-based alternatives relative to the Full haplotype-based approach. When power is high, we found strongly improved signal when the QTL is bi- or triallelic, and a reduction in signal otherwise. When power is low, we found only small differences in statistical signal between these alternatives. Under the null of no genetic effect, the lnBFs for Exponential were tightly distributed near zero, with a shorter positive tail than the other approaches. In the context of mapping QTL, when stringent significance thresholds are necessary, this conservative behavior may be beneficial.

We recommend using the Exponential prior distribution for the allele-based models. In many applied cases it will be reasonable to expect that QTL have only a few functional alleles, and when this is the case, the Exponential performs consistently better than the other alternatives across most metrics. In particular, when the number of alleles is small, the Exponential is less likely to overestimate the number of alleles than the Gamma. Additionally, the Exponential provides a decisive answer in the case of biallelic QTL. Haplotype effect estimates and statistical signal are similar for both the Gamma and Exponential, but the Exponential has potentially better, more conservative,

behavior under the null. For this reason, we used the Exponential as the basis for our second set of simulations, and it is the default approach for the remaining chapters.

Although the gain in statistical signal from using the allele-based models seems marginal, it is more compelling when considering the upper bound given by the Known allelic series. In this idealized case of perfect information, we found strong improvements in signal for all numbers of functional alleles, although this improvement does decrease with power and as the number of alleles increases. Together, these results suggest that complete knowledge of the allelic series improves statistical signal, but that there is often insufficient posterior certainty to fully realize this benefit.

One way to increase posterior certainty is to include additional prior information. Within our framework, additional information about the allelic series is introduced by conditioning on a coalescent tree that describes the phylogenetic relationship between the founder haplotypes. Across all our metrics, we found that including tree information improves the performance of the allelic series approach, even when that tree information is partially, but not completely, misspecified. The tree-informed approach behaves similarly to the tree-naive approach under the null. These promising results indicate that our tree-informed allelic series approach can improve inference about genetic architecture, phenotype prediction, and perhaps QTL detection.

CHAPTER 5

Application to Real Data

5.1 Overview

In this chapter, we apply our allele-based association approach to several real-data examples, each of which highlights a key point about our approach for allelic series inference. The first example, an analysis of a QTL for a red blood cell phenotype detected in incipient lines of the CC (PreCC) by Kelada et al. (2012), introduces allelic series inference and demonstrates how local phylogenetic information improves posterior inference of the allelic series. The second example, an analysis of whole lung cis-eQTL detected in the PreCC by Kelada et al. (2014), summarizes the distribution of allelic series over many QTL and identifies QTL which appear highly multiallelic. The final examples, an analysis of two whole head cis-eQTL detected in the Drosophila Synthetic Population Resource (DSPR) by King et al. (2014), an MPP with fifteen founder haplotypes rather than the typical eight, shows that our allele-based approach (without tree information) is applicable even when there are many founder haplotypes. Together, these examples showcase the potential usefulness of the allele-based association approach in future MPP studies.

The following sections address each of these examples in more detail. Each section begins with an overview of the data and example-specific methods, followed by a summary of the results. In all of the examples, we apply both the "Full" haplotype-based approach and the "Exponential" allele-based approach. These were defined in the previous chapter, and we now refer to the "Exponential" more generally as the "CRP". In the first example, we also apply the "Tree" approach, with an adjustment for uncertainty in the tree.

5.2 Inference with Tree Information

5.2.1 Data and Methods

In this section, we apply our allele-based approach to a QTL for mean red blood cell volume (MCV) previously reported in Kelada et al. (2012). This study of blood parameters used 131 genetically diverse mice that were partially-inbred (5-14 generations) during the early development of the CC. The authors identified a large-effect QTL for MCV on chromosome 7 along with a candidate causal gene, *Hbb-b1*. For our allelic series analyses, we used the phenotype values for MCV measured in this study and assumed that the genomic region around *Hbb-b1* is causal for the QTL.

Our allelic series approach requires full diplotype state probabilities for each mouse, as inferred by haplotype reconstruction. However, the published study focused on expected haplotype dosages, and the full diplotype state probabilities were discarded due to their size. Thus, for our analyses, we performed another haplotype reconstruction using the same genotype information, also using HAPPY (Mott et al., 2000) but with different settings. To speed computation, the published study averaged haplotype dosages from adjacent loci if there was no evidence of recombination across them in the PreCC population. To remain consistent with the published results of this study and Kelada et al. (2014), we averaged the diplotype state probabilities from our new haplotype reconstruction over the same regions.

To infer the phylogenetic tree of the founder haplotypes at the causal genomic region, we performed the following steps. First, we found the location of *Hbb-b1* (renamed *Hbb-bs*; Chr7: 103,826,523-103,827,928 in GRCm38/mm10), as reported by Mouse Genome Informatics (Bult et al., 2019). Next, we identified a larger 23kb nonrecombinant region surrounding the gene (Chr7: 103,807,679 103,831,178) by applying the four-gamete test (Hudson and Kaplan, 1985) to high-quality SNPs from Sanger Mouse Genomes Project (Keane et al., 2011). Then, we constructed a sequence alignment for the founder haplotypes using high quality SNPs and indels from the same source. Next, we used BEAST 1.8.3 (Drummond et al., 2012) to infer a coalescent phylogeny for this sequence alignment, assuming a constant mutation rate, constant population size and the HKY substitution model (Hasegawa et al., 1985). We generated one million MCMC samples from the posterior of coalescent trees, thinning every 1000 samples, yielding a total of 1000 posterior samples

Allelic Series	# of Alleles	Posterior Probability
0,1,0,1,1,0,0,0	2	0.5568
0,1,0,1,1,2,0,0	3	0.0801
0,1,2,1,1,2,0,0	3	0.0644
0,1,0,1,1,0,2,2	3	0.0278
0,1,0,1,1,0,2,0	3	0.0204
0,1,0,2,1,0,0,0	3	0.0189
0,1,2,1,1,0,0,0	3	0.0181
0,1,0,1,2,0,0,0	3	0.0178
0,1,0,2,2,0,0,0	3	0.0160
0,1,0,1,1,0,0,2	3	0.0132

Table 5.1: Top ten posterior allelic series for MCV QTL using the CRP approach.

of the tree. These trees are visualized using Densitree (Bouckaert and Heled, 2014). Finally, we computed the allelic series prior distribution for each sample of the tree and averaged the result in order to arrive at a final tree-informed allelic series prior distribution for this QTL.

We used the Full, CRP and Tree approaches to analyze this QTL, generating 100,000 posterior samples for each.

5.2.2 Results and Discussion

Figure 5.1 shows the MCV phenotype for the 94 of 131 mice with prior maximum diplotype states that are homozygous at the QTL, plotted by that haplotype. The phenotype clearly depends on the haplotype at the QTL, but intuitively, the number of functional alleles is not obvious. **Table 5.1** shows the top ten posterior allelic series inferred using the CRP approach. The top allelic series is biallelic and comprises 55.7% of the posterior probability, but there are several other multiallelic series with reasonable support (11 with $\geq$ 1%, together accounting for another 30.1% of the posterior). These multiallelic series preserve the biallelic contrast identified by the top allelic series, indicating that this is a high-confidence feature of the haplotype effects. The posterior expected number of alleles is 2.59, and the posterior distribution of the number of alleles is given in **Figure 5.2**. Overall, these results provide evidence in favor of a biallelic QTL but allelic series inference is still uncertain.

The left panel of **Figure 5.3** shows the posterior distribution of haplotype effects using both the Full and CRP approaches. As expected, the Full haplotype effect estimates are similar to the observed phenotypes in **Figure 5.1**. Relative to the Full, the CRP haplotype effects are more certain, with narrower 95% highest posterior density (HPD) intervals, as shown in **Table 5.2**. This increased
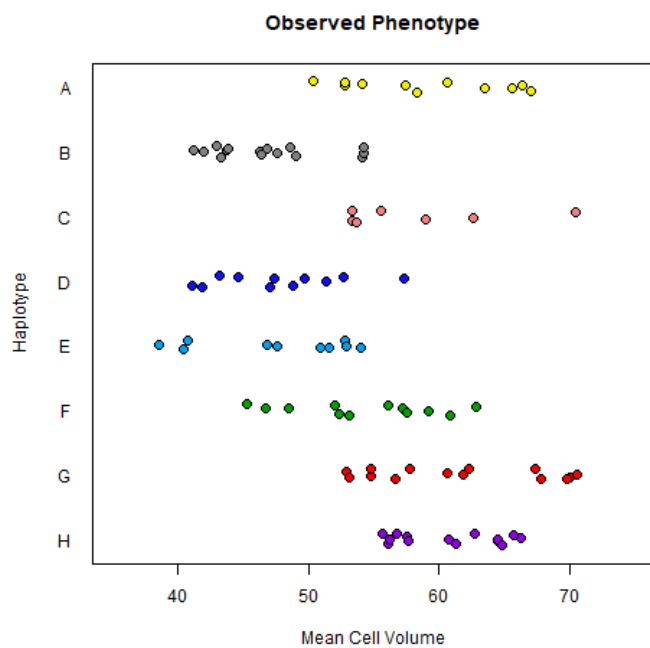
Figure 5.1: Mean cell volume by founder haplotype at the QTL for mice with homozygous prior maximum diplotype state.
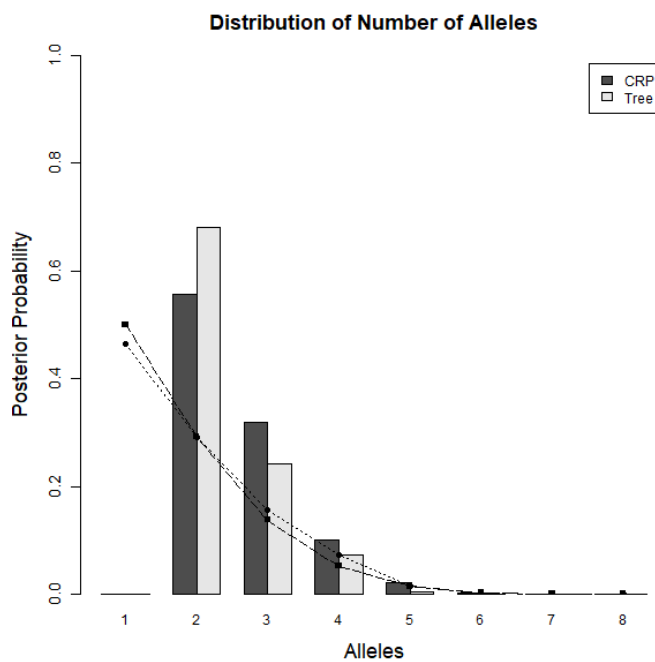
Figure 5.2: Posterior distribution of number of alleles for MCV QTL using the CRP and Tree approaches. The lines denote the corresponding prior distributions; square = CRP, circle = Tree.
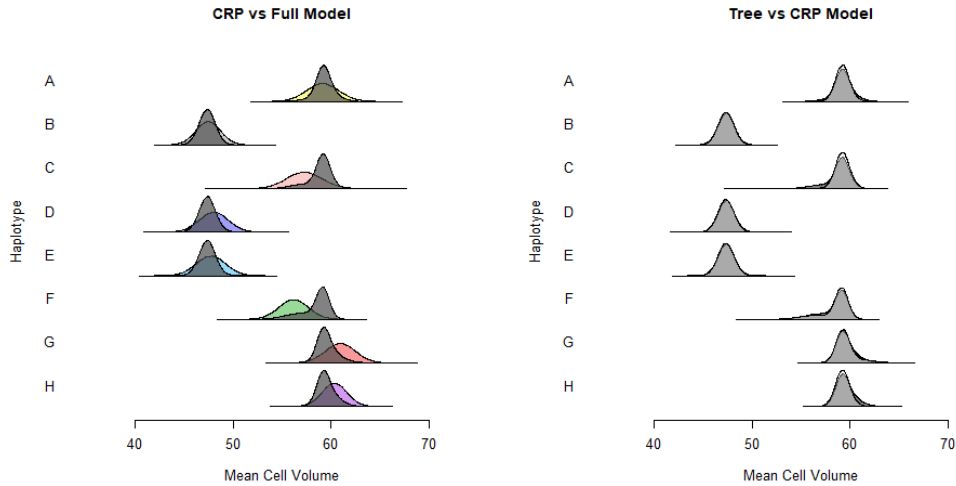
Figure 5.3: Posterior distribution of haplotype effects for MCV QTL using the Full, CRP, and Tree approaches. Full - multicolored; CRP - dark gray; Tree - light gray.

certainty is because the allelic series model allows information about the effects to be shared across haplotypes. This information sharing is particularly evident for haplotype F, which has its effect distribution pulled towards an allele effect that is shared with A, C, G, and H. Nonetheless, the haplotype effect distribution of F retains a long tail, covering much of the original range of the Full effect distribution. Comparing the Full and CRP approaches more broadly, the lnBF in favor of the CRP is 1.17, indicating positive evidence in favor of allele-based effects.

Figure 5.4 shows samples of phylogenetic trees that relate the founder haplotypes at the causal locus. For most samples, there are long branches separating haplotypes B, D, and E from the other five haplotypes. Among the others, haplotypes A, H, and C are more closely related than F and G. Relative to the coalescent, which is depicted in **Figure 5.5**, these trees are highly structured,

Haplotype	Full	CRP	Tree
A	6.60	4.06	2.99
B	5.03	3.36	3.23
C	7.14	5.04	3.04
D	5.87	3.44	3.23
E	5.96	3.43	3.23
F	5.96	5.57	5.72
G	6.08	3.96	4.43
H	5.07	3.74	2.99

Table 5.2: Width of the 95% highest posterior density interval for haplotype effects using the Full, CRP, and Tree approaches.
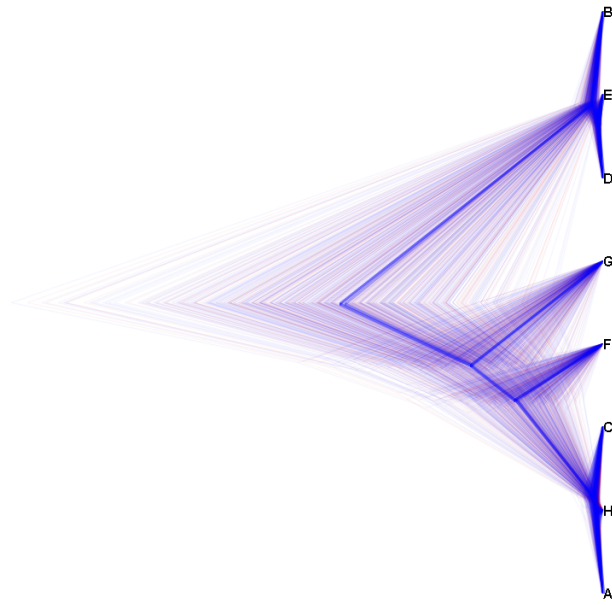
Figure 5.4: Samples of causal trees at the MCV locus. The consensus tree is in bold, and color denotes different tree topologies.

representing only three of 10,395 possible tree topologies. This structure informs the prior distribution of the allelic series in the Tree model, as shown in **Figure 5.6**. There are 720 allelic series with support using the Tree approach, compared with the full space of 4,140 using the CRP. The allelic series favored by the Tree approach reflect the relationships encoded by the causal trees; for example, the top non-null allelic series is biallelic and contrasts haplotypes B, D, and E against the others, and its prior probability is increased over 150-fold relative to the CRP. Notably, this allelic series is also the top result returned by the CRP approach.

Table 5.3 shows the top ten posterior allelic series inferred using the Tree approach. The top allelic series is unchanged from the CRP results, but its posterior probability is increased to 68.1%. There are fewer multiallelic series with reasonable support (5 with $\geq$ 1% posterior probability), and they have been informed by the phylogenetic distance between F, G and the other haplotypes. The posterior expected number of alleles is lower than the CRP at 2.40, and the posterior distribution of the number of alleles is given in **Figure 5.2**.

The right panel of **Figure 5.3** shows the posterior distribution of haplotype effects using both the CRP and Tree approaches. The Tree-informed haplotype effects are largely unchanged from the CRP
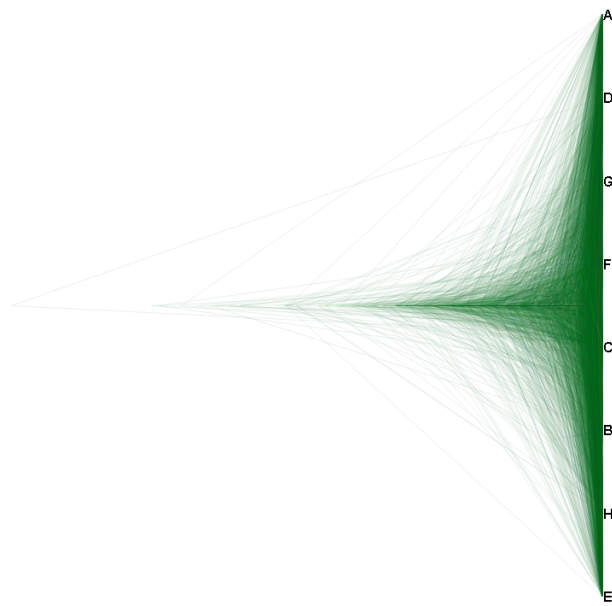
Figure 5.5: Samples of coalescent trees.
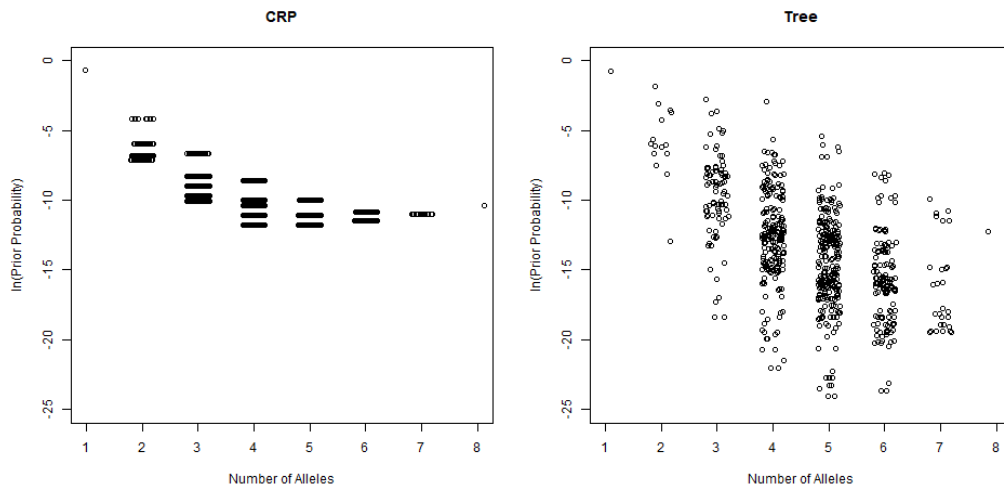
Figure 5.6: Prior probability of allelic series by number of alleles, using the CRP and Tree approaches.

114

Allelic Series	# of Alleles	Posterior Probability
0,1,0,1,1,0,0,0	2	0.6808
0,1,0,1,1,2,0,0	3	0.1335
0,1,0,1,1,0,2,0	3	0.0865
0,1,0,1,1,2,3,0	4	0.0639
0,1,0,1,1,2,2,0	3	0.0122
0,1,2,1,1,0,0,0	3	0.0027
0,1,0,2,2,0,0,0	3	0.0027
0,1,2,1,1,3,0,0	4	0.0021
0,1,2,1,1,3,4,0	5	0.0019
0,1,0,1,1,0,0,2	3	0.0015

Table 5.3: Top ten posterior allelic series for MCV QTL using the Tree approach.

haplotype effects, with only modest changes in HPD lengths (**Table 5.2**). Overall, there is strong positive evidence for the Tree approach relative to the CRP, with a lnBF of 4.81 in favor of the Tree.

In summary, this example demonstrates that our method can be used to infer the allelic series at a QTL, that this inference can improve haplotype effect estimation, and that including additional phylogenetic information can increase both posterior certainty and statistical signal.

5.3 Identifying Multiallelic QTL

5.3.1 Data and Methods

In this section, we apply our allele-based approach to a collection of gene expression QTL previously reported in (Kelada et al., 2014). This study analyzed rank-normalized gene expression, measured by microarray, in whole long tissue from PreCC mice. For our analyses, we focused on genes with cis-eQTL (within 10Mb of the gene), and we ignored eQTL for which array probes contained SNPs segregating between the founder strains, as these bias the microarray and are a potential source of false positive QTL (Alberts et al., 2007). These criteria left 4,516 cis-eQTL for our allelic series analyses.

Our allelic series approach requires full diplotype state probabilities for each mouse, but this study used many of the same mice from (Kelada et al., 2012), and these probabilities were also not retained. For our analyses, we used diplotype probabilities from the new haplotype reconstruction described in the previous section. For an unknown reason, seven previously-reported cis-eQTL no
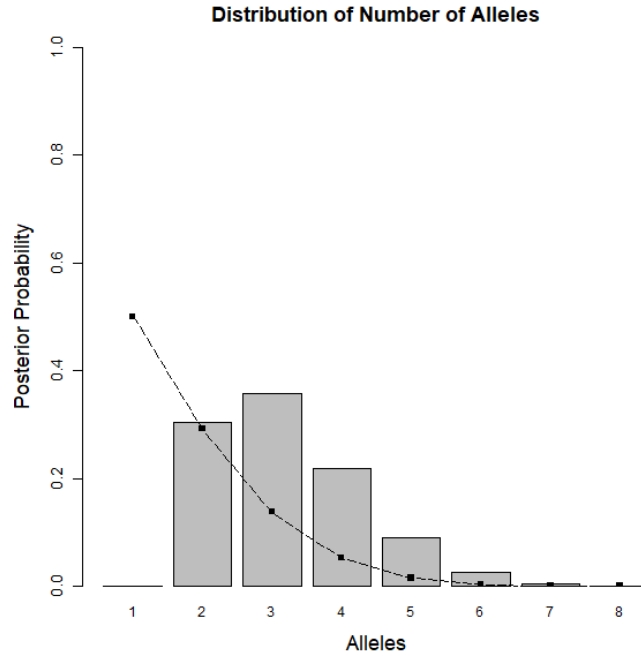
Figure 5.7: Posterior distribution of number of alleles, averaged over all cis-eQTL. The line denotes the corresponding prior distribution.

longer had corresponding genomic intervals in the new diplotype data. These were omitted from our analyses.

We used the CRP approach to analyze these 4,509 cis-eQTL, generating 100,000 posterior samples for each. We also used the Full approach when analyzing our most multiallelic result, with the same number of samples.

5.3.2 Results and Discussion

Figure 5.7 shows the posterior distribution of number of alleles, averaged over all cis-eQTL. This suggests that many QTL are multiallelic, with 35.7% and 21.8% posterior probability for three and four alleles, respectively. There is also substantial support for biallelic QTL, which has 30.3% posterior probability. Given that these QTL were genome-wide significant in (Kelada et al., 2014), it is not surprising that there is near-zero support for the null model of one allele. There is general support for 2-4 alleles, but there are also genes which appear highly multiallelic. **Table 5.4** highlights the most highly multiallelic QTL in this dataset.

116

	Probe	Gene	Chr	Position	Expected Alleles
1	ILMN_2667352	Glo1	17	30,729,806	6.7549
2	ILMN_2418957	BC032285	17	6,270,475	6.3900
3	ILMN_3023451	Zfp985	4	146,918,112	6.3694
4	ILMN_2880052	Xlr4b	X	70,459,704	6.2664
5	ILMN_2499598	AK009977	11	58,777,283	6.1449
6	ILMN_3004949	Fam55d	9	47,970,198	5.8813
7	ILMN_2643495	Fez1	9	36,640,394	5.8658
8	ILMN_2685581	H2-K1	17	34,132,957	5.7066
9	ILMN_1221376	Cyp4f39	17	32,589,668	5.6539
10	ILMN_3153940	Unc45b	11	82,724,831	5.6496
11	ILMN_2634905	Fbp2	13	62,938,245	5.6022
12	ILMN_2998406	Zfp979	4	146,986,048	5.5955
13	ILMN_1213056	Fez1	9	36,640,394	5.5861
14	ILMN_1236008	Isoc2a	7	4,828,740	5.5309
15	ILMN_2776728	Zfp979	4	146,986,048	5.4501
16	ILMN_2735046	Cml3	6	85,711,089	5.4235
17	ILMN_2527805	Wfdc10	2	164,481,546	5.4121
18	ILMN_2665266	H2-T22	17	36,175,354	5.4084
19	ILMN_2894678	H2-T22	17	36,175,354	5.4022
20	ILMN_2584887	Atp5f1	3	105,745,781	5.3849

Table 5.4: Highly multiallelic cis-eQTL; top twenty by posterior expected number of alleles. Gene positions from NCBI37/mm9.

The most multiallelic cis-eQTL in our dataset was *Glo1*. **Figure 5.8** shows *Glo1* expression for the 111 of 138 mice with prior maximum diplotype states that are homozygous at the QTL, plotted by that haplotype. Our approach finds over 95% posterior support for six to eight alleles at this QTL (**Figure 5.9**). **Figure 5.10** shows the posterior distribution of haplotype effects using both the Full and CRP approaches, which are similar in light of the high posterior number of alleles. Interestingly, previous studies have found that mouse strains have a complicated haplotype structure at *Glo1*, and that expression of this gene is associated with anxiety-like behavior in mice (Williams IV et al., 2009). This supports our finding that *Glo1* is highly multiallelic, and it also suggests that genes with multiallelic QTL may also be associated with other interesting phenotypes.

We also note that several highly multiallelic cis-eQTL are near the major histocompatibility complex on chromosome 17, which is consistent with high genetic diversity in this region (Lilue et al., 2019).

In conclusion, we find that whole lung cis-eQTL in the PreCC tend to have between two and four alleles. Given that the CC founders are comprised of three different subspecies of mice (Didion
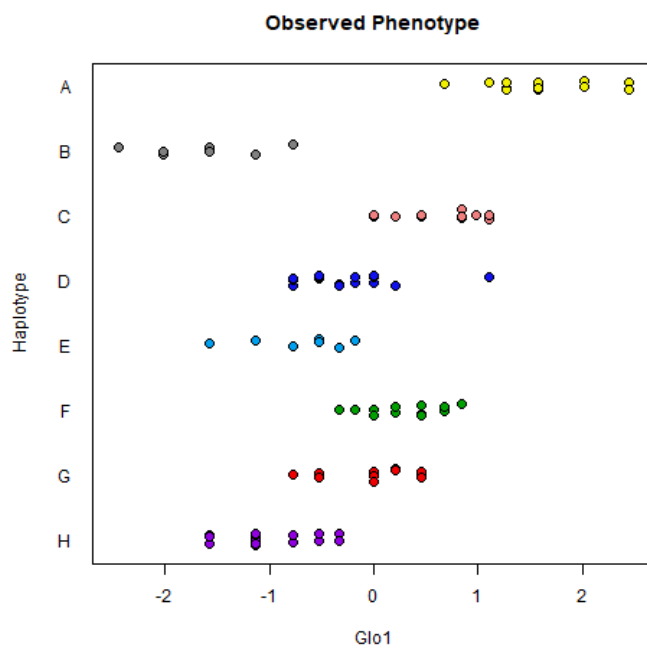
Figure 5.8: *Glo1* expression by founder haplotype at the QTL for mice with homozygous prior maximum diplotype state.

Figure 5.9: Posterior distribution of number of alleles for *Glo1* cis-eQTL. The line denotes the prior distribution.

Figure 5.10: Posterior distribution of haplotype effects for *Glo1* cis-eQTL using the Full and CRP approaches. Full - multicolored; CRP - dark gray.

and De Villena, 2013), this multiallelism is reasonable. It would be interesting to directly interrogate if the posterior allelic series for a gene corresponds to its subspecies origin. Additionally, we find highly multiallelic cis-eQTL in this dataset. In particular, *Glo1* is our most multiallelic gene, and its complicated haplotype structure has been implicated in anxiety-like behavior in mice. It would be interesting to investigate if other multiallelic genes are also associated with interesting phenotypes, or more generally, if multiallelic genes are enriched for any notable properties.

5.4 Inference with Many Founder Haplotypes

5.4.1 Data and Methods

In this section, we apply our allele-based approach to two gene expression QTL previously reported in (King et al., 2014). This study analyzed rank-normalized gene expression, measured by microarray, in pooled whole head tissue from 596 crosses of DSPR flies, with a total of fifteen founder haplotypes. One objective of this study was to estimate the number of alleles at eQTL. The authors highlighted

Allelic Series	# of Alleles	Posterior Probability
0,0,0,0,0,1,1,0,0,2,0,0,0,0,0	3	0.2536
0,0,0,0,0,1,2,0,0,3,0,0,0,0,0	4	0.0503
0,0,0,0,0,1,1,0,0,2,1,0,0,0,0	3	0.0482
0,0,0,0,1,1,1,0,0,2,0,0,0,0,0	3	0.0459
0,0,0,0,0,1,1,0,0,1,0,0,0,0,0	2	0.0361
0,1,1,1,1,0,0,1,1,2,1,1,1,1,1	3	0.0297
0,1,1,1,1,2,2,1,1,0,1,1,1,1,1	3	0.0253
0,0,0,0,0,1,1,0,0,2,2,0,0,0,0	3	0.0238
0,0,0,0,1,2,2,0,0,1,0,0,0,0,0	3	0.0229
0,0,0,0,1,1,1,0,0,2,1,0,0,0,0	3	0.0147

Table 5.5: Top ten posterior allelic series for CG4086 cis-eQTL using the CRP approach.

two examples, CG4086 and CG10245, as examples of biallelic and multiallelic QTL, respectively. We focused on these two examples for our analyses.

We used the Full and CRP approaches to analyze these QTL, generating 100,000 posterior samples for the Full and 1,000,000 samples for the CRP.

5.4.2 Results and Discussion

Figure 5.11 shows the posterior distribution of number of alleles for the CG4086 QTL using the CRP approach. Although the previous study found that this QTL was biallelic, we find a 61.7% posterior probability that the QTL has three functional alleles. **Table 5.5** shows the top ten posterior allelic series, which tend to contrast haplotypes A6, A7, and B2 against the others. This is consistent with the Full posterior haplotype effects, as shown in **Figure 5.12**. Relative to the Full, the allele-based haplotype effects of the CRP are more certain, with narrower 95% HPD intervals (**Table 5.6**). Notably, both the Full and CRP approaches make effect haplotype effect predictions for A1, A5, and B3, all of which are poorly represented at this QTL and were omitted in the previous study. Overall, there is very strong evidence in favor of the of the CRP relative to the Full approach, with a lnBF of 7.71.

Figure 5.13 shows the posterior distribution of number of alleles for the CG10245 QTL using the CRP approach. The previous study found that this QTL was highly multiallelic, a finding which we confirm, with an expected posterior number of alleles of 8.95. However, the posterior distribution of the allelic series is highly uncertain (**Table 5.7**), which is due to the large number of possible allelic series when there are fifteen founder haplotypes and many alleles. **Figure 5.14** shows the
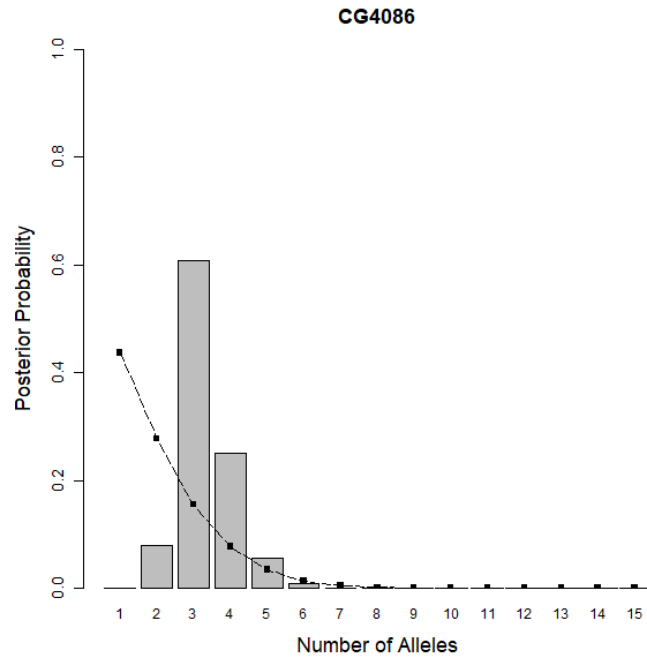
Figure 5.11: Posterior distribution of number of alleles for CG4086 cis-eQTL. The line denotes the prior distribution.
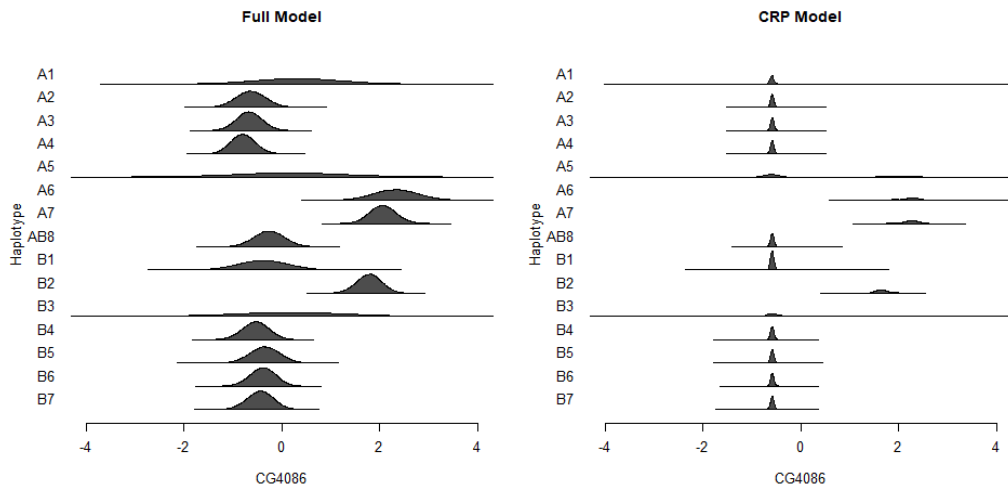
Figure 5.12: Posterior distribution of haplotype effects for CG4086 cis-eQTL using the Full and CRP approaches.

Haplotype	Full	CRP
A1	3.56	2.99
A2	1.19	0.17
A3	1.03	0.17
A4	0.99	0.17
A5	5.11	3.11
A6	1.81	1.31
A7	1.07	0.68
AB8	1.22	0.19
B1	2.03	0.19
B2	1.02	0.56
B3	5.11	3.11
B4	1.09	0.17
B5	1.19	0.18
B6	1.04	0.17
B7	1.06	0.17

Table 5.6: Width of the 95% highest posterior density interval for CG4086 haplotype effects using the Full and CRP approaches.

posterior distribution of haplotype effects using the Full and CRP approaches. Interestingly, many of the haplotype effect distributions for the CRP are multimodal, and the 95% HPD intervals for the CRP are generally wider than for the Full (**Table 5.8**). This is a consequence of the highly uncertain posterior allelic series. The intervals for A4 and B2, both of which are poorly represented at this QTL, are actually narrower, showing how the CRP can provide, in a sense, additional shrinkage to the haplotype effects. Consistent with extensive multiallelism, there is very strong evidence against the CRP relative to the Full approach, with a lnBF of -11.15.

Allelic Series	# of Alleles	Posterior Probability
0,1,2,1,2,1,3,4,1,1,5,6,1,7,8	9	0.002380
0,1,2,3,2,1,3,3,3,3,4,1,3,0,2	5	0.001602
0,1,2,3,2,4,5,3,3,3,1,6,3,5,7	8	0.001206
0,1,2,3,2,1,4,3,3,3,1,5,3,6,0	7	0.001068
0,1,2,3,2,1,4,5,1,1,6,7,1,8,9	10	0.001064
0,1,2,1,2,1,3,4,4,1,1,5,4,6,0	7	0.001050
0,1,2,2,2,1,3,3,3,3,4,1,3,0,2	5	0.001040
0,1,2,3,2,4,5,3,3,3,1,6,3,7,5	8	0.001034
0,1,2,1,2,1,3,4,1,5,6,7,1,8,9	10	0.001030
0,1,2,2,2,1,3,4,1,1,5,6,1,7,8	9	0.001021

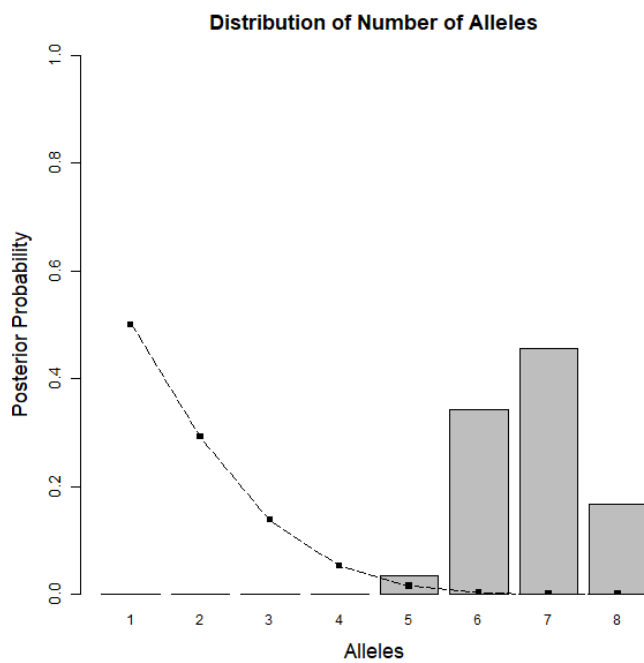Table 5.7: Top ten posterior allelic series for CG10245 cis-eQTL using the CRP approach.
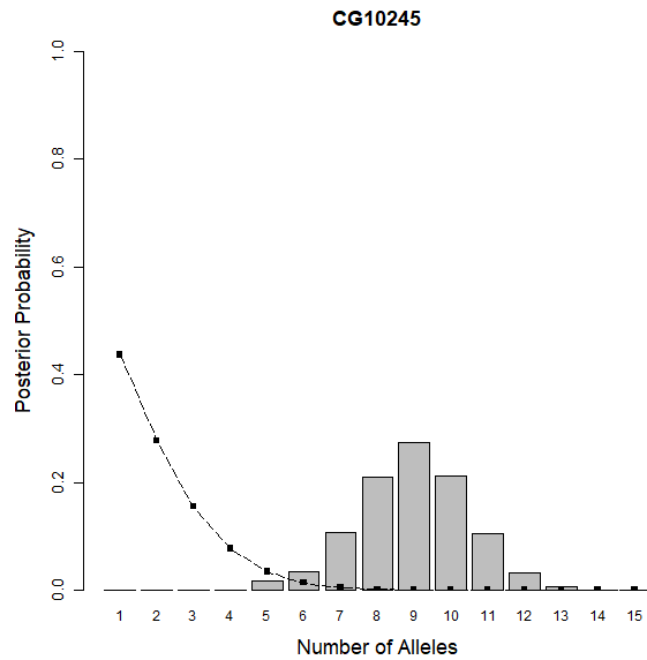
Figure 5.13: Posterior distribution of number of alleles for CG10245 cis-eQTL. The line denotes the prior distribution.

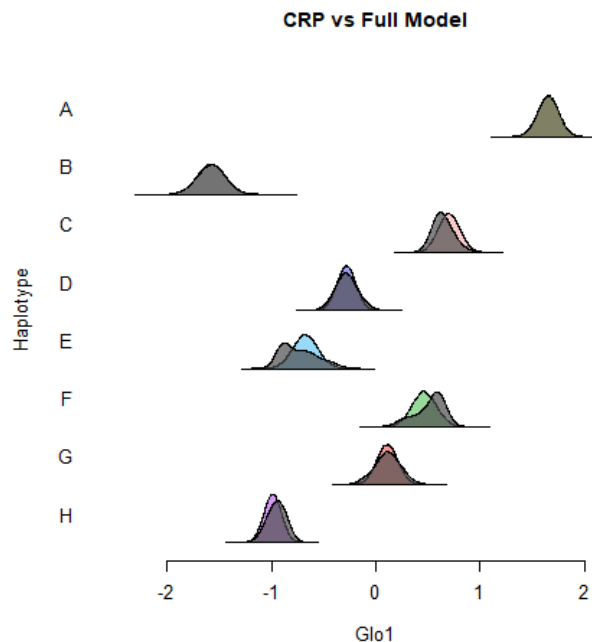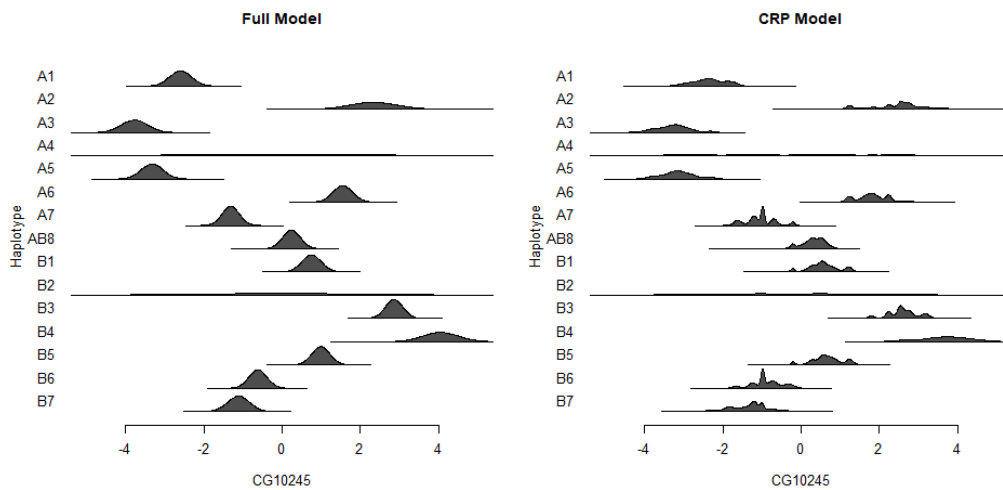Figure 5.14: Posterior distribution of haplotype effects for CG10245 cis-eQTL using the Full and CRP approaches.

Haplotype	Full	CRP
A1	1.10	1.56
A2	2.42	2.61
A3	1.33	1.83
A4	9.95	7.76
A5	1.11	1.72
A6	1.01	1.59
A7	0.88	1.53
AB8	0.90	1.12
B1	0.98	1.55
B2	9.98	7.80
B3	0.89	1.52
B4	1.80	2.53
B5	0.92	1.55
B6	0.91	1.53
B7	1.09	1.88

Table 5.8: Width of the 95% highest posterior density interval for CG10245 haplotype effects using the Full and CRP approaches.

In summary, these examples demonstrate that our approach can be used to infer the allelic series in MPPs with many founder haplotypes. Allelic series inference can be highly uncertain in this case, given the large number of possible allelic series. However, accounting for the allelic series can still improve haplotype effect estimation, provided that the number of functional alleles is small.

CHAPTER 6

Performance in QTL Mapping Simulations

6.1 Overview

In this chapter, we return to the topic of QTL mapping power in the Collaborative Cross (CC), reconsidered in the context of our allele-based association approach. **Chapter 2**, published as Keele et al. (2019), found that the power to detect QTL using a haplotype-based association approach depends on the allelic series. In particular, imbalanced allelic series, where most of the haplotypes are assigned to a single allele, are more difficult to detect than allelic series that are balanced. Given that power depends on the allelic series, it is reasonable to expect that an allele-based association approach could improve QTL mapping power relative to a haplotype-based approach. We introduced such an allele-based association approach in **Chapter 3**, and in **Chapter 4** we found cases where it increases statistical signal. In this chapter, we reevaluate QTL mapping power in the CC, again using simulation, and directly compare several haplotype-based association approaches with our allele-based approach.

We consider four alternative QTL mapping approaches:

- the method used in **Chapter 2**, a haplotype-based approach that estimates the haplotype effects as fixed effects and uses the regression on probabilities (ROP) approximation to account for diplotype uncertainty (termed "FE-ROP") (Haley and Knott, 1992),

- a haplotype-based approach that estimates the haplotype effects as random effects, which also uses ROP ("RE-ROP") (Wei and Xu, 2016),

- the "Full" approach from **Chapter 4**, which is our fully-Bayesian implementation of the haplotype-based approach that includes posterior diplotype sampling,

- the "CRP" approach from **Chapter 4** (previously Exponential), which is our allele-based approach, naive to additional prior tree information.

Comparing these approaches isolates several differences between the FE-ROP model of the first chapter and our allele-based model. First, our approach is more similar to a random effect model, which regularizes haplotype (or allele) effect estimates (highlighted by FE-ROP vs RE-ROP). Second, our haplotype-based approach performs posterior diplotype sampling, instead of approximating diplotype uncertainty using ROP (Full vs RE-ROP). Finally, and of of primary interest, our approach directly models the allelic series rather than assuming that all haplotypes are functionally distinct (CRP vs Full).

We do not consider prior information about haplotype phylogeny in this chapter. The local phylogeny of the CC founder strains is highly variable to recombination and introgression (Yang et al., 2011; Didion and De Villena, 2013), and inferring local phylogeny is a challenging problem, a topic we discuss in **Chapter 7**. Rather than make the strong assumption that the local phylogeny of the CC founders is known at every locus, our simulated QTL assume a uniform distribution on the allelic series, conditional on the number of alleles, as in Keele et al. (2019). Thus, the simulated QTL do not have an underlying true phylogeny, and we focus on the CRP approach, which assumes an unknown, latent coalescent tree. We point to the simulations in **Chapter 4** for the potential benefit of including additional prior tree information relative to the CRP approach.

In the remainder of this chapter, we outline the methods for our simulated QTL experiments, which are similar to **Chapter 2**. We then present the mapping power for each alternative approach. Finally, we conclude by discussing the utility of the allele-based association approach for QTL mapping.

6.2   Simulation Procedure

We simulated QTL using the same procedure from **Chapter 2** (Keele et al., 2019) but with a few modifications. The fully-Bayesian Full and CRP approaches require Gibbs sampling, making them computationally intensive (**Table 6.1**). To minimize computation time, we obtained genome-wide significance thresholds via null parametric bootstrap, as in Valdar et al. (2009), rather than the permutation procedure used in **Chapter 2**. Significance thresholds via permutation require analyzing

Method	Runtime (Min)
FE-ROP	1.42
RE-ROP	3.43
Full	42.63
CRP	891.64

Table 6.1: Runtime in minutes for a single simulated QTL mapping experiment using the FE-ROP, RE-ROP, Full and CRP approaches. All approaches are evaluated using the same hardware, analyzing a dataset of the size simulated in this chapter.

permuted (null) data over many iterations for each simulated QTL experiment. In contrast, the null parametric bootstrap also requires many iterations, but in our case it can be performed just once for all simulated QTL. The QTL we simulate do not include covariates and are standardized to have zero mean and unit variance, so the null parametric model is the standard normal distribution. Thus, we performed the following steps to obtain significance thresholds:

- Sample null phenotypes from the standard normal distribution, creating a null dataset that is the same size as the simulated data,

- Perform QTL mapping on the null data using each alternative mapping approach and collect the maximum observed test statistic. For the FE-ROP and RE-ROP approaches, the test statistic is $-\ln(p)$, and for the Full and CRP it is the lnBF relative to the null model,

- Iterate the previous steps 2,000 times to approximate the null distribution of the maximum test statistic using each approach,

- Fit generalized extreme value distributions for each null distribution and calculate the 95% quantiles (Dudbridge and Koeleman, 2004; Valdar et al., 2006a).

This procedure provides a 95% genome-wide significance threshold via null parametric bootstrap.

We also reduce computation by narrowly focusing this chapter on QTL mapping power. We do not assess the false positive rate (FPR), instead assuming that it is well-controlled by the procedure we just described. Mapping power only involves the threshold and the test statistic at the true locus (or a surrounding region, as in **Chapter 2**). If FPR is ignored and the threshold is known, it is only necessary to evaluate the mapping approaches at the true locus, not for every locus in the genome. This dramatically reduces computation time, allowing us to evaluate many more simulated QTL.

For the simulated QTL, we assume one observation ($r = 1$) from each of the CC strains used in the first chapter ($N = 72$). We specify either two, three or eight functional alleles, and the true allelic series is sampled uniformly conditional on the number of alleles. We also specifically evaluate biallelic series that are either balanced (4v4) or highly imbalanced (7v1). We specify QTL effect size, h^2_{QTL}, from 20% to 50% in increments of 10%. We assume no population structure, making background strain effects ignorable when there are no replicate observations. The QTL effect sizes used in this chapter are selected for illustration, as they span the full range of power for these approaches, given the selected sample size ($N = 72$) and number of replicates ($r = 1$).

For each of these combinations of settings, we simulate 10,000 QTL using the 'sparcc' R package developed by Keele et al. (2019), available on GitHub at `https://github.com/gkeele/sparcc`. We fit the FE-ROP and RE-ROP models using the 'miqtl' package developed by Greg Keele, available at `https://github.com/gkeele/miqtl`. We fit the Full and CRP models using our 'TIMBR' package, available at `https://github.com/wcrouse/TIMBR`. For each simulated QTL, we record if the true locus was detected (i.e. test statistic higher than threshold), and in the case of the CRP, if the MAP allelic series was correct ("0-1 Accuracy", defined in **Chapter 4**).

6.3 Results

Figure 6.1 shows QTL mapping power using the FE-ROP, RE-ROP, Full, and CRP approaches for varying numbers of true functional alleles. In general, differences in power between the four approaches are minor, but there are several clear trends. The allele-based CRP approach outperforms the Full haplotype-based approach when there are only two functional alleles, and it remains narrowly better when there are three functional alleles. When there are eight functional alleles, the CRP is narrowly worse. The Full and RE-ROP approaches are nearly identical, suggesting that there is not much uncertainty in the diplotype state probabilities used for these simulations. RE-ROP is consistently better than FE-ROP, indicating that regularizing haplotype effect estimates increases power. Consistent with the results of our simulation chapter, **Figure 6.2** shows that, when power is high, the CRP approach accurately identifies biallelic series, but this is not the case for multiallelic series.
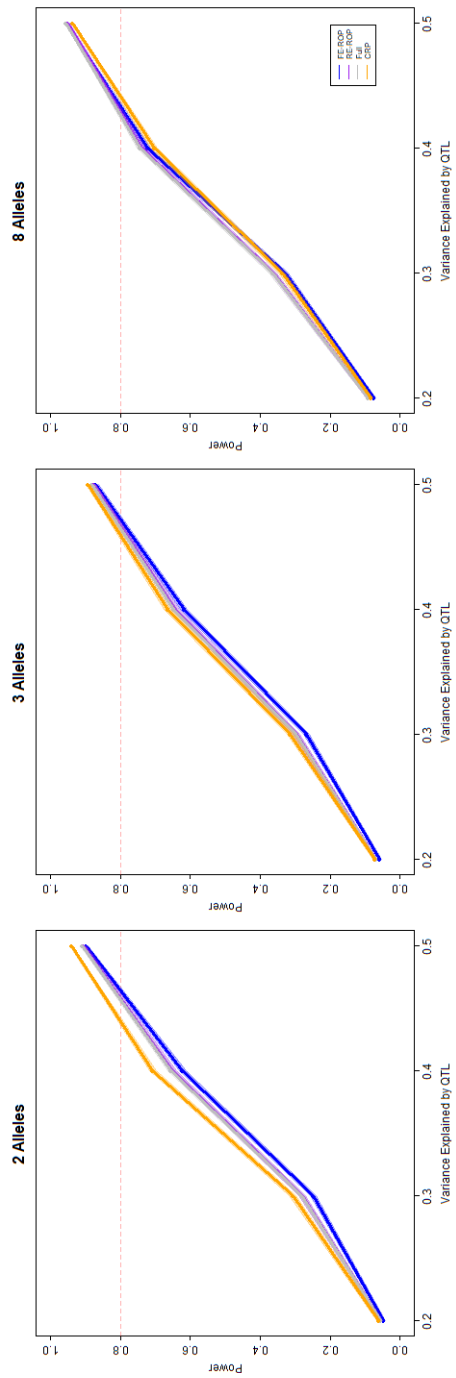
Figure 6.1: QTL mapping power by QTL effect size using the FE-ROP, RE-ROP, Full, and CRP approaches, for varying numbers of true functional alleles.
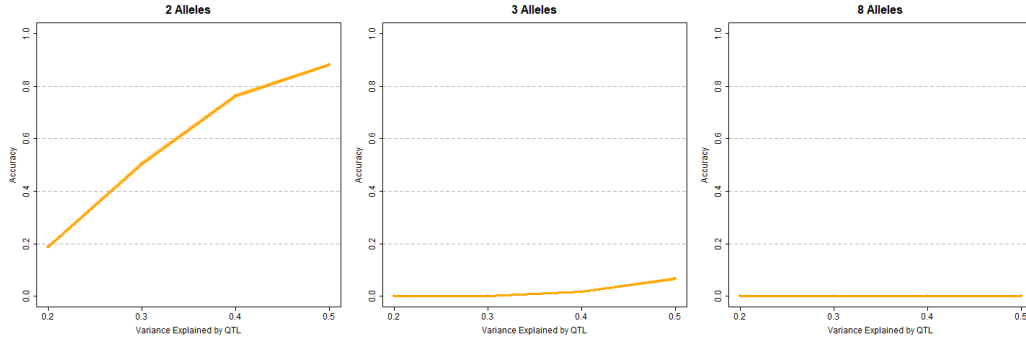
Figure 6.2: 0-1 accuracy of posterior allelic series inference by QTL effect size using the CRP approach, for varying numbers of true functional alleles.

Figure 6.3 shows QTL mapping power using the FE-ROP, RE-ROP, Full, and CRP approaches for balanced and imbalanced biallelic series. As observed in the first chapter, power is lower for all approaches when the allelic series is imbalanced. The CRP approach is the most powerful approach for all biallelic series, but especially when the allelic series is imbalanced.
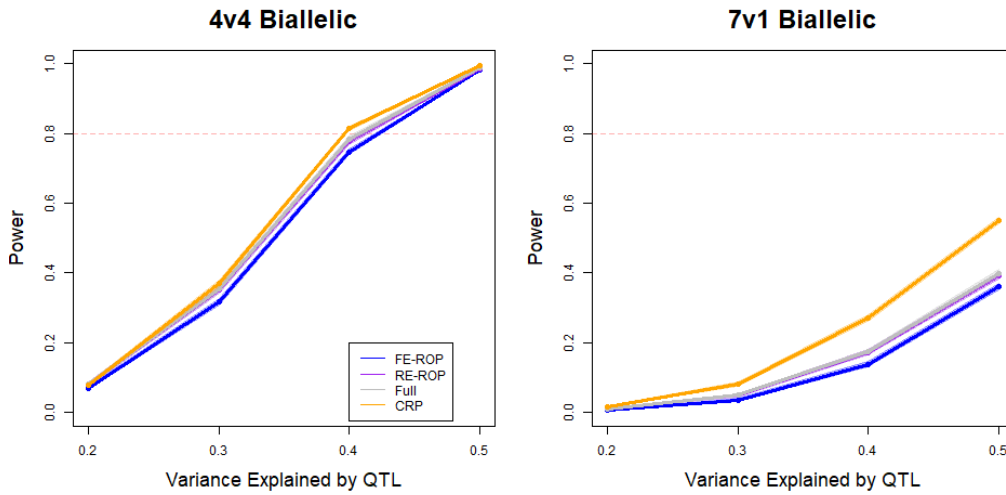
Figure 6.3: QTL mapping power by QTL effect size using the FE-ROP, RE-ROP, Full, and CRP approaches, for balanced and imbalanced biallelic series.

6.4 Discussion

In this chapter, we used simulations to evaluate various haplotype- and allele-based association approaches for their power in QTL mapping. We found that the allele-based CRP outperforms haplotype-based approaches when there are only two functional alleles. This is accompanied by

highly accurate allelic series inference when the power is also high. The CRP has low posterior allelic series accuracy when there are three functional alleles, but it still has marginally higher power than the other haplotype-based alternatives in this case. The haplotype-based approaches are more powerful than the allele-based approach when there are eight functional alleles. These QTL mapping results are broadly consistent with the single-locus results in **Chapter 4**.

We conclude that the allele-based association approach improves on existing haplotype-based approaches for QTL mapping, provided that there are only a few functional alleles at the locus. This improvement is marginal, however, and comes with a high computational cost. If computation is prohibitive, we recommend using the much faster RE-ROP random effect approach, which outperforms the similarly-fast FE-ROP fixed effect approach.

There may be situations which justify the increased computation of allele-based association approach. In particular, we found that the allele-based CRP substantially increases power in the case of imbalanced biallelic series. This is likely due to the "rich-get-richer" property of the CRP, which favors contrasts that are imbalanced *a priori* (Wallach et al., 2008). If genetic differences are driven by only one or a few founder haplotypes, as with the wild-derived founder strains of the CC and Diversity Outbred (DO) (Yang et al., 2011), we anticipate that our allele-based association approach would be more powerful than haplotype-based approaches.

CHAPTER 7

Conclusion

7.1 Overview

In this chapter, we summarize the previous chapters and comment on future directions for allele-based association approaches in MPPs. We consider phylogeny-informed QTL mapping using our approach, and more generally, we discuss the challenge of inferring local phylogeny with recombination. We also discuss how our allelic series model may be useful for identifying candidate causal variants. Finally, we comment on some of the limitations of our allele-based association approach and how these could be addressed.

7.2 Allelic series inference improves effect estimation despite posterior uncertainty

Our single-locus simulations in **Chapter 4** indicate that inference of the allelic series (in the absence of tree information) is generally uncertain, even in situations that we expect would have high QTL mapping power (the focus of **Chapter 2**). Posterior certainty is higher in the biallelic case, when there are relatively more observations to distinguish the allele effects. In combination with a prior allelic series that expects few functional alleles, the posterior allelic series is decisive in this case. Posterior certainty on the correct allelic series decreases, however, as the true number of functional alleles increases. This is because the space of possible allelic series configurations is larger (for an intermediate number of alleles), and there are relatively fewer observations to distinguish the allele effects. For these reasons, the allele-based association approach may be more useful for evaluating whether a QTL is more likely to be biallelic or multiallelic, rather than for identifying the allelic series *per se*. Our approach can also be used to determine the most-highly multiallelic QTL in a

dataset, as we did for eQTL (Kelada et al., 2014) in **Chapter 5**. Characterizing highly multiallelic QTL is an interesting topic for further investigation.

Despite uncertainty in the allelic series, the allele-based approach does improve haplotype effect estimation relative to the haplotype-based approach, provided there are only a few functional alleles at the locus. The allele-based approach allows the data to be represented using fewer parameters, and this reduction in parameters can still be beneficial even when the allelic series is only partially known. This improvement in effect estimation was particularly evident in the "biallelic" DSPR example (King et al., 2014) in **Chapter 5**. These results suggest that the allele-based approach will be useful in the context of phenotype prediction, or other applications that might benefit from improved effect estimation.

7.3 Allele-based approach can improve QTL mapping but is computationally intensive

Our simulated QTL mapping experiments in **Chapter 6** show that the tree-naive allele-based association approach improves QTL mapping power relative to the haplotype-based approach when the number of functional alleles is small. This is particularly true for imbalanced biallelic series, addressing a weakness of the haplotype-based approach identified in **Chapter 2**. In general, however, the improvement in power for the allele-based approach is modest, and it comes with a high computational cost. We expect that this cost will typically not be justified, except perhaps when imbalanced allelic series are anticipated *a priori*, or when additional phylogenetic information about the founder haplotypes can be included, which we discuss in the next section.

It may be reasonable to expect imbalanced biallelic series if genetic differences are driven by only one or a few founder haplotypes, as with the wild-derived founder strains of the Collaborative Cross (CC) and Diversity Outbred (DO) (Yang et al., 2011). If this is the case, it may be beneficial to consider the single-variant mapping approach as an alternative to the less-powerful haplotype-based approach and the much-slower allele-based approach, though this ignores the possibility of multiallelic, epistatic effects. Directly comparing the single-variant approach with the haplotype- and allele-based approaches would be a valuable direction for future research.

7.4 Local phylogeny improves inference but is uncertain

Accounting for founder haplotype relatedness in an allele-based association framework is the primary innovation of our research. The single-locus simulations in **Chapter 4** show that including prior information about haplotype relatedness, in the form of a coalescent tree, improves our allele-based association approach with respect to allelic series inference, haplotype effect estimation and statistical signal, although less so with reduced power. In combination with the tree-naive simulated QTL mapping experiments in **Chapter 6**, these results suggest that QTL mapping power could potentially be improved if local phylogeny was known throughout the genome.

The phylogenetic tree of the founder haplotypes is necessarily unknown, however, and can only be observed indirectly through genetic variation. Our framework is based on the coalescent (Kingman, 1982), which describes the phylogenetic relationship for a single nonrecombinant genomic region. The assumption of no recombination is necessary because, in recombinant systems, phylogeny can vary throughout the genome due to incomplete lineage sorting (Degnan and Rosenberg, 2009) and (particularly for the CC and DO founder strains) introgression (Yang et al., 2011; Didion and De Villena, 2013). This means that neighboring genomic regions can have distinct (but correlated) phylogenetic trees. The complex structure describing recombination events and varied local phylogeny is the "ancestral recombination graph" (ARG), and inferring it is the subject of active research (Kelleher et al., 2019; Rasmussen et al., 2014b). If the ARG were known exactly, variation in haplotype phylogeny throughout the genome could be a useful source of information for QTL mapping. In practice, though, the ARG will be uncertain, with regions that are poorly informed by mutations or biased due to errors. Due to this uncertainty, the inferred ARG will be less useful for QTL mapping than known local phylogeny, although it is unclear to what extent.

We did not consider haplotype phylogeny during the simulated QTL mapping experiments in **Chapter 6**. The simulation procedure, established in **Chapter 2**, uses real data from the CC. We did not want to make the strong assumption that the local phylogeny of the CC founders is known at every locus, described perfectly by an inferred ARG. Instead, we considered known phylogenetic trees in the single-locus simulations in **Chapter 4**. We found only small improvements in statistical signal from known trees in the lower power scenario, which is the situation where additional prior information would be most beneficial. We anticipate that any improvement in QTL

mapping power from local phylogenetic information is unlikely to outweigh the high computational cost that accompanies the allele-based approach.

Given uncertainty in the ARG, we recommend using our tree-naive approach by default when analyzing QTL in recombinant MPPs. There are situations, though, when local phylogeny can be accurately inferred, and in these cases, including tree information improves allelic series inference and haplotype effect estimation. We demonstrated this in **Chapter 5**, for a QTL with a known causal gene (Kelada et al., 2012), in combination with fully-Bayesian tree inference software (Drummond et al., 2012). We anticipate our tree-informed approach will be useful in haploid systems, as in Azim Ansari and Didelot (2016) and Cybis et al. (2018), because most haploids do not recombine, and thus have a single phylogenetic history for their entire genome.

7.5   Connecting the allelic series to causal variants

The allele-based association approach is useful for evaluating whether a QTL is more likely to be biallelic or multiallelic. It can be difficult, however, to connect information about the allelic series to causal variants. Evaluating evidence in favor of a single biallelic variant is straightforward, as our framework encompasses a fully-Bayesian implementation of merge analysis (Yalcin et al., 2005).

Multiallelic series are more challenging. When a QTL is multiallelic, we expect that multiple causal variants distinguish the haplotypes into functional alleles. Our allele-based association approach only considers haplotype effects for a single genomic interval (i.e. the diplotype state probabilities do not vary in this region). Thus, it implicitly assumes that all causal variants are on the same genomic interval. For this reason, results from our allele-based approach cannot strictly be used to evaluate combinations of variants from different (even adjacent) genomic intervals.

An alternative in our framework assumes that the allelic series is at least as complicated as a given biallelic variant. In this case, the prior distribution of the allelic series is restricted to exclude partitions that violate the functional distinctions given by a causal variant. For example, for $J = 3$ haplotypes and a causal biallelic variant that contrasts haplotypes A and C with B, the prior distribution for the allelic series is

$$p(\mathbf{M}|\mathbf{m}_1 \neq \mathbf{m}_2, \mathbf{m}_2 \neq \mathbf{m}_3) \propto p(\mathbf{m}_1 \neq \mathbf{m}_2, \mathbf{m}_2 \neq \mathbf{m}_3|\mathbf{M})p(\mathbf{M}).$$

The first term on the right-hand side is an indicator variable denoting whether the allelic series $\mathbf{M}$ satisfies the conditions given by the biallelic variant, $\mathbf{m}_1 \neq \mathbf{m}_2$ and $\mathbf{m}_2 \neq \mathbf{m}_3$. The second term is the prior distribution of the allelic series (marginalized over the concentration parameter). Using this, we can compute a "variant-consistent" prior distribution that allows for multiallelic effects, but only in combination with the causal variant under consideration (and implicitly, only other variants on the same interval, in proportion to the prior). This variant-consistent approach may be more useful than single-variant merge analysis for identifying candidate causal variants at multiallelic QTL. We have implemented this variant-consistent prior distribution for the allelic series in our R package 'TIMBR' (`https://github.com/wcrouse/TIMBR`), and evaluating it would be an interesting topic for future research.

7.6   Limitations of the allele-based approach

The allele-based association approach is limited by its computational speed. Our fully-Bayesian approach uses Gibbs sampling for posterior inference, which requires drawing many samples, at every locus, for every prior hypothesis. This limits the practical usefulness of the allele-based approach for mapping QTL and identifying causal variants. One solution could be approximate *maximum a posteriori* (MAP) inference (Raykov et al., 2016). MAP inference avoids sampling and would be considerably faster than full posterior inference, though presumably with reduced performance.

Our method for calculating the tree-informed allelic series prior distribution is also computationally expensive. This is because it involves precomputing the prior probability of all 2^{2J-2} possible configurations of branch mutations $\mathbf{b}$ on a tree T, recording the implied allelic series $\mathbf{M}$ for each. This approach is feasible for $J = 8$ founder haplotypes, the case for many MPPs, but not for $J = 15$, as in the DSPR. When J is large, it may be preferable to include the branch mutations $\mathbf{b}$ in the posterior sampling procedure, as in Azim Ansari and Didelot (2016), rather than integrating over them to compute the prior distribution. This will require mixing over the larger space of branch mutations $\mathbf{b}$, though, rather than the smaller space of allelic series $\mathbf{M}$. Another alternative is to disregard the tree structure, and instead use patristic distances between haplotypes as input for a distance dependent CRP (Blei and Frazier, 2009), as in Cybis et al. (2018). It would be interesting to

compare results from a distance dependent CRP with the tree-informed CRP that we have defined here.

Finally, the allele-based association approach only considers additive allele effects and unstructured error. As discussed in (Jannink and Wu, 2003), it would be possible to include effects for allelic dominance in our model, though it would be desirable to include these as an additional variance component, as in (Zhang et al., 2014). We did not consider error due to population structure in genetic background, which could also be included as an additional variance component (Eskin et al., 2008; Kang et al., 2010; Lippert et al., 2011; Zhou and Stephens, 2012). Adding model complexity via dominance effects or population structure may be useful, but it would increase the computational requirements of the allele-based association approach.

BIBLIOGRAPHY

Abramowitz, M. and Stegun, I. (1972). *Handbook of Mathematical Functions with Formulas and Mathematical Tables*. Courier Dover Publications.

Alberts, R., Terpstra, P., Li, Y., Breitling, R., Nap, J. P., and Jansen, R. C. (2007). Sequence polymorphisms cause many false cis eQTLs. *PLoS ONE*, 2(7).

Aylor, D. L., Valdar, W., Foulds-Mathes, W., Buus, R. J., Verdugo, R. a., Baric, R. S., Ferris, M. T., Frelinger, J. a., Heise, M., Frieman, M. B., Gralinski, L. E., Bell, T. a., Didion, J. D., Hua, K., Nehrenberg, D. L., Powell, C. L., Steigerwalt, J., Xie, Y., Kelada, S. N. P., Collins, F. S., Yang, I. V., Schwartz, D. a., Branstetter, L. a., Chesler, E. J., Miller, D. R., Spence, J., Liu, E. Y., McMillan, L., Sarkar, A., Wang, J., Wang, W., Zhang, Q., Broman, K. W., Korstanje, R., Durrant, C., Mott, R., Iraqi, F. a., Pomp, D., Threadgill, D., De Villena, F. P. M., and Churchill, G. a. (2011). Genetic analysis of complex traits in the emerging Collaborative Cross. *Genome Research*, 21(8):1213–1222.

Ayroles, J. F., Buchanan, S. M., O'Leary, C., Skutt-Kakaria, K., Grenier, J. K., Clark, A. G., Hartl, D. L., and de Bivort, B. L. (2015). Behavioral idiosyncrasy reveals genetic control of phenotypic variability. *Proceedings of the National Academy of Sciences*, 112(21):6706–6711.

Azim Ansari, M. and Didelot, X. (2016). Bayesian inference of the evolution of a phenotype distribution on a phylogenetic tree. *Genetics*, 204(1):89–98.

Baud, A., Hermsen, R., Guryev, V., Stridh, P., Graham, D., McBride, M. W., Foroud, T., Calderari, S., Diez, M., Ockinger, J., Beyeen, A. D., Gillett, A., Abdelmagid, N., Guerreiro-Cacais, A. O., Jagodic, M., Tuncel, J., Norin, U., Beattie, E., Huynh, N., Miller, W. H., Koller, D. L., Alam, I., Falak, S., Osborne-Pellegrin, M., Martinez-Membrives, E., Canete, T., Blazquez, G., Vicens-Costa, E., Mont-Cardona, C., Diaz-Moran, S., Tobena, A., Hummel, O., Zelenika, D., Saar, K., Patone, G., Bauerfeind, A., Bihoreau, M.-T., Heinig, M., Lee, Y.-A., Rintisch, C., Schulz, H., Wheeler, D. A., Worley, K. C., Muzny, D. M., Gibbs, R. A., Lathrop, M., Lansu, N., Toonen, P., Ruzius, F. P., de Bruijn, E., Hauser, H., Adams, D. J., Keane, T., Atanur, S. S., Aitman, T. J., Flicek, P., Malinauskas, T., Jones, E. Y., Ekman, D., Lopez-Aumatell, R., Dominiczak, A. F., Johannesson, M., Holmdahl, R., Olsson, T., Gauguier, D., Hubner, N., Fernandez-Teruel, A., Cuppen, E., Mott, R., and Flint, J. (2013). Combined sequence-based and genetic mapping analysis of complex traits in outbred rats. *Nature Genetics*, 45(7):767–775.

Beavis, W. D. (1994). The power and deceit of QTL experiments: lessons from comparative QTL studies. In *Proceedings of the forty-ninth annual corn and sorghum industry research conference*, pages 250–266. Washington, DC.

Belknap, J. K. (1998). Effect of within-strain sample size on GTL detection and mapping using recombinant inbred mouse strains. *Behavior Genetics*, 28(1):29–38.

Belknap, J. K., Mitchell, S. R., O'Toole, L. A., Helms, M. L., and Crabbe, J. C. (1996). Type I and type II error rates for quantitative trait loci (QTL) mapping studies using recombinant inbred mouse strains. *Behavior genetics*, 26(2):149–160.

Berestycki, N. (2009). Recent progress in coalescent theory. *Ensaios Matematicos*, 16:1–193.

Beyer, P., Morell, M., Mackay, I., and Powell, W. (2008). From mutations to MAGIC: resources for gene discovery, validation and delivery in crop plants.

Blackwell, D. (2007). Conditional Expectation and Unbiased Sequential Estimation. *The Annals of Mathematical Statistics*, 18(1):105–110.

Blei, D. M. and Frazier, P. I. (2009). Distance Dependent Chinese Restaurant Processes. *Journal of Machine Learning Research*, 12:2461–2488.

Bouchet, S., Olatoye, M. O., Marla, S. R., Perumal, R., Tesso, T., Yu, J., Tuinstra, M., and Morris, G. P. (2017). Increased Power To Dissect Adaptive Traits in Global Sorghum Diversity Using a Nested Association Mapping Population. *Genetics*, 206(June):573–585.

Bouckaert, R. and Heled, J. (2014). DensiTree 2: Seeing trees through the forest. *bioRxiv*, pages 1–11.

Broman, K. W., Gatti, D. M., Simecek, P., Furlotte, N. A., Prins, P., Sen, Ś., Yandell, B. S., and Churchill, G. A. (2019). R/qtl2: Software for Mapping Quantitative Trait Loci with High-Dimensional Data and Multiparent Populations. *Genetics*, 211(2):495–502.

Brown, L. D., Cai, T. T., and DasGupta, A. (2001). Interval Estimation for a Binomial Proportion. *Statistical Science*, 16(2):101–117.

Bult, C. J., Blake, J. A., Smith, C. L., Kadin, J. A., Richardson, J. E., Anagnostopoulos, A., Asabor, R., Baldarelli, R. M., Beal, J. S., Bello, S. M., Blodgett, O., Butler, N. E., Christie, K. R., Corbani, L. E., Creelman, J., Dolan, M. E., Drabkin, H. J., Giannatto, S. L., Hale, P., Hill, D. P., Law, M., Mendoza, A., McAndrews, M., Miers, D., Motenko, H., Ni, L., Onda, H., Perry, M., Recla, J. M., Richards-Smith, B., Sitnikov, D., Tomczuk, M., Tonorio, G., Wilming, L., and Zhu, Y. (2019). Mouse Genome Database (MGD) 2019. *Nucleic Acids Research*, 47(D1):D801–D806.

Chesler, E. J., Miller, D. R., Branstetter, L. R., Galloway, L. D., Jackson, B. L., Philip, V. M., Voy, B. H., Culiat, C. T., Threadgill, D. W., Williams, R. W., Churchill, G. A., Johnson, D. K., and Manly, K. F. (2008). The Collaborative Cross at Oak Ridge National Laboratory: developing a powerful resource for systems genetics. *Mammalian Genome*, 19(6):382–389.

Chib, S. (1995). Marginal Likelihood from the Gibbs Output. *Journal of the American Statistical Association*, 90(432):1313.

Churchill, G. A., Airey, D. C., Allayee, H., Angel, J. M., Attie, A. D., Beatty, J., Beavis, W. D., Belknap, J. K., Bennett, B., Berrettini, W., Bleich, A., Bogue, M., Broman, K. W., Buck, K. J., Buckler, E., Burmeister, M., Chesler, E. J., Cheverud, J. M., Clapcote, S., Cook, M. N., Cox, R. D., Crabbe, J. C., Crusio, W. E., Darvasi, A., Deschepper, C. F., Doerge, R. W., Farber, C. R., Forejt, J., Gaile, D., Garlow, S. J., Geiger, H., Gershenfeld, H., Gordon, T., Gu, J., Gu, W., de Haan, G., Hayes, N. L., Heller, C., Himmelbauer, H., Hitzemann, R., Hunter, K., Hsu, H. C., Iraqi, F. A., Ivandic, B., Jacob, H. J., Jansen, R. C., Jepsen, K. J., Johnson, D. K., Johnson, T. E., Kempermann, G., Kendziorski, C., Kotb, M., Kooy, R. F., Llamas, B., Lammert, F., Lassalle, J. M., Lowenstein, P. R., Lu, L., Lusis, A., Manly, K. F., Marcucio, R., Matthews, D., Medrano, J. F., Miller, D. R., Mittleman, G., Mock, B. A., Mogil, J. S., Montagutelli, X., Morahan, G., Morris, D. G., Mott, R., Nadeau, J. H., Nagase, H., Nowakowski, R. S., O'Hara, B. F., Osadchuk, A. V., Page, G. P., Paigen, B., Paigen, K., Palmer, A. A., Pan, H. J., Peltonen-Palotie, L., Peirce,

J., Pomp, D., Pravenec, M., Prows, D. R., Qi, Z., Reeves, R. H., Roder, J., Rosen, G. D., Schadt, E. E., Schalkwyk, L. C., Seltzer, Z., Shimomura, K., Shou, S., Sillanpää, M. J., Siracusa, L. D., Snoeck, H. W., Spearow, J. L., Svenson, K., Tarantino, L. M., Threadgill, D., Toth, L. A., Valdar, W., Pardo-Manuel de Villena, F., Warden, C., Whatley, S., Williams, R. W., Wiltshire, T., Yi, N., Zhang, D., Zhang, M., and Zou, F. (2004). The Collaborative Cross, a community resource for the genetic analysis of complex traits. *Nature Genetics*, 36(11):1133–1137.

Collaborative Cross Consortium, Iraqi, F. A., Mahajne, M., Salaymah, Y., Sandovski, H., Tayem, H., Vered, K., Balmer, L., Hall, M., Manship, G., Morahan, G., Pettit, K., Scholten, J., Tweedie, K., Wallace, A., Weerasekera, L., Cleak, J., Durrant, C., Goodstadt, L., Mott, R., Yalcin, B., Aylor, D. L., Baric, R. S., Bell, T. A., Bendt, K. M., Brennan, J., Brooks, J. D., Buus, R. J., Crowley, J. J., Calaway, J. D., Calaway, M. E., Cholka, A., Darr, D. B., Didion, J. P., Dorman, A., Everett, E. T., Ferris, M. T., Mathes, W. F., Fu, C. P., Gooch, T. J., Goodson, S. G., Gralinski, L. E., Hansen, S. D., Heise, M. T., Hoel, J., Hua, K., Kapita, M. C., Lee, S., Lenarcic, A. B., Liu, E. Y., Liu, H., McMillan, L., Magnuson, T. R., Manly, K. F., Miller, D. R., O'Brien, D. A., Odet, F., Pakatci, I. K., Pan, W., de Villena, F. P. M., Perou, C. M., Pomp, D., Quackenbush, C. R., Robinson, N. N., Sharpless, N. E., Shaw, G. D., Spence, J. S., Sullivan, P. F., Sun, W., Tarantino, L. M., Valdar, W., Wang, J., Wang, W., Welsh, C. E., Whitmore, A., Wiltshire, T., Wright, F. A., Xie, Y., Yun, Z., Zhabotynsky, V., Zhang, Z., Zou, F., Powell, C., Steigerwalt, J., Threadgill, D. W., Chesler, E. J., Churchill, G. A., Gatti, D. M., Korstanje, R., Svenson, K. L., Collins, F. S., Crawford, N., Hunter, K., Samir, N., Kelada, P., Peck, B. C., Reilly, K., Tavarez, U., Bottomly, D., Hitzeman, R., Mcweeney, S. K., Frelinger, J., Krovi, H., Phillippi, J., Spritz, R. A., Aicher, L., Katze, M., Rosenzweig, E., Shusterman, A., Nashef, A., Weiss, E. I., Houri-Haddad, Y., Soller, M., Williams, R. W., Schughart, K., Yang, H., French, J. E., Benson, A. K., Kim, J., Legge, R., Low, S. J., Ma, F., Martinez, I., Walter, J., Broman, K. W., Hallgrimsson, B., Klein, O., Weinstock, G., Warren, W. C., Yang, Y. V., and Schwartz, D. (2012). The genome architecture of the Collaborative Cross mouse genetic reference population. *Genetics*, 190(2):389–401.

Collins, F. S. and Varmus, H. (2015). A New Initiative on Precision Medicine. *New England Journal of Medicine*, 372(9):793–795.

Cowen, N. M. (1988). The use of replicated progenies in marker-based mapping of QTL's. *Theoretical and Applied Genetics*, 75(6):857–862.

Crowley, J. J., Kim, Y., Lenarcic, A. B., Quackenbush, C. R., Barrick, C. J., Adkins, D. E., Shaw, G. S., Miller, D. R., de Villena, F. P. M., Sullivan, P. F., and Valdar, W. (2014). Genetics of adverse reactions to haloperidol in a mouse diallel: A drug-placebo experiment and Bayesian causal analysis. *Genetics*, 196(1):321–347.

Crowley, J. J., Zhabotynsky, V., Sun, W., Huang, S., Pakatci, I. K., Kim, Y., Wang, J. R., Morgan, A. P., Calaway, J. D., Aylor, D. L., Yun, Z., Bell, T. A., Buus, R. J., Calaway, M. E., Didion, J. P., Gooch, T. J., Hansen, S. D., Robinson, N. N., Shaw, G. D., Spence, J. S., Quackenbush, C. R., Barrick, C. J., Nonneman, R. J., Kim, K., Xenakis, J., Xie, Y., Valdar, W., Lenarcic, A. B., Wang, W., Welsh, C. E., Fu, C.-P., Zhang, Z., Holt, J., Guo, Z., Threadgill, D. W., Tarantino, L. M., Miller, D. R., Zou, F., McMillan, L., Sullivan, P. F., and Pardo-Manuel de Villena, F. (2015). Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance. *Nature Genetics*, 47(4):353–360.

Cybis, G. B., Sinsheimer, J. S., Bedford, T., Rambaut, A., Lemey, P., and Suchard, M. A. (2018). Bayesian nonparametric clustering in phylogenetics: modeling antigenic evolution in influenza. *Statistics in Medicine*, 37(2):195–206.

Davies, R. W., Flint, J., Myers, S., and Mott, R. (2016). Rapid genotype imputation from sequence without reference panels. *Nature Genetics*, 48(8):965–969.

Degnan, J. H. and Rosenberg, N. A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology and Evolution*, 24(6):332–340.

Dell'Acqua, M., Gatti, D. M., Pea, G., Cattonaro, F., Coppens, F., Magris, G., Hlaing, A. L., Aung, H. H., Nelissen, H., Baute, J., Frascaroli, E., Churchill, G. A., Inzé, D., Morgante, M., and Pè, M. E. (2015). Genetic properties of the MAGIC maize population: a new platform for high definition QTL mapping in Zea mays. *Genome Biology*, 16(1):167.

Didion, J. P. and De Villena, F. P. M. (2013). Deconstructing Mus gemischus: Advances in understanding ancestry, structure, and variation in the genome of the laboratory mouse. *Mammalian Genome*, 24(1-2):1–20.

Doerge, R. W. and Churchill, G. A. (1996). Permutation tests for multiple loci affecting a quantitative character. *Genetics*, 142(1977):285–294.

Donoghue, L. J., Livraghi-Butrico, A., McFadden, K. M., Thomas, J. M., Chen, G., Grubb, B. R., O'Neal, W. K., Boucher, R. C., and Kelada, S. N. P. (2017). Identification of trans Protein QTL for Secreted Airway Mucins in Mice and a Causal Role for Bpifb1. *Genetics*, 207(2):801–812.

Druet, T. and Georges, M. (2010). A hidden Markov model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping. *Genetics*, 184(3):789–798.

Drummond, A. J., Suchard, M. A., Xie, D., and Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*, 29(8):1969–1973.

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. *The American Journal of Human Genetics*, 75(3):424–435.

Durrant, C. and Mott, R. (2010). Bayesian quantitative trait locus mapping using inferred haplotypes. *Genetics*, 184(3):839–852.

Durrant, C., Tayem, H., Yalcin, B., Cleak, J., Goodstadt, L., de Villena, F. P.-M., Mott, R., and Iraqi, F. A. (2011). Collaborative Cross mice and their power to map host susceptibility to Aspergillus fumigatus infection. *Genome Research*, 21(8):1239–1248.

Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H., and Nadeau, J. H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics*, 11(6):446–450.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588.

Eskin, E., Wade, C. M., Daly, M. J., Heckerman, D., Zaitlen, N. A., Kang, H. M., and Kirby, A. (2008). Efficient Control of Population Structure in Model Organism Association Mapping. *Genetics*, 178(3):1709–1723.

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology*, 3(1):87–112.

Falke, K. C. and Frisch, M. (2011). Power and false-positive rate in QTL detection with near-isogenic line libraries. *Heredity*, 106(4):576–584.

Ferris, M. T., Aylor, D. L., Bottomly, D., Whitmore, A. C., Aicher, L. D., Bell, T. A., Bradel-Tretheway, B., Bryan, J. T., Buus, R. J., Gralinski, L. E., Haagmans, B. L., McMillan, L., Miller, D. R., Rosenzweig, E., Valdar, W., Wang, J., Churchill, G. A., Threadgill, D. W., McWeeney, S. K., Katze, M. G., Pardo-Manuel de Villena, F., Baric, R. S., and Heise, M. T. (2013). Modeling Host Genetic Regulation of Influenza Pathogenesis in the Collaborative Cross. *PLoS Pathogens*, 9(2):e1003196.

Festing, M. F. W. (2010). Inbred strains should replace outbred stocks in toxicology, safety testing, and drug development. *Toxicologic Pathology*, 38(5):681–690.

Flint, J. and Mackay, T. F. C. (2009). Genetic architecture of quantitative traits in mice, flies, and humans. *Genome Research*, 19(5):723–733.

Fu, C.-P., Welsh, C. E., de Villena, F. P.-M., and McMillan, L. (2012). Inferring ancestry in admixed populations using microarray probe intensities. In *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine - BCB '12*, pages 105–112, New York, New York, USA. ACM Press.

Gatti, D. M., Svenson, K. L., Shabalin, A., Wu, L.-Y., Valdar, W., Simecek, P., Goodwin, N., Cheng, R., Pomp, D., Palmer, A., Chesler, E. J., Broman, K. W., and Churchill, G. A. (2014). Quantitative Trait Locus Mapping Methods for Diversity Outbred Mice. *G3: Genes, Genomes, Genetics*, 4(9):1623–1633.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–533.

Graham, J. B., Swarts, J. L., Mooney, M., Choonoo, G., Jeng, S., Miller, D. R., Ferris, M. T., McWeeney, S., and Lund, J. M. (2017). Extensive Homeostatic T Cell Phenotypic Variation within the Collaborative Cross. *Cell Reports*, 21(8):2313–2325.

Gralinski, L. E., Ferris, M. T., Aylor, D. L., Whitmore, A. C., Green, R., Frieman, M. B., Deming, D., Menachery, V. D., Miller, D. R., Buus, R. J., Bell, T. A., Churchill, G. A., Threadgill, D. W., Katze, M. G., McMillan, L., Valdar, W., Heise, M. T., Pardo-Manuel de Villena, F., and Baric, R. S. (2015). Genome Wide Identification of SARS-CoV Susceptibility Loci Using the Collaborative Cross. *PLoS genetics*, 11(10):e1005504.

Haley, C. S. and Knott, S. A. (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity*, 69(4):315–324.

Hamazaki, K. and Iwata, H. (2019). Haplotype-based genome wide association study using a novel SNP-set method : RAINBOW. *bioRxiv*, page 612028.

Hamblin, M. T. and Jannink, J.-L. (2011). Factors Affecting the Power of Haplotype Markers in Association Studies. *The Plant Genome Journal*, 4(2):145.

Hasegawa, M., Kishino, H., and aki Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22(2):160–174.

Huang, B. E., Verbyla, K. L., Verbyla, A. P., Raghavan, C., Singh, V. K., Gaur, P., Leung, H., Varshney, R. K., and Cavanagh, C. R. (2015). MAGIC populations in crops: current status and future prospects. *Theoretical and Applied Genetics*, 128(6):999–1017.

Hudson, R. R. and Kaplan, N. L. (1985). Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*, 111(1):147–164.

Jannink, J.-L. and Wu, X.-L. (2003). Estimating allelic number and identity in state of QTLs in interconnected families. *Genetical Research*, 81(2):133–44.

Kaeppler, S. M. (1997). Quantitative trait locus mapping using sets of near-isogenic lines: Relative power comparisons and technical considerations. *Theoretical and Applied Genetics*, 95(3):384–392.

Kafkafi, N., Agassi, J., Chesler, E. J., Crabbe, J. C., Crusio, W. E., Eilam, D., Gerlai, R., Golani, I., Gomez-Marin, A., Heller, R., Iraqi, F., Jaljuli, I., Karp, N. A., Morgan, H., Nicholson, G., Pfaff, D. W., Richter, S. H., Stark, P. B., Stiedl, O., Stodden, V., Tarantino, L. M., Tucci, V., Valdar, W., Williams, R. W., Würbel, H., and Benjamini, Y. (2018). Reproducibility and replicability of rodent phenotyping in preclinical studies. *Neuroscience and Biobehavioral Reviews*, 87(October 2016):218–232.

Kafkafi, N., Benjamini, Y., Sakov, A., Elmer, G. I., and Golani, I. (2005). Genotype-environment interactions in mouse behavior: a way out of the problem. *Proceedings of the National Academy of Sciences of the United States of America*, 102(12):4619–4624.

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S.-Y., Freimer, N. B., Sabatti, C., and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, 42(4):348–354.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

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.

Keele, G. R., Crouse, W. L., Kelada, S. N. P., and Valdar, W. (2019). Determinants of QTL Mapping Power in the Realized Collaborative Cross. *G3: Genes, Genomes, Genetics*, 9(5):1707–1727.

Keele, G. R., Prokop, J. W., He, H., Holl, K., Littrell, J., Deal, A., Francic, S., Cui, L., Gatti, D. M., Broman, K. W., Tschannen, M., Tsaih, S.-W., Zagloul, M., Kim, Y., Baur, B., Fox, J., Robinson, M., Levy, S., Flister, M. J., Mott, R., Valdar, W., and Solberg Woods, L. C. (2018). Genetic Fine-Mapping and Identification of Candidate Genes and Variants for Adiposity Traits in Outbred Rats. *Obesity*, 26(1):213–222.

Kelada, D. S. N. P., Carpenter, M. D. E., Aylor, D. D. L., Chines, M. P., Rutledge, M. H., Chesler, D. E. a., Churchill, D. G. a., Villena, D. F. P.-M. D., Schwartz, D. D. a., and Collins, D. F. S.

(2014). Integrative Genetics of Allergic Inflammation in the Murine Lung. *American Journal of Respiratory Cell and Molecular Biology*, 51(3):436–445.

Kelada, S. N. P. (2016). Plethysmography Phenotype QTL in Mice Before and After Allergen Sensitization and Challenge. *G3: Genes, Genomes, Genetics*, 6(9):2857–2865.

Kelada, S. N. P., Aylor, D. L., Peck, B. C. E., Ryan, J. F., Tavarez, U., Buus, R. J., Miller, D. R., Chesler, E. J., Threadgill, D. W., Churchill, G. A., Pardo-Manuel de Villena, F., and Collins, F. S. (2012). Genetic Analysis of Hematological Parameters in Incipient Lines of the Collaborative Cross. *G3: Genes, Genomes, Genetics*, 2(2):157–165.

Kelleher, J., Wong, Y., Wohns, A. W., Fadil, C., Albers, P. K., and McVean, G. (2019). Inferring whole-genome histories in large population datasets. *Nature Genetics*, 51(9):1330–1338.

King, E. G. and Long, A. D. (2017). The Beavis Effect in Next-Generation Mapping Panels in Drosophila melanogaster. *G3: Genes, Genomes, Genetics*, 7(6):1643 LP —- 1652.

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.

King, E. G., Sanderson, B. J., McNeil, C. L., Long, A. D., and Macdonald, S. J. (2014). Genetic dissection of the Drosophila melanogaster female head transcriptome reveals widespread allelic heterogeneity. *PLoS Genetics*, 10(5):e1004322.

Kingman, J. F. C. (1982). On the genealogy of large populations. *Journal of Applied Probability*, 19(A):27–43.

Kingman, J. F. C. (2006). Random partitions in population genetics. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 201(1143):217–217.

Kislukhin, G., King, E. G., Walters, K. N., Macdonald, S. J., and Long, A. D. (2013). The Genetic Architecture of Methotrexate Toxicity Is Similar in Drosophila melanogaster and Humans. *G3: Genes, Genomes, Genetics*, 3(8):1301–1310.

Klasen, J. R., Piepho, H. P., and Stich, B. (2012). QTL detection power of multi-parental RIL populations in Arabidopsis thaliana. *Heredity*, 108(6):626–632.

Knapp, S. J. and Bridges, W. C. (1990). Using molecular markers to estimate quantitative trait locus parameters: power and genetic variances for unreplicated and replicated progeny. *Genetics*, 126(3):769–777.

Kover, P. X., Valdar, W., Trakalo, J., Scarcelli, N., Ehrenreich, I. M., Purugganan, M. D., Durrant, C., and Mott, R. (2009). A multiparent advanced generation inter-cross to fine-map quantitative traits in Arabidopsis thaliana. *PLoS Genetics*, 5(7):e1000551.

Lander, E. S. and Botstein, D. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, 121(1):185–199.

Levy, R., Mott, R. F., Iraqi, F. A., and Gabet, Y. (2015). Collaborative cross mice in a genetic association study reveal new candidate genes for bone microarchitecture. *BMC Genomics*, 16(1):1013.

Li, H., Bradbury, P., Ersoz, E., Buckler, E. S., and Wang, J. (2011). Joint QTL linkage mapping for multiple-cross mating design sharing one common parent. *PloS ONE*, 6(3):e17573.

Lilue, J., Shivalikanjli, A., Adams, D. J., and Keane, T. M. (2019). Mouse protein coding diversity: What's left to discover? *PLOS Genetics*, 15(11):e1008446.

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. *Nature Methods*, 8(10):833–837.

Liu, E. Y., Zhang, Q., McMillan, L., de Villena, F. P.-M., and Wang, W. (2010). Efficient genome ancestry inference in complex pedigrees with inbreeding. *Bioinformatics*, 26(12):i199—-i207.

Lorè, N. I., Iraqi, F. A., and Bragonzi, A. (2015). Host genetic diversity influences the severity of Pseudomonas aeruginosa pneumonia in the Collaborative Cross mice. *BMC Genetics*, 16(1):106.

Lynch, M. and Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Sunderland, MA.

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 Research*, 45(D1):D896–D901.

Macdonald, S. J. and Long, A. D. (2007). Joint estimates of quantitative trait locus effect and frequency using synthetic recombinant populations of Drosophila melanogaster. *Genetics*, 176(2):1261–1281.

Mackay, T. F. C., 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. H., 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.

Madian, A. G., Wheeler, H. E., Jones, R. B., and Dolan, M. E. (2012). Relating human genetic variation to variation in drug responses. *Trends in Genetics*, 28(10):487–495.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., MacKay, T. F., McCarroll, S. A., and Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753.

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.

Mathes, W. F., Aylor, D. L., Miller, D. R., Churchill, G. A., Chesler, E. J., de Villena, F. P.-M., Threadgill, D. W., and Pomp, D. (2011). Architecture of energy balance traits in emerging lines

of the Collaborative Cross. *American Journal of Physiology-Endocrinology and Metabolism*, 300(6):E1124—-E1134.

McClurg, P., Janes, J., Wu, C., Delano, D. L., Walker, J. R., Batalov, S., Takahashi, J. S., Shimomura, K., Kohsaka, A., Bass, J., Wiltshire, T., and Su, A. I. (2007). Genomewide association analysis in diverse inbred mice: Power and population structure. *Genetics*, 176(1):675–683.

Meuwissen, T. H., Odegard, J., Andersen-Ranberg, I., and Grindflek, E. (2014). On the distance of genetic relationships and the accuracy of genomic prediction in pig breeding. *Genetics Selection Evolution*, 46(1).

Molenhuis, R. T., Bruining, H., Brandt, M. J. V., van Soldt, P. E., Abu-Toamih Atamni, H. J., Burbach, J. P. H., Iraqi, F. A., Mott, R. F., and Kas, M. J. H. (2018). Modeling the quantitative nature of neurodevelopmental disorders using Collaborative Cross mice. *Molecular Autism*, 9(1):63.

Morgan, A. P., Fu, C.-P., Kao, C.-Y., Welsh, C. E., Didion, J. P., Yadgary, L., Hyacinth, L., Ferris, M. T., Bell, T. A., Miller, D. R., Giusti-Rodriguez, P., Nonneman, R. J., Cook, K. D., Whitmire, J. K., Gralinski, L. E., Keller, M., Attie, A. D., Churchill, G. A., Petkov, P., Sullivan, P. F., Brennan, J. R., McMillan, L., and Pardo-Manuel de Villena, F. (2016). The Mouse Universal Genotyping Array: From Substrains to Subspecies. *G3: Genes, Genomes, Genetics*, 6(2):263–279.

Morgan, A. P. and Welsh, C. E. (2015). Informatics resources for the Collaborative Cross and related mouse populations. *Mammalian Genome*, 26(9-10):521–539.

Mosedale, M., Kim, Y., Brock, W. J., Roth, S. E., Wiltshire, T., Eaddy, J. S., Keele, G. R., Corty, R. W., Xie, Y., Valdar, W., and Watkins, P. B. (2017). Candidate risk factors and mechanisms for tolvaptan-induced liver injury are identified using a collaborative cross approach. *Toxicological Sciences*, 156(2):438–454.

Mott, R., Talbot, C. J., Turri, M. G., Collins, A. C., and Flint, J. (2000). A method for fine mapping quantitative trait loci in outbred animal stocks. *PNAS*, 97(23):12649–12654.

Müller, P., Quintana, F. A., Jara, A., and Hanson, T. (2015). *Bayesian Nonparametric Data Analysis*. Springer Series in Statistics. Springer International Publishing, Cham.

Najarro, M. A., Hackett, J. L., Smith, B. R., Highfill, C. A., King, E. G., Long, A. D., and Macdonald, S. J. (2015). Identifying Loci Contributing to Natural Variation in Xenobiotic Resistance in Drosophila. *PLoS Genetics*, 11(11):1–25.

Neal, R. M. (2000). Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, 9(2):249–265.

Noble, L. M., Chelo, I., Guzella, T., Afonso, B., Riccardi, D. D., Ammerman, P., Dayarian, A., Carvalho, S., Crist, A., Pino-Querido, A., Shraiman, B., Rockman, M. V., and Teotónio, H. (2017). Polygenicity and Epistasis Underlie Fitness-Proximal Traits in the Caenorhabditis elegans Multiparental Experimental Evolution (CeMEE) Panel. *Genetics*, 207(4):1663–1685.

Orgel, K., Smeekens, J. M., Ye, P., Fotsch, L., Guo, R., Miller, D. R., Pardo-Manuel de Villena, F., Burks, A. W., Ferris, M. T., and Kulis, M. D. (2019). Genetic diversity between mouse strains allows identification of the CC027/GeniUnc strain as an orally reactive model of peanut allergy. *The Journal of Allergy and Clinical Immunology*, 143(3):1027—-1037.e7.

Park, T. and Van Dyk, D. A. (2009). Partially collapsed Gibbs samplers: Illustrations and applications. *Journal of Computational and Graphical Statistics*, 18(2):283–305.

Peirce, J. L., Lu, L., Gu, J., Silver, L. M., and Williams, R. W. (2004). A new set of BXD recombinant inbred lines from advanced intercross populations in mice. *BMC Genetics*, 5(7):7.

Pfaff, B. and McNeil, A. (2018). *evir: Extreme Values in R*.

Philip, V. M., Sokoloff, G., Ackert-Bicknell, C. L., Striz, M., Branstetter, L., Beckmann, M. a., Spence, J. S., Jackson, B. L., Galloway, L. D., Barker, P., Wymore, A. M., Hunsicker, P. R., Durtschi, D. C., Shaw, G. S., Shinpock, S., Manly, K. F., Miller, D. R., Donohue, K. D., Culiat, C. T., Churchill, G. a., Lariviere, W. R., Palmer, A. a., O'Hara, B. F., Voy, B. H., and Chesler, E. J. (2011). Genetic analysis in the Collaborative Cross breeding population. *Genome Research*, 21(8):1223–1238.

Phillippi, J., Xie, Y., Miller, D. R., Bell, T. A., Zhang, Z., Lenarcic, A. B., Aylor, D. L., Krovi, S. H., Threadgill, D. W., Pardo-Manuel de Villena, F., Wang, W., Valdar, W., Frelinger, J. a., de Villena, F. P.-M., Wang, W., Valdar, W., and Frelinger, J. a. (2014). Using the emerging Collaborative Cross to probe the immune system. *Genes & Immunity*, 15(1):38–46.

Pook, T., Schlather, M., de los Campos, G., Mayer, M., Schoen, C. C., and Simianer, H. (2019). HaploBlocker: Creation of Subgroup-Specific Haplotype Blocks and Libraries. *Genetics*, 212(4):1045–1061.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ram, R., Mehta, M., Balmer, L., Gatti, D. M., and Morahan, G. (2014). Rapid identification of major-effect genes using the collaborative cross. *Genetics*, 198(1):75–86.

Rasmussen, A. L., Okumura, A., Ferris, M. T., Green, R., Feldmann, F., Kelly, S. M., Scott, D. P., Safronetz, D., Haddock, E., LaCasse, R., Thomas, M. J., Sova, P., Carter, V. S., Weiss, J. M., Miller, D. R., Shaw, G. D., Korth, M. J., Heise, M. T., Baric, R. S., de Villena, F. P.-M., Feldmann, H., and Katze, M. G. (2014a). Host genetic diversity enables Ebola hemorrhagic fever pathogenesis and resistance. *Science*, 346(6212):987–991.

Rasmussen, M. D., Hubisz, M. J., Gronau, I., and Siepel, A. (2014b). Genome-Wide Inference of Ancestral Recombination Graphs. *PLoS Genetics*, 10(5):e1004342.

Raykov, Y. P., Boukouvalas, A., and Little, M. A. (2016). Simple approximate MAP inference for Dirichlet processes mixtures. *Electronic Journal of Statistics*, 10(2):3548–3578.

Relling, M. V. and Evans, W. E. (2015). Pharmacogenomics in the clinic. *Nature*, 526(7573):343–350.

Rogala, A. R., Morgan, A. P., Christensen, A. M., Gooch, T. J., Bell, T. A., Miller, D. R., Godfrey, V. L., and de Villena, F. P.-M. (2014). The Collaborative Cross as a resource for modeling human disease: CC011/Unc, a new mouse model for spontaneous colitis. *Mammalian Genome*, 25(3-4):95–108.

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

Rota, G.-C. (1964). The Number of Partitions of a Set. *The American Mathematical Monthly*, 71(5):498–504.

Rutledge, H., Aylor, D. L., Carpenter, D. E., Peck, B. C., Chines, P., Ostrowski, L. E., Chesler, E. J., Churchill, G. A., de Villena, F. P. M., and Kelada, S. N. P. (2014). Genetic regulation of Zfp30, CXCL1, and neutrophilic inflammation in murine lung. *Genetics*, 198(2):735–745.

Servin, B. and Stephens, M. (2007). Imputation-based analysis of association studies: Candidate regions and quantitative traits. *PLoS Genetics*, 3(7):1296–1308.

Shim, H., Chun, H., Engelman, C. D., and Payseur, B. A. (2009). Genome-wide association studies using single-nucleotide polymorphisms versus haplotypes: an empirical comparison with data from the North American Rheumatoid Arthritis Consortium. *BMC Proceedings*, 3(Suppl 7):S35.

Shorter, J. R., Odet, F., Aylor, D. L., Pan, W., Kao, C.-Y., Fu, C.-P., Morgan, A. P., Greenstein, S., Bell, T. A., Stevans, A. M., Feathers, R. W., Patel, S., Cates, S. E., Shaw, G. D., Miller, D. R., Chesler, E. J., McMillian, L., O'Brien, D. A., and de Villena, F. P.-M. (2017). Male Infertility Is Responsible for Nearly Half of the Extinction Observed in the Mouse Collaborative Cross. *Genetics*, 206(2):557–572.

Shusterman, A., Salyma, Y., Nashef, A., Soller, M., Wilensky, A., Mott, R., Weiss, E. I., Houri-Haddad, Y., and Iraqi, F. A. (2013). Genotype is an important determinant factor of host susceptibility to periodontitis in the Collaborative Cross and inbred mouse populations. *BMC Genetics*, 14:68.

Soller, M. and Beckmann, J. S. (1990). Marker-based mapping of quantitative trait loci using replicated progenies. *Theoretical and Applied Genetics*, 80(2):205–208.

Song, Y. S. (2006). Properties of subtree-prune-and-regraft operations on totally-ordered phylogenetic trees. *Annals of Combinatorics*, 10(1):147–163.

Srivastava, A., Morgan, A. P., Najarian, M. L., Sarsani, V. K., Sigmon, J. S., Shorter, J. R., Kashfeen, A., McMullan, R. C., Williams, L. H., Giusti-Rodríguez, P., Ferris, M. T., Sullivan, P., Hock, P., Miller, D. R., Bell, T. A., McMillan, L., Churchill, G. A., and De Villena, F. P. M. (2017). Genomes of the Mouse Collaborative Cross. *Genetics*, 206(2):537–556.

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.

Takuno, S., Terauchi, R., and Innan, H. (2012). The power of QTL mapping with RILs. *PloS ONE*, 7(10):e46545.

Thompson, K. L. and Kubatko, L. S. (2013). Using ancestral information to detect and localize quantitative trait loci in genome-wide association studies. *BMC Bioinformatics*, 14(1):200.

Threadgill, D. W. and Churchill, G. A. (2012). Ten Years of the Collaborative Cross. *Genetics*, 190(2):291–294.

Threadgill, D. W., Hunter, K. W., and Williams, R. W. (2002). Genetic dissection of complex and quantitative traits: from fantasy to reality via a community effort. *Mammalian Genome*, 13(4):175–178.

Valdar, W., Flint, J., and Mott, R. (2006a). 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.

Valdar, W., Holmes, C. C., Mott, R., and Flint, J. (2009). Mapping in structured populations by resample model averaging. *Genetics*, 182(4):1263–1277.

Valdar, W., Solberg, L. C., Gauguier, D., Burnett, S., Klenerman, P., Cookson, W. O., Taylor, M. S., Rawlins, J. N. P., Mott, R., and Flint, J. (2006b). Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nature Genetics*, 38(8):879–887.

van Dyk, D. A. and Park, T. (2008). Partially Collapsed Gibbs Samplers: Theory and Methods. *Journal of the American Statistical Association*, 103(482):790–796.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition.

Venkatratnam, A., Furuya, S., Kosyk, O., Gold, A., Bodnar, W., Konganti, K., Threadgill, D. W., Gillespie, K. M., Aylor, D. L., Wright, F. A., Chiu, W. A., and Rusyn, I. (2017). Collaborative Cross Mouse Population Enables Refinements to Characterization of the Variability in Toxicokinetics of Trichloroethylene and Provides Genetic Evidence for the Role of PPAR Pathway in Its Oxidative Metabolism. *Toxicological Sciences*, 158(1):48–62.

Vered, K., Durrant, C., Mott, R., and Iraqi, F. A. (2014). Susceptibility to Klebsiella pneumonaie infection in collaborative cross mice is a complex trait controlled by at least three loci acting at different time points. *BMC Genomics*, 15(1):865.

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., and Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *American Journal of Human Genetics*, 101(1):5–22.

Wallach, H. M., Jensen, S. T., Dicker, L., and Heller, K. A. (2008). An Alternative Prior Process for Nonparametric Bayesian Clustering. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 892–899.

Wei, J. and Xu, S. (2016). A Random-Model Approach to QTL Mapping in Multiparent Advanced Generation Intercross (MAGIC) Populations. *Genetics*, 202(2):471–86.

Wei, W. H., Hemani, G., and Haley, C. S. (2014). Detecting epistasis in human complex traits. *Nature Reviews Genetics*, 15(11):722–733.

Welling, M. (2006). Flexible Priors for Infinite Mixture Models. In *Proceedings of the Workshop on Learning with Nonparametric Bayesian Methods, 23rd ICML*.

Welsh, C. E., Miller, D. R., Manly, K. F., Wang, J., McMillan, L., Morahan, G., Mott, R., Iraqi, F. A., Threadgill, D. W., and de Villena, F. P.-M. (2012). Status and access to the Collaborative Cross population. *Mammalian Genome*, 23(9-10):706–712.

Wilke, R. A., Lin, D. W., Roden, D. M., Watkins, P. B., Flockhart, D., Zineh, I., Giacomini, K. M., and Krauss, R. M. (2007). Identifying genetic risk factors for serious adverse drug reactions: Current progress and challenges. *Nature Reviews Drug Discovery*, 6(11):904–916.

Williams IV, R., Lim, J. E., Harr, B., Wing, C., Walters, R., Distler, M. G., Teschke, M., Wu, C., Wiltshire, T., Su, A. I., Sokoloff, G., Tarantino, L. M., Borevitz, J. O., and Palmer, A. A. (2009). A common and unstable copy number variant is associated with differences in Glo1 expression and anxiety-like behavior. *PLoS ONE*, 4(3).

Xu, S. (2003). Theoretical basis of the Beavis effect. *Genetics*, 165(4):2259–2268.

Yalcin, B., Flint, J., and Mott, R. (2005). Using progenitor strain information to identify quantitative trait nucleotides in outbred mice. *Genetics*, 171(2):673–681.

Yamamoto, E., Iwata, H., Tanabata, T., Mizobuchi, R., Yonemaru, J.-i., Yamamoto, T., and Yano, M. (2014). Effect of advanced intercrossing on genome structure and on the power to detect linked quantitative trait loci in a multi-parent population: a simulation study in rice. *BMC Genetics*, 15:50.

Yang, H., Wang, J. R., Didion, J. P., Buus, R. J., Bell, T. A., Welsh, C. E., Bonhomme, F. F., Yu, A. H.-T. T., Nachman, M. W., Pialek, J., Tucker, P., Boursot, P., McMillan, L., Churchill, G. A., and De Villena, F. P.-M. M. (2011). Subspecific origin and haplotype diversity in the laboratory mouse. *Nature Genetics*, 43(7):648–655.

Yano, K., Yamamoto, E., Aya, K., Takeuchi, H., Lo, P. C., Hu, L., Yamasaki, M., Yoshida, S., Kitano, H., Hirano, K., and Matsuoka, M. (2016). Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. *Nature Genetics*, 48(8):927–934.

Yu, J., Holland, J. B., McMullen, M. D., and Buckler, E. S. (2008). Genetic design and statistical power of nested association mapping in maize. *Genetics*, 178(1):539–551.

Yuan, Z., Zou, F., and Liu, Y. (2011). Bayesian multiple quantitative trait loci mapping for recombinantinbred intercrosses. *Genetics*, 188(1):189–195.

Zhang, Z., Guillaume, F., Sartelet, A., Charlier, C., Georges, M., Farnir, F., and Druet, T. (2012a). Ancestral haplotype-based association mapping with generalized linear mixed models accounting for stratification. *Bioinformatics*, 28(19):2467–2473.

Zhang, Z., Wang, W., and Valdar, W. (2014). Bayesian Modeling of Haplotype Effects in Multiparent Populations. *Genetics*, 198(1):139–156.

Zhang, Z., Zhang, X., and Wang, W. (2012b). HTreeQA: Using Semi-Perfect Phylogeny Trees in Quantitative Trait Loci Study on Genotype Data. *G3: Genes, Genomes, Genetics*, 2(2):175–189.

Zheng, C., Boer, M. P., and van Eeuwijk, F. A. (2015). Reconstruction of Genome Ancestry Blocks in Multiparental Populations. *Genetics*, 200(4):1073–1087.

Zhou, X. and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, 44(7):821–824.

Zollner, S. and Pritchard, J. K. (2007). Overcoming the winner's curse: estimating penetrance parameters from case-control data. *American Journal of Human Genetics*, 80(4):605–615.

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

Zuk, O., Hechter, E., Sunyaev, S. R., and Lander, E. S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences*, 109(4):1193–1198.