Experimental Design & Analysis with Multiparental Populations

Gregory R. Keele

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 in Bioinformatics and Computational Biology.

Chapel Hill
2018

Approved by:

William Valdar

Fernando Pardo-Manuel de Villena

Samir Kelada

Michael Love

Leonard McMillan

Daniel Pomp

ABSTRACT

Gregory R. Keele: Experimental Design & Analysis with Multiparental Populations
(Under the direction of William Valdar)

Multiparental populations (MPP) are experimental populations descended from more than two founder or parental inbred strains. They generally possess far greater genetic variation and phenotypic variability than simpler bi-parental populations, and are thus powerful resources for genetic studies,. MPP have been developed in numerous model systems, and have been successfully utilized in genetic association or quantitative trait locus (QTL) mapping studies for identifying candidate genes and variants that modulate complex phenotypes. Statistical methods developed for simpler populations have been extended successfully for analyses of MPP, though problems can arise, such as dubious QTL that occur at positions with imbalanced founder haplotype contributions. This shortcoming reflects the potential value of statistical tools designed specifically for MPP that can better leverage the abundant genetic and phenotypic variation for design and analyses of powerful experiments.

This dissertation has two main topics: 1) experimental design and 2) genetic association and related analyses in MPP. Within the topic of experimental design, the use of the diallel, a specific form of MPP, to inform selection of powerful follow-up bi-parental crosses for QTL mapping is explored. More broadly, this approach represents a Bayesian decision theoretic approach and is found to provide a quantitative, principled procedure for leveraging information in the pilot data towards follow-up experiments. The second design subject is a power calculation tool for the Collaborative Cross (CC), a panel of recombinant inbred strains of mice, providing highly tailored power estimates for the design of mapping studies in the realized CC strains. Additionally, the tool is used to investigate how various aspects of experimental design and features of the underlying QTL affect the power to map QTL broadly.

The first subject for the topic of genetic association is a multiple imputation approach to QTL mapping in MPP that is shown to reduce false QTL that result from founder haplotype uncertainty and imbalanced founder haplotype contributions. Next, an analysis of heterogeneous stock rats, an

outbred MPP, is presented, in which imputed SNP association and fine-mapping approaches, including an integrative mediation procedure, are used to identify candidate variants influencing adiposity phenotypes. Finally, QTL mapping is performed on gene expression and chromatin accessibility outcomes in the CC, which largely detect local signals (within 5 Mb upstream or downstream of target outcome). These analyses are followed by a genome-wide integrative mediation approach, that detects local signatures of mediation of gene expression through chromatin accessibility, in a limited sample of CC mice.

To J. W. Keele

My dad and favorite scientist

ACKNOWLEDGEMENTS

There are many people and institutions to acknowledge and thank for supporting me in my development as a scientist capable of pursuing this research.

First and foremost, I thank my mentor and advisor, Will Valdar. We are both self-effacing introverts, and I appreciated that we could minimally acknowledge each other in the hallway without worry that either one of us would take offense or fear that our relationship was deteriorating. In Will I found a caring advisor who fosters an environment in which his students can feel comfortable to learn and to develop, without a harsh, uncompromising focus on production. He emphasizes a comprehensive training experience, including foundations in statistics, programming, and quantitative genetics, writing, engagement within our scientific community, and how to best communicate with a given audience. I know this is not the universal graduate student experience, and am all the more grateful for it.

I am thankful for the members of the Valdar lab; undoubtedly we are an interesting cohort. Past members include Alan Lenarcic, who helped me with the analysis of diallel data, Jeremy Sabourin, whom I worked with on DiploLASSO, a cool project not included in this disseration, Zhaojun Zhang, who handed Diploffect off to me, and Yunjung Kim, who assisted in the development of the multiple imputation project. Finally Yuying Xie, who acted as my "tiger mom" during my rotation. I miss him and his lovely family.

As for my concurrent Valdar lab peers, they truly feel like siblings to me, meaning I care for them dearly, but they also drive me crazy. It is an injustice to our relationships to summarize things with a single quip, but alas, I must. Dan Oreper, who seems to actually appreciate my extraneous knowledge and strong feelings about animals. Robert Corty, who is the extraverted yin to my introverted yang, and can always find a pun, even if it is not really there. Paul Maurizio, who kept me safe in Italy and allowed me to be "Tito Greg" to his son Lucas. Wes Crouse, a trophy husband who can eat a sickening quantity of hotdogs. Yanwei Cai, whom I bonded with through Pokemon, which his dog

Goki basically is. Last but not least, Kathie Sun, who can really tackle an all-you-can-eat salad bar, and gives great book recommendations. I hope to keep in touch with all of them.

I would also like to thank fellow students that I interacted with while at UNC, a few whom I wish to mention further. Austin Hepperla, who survived many years of Thanksgiving turkey fry and is always willing to talk sports, politics, video games, pop culture, etc. Bryan Quach, who also survived the previously mentioned frying events, and is a wonderful collaborator on the gene expression and chromatin accessibility project, never growing angry with me as I fiddled with the QTL mapping and mediation. Natalie Stanley, who I suspect converted to Bayesian for the jokes (Seinfeld allusion). Nur Shahir, who often shares her baking with the lab, and whose laugh can be heard across the building. Lauren Donoghue, for conversations on science and shared interests (turtles, non-turtle animals, etc. and probably topics only I have interest in but she is too polite to tell me are boring). The value of friendship cannot be overstated, particularly during challenging times.

I am grateful for the education that I received at my undergraduate institution, the University of Nebraska-Lincoln, and in particular wish to thank my honors thesis advisor, John Janovy Jr. I have fond memories of listening to tales of his parasitological exploits. I am also grateful for my experiences during my Master's degree in biostatistics at the University of Michigan and in Mike Boehnke's statistical genetics group. Those years helped build the foundation in statistics and computation that have been of great value during my PhD. I also thank Jennifer Beebe-Dimmer and Ann Schwartz for the opportunity to work with them and their support at Wayne State University.

I need to thank my committee members for their interest, support, and expertise. I sincerely thank Fernando Pardo-Manuel de Villena as chair, Samir Kelada, Mike Love, Leonard McMillan, Daniel Pomp, and former members Brian Bennett and Wei Sun.

I thank my collaborators in the projects that make up this work. First I thank Leah Solberg Woods at Wake Forest University, for allowing me to analyze her heterogeneous rat data, and putting up with my mistakes and the many adventures they brought. I am grateful to Terry Furey at UNC for the opportunity to work on the gene expression and chromatin accessibility data in the CC.

I also wish to thank the Bioinformatics and Computational Biology Curriculum and the Department of Genetics, especially Cara Marlowe and John Cornett. Without them, my family may have gone uninsured, I would have unintentionally committed tax fraud, and certainly failed to file the paperwork necessary to graduate.

The studies underlying my work were made possible through funding from the Biological and Biomedical Sciences Program, the Bioinformatics and Computational Biology Training Grant, and the National Institute of General Medical Sciences.

I have also been the beneficiary of a loving family for the entirety of my life. I express love and gratitude for my siblings: Ben and his wife Christy, Marcia and her husband Nick and my niece Nadia, Alex and his wife Abby, Emma, Jeff, and Joe. Being a member of seven siblings has indelibly influenced who I have become, and though obviously I had no choice in this aspect of my identity, I cannot imagine my life without them.

My parents have been a constant source of love and support. I am grateful for my mom, who is the nucleus of a large dispersing family. I appreciate the effort she exherts to visit regularly, and maintain a strong loving relationship with her granddaughter despite the physical distances I keep placing between them.

I am grateful for my dad, to whom this work is dedicated and with whom it is inexorably connected in my mind. Science means many things to me. Science is a career. Science incites passion and frustration. Science has helped two quiet men, lost in their heads, talk and relate to each other. Science makes me feel like a reflection of him, reminds me that I am his son.

I express complete and utter gratitude for the love and support of my wife Lindsey. Through all my periodic anxieties and agitations, she is a steady and constant source of love and support. It seems unlikely that I will get easier to deal with, so to Lindsey, please remember me at my best and forget the exhausting moments.

Finally, I am grateful for my wonderful daughter Cori. I am proud of the work in this dissertation, but it simply cannot compare to how proud of her I am. May she always know how I adore her.

PREFACE

Here I provide additional details and brief descriptions of the chapters within the context of the overall dissertation for the published and in-preparation manuscripts that make up the chapters of this dissertation.

Chapter 2: This chapter began as a course project in a Bayesian statistics course, BIOS 779, in Fall 2013 at UNC, taught by Professor Amy Herring, now at Duke University. This work is an extension of Bayesian methods developed for the analysis of diallel data within the Valdar lab (Lenarcic et al., 2012; Phillippi et al., 2014; Crowley et al., 2014; Maurizio et al., 2018; Turner et al., 2018). In this work, strain-level effects are characterized from diallel data, as in (Lenarcic et al., 2012), but here these effects are used as inputs into utility functions, such that they are meaningful for QTL mapping, in order to prioritize and select bi-parental crosses. It is uniquely qualified to be the first chapter of this dissertation due to the unique intermediary position the diallel occupies between bi-parental populations and multiparental ones.

Chapter 3: This chapter began after discussions with members of the lab of Professor Samir Kelada at UNC about calculating QTL mapping power specific to the CC. Highly efficient code was developed for the QTL mapping in **Chapter 6**, which was adapted to simulated CC data for the power calculations. Additionally, we investigated the effect of a range of genetic architectures (QTL to background strain effect sizes and allelic series) and experimental characteristics (number of CC strains and number of replicates) on power from a large scale perspective. Our goal was to provide a tool that could provide a highly tailored power for a given experiment, as well as some general power curves that can be used as reference for labs designing experiments in the CC. This work represents a bridge between the two topics of this dissertation of experimental design and genetic association in multiparental populations.

Chapter 4: The multiple imputation method described in detail has already been used in (Mosedale et al., 2017) for QTL mapping in CC mice. This chapter is also the first chapter of this dissertation wholly focused on genetic association analyses of multiparental populations, rather than

experimental design. In this work, a conservative multiple imputation approach to QTL mapping is used to avoid false associations that results from founder haplotype uncertainty and founder haplotype frequency imbalance. **Chapter 5** relates to this one, as an alternative approach to QTL mapping based on the challenges that this work revealed.

Chapter 5: In this chapter we tried multiple analytical approaches, some of which is described in **Chapter 4**, before arriving at the final process. The primary issue is that the population of heterogeneous stock (HS) rats had relatively high levels of uncertainty in terms of distinguishing founder haplotypes, as well as poor balance in terms of founder haplotype contributions. For example, at certain positions, more than half of the individuals could have inherited an allele from a single founder out of eight. We found that these joint issues led to particularly unstable haplotype-based associations (**Chapter 4**). To reduce these issues, we used an imputed SNP approach, in which we used the founder haplotype probabilities to impute SNP alleles, which effectively stabilized the associations, and even increased power by reducing the number of allele effect parameters that were being estimated. This chapter also introduces the use of mediation, which will be further used and developed in **Chapter 6**, to better understand the biology underlying a QTL.

Chapter 6: This chapter began more as a consultation on QTL mapping for collaborators in the lab of Professor Terry Furey at UNC, but became more involved as it became clear that more efficient mapping software for the CC would be required to accommodate having thousands of phenotypes (gene expression and chromatin accessibility). The work was further expanded to include assessment of the evidence for mediation of the eQTL effect on gene expression through chromatin accessibility, using a similar approach as (Chick et al., 2016) used for gene expression and protein abundance. In terms of this chapter's place within the arc of this dissertation, it represents progress beyond traditional QTL mapping in multiparental populations, and additionally provides a demonstration of the value of the systems genetics approach that is possible with the CC. As the overall work is highly collaborative and unfinished, the introduction, preliminary results, and discussion will be briefer than previous chapters, and focus on the portions relevant to this dissertation.

TABLE OF CONTENTS

LIST OF TABLES xix

LIST OF FIGURES xx

LIST OF ABBREVIATIONS xxiii

1 Introduction 1

1.1 Experimental populations 2

1.1.1 Genetic reference populations 3

1.1.2 Bi-parental populations 3

1.1.3 Multiparental populations 4

1.1.3.1 Heterogeneous Stock 5

1.1.3.2 Collaborative Cross and related populations 5

1.1.3.3 Multiparental populations in non-rodent model systems 6

1.1.4 Diallel 7

1.2 Experimental design 9

1.2.1 Diallel-informed bi-parental cross selection 9

1.2.1.1 Quantitative analysis of the diallel 9

1.2.1.2 DIDACT 11

1.2.2 Power to detect QTL in the realized Collaborative Cross 11

1.2.2.1 Realized Collaborative Cross 12

1.2.2.2 SPARCC 12

1.3 Genetic association and related analyses 12

1.3.1 Multiple imputation approach to QTL mapping in multiparental populations . 14

1.3.1.1 Developments in interval mapping 14
1.3.1.2 False associations and problematic uncertainty and founder allele frequency 15
1.3.1.3 Multiple imputation approach 15
1.3.2 Analysis of heterogeneous stock rats 17
1.3.2.1 Imputed SNP association 17
1.3.2.2 Fine-mapping approaches 17
1.3.2.3 Gene expression as mediator of QTL effect on phenotype 17
1.3.3 Integrative mediation analysis of gene expression and chromatin accessibility 19
1.3.3.1 Description of CC data and analyses 19
1.4 Summary 21
2 Using the diallel to select optimal bi-parental crosses to map QTL 22
2.1 Introduction 22
2.2 Statistical Models and Methods 27
2.2.1 Power to map QTL 27
2.2.1.1 Single QTL model of bi-parental cross 27
2.2.1.2 Power calculations 30
2.2.2 Characterization of strain-level genetic effects from pilot data 32
2.2.2.1 Bayesian modeling of diallel data 32
2.2.2.2 Prior specification 33
2.2.2.3 Strain-level effect to QTL effect 33
2.2.3 Decision theoretic approach 34
2.2.3.1 Power as utility function 34
2.2.4 Availability of data and software 37
2.3 Results 38
2.3.1 Mendelian phenotype 38
2.3.1.1 *Mx1* as a critical host-resistance factor in mice: 38
2.3.1.2 Expectations of DIDACT with a Mendelian trait 40

2.3.2 Complex trait 40
2.3.2.1 Calculated hemoglobin (cHGB): 40
2.3.3 Additional summaries of information 43
2.3.4 Parent-of-origin effects and RBC 43
2.4 Discussion 45
2.4.1 Assumptions connecting strain-level effect to QTL effect are wrong 45
2.4.2 Genetic similarity between strains 47
2.4.3 Extension to multiparental populations 47
2.4.4 Summary 48
3 SPARCC: An R package for estimating power to detect QTL through simulated experiments in the realized Collaborative Cross 49
3.1 Introduction 49
3.2 Methods 50
3.2.1 Data simulation 50
3.2.1.1 QTL effect 51
3.2.1.2 Strain effect 52
3.2.1.3 Noise effect 52
3.2.1.4 Robust power estimation 53
3.2.2 Mapping procedure 53
3.2.2.1 Regression model 53
3.2.2.2 Significance thresholds and power 54
3.2.2.3 QR decomposition for fast regression 55
3.2.2.4 Performing genome scans 56
3.2.3 Availability of data and software 56
3.2.3.1 R package 56
3.2.3.2 CC haplotype pair probabilities 56
3.2.3.3 Haplotype data reduction 57
3.3 Results and Discussion 57

3.3.1 Simple SPARCC example . . . 57
3.3.1.1 Run-time performance . . . 59
3.3.2 Large scale power dynamics . . . 59
3.3.2.1 Computing environment . . . 60
3.3.2.2 Experiment size and power . . . 60
3.3.2.3 Allelic series and power . . . 61
3.3.2.4 Statistical procedure assumes eight alleles . . . 61
3.3.2.5 Observed functional allele frequency imbalance . . . 61
3.3.3 CC as a mapping population . . . 62
3.3.4 Limitations . . . 62
3.3.5 Conclusion . . . 63
3.4 Simulation Documentation: Detailed description of `sim.CC.data()` options . . . 63
3.4.1 QTL effect . . . 63
3.4.2 Strain effect . . . 65
3.4.3 Additional options . . . 65
4 Accounting for haplotype uncertainty in QTL mapping of multiparental populations using multiple imputation . . . 71
4.1 Introduction . . . 71
4.2 Statistical Models and Methods . . . 78
4.2.1 General Framework . . . 78
4.2.2 Incorporating uncertainty in genetic state . . . 79
4.2.3 Modeling and sampling genetic state . . . 80
4.2.4 Conservative multiple imputation procedure . . . 81
4.2.5 Median as aggregate statistic . . . 82
4.2.6 Assessing genome-wide significance . . . 83
4.2.7 Availability of data and software . . . 83
4.3 Simulations . . . 84
4.3.1 Simulated populations . . . 84

4.3.2 Tested mapping procedures 84
4.3.3 Simulation of uncertainty 85
4.3.3.1 Probability dilution: 85
4.3.3.2 Dirichlet sampling: 86
4.4 Data Sets 87
4.5 Results 88
4.5.1 Illustration of false association with ROP 88
4.5.2 Simulated results 88
4.5.2.1 Dirichlet sampling: 90
4.5.2.2 Probability dilution: 92
4.5.3 More examples of results in real populations 93
4.5.4 Founder haplotype frequency and haplotype uncertainty 95
4.6 Discussion 99
4.6.1 SNP association as an alternative to ROP 100
4.6.2 Shrinkage as an alternative to ROP fixef 100
4.6.3 Disparity between ROP in simulated and real data 101
4.6.4 Summary 102
4.7 Additional Figures 104
5 QTL mapping in outbred rat population with imbalanced founder allele frequencies 106
5.1 Introduction 106
5.2 Methods and Procedures 107
5.2.1 Animals 107
5.2.1.1 Heterogeneous stock colony 107
5.2.1.2 Founding inbred sub-strains 107
5.2.2 Phenotyping protocol 107
5.2.3 Genotyping 108
5.2.4 RNA-Seq 108

5.3 Statistical Analysis 109
5.3.1 Estimating heritability of adiposity traits 109
5.3.2 Genome-wide association 109
5.3.3 Fine-mapping and haplotype effect estimation at detected QTL 109
5.3.4 Candidate gene identification 110
5.4 Results 110
5.4.1 HS founder strains exhibit large variation in adiposity traits 110
5.4.2 Adiposity traits are highly correlated with measures of metabolic health in HS rats 111
5.4.3 Adiposity traits are highly heritable 112
5.4.4 RetroFat QTL on chromsomes 1 and 6 112
5.4.5 Identification of *Adcy3*, *Krtcap3*, *Slc30a3* within the chromosome 6 RetroFat QTL 114
5.4.6 Identification of *Prlhr* within the chromosome 1 RetroFat QTL 114
5.5 Body weight QTL on chromosome 4 and identification of *Grid2* 118
5.6 Discussion 118
5.7 Detailed Methods 123
5.7.1 Animals 123
5.7.1.1 Housing 123
5.7.2 Statistical genetic analysis 123
5.7.2.1 Modeling genetic effects on adiposity 123
5.7.2.2 Heritability estimation 124
5.7.2.3 QTL mapping 124
5.7.2.4 Fine-mapping through Group-LASSO with fractional resample model averaging 125
5.7.2.5 Estimating haplotype substitution effects at detected QTL 126
5.7.2.6 Analysis of RNA-Seq data 126
5.7.3 Mediation analysis of phenotype, expression, and QTL 127
5.7.4 Mediation analysis results 129

5.7.4.1 Body weight chromosome 4 locus 130
5.7.4.2 RetroFat chromosome 1 locus 130
5.7.4.3 RetroFat chromosome 6 locus 130
5.8 Additional Figures 130
6 Detecting chromatin accessibility as a mediator of gene expression in Collaborative Cross mice 144
6.1 Introduction 144
6.2 Materials and Methods 146
6.2.1 Animals 146
6.2.2 Collaborative Cross reference genomes and transcriptomes 146
6.2.3 mRNA sequencing and processing 147
6.2.4 ATAC-Seq data processing 147
6.2.5 Chromatin accessibility quantification and windowing 148
6.2.6 Outcome filtering for eQTL and cQTL mapping 149
6.2.7 Founder haplotype data reduction 149
6.2.8 Differential expression and chromatin accessibility analysis 150
6.2.9 Gene set association analysis 150
6.2.10 QTL mapping 151
6.2.11 QTL mapping family-wide error rate (FWER) control 152
6.2.12 eQTL and cQTL false discovery rate (FDR) control 153
6.2.13 Detection of multiple QTL per outcome 153
6.2.14 Genome-wide and local chromosome-wide significance 154
6.2.15 Formal mediation analysis 155
6.2.16 Genome-wide mediation analysis 156
6.2.17 Mediation scan significance thresholds 158
6.3 Preliminary Results and Discussion 158
6.3.1 Summaries of the number of associations 158
6.3.1.1 eQTL 158

6.3.1.2 cQTL 158
6.3.1.3 Mediation 159
6.3.2 eQTL and cQTL mapping results 160
6.3.3 Mediation results 160
6.3.3.1 Identifying co-localizing eQTL and cQTL with mediation and without 163
6.3.3.2 eQTL, cQTL, and mediation are highly local 163
6.3.3.3 Detection of alignment issue in chromatin accessibility data 163
6.3.4 Distance from QTL or mediator to outcome 165
6.3.5 Buffering of eQTL effect from cQTL effect 166
6.3.6 Frequency of distal-QTL signal in comparison to local-QTL 169
6.3.7 Complexity of the underlying mediation 170
6.3.8 Summary 170
6.4 Additional Figures 171
7 Concluding remarks 172
7.1 Experimental design 173
7.1.1 Using the diallel to select optimal bi-parental crosses to map QTL 173
7.1.2 Simulated power to map QTL in the realized Collaborative Cross 173
7.2 Genetic association and related analyses 174
7.2.1 Accounting for haplotype uncertainty in QTL mapping of multi-parental populations using multiple imputation 174
7.2.2 QTL mapping in outbred rat population with imbalanced founder allele frequencies 175
7.2.3 Detecting chromatin accessibility as a mediator of gene expression in Collaborative Cross mice 176
7.3 Final conclusion 177
BIBLIOGRAPHY 178

LIST OF TABLES

2.1 Model of QTL effect on the mean for F2 and BC 29

2.2 Variance attributable to QTL effect for F2 and BC 29

4.1 Mapping procedures for simulated data 93

5.1 Correlations between adiposity and measures of metabolic health in HS rats 113

5.2 Genes in RetroFat chromosome 6 QTL interval 135

5.3 Genes in RetroFat chromosome 6 QTL interval (continued) 136

5.4 Genes in RetroFat chromosome 6 QTL interval (continued) 137

5.5 Genes in RetroFat chromosome 6 QTL interval (continued) 138

5.6 Genes in RetroFat chromosome 1 QTL interval 139

5.7 Genes in RetroFat chromosome 1 QTL interval (continued) 140

5.8 Genes in body weight chromosome 4 QTL interval 141

5.9 Potential mediators in RetroFat chromosome 6 QTL interval 142

5.10 Candidate mediators of the RetroFat chromosome 6 QTL 143

6.1 Number of genes with eQTL detected (q-value < 0.1) in lung, liver, and kidney tissues 159

6.2 Number of chromatin accessibility sites with cQTL detected (q-value < 0.1) in lung, liver, and kidney tissues 160

6.3 Number of chromatin mediators of gene expression detected (q-value < 0.1) in lung, liver, and kidney tissues 161

LIST OF FIGURES

1.1 Simplified bi-parental populations 4

1.2 Simplified depiction of the Collaborative Cross and Diversity Outbred stock in mice 6

1.3 Simplified depiction of Heterogeneous Stocks 7

1.4 Simplified representation of diallel 8

1.5 Simplified representation of DIDACT 10

1.6 SPARCC power curves comparing eight and two allele models 13

1.7 Example of artificial QTL with ROP 16

1.8 Example of HS rat analysis 18

1.9 Mediation model and example of genome-wide significant mediation of gene expression through chromatin accessibility 20

1.1 Diagrams of bi-parental crosses 24

1.2 Diagram of the diallel 26

1.3 Illustration of DIDACT 35

1.4 Response to Influenza A infection in diallel and its strain-level effects 39

1.5 Mean posterior utility for day 4 weight loss % post Influenza A infection in F2 crosses 41

1.6 Mean posterior utility for day 4 weight loss % post Influenza A infection in BC 42

1.7 DIDACT analysis of a complex trait, calculated hemoglobin 44

1.8 Detailed posterior utility visualization 45

1.9 Differences in mean posterior utility due to parent-of-origin effects 46

1.1 Simulated genome scans using SPARCC 66

1.2 Panel of power curves with respect to number of CC strains 67

1.3 Panel of power curves with respect to number of replicate observations 68

1.4 Heatmap of power by number of replicate observations and total mice in experiment 69

1.5 Comparison of power curves for varying allelic series with two functional alleles 70

1.1 Comparison of ROP and MI in HS rat data 89

1.2 Simulated CC-like population with no uncertainty 90

1.3 Effect of uncertainty modeled through Dirichlet sampling on genome scan in a single simulated CC-like RI panel 91

1.4 Effect of uncertainty modeled through Dirichlet sampling on association at QTL in simulated data 94

1.5 Effect of uncertainty modeled through Dirichlet sampling on association at null locus in simulated data 94

1.6 Comparison of ROP and MI in CC data 96

1.7 Comparison of ROP and MI in large HS rat data set 97

1.8 Founder haplotype allele frequencies in experimental populations 98

1.9 Alternatives to fixef ROP and fixef MI 99

1.10 Effect of uncertainty modeled through probability dilution on genome scan in a single simulated CC-like RI panel 104

1.11 Effect of uncertainty modeled through probability dilution sampling on association at QTL in simulated data 105

1.12 Effect of uncertainty modeled through probability dilution sampling on association at null locus in simulated data 105

1.1 Adiposity traits in the HS rats and the inbred founders 111

1.2 Significant correlations between adiposity and metabolic traits in HS rats 112

1.3 RetroFat chromosome 6 QTL and subsequent fine-mapping analyses 115

1.4 Co-localizing eQTL and their effect on the QTL association with RetroFat when included in the model 116

1.5 Causal graph model for the candidates underlying the RetroFat chromosome 6 QTL 117

1.6 RetroFat chromosome 1 QTL and subsequent fine-mapping analyses 119

1.7 Body weight chromosome 4 QTL and subsequent fine-mapping analyses 120

1.8 First ten principal components comparing genotypes between two genotyping centers 131

1.9 Fine-mapping of the RetroFat chromosome 6 QTL with LLARRMA-dawg 132

1.10 Gene annotations for the LLARRMA-dawg fine-mapping interval 133

1.11 Scatterplot of RetroFat by *Krtcap3* shows signs of direct mediation 133

1.12 Scatterplot of RetroFat by *Slc30a3* suggests a suppressor 134

1.1 Simple model for mediation of eQTL effect on gene expression through chromatin accessibility 156

1.2 eQTL and cQTL signals for lung, liver, and kidney 162

1.3 Example of evidence consistent with chromatin accessibility mediating between eQTL and gene expression 164

1.4 Co-localizing eQTL and cQTL not sufficient for mediation 165

1.5 eQTL, cQTL, and mediation signal is primarily local in lung, liver, and kidney 166

1.6 Genome-wide association signal increases as distance decreases from gene, chromatin region, and eQTL 167

1.7 Local chromosome-wide association signal increases as distance decreases from gene, chromatin region, and eQTL 168

1.8 Principal components of gene expression and chromatin accessibility 171

LIST OF ABBREVIATIONS

AIL	Advanced Intercross Line
ATAQ	Assay for Transposase Accessible Chromatin
BC	Backcross
BH	Benjamini-Hocheberg
bp	base pair
BXD	C57BL/6J $\times$ DBA/2J recombinant inbred strains
CC	Collaborative Cross
CC-RIX	CC F1 intercrosses
cHGB	calculated Hemoglobin
CI	Confidence Interval
cM	centi-Morgan
CPM	Counts Per Million
cQTL	chromatin accessibility Quantitative Trait Locus/Loci
DIDACT	Diallel Informed Decision theoretic Approach for Crosses Tool
DNA	Deoxyribonucleic Acid
DO	Diversity Outbred stock
DSPR	*Drosophila* Synthetic Population Resource
EM	Expectation-Maximization
eQTL	expression Quantitative Trait Locus/Loci
EVD	Extreme Value Distribution
FDR	False Discovery Rate
FWER	Family-Wide Error Rate
GRP	Genetic Reference Population
GWAS	Genome-Wide Association Study
HK	Haley-Knott
HMM	Hidden Markov Model
HPD	Highest Posterior Density
hQTL	histone modification Quantitative Trait Locus/Loci

HS	Heterogeneous Stock
IAV	Influenza A Virus
IM	Interval Mapping
LASSO	Least Absolute Shrinkage and Selection Operator
LD	Linkage Disequilibrium
LMM	Linear Mixed Model
LOD	Logarithm of Odds
LRT	Likelihood Ratio Test
MAF	Minor Allele Frequency
MAGIC	Multiparent Advanced Generation Inter-Cross
Mb	Megabase
MC	Monte Carlo
MCMC	Markov Chain Monte Carlo
MI	Multiple Imputation
MLE	Maximum Likelihood Estimate
MPGRP	Multiparental Genetic Reference Population
MPP	Multiparental Population
mRNA	messenger RNA
NAM	Nested Association Mapping
POE	Parent-of-Origin Effect
pQTL	protein abundance Quantitative Trait Locus/Loci
QTL	Quantitative Trait Locus/Loci
RetroFat	Retroperitoneal Fat pad
RLE	Relative Log Expression
RI	Recombinant Inbred
RMIP	Resample Model Inclusion Probability
RNA	Ribonucleic Acid
ROD	Regression-On-Diplotype
ROP	Regression-On-Probabilities

rQTL	ribosome occupancy Quantitative Trait Locus/Loci
rRNA	ribosomal RNA
Seq	Sequence
SNP	Single Nucleotide Polymorphism
SPARCC	Simulated Power Analysis for the Realized Collaborative Cross
TPM	Transcripts Per Million
TSS	Transcription Start Site
WLS	Weighted Least Squares

CHAPTER 1
Introduction

Quantitative trait locus (QTL) mapping in model organisms is a form of genetic association that has successfully identified genetic variants associated with medically-relevant phenotypes, including but certainly not limited to these examples in rodents: muscle malformation (Hartmann et al., 2008; Kelly et al., 2013), cocaine response (Kumar et al., 2013), asthma (Kelada, 2016; Donoghue et al., 2017), diabetes (Solberg Woods et al., 2010, 2012; Keele et al., 2018), and drug response in liver disease (Mosedale et al., 2017). Compared with epidemiological studies of naturally occurring human populations, sometimes referred to with the more general term genome-wide association studies (GWAS) (McCarthy et al., 2008; Teslovich et al., 2010; Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014), crosses of organisms allow for reduced population structure, better control of unobserved environmental factors, and more extensive or invasive phenotyping of samples than is possible with humans.

Traditionally experimental crosses involved two inbred strains as founders or parents, which can be referred to as bi-parental crosses. Recently, experimental populations in which individuals descend from more than two inbred founder strains, or multiparental populations (MPP), have been developed in a number of model organism systems. These populations, particularly replicable ones, act as rich reservoirs of genetic variants and phenotypic variability that provide the raw material that investigators must sift through for interesting biology relating to their system and model. Their additional complexity also pose unique statistical challenges, for which specialized and bespoke analytical methods could more effectively and efficiently draw insights and inferences. This dissertation is organized into two main topics relating to:

1. Experimental design approaches designed around MPP data and related experiments.

2. Genetic association approaches for MPP data and related analyses.

Together these sections provide novel advancements in the use of MPP data, both towards the design of experiments and the genetic association analyses, which will allow for these resources to be more effectively utilized. Examples and analyses will generally focus on rodent models, primarily laboratory mice. However, these organisms are not fundamental to the methodology, which could have application to any organism with an MPP. Prior to describing the projects that make up this dissertation in detail, background information will be presented on the various forms of experimental populations, both to provide context and justification for this work.

1.1 Experimental populations

This work generally focuses on studies, that at least in part, seek to associate positions in the genome, or more ideally, genes or variants with phenotypes, primarily within experimental populations, particularly MPP. Genetic association studies fundamentally require the genomes amongst the study samples to be randomized at loci across the genome, thus allowing the effect at one position to be separated from the effect at another. When this randomization is flawed, loci that are not physically linked can become correlated, representing non-syntenic associations, and possibly result in false positive associations. This is the process which underlies population structure, in which genetic drift and non-random mating create the non-syntenic associations that correlate to some extent with unobserved population factors. Whereas population structure must be recognized and accounted for in observational epidemiological populations of humans (Devlin and Roeder, 1999; Hoffman, 2013), the breeding design in experimental populations of organisms can greatly minimize the issue, as well as strongly controlling other influential factors, resulting in individuals with more perfectly randomized genomes. Alternatively, such individuals can be referred to as exchangeable. It is also important to acknowledge that population structure, and other unobserved confounders, can still occur in experimental populations, and is indeed likely with certain breeding and experimental designs. Still, the potential to have a more ideally controlled population is an appealing feature for using experimental populations as opposed to observational ones.

1.1.1 Genetic reference populations

One useful form of experimental population are collections, or panels, of inbred lines or strains, which began to be developed in full force in the early 20th century in a number of organisms, including the mouse (Casellas, 2011). An inbred animal results from multiple generations of inbreeding through sib-sib matings, and are predominantly homozygous at positions across the genome. Whereas the inbred state can be challenging to animals and result in line extinctions (Shorter et al., 2017), plant models can be more amenable, particularly with self-pollinators (Allard, 1999), and even cross-pollinators (Robsa Shuro, 2017). These panels provide researchers a renewable source of replicable genomes, ignoring *de novo* mutations and genetic drift (Keane et al., 2011), and powerfully minimize external sources of error outside of genetic effects specific to the strains. Phenotype surveys across a panel of inbred strains represent stable references within model organism systems (Phillippi et al., 2014; Rasmussen et al., 2014; McMullan et al., 2016; Roberts et al., 2018). For these reasons, inbred panels represent a class of experimental population, the genetic reference population (GRP), primarily providing stable, replicable genomes and phenotypes. For QTL mapping to be possible, the genomes of a population must be randomized through recombination events in meiosis is necessary, thus leading to experimental crosses of inbred strains.

1.1.2 Bi-parental populations

The simplest experimental crosses involve two strains, which will be referred to as bi-parental crosses or populations (Broman, 2001). The simplest bi-parental crosses are F2 intercrosses and backcrosses (BC), which will be discussed in **Chapter 2**, in which only a single generation of genetic recombination occurs between the parental haplotypes, resulting in mapping populations with little population structure but poor mapping resolution. The mapping resolution can be improved through additional generations of intercrosses, often referred to as advanced intercross lines (AIL) (Darvasi and Soller, 1995; Parker et al., 2011, 2012, 2014), though population structure can become an issue. These previous bi-parental experimental populations are all outbred and non-replicable. GRP populations that can also function as mapping populations are possible through the development of recombinant inbred (RI) lines or strains, as well as their intercrosses (RIX) (Zou et al., 2005), which are particularly common in plants (Lister and Dean, 1993; Mansur et al., 1996; Monforte

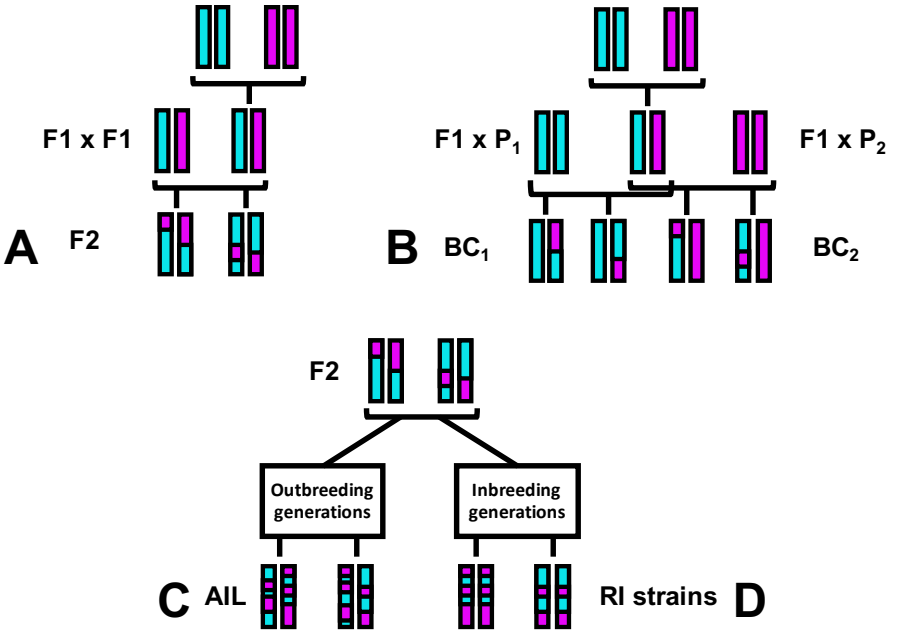

Figure 1.1: Simplified representations of F2 (A), BC (B), AIL (C), and RI strains (D). Each genome is simplified to a single pair of chromosomes, colored with respect to parental haplotypes. The F2, BC, and AIL represent outbred populations. The BC is unique in that at any given locus, only two potential genetic states are observed, rather than three. The RI strains are inbred and replicable, and thus as a panel, represent a GRP. Although these populations are powerful tools for genetic experiments, they are constrained in terms of their genetic variation, as only two founder haplotypes are present.

and Tanksley, 2000), though also in mice, such as the BXD lines (Peirce et al., 2004; Carbonetto et al., 2014), in which inbreeding generations follow the initial outbreeding generations as in F2 crosses or BC until an inbred state is established, resulting in individuals with genomes that are both inbred and mosaics of the two parental haplotypes. These forms of experimental populations are powerful genetic tools, though they are generally constrained in terms of the overall genetic variation phenotypic variability they can possess from the natural populations from which they descend. Simplified visual representations of these bi-parental populations are in **Figure 1.1**.

1.1.3 Multiparental populations

MPP address this issues of reduced genetic variation in bi-parental populations by incorporating more inbred strains, and thus likely genetic variation, into the populations. A practical challenge involved in the development of MPP is the greater complexity in breeding design; ideally additional

lines of inheritance are incorporated in such a way as to avoid population structure as well as maintain balance in terms of founder contributions.

1.1.3.1 Heterogeneous Stock

Heterogeneous stock (HS) populations in mice (Valdar et al., 2006b) and rats (Hansen and Spuhler, 1984) represent outbred MPP that, due to additional generations of recombinations through outbreeding, have finer mapping resolution. Alternatively, due to the rotational breeding design used, the HS can have greater levels of population structure and founder allele frequency imbalances. These populations can be viewed as an MPP analogue to the bi-parental AIL. HS rat data will be analyzed and discussed in **Chapters 4** and **5**. As outbred populations, the HS genomes are not replicable, and as such, are not GRP. Recently multiparental genetic reference populations (MPGRP) have been developed in a number of animal and plant models, which bring together the powerful experimental control of panels of inbred strains and the increased genetic diversity of MPP.

1.1.3.2 Collaborative Cross and related populations

MPGRP represent an MPP generalization of bi-parental RI strains and their intercrosses. The Collaborative Cross (CC) (Churchill et al., 2004; Collaborative Cross Consortium, 2012; Srivastava et al., 2017), which will be a focus in **Chapters 3**, **4**, and **6**, is an multiparental panel of RI strains in mouse, descended from five traditional inbred strains (A/J, C57BL/6J, 129S1/SvImJ, NOD/LtJ, NZO/H1LtJ) and three wild derived strains (CAST/EiJ, PWK/PhJ, WSB/EiJ), representing three subspecies of the house mouse, *Mus musculus*, and thus collectively possessing a high level of genetic variation (Yang et al., 2007, 2011), particularly in comparison to bi-parental populations. Alhough subspecies incompatibilities (Shorter et al., 2017) limited the number of strains produced, the CC, and its incipient lines, have been a valuable tool for QTL mapping (Aylor et al., 2011; Phillippi et al., 2014; Kelada, 2016; Mosedale et al., 2017; Donoghue et al., 2017). The CC is also a source of better murine models of human disease, likely the result of interesting allelic combinations being fixed across the genome, for example, of colitis (Rogala et al., 2014), ebola infection (Rasmussen et al., 2014), and West Nile Virus infection (Graham et al., 2015).

Related MPP have developed out of the CC. The CC F1 intercrosses (CC-RIX) (Rasmussen et al., 2014; Graham et al., 2015) allow for replicable outbred genomes, which generally produce more

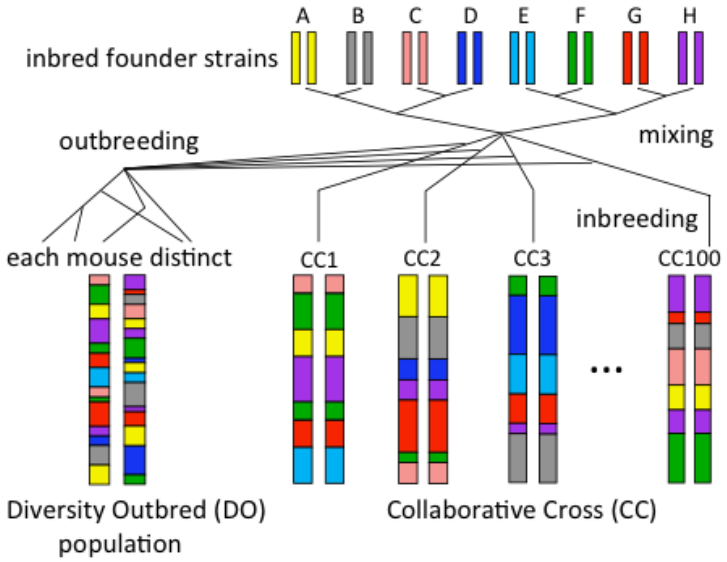

Figure 1.2: Simplified representations of the CC and DO, with each genome being simplified to a single pair of chromosomes, colored with respect to founder haplotype. The CC is a panel of MPP RI strains. The DO are outbred and have finer-grain founder haplotype blocks than the CC. CC data are analyzed in **Chapters 5** and **6**. Figure courtesy of William Valdar.

robust progeny, representing large scale heterosis (Birchler et al., 2006), and thus better approximating natural populations. Similarly, the Diversity Outbred stock (DO) (Churchill et al., 2012; Svenson et al., 2012; Gatti et al., 2014) represents an outbred population that shares the same founders as the CC, sacrificing replicability but providing fine scale mapping resolution with relatively little population structure. These related populations have the potential to be jointly analyzed, or used to replicate or confirm findings amongst one another, as was done in (Chick et al., 2016) in which allele effects at QTL detected in the DO were confirmed in the CC. Together these populations provide a strong foundation for systems genetics in mouse models. MPP and MPGRP have been developed in other organisms, though characteristics vary in terms of number of founder strains, number of resulting RI strains, and breeding design. Simplified representations of the CC and the DO together and the HS are presented in **Figures 1.2** and **1.3** respectively.

1.1.3.3 Multiparental populations in non-rodent model systems

Some animal species reproduce rapidly in comparison to rodents, providing the potential for more complex, extensive MPP. Examples in animals include the *Drosophila* Synthetic Population Resource (DSPR) in flies (King et al., 2012b,a; Long et al., 2014; King and Long, 2017; Najarro

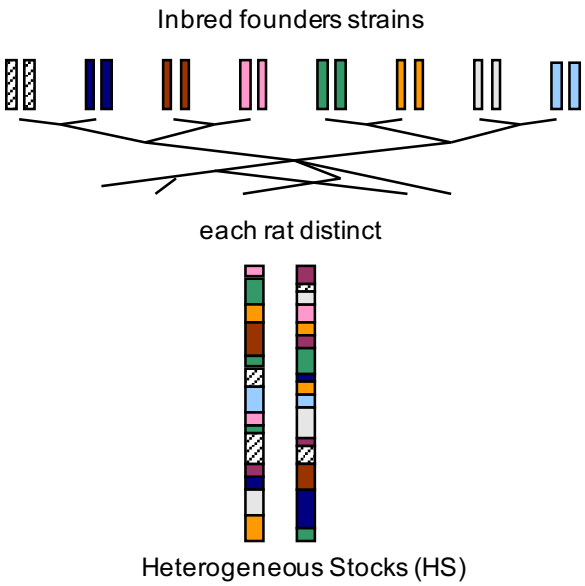

Figure 1.3: Simplified representation of the HS, with each genome being simplified to a single pair of chromosomes, colored with respect to founder haplotype. The HS are outbred and have fine-grain founder haplotype blocks, due to many generations of outbreeding. HS populations are similar to the DO (**Figure 1.2**), though likely less balanced in terms of founder haplotype contributions. HS data are analyzed in **Chapters 4** and **5**. Figure courtesy of William Valdar.

et al., 2017; Stanley et al., 2017), round worm (Noble et al., 2017), and yeast (Cubillos et al., 2017). Certain plant models can also reproduce prodigiously, as well as being more amenable to inbreeding. Examples of MPP in plants include multiparent advanced generation intercross lines (MAGIC) in *Arabidopsis* (Kover et al., 2009; Huang et al., 2011) and rice (Bandillo et al., 2013; Raghavan et al., 2017) and nested association mapping (NAM) populations in maize (Buckler et al., 2009), sorghum (Bouchet et al., 2017), strawberry (Mangandi et al., 2017), and oil palm (Tisné et al., 2017). Though the work presented here will focus completely on data from mice and rats, the ideas and methodology should generalize to these populations as well.

1.1.4 Diallel

The diallel, as a collection of inbred strains and the full set of F1 hybrids, including reciprocal hybrids that distinguish between maternal and paternal strain identities of the parents (A mat $\times$ B pat and B mat $\times$ A pat are reciprocal F1 hybrids with respect to each other), represents an experimental population that is intermediary to bi-parental populations and MPP. Any given individual will descend from at most two inbred strains, but in aggregate, multiple inbred strains are represented. One way to

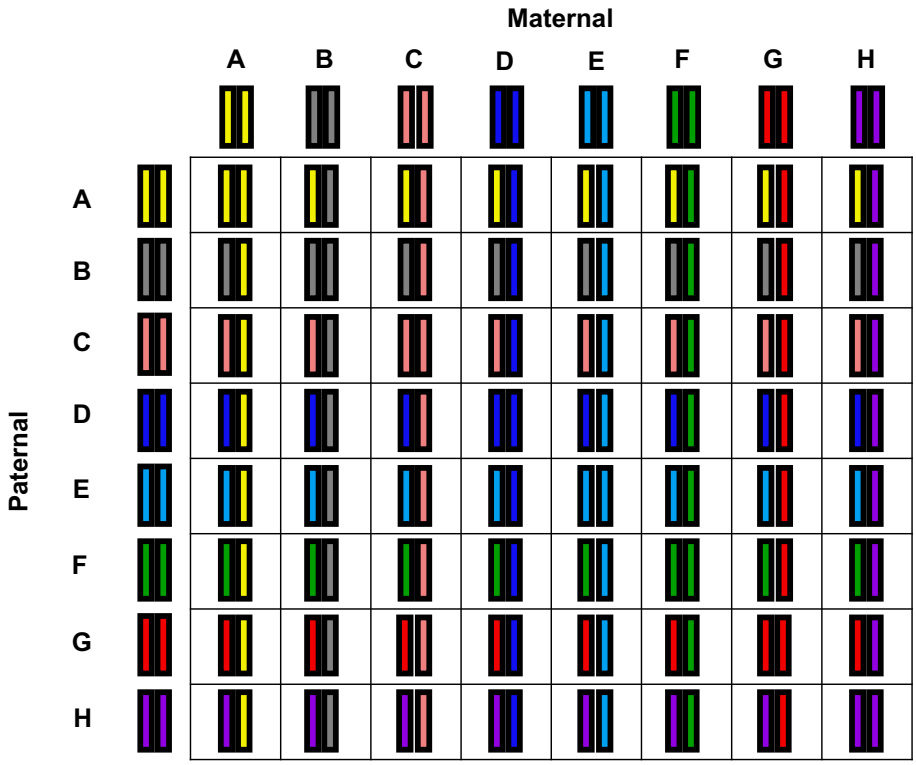

Figure 1.4: Simple representation of a diallel. The paternal strains are listed in the rows, maternal strains in the columns, and the offspring in their intersections. Here a unique genome is presented as a single pair of colored chromosomes. Cells along the diagonal represent the inbred individuals, and the off-diagonal are the F1 hybrids. Cells in mirror positions of each other with respect to the diagonal are reciprocal F1 hybrids with respect to each other, in which maternal and paternal strains are reversed. The diallel population is not a mapping population because recombination events are not observed between the founder haplotypes.

view the diallel with respect to MPP is as the full grid of potential crosses that produce the individuals in the initial outbreeding crosses necessary for the development of an MPP. The diallel is not a mapping population, due to no recombination events between the founder haplotypes; however, it can be used to characterize aggregate strain-level effects on phenotypes (Lenarcic et al., 2012). The diallel will be discussed in greater detail in **Chapter 2**. A simple depiction of the diallel is present in **Figure 1.4**.

Having described the experimental populations that will be used, the focus now shifts to the primary topics of interest for this dissertation: experimental design and genetic association and related analyses.

1.2 Experimental design

Experimental design is a broad topic, heavily dependent on the specific field of science and its range of experiments. The primary focus will be on the experimental design of QTL mapping experiments, though certain concepts can be extended to other types of experiments. This portion on design is organized into two parts:

1. A unique and novel approach to using diallel data as pilot data for selecting bi-parental crosses for QTL mapping (**Chapter 2**).

2. A focused simulation approach to power calculation for QTL mapping with the realized CC genomes that can assist in choosing the number of strains and replicate observations (**Chapter 3**).

1.2.1 Diallel-informed bi-parental cross selection

The selection of a breeding strategy or design for the purpose of QTL mapping has generally involved crossing inbred strains that strongly contrast with respect to the phenotype of interest, as the resulting mapping population should possess segregating variants that influence the phenotype. An Inbred strain survey (Phillippi et al., 2014; Rasmussen et al., 2014; McMullan et al., 2016; Roberts et al., 2018) that provides the phenotype information necessary to select promising crosses can be viewed as a partial diallel, thus suggesting that quantitative approaches could be used to leverage information in the diallel for the design of downstream bi-parental crosses. This concept, implemented in an R package called DIDACT (Diallel Informed Decision theoretic Approach for Crosses Tool), is represented in **Figure 1.5**.

1.2.1.1 Quantitative analysis of the diallel

The diallel was originally put forth in the early 20^{th} century, and has seen a steady advancement in methodology, from estimation of general combining ability (Griffing, 1956) with related F2 populations, to harnessing shrinkage through use of random effects (Zhu and Weir, 1996; Tsaih et al., 2005), and finally the use of Bayesian methods (Greenberg et al., 2010; Lenarcic et al., 2012). (Verhoeven et al., 2006) explored jointly analyzing a partial diallel with observed downstream F2

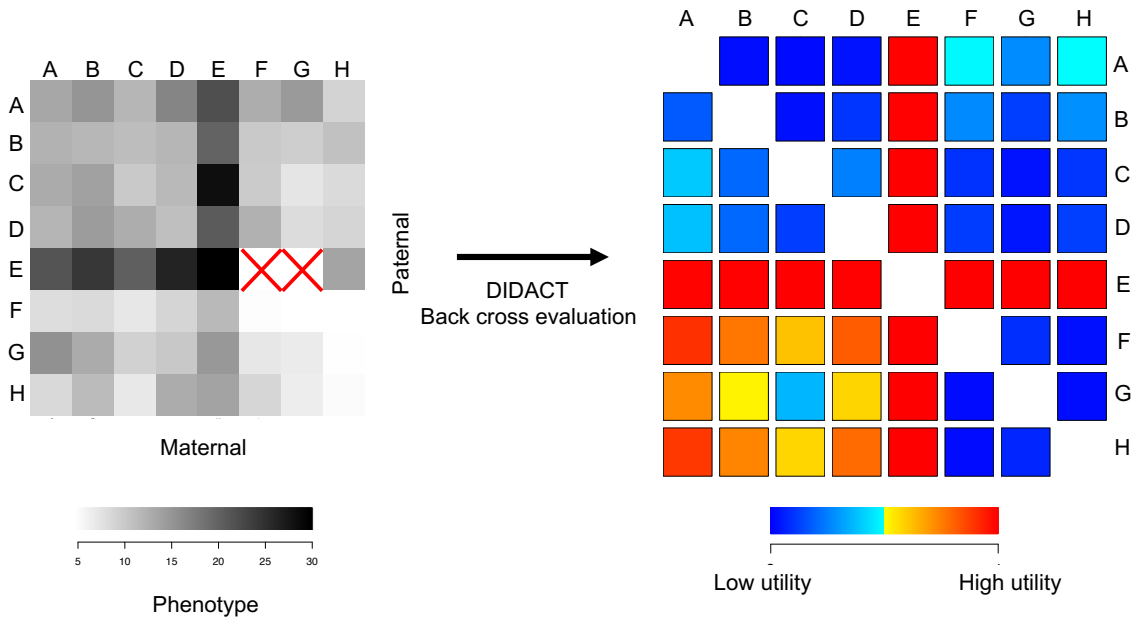

Figure 1.5: DIDACT seeks to evaluate potential bi-parental crosses with respect to a utility function based on pilot data from the diallel. The process involves connecting estimated strain-level effects from the diallel data, represented here as the gray scale grid of mean phenotype level per diallel cell, to a specified utility functions that use these strain-level effect estimates as inputs. A red "X" indicates that no progeny were produced for that cell of the diallel. In this example, a max utility BC is red and a minimum utility BC is blue. One possible utility function is the power to detect some number of QTL underlying the estimated strain-level effects. Alternatively, a simpler utility function would be the difference in expected phenotype of the progeny based on the strain-level effects.

populations, and found that such an approach could improve generalizing QTL effects from their narrow F2 mapping population into the broader diallel or panel of inbred strains.

1.2.1.2 DIDACT

DIDACT (**Chapter 2**) is ignorant of the downstream F2 populations or any mapping populations, and instead seeks to evaluate a specified utility function for each potential downstream cross, which consist of F2, BC, and parent-of-origin effect reciprocal BC, based on strain-level effects as inputs.

DIDACT is flexible to different utility functions, with one example being the power to map a QTL underlying the estimated strain-level effects. Though the QTL power utility function is dependent on the strong and unlikely assumption that the strain-level effects are attributable to some specified number of QTL, in practice the power tracks with crosses that match strains with contrasting phenotypes. This is similar to what has been done previously, however, now in a principled way that can incorporate complex strain-level effects. Alternatively, utility functions that require less assumptions can be used, such as the expected difference in phenotype based on the strain-level effects, though the utility may be less interpretable in comparison to a quantity like power. More generally, DIDACT is a demonstration of a fundamental Bayesian decision theoretic approach that is extendable to other experimental settings as well.

1.2.2 Power to detect QTL in the realized Collaborative Cross

With **Chapter 3**, the focus of the dissertation begins to transition towards genetic association in MPP, though still within the context of experimental design, specifically QTL mapping experiments with the CC. Panels of RI strains are particularly valuable tools for QTL mapping because of their status as GRP, allowing for the potential of QTL results to be replicated across experiments, labs, and related populations (Belknap and Atkins, 2001). Their stable nature as reference populations also allows for highly specific QTL mapping power calculations that can assist researchers in designing efficient but powerful experiments. Previous literature has focused on analytical power estimation within bi-parental RI strains (Kaeppler, 1997). Within plant models, QTL mapping power calculations has been performed through simulation, in which both RI genome and phenotype were simulated (Falke and Frisch, 2011; Takuno et al., 2012). However, their simulation are tailored for QTL mapping experiments in plants, with particularly large QTL effect sizes and elaborate multiple

QTL models, whereas in many phenotypes in animal models, the expectation will be for smaller QTL effects and a preference for single locus models.

1.2.2.1 Realized Collaborative Cross

At the onset of the development of the CC, power calculations were performed through simulations of the RI strain genomes and phenotypes (Valdar et al., 2006a). However, such power estimates are not necessarily representative of the resulting population, which fell short of the stated 1000 goal of RI strains (Churchill et al., 2004) due to line extinctions, likely as a result of allelic incompatibilities (Shorter et al., 2017). With the finalized strains (around 75), power calculations can be based on the actual, or realized, CC genomes, and can thus reflect slight deviations from the expected balance in founder contributions (Srivastava et al., 2017).

1.2.2.2 SPARCC

The R package SPARCC (Simulated Power Analysis in the Realized Collaborative Cross) allows for power calculations that can be highly tailored to a specific experiment with a specific set of strains. Alternatively, it can also perform robust power calculations by varying the set of CC strains, the position of the simulated QTL, and even the allelic series (Yalcin et al., 2005). In **Chapter 3**, SPARCC is used to investigate how QTL effect size, background strain effect size, number of CC strains, number of replicate observations, and the allelic series affect QTL mapping power in the CC. **Figure 1.6** is an example of how SPARCC can interrogate aspects of the experimental design as well as the underlying biology as modulators of QTL mapping power in the CC.

1.3 Genetic association and related analyses

The focus now shifts fully from experimental design to the actual genetic association analysis in MPP, fine-mapping analyses, and finally a genome-wide mediation approach that statistically integrates multiple levels of data on the same individuals. Specifically, **Chapters 4** focuses entirely on QTL mapping in MPP populations, **Chapter 5** transitions between QTL mapping and fine-mapping approaches to identify candidate variants or genes, particularly emphasizing a mediation approach, and finally **Chapter 6** is primarily focused on genome-wide mediation in the CC.

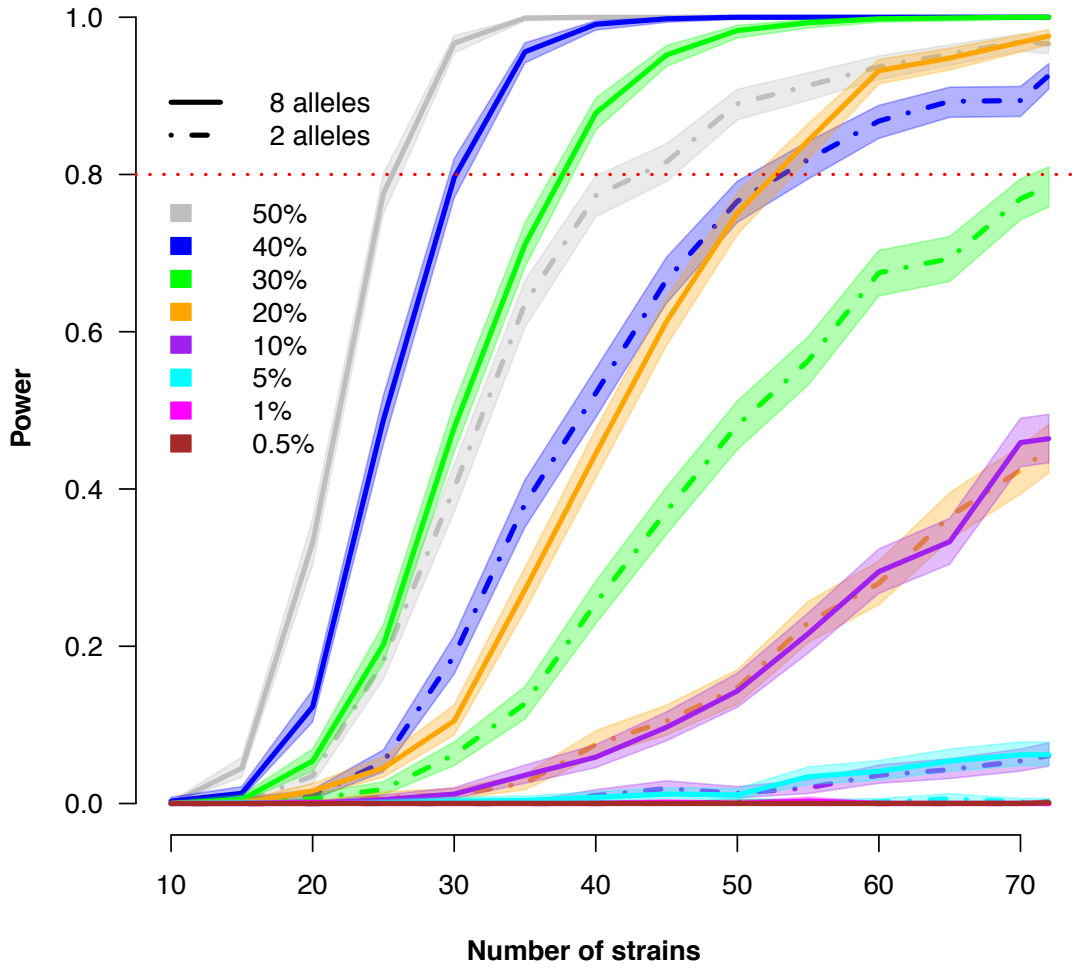

Figure 1.6: Power curves from SPARCC with power on the y-axis and number of CC strains on the x-axis. Five replicate observations per CC strain were simulated for this example. Colors represent different QTL effect sizes in terms of proportion of total variance. Solid lines represent an eight allele model for the simulation, and dashed lines represent an allelic series with two functional alleles. The QTL mapping procedure uses an eight allele alternative model, which is standard practice. Power is significantly worse for allelic series with two functional alleles, mostly due to imbalanced observations of each functional allele.

In general, QTL mapping approaches developed in simpler bi-parental populations have been extended for use in MPP and been successful (Valdar et al., 2006c,b, 2009; Svenson et al., 2012; Baud et al., 2013, 2014; Gatti et al., 2014; Phillippi et al., 2014). **Chapter 4** focuses on examples when the use of recycled methods from bi-parental populations can be problematic.

1.3.1 Multiple imputation approach to QTL mapping in multiparental populations

1.3.1.1 Developments in interval mapping

QTL mapping through interval mapping (IM) (Lander and Botstein, 1989) models the association between founder haplotype and phenotype, as opposed to the association between a variant genotype and phenotype, as is more common in human GWAS. Founder haplotype identities are not directly observed, but rather probabilistically inferred, commonly with a hidden Markov model (HMM) (Lander and Green, 1987; Mott et al., 2000; Liu et al., 2010; Fu et al., 2012; Gatti et al., 2014; Zheng et al., 2015) using genotype data. IM, in its original form, acknowledged this uncertainty through the use of a mixture of Gaussians model (Broman and Sen, 2009), which required an expectation-maximization (EM) algorithm (Dempster et al., 1977) in order to fit maximum likelihood parameters (MLE). The EM is an iterative procedure and thus computationally expensive on a large scale, which becomes more problematic with denser genome scans that involve more tests of association. Additionally, the MLE estimates can be unstable in the presence of little information distinguishing the haplotype states, instead becoming stuck in local maxima. (Haley and Knott, 1992) and (Martínez and Curnow, 1992) proposed a computationally efficient regression approximation in which the phenotype is simply regressed on the probabilities, or dosages for an additive model, of the haplotypes. This approximation to IM is sometimes called Haley-Knott (HK) regression or regression-on-probabilities (ROP), and has generally proven accurate, efficient, and thus highly successful. Compared to formal IM, ROP is easily extendible to modeling considerations such as MPP, essentially estimating more allele parameters, as well as other factors such as covariates, and mixed effect models.

1.3.1.2 False associations and problematic uncertainty and founder allele frequency

Previous work from the Valdar lab has shown that the approximate nature of ROP could produce unstable and uninterpretable regression coefficients, which are often used as allele effect estimates (Zhang et al., 2014), with an extreme example in **Figure 1.7B**. They correct for this issue with the Diploffect model, a Bayesian procedure that involves multiply imputing the haplotype pairs, or diplotypes, from their probabilities. The statistical score of association can also be greatly inflated by the ROP approximation, particularly when at a locus with founder haplotype frequencies that are highly imbalanced, as in **Figure 1.7C**. It is possible that founder haplotypes will be completely lost at random loci simply through genetic drift. If there were no uncertainty, simply no parameter for that founder would be fit at that locus in the genome scan. However, when there is uncertainty, there is the potential that some minute probability mass happens to correlate strongly with the phenotype, resulting in a strong, but artificial, association score, as in **Figure 1.7A**.

1.3.1.3 Multiple imputation approach

There have been Bayesian QTL mapping procedures proposed that also involve multiply imputing the diplotypes from the probabilities (Sen and Churchill, 2001; Durrant and Mott, 2010). **Chapter 4** describes a conservative multiple imputation approach that foregoes a fully Bayesian approach for the sake of computational efficiency. A related problem is also described, in which a founder allele is rarely observed but now with strong certainty. In this situation of unbalanced certain data, shrinkage approaches (Wei and Xu, 2016) should be used, for which two different approaches are discussed.

Variant association, similar to methods used in human GWAS, is also an alternative to IM, or what could also be called haplotype-based association. Haplotype-based association has some advantages to variant association, such as implicitly modeling a more complex system, such as the local epistasis in the region. However, these strengths are contingent on the stable presence of the various haplotypes, and that they are reasonably estimable. When this is strongly violated, the simpler variant association model can be more stable and powerful, which is a topic of **Chapter 5** and published as (Keele et al., 2018).

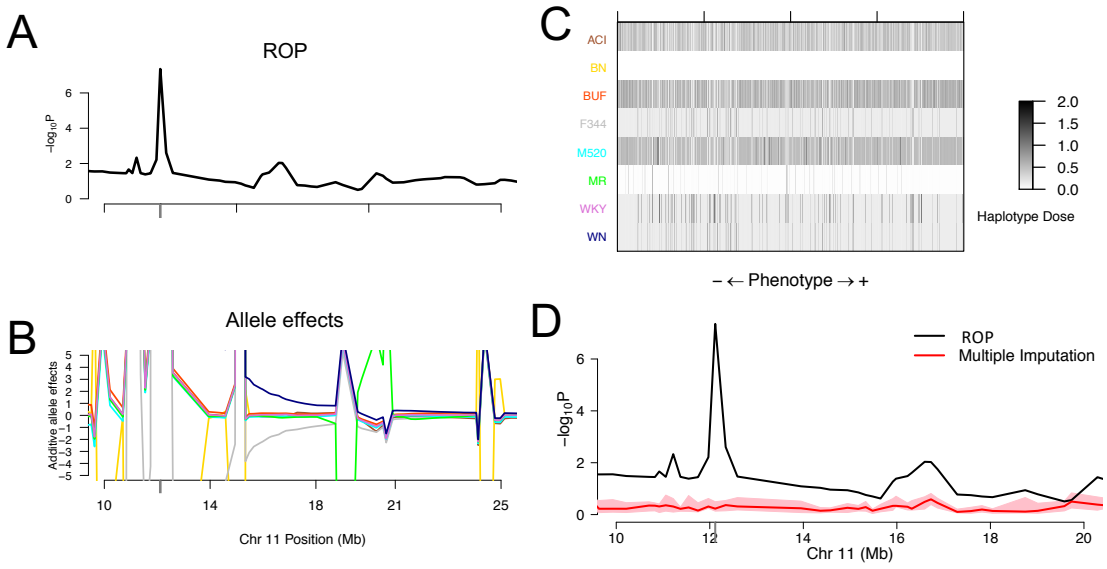

Figure 1.7: A surprisingly sharp association signal is observed for a phenotype in a large population (> 700) of HS rats using ROP (A). The allele effects are highly unstable, particularly around the QTL peak (B). In particular, the BN allele (yellow) appears to approaching negative infinity. A representation of the haplotype dosages at the peak reveals that at the putative QTL there are issues with uncertainty (founders F344, WKY, and WN are largely indistinguishable for many individuals) and imbalance in founder contributions (MR is rarely observed and BN appears to have been lost) (C). A vertical column of the grid represents the haplotype dosage vector of a single rat, with rats being ordered horizontally with respect to phenotype. No substantial founder effects are visually distinguishable, with the extreme BN effect appearing to be an artifact of the problematic uncertainty at the locus. Comparison of ROP (black line) and MI (red line and 95% confidence interval on the median association across imputations in pink) in the region of the sharp association peak. The signal is completely removed, likely because the BN allele is never sampled.

1.3.2 Analysis of heterogeneous stock rats

1.3.2.1 Imputed SNP association

The HS rat population that produced the data analyzed in **Figure 1.7** is highly unbalanced with respect to founder haplotype dosage cumulatively across all loci (**Figure 1.8A**). Though haplotype reconstruction poorly distinguished certain founders at some loci, the information content on the simpler SNP genotype can be more complete, resulting in stable association scans (**Figure 1.8B**), and ultimately produced three QTL regions for two different phenotypes, retroperitoneal fat pads (RetroFat) and body weight. The causal variants that induce QTL are usually not obvious, the region instead potentially containing from a handful of genes and variants to hundreds, thus **Chapter 5** also focuses with on quantitative fine-mapping approaches used in order to identify and prioritize candidate genes and variants under the QTL.

1.3.2.2 Fine-mapping approaches

A variety of approaches were used to assess variants within candidate genes that fell in the QTL regions. The Diploffect model (Zhang et al., 2014) was used to characterize founder haplotype effects at the QTL, which are useful for potentially identifying variants with alleles that are distributed amongst the founders such that they match these effects patterns. LLARRMA-dawg (Sabourin et al., 2015), a tool designed to simultaneously model and select important SNPs from within a GWAS hit region, significantly reduced a wide QTL region. Protein modeling (Prokop et al., 2017) was performed on candidate genes with variants that corresponded with the allele effects and fell within or near the QTL regions to assess the predicted effect of the variant alleles on protein function. Finally, gene expression as a possible mediator of the QTL effect on phenotypes was also investigated.

1.3.2.3 Gene expression as mediator of QTL effect on phenotype

Mediation approaches have recently been applied to genomic data (Battle et al., 2014; Chick et al., 2016; Roytman et al., 2018), providing avenues for confirming signals as well as potentially teasing apart the underlying relationships between the levels of the biological data. In the context of the HS rats, collaborators collected gene expression data from the liver on a large subset of the sample populations. Only expression levels of genes local to the QTL signal were considered, greatly

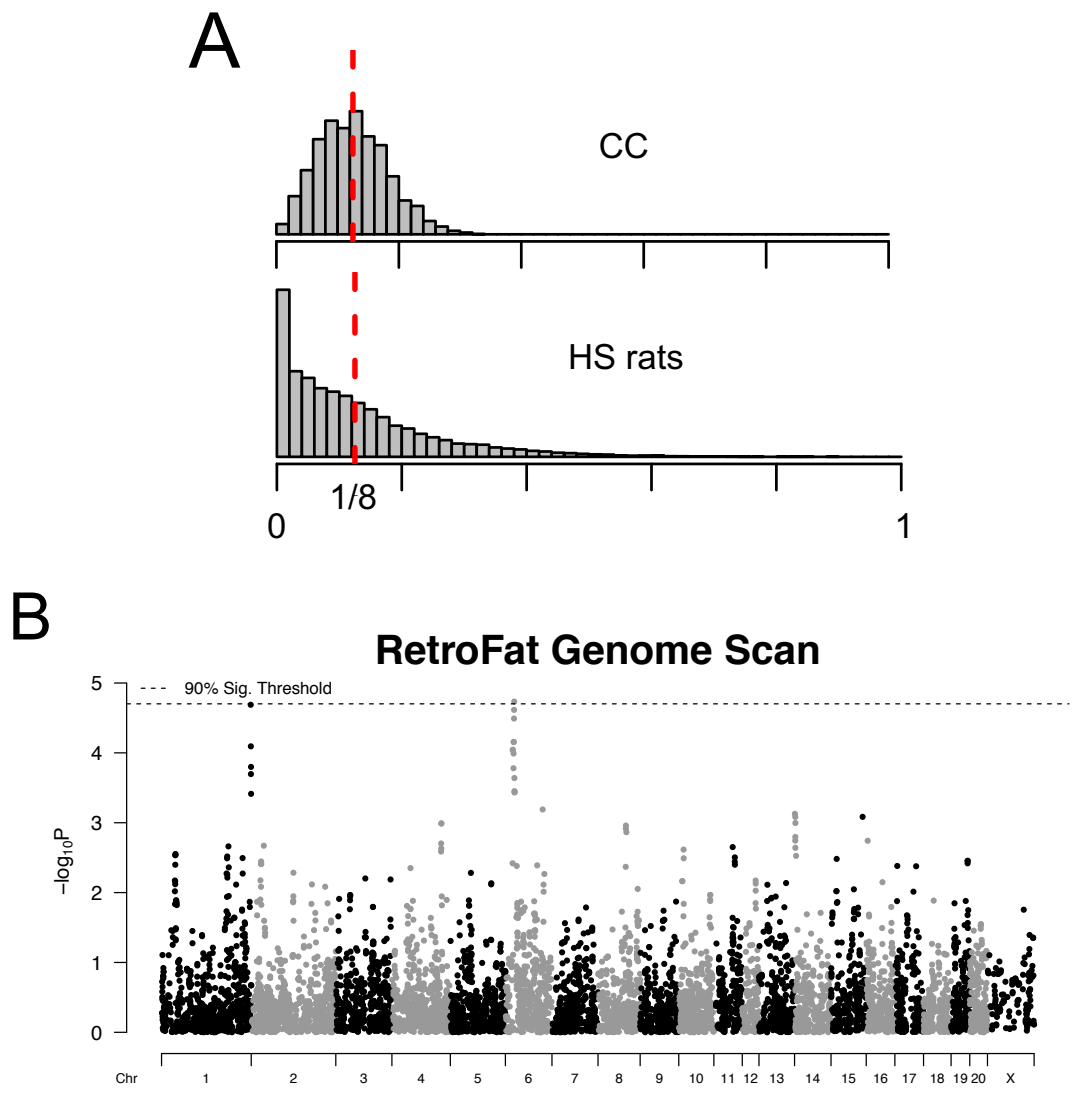

Figure 1.8: The histograms of founder haplotype allele frequencies for the CC and the HS rat population (A). The distribution in the CC is centered around 1/8, the expectation for a balanced population descended from eight founders. In contrast, the HS rats, also descended from eight founders, have a large enrichment in small frequencies, particularly near-zero, consistent with the observation that founder haplotypes have been routinely lost at loci across the genome. This pattern of haplotype distribution is problematic to haplotype-based association, a topic discussed in **Chapter 4**. An alternative to the conservative multiple imputation procedure is to impute SNP genotypes from the haplotype probabilities, potentially correcting for genotyping errors and no calls, and do variant association. Whereas there may be poor information to distinguish haplotypes, the simpler SNP imputation may be well-informed, resulting in stabler and more powerful association scans (B).

reducing the computational and testing burden, and a simple model of mediation was used (Baron and Kenny, 1986).

1.3.3 Integrative mediation analysis of gene expression and chromatin accessibility

The latter portion of **Chapter 5** demonstrates a range of ideas as well as quantitative tools for delving further into QTL findings. Mediation, and other causality-oriented approaches such as Mendelian Randomization (Smith and Ebrahim, 2003; Lawlor et al., 2008), represent exciting areas of research that leverage the big data that are being collected, with multiple dimensions per individual, sometimes referred to as multi-omics, to answer questions about and better understand the relationships between the levels of data, with particular focus on the relationships at play in the flow of information from gene to phenotype (Degner et al., 2012; Pai et al., 2015; Battle et al., 2015; Alasoo et al., 2018; Wu et al., 2018). **Chapter 6** further explores this topic, by investigating genetic regulation of gene expression and chromatin accessibility, as well as the potential relationship between them, genome-wide, in CC mice through an integrative mediation analysis.

1.3.3.1 Description of CC data and analyses

This project is highly collaborative with members of the Furey Lab, as well as collaborators at Texas A&M (Ivan Rusyn) and NC State (Fred Wright) and the results are preliminary. As such, the description in **Chapter 6** will be brief, and focus on the methodology, which relates to the research focus of this dissertation. The data for these analyses consist of RNA-Seq (gene expression) and ATAC-Seq (chromatin accessibility) in three tissues (lung, liver, and kidney) for only 47 CC strains with a single observation per strain. QTL mapping through a multi-stage conditional fitting approach (Jansen et al., 2017) was performed for both expression (eQTL) and chromatin accessibility (cQTL) in each tissue, allowing for the potential of multiple QTL per phenotype. A genome-wide mediation analysis was developed and used that draws from the approach used in (Chick et al., 2016) for jointly modeling gene expression and protein levels. Despite mediation not being equivalent to causality and the undoubtedly complex and multifactorial nature of the underlying biology involved in the steps of the regulation of the flow of information from genomic DNA to protein, simplistic mediation models can detect evidence that is consistent with broad hypotheses of how the levels relate. MPP, such as the CC or DO, can be particularly powerful tools for these genome-wide mediation approaches due

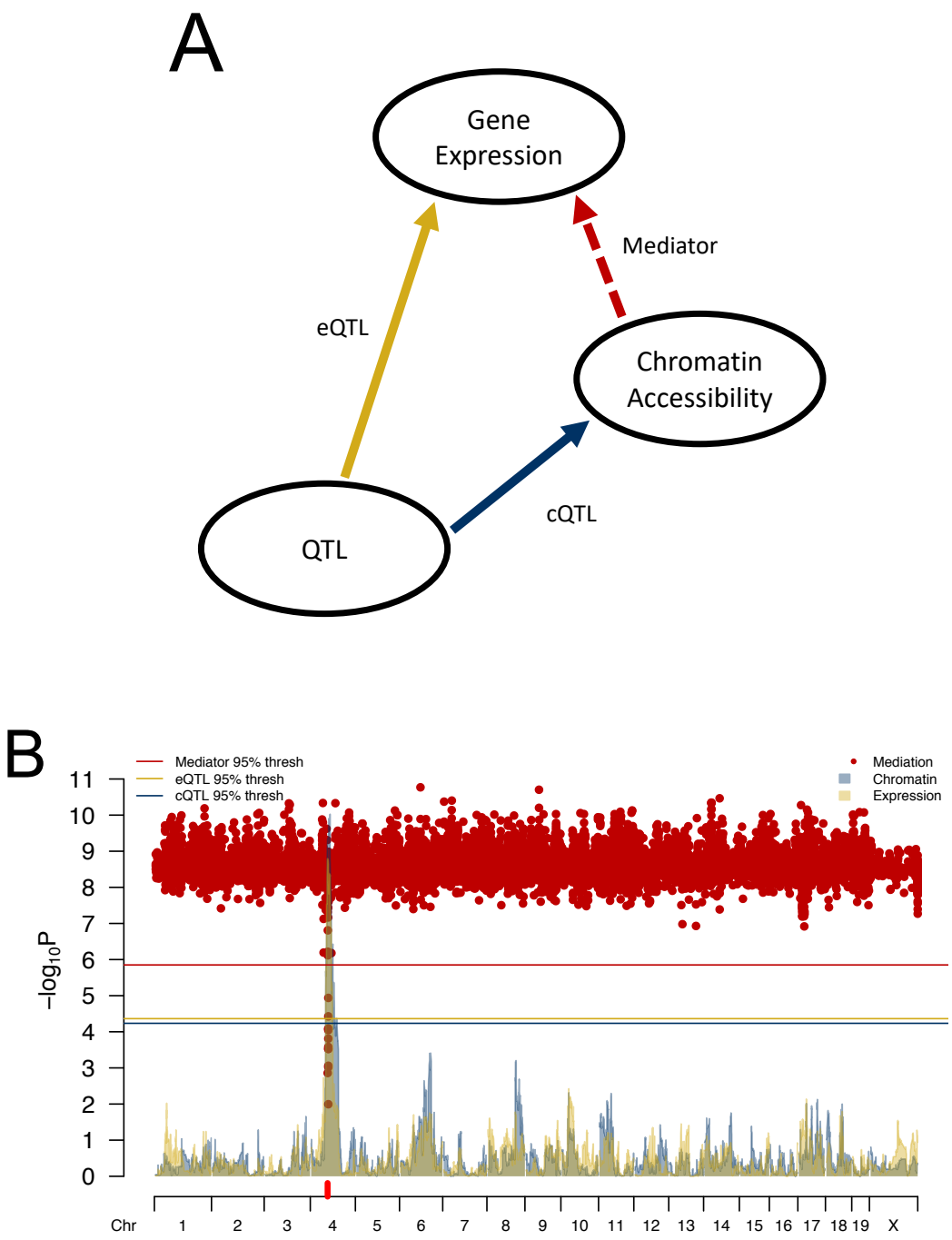

Figure 1.9: The simplistic model of chromatin accessibility as a possible mediator of the eQTL effect on gene expression (A). This model was tested for all genes with detected eQTL in lung, liver, and kidney tissue for 47 CC strains with only a single observation per strain. Example of genome-wide significant mediation (B). The gene *Alad* has a significant eQTL, local to its transcription start site (TSS) (yellow) on chromosome 4, in kidney tissue. The chromatin accessibility in the region also has a significant local cQTL (blue). Significant mediation is detected in the area (red), by the drop in p-value that reflects chromatin accessibility out-competing the eQTL when included in the null and alternative models of the genome scan of *Alad* expression.

to the ability to characterize associations with respect to the founder haplotypes (allele effects), and potentially replicate findings or design downstream experiments in related populations.

1.4 Summary

This dissertation describes a range of methods and analyses for use with MPP. **Chapters 2** and **3** describe approaches for designing experiments, first using MPP pilot data in the form of the diallel to select promising bi-parental crosses, and second to design adequately powered mapping studies of the finalized CC strains. **Chapters 4**, **5**, and **6** collectively focus on genetic association analyses for MPP: QTL mapping in MPP populations with problematic founder haplotype uncertainty, imputed SNP association and fine-mapping approaches in an HS rats population, and an integrative genetic mediation analysis of gene expression and chromatin accessibility in the CC, respectively. Taken together broadly, this research presents novel methodologies for accommodating and thus harnessing MPP resources for powerful genetic experiments.

CHAPTER 2

Using the diallel to select optimal bi-parental crosses to map QTL [1]

2.1 Introduction

Geneticists commonly conduct experiments with the goal of identifying quantitative trait loci (QTL) using crosses of inbred strains of model organisms. These experiments can be costly in terms of resources, due to the organisms, their care, genotyping or sequencing, as well as the time and energy required for the experiment itself. In the face of these constraints, procedures that explore the potential set of experimental cross designs and allow researchers to select experiments with greater potential to be successful are beneficial to the field of complex traits.

Although the goals for a given experiment will be nuanced and unique to each study, the mapping portion is successful if a QTL is detected with a statistically significant signal, using established methodologies (Lander and Botstein, 1989; Haley and Knott, 1992; Dupuis and Siegmund, 1999; Broman, 2001). This outcome is not guaranteed simply due to the presence of segregating QTL in the mapping population: the experimental design may not be sufficiently powered to identify them. The power of an experiment, the probability that a non-zero effect will be recognized given that it is present, is influenced by a number of biological factors, some of which can be more easily manipulated and optimized through experimental design choices. These factors include genetic architecture, mode of action, and the variation in the population due to noise. If the genetic architecture of the trait is highly polygenic with many loci of small effect, power will be reduced compared to tests for QTL of larger effect. Similarly, mode of action (e.g. additive, dominant), for a QTL will also influence power because certain experimental designs will have differential ability to detect a given effect type. For example, a backcross (BC) cannot identify a QTL underlying a fully recessive effect when the homozygote of the recessive allele is never observed. Finally, an increase

[1] This chapter represents a mature draft of a manuscript currently in preparation, with slight modifications made for the format. Current author line and title are: Keele, G. R., Maurizio, P. L., Oreper, D., Valdar, W. Diallel-informed experimental cross selection for QTL mapping.

in variation due to noise will decrease power because the noise drowns out the true signal. Ideally, investigators would select the experiment that can best handle these factors in the given setting.

In the context of crosses of inbred organisms, one major component of the experimental design is the founder or parental strains. The selection of parental strains allows the investigator to control the genetic background of the experimental population, which can greatly influence the previously mentioned biological factors, and ultimately influence the potential for mapping success. For example, a trait could be highly polygenic and have loci with complex modes of action within natural populations, but much of the genetic and phenotypic variation becomes fixed within two closely related inbred strains. The reduced genetic variability can impact all of the biological factors: the complexity or polygenic nature of the genetic architecture by fixing many of the loci, the mode of action by limiting the potential for epistatic effects through less segregating variants, and the variance attributable to noise through the reduction in phenotypic variability.

The ability to strongly influence the sources of variation in the population is important to consider. If the QTL explains a large proportion of the variance in the population, a simple cross will be well-powered to identify the QTL, even if its effect is small. The balance between the variance attributable to the QTL versus how generalizable the experiment is to natural populations is important to consider when making decisions about experimental design. Ultimately a finding that is characteristic of only a very unnatural experimental population and does not generalize well to more natural, outbred ones, will greatly reduce the impact of such an experiment and even undermine the purpose of experiments of model organism in general. The ideal experiment will be well-powered to identify QTL, but also generalizable to natural populations.

The power of an experiment cannot be directly assessed because it requires knowledge of the true effect, which is unknown. Instead power calculations are performed for a range of plausible parameters, usually over varying effect sizes or sample sizes, given some type I error level and error variance, which can then be represented as power curves. Power calculations have been specifically developed and refined for simple cross designs such as F2 intercross, BC, and recombinant inbred (RI) strain panels, using an information perspective approach, which posits that the complete information is composed of the observed information and the missing information (Sen et al., 2005). These power calculations are still dependent on assumed parameters, in this case QTL effect sizes and error variances. As a result, meaningful and useful power calculations still depend on the consideration of

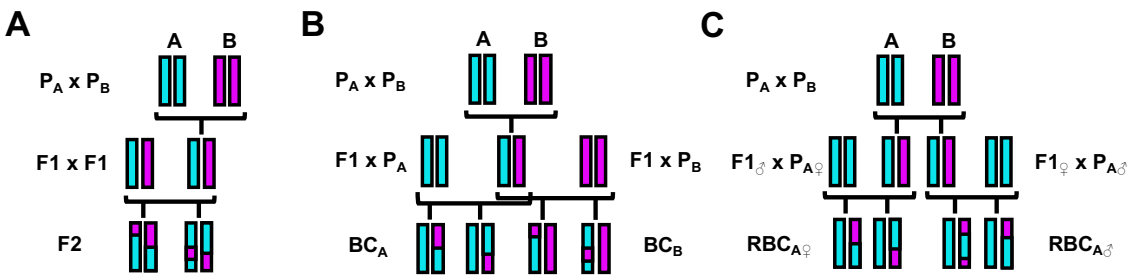

Figure 1.1: Diagrams of potential bi-parental crosses that we consider with DIDACT: F2 (A), BC (B), and RBC (C). An parental haplotype is represented as a single colored chromosome. The P and F1 generations are replicable, whereas the mapping populations are not. Of these three cross designs, only the F2 mapping population has potentially all three genotypes at a locus (A/A, A/B, and B/B), which allows for additive and dominance effects to be estimated. With traditional BC, one parental homozygote is possibly observed, depending on which parent is back crossed. By jointly analyzing RBC, it is possible to detect effects from heterozygous sites in which the parent-of-origin differs for the back crossed parental allele.

an appropriate set of values for these unknown quantities, otherwise the power estimates could be uninformative or even misleading.

Pilot data can provide information about the underlying genetic signals present in potential experiments. One source of pilot data is the inbred founder strains themselves as well as their hybrid crosses (F1). Comparisons of F1 individuals to the inbred strains can provide estimates of various genetic effects for given strains, aggregated from causal variants across the entire genome. These effects can include additive, inbred, and epistatic. An additive effect for a given strain can be estimated from averages of F1 that do not have the strain as a parent (0 copies), to averages of F1 that do have the strain as a parent (1 copy), and finally to the inbred strain itself (2 copies). An inbred effect is estimated from these same sets of crosses, but represent the average departures observed from the expectation of the hybrid according to the additive effect to its actual observed value. An epistatic effect represents departures from expectation for a specific cross of two strains, thus it is an interaction effect of the two strains.

Additional information is contained in the reciprocal crosses that compose the F1 hybrids, and can be characterized as parent-of-origin effects (POE). Reciprocal F1 crosses have the same parental strains, but the dam-sire identities are switched. The average differences between reciprocal crosses can be used to estimate the POE. QTL underlying these POE effects can be mapped using a unique BC design that we will refer to as RBC (Gonzalo et al., 2007). RBC subtly differs from what is

traditionally known as reciprocal BC, in which the F1 is the same but back crossed to the alternative parental strain. RBC have the same F1 and back crossed parent, but the dam and sire strains are reversed between reciprocal pairs; thus the parent-of-origin for each allele is known at heterozygous sites, and differences in the trait that correlate to genotype and parent-of-origin can be detected. The estimation of POE through reciprocal crosses allows researchers to add RBC to their collection of potential experiments. Though RBC are not as commonly used as F2 and BC, interest in POE has increased (Lawson et al., 2013; Bérénos et al., 2014; Connolly and Heron, 2015; Harper et al., 2014; Zou et al., 2014). Pilot data that distinguishes between reciprocal F1s allow for an even larger number of experiments to be explored and considered. These potential bi-parental mapping populations, F2, BC, and RBC, are depicted in **Figure 1.1**.

These experiments can best be explored with the full set of potential founder lines and their F1 hybrids, which represent a classic genetic experiment, the diallel. Diallel crosses have been performed in a number of traits and across a diverse set of organisms, including mating speed, female receptivity, and temperature preference in fruit fly (Parsons, 1964; Casares et al., 1992; Yamamoto, 1994); immune function, polyandry, and genetic-environment interactions in crickets (Rantala and Roff, 2006; Ivy, 2007; Nystrand et al., 2011); and heterosis and reciprocal effects in poultry (Fairfull et al., 1983). Additionally, the diallel has a long history in plant breeding (Gilbert, 1958) and numerous recent applications (Bahari et al., 2012; Ghareeb Zeinab and Helal, 2014; Dos Santos et al., 2016).

Since being described in the early 20^{th} century, statistical methodology for the diallel has seen steady advancements, from estimating the general combining ability with related F2 populations (Griffing, 1956), the use of random effects (Zhu and Weir, 1996; Tsaih et al., 2005), and the use of a Bayesian hierarchical model for a sparse diallel (Greenberg et al., 2010). Recently, (Lenarcic et al., 2012) used Bayesian hierarchical modeling of diallel data to allow for stable estimation of a large number of strain-level genetic effects (such as additive, inbred, epistatic, and maternal), and has been used to analyze a number of phenotypes and organisms, such as cranial shape (Gonzalez et al., 2016), response to treatment and infection (Crowley et al., 2014; Maurizio et al., 2018) in mice, and shoot growth in carrots (Turner et al., 2018). Even incomplete or sparse diallel data can be used for the characterization of some of the underlying strain-level genetic signals, which can then be used to evaluate the potential space of experiments, and allow for the selection of a favorable one. A

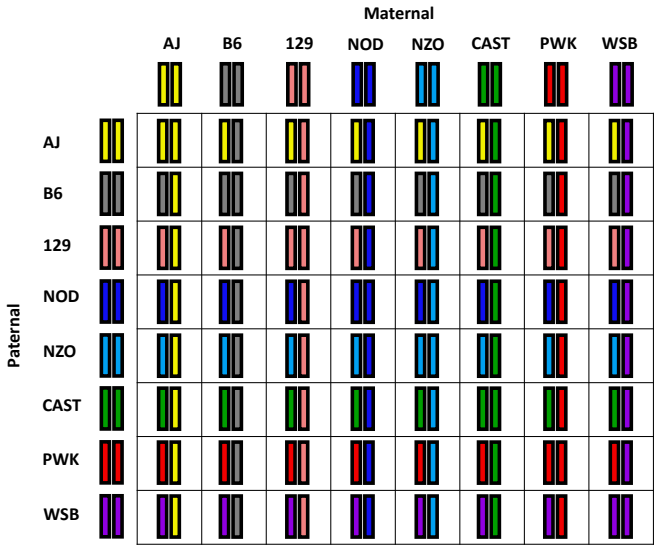

Figure 1.2: A cartoon representation of a diallel of the CC founders. Each unique strain genome is represented as a single colored chromosome. Genomes along the diagonal represent the inbred founders themselves. Off-diagonal genomes are the F1 hybrids of a pair of founders. Mirrored genomes across the diagonal represent reciprocal F1 genomes, in which the genotypes will be identical, but parent-of-origin for each chromosome will be flipped. All the genomes in a diallel are replicable, and can thus be measured on multiple individuals. Some cells of the diallel may not be observed, which reduces the ability to estimate certain strain effects.

simplified representation of a diallel, in the founders of the Collaborative Cross (CC), a multiparental recombinant inbred panel in laboratory mouse, is shown in **Figure 1.2**.

(Verhoeven et al., 2006) investigated jointly modeling diallel data with the related downstream F2 populations, and found that it allowed for the simultaneous dissection of the trait across all the populations, or characterization of strain-level effects, as well as the ability to generalize the QTL findings from the mapping populations in terms of the multiparental diallel population. We focus on the situation in which none of the F2 populations, or any such downstream cross populations, are observed, and attempt to evaluate the utility of potential crosses in terms of QTL mapping. Herein we bring together three lines of research:

1. The estimation of the power to map putative QTL of given effect sizes.

2. The characterization of genetic effects from pilot data.

3. The selection of optimal experiments through a decision theoretic approach.

We use a Bayesian hierarchical model to characterize the genetic information contained in pilot data as aggregate strain effects (Lenarcic et al., 2012). This Bayesian approach allows us to stably estimate a large number of genetic effects through the sharing of information across strains, as well as assess the uncertainty around these effects. This uncertainty is then propagated through to power calculations of potential experimental crosses, which is generally ignored in power calculation and experiment selection. Our approach will aid researchers in selecting better experiments with greater potential according to pilot data over ineffective or inefficient options. These opportunities include not only favorable experiments for mapping additive traits, which have commonly been studied, but also for mapping the QTL responsible for less well-understood effects such as POE.

2.2 Statistical Models and Methods

Our approach builds on three separate areas of research. Firstly we consider the calculation of power to map QTL given that the QTL effect θ is known. This will require the review of general concepts in quantitative genetics and statistics in the context of crosses of two inbred strains. Because in reality θ is never actually observed, we next consider the characterization of θ from pilot data. Finally we discuss the selection of optimal experimental crosses through the maximization of a chosen utility function.

2.2.1 Power to map QTL

2.2.1.1 Single QTL model of bi-parental cross

Here we review the general concepts in quantitative genetics and statistics that support the method used by (Sen et al., 2005) for power calculations of traditional crosses like the F2 and BC. Consider this model:

$$y_i = \text{QTL}_i + G_i + E_i + \epsilon_i \tag{2.1}$$

where y_i is the phenotype of individual i, QTL_i is the effect of the QTL for individual i, G_i is the effect of other genetic elements for individual i, E_i is the effect of environmental factors for individual i, and ϵ_i is the random noise for individual i. G_i and E_i are un-modeled, and can thus be

collapsed with ϵ_i into a single error term ε_i.

$$y_i = \text{QTL}_i + \varepsilon_i \tag{2.2}$$

where $\varepsilon_i \sim \text{N}(0, \sigma^2)$ with σ^2 representing the error variance in the data. The QTL effect is a vector, traditionally parameterized as additive and dominant effects (Lynch and Walsh, 1998). This can be formulated in a traditional regression framework:

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ &= \mathbf{X} \begin{bmatrix} \mu \\ \alpha \\ \delta \end{bmatrix} + \boldsymbol{\varepsilon}, \end{aligned} \tag{2.3}$$

where $\mathbf{y}$ is the phenotype vector, $\mathbf{X}$ is the design matrix that we will define further, $\boldsymbol{\beta}$ is the vector of effects composed of μ, the overall phenotypic mean, α, the additive effect of the QTL, and δ, the dominance effect for the QTL, and $\boldsymbol{\varepsilon}$ is the vector of errors. Consider an F2 or BC of strains A and B, with the genotype of an individual represented in terms of strain identity, denoted in the subscript. α is the midpoint of the difference between the homozygotes:

$$\alpha = \frac{\text{E}(y_{AA}) - \text{E}(y_{BB})}{2} \tag{2.4}$$

δ is the deviation of the heterozygote from the average of the homozygotes:

$$\delta = \text{E}(y_{AB}) - \frac{\text{E}(y_{AA}) + \text{E}(y_{BB})}{2} \tag{2.5}$$

Table 2.1 lists Eq 2.3 parameterized in terms of these QTL effects. This parameterization maintains the identifiability of all the effects, though it may not be as intuitive to researchers accustomed to more traditional regression models used commonly in genome-wide association studies.

Returning to the formulation of the model in Eq 2.6, the variance of the model can be characterized as follows with the assumption that there is no covariance between the QTL effect and the

Table 2.1: Model of QTL effect on the mean for F2 and BC

			Probability[a]		
Genotype	E(y)[b]	$\mathbf{x}$[c]	F2	BC$_A$	BC$_B$
AA	$\mu+\alpha-\frac{\delta}{2}$	$\begin{bmatrix} 1 & 1 & -\frac{1}{2} \end{bmatrix}$	$\frac{1}{4}$	$\frac{1}{2}$	0
AB	$\mu+\frac{\delta}{2}$	$\begin{bmatrix} 1 & 0 & \frac{1}{2} \end{bmatrix}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
BB	$\mu-\alpha-\frac{\delta}{2}$	$\begin{bmatrix} 1 & -1 & -\frac{1}{2} \end{bmatrix}$	$\frac{1}{4}$	0	$\frac{1}{2}$

[a]Mendelian inheritance probabilities based on independent assortment of alleles A and B for specified bi-parental cross.

[b]Parameters as defined in Eq 2.3, Eq 2.4, and Eq 2.5.

[c]Row vector of the design matrix $\mathbf{X}$ in Eq 2.3.

Table 2.2: Variance attributable to QTL effect for F2 and BC

Model	Parameter[a]	F2	BC$_A$	BC$_B$
General		$\frac{1}{4}\delta^2+\frac{1}{2}\alpha^2$	$\frac{1}{4}(\alpha+\delta)^2$	$\frac{1}{4}(\alpha+\delta)^2$
Fully additive	$\delta=0$	$\frac{1}{2}\alpha^2$	$\frac{1}{4}\alpha^2$	$\frac{1}{4}\alpha^2$
A dominant	$\delta=\alpha$	$\frac{3}{4}\alpha^2$	0	α^2
B dominant	$\delta=-\alpha$	$\frac{3}{4}\alpha^2$	α^2	0
Fully dominant	$\alpha=0$	$\frac{1}{4}\delta^2$	$\frac{1}{4}\delta^2$	$\frac{1}{4}\delta^2$

[a]Parameters as defined in Eq 2.3, Eq 2.4, and Eq 2.5.

error,

$$\text{Var}(y) = \text{Var}(\text{QTL}) + \sigma^2 \tag{2.6}$$

The background genetic and environmental variation are captured in σ^2; here we focus on the variability due to the QTL. E(y) will vary depending on the genotype, which will vary probabilistically according to the type of cross, as described in **Table 2.1**. As example, for an F2 cross, the Pr$(AA) = \frac{1}{4}$, Pr$(AB) = \frac{1}{2}$, and Pr$(BB) = \frac{1}{4}$. The variance of a random variable X is defined as Var$(X) =$ E$(X - \text{E}(X))^2$. The variable X in this setting is QTL, which is the categorical genetic state at the QTL. The expectation of X is E$(X) = \sum_{x\in\mathcal{X}} x\text{Pr}(X = x)$. Based on the genotype probability for a given cross, the variances due to the QTL in terms of the QTL effects are presented in **Table 2.2**.

The mode of action of the locus impacts the variability in phenotype due to QTL within a cross type, as seen in **Table 2.2**. This is particularly noticeable in the BC experiments, where certain modes of action produce no variance. If the locus is recessive (or conversely dominant), the genotype with differing phenotype will not be observed, and nor will variation due to QTL. Finally, cross type also impacts the QTL variance, which is also clear in **Table 2.2**. Increasing the variance attributable to the QTL will increase power to map the QTL; in contrast, increasing the overall variance that is

attributable to noise (un-modeled background genetic factors or environmental factors) will reduce the significance of statistical tests, and thus decrease the power.

2.2.1.2 Power calculations

Analytical power calculations are generally based upon some null distribution for a statistic of interest as well as some range of values for the statistic that will be observed in the experiment. Consider θ, some function of the QTL effects α and δ, as the parameter of interest. We wish to calculate the probability of mapping the QTL that results in θ. In terms of the association modeling, a natural null hypothesis is $H_0 : \theta = \theta_0$ with $\theta_0 = 0$, that there is no QTL effect. The alternative hypothesis is $H_A : \theta \neq \theta_0$. By specifying a model for the data, or more precisely the distribution of the error term of the model, the likelihood $\mathcal{L}(\theta)$ can be evaluated. The likelihood ratio test (LRT) statistic, $T = -2 \log \frac{\mathcal{L}(\theta=0)}{\mathcal{L}(\theta=\widehat{\theta})}$, where $\widehat{\theta}$ is a proposed estimate of θ, can be used to perform power calculations.

To use the LRT statistic for power calculations, a significance threshold and corresponding statistic distribution for T are necessary. The traditional scale of significance used in the linkage and QTL fields is the $\log_{10}$ likelihood ratio or LOD (logarithm of odds) score. Historically a LOD score of 3 ($2\log(10) \times 3$ on the likelihood ratio scale) has been used as a significance threshold, meaning approximately that the data support the alternative model over the null model 1000 to 1. A more stringent significance threshold than 3 can be used to further reduce the risk of false positives or possibly account for a multitude of tests (though it is worth noting these tests will not be fully independent). Given some significance threshold C is chosen to determine genome-wide significance; if $T \geq C$ for some locus, the null hypothesis is rejected. The threshold C will affect the the true positive and false positive rates, and more important to our topic, the power.

Statistically, power is the probability that the null hypothesis is rejected given that alternative hypothesis is true. The LRT T is the statistic upon which the power calculations are drawn, thus the power will be $\Pr(T \geq t | \theta \neq \theta_0)$ where t is the observed statistic produced by the data. With the LRT statistic, when the models are nested and the maximum likelihood estimate (MLE) is used ($H_A : \theta = \widehat{\theta}_{\text{MLE}}$), as they are in this case, and the null model is true, T is asymptotically χ^2_k distributed, where k is the degrees of freedom, the difference in number of parameters between the models. A power calculation from this distribution would not be useful because it would represent

the probability that the null hypothesis is rejected when there is no genetic effect, or the false positive probability. The power is rather based on the alternative hypothesis being true, $\theta \neq \theta_0$, and thus χ^2_k distribution is inappropriate. When the alternative hypothesis is true rather than the null, that $\theta = \widehat{\theta}_{\text{MLE}}$, T is proportional to the noncentral χ^2 distribution with noncentrality parameter $(\theta - \theta_0)^T \mathcal{I}(\theta)(\theta - \theta_0)$ where $\mathcal{I}(\theta)$ is the expected Fisher information matrix. We model the data with a Gaussian mixture distribution with a shared residual variance, which naturally extends from the bi-parental cross statistical model. A key feature of this model is that the LRT reduces to the variance attributable to the QTL as a function of effects that we presented in table 1. This variance parameter is scaled by σ^2, which sets the variance of each Gaussian component to 1. Thus the power calculations are intuitively a function of the effect size, the proportion of the variance explained by the QTL (effect size combined with residual error variance), and the sample size.

It is important to note that the actual $\widehat{\theta}_{\text{MLE}}$ cannot be calculated because no actual cross data for QTL mapping is observed, but the underlying theory of the method assumes that the alternative θ is the MLE estimator. σ^2 is also never actually known, but we estimate it from the information present in the pilot data. The final interpretation of this power calculation is the probability that a significant result is found ($T \geq t$) given that there is some QTL effect specified in the proposed MLE estimator $\widehat{\theta}_{\text{MLE}}$ with an error variance of σ^2.

(Sen et al., 2005) develop the theory further to account for the fact that the information is generally never complete in QTL studies. The true QTL variant is most likely not observed (genotyped), but rather loci in linkage disequilibrium are, and thus contain some of the information from the QTL. They develop the theory to take into account this missing information from sparse markers (as previously described), as well as selective genotyping (genotyping study individuals on the tails of the phenotype distribution). As a result of this, power can be reduced by not only greater error variance, but also missing information. The advancement in genotyping technology is generally leading to denser markers in QTL studies, leading us to make the assumption of complete information. We directly incorporate the R package qtlDesign (Sen et al., 2007) into our method, so missing information can be specified in the power calculations. See (Sen et al., 2005) for a description of the missing information theory used.

2.2.2 Characterization of strain-level genetic effects from pilot data

The power calculations described above are dependent on known QTL effects θ, but in reality, θ is not observed. However, information about θ is contained in pilot data, which can be exploited to characterize plausible distributions for θ.

2.2.2.1 Bayesian modeling of diallel data

One potential convenient source of pilot data are the parental strains and some subset of their F1 hybrids. Direct estimation of θ is not possible because no recombinations occur between the parental haplotypes within F1 individuals, but rather strain effects that represent the accumulated effect of the segregating variants within each inbred strain can be estimated. Denote these strain effects, the vector of effects that will be defined in Eq 2.7, as ϕ to distinguish them from θ, the effect of a single QTL.

The strain-level vector ϕ can encompass effects of different modes of actions based on the strain identities of the dam and sire of an individual. These strain-level effects include additive, inbred, epistatic, and maternal. The additive effects characterize the average effect of a strain constrained to a dosage-like model. Such a simple model is not always sufficient to accurately model data, such as the situation that an F1 hybrid is not approximately the midpoint between the parental strain phenotypes. We account for this potential deviation from additivity with an inbred effect, which is in contrast to the more traditional view of non-additivity as dominance. This parameterization of the model is appropriate for our pilot data because, considering J parental strains, there will be $J(J-1)$ possible F1 hybrids, and only J inbreds. When J is greater than 2, which is likely, the number of possible hybrid F1 will outnumber the J strains. Thus modeling the state of being outbred as the default state more intuitively matches the structure of our data.

Epistatic and maternal effects represent other potential sources of deviation from strict additivity. Epistatic effects are essentially an interaction between strains, thus allowing a specific F1 hybrid to deviate from its additive expectations. Maternal effects can capture strain-specific POEs where there is an average difference between reciprocal F1. As demonstrated in (Lenarcic et al., 2012), consider pilot data that are some subset of the J inbred strains and their F1. The strain identities of dam, sire, and dam-sire pair for individual i are indexed as $j[i]$, $k[i]$, and $(j,k)[i]$, respectively. We model the

pilot data as

$$y_i = \mu + \underbrace{a_{j[i]} + a_{k[i]}}_{\text{additive}} + \underbrace{I_{\{j=k\}}(b_j + \beta_{\text{inbred}})}_{\text{inbred}} + \underbrace{I_{\{j\neq k\}} v_{(j,k)[i]}}_{\text{epistatic}} + \underbrace{m_{j[i]} - m_{k[i]}}_{\text{maternal}} + \varepsilon_i, \tag{2.7}$$

where y is the continuous phenotype value, μ is the intercept, a is a strain-specific additive or dose effect, β_{inbred} and b are respectively a general inbred effect and a strain-specific inbred effect that are included only if individual i is inbred, v is a strain-by-strain interaction effect that we will call an epistatic effect and is only included if individual i is outbred, m is a strain-specific maternal effect, and ε_i is the individual-specific noise (deviation from the model expectation) and is distributed: $\varepsilon_i \sim \text{N}(0, \sigma^2)$. The model can also include important covariates, such as sex, that need to be adjusted for as fixed effects. The complete set of founder strains and all their reciprocal F1 hybrids represent what is called a diallel, which would allow for the estimation of the full set of strain effects described. Although an incomplete diallel cannot estimate all the strain effects, it still provides information that can be used to estimate $\boldsymbol{\phi}$.

2.2.2.2 Prior specification

Following the lead of (Lenarcic et al., 2012), we use conjugate priors for the parameters in the model. For example, the strain-level additive effects are distributed following $a \sim \text{N}(0, \tau_a^2)$. For fixed effect terms, such as β_{inbred}, τ^2 is set to 10^3. For the variance parameters, consider σ^2 which is distributed following $\sigma^2 \sim \text{IG}(\nu/2, \psi/2)$. We set the hyper parameters ν and ψ to 0.02 and 2 respectively. These represent diffuse priors, with the intention of allowing the information in the data to inform the estimates. The hyper parameter values can be adjusted within DIDACT.

2.2.2.3 Strain-level effect to QTL effect

Transitioning from strain-level genetic effects $\boldsymbol{\phi}$ to the effect of a single QTL θ requires some strong assumptions. Pilot data consisting solely of F1 individuals cannot provide information about specific loci or the number of loci contributing to a strain effect; there are an infinite number of genetic architectures that can explain a given strain effect. It is possible that conducting a small set of F2 crosses and investigating the variability in phenotype for the resulting population could provide information about the trait genetic architecture, such as distinguishing between highly polygenic

and oligogenic traits, but here we focus on using only F1. We make the assumption that the strain effects represent the effect of a single QTL, which is somewhat biologically unlikely but provides a straightforward approach to connect information in the pilot data to the power calculations. We use Eq 2.7 to produce expected phenotype values for a given cross of two strains, assuming the trait is controlled by a single QTL. Consider comparing strains A and B, Eq 2.7 produces E(y_{AA}), E(y_{AB}), and E(y_{BB}). From these expected values, we can estimate traditional single QTL additive and dominant effects, α and δ respectively, using Eq 2.4 and Eq 2.5. These estimates along with estimates of σ^2 can then be used with the power calculation machinery described before. Different QTL effects will be estimated from the model in Eq 2.7 for different potential crosses of inbred strains.

2.2.3 Decision theoretic approach

Different inbred strains will possess differing segregating variants to potentially identify. We use our model of pilot data to make predictions for some set of possible experiments, which can be viewed from a decision theoretic (Raiffa and Schlaifer, 2000) perspective as a decision space. Let us define $\mathcal{A}$ as the set potential experimental crosses. Considering n inbred strains, $\mathcal{A}$ could contain all of or some subset of the $\binom{n}{2}$ potential F2 crosses and $3n(n-1)$ potential BC.

2.2.3.1 Power as utility function

Let an element of $\mathcal{A}$ represent a specific action $\mathfrak{a}$, in this setting, a cross experiment that has corresponding single QTL effect composed of α and δ. If we define Q to be a binary variable that the QTL that causes θ (α and δ) is successfully mapped:

$$Q : \begin{cases} q = 1 & \text{QTL is mapped} \\ q = 0 & \text{QTL is not mapped} \end{cases}$$

Pr($Q = 1|\mathfrak{a}$) represents the power that the QTL is successfully mapped, and can be calculated using the noncentral χ^2 distribution described previously. We next define $\mathcal{C}$ to be the consequence or experiment outcome space for a QTL mapping experiment, where $c = \{q_1, \ldots q_p\}$ is the specific joint mapping outcome of the p QTL that the cause the strain effects. This step generalizes the

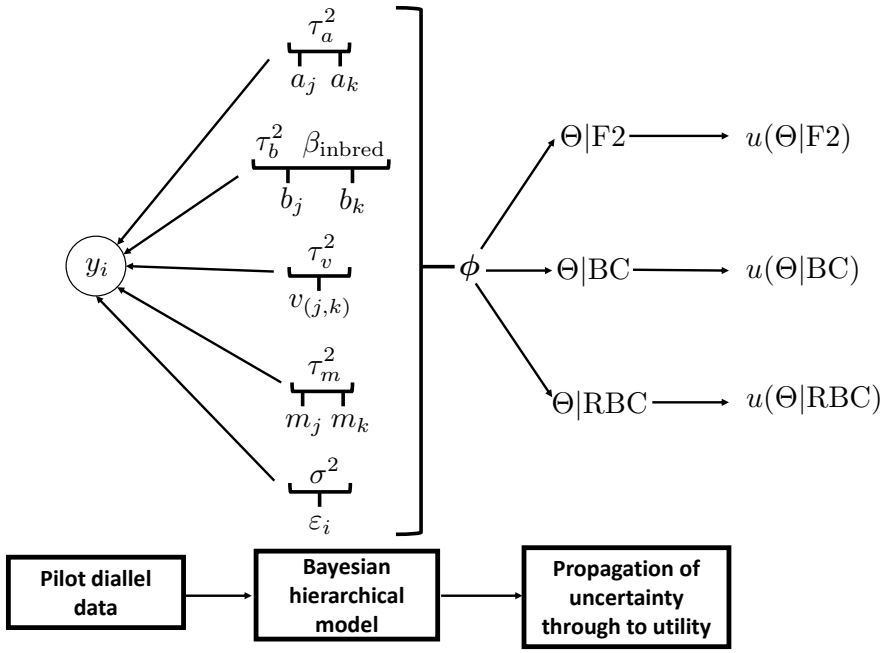

Figure 1.3: Illustration of the Bayesian hierarchical model that is fit within DIDACT, and then propagated through to a utility function. The diagram represents a single sample from a Gibbs sampler, though our decision theoretic approach would be compatible generally with other MCMC procedures. Strain-level effects are sampled based on the pilot diallel data, collectively referred to as ϕ. A sample of ϕ is then mapped using functions that draw from Eq 2.4 and Eq 2.5 and **Tables 2.1** and **2.2** to $\theta|\mathfrak{a}$, with θ representing the effect of a single putative QTL in a bi-parental cross and $\mathfrak{a}$ a specific type of cross of two specific founder strains. We collectively refer to all θ from the possible F2 crosses as $\Theta|$F2, as well as for BC and RBC. Effectively Θ are functions of ϕ, which are then used as inputs into the utility function $u(.)$, in our case, a putative QTL power estimate. This process is repeated for s samples from the MCMC procedure, which allows for posterior estimates on utility.

problem to multiple QTL rather than a single one, thus allowing us to reduce the assumption of a single QTL causing the strain effect.

A utility function is an important concept in decision theory. It provides a common scale to compare potential experimental outcomes, and select optimal experiments. Alternative utility functions can be devised and easily swapped to place value on differing aspects that investigator want to prioritize. We define a utility function, $u(.)$, to map from $\mathcal{C}$ to the reduced utility space, $\mathcal{U}$, which we pose as a function of power, a natural quantity to prioritize. Consider the probability of a specific consequence, which will be a product of a function of the individual power for each QTL: $\Pr(c = \{q_1, \ldots, q_p\}|\mathfrak{a}) = \prod_{i=1}^{p} \Pr(Q_i = q_i|\mathfrak{a})$. We define $u(.)$ to be the count of p QTL that were successfully mapped: $u(c) = \sum_{i=1}^{p} q_i$. The probability of a utility υ can be calculated from subsets of $\mathcal{C}$:

$$\begin{aligned} \Pr(\upsilon|\mathfrak{a}) &= \sum_{c \in \mathcal{C}: \upsilon = u(c)} \Pr(c|\mathfrak{a}) \\ &= \sum_{c \in \mathcal{C}} I_{\{\upsilon = u(c)\}} \Pr(c|\mathfrak{a}) \end{aligned} \tag{2.8}$$

Strictly speaking, the probability of a utility is also dependent on QTL effect θ: $\Pr(\upsilon|\mathfrak{a}, \theta) = \sum_{c \in \mathcal{C}} I_{\{\upsilon = u(c)\}} \Pr(c|\mathfrak{a}, \theta)$. θ can be marginalized out through integration: $\Pr(\upsilon|\mathfrak{a}) = \int_{\theta} \Pr(\upsilon|\mathfrak{a}, \theta) \Pr(\theta|\mathcal{D}) d\theta$, where $\mathcal{D}$ represents the pilot data.The probability of this utility function provides an evaluation of the uncertainty of mapping QTL of a given effect size, but does not take into account the uncertainty of α, δ, and σ^2, which are produced from the Bayesian model. Through Gibbs sampling or some other Markov Chain Monte Carlo (MCMC) method, a Bayesian model can produce S draws from the posterior distribution of these parameters. Monte Carlo (MC) averaging allows us to take into account this extra source of variability, resulting in the posterior expected utility for cross $\mathfrak{a}$:

$$\begin{aligned} \text{PEU}(\mathfrak{a}) &= \int_{\theta} \int_{\upsilon} \upsilon \Pr(\upsilon|\mathfrak{a}, \theta) d\upsilon d\theta \\ &= \int_{\theta} \sum_{\upsilon \in \mathcal{U}} \upsilon \sum_{c \in \mathcal{C}} I_{\{\upsilon = u(c)\}} \Pr(c|\mathfrak{a}, \theta) d\theta \end{aligned} \tag{2.9}$$

where θ is the vector function of α, δ, and σ^2. The quantity $\sum_{\upsilon \in \mathcal{U}} \upsilon \Pr(\upsilon|\mathfrak{a}, \theta)$ within the PEU($\mathfrak{a}$) is the expected utility for a single draw s from the Bayesian model. This quantity is then be averaged over the QTL effect space of the posterior distribution, traversed through the MC samples. This can be summarized as a point estimate such as the posterior mean or median, or the posterior distribution of expected utilities can be plotted for a given cross $\mathfrak{a}$. Interpretations of the PEU($\mathfrak{a}$) will vary amongst utility functions, but we will focus our discussions on power as the utility being maximized.

If we assume all p QTL have the same effect size, our utility function $u(c)$, the number of p QTL that were successfully mapped, follows a binomial distribution. Consider simple case of a single QTL (p = 1), in which the binomial reduces to the Bernoulli distribution. In this setting, the PEU($\mathfrak{a}$) reduces to the posterior probability of mapping the QTL. When p is greater than one, as with a binomial variable, PEU($\mathfrak{a}$) now represents the expected number of QTL to be mapped. Our approach should be flexible to any reasonable utility function investigators can define, but we emphasize power because its PEU($\mathfrak{a}$) are easy to interpret.

2.2.4 Availability of data and software

All analyses were conducted in the statistical programming language R (R Core Team, 2018). Our R package DIDACT (Diallel Informed Decision theoretic Approach for Crosses Tool), which is available on GitHub at `https://github.com/gkeele/DIDACT`, can estimate strain-level effects from diallel data using a Bayesian hierarchical model, and then perform the posterior utility analysis. The R package BayesDiallel can alternatively be used to estimate the strain-level effects, and used as inputs to DIDACT.

DIDACT includes three diallel data sets from the CC founders (Churchill et al., 2004; Collaborative Cross Consortium, 2012; Srivastava et al., 2017), each with a number of phenotypes, described in detail in (Lenarcic et al., 2012). The CC founders represent the following inbred strains of mouse (abbreviated names in parentheses): A/J (AJ), C57BL/6J (B6), 129S1/SvImJ (129), NOD/LtJ (NOD), NZO/H1LtJ (NZO), CAST/EiJ (CAST), PWK/PhJ (PWK), and WSB/EiJ (WSB).

We also make use of an additional diallel data set in the CC founders of response to Influenza A virus (IAV) infection phenotypes, and is available at `https://github.com/mauriziopaul/flu-diallel`.

2.3 Results

We provide example analyses from diallel data of the CC founders to demonstrate our decision theoretic procedure used in the DIDACT package. Our approach depends on assumptions about the effect of a single putative QTL in a bi-parental cross (described in **Table 2.1**) given strain-level effects estimated from diallel data based on the parameterization described in Eq 2.7. This assumption is most straightforward in the case of a largely Mendelian phenotype, in which a single locus modulates the variation observed in a relatively deterministic manner, and as such, the QTL effect θ can draw from the strain-level effect ϕ wholly.

2.3.1 Mendelian phenotype

To demonstrate a straightforward application of DIDACT to a phenotype largely driven by a single locus, we use resistance to IAV infection and the *Mx1* gene. In previous work (Maurizio et al., 2018), we investigated strain-level effects in day four post-infection (D4 p.i.) body weight loss percentage in a diallel of the CC founders. The phenotype of interest is a response to infection, in which three infected animals were compared to a single mock-infected animals. Occasionally three infected animals were not observed at later time points, which we accounted for through a multiple imputation procedure that imputed unobserved animals from the posterior predictive distributions of the BayesDiallel model (Lenarcic et al., 2012). Here we use only a single imputed data set of 131 outcomes, as this example is only a proof of principle for DIDACT, and not a rigorous investigation of strain-level effects.

2.3.1.1 *Mx1* as a critical host-resistance factor in mice:

It has previously been shown that *Mx1* largely drives IAV-resistance in the CC founders, and has three major functional classes corresponding to the three subspecies of *Mus musculus*: *domesticus* (hereafter *dom*; CC founders with *dom* allele are AJ, B6, 129, NOD, and WSB), *castaneus* (*cast*; CAST), and *musculus* (*mus*; PWK and NZO) (Ferris et al., 2013). The *dom* allele of *Mx1* ($Mx1^{dom}$) was found to be null and those individuals susceptible to IAV infection, whereas $Mx1^{mus}$ and $Mx1^{cast}$ confer degrees of resistance.

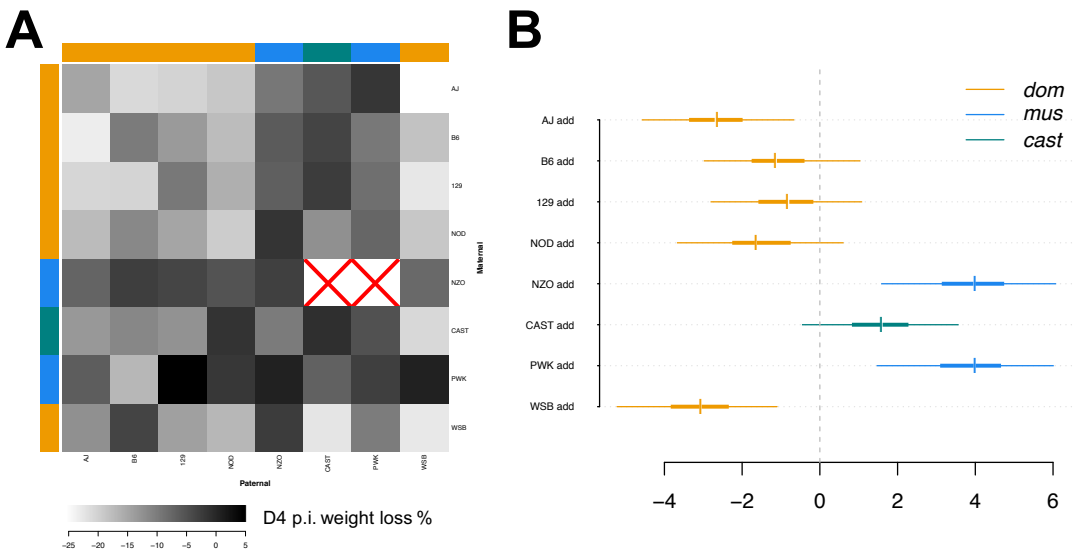

Figure 1.4: *Mx1* as a driver of IAV-resistance can be seen in the raw data, as mean day four post-infection (D4 p.i.) body weight loss percentage in a diallel of the CC founders and their hybrids (n=381 mice) (A). Squares with a red "X" represent crosses that produced no offspring. Resistance to IAV infection through the functional alleles of *Mx1* is visible and highlighted with blue ($Mx1^{dom}$) and teal ($Mx1^{cast}$) bars. Reduced to no body weight loss is observed in mice with $Mx1^{dom}$ (NZO and PWK) and $Mx1^{cast}$ (CAST), reflected in the comparatively dark horizontal and vertical bars corresponding to these founders in the diallel grid. The strain-level additive effects estimated from the Bayesian hierarchical model with DIDACT reflect the possession of a functional *Mx1* allele, with $Mx1^{dom}$ (NZO and PWK) conferring more resistance than $Mx1^{cast}$ (CAST) (B). The DIDACT-estimated strain-level additive effects are presented as highest posterior density (HPD) intervals with 95% HPD as thin lines and 50% HPD as thick lines, and posterior means and medians represented as colored ticks and white ticks respectively. The effects closely match those estimated in (Maurizio et al., 2018), which used the more complex BayesDiallel model, and also summarized over many imputed data sets.

Though IAV-resistance is largely Mendelian in that it is driven by *Mx1*, the genetic architecture of the trait in the diallel of CC founders is more complicated than a bi-allelic locus, but rather has multiple functional alleles, $Mx1^{mus}$ and $Mx1^{cast}$ in comparison to the null allele $Mx1^{dom}$. $Mx1^{mus}$ has a dominant mode of action, conferring the same resistance in $Mx1^{dom}/Mx1^{mus}$ individuals as in $Mx1^{mus}/Mx1^{mus}$, whereas $Mx1^{cast}$ is additive with $Mx1^{cast}/Mx1^{mus}$ being intermediate in IAV-resistance to $Mx1^{mus}/Mx1^{mus}$ and $Mx1^{cast}/Mx1^{cast}$. The increased IAV-resistance of $Mx1^{mus}$ and $Mx1^{cast}$ is noticeable and in the raw data and estimated strain-level effects estimated through DIDACT, highlighted in **Figure 1.4**.

2.3.1.2 Expectations of DIDACT with a Mendelian trait

Our primary expectation for the performance of DIDACT with a Mendelian phenotype is that it should favor crosses that will have segregating variants at the locus, in this case *Mx1*, in particular crosses that match $Mx1^{dom}$ with $Mx1^{mus}$ or $Mx1^{cast}$. Crosses that fix a homozygous genotypes at *Mx1* should fix much of the trait variation, and ultimately cannot detect the Mendelian locus. As expected, DIDACT largely favors crosses that result in multiple segregating *Mx1* alleles, $Mx1^{dom}$ with $Mx1^{mus}$ or $Mx1^{cast}$, shown for potential F2 experiments in **Figure 1.5** and BC experiments in **Figure 1.6**. Crosses that DIDACT predicts to be more successful than our knowledge of *Mx1* would support, such as the WSB $\times$ B6 F2 cross, likely reflect effects from the genetic background of various strains that are independent of *Mx1* (Maurizio et al., 2018). It is also important to note that this analysis of *Mx1* represents a single imputation of the multiply imputed data.

2.3.2 Complex trait

We next consider a trait that is not known to be Mendelian, but instead likely complex.

2.3.2.1 Calculated hemoglobin (cHGB):

As reported in (Lenarcic et al., 2012), blood phenotypes were measured on 626 mice, which included cHGB, an estimate of the quantity of hemoglobin in the blood (**Figure 1.7A**). The means of the raw data do not suggest clear strain-level effects like D4 p.i. weight loss % did; however, DIDACT estimates stable strain-level effects, as well as various non-zero effects across all the effect types (**Figure 1.7B**). On closer inspection, the posterior utility estimates for potential experimental

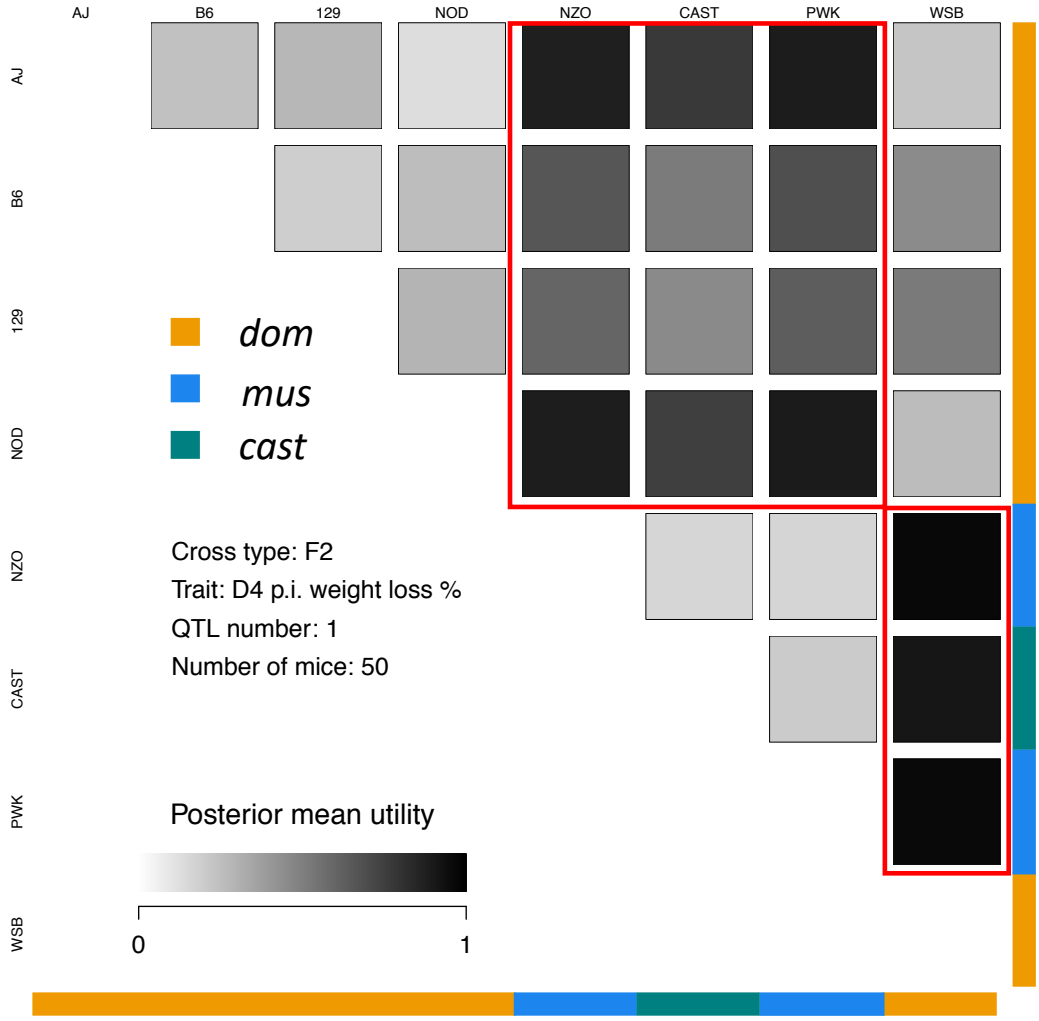

Figure 1.5: Posterior mean utility, here set to be the power to map a single QTL in 50 individuals, for the 28 possible F2 crosses of the CC founder strains. DIDACT generally estimates higher posterior mean power for F2 crosses that match a founder strain with $Mx1^{dom}$ with either $Mx1^{mus}$ or $Mx1^{cast}$, which maintains the genetic variability at *Mx1* that correlates with D4 p.i. weight loss %, and thus represent potentially powerful mapping crosses. F2 crosses of WSB with B6 and 129 have higher posterior mean power than other $Mx1^{dom}/Mx1^{dom}$ pairings, likely representing the influence of other factors specific to the WSB genetic background. Posterior mean utility for BC experiments can be seen in **Figure 1.4A**.

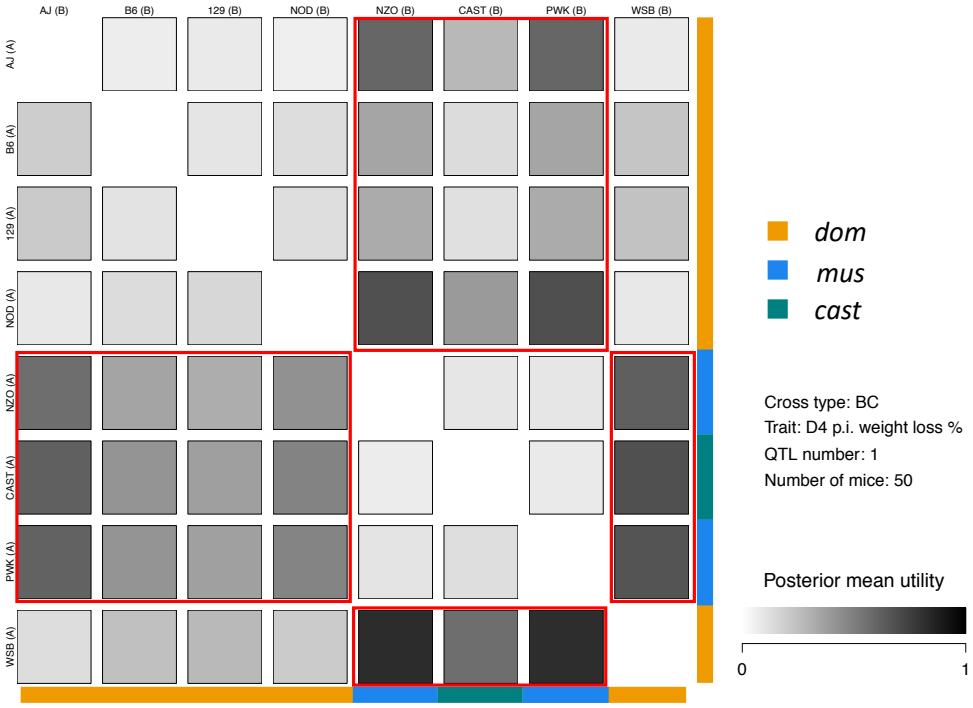

Figure 1.6: Posterior mean utility, here set to be the power to map a single QTL in 50 individuals, for the 56 possible BC experiments of the CC founder strains. The A strain, corresponding to row, is the strain that is backcrossed with the F1 in the BC, therefore the homozygous genotypes of the A strain are observed along with heterozygotes. Though less obvious than in F2 crosses (**Figure 1.5**) DIDACT generally estimates higher posterior mean power for BC experiments that match a founder strain with $Mx1^{dom}$ with either $Mx1^{mus}$ or $Mx1^{cast}$, which maintains the genetic variability at *Mx1* that correlates with D4 p.i. weight loss %, and thus represent potentially powerful mapping crosses. BC with CAST are less powerful than NZO or PWK because $Mx1^{cast}$ is additive in comparison to the dominance of $Mx1^{mus}$. BC of a strain that carries $Mx1^{dom}$ with NZO or PWK in which the $Mx1^{dom}$ strain is the backcrossed strain are more powerful than when either NZO or PWK are backcrossed, particularly with AJ and NOD. This is consistent with the dominant effect of $Mx1^{mus}$, or that $Mx1^{dom}/Mx1^{mus}$ will be more similar to $Mx1^{mus}/Mx1^{mus}$ in comparison to $Mx1^{dom}/Mx1^{mus}$ to $Mx1^{dom}/Mx1^{dom}$.

crosses, F2 and BC (**Figures 1.7C** and **1.7D** respectively), correspond to the strain-level effects. For example, the strongly negative CAST inbred effect is likely responsible for DIDACT estimating higher posterior power for BC in which the CAST parent is backcrossed with the F1. DIDACT is also estimating several non-zero strain-level maternal effects, which include AJ, B6, and PWK, suggesting that RBC may have differential posterior power.

2.3.3 Additional summaries of information

DIDACT can provide more detailed descriptions of predicted bi-parental crosses than shown in **Figures 1.5**, **1.6**, and **1.7**. At its core, DIDACT is a Bayesian hierarchical model of strain-level effects that propagates uncertainty to predetermined QTL-level utility functions, and as such, posterior intervals can be produced in addition to the point estimates. Three potential F2 crosses were selected from the full panel for cHGB (**Figure 1.7C**), and are presented in **Figure 1.8**. Posterior summaries of the distribution of utility, in this case power, median utility, predicted phenotypes per QTL genotype, and variance attributable to QTL are overlayed onto the posterior mean. Unsurprisingly, DIDACT attributes higher posterior power with crosses in which the QTL explains more of the overall variability, and in which the phenotype separate more by QTL genotype.

2.3.4 Parent-of-origin effects and RBC

There is not currently a satisfactory approach and solution for parameterizing QTL effects that contain a POE mode of action, such as exists for additivity and dominance as described in Eq 2.4 and 2.5 as well as in **Tables 2.1** and **2.2**, which ultimately limits the ability of DIDACT to make power calculations for RBC as described in **Figure 1.1C**. However, it is possible for DIDACT to characterize the utility in terms of predicted BC, but with the maternal and paternal identities fixed as in the RBC. Though the power calculation will not correspond to the design specified in **Figure 1.1C**, in which three genetic states are observed in comparison to two for BC, differences in QTL mapping power for BC that are equivalent except for the maternal and paternal statuses of the backcrossed parental strain and F1 are potentially interesting, shown in **Figure 1.9** for cHGB. Corresponding BC that have markedly different posterior utility match pairings of strains with non-zero strain-level maternal effects in **Figure 1.7B**, such as B6 $\times$ PWK, with B6 backcrossed. For this approach to RBC, DIDACT is still dependent on assumptions connecting strain-level effects in the diallel to

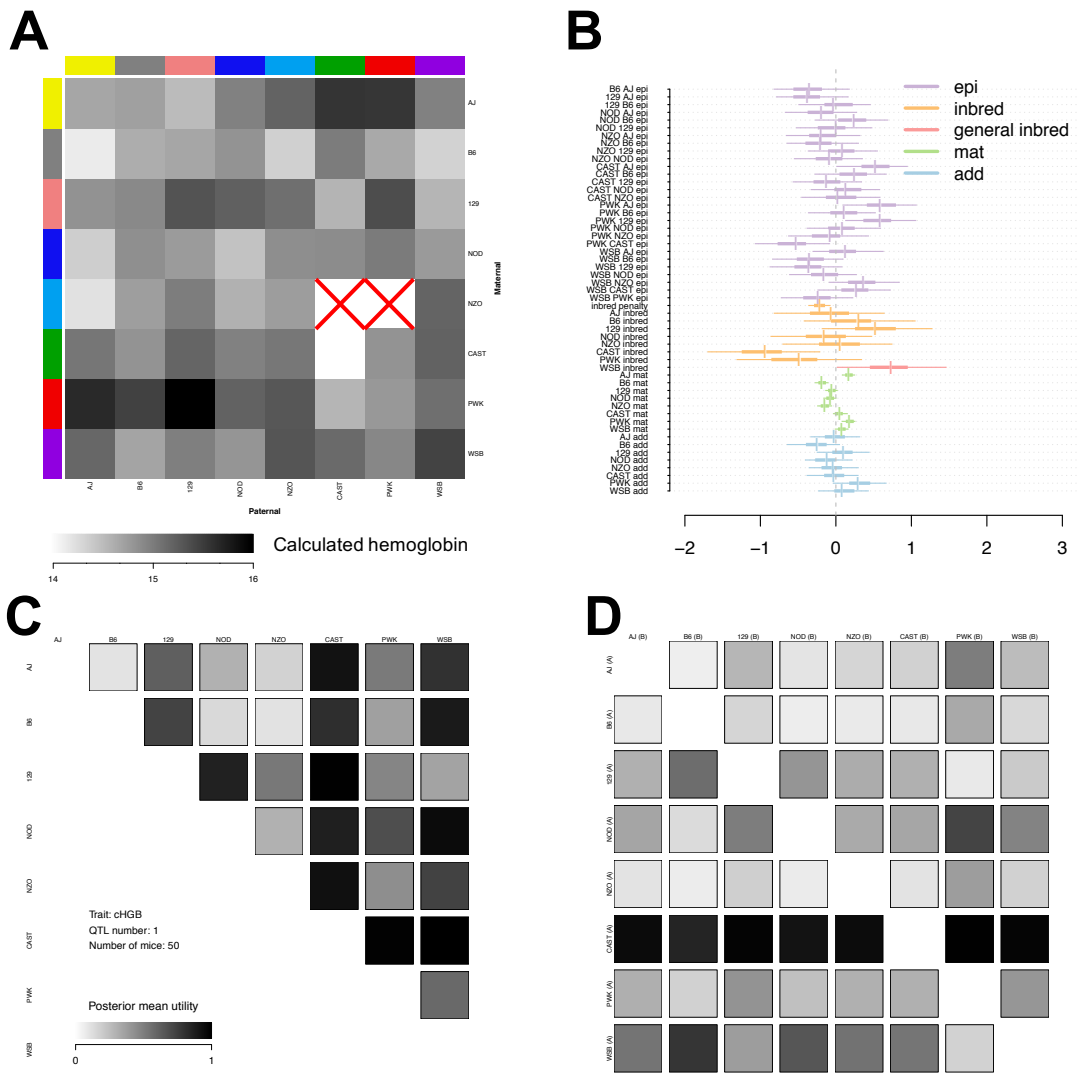

Figure 1.7: Calculated hemoglobin (cHGB) (g/dl) in a diallel of the CC founder strains composed of 626 mice. The cHGB diallel cell means of the raw data do not show the clear, consistent additive strain-level effects seen in the body weight response to IAV infection data (**Figure 1.4A**) (A). The red "X" represents crosses that did not produce viable offspring. Because the effects do not appear to correlate with subspecies, we use separate colors to label dam and sire strains. Despite the reduced level of visual clarity in the raw data, DIDACT is able to stably estimate strain-level effects, many of which are non-zero, and present across the various effect types (B). Effects are represented as HPD intervals, with 95% as thin lines and 50% as thick, and colored ticks and white ticks representing posterior means and medians respectively. This pattern of strain-level effects suggests potential complex genetic architectures underly cHGB in these strains. When DIDACT includes all of the strain-level effects into a single putative QTL effect, it results in some F2 (C) and BC (D) experiments with high posterior power. For example, the strongly negative CAST inbred effect is reflected in BC in which CAST is backcrossed having high posterior power, and low power when CAST is not backcrossed. The cHGB strain-level effects are certainly not the result of a single QTL and that assumption false, but the DIDACT results still support crosses that are likely to pair founders with highly divergent phenotypes.

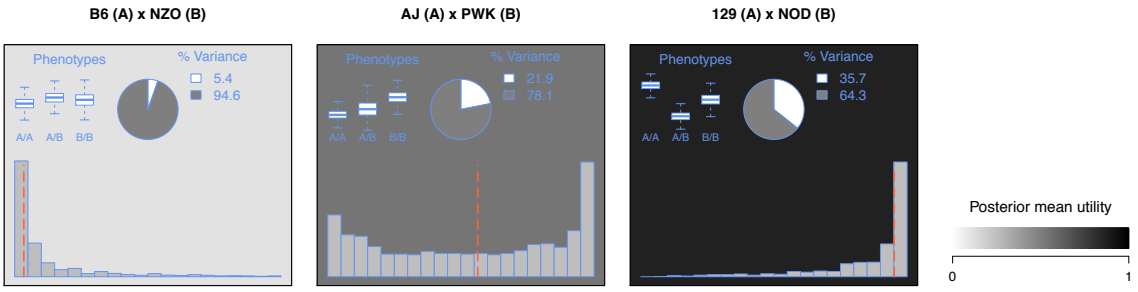

Figure 1.8: Summary plots of potential F2 crosses from DIDACT for cHGB. The full panel of potential F2 crosses are presented in **Figure 1.7C**. DIDACT allows for additional information to be overlayed on the posterior mean of the utility, which is represented by the background color. These plots include the histogram of the posterior distribution of the utility function, in this case power to detect a single QTL, the posterior median utility as a red dashed line, the posterior median variance explained by the QTL as a pie chart as well as point estimates, and posterior five point summaries of the phenotype per QTL genotype.

putative QTL segregating in a bi-parental cross; the default behavior attributes the entirety of the strain-level effects, in this case including maternal effects, to a QTL.

2.4 Discussion

We propose an experimental design approach that uses diallel data as input pilot data to characterize strain-level genetic effects with a Bayesian hierarchical model, which are then mapped with some user-defined utility function that can be used to identify promising bi-parental crosses for mapping QTL. Herein, we define utility to be QTL mapping power, though other functions could be used, so long as the strain-level effects are their inputs.

2.4.1 Assumptions connecting strain-level effect to QTL effect are wrong

DIDACT requires a strong assumption in connecting the strain-level diallel effects to putative QTL effects. We show that DIDACT performs well in a mostly Mendelian phenotype in which we can directly connect the strain-level effects to a single putative QTL. However, phenotypes that are highly heritable are often modulated through many loci, often with few or none with large effects, such as with height in humans being a clear example (Wood et al., 2014). In the vast majority of complex traits, the assumption of a single QTL absorbing all or most of the strain-level effects is

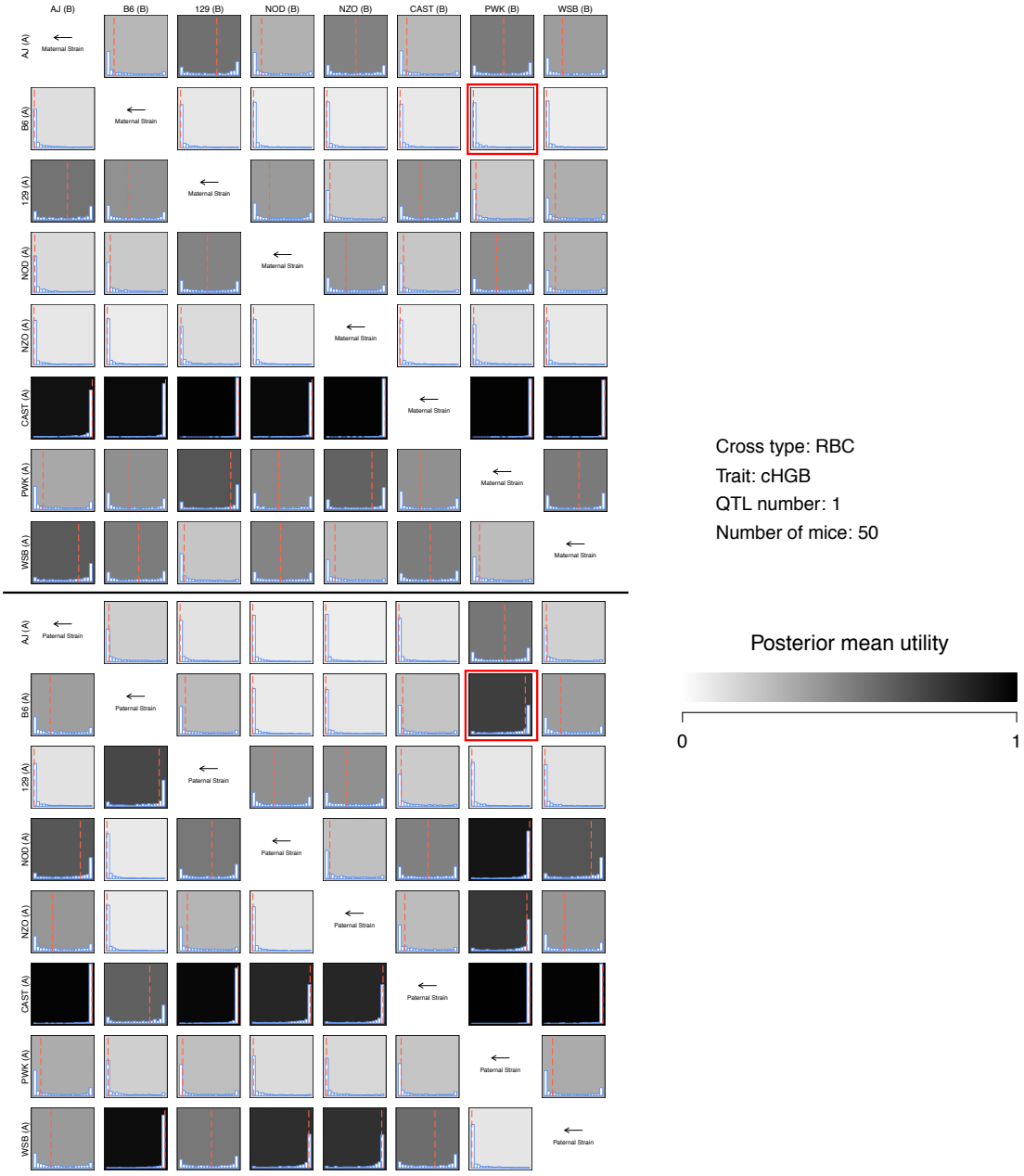

Figure 1.9: A panel of DIDACT posterior power for all possible BC in which the maternal and paternal strain identities of the F1 and backcrossed generation are fixed. Posterior histograms are included as well as posterior medians as red dashed lines. Though not a direct calculation of power for the RBC design in **Figure 1.1C**, our approach highlights the potential that strain-level maternal effects can contribute to differences in predicted RBC. The B6 $\times$ PWK BC, with B6 as the backcrossed parent, are marked with red squares, for which DIDACT predicts the BC with the backcrossed B6 as the sire (lower) as being far more likely to be successful than that with a dam (upper). These predictions reflect the non-zero maternal effects estimated for B6 and PWK in **Figure 1.7B**.

wildly optimistic. However, we posit that though the assumption is unlikely, its use as a utility function can still produce a useful analysis of potential bi-parental crosses.

We make this claim because the power calculation underlying DIDACT favors QTL that explain a large proportion of the variability in the phenotype. In fact, the power function should track closely with variability explained by QTL, which will relate to the variability explained by strain identity in the diallel in this context, and could even be used as the utility function itself. Though the interpretation of the posterior utility as an accurate power may be highly unrealistic, it will select pairings that are phenotypically distinct, which is a common criterion for selecting crosses. And, it will do so in a highly principled approach that intuitively accounts for uncertainty.

2.4.2 Genetic similarity between strains

DIDACT, in its current form, does not make use of any information regarding the similarity of the inbred strains in the diallel, which could also inform how appealing an experimental cross is in terms of fine-mapping the identified QTL. The reduced complexity cross (RCC) is a developing approach in systems genetics (Williams and Williams, 2017) in which strains that are phenotypically divergent but genetically similar are crossed, such as C57BL/6J and C57BL/6N substrains (Khisti et al., 2006; Mulligan et al., 2008; Kumar et al., 2013; Simon et al., 2013; Kirkpatrick and Bryant, 2014). RCC provide a powerful tool for fine-mapping causal variants because the genetic variability between strains are greatly reduced, restricting the set of possible causal variants to be considered.

There are a number of ways that DIDACT could be modified to incorporate genetic similarity information, probably most simply through the utility function. The utility function could be expanded to flexibly weight potential experimental crosses by the genetic similarity, resulting in posterior utilities that are informed by both phenotype and genetic similarity. We believe this highlights the potential of DIDACT, and its underlying concept in general, to be flexible to the context of the experimental system, at the hierarchical model, but particularly at the utility function.

2.4.3 Extension to multiparental populations

We present DIDACT analyses from diallels of the CC founders, which poses the question of designing experiments involving the CC themselves based on their diallel data. The CC are a multiparental (MPP), meaning that individuals descent from multiple well-characterized founders.

Extending the philosophy of DIDACT to experiments of an MPP RI panel is challenging, as the recombination events that randomize segments of the genome to allow for QTL mapping have already occurred during the generation of the recombinant inbred strains, and all strains have contributions from each founder at locations across the genome. Another approach to extending DIDACT to an MPP RI panel like the CC would be to consider the CC panel as a large sparse diallel, potentially with some off-diagonal cells representing F1 hybrids of the CC (CC-RIX) (Bogue et al., 2015) observed. DIDACT could then be adapted to select potentially interesting but unobserved CC-RIX based on the CC-specific strain-level effects. Effectively adapting DIDACT for design of MPP experiments is an area of interest for future research.

2.4.4 Summary

We describe a novel approach to using prior collected diallel data from a panel of inbred strains to inform the selection of potential downstream experiments according to a user-specified utility function, in our case, power to map QTL in bi-parental cross experiments, consisting of F2, BC, and RBC. The core of this approach, DIDACT, is to propagate the uncertainty characterized through the Bayesian hierarchical model through to the utility functions, which can be customized to the needs and constraints of the system at hand.

As proof of principle, we evaluated DIDACT in a phenotype known to be Mendelian: resistance to IAV-infection, which is largely modulated by the *Mx1* gene with a null (susceptible) and two non-null (resistant) alleles. DIDACT largely evaluated bi-parental crosses of null with non-null *Mx1* strains as having higher posterior power to map the QTL. For the non-Mendelian calculated hemoglobin, DIDACT favors crosses that pair strains with contrasting phenotypes. Though the posterior power as utility, in the sense of its nominal interpretation as power, is highly optimistic, still provides a reasonable metric for comparing potential experiments, given the available pilot data. We believe our approach can be extended in many ways, in terms of both the utility function that is being optimized and the model system, many which have sparse diallel data available in the form of strain surveys. We believe DIDACT represents a philosophical advancement in terms of good experimental design and efficient use of available resources.

CHAPTER 3

SPARCC: An R package for estimating power to detect QTL through simulated experiments in the realized Collaborative Cross [1]

3.1 Introduction

The Collaborative Cross (CC) is a panel of multiparental (MPP) recombinant inbred (RI) strains of laboratory mouse, descended from eight inbred founder strains (Threadgill et al., 2002; Churchill et al., 2004). These founders represent three subspecies of the domesticated house mouse *Mus musculus* (Yang et al., 2011), imbuing the CC panel with far greater genetic variation than traditional inbred strains, in particular, the presence of alleles inherited from wild-derived strains. The CC panel is a powerful tool that provides a genetically diverse set of reproducible genomes (Collaborative Cross Consortium, 2012; Srivastava et al., 2017).

The genetic diversity present within the CC makes it ideal for modern genetic studies of genetically diverse populations, such as for modeling complex disease in humans. Drawing from the presence of unique allelic combinations across the genomes, the CC can often provide better models of human disease, usually not possible in traditional inbred mouse strains (Rogala et al., 2014; Gralinski et al., 2015). The CC are also valuable for joint analyses with its outbred sister population, the Diversity Outbred (DO) stock (Churchill et al., 2012; Chick et al., 2016). Finally, the genetic diversity can be interrogated through quantitative trait loci (QTL) mapping (Aylor et al., 2011; Kelada, 2016; Donoghue et al., 2017; Maurizio et al., 2018), including the ability to map phenotypes that can only be measured from counter-factual observations, such as drug response, due to its genetic reproducibility (Mosedale et al., 2017).

At the time of this writing, 72 CC strains were available, falling well below the initial stated goal of 1000 RI strains. The reduced number of strains is due to extinctions during the inbreeding

[1] This chapter represents a mature draft of a manuscript currently in preparation, with slight modifications made for the format. Current author line and title are: Keele GR*, Crouse WL*, Kelada SNP, Valdar W. SPARCC: An R package for calculating power through simulation of QTL mapping experiments in the realized Collaborative Cross. Co-first authors*.

phase, likely because of allelic incompatibilities across subspecies (Shorter et al., 2017). Although these extinctions have provided an unexpected source of insight into the genetics of fertility-related traits, it is unclear to what extent the reduction in CC strains reduces the power to map QTL. Initial power estimates were based on many simulated CC genomes (Valdar et al., 2006b); however, these calculations do not reflect the number of available strains or the actual founder mosaics realized in the genomes of the currently available strains.

Power dynamics have been investigated in genome-wide association studies (GWAS) in humans (Purcell et al., 2003; Klein, 2007), though they do not assess important experimental design considerations for reproducible experimental populations like the CC, for which the same genomes are used across many studies, and possibly include replicate observations. (Kaeppler, 1997) performed power calculations in RI strains analytically, though these estimates will not reflect the specific genomes of the CC, nor the specific statistical procedures used to map in the CC. (Falke and Frisch, 2011) and (Takuno et al., 2012) do use simulations to estimate QTL mapping in RI panels and near-isogenic lines (NIL), but their focus is more within the context of plant RI panels, resulting in simulations that reflect those model systems more than those of animal models. This supports the need to explore QTL mapping power in the realized CC.

Our R package, Simulated Power Analysis of the Realized Collaborative Cross (SPARCC), is a tool that evaluates power to map QTL by performing efficient regression-based association analysis of simulated QTL using the currently-available CC genomes. SPARCC is highly flexible, allowing researchers to tailor their calculations based on the CC strains available to them and the genetic architecture of their phenotypes.

3.2 Methods

3.2.1 Data simulation

SPARCC allows the user to simulate CC phenotypes that reflect a range of underlying genetic architectures. These are controlled by various input parameters to the `sim.CC.data()` function which we will describe in detail. The data-generating model is:

$$\mathbf{y} = \mathbf{1}\mu + \underbrace{\mathbf{ZX}\boldsymbol{\beta}}_{\text{QTL effect}} + \underbrace{\mathbf{Z}\boldsymbol{\delta}}_{\text{Strain effect}} + \underbrace{\boldsymbol{\varepsilon}}_{\text{Noise effect}}, \tag{3.1}$$

where $\mathbf{y}$ is the phenotype, μ is the intercept, $\mathbf{Z}$ is the strain design matrix, $\mathbf{X}$ is the QTL allele dosage matrix, $\boldsymbol{\beta}$ is the QTL effect vector, $\boldsymbol{\delta}$ is the background strain effect, and $\boldsymbol{\varepsilon}$ is an unstructured, random error term. We will describe each component of Eq 3.1 in greater detail, with additional options for `sim.CC.data()` described in **Simulation Documentation**.

3.2.1.1 QTL effect

The QTL effect represents the component of the phenotype that is determined by the founder haplotype states at a locus. By default, this locus is sampled from the genome, but it can also be user-specified. The QTL effect in Eq 3.1 is simplified relative to the actual simulation procedure, which specifies a number of functional alleles and the assignment of founders haplotypes to these alleles (Yalcin et al., 2005) through

$$\mathbf{X} = \mathbf{DAM}, \tag{3.2}$$

where $\mathbf{D}$ is the matrix of haplotype pairs, also referred to as diplotypes, at the QTL, $\mathbf{A}$ is an additive model matrix that maps diplotypes to haplotype dosages, and $\mathbf{M}$ is a founder-to-allele mapping matrix. Though the haplotype is described in terms of eight "alleles" corresponding to the founder haplotypes, it may be expected that the QTL effect results from fewer functionally distinct alleles than the number of founders, for example, via an underlying bi-allelic single nucleotide polymorphism (SNP). Thus, the matrix $\mathbf{M}$ encodes a mapping between the founder haplotypes and a specified number of functional alleles, termed the allelic series. The allelic series can be randomly sampled within SPARCC or specified by the user. Assumptions about the allelic series may have a substantial influence on power, as some allelic configurations may be highly imbalanced or poorly-represented by the founder haplotypes at the QTL locus.

The allelic series, the functional allele effect vector $\boldsymbol{\beta}$, and the particular population will together determine the proportion of the phenotypic variance that the QTL controls, which we refer to as the QTL effect size. It follows that CC data can be simulated with respect to the QTL effect size in differing populations, for example the mapping population itself, with resulting powers highly specific to a given sample of CC strains. Alternatively, the CC data could be simulated with respect to a theoretical, more natural population that possesses the genetic variants present in the founder

strains, which would represent the power to map variants segregating in the initial founders. We will present results from both approaches to simulating data based on QTL effect size.

The function for simulating CC data is `sim.CC.data()`, which has a number of arguments that control the various components of this effect, which are described in detail in **Simulation Documentation**. The primary components of interest that can be controlled have been described, such as the QTL effect size, the allelic series ($\mathbf{M}$), the set of CC strains, and the locus from which to simulate. The current version of SPARCC assumes a single QTL, though the procedure could be generalized to multiple QTL with few modifications.

3.2.1.2 Strain effect

The background strain effect represents the aggregate genetic effect present in each strain, not including the simulated QTL. Many complex traits, such as height in humans (Wood et al., 2014), have highly polygenic and complex genetic architectures (Phillippi et al., 2014). It is possible, even expected, that any given QTL will have an individually small effect, despite the phenotype being highly heritable.

The phenotypic variability that results from strain background presents as additional noise with respect to identifying QTL. This is particularly clear in the the situation in which only a single observation of each strain is observed, at which point, $\mathbf{Z} \rightarrow \mathbf{I}$ in Eq 3.1, and $\boldsymbol{\delta}$ and $\boldsymbol{\varepsilon}$ are indistinguishable from each other. Replicate observations, an important feature of the CC, allow for the unstructured, individual error ($\boldsymbol{\varepsilon}$) to be reduced, potentially improving the power to detect QTL. However, replicate observations will not reduce variability due to the background strain effect, and thus will not improve QTL mapping power in phenotypes with a large background strain effect. The options for `sim.CC.data()` that control the strain.effect are described in **Simulation Documentation**.

3.2.1.3 Noise effect

The noise effect, or variation due to random error, will automatically be calculated as

$$1 - \text{QTL effect} - \text{Strain effect},$$

which is used as the value of σ^2 for sampling $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}\sigma^2)$. The error variance must be greater than zero.

3.2.1.4 Robust power estimation

Our intention is that SPARCC will serve as a flexible tool for calculating power in many different experimental and genetic contexts. The output of the CC data simulation is a matrix of outcomes, with each column $\mathbf{y}^{(s)}$, the s^{th} simulation from equation 3.1. By default, we vary the set of CC strains, loci, and allelic series, producing power estimates that take into account many sources of uncertainty, and are thus broadly interpretable. Alternatively, an investigator may be interested in a more focused power calculation, for instance, on a set of CC strains that have already been chosen. Similarly, it could be used to estimate power for specific loci and allelic series configurations.

3.2.2 Mapping procedure

QTL mapping or genome-wide association involves testing the association of a phenotype with the genetic information at positions across the genome, often through a linear model. In mapping populations with well-characterized haplotypes, which can be probabilistically inferred (Lander and Green, 1987; Mott et al., 2000; Liu et al., 2010; Fu et al., 2012; Gatti et al., 2014; Zheng et al., 2015), rather than association on typed variants, such as SNPs, the association between phenotype and haplotype is possible, a procedure called interval mapping (Lander and Botstein, 1989), which formally takes into account the uncertainty in haplotype, requiring a computationally costly expectation-maximization (EM) procedure (Dempster et al., 1977). An approximation to interval mapping was proposed by (Haley and Knott, 1992; Martínez and Curnow, 1992) that uses standard linear regression (HK regression), which is computationally efficient, generally stable, and accurate when the information content on diplotype probabilities is high.

3.2.2.1 Regression model

The DEFAULT for SPARCC is to use HK regression on strain means across replicates (Zou et al., 2006). Thus, we fit each simulated phenotype $\mathbf{y}^{(s)}$ using the following model:

$$\bar{\mathbf{y}}^{(s)} = \mathbf{1}\mu + \mathbf{PA}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{3.3}$$

where $\bar{\mathbf{y}}^{(s)}$ is the vector of strain means, and $\boldsymbol{\epsilon}$ is the residual on the means, which is expected to be distributed following $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}(\tau^2 + \frac{\sigma^2}{r}))$. The allelic series, encoded in $\mathbf{M}$, is fixed to the eight allele model (Gatti et al., 2014). We note that this could lead to lower power when there are fewer functional alleles, particularly at loci in which the functional alleles are not well represented.

We compare the model fit of Eq 3.3 to the fit of a null model with just the intercept and obtain a p-value based on an F-test from the residual sums of squares (RSS) for the two models. Alternatively, the likelihood ratio test could be used, although the F-test is preferable given the relatively low number of CC strains. This procedure is performed for all loci, resulting in a genome scan for $\mathbf{y}^{(s)}$. Power to map the simulated QTL is the proportion of $\mathbf{y}^{(s)}$ for which the QTL was detected at a genome-wide significance threshold.

3.2.2.2 Significance thresholds and power

The CC panel is a balanced population with respect to founder genomic contributions, with limited levels of population structure, which supports the assumption of exchangeability. As such, we use permutations of $\mathbf{y}^{(s)}$ to assess genome-wide significance based on controlling the family-wide error rate (FWER) (Doerge and Churchill, 1996). Briefly, we sample p permutations, which can be represented as $\mathbf{U}_p\mathbf{y}^{(s)}$, where $\mathbf{U}$ is a permutation matrix that re-orders $\mathbf{y}^{(s)}$ accordingly. We select the maximum $-\log_{10}$ p-value ($\mathrm{logP}_p^{(s)}$) from each p genome scan of simulation s, which are used to fit an extreme value distribution (EVD) (Dudbridge and Koeleman, 2004; Valdar et al., 2006a), which represents a FWER-controlled null distribution from which to draw a significance threshold $T_\alpha^{(s)}$ for genome-wide FWER false positive probability α. The QTL is mapped if the $\mathrm{logP}^{(s)} \geq T_\alpha^{(s)}$. Power, the probability that the simulated QTL is mapped follows:

$$\text{Power} = \frac{\sum_{s=1}^{S} 1_{\mathrm{logP}^{(s)} \geq T_\alpha^{(s)}}}{S} \tag{3.4}$$

where 1_A is the indicator function on whether $A : \mathrm{logP}^{(s)} \geq T_\alpha^{(s)}$ is satisfied. The power estimate in Eq 3.4 is a point estimate specific to the simulation procedure. Investigators should consider ranges of simulation settings, in particular varying the QTL effect size, number of strains, and number of replicates. Running SPARCC for multiple settings can then be used to produce useful power curves or tables.

3.2.2.3 QR decomposition for fast regression

Because the genome-wide scans SPARCC performs on simulated data from the realized CC genomes require permutations for determination of statistical significance, 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 3.3, with $m = 8$ for SPARCC. 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 RSS, and ultimately logP, can be rapidly calculated. Because the SPARCC uses a simulation approach that involves regressing many permuted outcomes ($\mathbf{U}_p\mathbf{y}^{(s)}$) 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 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.

3.2.2.4 Performing genome scans

The SPARCC function for running genome scans from the simulated data is `run.sim.scans()`, The primary argument is `sim.data`, which expects simulated data output from `sim.CC.data()`. There are additional arguments to restrict the scans to a subset of the chromosomes, to a subset of the simulated phenotypes, or to a subset of loci. Finally, the user can provide the precomputed QR decompositions and specify whether the output should return those decompositions, which can be expensive in terms of memory.

3.2.3 Availability of data and software

3.2.3.1 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`. SPARCC also depends upon QTL mapping functionality present in the R package miqtl, which is also available on GitHub at `https://github.com/gkeele/miqtl`.

3.2.3.2 CC haplotype pair probabilities

Founder haplotype probabilities for each CC strain are available on the CC resource website (`http://csbio.unc.edu/CCstatus/index.py?run=FounderProbs`). The diplotype probabilities were constructed using a hidden Markov model (HMM) for haplotype inference as previously described (Liu et al., 2010). The probabilities are based on genotype calls from the MegaMUGA SNP array that contains 77,800 genotype markers. We use probabilities corresponding to build 37 of the mouse genome, though build 38 is also available at the previously mentioned website.

3.2.3.3 Haplotype data reduction

We reduce the size of the CC haplotype probability data by averaging adjacent intervals that are similar in probabilities, in order to reduce the computational expense of scans. Adjacent sites were averaged if the maximum L2 norm between the probability vectors of all individual is less than 10% of the maximum possible L2 norm ($\sqrt{2}$), ultimately reducing the cumulative storage from 610 MB to 288 MB. We store these data in a directory with a structure with which SPARCC is designed to interact. These data are available on GitHub at `https://github.com/gkeele/sparcc_cache`.

3.3 Results and Discussion

3.3.1 Simple SPARCC example

We provide a simple demonstration of simulating a data set, performing genome scans, determining thresholds of significance, and ultimately QTL mapping power. This example is computationally efficient because CC strains are not varied across simulations, though the locus is. We also provide run-time estimates for the main steps.

```
################################################
### Useful functions for parsing haplotype data
> library(miqtl)
> h <- DiploprobReader$new("./sparcc_cache/")
> set.seed(10)

### Grabbing random sample of 65 CC strains
> these.cc.lines <- sample(h$getSubjects(), size=65)

> library(sparcc)
### Simulate 5 data sets:
#### Specified 65 CC strains
#### 5 replicate observations of each
#### 2 functional alleles, allelic series not specified
```

```
#### QTL effect size of 30%
#### Background Strain effect of 10%
> simple.data <- sim.CC.data(genomecache="./sparcc_cache/",
                             CC.lines=these.cc.lines,
                             num.replicates=5,
                             num.sim=5,
                             num.alleles=2,
                             qtl.effect.size=0.3,
                             strain.effect.size=0.1)

### Genome scans
> simple.scans <- run.sim.scans(sim.data=simple.data,
                                return.all.sim.qr=TRUE)

### Generating permutation index
> perm.index <- generate.perm.matrix(num.lines=65,
                                     num.perm=100)

### Permutation scans
> thresh.scans <- run.perm.scans(perm.matrix=perm.index,
                                 sim.CC.object=simple.data,
                                 sim.CC.scans=simple.scans)

### Calculating significance thresholds
> all.thresh <- get.thresholds(thresh.scans=thresh.scans)

### Power estimate
> pull.power(sim.scans=simple.scans,
 thresh=all.thresh)
[1] 0.8
```

```
### Plot a genome scan of a single simulated phenotype
> single.sim.plot(simple.scans,
                  thresh=all.thresh,
                  phenotype.index=1)
##############################################
```

Plots of the simulated CC genome scans produced by the above code are in **Figure 1.1**.

3.3.1.1 Run-time performance

The simple example was run locally on an Early 2015 MacBook Pro with a 2.9 GHz Intel Core i5 processor and 8 GB of RAM. The data simulation and genome-wide scans for five phenotypes took 32.3781 seconds. Computational time increases linearly with number of phenotypes simulated. Computational times will also decrease for lower numbers of CC strains. 100 permutations for 5 simulated phenotypes took 9.315485 minutes. Although the time expense for SPARCC is not trivial, the overall process is highly optimized; this simple example involves fitting 5 phenotypes $\times$ 17900 loci $\times$ 100 permutation alternative models. The process can be sped up using a parallel computing environment, as we do with the following large scale analysis. Highly specific power calculations for an experiment are feasible on a local computer using a single core.

3.3.2 Large scale power dynamics

We have run SPARCC with different combinations of various parameters in order to provide a resource for QTL mapping power in the CC that can be broadly referenced. The specific parameter settings follow:

- Number of strains: [{10-70 by 5}, 72]
- Number of replicates: [1-10, 20, 50]
- QTL effect size (%): [0.5, 1, 5, {10-50 by 10}]
- Number of functional alleles: [2, 3, 8]

- Background strain effect size: [0]

CC lines and the position of the QTL were sampled for each simulation, providing estimates of power that are effectively averaged over the CC population.

3.3.2.1 Computing environment

We performed 1000 simulations (in batches of 100) for each combination of the parameters, resulting in 40,320 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 7 hours to complete.

3.3.2.2 Experiment size and power

We used the results of these simulations to produce power curves that illustrate the relationship between power and the number of strains (**Figure 1.2**) or number of replicate observations (**Figure 1.3**), for a variety of QTL effect sizes, holding other variables fixed. These power curves provide several insights regarding the power to detect QTL in the CC. In general, we find that studies with small-to-moderate sample sizes are well-powered to detect large effect QTL, but that detecting smaller effect QTL could require many replicates. Detecting QTL with effect sizes $\leq 5\%$ is challenging in the CC, reaching 80% power to detect an effect size of 5% when all 72 CC strains are used with greater than 15 replicate observations (**Figure 1.3 [bottomright]**). Detecting 1% or 0.5% QTL would require even higher numbers of strains and replicates. For certain patterns of functional alleles, these curves suggest that mapping QTL with effect sizes $\geq 5\%$ are attainable through the use of more CC strains or more replicate observations.

We also investigated the relationship between power and the total number of mice, particularly focusing on whether additional CC strains or additional replicate observations are more valuable in terms of QTL mapping power. To do this, we calculated the number of mice used in each simulated experiment and interpolated the power at regular grid of values for number of replicates and number of mice. SPARCC generally finds that additional CC strains improve mapping power more than

replicate observations, indicated by higher power values for lower numbers of replicate observations while holding number of mice constant in **Figure 1.4**.

3.3.2.3 Allelic series and power

We emphasize that the overall power depends on the assumed number of functional alleles underlying the QTL. The reasonableness of an assumed number of alleles for a simulation depends on the phenotype. For instance, if the expected causal variant is a single SNP, biallelic QTL are most appropriate, and multiallelic QTL simulations could be overly optimistic. However, a multiallelic QTL can result from local epistatic interactions in the region, which may be more likely with phenotypes closer to the genome, such as gene expression, than physiological phenotypes.

3.3.2.4 Statistical procedure assumes eight alleles

Several factors contribute to dependency of power on the number of functional alleles. One component to the reduction in power for QTL with fewer than eight alleles is that the fit alternative model assumes that each founder strain is an allele. For QTL with fewer than eight alleles, some degrees of freedom are being wasted on estimating redundant allele parameters. Power would likely improve for bi-allelic QTL were simpler models used, such as bi-allelic genotypes (Yalcin et al., 2005). The development of alternative mapping approaches that specifically account for the allelic series remains to be adequately addressed, though such approaches will not be trivial and amenable to power calculations. Still, it stands to reason that if the QTL has less than eight functional alleles, a corresponding allelic genome scan would be more powerful than the eight allele model used in SPARCC.

3.3.2.5 Observed functional allele frequency imbalance

Also contributing to reduction in power for QTL with fewer functional alleles than the statistical procedure is the observed allele frequency balance in the data set. While the CC is generally balanced with respect to inheritance from all eight founders across the genome, certain allelic series will result in data that are potentially highly imbalanced in terms of the observed functional alleles. For example, a functional bi-allelic SNP with one allele present in only one of the founder haplotypes will have a minor allele frequency of 12.5% at a locus that is perfectly balanced in CC. This reduces

the variance explained by the QTL effect in the population, and correspondingly, the power to detect that effect.

Taking the allele frequency balance issue to the extreme, though the CC has good average balance with respect to the founder haplotypes, at any given specific loci, one or more alleles may be lost, and thus their functional alleles unobservable. By posing the problem of power estimation in the context of the CC founders and the realized CC strains, the power estimates from SPARCC can reflect the reduction in power to map QTL at loci where potential functional alleles have been lost, which we view as a strength of our approach. SPARCC can produce more optimistic bi-allelic power calculations by fixing the allelic series to be balanced (example: `M.ID="0,0,0,0,1,1,1,1"`), but in reality, such power calculations are themselves overly-optimistic in assuming that bi-allelic QTL will be balanced across the founder haplotypes. **Figure 1.5** illustrates the effect that imbalance of the allelic series can have on the power to map QTL in the CC.

3.3.3 CC as a mapping population

SPARCC demonstrates that the CC can be used to effectively map QTL. Though the power calculations in the realized CC are not as optimistic as the simulated expectations of 1000 lines (Valdar et al., 2006a), successful mapping experiments can still be designed, particularly harnessing the ability to have replicate observations. It also bears emphasizing that, aside from mapping, the CC is a powerful tool for new disease models (Rogala et al., 2014; Gralinski et al., 2015) and as a means of validating results from the DO (Chick et al., 2016).

3.3.4 Limitations

Any analysis of power is subject to the assumptions underlying that analysis. One of the advantages of SPARCC is that its flexibility allows the impact of many of these assumption to be explored. For example, assumptions about how well the strain effect is modeled or the number of independent QTL signals may provide valuable insight into how genetic architecture determines power in the CC. In addition, SPARCC could be used to investigate many related questions, including the power for specific combinations of CC strains or experimental designs, exploring genome-wide false positive rates, or assessing how the power to detect QTL varies depending on genomic position.

In terms of future work, the simulation procedure within SPARCC could be expanded to investigate how problems like variance heterogeneity or model mis-specification influence power.

3.3.5 Conclusion

SPARCC is a useful software tool for exploring the power to detect QTL in the CC. This software leverages an efficient model fitting approach in order to explore power in a level of detail that has previously been impractical. This simulation-based approach improves on previous attempts to characterize power in the CC by using the realized CC genomes currently available. We intend that SPARCC will be a useful and flexible tool for researchers designing CC experiments.

3.4 Simulation Documentation: Detailed description of `sim.CC.data()` options

3.4.1 QTL effect

- `qtl.effect.size`
  - $0 \leq$ `qtl.effect.size` $< 1 -$ `strain.effect.size`
  - This argument represents ϕ^2, such that $\boldsymbol{\beta} \sim \text{N}(\mathbf{0}, \mathbf{I}\phi^2)$.
  - A specific $\boldsymbol{\beta}$ can specified with the `beta` argument, though it will be scaled to match `qtl.effect.size`. If `beta=NULL`, then $\boldsymbol{\beta}$ is sampled accordingly.
- `num.alleles` (DEFAULT = 8)
  - $2 \leq$ `num.alleles` ≤ 8
- `M.ID`
  - Rather than specifying `num.alleles` and then sampling $\mathbf{M}$, these can be fixed with the `M.ID` argument.
  - Expects strings of the form `"A,B,C,D,E,F,G,H"`, with each letter corresponding to a founder strain, taking an integer value 0-7, representing functional alleles.
  - Example: `M.ID="0,0,0,0,1,1,1,1"` represents a biallelic causal variant, in which the first four strains have one allele, and the last four having the other.

- `CC.lines` (DEFAULT = NULL)
  - This argument allows the user to provide a vector of CC line IDs on which to base the power calculation. The CC genomes, along with `locus`, will determine $\mathbf{D}$ in Eq 3.2.
  - If `CC.lines = NULL`, then SPARCC will sample `num.lines` from all available lines.
    * `vary.lines` (DEFAULT = TRUE)
      · If `vary.lines = TRUE`, the set of lines for each simulation will be sampled and vary.
      · If `vary.lines = FALSE`, the set of lines will be sampled once, and used for each simulated outcome.
- `locus` (DEFAULT = NULL)
  - This argument allows the user to specify a specific locus for the simulated QTL, in effect determining the haplotype dosage matrix $\mathbf{D}$.
  - If the argument is left empty, SPARCC will sample loci uniformly from the CC genomes, thus providing power estimates averaged over genomic positions.
- `impute` (DEFAULT = TRUE)
  - If `impute=TRUE`, then $\mathbf{D}$ in Eq 3.2 is sampled from the probabilistically reconstructed diplotypes at the QTL

$$\mathbf{D}_i \sim \text{Cat}(\mathbf{P}_i) \tag{3.5}$$

    where Cat$(.)$ is a categorical distribution and $\mathbf{P}$ is a matrix of diplotype probabilities for the CC genomes at the QTL.
  - If `impute=FALSE`, then $\mathbf{D} = \mathbf{P}$ in the simulation procedure.
- `scale.qtl.mode` (DEFAULT = "B")
  - If `scale.qtl.mode="B"`, var$(2\boldsymbol{\beta})$ is scaled to `qtl.effect.size`, setting the QTL effect size with respect to a theoretical population that is balanced with respect to functional alleles, from which the CC mapping population developed.

- If `scale.qtl.mode="MB"`, var($2\mathbf{M}\boldsymbol{\beta}$) is scaled to `qtl.effect.size`, setting the QTL effect size to a theoretical natural-like population with a specific allelic series.
- If `scale.qtl.mode="DAMB"`, var($\mathbf{DAM}\boldsymbol{\beta}$) is scaled to `qtl.effect.size`, setting the QTL effect size to a specific set of CC strains and allele series.
- If `scale.qtl.mode="ZDAMB"`, var($\mathbf{ZDAM}\boldsymbol{\beta}$) is scaled to `qtl.effect.size`, setting the QTL effect size with respect to the specific set of CC strain, allelic series, and number of replicate observations.
- If `scale.qtl.mode="none"`, $\boldsymbol{\beta}$ is not scaled, allowing the user to specify an effect vector without it being scaled.

3.4.2 Strain effect

- `strain.effect.size`
  - $0 \leq$ `strain.effect.size` $\leq 1 -$ `qtl.effect.size`
  - This argument specifies τ^2, such that $\boldsymbol{\delta} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}\tau^2)$.

The actual sampled strain effect are scaled in the same manner as the QTL effect, which is specified with `scale.by.var`.

3.4.3 Additional options

- `num.sim`
  - This argument specifies SPARCC to simulate s samples of $\mathbf{y}$ from Eq 3.1.
- `num.replicates`
  - This argument allows the user to set the number r of replicate observations of each CC line. The reproducibility of CC genomes is an important feature, allowing noise variation to be reduced.
  - SPARCC currently requires all CC lines to have the same number of replicates.

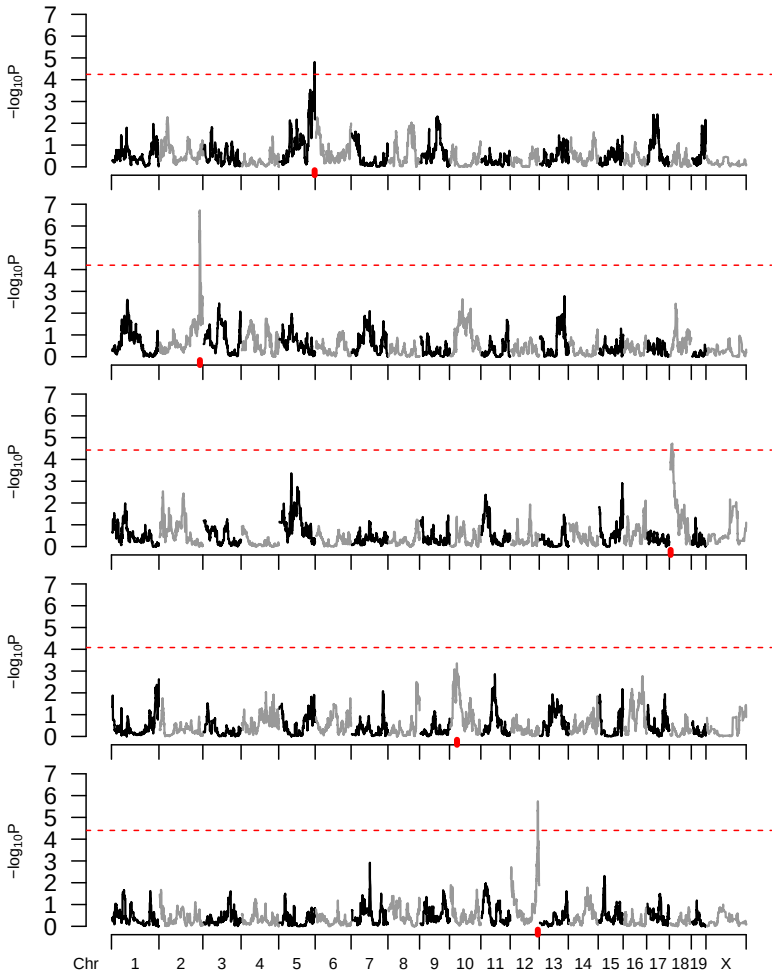

Figure 1.1: Five simulated genome scans generated by the code provided in the simple SPARCC example. Red dashed lines represent 95% significance thresholds based on 100 permutation scans. The red tick represents the simulated QTL position. These simulations were based on a specified set of 65 CC strains, five replicate observations of each strain, two functional alleles, 30% QTL effect, and 10% strain effect. The QTL is not mapped in the fourth simulation, ranked top to bottom. Actual power calculations should be based upon a greater number of simulations.

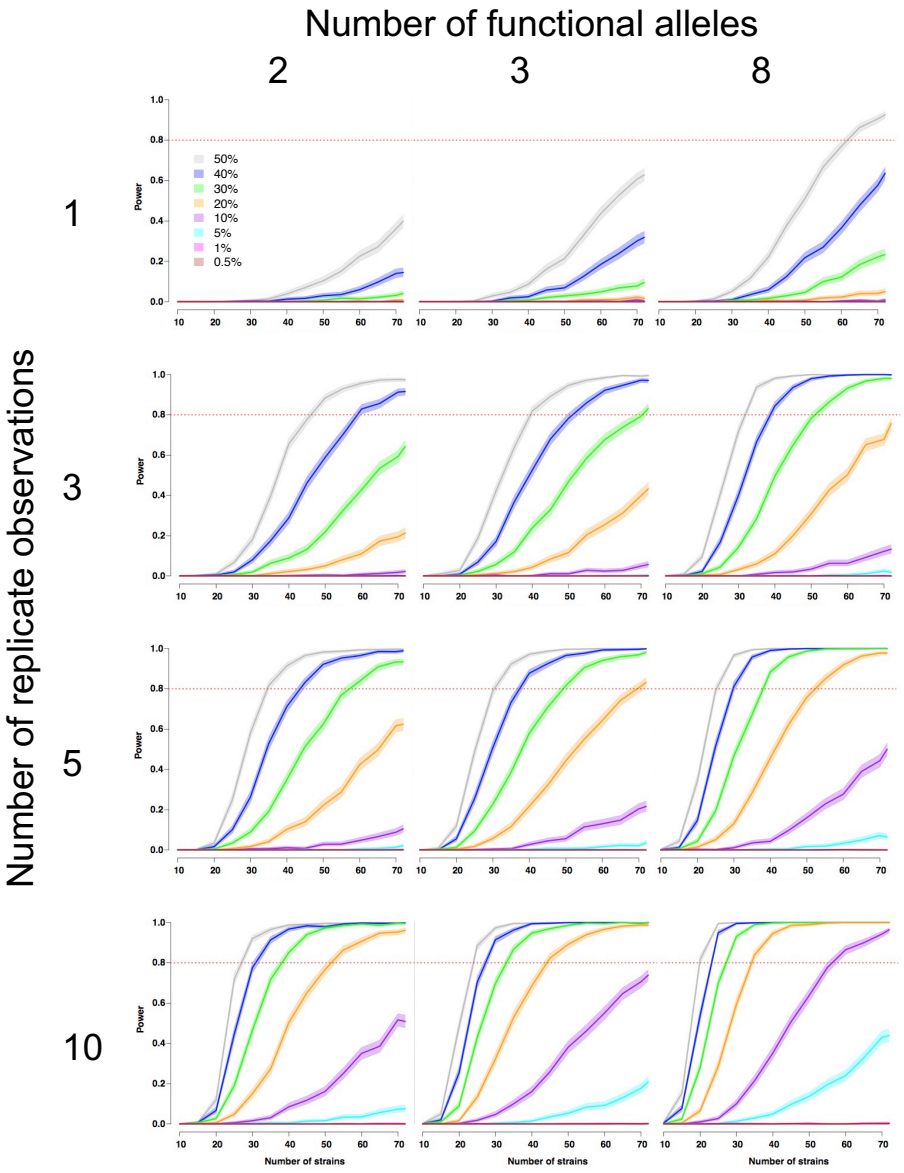

Figure 1.2: Power curves based on a thousand simulations per setting with respect to number of CC strains, stratified by number of replicates and the number of functional alleles. The red dashed line emphasizes 80% power. CC strains and loci were varied in simulations, resulting in powers that average over loci and strain combinations. Confidence intervals were calculated based on Jeffreys interval for a binomial proportion. The columns, left to right, correspond to two functional alleles, three functional alleles, and eight functional alleles. The alternative model fit at each locus is an eight allele model, parameterized with respect to the eight inbred founders. The rows, top to bottom, correspond to a one, three, five, and ten observations of each CC strain. Better power tracks with increased numbers of strains, numbers of replicate observations, and functional alleles. **Figure 1.3** has power curves with respect to number of replicate observations rather than number of CC strains. The allelic series for the two and three allele simulations were sampled uniformly, meaning any distribution of functional alleles to founders was given equal probability weight.

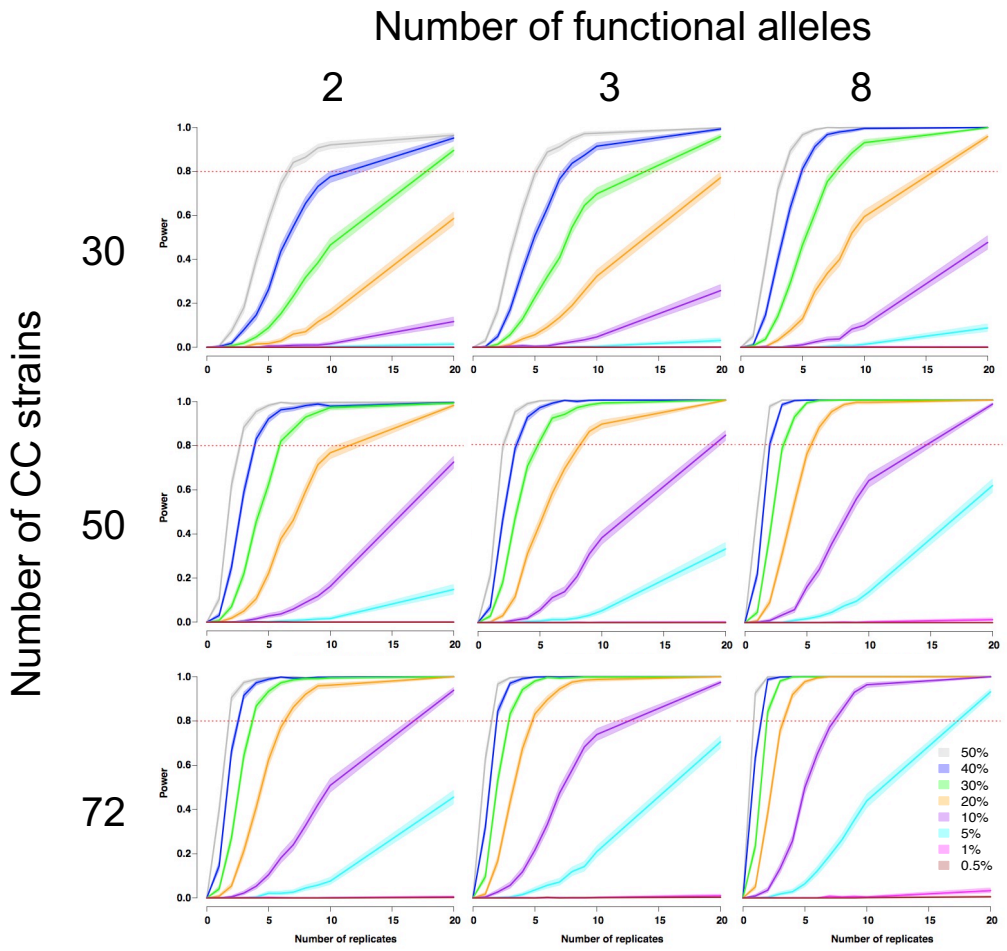

Figure 1.3: Power curves based on 1000 simulations per setting with respect to number of replicates per CC strain, stratified by number of CC strains and the number of functional alleles. The red dashed line emphasizes 80% power. CC strains and loci were varied in simulations, resulting in powers that average over loci and strain combinations. Confidence intervals were calculated based on Jeffreys interval for a binomial proportion. The columns, left to right, correspond to two functional alleles, three functional alleles, and eight functional alleles. The alternative model fit at each locus is an eight allele model, parameterized with respect to the eight inbred founders. The rows, top to bottom, are 30, 50, and 72 CC strains. As seen in **Figure 1.2**, better power tracks with increased numbers of strains, numbers of replicate observations, and functional alleles.

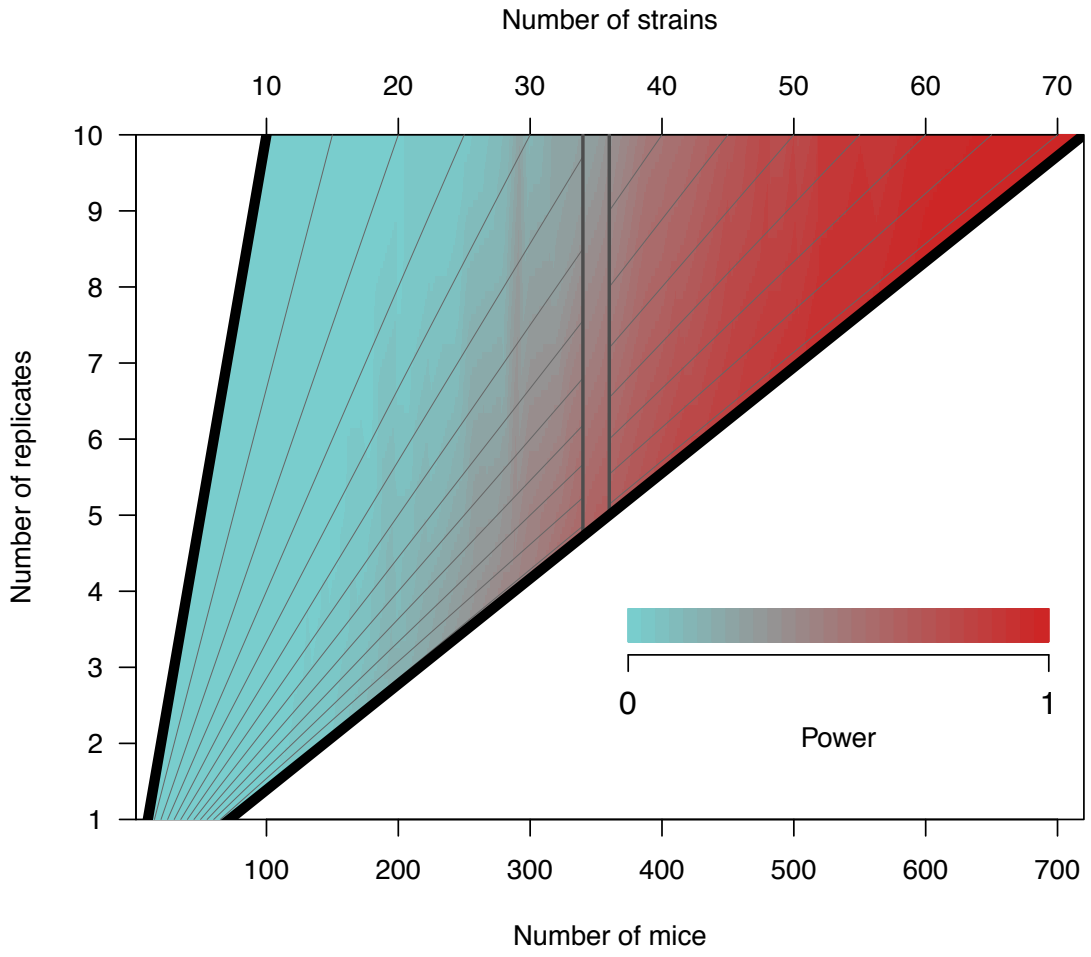

Figure 1.4: A heatmap of QTL mapping power for number of replicate observations by total number of mice in the experiment. This figure assumes a QTL effect size of 20%, no background strain effect, and two functional alleles, though varying these parameters should not change the inference. Power was interpolated at regular intervals across a grid of values for number of replicates and number of mice to facilitate plotting, approximating power for numbers of mice that were not directly assessed. Note that some combinations of number of replicate observations and total number of mice are not defined because the CC is limited to 72 strains and we only considered equal numbers of replicates for all strains. 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, the power reduces as the percentage of the sample that are replicates increases, suggesting that observations of new genomes are more important to QTL mapping power than replicate observations. This is illustrated with a cutout band centered on 350 mice. Power is lower at the top of the band where replicate mice are a relatively higher proportion of the total number of mice. Thus, prioritizing experiments with higher numbers of CC strains rather than higher numbers of replicates is ideal. Increasing the number of replicate observations does benefit QTL mapping power, but not as effectively as additional strains.

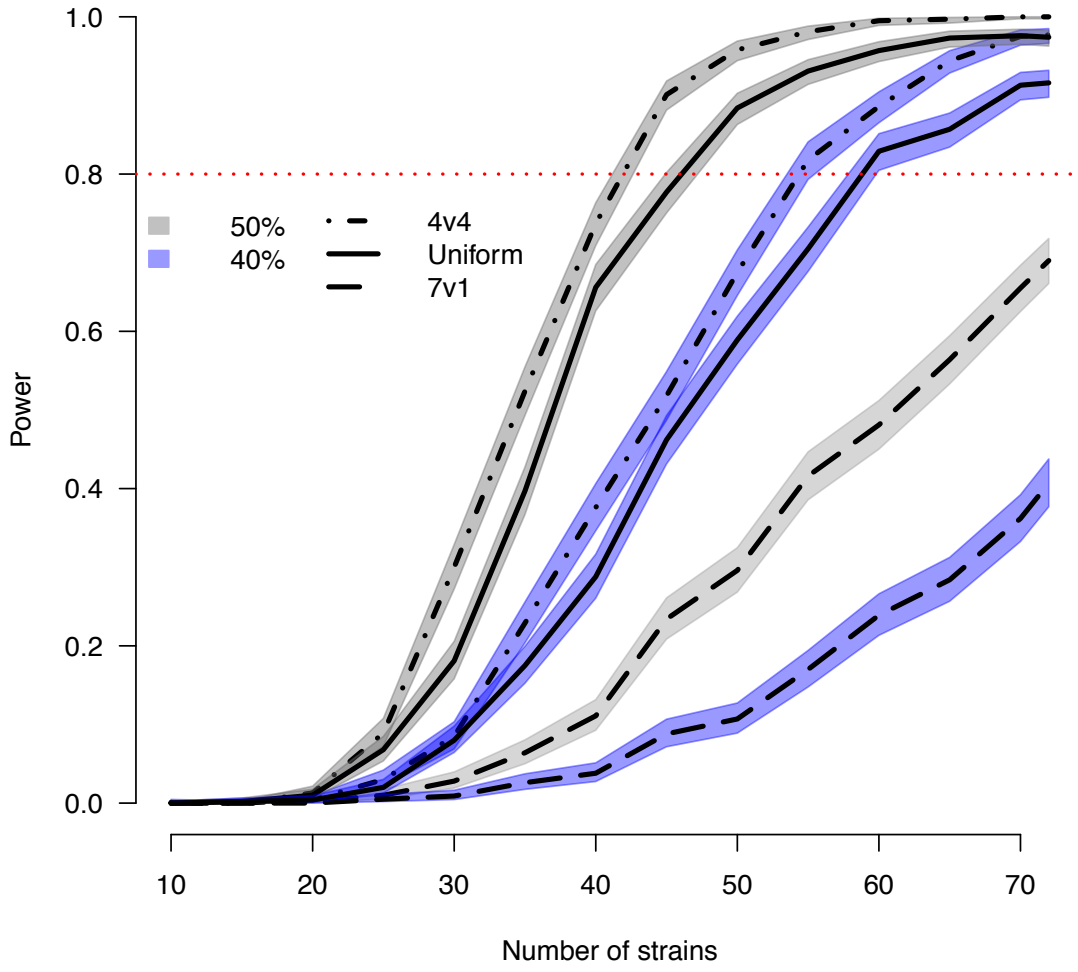

Figure 1.5: Power curves comparing two QTL effect sizes and four different settings for an allelic series with two functional alleles. These simulations are based on three replicate observations per genome. A balanced representation of the functional alleles, with each allele corresponding to four of the founders (4v4), produces the best power. This is followed closely by uniform sampling of allelic series, in which any bi-allelic allelic series is equally likely (Uniform; the default for SPARCC). Finally, fixing the allelic series at a highly imbalanced setting, one functional allele corresponding to only a single founder (7v1), results in greatly reduced power.

CHAPTER 4

Accounting for haplotype uncertainty in QTL mapping of multiparental populations using multiple imputation [1]

4.1 Introduction

Genetic association studies have been extraordinarily successful at identifying genes and regions of the genome that are important to the underlying biological mechanisms modulating complex phenotypes that are highly relevant to medicine and agriculture. Within the context of human studies, genome-wide association studies (GWAS) have been prolific in their ability to identify common variants as candidates for further study (Lee et al., 2016). However, such studies are constrained by their observational nature, complex population structure, and potentially unobserved confounding factors. These challenges, along with constraints in the phenotypes that can be reasonably measured in humans, provide support for controlled experiments in model organism systems, both as models of complex human phenotypes and diseases, as well as agriculturally relevant traits.

Many traditional experimental designs for model organisms result in individuals descended from two founders, or bi-parental populations. These populations have been highly fruitful for QTL mapping, and thus many statistical methodologies have been developed (Broman, 2001). A limitation of these simpler populations is that they do not possess as much genetic variation as is in naturally occurring ones, thus limiting their ability to model humans adequately for certain biological systems. Addressing this limitation of bi-parental populations, multiparental populations (MPP) possess greater phenotypic and genetic diversity through the incorporation of additional founders, while often maintaining reproducibility. MPP populations have been developed in a number of species, such as the Collaborative Cross (CC) (Collaborative Cross Consortium, 2012; Srivastava et al., 2017) and Diversity Outbred (DO) stock (Churchill et al., 2012) in laboratory mouse; the

[1] This chapter represents a mature draft of a manuscript currently in preparation, with slight modifications made for the format. Current author line and title are: Keele, G.R., Valdar, W. Accounting for haplotype uncertainty in QTL mapping of multiparental populations using multiple imputation.

Drosophila synthetic population resource (DSPR) in flies (King et al., 2012a; Long et al., 2014; King and Long, 2017; Najarro et al., 2017; Stanley et al., 2017); round worm (Noble et al., 2017); yeast (Cubillos et al., 2017); multi-parent advanced generation inter-cross lines (MAGIC) in *Arabidopsis* (Kover et al., 2009; Huang et al., 2011) and rice (Bandillo et al., 2013; Raghavan et al., 2017); nested association mapping population (NAM) in maize (Buckler et al., 2009) and sorghum (Bouchet et al., 2017); strawberry (Mangandi et al., 2017); and oil palm (Tisné et al., 2017). The well-characterized founder haplotypes allows for QTL mapping approaches that, rather than modeling phenotype in terms of genotyped variants, models phenotype with haplotype descent. This haplotype approach allows un-genotyped loci to be tested as putative QTL positions. Although haplotype association will not necessarily outperform genotype association, particularly if a genotyped variant, such as a single nucleotide polymorphism (SNP), is causal or strongly tags the causal variant, it will implicitly model all local variants, possibly including local epistatic interactions specific to a haplotype block that would be challenging to model in a principled way through genotypes. Haplotype association does, however, complicate the statistical methodology due to the fact that haplotypes are not directly observed, but rather probabilistically inferred.

This uncertainty surrounding haplotype, or more generally, genetic state, is formally addressed with interval mapping (IM) (Lander and Botstein, 1989). IM models the data as a mixture of normal distributions, a result of the genetic state uncertainty at an interval or position in the genome, and fits maximum likelihood estimates (MLE) of parameters through an Expectation-Maximization (EM) algorithm (Dempster et al., 1977) for Frequentist inference. Genetic state probabilities are estimated for intervals that span the entire genome, either as pseudomarkers (often regularly spaced) or at the genotyped marker positions (Lander and Green, 1987). There are a number of Hidden Markov models (HMM) that can be used to construct the probabilistic reconstructions of genetic state, allowing for the incorporation of multiple sources of information and uncertainty, such as genotyping error (Mott et al., 2000; Liu et al., 2010; Fu et al., 2012; Zheng et al., 2015) or even incorporate information from genotype probe intensities (Gatti et al., 2014). Although IM was initially proposed in bi-parental populations, first in backcrosses (BC) (Lander and Botstein, 1989) and extended for F2 intercrosses and other designs (Dupuis and Siegmund, 1999), it has also been generalized to multi-allelic settings (Liu and Zeng, 2000), such as in MPP. Though IM models the uncertainty in haplotype, in cases of low information distinguishing genetic states, it can become unstable. This issue can be compounded

in MPP where there are more genetic states to distinguish at a locus. It is also possible that the EM procedure will become stuck in local maxima, which is also more challenging in MPP where the likelihood is more complex and multi-dimensional. Finally, the EM is an iterative method, and thus potentially computationally intensive, particularly for dense scans in populations with fine mapping resolution.

The problems of stability and computational efficiency of IM were addressed with an approximate regression approach, sometimes referred to as Haley-Knott regression, though which we will refer to as regression on probabilities (ROP) (Haley and Knott, 1992; Martínez and Curnow, 1992), that involves regressing the phenotype directly on the genetic state probabilities, or some function of the probabilities, such as the additive dosages of an allele. By dosage, we mean a probabilistic generalization of a count of alleles. ROP, also referred to has Haley-Knott regression, was initially proposed for bi-parental populations in which there are only two founder haplotypes, similar to bi-allelic SNPs, and has been extended to multi-allelic populations (Rebai and Goffinet, 1993), and is commonly used (Mott et al., 2000; Valdar et al., 2006b, 2009; Kover et al., 2009; Svenson et al., 2012; Gatti et al., 2014). Although ROP does not directly model the uncertainty present in the genetic state, the expectations of the modeled data are equivalent in certain settings from the mixture of normals model (IM) and the standard ROP regression. Although it has been known that ROP can produce unstable allele effect estimates (Zhang et al., 2014), it has thus far been considered reliable for hypothesis testing.

ROP-like approximations have commonly been used in human GWAS, in which it is common practice to use SNP probabilities or dosages for unobserved variants based on probabilistic reconstructions from densely genotyped reference samples (like HapMap (Gibbs et al., 2003)). A very simplistic approach is to take the most likely genotype and completely ignore the uncertainty present in the genotype, although ROP procedures have been found to outperform such an approach (Li et al., 2009; Aulchenko et al., 2010; Marchini and Howie, 2010; Zheng et al., 2011). Though an improvement over completely ignoring uncertainty in the genotypes, ROP does not directly model it and (Kutalik et al., 2011) notes that this can lead to an increased false positive rate (FPR) in SNP-based GWAS. This increased FPR is the result of small probabilities correlating strongly with the phenotype, which the ROP procedure will not treat as highly unlikely SNP alleles, but rather as a small SNP dosage that strongly predicts phenotype, resulting in an entirely artificial association.

In response to these problems that can occur with the ROP approach, there have been proposed statistical methods, primarily for SNP-based analysis, to directly model the uncertainty (Kutalik et al., 2011; Acar and Sun, 2013), which are similar to IM in model organisms, through the use of an EM procedure.

This problem of observed false positives resulting from low probability alleles is more avoidable in SNP association than haplotype association, in which it is common practice to filter out SNPs with very low minor allele frequencies (MAF), which are considered likely genotyping errors. This filtering step will also most likely remove the markers that are prone to ROP significance inflation. However, in the multi-allelic setting of MPP, depending on the allelic balance of the population, almost all loci may possess alleles with low allele frequencies, and are thus prone to producing an artificial ROP signal. One approach to countering this issue is to fit the haplotype effects as random effects with a single variance component, thus harnessing shrinkage (Verbyla et al., 2014; Wei and Xu, 2016). Though it is computationally more intensive to fit the QTL effect as a random effect, possibly prohibitively slow in certain data sets and particularly in the presence of population substructure, this approach is preferable to fixed effects. However, fitting the QTL effect as a random term through ROP does not directly address the underlying issue of potential correlations between outcome and probabilities or dosages, but does happen to greatly restrict the problem by dynamically shrinking the potential effects. As such, statistical approaches that more directly model the uncertain nature of inferred genetic state are needed.

Bayesian approaches offer alternatives to Frequentist QTL mapping methods, and with advances in computing, are becoming increasingly appealing due to their ability to fit complex, highly parametric models, including stably fitting multi-locus models with shrinkage, handling multiple outcome models, and naturally incorporating additional sources of uncertainty through the hierarchical model. Genomic prediction is a natural application of Bayesian models to genetic data due to its focus on optimally fitting and predicting data, thus harnessing the potential of Bayesian statistics for stable yet highly parametric models, such as potentially including all loci (Meuwissen et al., 2001; Xu, 2003; Yi and Xu, 2008). These ideas can also extend to Bayesian QTL mapping, particularly a fully multilocus approach (Crawford et al., 2017). Here we instead focus on the Bayesian modeling of the genetic state uncertainty jointly with other model parameters in the context of populations descended

from inbred founders, allowing this uncertainty to be directly modeled rather than approximated as in ROP.

A fully Bayesian approach would include genetic state as an unobserved variable in the hierarchical model, allowing the genetic state probabilities to be updated through sampling in response to the other parameters in the model, particularly phenotype and QTL effects. For normal data, often the hierarchical model is specified such that the QTL effect is dependent on the noise variance parameter, resulting in conjugate priors (Servin and Stephens, 2007) and a factorizable posterior, allowing Markov Chain Monte Carlo (MCMC) sampling to be avoided, which can be prohibitively slow and fail to mix with complicated models.

(Sen and Churchill, 2001) propose a comprehensive and generalizable Bayesian hierarchical model for QTL mapping that simultaneously models multiple loci based on a pre-specified grid of pseudomarker locations. A binary vector of QTL status is sampled, and its posterior used for hypothesis testing. To avoid MCMC and still acknowledge that the phenotype can inform the estimate of genetic state, they use an importance sampling scheme with weights calculated based on how well the genetic states at the QTL explain the phenotype, in essence updating the genetic state probabilities, and allowing weighted Monte Carlo (MC) sampling from initial joint multipoint imputations of the genetic states across the pseudomarker grid. Though the model is broadly proposed, it is applied in simpler bi-parental populations assumed to have no population structure. Generalizing the method to MPP is possible, likely requiring the inclusion of a polygenic effect with corresponding variance component as well as imposing shrinkage on QTL effects with variance components. This additional model complexity would likely require MCMC sampling that include computationally expensive matrix operations, and would thus likely be infeasible without further assumptions or approximations.

A Bayesian mapping approach could be simplified to a single locus perspective, and potentially allow for computationally feasible mapping in MPP. (Durrant and Mott, 2010) proposed a single locus Bayesian QTL mapping approach for MPP that has some similarities to (Sen and Churchill, 2001), such using the conjugate prior for QTL effects as dependent on the noise variance. They also make the assumption that genetic state is independent of phenotype and QTL effect, thus not require updating of the genetic state probabilities and allowing MC sampling. This assumption should be conservative, and greatly reduces the computational burden through the avoidance of MCMC. Notably an important contribution is made through the inclusion of a variance component

on the effect of a single locus, imposing shrinkage, and potentially crucial for the more complicated genetic models that result from MPP haplotype alleles. Their method still does not generalize in a computationally feasible way to an MPP with population structure. An additional variance component corresponding to overall relatedness would require either MCMC sampling, which would likely mix poorly and require matrix operations, or a more challenging and extensive MC sampling from the joint non-standard distribution of the variance components.

Features from the previous methods are shared with other Bayesian models of MPP data. (Zhang et al., 2014) proposed the Diploffect model for estimating MPP haplotype allele effects at a locus while taking into account the haplotype uncertainty. Similar to (Durrant and Mott, 2010) its focus is a single locus model for an MPP population; however, it allows for the genetic state parameters to be updated from information in the phenotype and QTL effect through importance sampling, as in (Sen and Churchill, 2001). Briefly, Diploffect is more flexible to modeling of population structure when implemented through integrated nested Laplace approximation, but is not feasible nor intended for a QTL scan across the genome.

The previously described Bayesian methods demonstrate how Bayesian statistics can naturally model and account for many levels of uncertainty. However, they also reflect the computational burden of increasingly complicated genetic models, particularly within a genome-wide context. Though jointly modeling uncertainty on genetic state and parameters would be ideal, incorporating the sampling process on the genetic states with traditional Frequentist likelihood-based inference could account for the genetic state uncertainty and be computationally feasible. In terms of the described Bayesian methods, if the locus effect is fit as a random effect, this would represent a multiple imputation Frequentist version of (Durrant and Mott, 2010), in which the genetic state probabilities are not updated; essentially the prior is being treated like a posterior and averaged over in the multiple imputation process. And though sacrificing the ability to propagate and characterize the parameter uncertainty that is an important feature of Bayesian statistics, we gain computational efficiency to feasibly model population structure.

Although multiple imputation can intuitively be viewed through Bayesian lens as part of the overall sampling process, with each imputation representing a sample from the prior distributions of the missing data, here modeled as parameters in the hierarchical model, inferences can still be drawn using Frequentist methods. This process involves repeating the statistical procedure on each imputed

data set, aggregating results, and drawing inference from the summary, therefore incorporating the uncertainty due to missing information, in this case genetic state uncertainty. There have been numerous proposed methods for aggregating statistics over imputations (Li et al., 1991). Commonly regression coefficients are averaged, and incorporated into a multiple imputation version of a Wald statistic. This approach would not naturally generalize to fitting the QTL effect as random effect. (Meng and Rubin, 1992) propose an aggregate likelihood ratio test, though it would also require a fixed effect QTL model. (Li et al., 1991) also propose aggregating over p-values as an approximate approach for drawing inference across imputations, which would generalize whether the QTL was fit as either a random or fixed effect.

It is important to note that in the context of GWAS, the term imputation is commonly used in reference to estimating variant allele probabilities at unobserved loci. These are related concepts, though we are using its original meaning from the missing data statistics field, whereby an imputation is a sample or realization drawn from a data probability distribution. ”Imputed” SNP variants in GWAS usually represent a ROP-like analysis, as is common in SNP-based GWAS, though multiple imputation analyses has been used with SNP data (Ramstein et al., 2015). Here, we propose a multiple imputation approach to haplotype association, in which genetic state, in this context, diplotype state, is imputed from estimated diplotype probabilities.

A multiple imputation with Frequentist inference approach would control false positives that can occur with ROP while also being computationally convenient. Furthermore, regression-based procedures are easily extended to important modeling considerations such as additional covariates and confounders, population substructure, batch effects, as well as alternative parameterizations of the genetic model (*e.g.* additive model). As such, a Frequentist multiple imputation procedure provides a flexible approach that extends many of the appealing features of ROP, while also accounting for genetic state uncertainty. In addition, such an approach remains computationally feasible for a genome-wide procedure in comparison to fully Bayesian approaches.

4.2 Statistical Models and Methods

We will describe various methods for mapping QTL based on testing how well an individual's genetic state at a locus predicts the phenotype with a linear model. First we will briefly describe our framework.

4.2.1 General Framework

Define y_i to be the observed phenotype value of an individual. The genetic state for an individual at a locus is encoded in $\mathbf{g}_i$, a K-element vector. The number of genetic states K is determined by the number of allele J such that $K = J + \binom{J}{2}$. For bi-allelic variants, such as is common with SNPs, $J = 2$ and $K = 3$. For an eight founder MPP, such as the DO, $J = 8$ and $K = 36$, assuming the population includes individuals that are heterozygous at loci. If a population were completely inbred, $K = J$. The statistical procedures we describe and propose use a single locus approach, with a general model of the form

$$y_i = \text{QTL}_i + \varepsilon_i,$$

in which $\text{QTL}_i = \mathbf{x}_i^{\text{T}}\boldsymbol{\beta}_{\text{QTL}}$ represents the additive linear component of the phenotype y_i attributable to the modeled locus, $\mathbf{x}_i^{\text{T}} = \mathbf{g}_i^{\text{T}}\mathbf{M}$ is the ith row of the QTL design matrix $\mathbf{X}$ representing $\mathbf{g}_i$ rotated according to the model matrix $\mathbf{M}$, and ε_i as the remainder or residual of y_i as an un-modeled error term.

This simple linear model is appealing in its flexibility, such as allowing alternative parameterizations of the QTL term through $\mathbf{M}$, which maps between the K genetic states and some linear function of them, often with the intent to simplify the genetic model. One commonly used $\mathbf{M}$, $\mathbf{M}_{\text{Add}}$ rotates the genetic state probabilities, which include heterozygous pairings of alleles, to the additive allele dosages. With no uncertainty, $\mathbf{M}_{\text{Add}}$ maps genetic states to counts of the alleles; in the case of MPP, diplotypes to counts of founder haplotypes. In addition, the model can be expanded to incorporate covariates as fixed effects, as well as split ε_i into multiple components with corresponding variance components, both structured and unstructured. Mapping consists of, for each locus over the genome, testing whether a model with the genetic state at the given locus predicts the modeled phenotype

better than a model with no genetic information, or equivalently:

$$H_0 : \boldsymbol{\beta}_{\text{QTL}} = \mathbf{0}$$

When $\mathbf{g}_i$ are directly observed with complete certainty for all n individuals (represented as an $n \times K$ matrix of genetic states $\mathbf{G}$), this process becomes straightforward. Though these methods could be extended to generalized linear models for non-normal outcomes, we assume that y is a normally distributed trait:

$$\mathbf{y}|\mathbf{G} \sim \text{N}(\mathbf{X}\boldsymbol{\beta}_{\text{QTL}}, \mathbf{V}(\boldsymbol{\theta}))$$

where $\mathbf{y}$ is the n-element vector of outcomes, $\mathbf{X} = \mathbf{GM}$, and $\mathbf{V}(\boldsymbol{\theta})$ is the variance-covariance matrix that includes a parameter vector specifying the variance parameter(s) of the normal distribution. Fitting complex $\mathbf{V}(\boldsymbol{\theta})$ can become computationally prohibitive. There are established approaches to fitting $\mathbf{V}(\boldsymbol{\theta}) = \mathbf{K}\theta_1 + \mathbf{W}^{-1}\theta_2$, such that $\mathbf{K}$ is a symmetric relationship matrix and $\mathbf{W}$ is a diagonal matrix, often the $\mathbf{I}$ when individuals are weighted equally. Consider the simple situation in which y are independent, then $\theta_1 = 0$, $\theta_2 = \sigma^2$, and $y_i|\mathbf{g}_i \sim \text{N}(\mathbf{x}_i^{\text{T}}\boldsymbol{\beta}_{\text{QTL}}, \sigma^2)$. This is equivalent to standard linear regression in which the phenotype is regressed on a linear function of the genetic state. Maximum likelihood estimators (MLE) of the regression parameters ($\boldsymbol{\beta}_{\text{QTL}}$, σ^2) can be calculated, and used to conduct hypothesis tests. $\boldsymbol{\theta}_1 \neq 0$ is often included to model correlations between y, such as when population structure is present, at which point a linear mixed effect model is being fit.

Modeling the outcome in terms of genetic state becomes more challenging when the genetic state is not directly observed and thus not known with complete certainty. This uncertainty can arise in the context of assayed genotypes, due to genotyping errors or no-calls. Genetic state can also represent haplotype descent or un-assayed variants, which are not directly observed, but rather probabilistically reconstructed based on LD in nearby genotyped variants.

4.2.2 Incorporating uncertainty in genetic state

The resulting uncertainty in $\mathbf{G}$ can be described through a probability distribution function $\Pr(\mathbf{G})$. Formally acknowledging this uncertainty in the association analysis would involve integrating or

averaging the likelihood over $\Pr(\mathbf{G})$:

$$\sum_{\mathbf{G}} \Pr(\mathbf{y}|\mathbf{G})\Pr(\mathbf{G})$$

Accounting for this additional uncertainty lends itself to Bayesian approaches, which allows for hierarchical models to easily be specified. Although additional sources of uncertainty can be intuitively incorporated in Bayesian procedures, the computational cost can be great, and unfeasible particularly when the model becomes complex and includes multiple variance components. An approximate approach that sidesteps the full sampling process of Bayesian statistics is to marginalize out $\mathbf{G}$, producing a marginal distribution/likelihood, which will be of the form of a normal mixture distribution, and then use hypothesis testing procedures.

Hypothesis testing is more challenging due to analytic solutions not existing for the MLE of the mixture distribution likelihoods. Instead they must be estimated using an iterative expectation-maximization algorithm (EM) method, which alternates between updating the parameter MLE conditioned on an estimate of the expected value of $\mathbf{G}$ ($[\widehat{\boldsymbol{\beta}}_{\text{QTL}}, \widehat{\boldsymbol{\theta}}]^{(t)}|\tilde{\mathbf{G}}^{(t-1)}$) then re-estimating the expected value of $\mathbf{G}$ conditioned on the parameter MLE ($\tilde{\mathbf{G}}^{(t)}|[\widehat{\boldsymbol{\beta}}_{\text{QTL}}, \widehat{\boldsymbol{\theta}}]^{(t)}$), with t signifying the t^{th} iteration. This process is repeated until convergence in the MLE is reached ($[\widehat{\boldsymbol{\beta}}_{\text{QTL}}, \widehat{\boldsymbol{\theta}}]^{(t)} = [\widehat{\boldsymbol{\beta}}_{\text{QTL}}, \widehat{\boldsymbol{\theta}}]^{(t-1)}$). This mixture model, marginalized over the genetic state probability space, is the statistical procedure used in standard interval mapping (IM) (Lander and Botstein, 1989).

Though IM accounts for uncertainty in $\mathbf{G}$, it does not directly jointly model it along with the phenotype, which can result in issues. IM can be unstable (failing to converge or falling into local maxima) particularly if there is little information (high uncertainty) on $\mathbf{G}$, which can be more likely in MPP (as J, the number of founder alleles, increases). One possible alternative to IM draws from the previously mentioned Bayesian perspective which is to explicitly explore $\Pr(\mathbf{G})$ through sampling. Sampling will require a more complete definition of $\Pr(\mathbf{G})$.

4.2.3 Modeling and sampling genetic state

We assume that genetic states of individuals, the rows of $\mathbf{G}$, are independent of each other, thus $\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_n$ can be sampled independently. Violation of this assumption would occur in populations with some level of population structure, but it should not result in a bias even in

populations containing close relatives. An intuitive model for $\Pr(\mathbf{g}_i)$ is

$$\mathbf{g}_i \sim \text{Multinomial}(\text{size} = 1, \text{probs} = \boldsymbol{\phi}_i)$$

with $\boldsymbol{\phi}_i$ representing the K-element vector of genotype state probabilities, for which $\sum_{k=1}^{K} \phi_{ik} = 1$. $\boldsymbol{\Phi}$, an $n \times K$ probability matrix with individual $\boldsymbol{\phi}$ as rows, can be estimated with an HMM, using the information contained in the LD in a window of markers that surround a locus, as well as incorporating additional sources of noise (Mott et al., 2000; Fu et al., 2012). We sample genetic state for loci independently, though a multilocus approach could also possible (Sen and Churchill, 2001), based on sampling directly from the HMM (essentially sampling $\boldsymbol{\mathcal{G}}$, an $n \times K \times P$ tensor, P representing the number of loci). A full Bayesian approach would involve conducting the association procedure on each imputation s to produce an association score statistic, $u^{(s)}$. Inference would then be drawn from the posterior distribution of u over many imputations (many samples $\tilde{\mathbf{G}}^{(s)}$ from $\boldsymbol{\Phi}$) or alternatively through an importance sampling weighting scheme to reduce the sampling burden, as done in (Sen and Churchill, 2001) for simpler bi-parental populations. Extensively sampling from $\Pr(\mathbf{G})$ could require prohibitively large numbers of imputations, due to the complex probability space of $\Pr(\mathbf{G})$ that is a result of both the information content (quantified in $\boldsymbol{\phi}_i$) and the number of genetic states K. If the data include a large number of individuals and/or the model includes random effects, the complete Bayesian sampling approach can become unfeasible computationally.

4.2.4 Conservative multiple imputation procedure

To avoid a heavy sampling burden, our method is intentionally conservative and not formally Bayesian, primarily targeted at reducing the risk of false positive QTL signals stemming from uncertainty in genetic state. We do not seek to completely or approximately draw inference from posterior distributions, but rather assess the fragility of the association measured through hypothesis tests that results from the uncertainty around $\mathbf{G}$. We continue to use the single locus model previously described, now sampling imputations for the QTL term. The procedure uses the following steps:

1. Sample $\tilde{\mathbf{G}}_p^{(s)} \sim \text{Cat}(\Phi_p)$

- For $i = 1, 2, \ldots, n$:

  $\tilde{\mathbf{g}}_{ip}^{(s)} \sim \text{Multinomial}(\text{size} = 1, \text{probs} = \boldsymbol{\phi}_{ip})$

2. Regress $\mathbf{y}$ onto $\tilde{\mathbf{X}}_p^{(s)} = \tilde{\mathbf{G}}_p^{(s)}\mathbf{M}$ for which $\text{QTL}_i = \tilde{\mathbf{x}}_p^{\text{T}(s)}\boldsymbol{\beta}_{\text{QTL}}$

3. Conduct statistical association test for the presence of a QTL effect, resulting in statistic $u_p^{(s)}$

   - Compare H_0: $\boldsymbol{\beta}_{\text{QTL}} = \mathbf{0}$ versus H_A: $\boldsymbol{\beta}_{\text{QTL}} \neq \mathbf{0}$
   - The multiple imputations procedure is flexible to different statistical tests of association. An F test can be used when there is just one variance component present. With additional variance components, approximate F tests (Halekoh and Højsgaard, 2014) and likelihood ratio tests (LRT) are options.

4. Repeat steps 1-3 for $s = 1, 2, \ldots, S$ imputations

5. Summarize over S $u_p^{(s)}$ with an aggregate function: $\text{summary}(\mathbf{u}_p) = \bar{u}_p$ to produce score of association across the imputations

   - As summary, we use the median ($\bar{u}_p = \text{median}(\mathbf{u}_p)$).

6. Repeat steps 1-5 for $p = 1, 2, \ldots, P$ loci

4.2.5 Median as aggregate statistic

There has been substantial work on how to aggregate across imputations in terms of Frequentist inference (Li et al., 1991). Tradition, with respect to Wald statistics and likelihood ratio tests (Meng and Rubin, 1992), aggregate test statistics are estimated from averages of the regression coefficients. This is inconvenient for multiple imputation of genetic state because technically it all the data are observed with some level of uncertainty. If a genetic state is likely unobserved, it becomes like that an imputation of the data will not estimate certain genetic state effects, making it awkward to handle averages based on varying numbers of observations. It also does not easily generalize to fitting the locus effect as a random effect.

(Li et al., 1991) does mention an approximate approach of aggregating over p-values. Though such a multiple imputation statistic does not perform as optimally in terms of statistical properties, our goal is a conservative, computationally efficient method that will reduce false positives that

result from problematic uncertainty in genetic state. Towards this goal, summarizing over p-values is appealing due to its flexibility over differing underlying statistical models.

In terms of summaries of p-values, we find that the median has a number of appealing characteristics in comparison to an arithmetic mean. With an odd number of imputations, the median is scale independent, even over non-linear, monotonic transformations. This property removes all questions of which scale of a statistic to summarize over. In addition, the mean is sensitive to extreme values, which would emphasize the need for a more complete Bayesian approach with increased sampling $\Pr(\mathbf{G})$ or weighting through an importance sampling scheme. Therefore, the median is an intuitive and simple approach that reduces the influence of extreme statistics that result from a particular imputations from $\Pr(\mathbf{G})$, and thus gives a more stable point summary of the association across multiple imputation. Interval summaries can be estimated as confidence intervals on the median as well, based on the binomial distribution with probability parameter $\pi = 0.5$ (Ott and Longnecker, 2006), providing a clear way to characterize as well as visualize the stability of the association over imputations.

4.2.6 Assessing genome-wide significance

Genome-wide statistical significance can be assessed through repeating the procedure on permutations of the data when exchangeability can be reasonably assumed (Doerge and Churchill, 1996). Alternatively, if exchangeability cannot be assumed, there are alternatives that do not require it, such as parametric bootstrap samples from the null model. Maximum statistics from the permutation or bootstrap scans are used to fit an extreme value distribution, which is used to specify significance thresholds (Dudbridge and Koeleman, 2004; Valdar et al., 2006a).

4.2.7 Availability of data and software

All analyses were conducted using the R statistical programming language (R Core Team, 2018). The various modeling approaches described for QTL mapping, in particular ROP and MI, and plotting functions can be performed with the R package miqtl, which we make available through a GitHub repository at `https://github.com/gkeele/miqtl`.

4.3 Simulations

4.3.1 Simulated populations

Simulations were performed to evaluate how various MPP mapping procedures performed when the causal QTL was known. We simulated 100 samples or realizations of a panel of 200 recombinant inbred (RI) strains, based on the breeding scheme (20 generations of inbreeding) of the CC, using software as used in (Valdar et al., 2006a). We simulated only two chromosomes per individual: chromosome 1 consists of 101 markers, each equally spaced by 1 centi-Morgan (cM), with a single QTL that explains 10% of the phenotypic variation in the founders at the 55.5 cM position; chromosome 2 contains 201 markers, each spaced by a single cM, and carries no QTL. Both chromosomes allow us to observe the performance of the methods with and without a signal.

4.3.2 Tested mapping procedures

We used ROP and MI to analyze the simulated CC data, testing the locus effect as either a fixed effect or a random effect (Wei and Xu, 2016). There mapping approaches are flexible to other modeling considerations such as population structure and nuisance covariates. This flexibility is an valuable feature for many mapping populations, however, our use of simulated CC, which are approximately exchangeable, allow us to consider and compare the following less flexible methods.

We implemented traditional interval mapping through the EM algorithm, in which the probabilistic nature of the data are formally acknowledged by iteratively marginalizing over the genetic state's probability space. IM requires starting parameter values, and we use two approaches for initializing them. We used ROD coefficients and noise variance estimates, which would never be known in real analyses, as the starting values in what we refer to as the oracleIM. For standard IM, we used the sample variance of $\mathbf{y}$ as the starting value of σ^2 and set the starting founder haplotype effect parameters at 0.

Finally, we evaluated a maximally expanded data weighted least squares method that we call complete WLS. (Durrant and Mott, 2010) briefly mention expanded data approaches as an approximate Bayesian approach. Maximally expanded means that the data are expanded from n observations to $n \times K$: $\{\mathbf{y}_{\text{aug}}\}_i = y_i \times \mathbf{1}_{K\times 1}$. The design matrix is similarly expanded: $\{\mathbf{X}_{\text{aug}}\}_i = \mathbf{I}_{K\times K}\mathbf{M}$.

Finally the weights, $\{\mathbf{w}_{\text{aug}}\}_i = \phi_i$, match the corresponding K genetic states encoded in $\{\mathbf{\Phi}\}_i$ and $\{\mathbf{X}_{\text{aug}}\}_i$. When there is no uncertainty around $\mathbf{G}$, complete WLS will converge to ROD just as ROP, MI, and IM do, as the $K-1$ false genetic states will be given a weight of zero. Complete WLS has the unappealing characteristic of making the data very large, and thus potentially challenging computationally for large data sets. It also does not generalize in a reasonable way to allow for an additional variance component to account for population structure. We wished to see how it performed in comparison to the other methods in the setting in which these shortcomings are not present. **Table 4.1** summarizes the evaluated methods.

4.3.3 Simulation of uncertainty

We simulated the underlying haplotype states rather than marker genotypes, as these methods in MPP are primarily focused on haplotype-based association. Simulating haplotypes allows us to compare the performance of the methods when there is complete certainty to when there is uncertainty. We use two approaches to sample simulated probabilities from the true genetic states: probability dilution and Dirichlet sampling.

4.3.3.1 Probability dilution:

The probability dilution process converts a genetic state vector to a probability vector ($\mathbf{g} \rightarrow \tilde{\phi}$) in a deterministic manner. Let $\mathbf{g}_{i,p}$ be the K-element genetic state vector for individual i and locus p. The true genetic state, corresponding to a $k = t$ element of $\mathbf{g}_{i,p}$: $g_{itp} = 1$, and all other elements ($k \neq t$) are 0. We perform probability dilution by setting the probability of the true genetic state, α, to some value in $[0, 1]$ and all other elements to $\frac{1-\alpha}{K-1}$. Of note, for any specification of α, any individuals with the same genetic state vector $\mathbf{g}$ will also have the same realized probability vector $\tilde{\phi}$.

From a technical perspective, probability dilution does output probabilities, as they are non-negative and sum to 1. However, as simulations of uncertainty around genetic state, they are unrealistically clean, as a predetermined pattern of uncertainty will perfectly correlate with true genetic state. To break this hard correlation between uncertainty and genetic state, we use Dirichlet draws from the genetic state vector of an individual, which is the reverse process of our multiple imputation procedure.

4.3.3.2 Dirichlet sampling:

As with the probability dilution process, we have simulated individual i with a K-element genetic state vector $\mathbf{g}_{ip}$ for locus p. We sample a probability vector according to

$$\tilde{\boldsymbol{\phi}}_{ip}^{(s)} \sim \text{Dirichlet}\left(\tilde{n}\alpha\mathbf{g}_{ip} + \tilde{n}\frac{(1-\alpha)}{K-1}(\mathbf{1} - \mathbf{g}_{ip})\right),$$

such that $\mathbf{1}$ is a K-element vector of 1's, and $\tilde{n}$ is a total number of pseudocounts. Similar to our probability dilution specification, α is used to control the probability mass placed on the true and false genetic states; however, now the process is stochastic and thus described in terms of probabilistic expectations. The expected probability for the true genetic state and the false states follows from the Dirichlet: $\text{E}(\tilde{\phi}_i^{\text{true}}) = \alpha$ and $\text{E}(\tilde{\phi}_i^{\text{false}}) = \frac{1-\alpha}{K-1}$. The variance of the true probability states follows: $\text{Var}(\tilde{\phi}_i^{\text{true}}) = \frac{\tilde{n}\alpha(\tilde{n}-\tilde{n}\alpha)}{\tilde{n}^2(\tilde{n}+1)}$. We can set the expected uncertainty around the true genetic state to be equivalent to the pattern of uncertainty produced by probability dilution, but allowing for more realistic levels of noise. How far samples deviate from expectation due to the noise can be manipulated with the pseudocount parameter ($\tilde{n}$), with lower values having higher noise and higher values approaching probability dilution. A similar Dirichlet sampling scheme was used for SNP genotypes (Acar and Sun, 2013), however the the probability is spread more thinly in inbred MPP setting with $K = 8$ compared to $K = 3$ with bi-allelic SNPs.

It is challenging to simulated probabilistic genetic data as would likely be observed in real data, such as sets of genetic states being less distinguishable at certain loci, which is already known to hamper allele effect estimation (Zhang et al., 2014). Another potential realistic issue is differential levels of uncertainty for individuals in the same dataset, which will not be captured in these simulations. The space of potential patterns of uncertainty that could occur in MPP genetic states makes it effectively impossible to explore all of them in a systematic way. While these probability simulation schemes generally place the most probability mass on the true genetic state, which will generally favor ROP, we can still observe how the various mapping procedures perform with varying specifications of the simulation parameters.

4.4 Data Sets

We include examples from three different data sets that come from different types of MPP populations to demonstrate the differences between MI and ROP associations that can occur in actual data, which can differ strikingly from idealized simulated data.

The first example population is composed of 989 individuals from an outbred rat heterogeneous stock (HS) (Solberg Woods et al., 2012; Keele et al., 2018), referred to as HS1. The rats were measured on various diabetes and obesity phenotypes. The second population consists of 1407 individuals from a rat HS, independent from HS1 but derived from the same founder strains, that were measured for a large number of phenotypes, and are described in greater detail in (Baud et al., 2013, 2014). This population will be referred to as HS2.

The third population is from the CC. A more thorough description can be found in (Mosedale et al., 2017). Briefly, the experimental design involved treating four male mice from each of 45 CC lines with tolvaptan, a candidate treatment for kidney disease, while another four male mice from each line received vehicle (control) instead. In terms of modeling, CC lines from separate breeding funnels are approximately independent from each other, and thus do not require a random effect and associated genetic relationship matrix to model population structure. However, replicate observations from the same CC line due require special modeling considerations, such as a random effect with independent levels, or regression on strain means (Zou et al., 2006), as was used in (Mosedale et al., 2017).

We highlight differences between these populations to emphasize the need for statistical mapping procedures that can flexibly accommodate multiple sources of variation. For instance, experiments designed for the CC can have genetic replicates, whereas each HS individual has a unique genome. Although the HS will not contain the strongly structured correlations between individuals as expected from genetic replicates in the CC, the rotational breeding scheme produces more subtle population structure as a result of individuals being differentially related to each other rather than approximately equally related (as in CC lines from independent breeding funnels, like our simulations). ROP and MI are much more accommodating to these features than traditional IM, and for this reason, we used only them for analyses of real populations.

4.5 Results

4.5.1 Illustration of false association with ROP

Before presenting simulation results, we provide an example that demonstrates a need for our MI procedure in real data, in this case the HS1 rat population. In populations with high degrees of genetic state uncertainty and allele frequency imbalance, the genome scans from ROP and MI can strikingly differ, as in **Figure 1.1** which depicts analyses of serum cholesterol levels in HS1.

A striking characteristic of spurious QTL that occur with ROP is an extremely narrow association peak, depicted in **Figure 1.1A**, particularly the peak on chromosome 11. Many of these associations completely disappear with MI (**Figure 1.1E**), suggesting that the associations are completely the result of the ROP approximation. This conclusion is further strengthened by closer inspection of the uncertainty present at the sharp peak on chromosome 11, presented in **Figure 1.1F**. Notably, the B founder allele is highly unlikely to have been observed in the population at this locus, though is still present in the model due to uncertainty. Ultimately a highly inflated association score is produced due to very small dosages that strongly correlate with the phenotype. MI corrects for these occurrences because these rare alleles are rarely or never observed in the imputations. This inflation in significance can be seen in association studies with SNP dosages; however bi-allelic SNPs can be easily screened for low MAF, whereas many loci, even most in a population with high founder haplotype frequency imbalance, such as an HS, have founder haplotype alleles with very low frequencies.

4.5.2 Simulated results

Considering the strikingly different genome scans seen in the HS1 data, we assessed the performance of various mapping procedures in simulated MPP data. We simulated 100 realizations of breeding funnels corresponding to the strategy used for the CC, producing a panel of 200 RI strains. We then evaluated the performance of ROP fixef, ROP ranef, MI fixef, MI ranef, IM, oracleIM, and complete WLS (**Table 4.1**) on these simulated populations. We were primarily interested in how the methods responded to increasing level of uncertainty in the genetic state at the simulated QTL as well as at unassociated loci. The genome scan of a single realization of a simulated population,

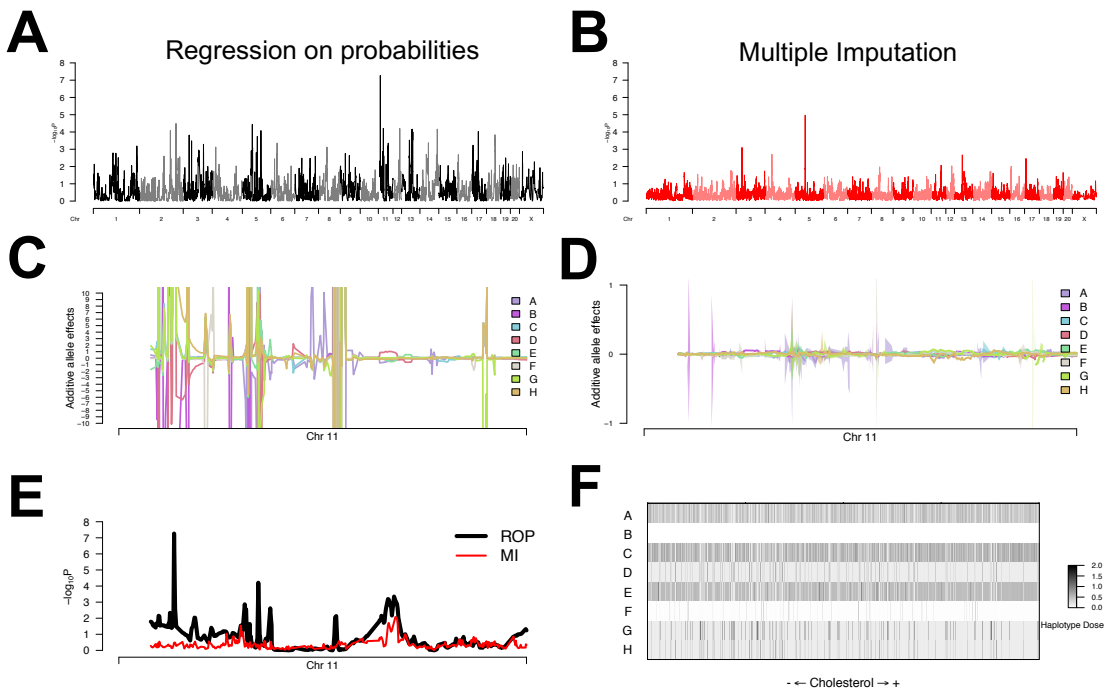

Figure 1.1: Analysis of blood cholesterol levels in the HS1 rat population. Genome scans with ROP (A) and MI (B) show drastically different patterns of association. Additive founder haplotype effects, labeled A through H, estimated from regression coefficients for ROP (C) and MI (D), across chromosome 11. The ROP effects are highly unstable, as a result of the highly unbalanced founder haplotype frequencies. For MI effects, transparent color bands represent the 95% confidence interval for the mean additive haplotype effect over 11 imputations, highlighting regions in which effects are unstable over imputations and in which an allele is unobserved. A comparison of the associations on chromosome 11 reveals that the sharp QTL associations that occur with ROP are almost completely reduced with MI (E). The uncertainty in haplotype dosages observed at the chromosome 11 locus is problematic for ROP and results in inflated associations that are not observed with MI (F). The rows of the probability grid plot represent the haplotypes of the founders. A column of the grid represents the genetic state of a single individual at the a single locus, with the shading of each cell representing the magnitude of haplotype dose. The individuals (columns) are ordered left to right by phenotype rank, allowing for potential haplotype effects to be seen from the raw data, which will appear as cluster in the founder haplotype rows. No clusters are immediately obvious, and more so, the uncertainty present in the dosages is extensive with founders D, G, and H being poorly distinguishable. Additionally there is broad founder allele frequency imbalance, with essentially no haplotype B alleles observed, and very little of the F allele. The strong association is a result of a strong correlation between near-zero probabilities in allele B and the phenotype.

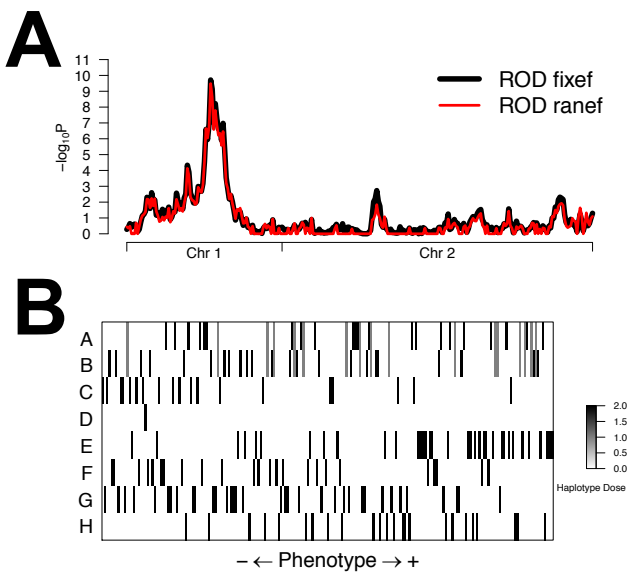

Figure 1.2: The first and second chromosomes of a simulated panel of 200 CC-like RI strains with a single QTL that explains 10% of phenotypic variance were simulated. The QTL is located on chromosome 1, and no QTL are present on chromosome 2. Association scans of a single simulated population (A). The fixed effect procedure (fixef) and random effect procedure (ranef) generally track similarly, though ranef produces a lower significance score of association due to shrinkage. The simulated haplotype counts represented as an 8×200 grid, ordered along the x-axis by phenotype (B). A single column of the grid represents the genetic state of an individual at the QTL, with the shading of i, j-th cell corresponding to count of haplotype j for individual i. As there is no uncertainty of genetic state, all shaded cells represent counts of 0, 1, and 2, which are white, gray, and black respectively. A Non-zero effect is clear in E, and potentially other alleles. The grid gives a clear visual representation of the level of uncertainty at a locus for a given population.

as well as a depiction of the underlying genetic state at the QTL position, when no uncertainty is obscuring genetic state can be seen in **Figure 1.2**.

Given a simulated population with known genetic states for all individuals, we then simulated uncertainty in genetic state through either probability dilution or Dirichlet sampling from the vector of true genetic state.

4.5.2.1 Dirichlet sampling:

For a single simulated population, as expected, as the level of uncertainty increases, the association at the QTL is reduced (**Figure 1.10**). Complete WLS seems the most penalized by increasing uncertainty, then MI, and finally ROP and IM do the best. We include only the fixed effect models of ROP and MI to both minimize visual clutter and because the fixed effects models are more

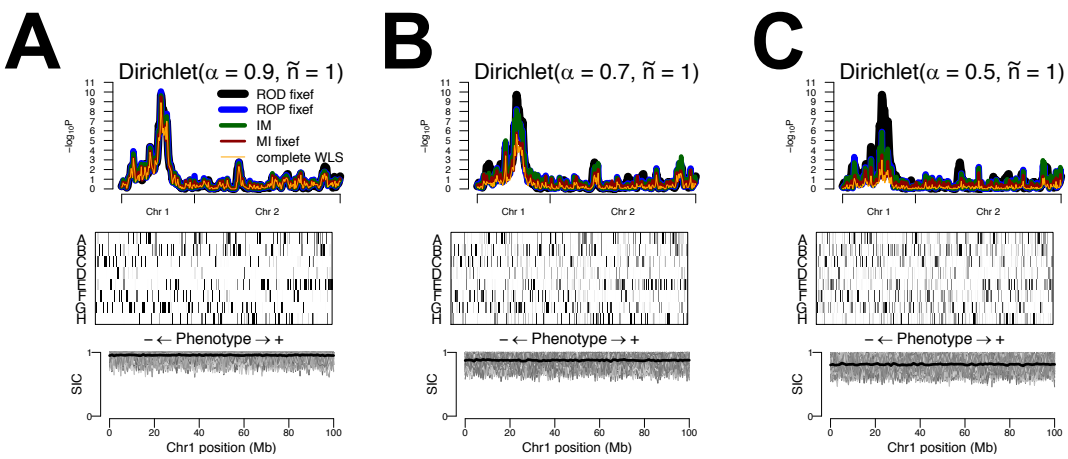

Figure 1.3: Comparison of mapping approaches with varying levels of Uncertainty in genetic state simulated through Dirichlet sampling with $\alpha = 0.9$ (A), $\alpha = 0.7$ (B), and $\alpha = 0.5$ (C), all with pseudocount $\tilde{n} = 1$ in 200 simulated CC-like RI strains with a QTL that explains 10% of phenotypic variation on chromosome 1. With Dirichlet sampling, α is the expected probability mass on true genetic state and $\tilde{n}$ determines the variability over Dirichlet samples. For a given α, the expected probability vector is equivalent to the probability vector from the probability dilution process, as in Figure 1.10 ($\mathrm{E}(\boldsymbol{\phi}_{i,\alpha}^{\text{Dirichlet}}) = \boldsymbol{\phi}_{i,\alpha}^{\text{Dilution}}$). The middle plot of each subfigure depicts the simulated uncertainty at the QTL for the three scenarios. The bottom plot of each subfigure is the the SIC across chromosome 1. Described in greater detail in (Rönnegård and Valdar, 2011), briefly, SIC is a standardized Kullback-Leibler divergence, which we use as a measure of the information content on the genetic state probabilities of an individual ($\boldsymbol{\phi}_i$) at a locus. SIC $\in [0, 1]$, with the boundaries corresponding to genetic states being completely indistinguishable ($\boldsymbol{\phi}_i = \frac{1}{K} \times \mathbf{1}_{K\times 1}$) and complete certainty ($\boldsymbol{\phi}_i \rightarrow \mathbf{g}_i$), respectively. It is important to note that SIC does not reflect whether the information in the genetic state probabilities are correct.

comparable to IM and complete WLS. In the single simulation, we do not see inflated spurious associations with ROP as seen in real data, though it is important to look across all the simulated data sets.

Across all simulated data, we see similar trends to the single data set (**Figure 1.4**). To assess the mapping procedures across many simulated populations, we evaluated the change in p-value at the QTL and unassociated locus. The associations from ROP and IM degrade the least as uncertainty increases, and MI performs better than complete WLS. At an unassociated locus, all the procedures performed similarly and did not produce false associations. The conservative nature of MI and complete WLS compared to ROP and IM was consistent at the QTL and an unassociated locus.

4.5.2.2 Probability dilution:

The results for simulations of genetic state uncertainty through probability dilution, a deterministic process, were consistent with Dirichlet sampling, though with a few notable exceptions. For the single simulated population, ROP performs as well as ROD, which can be seen in **Figure 1.10**. The performance of IM at the QTL is reduced, but to a lesser extant than seen with Dirichlet sampling, despite the SIC content being lower with dilution, suggesting that the noisier Dirichlet sampling obscures the signal more. MI and complete WLS are both penalized similarly with either Dirichlet sampling or dilution.

Across all 100 simulated populations, we see these same trends. ROP loses none of the statistical signal, except at the point in which each genetic state has the exact same probability (**Figure 1.11**). IM performs better than with Dirichlet sampling, though at the unassociated locus, some false positives occur when uncertainty is very high (**Figure 1.12**). MI and complete WLS are consistent across Dirichlet sampling and probability dilution, at the QTL and unassociated locus.

The probability grid plots in **Figures 1.11, 1.12 [bottom row]** reveal that probability dilution is a very artificial form of uncertainty simulation, as no noise is incorporated to distort the true signal. Thus ROP is able to handle the probabilities as artificial doses that perfectly track with the simulated truth. Similarly, IM handles the uncertainty from probability dilution very well, with only minor reduction in the association at the QTL, though the association does become more unstable across simulations and potentially inflated at the null locus where there is no signal. Likely, this strong performance results from the fact that though IM is probability aware, it is not sampling over the

Table 4.1: Mapping procedures for simulated data (**Figures 1.3, 1.4, 1.5, 1.10, 1.10, 1.11, 1.12**)

Mapping procedure	Color	Uncertainty status	Description
ROD fixef		No uncertainty	Regression with complete certainty on genetic state and locus effect fit as a fixed effect
ROP fixef		Unaware	ROP with locus effect fit as a fixed effect
ROP ranef		Unaware	ROP with locus effect fit as a random effect
MI fixef		Aware	MI with locus effect fit as a fixed effect
MI ranef		Aware	MI with locus effect fit as a random effect
IM		Aware	IM with initial parameter values set to sample founder means and sample variance
oracle IM		Aware	IM with initial parameters set to ROD fixef parameter estimates
complete WLS		Aware	Weighted regression with full data expansion of genetic states

space, but rather iteratively marginalizing over the genetic state space. When no noise is incorporated into the system, as with probability dilution, this marginalization is very effective at capturing the signal. MI involves a full sampling process, and is thus heavily penalized by the greater uncertainty, which it explores despite the signal not being reduced through dilution. Complete WLS is similarly harshly penalized.

We find that Dirichlet sampling of genetic state uncertainty more equally affects all mapping procedures, and more realistically represents realistic patterns of uncertainty. From the probability grids and SIC plots, it is clear that the level of uncertainty is actually lower in the Dirichlet simulations, but that it importantly does not track perfectly with the true genetic state. This leads to ROP and IM being penalized as well, though they still perform better than MI and complete WLS. We also re-emphasize that given an α, as $\tilde{n} \to \infty$, then Dirichlet sampling will converge to probability dilution.

4.5.3 More examples of results in real populations

The analyses of simulated data found ROP be an effective approximate approach that performs as well as IM when genetic state uncertainty is realistically simulated through probability dilution. Despite the strong performance of ROP in simulated data, we return to real data to focus on populations with lower levels of genetic state uncertainty and more balanced founder haplotype

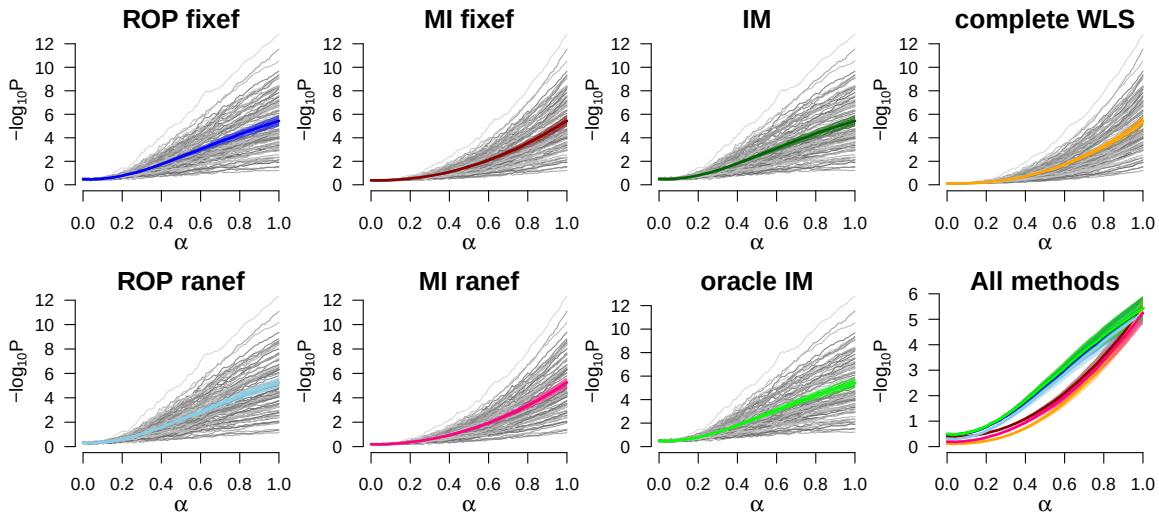

Figure 1.4: The change in association at 10% QTL with varying levels of uncertainty, simulated through Dirichlet sampling. α is the expected probability mass of the true genetic state, with the pseudocount parameter ($\tilde{n}$) determining sampling variance. We have set $\tilde{n} = 1$. Colored lines and transparent bands represent the mean p-value and 95% confidence interval on the mean p-value for the various mapping procedures (**Table 4.1**) over the 100 populations. Gray lines represent the $-\log_{10}$ P for a single population. Because the Dirichlet sampling process is random, the gray lines represent mean $-\log_{10}$ P from 100 Dirichlet sampling steps of the simulated population.

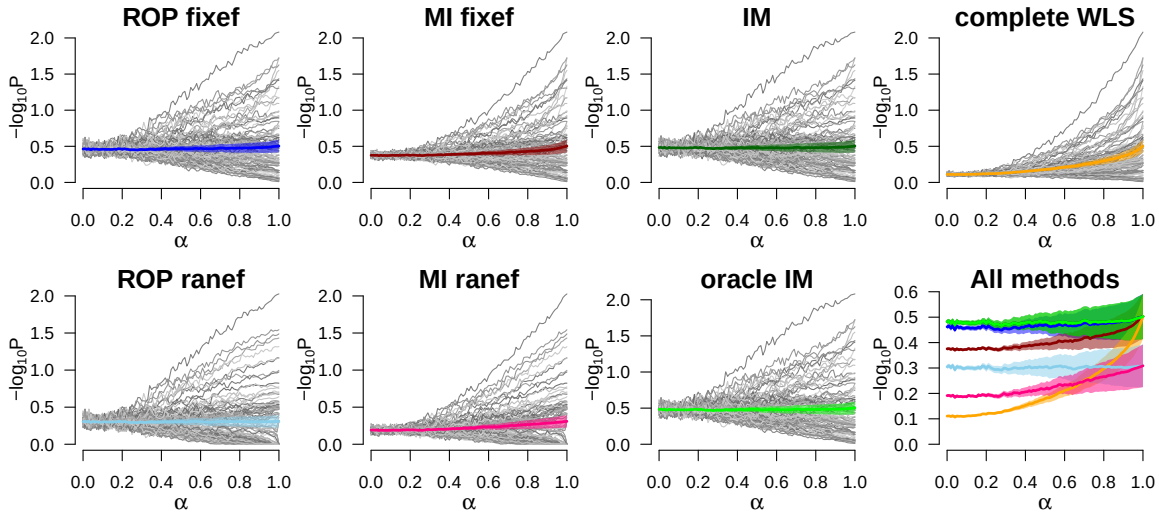

Figure 1.5: The change in association at a null locus with varying levels of uncertainty, simulated through Dirichlet sampling. α is the expected probability mass of the true genetic state, with the pseudocount parameter ($\tilde{n}$) determining sampling variance. We have set $\tilde{n} = 1$. Colored lines and transparent bands represent the mean p-value and 95% confidence interval on the mean p-value for the various mapping procedures (**Table 4.1**) over the 100 populations. Gray lines represent the $-\log_{10}$ P for a single population. Because the Dirichlet sampling process is random, the gray lines represent mean $-\log_{10}$ P from 100 Dirichlet sampling steps of the simulated population.

frequencies compared to HS1 to assess whether ROP could still produce inflated associations in real data that are more similar to the simulated data. In actual CC data, from which the simulated populations were modeled, we still see less extreme examples of the inflated associations (as in HS1) that are reduced though multiple imputation (**Figure 1.6**). In particular, the chromosome 14 peak is stable over multiple imputation, whereas chromosomes 12 and 16 have narrow peaks that are reduced (**Figure 1.6C**).

In real CC data, false associations can be inflated beyond what was observed with the clean simulated data. Though more conservative than ROP, MI can reduce these associations while detecting stable associations. The ability of MI to detect QTL is further demonstrated in the larger HS rat population HS2 (**Figure 1.7**). The similarity of the genomes scans of HS2 through ROP and MI suggest there is less genetic state uncertainty and founder haplotype imbalance in HS2. With MI the strong associations at chromosomes 4 and 8 are maintained in MI, and are even strengthened compared to the other associations due to the less inflated MI associations. Lesser ROP significant associations on chromosomes 4 and 11 drop below significance with MI. MI also appears to support a peak on chromosome 14 as being near significance compared to its association in ROP. These results show that MI can reliably capture similar associations as ROP, while generally reducing or removing questionable ones.

4.5.4 Founder haplotype frequency and haplotype uncertainty

The previous results highlight the potentially drastic disparities in performance of MI compared to ROP, particularly in the HS1 population. As previously stated, this is in large part due to founder haplotype frequency imbalances compounded with haplotype uncertainty. The rotational breeding scheme used for the HS populations is expected to result in more founder allele imbalances than in the CC, as is seen in **Figure 1.8**, which depicts histograms of the founder haplotype allele frequencies across the genome for each population. The founder haplotype imbalance in HS1 appears to be more extreme than in HS2, as seen in the even greater enrichment for very low founder haplotype allele frequencies, thus explaining the comparatively less stable ROP genome scans.

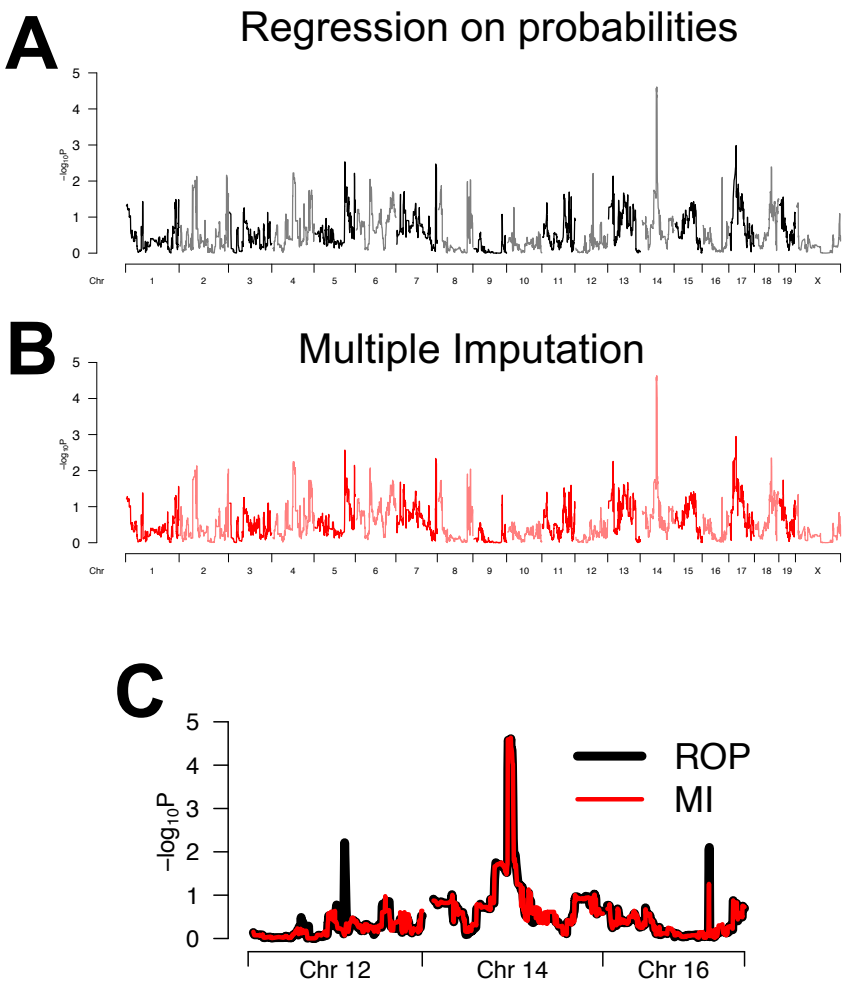

Figure 1.6: Example analysis of response to kidney drug in a sample population of 45 CC strains with a total of 159 individuals. Genome scans through ROP (A)and MI (B) show similar associations, though MI lowers some narrow signals. Notable QTL associations for chromosomes 12, 14, and 16 for ROP and MI (C). A notable signal that is consistent across ROP and MI is on chromosome 14. The chromosome 12 and 16 signals are only present with ROP.

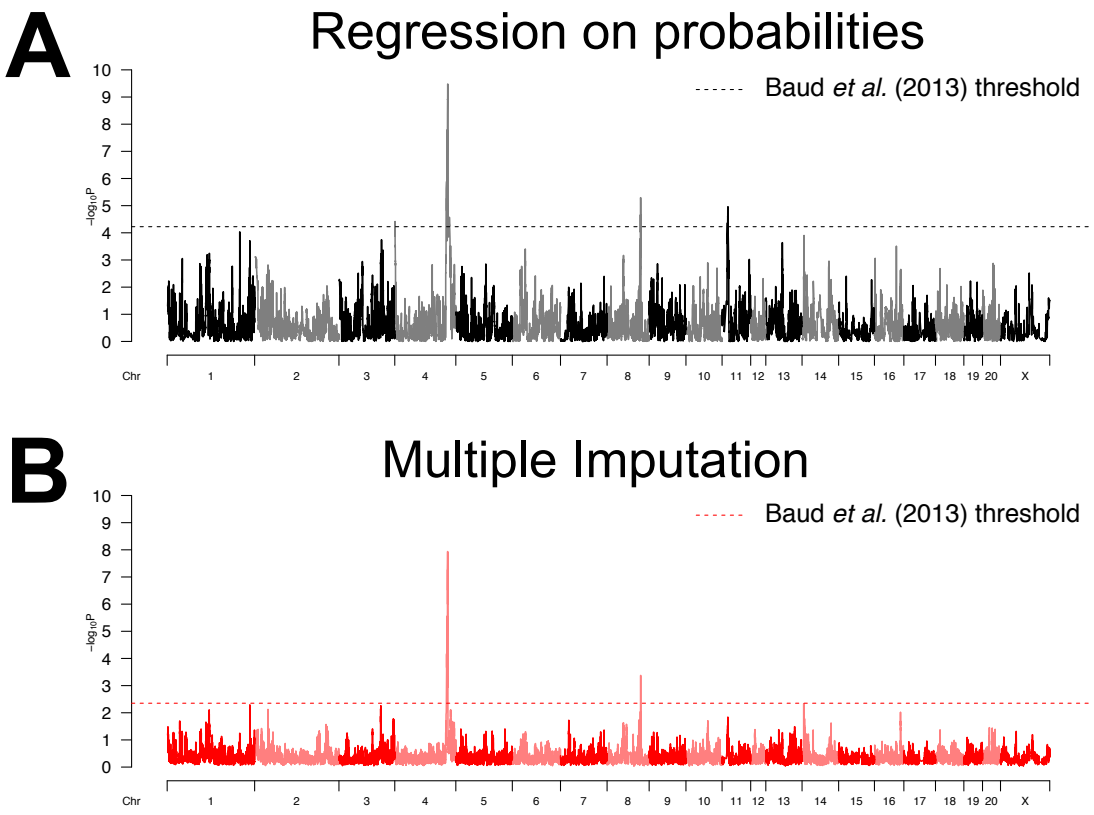

Figure 1.7: Genome scans of platelet counts, using ROP (A) and MI (B) in the HS2 rat population, with thresholds for both methods as described in (Baud et al., 2013). MI is more stringent and prioritizes the QTL observed on chromosomes 4 and 8.

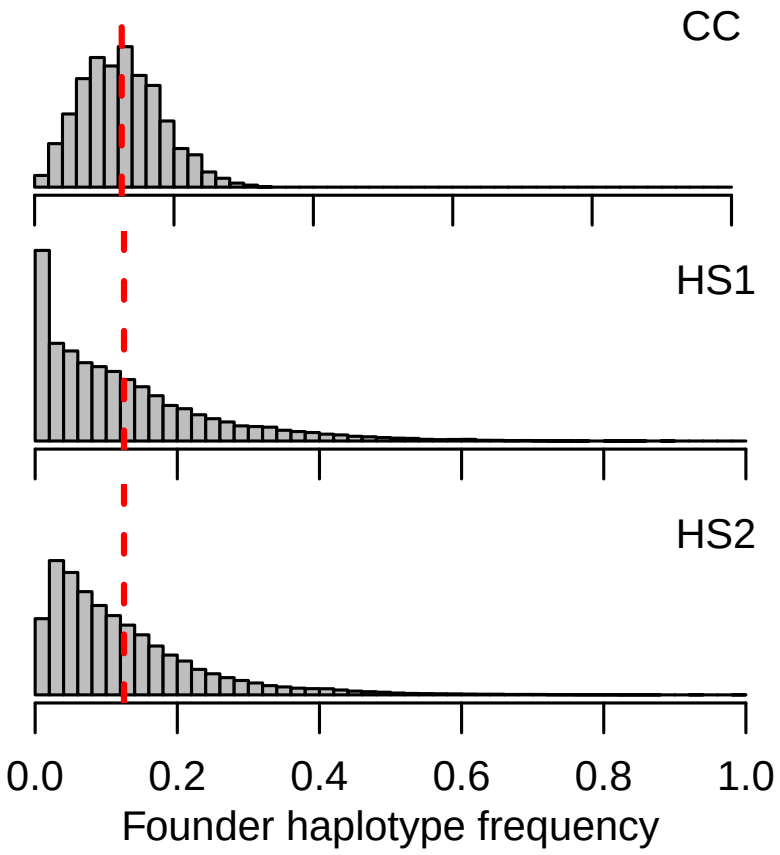

Figure 1.8: Histograms of founder haplotype allele frequencies of loci across the genome for the CC, HS1 and HS2 populations. The vertical dashed red line represents an allele frequency of $1/8$, which would be the mean allele frequency of a perfectly balanced MPP with eight founders. The CC have nicely balanced allele frequencies across the genome, whereas, as expected from rotational breeding, the HS populations have more imbalances, particularly in HS1. These frequencies are based on allele dosages, thus incorporating uncertainty into their estimates.

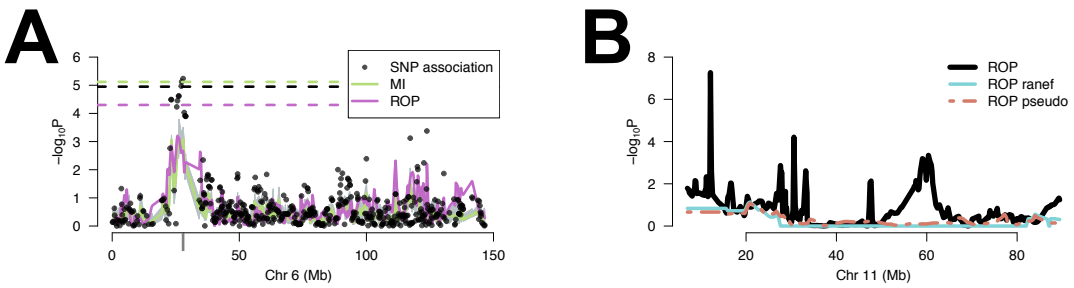

Figure 1.9: Alternative approaches to ROP fixef haplotype association in HS1. Chromosome 6 scans in HS1 of retroperitoneal fatpad mass (A). ROP, MI, and SNP association are included. Although the there is an increased association signal in the haplotype-based methods, the estimation of extra haplotype parameters results in a statistical burden and the associations do not rise above statistical significance. At this locus, one founder possesses a unique SNP allele, and along with the collapsing of the other founders into the other SNP allele, a strong signal is detected. Note the band around the MI association line, which is the 95% confidence interval on the median p-value across imputations. If the founder haplotype alleles can be captured in a model with less parameters, an increase in power is expected. Comparisons of ROP with the QTL effect fit as a fixed effect with ROP procedures that use shrinkage approaches, either by fitting the QTL as a random effect with a corresponding variance component or through null pseudo-observations, cumulatively weighted to be a single data point (B). Both shrinkage approaches remove the sharp association seen in standard ROP by harshly down-weighting the signal from the near-zero probabilities of the founder B allele (**Figure 1.1F**).

4.6 Discussion

In the context of standard and generalizable regression-based QTL mapping in MPP, our multiple imputation approach provides an intuitive approach to incorporate genetic state uncertainty. Theses patterns of uncertainty could present in different ways, and is a more likely issue in MPP where K, the number of genetic states, becomes large, particularly in comparison to SNP association ($K = 3$). With higher K and certain breeding designs, it becomes more likely that founder haplotype alleles will not be observed at a locus; however, some probability mass is likely still attributed to the founder allele that has been lost because the genetic states are being inferred. This results in the situation highlighted in **Figure 1.1F** in which near-zero probabilities induce an artificial association. In addition to MI, we considered other approaches to counter spurious associations, but also more powerfully map in these populations than may be possible with standard haplotype-based association.

4.6.1 SNP association as an alternative to ROP

One potential alternative to haplotype-based association is SNP association, which is similar to what is commonly used in human GWAS. The founder allele imbalance seen across the genome in HS1 (**Figure 1.8**), as well as the pattern of uncertainty exhibited in **Figure 1.1F**, in which allele B is not observed, F is very rarely observed, and D, G, and H are poorly distinguishable, exemplify a population that is extremely problematic for ROP, or any form of haplotype association. These problems can be greatly reduced through a simpler genetic model with less genetic states, as in SNP association. In effect, a SNP genetic model implicitly reduces the number of genetic states K at a locus, reducing the number of alleles from the J founders to two in a bi-allelic SNP, making it unlikely that an allele is unobserved or very rare. ROP-like SNP-based procedures, as described earlier, could be used, and thus not requiring the additional computational burden of multiple imputation. In MPP with poor founder haplotype reconstructions, SNP association may have greater statistical efficiency compared to MI, whereas ROP would be prone to spurious associations. **Figure 1.9A** presents a case in HS1 in which SNP association was found to be more powerful for detecting a QTL (Keele et al., 2018). We do not suggest that SNP association is superior for QTL mapping in MPP, but that in situations where founder haplotypes are imbalanced and reconstruction problematic, SNP association can provide a simplified genetic model and avoid spurious associations from ROP. Conversely, SNP association will be less powerful for detecting associations that track with a specific founder haplotype, potentially do to local epistatic interaction, and thus does strongly correlate with a genotyped variant.

4.6.2 Shrinkage as an alternative to ROP fixef

Shrinkage presents a particularly attractive statistically-oriented approach for dealing with issues that result in MPP due to unbalanced founder haplotype frequencies. Rather than fitting an allele parameter wholly-based on very few observations (or even near-zero probabilities), information is shared across genetic states, resulting in predictors of the allele effects that are shrunk toward the overall mean, with more shrinkage present in poorly represented allele, and ultimately resulting in a more conservative modeling approach. This borrowing of information is accomplished by specifying a variance component on the QTL effect; consider the QTL term in the regression model:

$\text{QTL}_i = \mathbf{x}_i^{\text{T}}\boldsymbol{\beta}_{\text{QTL}}$, then $\boldsymbol{\beta}_{\text{QTL}} \sim \text{N}(\mathbf{0}, \mathbf{I}\tau^2)$. Rather than the more conventional fixed effect test of $H_A : \boldsymbol{\beta}_{\text{QTL}} \neq \mathbf{0}$, the variance component can be tested, $H_A : \tau^2 \neq 0$ (Wei and Xu, 2016). This approach was also included in the analyses of simulated data (as seen in **Figures 1.2, 1.3, 1.10, 1.4, 1.5, 1.11, 1.12**).

A random effect fitting of the QTL term presents computational challenges compared to fixed effects model, due to need to optimize the likelihood with respect to multiple variance components. This approach could become unfeasible in large samples, particularly in terms of determining significance thresholds through permutations or null bootstraps.

An approximate approach to shrinkage is to include pseudo-observations in the data set. These observations $\tilde{\mathbf{y}}$ represent expectations from H_0, the model of no QTL effect. Generally $\tilde{\mathbf{y}}$ will contain between K and j elements, depending on the model being fit. Furthermore, these pseudo-observations can be given fractional weights, allowing for the cumulative amount of null pseudo-data to be less than the number of elements of $\tilde{\mathbf{y}}$.

The pseudo-data approach to shrinkage can be made as computationally efficient as standard ROP, making large data sets and computationally expensive procedures like permutations feasible. An unappealing feature of the approach is that selecting the portion of null data to add is arbitrary. In addition, drawing null observations from H_0 can be done with varying degrees of sophistication, and becomes more complicated with increasingly complex models, such as when covariates are included and a random effect is used to model a polygenic term. Both approaches to shrinkage completely remove the sharp association observed on chromosome 11 for the HS1 rats (**Figure 1.9B**).

4.6.3 Disparity between ROP in simulated and real data

Based on the simulated data with Dirichlet sampling, ROP performs as well as IM, and is also computationally more efficient and generalizable to other populations. In part we chose simulations of a balanced population like the CC to allow for a greater number of methods to be easily compared with ROP and MI, in particular IM and complete WLS. Our findings suggest that ROP performs well when the data are well-balanced and the genetic state is well-behaved, as with probability dilution; however, deviations from such a setting can result in the inflated associations, which will be pervasive in certain populations (HS1) and still present to a lesser degree in realized balanced populations (CC).

As such, MI provides a mapping approach that can restrict these inflated association scores in real data.

4.6.4 Summary

We propose a multiple imputation linear regression procedure for QTL mapping in MPP that accounts for uncertainty in genetic state, which in practice protects against detection of spurious signals caused by unexpected correlations between the phenotype and near-zero allele probabilities or dosages. Our method is flexible to many modeling features, such as population structure modeled as polygenic effect with a corresponding variance component, which is an important consideration in many MPP. The procedure as currently specified uses a single locus model and can easily be used for data being analyzed through ROP, commonly done within software such as the R packages DOQTL (Gatti et al., 2014) and qtl2 (Broman, 2017). Its computation scales with ROP linearly in terms of the number of imputations performed.

We found the standard ROP approach performed exceptionally well in simulated data, both simulated through probability dilution and Dirichlet sampling, and generally failed to capture or reflect the pattern of associations seen in many of the real data sets we analyzed. This led us to realize that the probability space of the genetic state in MPP is particularly large and complex. Although ROP performs well as a computationally efficient and stable approximation of uncertainty-aware statistical procedures like interval mapping, it clearly struggles with faced with particularly problematic patterns of uncertainty that are also challenging to reliably simulate in practice. Our simulations reveal that the MI procedure is conservative compared to ROP, particularly as uncertainty in genetic state increases. However, we find an alternative conservative procedure preferable when in real data the standard is producing clearly artificial associations.

We have proposed multiple approaches to limiting false positives that result from founder allele frequency imbalances and haplotype uncertainty in MPP, describing a multiple imputations procedure in great detail. MI is easily understood as well as implement, and also flexible to many modeling considerations (alternative models of genetic state, covariates, polygenic term, etc.). Our multiple imputation approach can also incorporate shrinkage methods through a formal random effect fitting of the QTL, which is computationally intensive, and through specification of pseudo-observations. These described approaches provide options for addressing the problems that result

from founder haplotype allele frequency imbalances, which can be further compounded though haplotype uncertainty. Though the burden of haplotype uncertainty should lessen as sequencing technologies improve, become less expensive, and are even more extensively used, founder alleles imbalances can still occur in MPP due to genetic drift, certain breeding schemes, and small sample sizes. Methods such as we describe here will be important for reducing the number of false positive associations that are reported, and ultimately feed negative narratives about genetic association studies producing findings that do not replicate.

4.7 Additional Figures

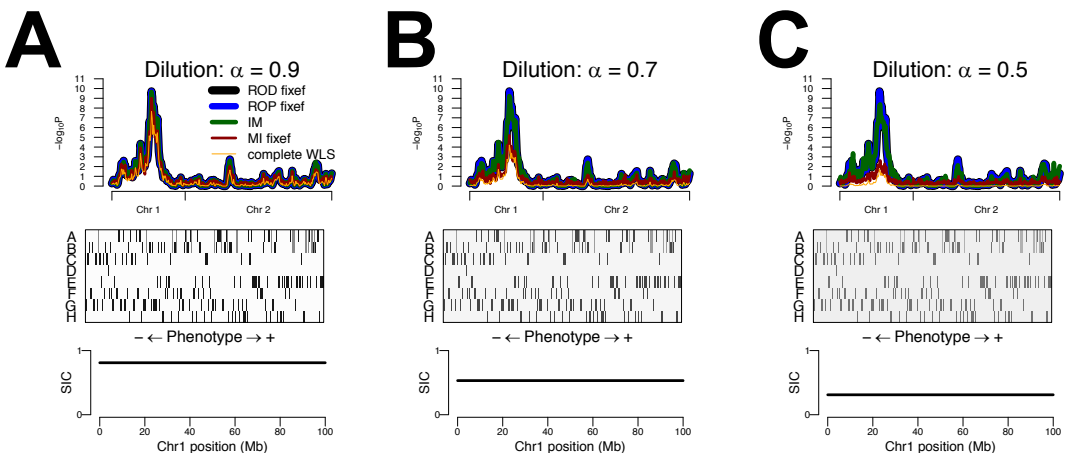

Figure 1.10: Comparison of mapping approaches with varying levels of Uncertainty in genetic state simulated through probability dilution with $\alpha = 0.9$ (A), $\alpha = 0.7$ (B), and $\alpha = 0.5$ (C) in 200 simulated CC-like RI strains with a QTL that explains 10% of phenotypic variation on chromosome 1. Probability dilution is deterministic, with α representing the probability placed on the true genetic state and the remaining mass being evenly allocated to the remaining genetic state categories. The top panel for each subfigure is the genome scan comparing four mapping procedures as well ROD. The middle plot of each subfigure depicts the simulated uncertainty at the QTL for the three scenarios. The bottom plot of each subfigure is the the SIC across chromosome 1, which is summary of information content.

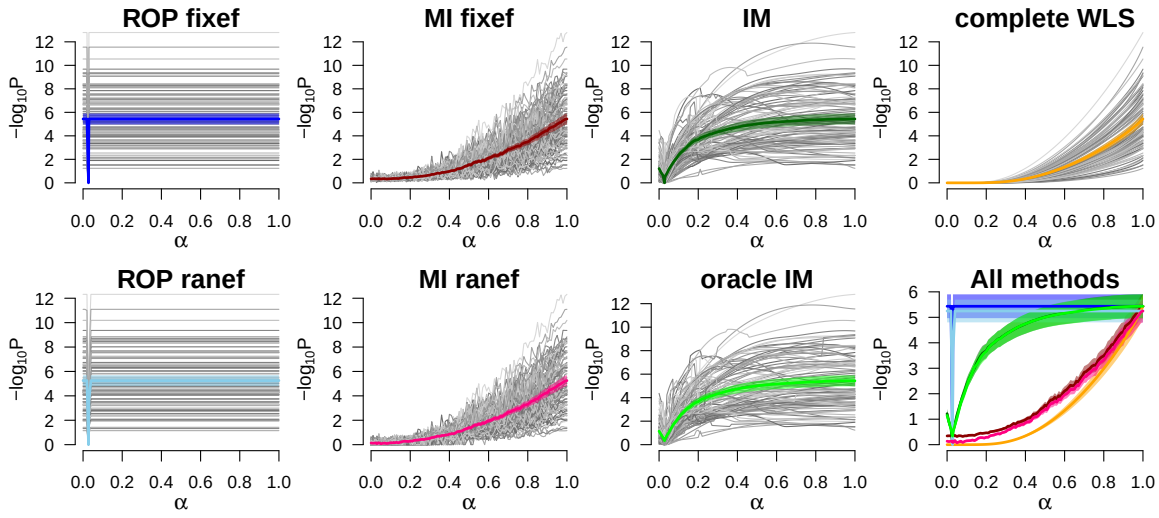

Figure 1.11: The change in association at 10% QTL with varying levels of uncertainty, simulated through probability dilution. α is the probability mass of the true genetic state, with the remaining probability mass being split evenly across the other genetic states. Colored lines and transparent bands represent the mean p-value and 95% confidence interval on the mean p-value for the various mapping procedures (**Table 4.1**) over the 100 populations. Gray lines represent the $-\log_{10}$ P for a single population.

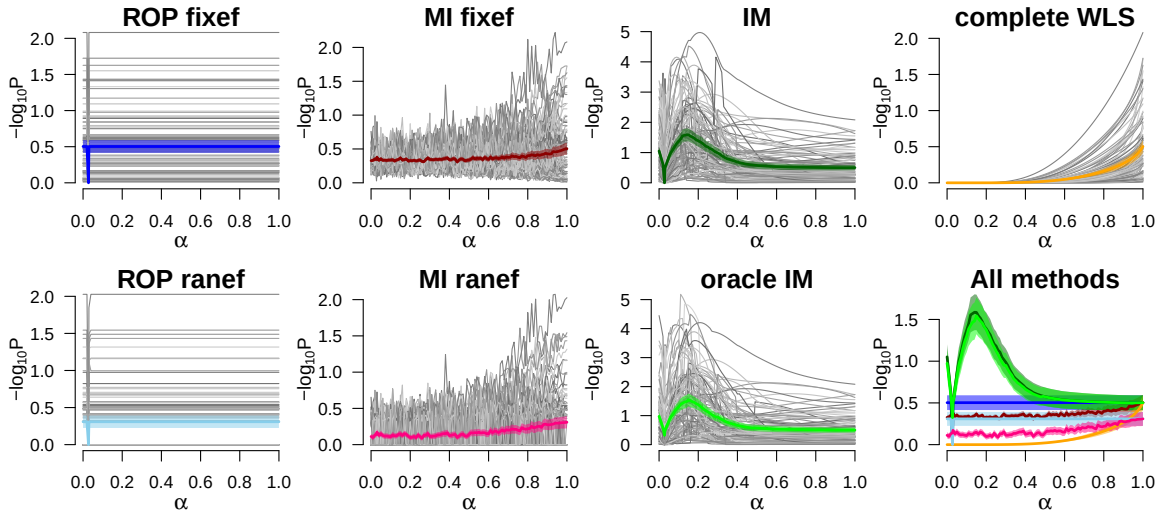

Figure 1.12: The change in association at a null locus with varying levels of uncertainty, simulated through probability dilution. α is the probability mass of the true genetic state, with the remaining probability mass being split evenly across the other genetic states. Colored lines and transparent bands represent the mean p-value and 95% confidence interval on the mean p-value for the various mapping procedures (**Table 4.1**) over the 100 populations. Gray lines represent the $-\log_{10}$ P for a single population.

CHAPTER 5

QTL mapping in outbred rat population with imbalanced founder allele frequencies [1]

5.1 Introduction

Obesity and overweight are major risk factors for multiple cardiovascular and metabolic diseases (Wang et al., 2009). Of particular importance is visceral, or abdominal, adipose tissue, which is strongly predictive of metabolic health (Emdin et al., 2017). Multiple environmental (e.g., lifestyle) and genetic factors contribute to obesity with genetics accounting for up to 70% of the population variance for human body mass index (BMI) and obesity (Stunkard et al., 1986) and visceral adiposity (Katzmarzyk et al., 2000). To date, human genome-wide association studies have identified many genes for anthropomorphic traits (Locke et al., 2015; Lu et al., 2016; Ng et al., 2017; Speliotes et al., 2010), but these genes explain only a small proportion of the heritable variation (Locke et al., 2015), indicating many genes are yet unidentified. Identification of additional genes is particularly important because there has been a steady increase in prevalence of overweight and obesity since the 1970's (Wang et al., 2009), with over one third of adults and almost one fifth of all children in the United States being classified as obese (Flegal et al., 2016).

One strategy for identifying the heritable modifiers of obesity is to control for exogenous environmental factors using experimental genetic mapping strategies such as the outbred heterogeneous stock (HS) rats. HS rats descend from eight inbred founder strains and have been out-bred for over 70 generations, such that the fine recombination block structure allows genetic mapping to identify regions that are only a few Mb (Solberg Woods, 2014). In previous work, we used HS rats to fine-map a single region on rat chromosome 1 previously identified for glucose and insulin traits (Solberg

[1] This chapter has been adapted from a paper published in Obesity. The citation will be as follows: 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.

Woods et al., 2010, 2012), and identified *Tpcn2* as a likely causal gene at this locus (Tsaih et al., 2014). Here, we demonstrate that HS rats vary for adiposity traits including body weight and visceral fat pad weight, and that these measures correlate with metabolic health. We then detect and fine-map QTL for these traits genome-wide and identify five likely causal genes within these loci.

5.2 Methods and Procedures

5.2.1 Animals

5.2.1.1 Heterogeneous stock colony

The NMcwi:HS colony, hereafter referred to as HS, was initiated by the NIH in 1984 using the following eight inbred founder strains: ACI/N, BN/SsN, BUF/N, F344/N, M520/N, MR/N, WKY/N and WN/N (Hansen and Spuhler, 1984). This colony has been maintained at the Medical College of Wisconsin since 2006 and has been through over 70 generations of breeding. Rats were given ad libitum access to Teklad 5010 diet (Harlan Laboratories). Additional housing conditions are detailed in Detailed Methods.

5.2.1.2 Founding inbred sub-strains

Other than M520/N (now maintained at MCW), phenotyping of the founders was conducted in the following sub-strains (abbreviated names to be used throughout manuscript in parentheses): ACI/Eur or ACI/Seg (ACI), BN/SsnHsd (BN), BUF/NHsd (BUF), F344/NHsd (F344), and WKY/NHsd (WKY). We tested 8-19 male rats per inbred strain.

5.2.2 Phenotyping protocol

We measured body weight at 16 weeks of age in 989 male HS rats. Rats underwent an intraperitoneal glucose tolerance test (IPGTT) as described previously (Solberg Woods et al., 2010, 2012). We used the Ascensia Elite system for reading blood glucose values (Bayer, Elkhart, IN). Plasma insulin levels were determined using an ultrasensitive ELISA kit (Alpco Diagnostics, Salem, NH). The following metabolic measures were calculated: area under the curve for glucose (glucose_AUC) and insulin (insulin_AUC) during the IPGTT, the quantitative insulin sensitivity check (QUICKI) as

a measure of insulin sensitivity, and the insulinogenic index (IGI) as a measure of beta cell sensitivity to glucose (Solberg Woods et al., 2012).

Inbreds and 743 HS rats were euthanized after an overnight fast at 17 weeks of age. Body weight and two measures of body length (from nose to base of the tail and from nose to end of tail) were collected, allowing us to calculate two measures of body mass index: BMI_Tail_Base and BMI_Tail_End. BMI was calculated as: (body weight/body length2)$\times$10. Rats were euthanized by decapitation and trunk blood was collected. Fasting cholesterol and triglycerides were determined from fasting serum on an ACE Alera autoanalyzer using an enzymatic method for detection. Several tissues were dissected and weighed including retroperitoneal and epididymal visceral fat pads, hereafter referred to as RetroFat and EpiFat, respectively. Liver and adipose tissues were snap-frozen in liquid nitrogen for subsequent expression analysis. All protocols were approved by the IACUC committee at MCW. Phenotyping data have been deposited in RGD (`www.rgd.mcw.edu`).

5.2.3 Genotyping

We extracted DNA from tail tissue from HS and the original eight inbred founder strains (tissue obtained from the NIH) using either the Qiagen DNeasy kit (Valencia, CA) or a phenol-chloroform extraction. Founder and HS rats were genotyped using the Affymetrix GeneChip Targeted Genotyping technology on a custom 10K SNP array panel as previously described (STAR Consortium et al., 2008), with marker locations based on rat genome assembly 6.0. 147 samples were genotyped by the Vanderbilt Microarray Shared Resource center at Vanderbilt University in Tennessee (currently VANTAGE: `http://www.vmsr.edu`) and the remaining 842 by HudsonAlpha Institute (`http://hudsonalpha.org`). From the 10,846 SNPs on the array, 8,218 were informative and produced reliable genotypes in the HS rats. From these final informative markers, the average SNP spacing was 284 Kb, with an average heterozygosity of 25.68%. Principle Component Analysis was used to confirm there were no systematic genotyping differences between the two centers (**Figure 1.8**).

5.2.4 RNA-Seq

RNA was extracted from liver of 398 HS rats using Trizol. Illumina kits were used to create library preps and RNA-Seq was run on the Illumina HiSeq 2500. RSEM and Bowtie were used to align reads and compute transcript level expression abundance (Detailed Methods).

5.3 Statistical Analysis

5.3.1 Estimating heritability of adiposity traits

Narrow-sense heritability was estimated for each transformed phenotype using a Bayesian linear mixed model (LMM) implemented in INLA (Holand et al., 2013; Rue et al., 2009). The LMM included fixed effects representing time food deprived, order of tissue harvest, and dissector (notably, dissector significantly affected EpiFat and BMI_Tail_Base), and a random "polygenic" effect, which represented the effect of overall relatedness (calculated as in (Gatti et al., 2014)). Heritability, h^2, was defined as the proportion of variance attributed to polygenic effects vs residual noise (Detailed Methods).

5.3.2 Genome-wide association

QTL were identified by genome-wide association of imputed allele dosages of genotyped SNPs. A hidden Markov model (Broman, 2016) was used to infer each HS rat's haplotype mosaic and thereby obtain robust estimates of the genotype of each SNP. Association tests were then performed, SNP-by-SNP, on each trait using a frequentist version of the LMM described for estimating heritability but with an added SNP effect term. Tests of the SNP effect yielded p-values that are here reported as negative log to the base 10, or "logP". Genome-wide significance thresholds for logP scores were estimated by parametric bootstrap samples from the fitted null (Solberg Woods et al., 2010; Valdar et al., 2009). Linkage Disequilibrium (LD) intervals for the detected QTL were defined by including neighboring markers that met a set level of LD, measured with the squared correlation coefficient r^2; we used $r^2 = 0.5$ to define intervals based on the plots of the SNP associations overlaid with LD information (Detailed Methods).

5.3.3 Fine-mapping and haplotype effect estimation at detected QTL

SNP variants within the LD interval were prioritized used the multi-SNP method LLARRMA-dawg (Sabourin et al., 2015), which calculates for each SNP a resample model inclusion probability (RMIP): SNPs with high RMIPs represent strong, independent signals, and the existence of multiple SNPs with a high RMIP is consistent with the presence of multiple independent signals. To char-

acterize each QTL signal, we used the Diploffect model (Zhang et al., 2014), which estimates the relative contributions of alternate founder haplotypes (Detailed Methods).

5.3.4 Candidate gene identification

Two parallel approaches were used: 1) bioinformatic analysis and protein modeling of known sequence variants; and, 2) mediation analysis of expression levels. For (1), we used HS founder sequence (`www.rgd.mcw.edu`; genome build Rn6) to identify highly conserved, non-synonymous coding variants within each QTL that were predicted to be damaging by Polyphen (`http://genetics.bwh.harvard.edu/pph/`) and/or SIFT, focusing on variants in founder strains that showed haplotype effects at the locus. Variants were confirmed using Sanger sequencing and then analyzed in the Sequence-to-Structure-to-Function analysis as previously described (Prokop et al., 2017). Briefly, proteins were assessed with codon selection analysis of multiple species open reading frames, inspected for linear motif impact near variants of interest, and modeled with I-TASSER (Roy et al., 2010) and YASARA (Krieger et al., 2009). Models were then assessed for likely impact on protein folding and/or function based on model confidence, phylogenetic sequence alignment, conservation, and whether or not the variant altered structural packing, molecular dynamic simulations, binding partners, linear motifs or post-translational modifications. For (2), transcript abundance levels of genes within HS liver were evaluated as potential causal mediators of the physiological QTL through mediation analysis (Baron and Kenny, 1986) (Detailed Methods).

5.4 Results

5.4.1 HS founder strains exhibit large variation in adiposity traits

All phenotypes were rank-inverse normal transformed except EpiFat, which instead was log transformed based on the Box-Cox procedure. All traits differed significantly between the inbred founder strains: body weight ($F_{5,76} = 15.492$, $p = 1.74\text{e}{-10}$), BMI_Tail_End ($F_{5,73} = 25.024$, $p = 1.34\text{e}{-14}$), BMI_Tail_Base ($F_{5,73} = 9.683$, $p = 4.02\text{e}{-7}$), EpiFat ($F_{5,78} = 69.541$, $p < 2.2\text{e}{-16}$) and RetroFat ($F_{5,78} = 38.157$, $p < 2.2\text{e}{-16}$; **Figure 1.1**). The BUF inbred strain had significantly more EpiFat mass (Tukey-Kramer $p < 0.05$) and BMI_Tail_End (Tukey-Kramer $p < 0.05$) relative to all other strains. BUF also had significantly more RetroFat mass compared to all strains (Tukey-

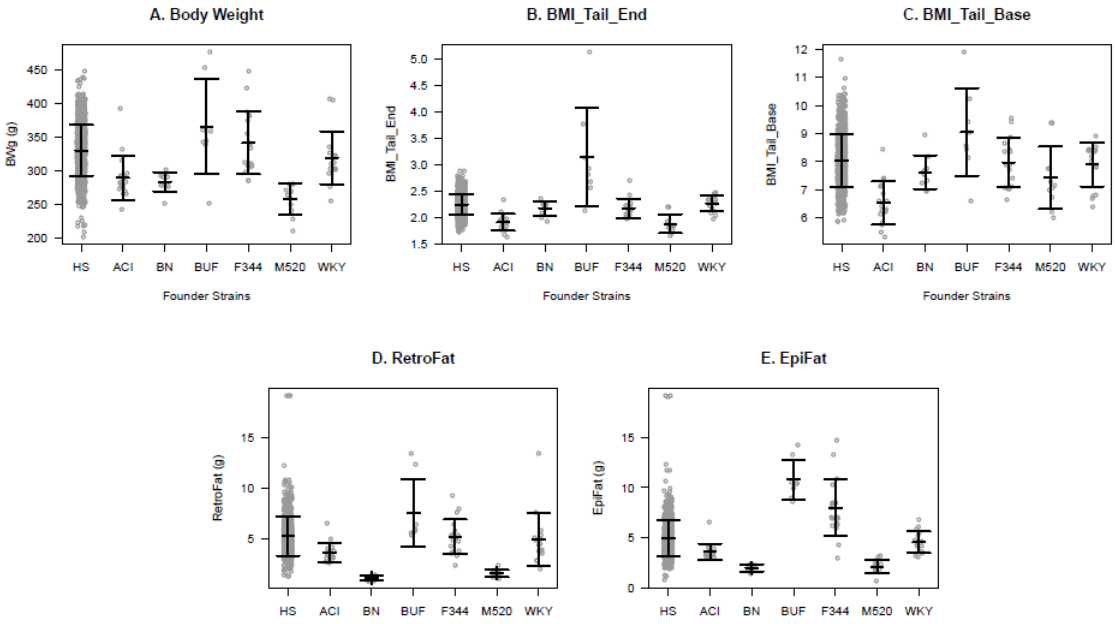

Figure 1.1: Adiposity traits in inbred founders and HS rats. Mean + SD are shown. BMI is body mass index from nose to end of tail (BMI_Tail_End) and from nose to base of the tail (BMI_Tail_Base). EpiFat and RetroFat are epididymal and retroperitoneal fat pad weight, respectively. Gray circles represent individual animals from 8-19 individuals from 6 of the founder strains, and the HS rats (989 in body weight; 741 in RetroFat, EpiFat, and BMI_Tail_End; and 740 in BMI_Tail_Base). See text for statistical differences between founder strains.

Kramer $p < 0.001$) except F344 (Tukey-Kramer $p = 0.06175$), higher body weight relative to ACI, BN, and M520 (Tukey-Kramer $p < 0.01$), and higher BMI_Tail_Base than ACI and M520 (Tukey-Kramer $p < 0.05$). ACI, BN and M520 were the lightest strains, with BN and M520 showing significantly lighter EpiFat (Tukey-Kramer $p < 1e{-}5$) and RetroFat (Tukey-Kramer $p < 0.001$) relative to other strains.

5.4.2 Adiposity traits are highly correlated with measures of metabolic health in HS rats

Variation between the founder strains is represented within the HS colony (**Figure 1.1**). Adiposity measures were highly correlated with several measures of metabolic health (**Table 5.1**, **Figure 1.2**). EpiFat significantly correlated with every measure of metabolic health and RetroFat correlated with all but fasting glucose. Body weight significantly correlated with all measures except fasting triglycerides. BMI_Tail_End significantly correlated with fasting total cholesterol, fasting tricglyc-

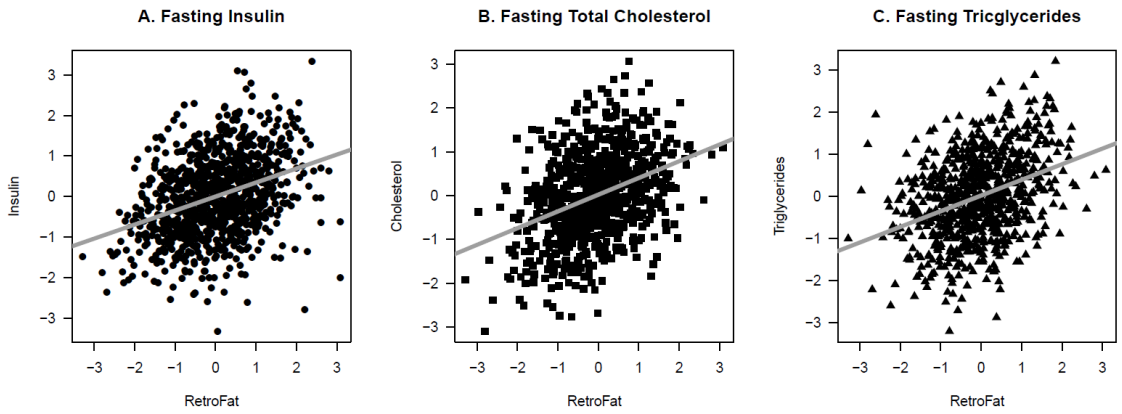

Figure 1.2: Significant correlations between RetroFat (retroperitoneal fat pad weight) and A) fasting insulin ($p = 4.75\text{e}{-27}$), B) fasting total cholesterol ($p = 1.02\text{e}{-20}$) and C) fasting triglycerides ($p = 2.55\text{e}{-20}$) in HS rats. Plots show the residuals of rank-inverse normal transformed phenotypes with nuisance factors regressed out to restrict correlation estimates to that between RetroFat and these metabolic traits. Significant correlations were also found between RetroFat and several other measures of metabolic health (see **Table 5.1**).

erides, glucose AUC, insulin AUC, and IGI, whereas BMI_Tail_Base did not significantly correlate with any of the measures of metabolic health.

5.4.3 Adiposity traits are highly heritable

Adiposity traits were highly heritable in HS rats: body weight (posterior mode of $h^2 = 75.3\%$; 95% highest posterior density interval$= 67.0 - 81.7\%$), EpiFat (54.1%; $40.1 - 66.0\%$), RetroFat (53.9%; $39.7 - 66.7\%$), BMI_Tail_End (45.0%; $32.3 - 57\%$) and BMI_Tail_Base (25.4%; $13.6 - 41.8\%$).

5.4.4 RetroFat QTL on chromsomes 1 and 6

Two 90% significant QTL were identified for RetroFat, a QTL on rat chromosome 6: 22.79-28.93 Mb (6.14 Mb, logP $= 4.73$) and a QTL on chromosome 1: 280.63 - 281.82 Mb (1.19 Mb, logP $= 4.69$; **Figures 1.3ABC**, **1.6ABC**). The LLARRMA-dawg multi-SNP fine-mapping analysis narrowed the most likely region of the broader chromosome 6 QTL to 1.46 Mb region (27.17 - 28.63 Mb; **Figure 1.9**) narrowing the number of the genes from 130 to 30 (**Tables 5.2-5.5**, **Figure 1.10**). Estimating founder haplotype effects at the chromosome 6 QTL gave an effect size (posterior median) of 11.05% and showed that at this locus, decreased fat pad weight is associated with the

	Body Weight	BMI_Tail_End	BMI_Tail_Base	EpiFat	RetroFat
Fasting Glucose	**0.1453**	0.0952	0.0886	**0.1931**	0.1132
	(0.0012)	(0.76)	(1)	**(3.24e-05)**	(0.1009)
Fasting Insulin	**0.1936**	0.1153	-0.0248	**0.4314**	**0.3516**
	(3.48-07)	(0.091)	(1)	**(1.35e-31)**	**(4.75e-27)**
Fasting Total Cholesterol	**0.2644**	**0.2428**	0.1172	**0.2535**	**0.3529**
	(7.10e-11)	**(5.75e-09)**	(0. 39)	**(6.85e-10)**	**(1.02e-20)**
Fasting Triglycerides	0.1291	**0.2426**	0.0950	**0.3096**	**0.3499**
	(0.12)	**(6.07e-09)**	(1)	**(1.75e-15)**	**(2.55e-20)**
Glucose_AUC	**0.1378**	**0.1259**	0.0652	**0.1663**	**0.1718**
	(0.0040)	**(0.0214)**	(1)	**(0.0016)**	**(1.71e-05)**
Insulin_AUC	**0.2937**	**0.2238**	0.0312	**0.4670**	**0.4397**
	(5.67e-18)	**(7.99e-10)**	(1)	**(2.71e-37)**	**(8.79e-44)**
QUICKI	**-0.2042**	-0.1198	0.0211	**-0.4338**	**-0.3544**
	(3.91e-08)	(0.0523)	(1)	**(5.24e-32)**	**(1.57e-27)**
IGI	**0.1629**	**0.1220**	0.0001	**0.2475**	**0.2038**
	(0.0001)	**(0.0430)**	(1)	**(5.46e-09)**	**(5.46e-08)**

Spearman's rank correlation with Bonferonni-adjusted p-values in parentheses (bold if <0.05). To mitigate potentially confounding effects of experimental covariates, correlations are performed on phenotypic residuals (ie, on the residuals after regressing out covariate effects from the rank-inverse normal transformed phenotype).

Table 5.1: Correlations between adiposity and measures of metabolic health in HS rats

WKY haplotype (**Figure 1.3D**). For the chromosome 1 QTL the effect size was 13.33%, with increased fat pad weight associated with BUF, MR and WKY haplotypes (**Figure 1.6D**).

5.4.5 Identification of *Adcy3*, *Krtcap3*, *Slc30a3* within the chromosome 6 RetroFat QTL

Within the chromosome 6 RetroFat QTL, bioinformatic analysis revealed only one gene, *Adcy3*, that had a highly conserved, potentially damaging, non-synonymous variant in the WKY rat, the founder haplotype associated with decreased Retrofat at this locus. *Adcy3* also falls within the fine-mapped support interval of the QTL (**Figure 1.3C**). The WKY founder strain harbors a C at position 28,572,363 bp within *Adcy3* while all other strains harbor a T, resulting in a leucine-to-proline substitution at amino acid 121. Based on DNA information from 86 nucleotide sequences for ADCY3, this variant is highly conserved with evidence for selective pressure. Protein modeling indicated that amino acid 121 is located within the first transmembrane region, with a proline likely causing a bend in the helix and thus altered transmembrane packing (**Figure 1.3E**).

Because fine-mapping supported multiple independent signals at the QTL, we investigated potential mediators among the cis-expressed genes. Mediation analysis identified six potential mediators, all driven by the WKY haplotype (**Figure 1.4**; **Tables 5.9** and **5.10**), suggesting that multiple genes may influence RetroFat at this locus. Two in particular were strongly supported: *Krtcap3* as a full mediator and *Slc30a3* as a partial mediator, remaining significant after controlling for *Krtcap3*. Under the proposed model (**Figure 1.5**), the WKY haplotype increases expression of *Krtcap3*, which is itself negatively correlated with RetroFat (**Figure 1.11**), and thus the causal path is consistent with the negative WKY effect on RetroFat at the locus; meanwhile, WKY decreases expression of *Slc30a3*, which is also negatively correlated RetroFat, suggesting *Slc30a3* is a suppressor of the QTL/*Krtcap3* effect.

5.4.6 Identification of *Prlhr* within the chromosome 1 RetroFat QTL

Within the chromosome 1 RetroFat QTL, there are 15 genes, ten of which are uncharacterized (**Figure 1.6BC**, **Table 5.6 - 5.7**). One gene, *Prlhr*, contains a non-synonymous variant in the BUF and WKY founder strains, two founder haplotypes associated with an increase fat pad weight at this locus. The variant falls within the start codon, changing methionine to isoleucine. The next methionine falls

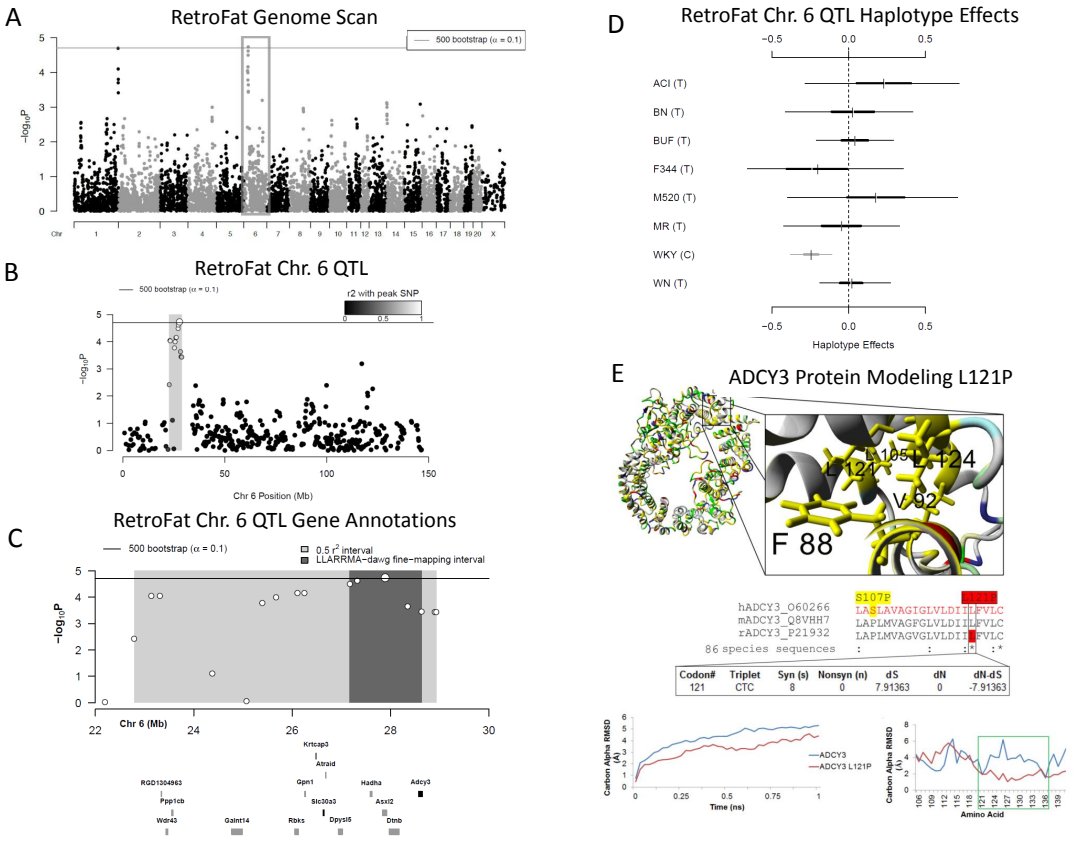

Figure 1.3: Genome scan of RetroFat (A). X-axis is position on chromosome and y-axis is the logP level of association. Genome-wide significance thresholds were calculated using parametric bootstraps from the null model ($\alpha = 0.1$, logP $= 4.70$). The grey region highlights the 6.14 LD support interval of the chromosome 6 QTL showing neighboring markers that are correlated with the peak marker, representing genomic regions likely to contain the causal variant underlying the statistical signal (B). Annotation of genes that fall within the support interval (C). The entire 6.14 region is shaded in grey, with the fine-mapped 1.46 Mb region shaded in dark grey. Only genes that have a cis eQTL are shown. All 130 genes within the region are listed in **Tables 5.2-5.5**. Additive haplotype effects were estimated using the Diploffect model, which takes into account uncertainty in haplotype state (D). SNP allele information is overlaid on the haplotype effects, and are distinguished by black or gray. The WKY haplotype, the only haplotype with the C allele at the chromosome 6 locus, has a significantly negative effect on phenotype. Protein modeling for ADCY3 (E). Variant L121P of ADCY3 is found with the conserved hydrophobic core of the transmembrane helices. A zoomed in view is shown to the right. The middle panel shows sequence alignments of amino acids. ADCY3 amino acid 121 is also 100% conserved (red) as a leucine in 86 analyzed vertebrate species. A human SNP is known at amino acid 107 (yellow). Using the DNA information from the 86 nucleotide sequences for ADCY3, there is also evidence of selective pressure in the DNA sequence to conserve the amino acid. Bottom panel shows molecular dynamic simulations for ADCY3. Simulations performed on the protein dimer for wild type (WT blue) or the mutant (ADCY3 L121P, red) suggests that the models' average movement over time is altered. Altered movement is seen in the simulations for ADCY3 with fluctuation of amino acids found near amino acid 121 when mutated.

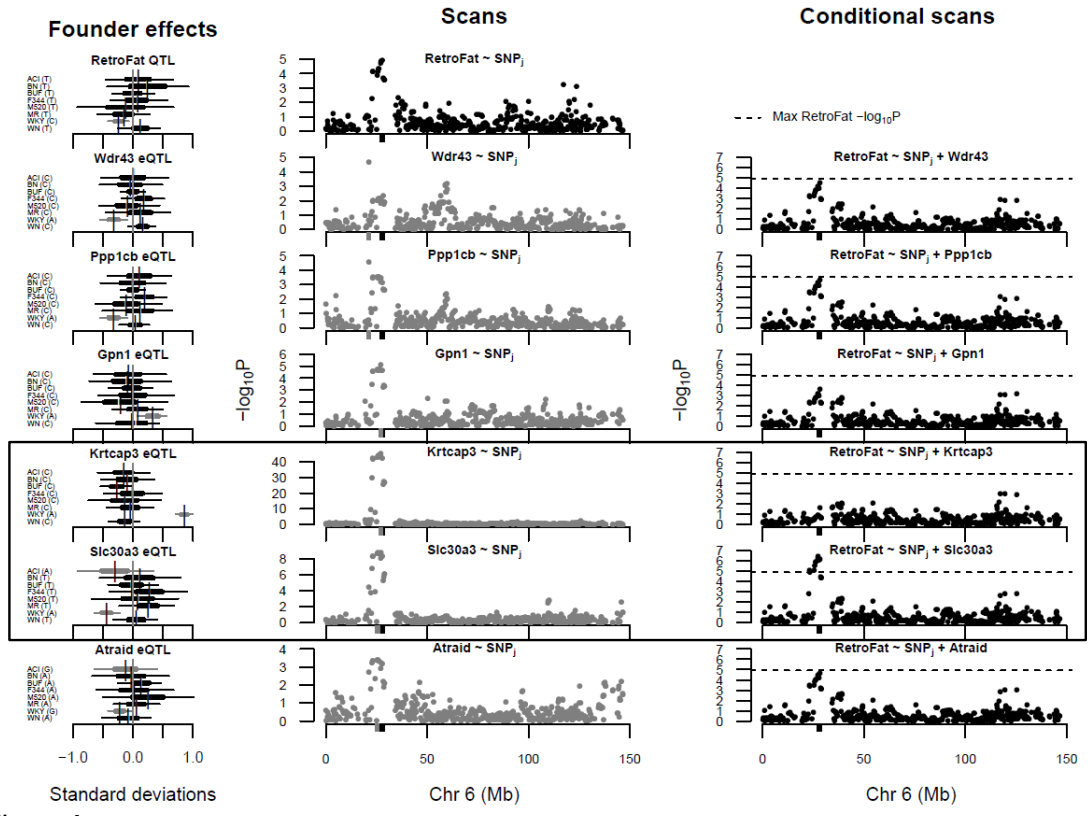

Figure 1.4: Mediation analysis identified the expression levels of six genes (*Wdr43*, *Ppp1cb*, *Gpn1*, *Krtcap3*, *Slc30a3*, and *Atraid*; **Table 5.10**) in the RetroFat chromosome 6 QTL interval as potential mediators of the QTL effect on the phenotype. [Middle column] Comparisons of the RetroFat chromosome 6 association scan with association scans for the potential mediators reveals them to likely have co-localizing cis eQTL with the RetroFat QTL. [Left column] The haplotype effects on RetroFat at the QTL and on the mediators at the eQTL reveals that in this region, the WKY haplotype is largely driving the differences in RetroFat and mediator gene expression, suggesting a possible connection between RetroFat and local gene expression. [Right column] RetroFat chromosome 6 association scans, conditioned on candidate gene expression, is consistent with the mediation analysis finding that *Krtcap3* is a strong candidate as full mediator of the effect of QTL on RetroFat. When *Krtcap3* expression is included in the model, the QTL is largely removed. *Slc30a3*, as a potential suppressor of the QTL effect on RetroFat, actually increases the significance seen at the QTL.

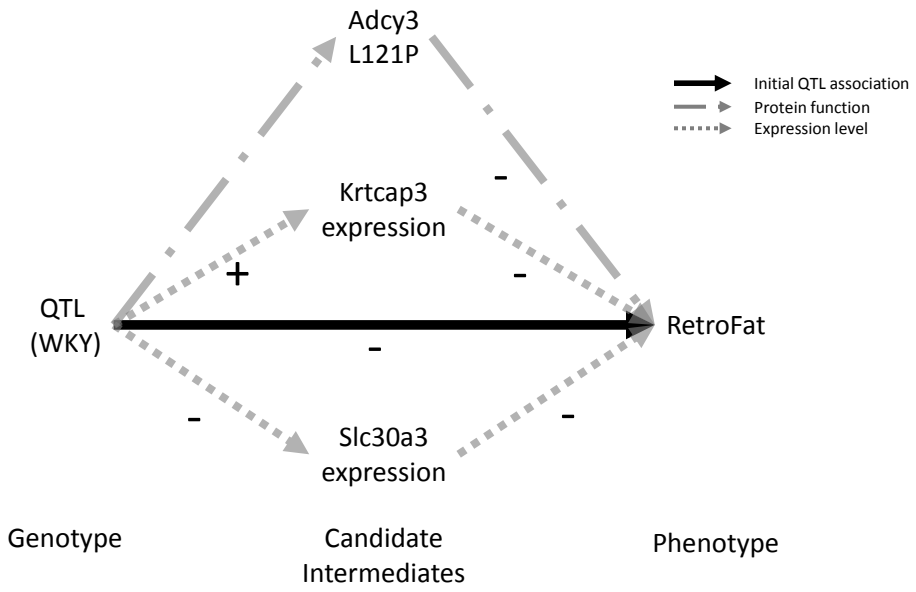

Figure 1.5: Model demonstrating role of *Adcy3*, *Krtcap3* and *Slc30a3* on RetroFat. WKY haplotype increases expression of *Krtcap3*, which is itself negatively correlated with RetroFat (**Figure 1.11**), and thus the causal path is consistent with the negative WKY effect on RetroFat at the locus. In contrast, WKY decreases expression of *Slc30a3*, which is also negatively correlated RetroFat, suggesting *Slc30a3* is a suppressor of the QTL/*Krtcap3* effect. Finally, the non-synonymous variant with *Adcy3* causes amino acid change L121P leading to lower RetroFat.

at amino acid 65, such that the conserved N-terminal region and half of transmembrane helix 1 would be deleted with the variant (**Figure 1.6E**, `https://youtu.be/vRTIkITXRbw`). A molecular dynamic simulation of the PRLHR protein with and without the first 64 amino acids showed strong changes to the entire GPCR transmembrane region. *Prlhr* is expressed mainly in adrenal and brain such that expression levels could not be determined in liver tissue. None of the liver-expressed genes local to the QTL map as cis-eQTL, thus *Prlhr* remains the strongest candidate within this region.

5.5 Body weight QTL on chromosome 4 and identification of *Grid2*

A 95% significant QTL for body weight was also detected on rat chromosome 4: 91.35Mb to 94.7Mb (3.35 Mb, logP $= 5.32$) (**Figure 1.7ABC**). At this locus, whose effect size was 12.33%, decreases in body weight were associated with ACI, BUF, F344 and MR haplotypes, and increases with BN (**Figure 1.7D**).

Within the body weight QTL, there are only 11 genes, nine of which are pseudogenes or uncharacterized LOC proteins, leaving only *Ccser1* and *Grid2* (**Figure 1.7C**, **Table 5.8**). None of the genes at this locus contained highly conserved potentially damaging non-synonymous variants. Both *Ccser1* and *LOC108350839* are expressed in liver, but the expression of neither was significantly associated with the body weight QTL, ruling these out as candidate mediators. The brain-specific *Grid2* is the only gene that has previously been linked to body weight (Nikpay et al., 2012) and *Grid1* was recently associated with BMI in human GWAS (Locke et al., 2015), implicating *Grid2* as the most likely candidate at this locus.

5.6 Discussion

This is the first study to map adiposity traits genome-wide using HS rats and demonstrates their utility for uncovering genes and variants likely to impact human adiposity. We identified QTL for RetroFat on rat chromosomes 1 and 6 and a QTL for body weight on chromosome 4. Using various fine-mapping procedures, we identified three likely candidate genes within the chromosome 6 RetroFat locus: a protein-coding variant within *Adcy3*, and transcriptional regulation of *Krtcap3* and *Slc30a3* that mediate between the QTL and RetroFat. Within the chromosome 1 RetroFat QTL, we identified a variant within *Prlhr* that increases fat pad weight. Lastly, *Grid2* was identified as

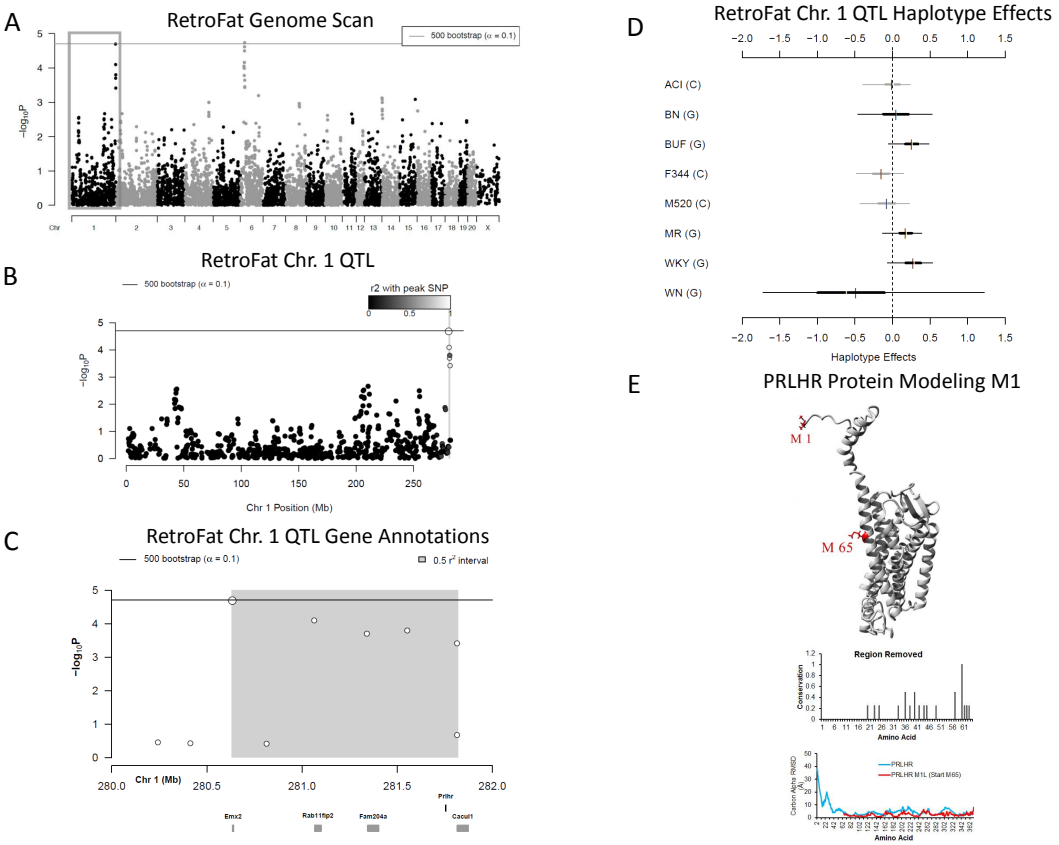

Figure 1.6: Genome scan of RetroFat as described in **Figure 1.3A** (A). The grey region highlights the 1.19 Mb LD support interval for the chromosome 1 locus representing neighboring markers that are correlated with the peak marker, representing genomic regions likely to contain the causal variant underlying the statistical signal (B). Annotation of the five characterized genes that fall within the support interval (C). Additive founder haplotype effects for the chromosome 1 RetroFat locus (D). Additive haplotype effects were estimated using the Diploffect model, which takes into account uncertainty in haplotype state. SNP allele information is also overlaid on the haplotype effects. The C allele is shared by ACI, F344, and M520, that possesses a variant with a negative effect on RetroFat, whereas BUF, MR and WKY haplotypes result in increased RetroFat at this locus. Protein modeling for PRLHR (E). Variant M1I of PRLHR is found within the methionine start site. The next start site is at position 65 leading to removal of the conserved N-terminal region and half of transmembrane helix 1. 16 amino acids removed are under selective pressure (middle panel) and the deletion of the first 64 amino acids causes a destabilization of the entire proteins dynamics as seen by the molecular dynamic simulations (bottom panel).

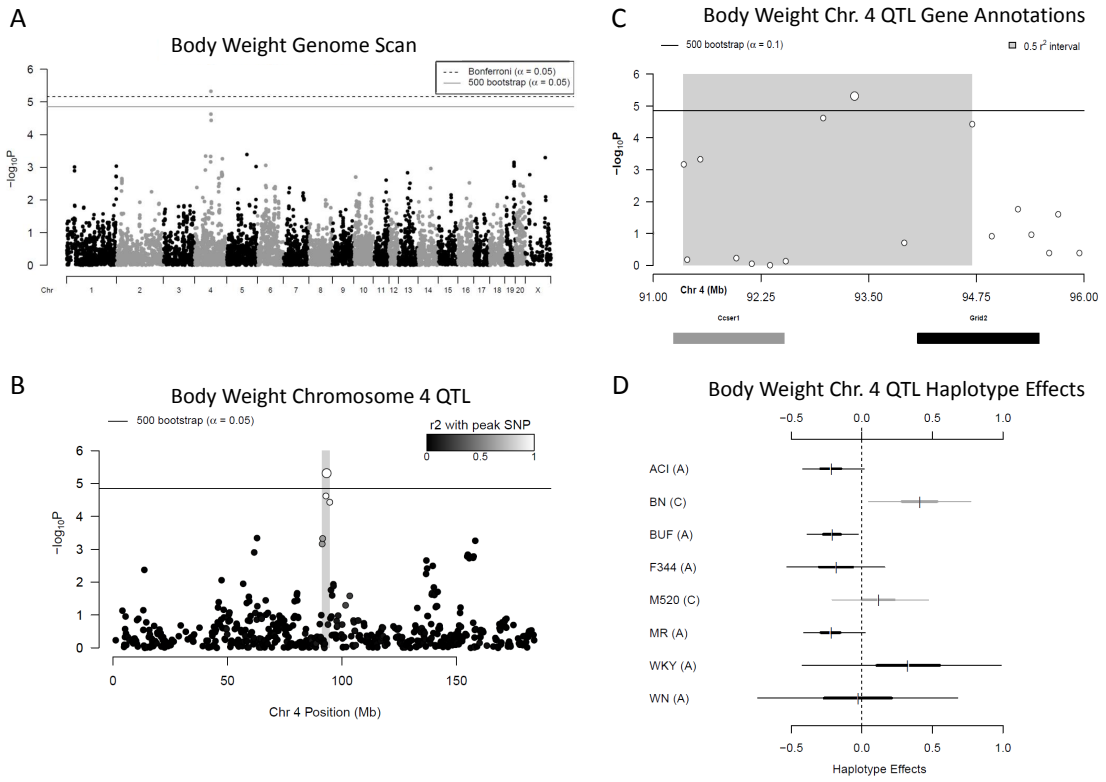

Figure 1.7: Genome scan of body weight (A). X-axis is position on chromosome and y-axis is the logP level of association. Genome-wide significance thresholds were calculated using parametric bootstraps from the null model (significant: $\alpha = 0.05$, logP $= 4.86$) and conservative $\alpha = 0.05$ Bonferroni thresholds (logP $= 5.16$). Linkage disequilibrium support interval in grey is 3.35 Mb (B). Annotation of the two characterized genes that fall within the support interval (C). Additive haplotype effects for chromosome 4 body weight QTL (D). The C allele at the marker could represent shared haplotype descent between BN and M520, both which have an increasing effect on body weight at this locus. ACI, BUF, F344 and MR haplotypes have a decreasing effect of body weight at this locus, all of which share the A allele. The WKY and WN also have an A allele and the WKY haplotype has an increasing effect on body weight, while the WN haplotype appears not to effect body weight, although the credible interval of both is fairly large and not well represented in the data at this locus.

the most likely candidate gene within the body weight locus. It is of interest that several of these candidate genes play a role in neural regulation of energy metabolism and/or feeding behavior.

As expected, both the HS founders and the HS population varied for adiposity traits, with BUF showing the highest adiposity and ACI, BN and M520 showing the lowest. As seen in humans (Stunkard et al., 1986; Katzmarzyk et al., 2000), these traits were highly heritable. Also similar to humans (Emdin et al., 2017), increased body weight, particularly visceral fat pad weight (RetroFat and EpiFat), was significantly associated with several measures of metabolic health in the HS rats, indicating that genes underlying QTL for adiposity traits are likely to contribute to overall metabolic health in HS rats.

Despite high heritability for adiposity traits in this model, we found only three QTL for the five traits that were studied. The remaining heritability could be attributable to loci of small effect and/or complex genetic architecture that lie below the limit of detection in this study, and this accords with the fact that the QTL we identified were each of relatively large effect (12.33%, 13.33% and 11.05% respectively). The identified QTL also had relatively small LD support intervals (3.35 Mb, 1.19 Mb and 6.14 Mb, respectively), significantly decreasing the number of potential candidate genes within each QTL relative to traditional QTL studies using F2 intercross or backcross animals; the map density, however, was too low for high resolution mapping of the genes within the interval. We expect that increasing both the number of animals used as well as the density of genotyping would serve to uncover additional loci.

The chromosome 6 RetroFat locus encompassed 6.14 Mb, contained 130 genes and was driven by the WKY haplotype. Using a fine-mapping procedure that allowed for the presence of multiple signals, we identified a 1.46 Mb plausible region for the QTL. *Adcy3* was the only gene in this region to contain a highly conserved, non-synonymous variant in the WKY founder strain that is predicted to be damaging: the leucine-to-proline switch at amino acid 121 would likely induce a bend in the helix leading to altered membrane interactions and binding. In addition, we found that multiple genes within the locus map as eQTL. Subsequent mediation analysis supported roles for *Krtcap3* and *Slc30a3*, with *Krtcap3* expression presenting as a full mediator of the QTL and *Slc30a3* expression as a partial/suppressor mediator. Although little is known about *Krtcap3*, *Slc30a3* is a zinc transporter that plays a role in glucose transport and metabolism (Smidt et al., 2009), and *Adcy3* is an enzyme that catalyzes the cAMP second messenger system and is likely involved in energy homeostasis (Wu

et al., 2016). The POMC/RBJ/ADCY3 region has previously been identified in multiple human GWAS for BMI and obesity (Speliotes et al., 2010; Nordman et al., 2008; Stergiakouli et al., 2014; Wen et al., 2012). Interestingly, a non-synonymous amino acid change (Ser107Pro) in the human *Adcy3* gene (Speliotes et al., 2010), which falls within the same transmembrane helix as the rat variant, has been identified as the causal variant in height-adjusted childhood BMI (Stergiakouli et al., 2014), indicating the same likely causal variant between rat and human. *Adcy3* knock-out and haplo-insufficient mice become obese with age, exhibiting increased food intake and decreased locomotion (Tong et al., 2016; Wang et al., 2011). In addition, gain of function in *Adcy3* protects against diet-induced obesity (Pitman et al., 2014), further supporting a causal role for this gene.

The chromosome 1 RetroFat locus encompassed 1.19 Mb, contained 15 genes, with BUF, MR, and WKY haplotypes increasing RetroFat. *Prlhr*, containing a non-synonymous variant in both the BUF and WKY strains, stood out as the most likely candidate gene: the variant fell within the methionine start site and leads to removal of the conserved N-terminal region and half of transmembrane helix 1, likely having a large impact on protein function. This variant is found in several other rat strains including FHH, GK, LEW and SD. *Prlhr* is known to play a role in feeding behavior, with ICV administration in the hypothalamus leading to decreased food intake (Lawrence et al., 2000), and *Prlhr* knock-out mice exhibiting increased food intake, body weight and fat pad weight (Gu et al., 2004). Interestingly, this specific variant did not alter feeding behavior in outbred Sprague-Dawley rats (Ellacott et al., 2005), indicating that the effect of the variant on fat pad weight may be independent of food intake, although additional studies are needed to confirm this.

The body weight locus encompassed 3.35 Mb and contained 11 genes, none of which contained highly conserved non-synonymous variants predicted to be damaging between the two haplotype effect groups: ACI, BUR, F344, MR versus BN. Only one gene in the region, *Grid2*, has previously been linked to obesity, jointly with tobacco use, in a family-based study (Nikpay et al., 2012), making it the most likely candidate gene. Interestingly, *Grid1* was associated with BMI in a recent human GWAS (Locke et al., 2015), further supporting a potentially causal role for *Grid2* within the rat body weight locus. *Grid2* encodes the glutamate ionotropic receptor delta type subunit 2 and is known to play a role in synapse formation, particularly within the cerebellum (Hirai et al., 2003). Synaptic formation and plasticity are increasingly being recognized as playing a role in metabolism and energy

balance (Dietrich and Horvath, 2013). Additional work, including assessing *Grid2* expression levels in brain, is needed to confirm or eliminate *Grid2* as the causal gene at this locus.

In summary, we have used HS rats to identify QTL for adiposity traits, leading to identification of five candidate genes and two likely causal variants. Some genes have previously been identified in human GWAS or linkage studies (*Adcy3*, *Grid2*) or implicated in rodent models of obesity (*Adcy3*, *Prlhr*), while two genes are novel (*Krtcap3*, *Slc30a3*). The *Adcy3* variant falls within the same transmembrane helix as that found in humans indicating direct human relevance of this work. It is also of interest that *Adcy3*, *Prlhr* and *Grid2* have previously been found to impact feeding behavior and/or neural regulation of metabolism. This work demonstrates the power of HS rats for genetic fine-mapping and identification of underlying candidate genes and variants that will likely be relevant to human adiposity.

5.7 Detailed Methods

5.7.1 Animals

5.7.1.1 Housing

Rats were housed two per cage in micro-isolation cages in a conventional facility using autoclaved bedding (sani-chips from PJ Murphy). They had *ad libitum* access to autoclaved Teklad 5010 diet (Harlan Laboratories) and were provided reverse osmosis water chlorinated to 2-3 ppm.

5.7.2 Statistical genetic analysis

5.7.2.1 Modeling genetic effects on adiposity

All statistical genetic analyses described used the same general model (or approximations to it) for linking the genetics of a given rat to its measured phenotypic outcome. This was the linear mixed effect model (LMM)

$$f(y_i) = \text{covariates}_i + \text{QTL}_i(m) + u_i + \text{residual}_i, \tag{5.1}$$

where, in brief: $f(y_i)$ is the phenotype subject to a normalizing transformation, specifically, as a conservative measure to rein in high influence data points, we used the rank inverse normal

transformation; covariates$_i$ is a fixed effects term that includes variables representing time food deprived, order of tissue harvest, and dissector (notably, dissector significantly affected EpiFat and BMI_Tail_Base); QTL$_i(m)$ represents the effect of the quantitative trait locus (QTL) at genomic locus m, and is defined in more detail below; and residual$_i$ models the remaining individual-to-individual variation as a normal deviate with variance σ^2. The u_i term is a random polygenic effect representing the effect of overall genetic relatedness, modeled as vector $\mathbf{u} = (u_i, \ldots, u_n)$ drawn from a multivariate normal with covariance matrix $\mathbf{G}\tau^2$, where τ^2 is unknown and $\mathbf{G}$ is the realized genetic relationship matrix, estimated as the pairwise distance in allelic dosages defined by the identity by descent (IBD) probabilities from founder haplotypes, standardized by allele frequency and averaged over loci across the genome, calculated using the `kinship.probs` function in the DOQTL R package (Gatti et al., 2014). The LMM in Eq 5.1 with QTL$_i(m)$ omitted is hereafter referred to as "the null model".

5.7.2.2 Heritability estimation

Narrow-sense heritability,

$$h^2 = \frac{\tau^2}{\tau^2 + \sigma^2} \times 100\% \ ,$$

was estimated for each phenotype by fitting the null model as a Bayesian LMM using INLA (Rue et al., 2009; Holand et al., 2013), which gives a complete posterior distribution of h^2, along with point and interval estimates. Phenotypes were scaled to have a mean of 0 and standard deviation of 1, and a uniform prior on h^2 was obtained by setting priors on τ^{-2} and σ^{-2} to Ga$(1, 1)$, with other settings being default.

5.7.2.3 QTL mapping

QTL were identified by genome-wide association of imputed SNPs. This was performed in three steps. First, as in previous work (Solberg Woods et al., 2012), we obtained a probabilistic reconstruction of each rat's haplotype mosaic, that is, the configuration of inherited founder haplotypes that compose its genome, using a hidden Markov model (HMM), implemented in R/qtl2geno (Broman, 2016), applied to the genotype data on HS rats and their founders. This HMM was used to calculate for each individual $i = 1, \ldots, n$, at each marker position $m = 1, \ldots, 8218$, a vector

of 36 descent probabilities, $\mathbf{p}_{im}$, containing the posterior probability of descent from each of the possible $\frac{8(8+1)}{2} = 36$ haplotype pairs (diplotypes). n, the sample size, varies between phenotypes, with $n = 989$ for those irrespective of tissue harvest age, such as body weight, and $n = 743$ for those that include only individuals with tissue harvested at 17 weeks of age, such as RetroFat (two rats did not have RetroFat measurements, resulting in $n = 741$). Second, these descent probabilities were used to re-estimate the original SNP genotypes, that is, each $\mathbf{p}_{ij}$ was used to infer a 3-vector of imputed genotype probabilities $\mathbf{g}_{ij}$; these imputed genotypes, which, unlike their raw counterparts, were both complete and relatively robust to genotyping error, were carried forward into subsequent analyses. Third, at each SNP, we fitted the LMM in Eq 5.1, setting $\text{QTL}_i(m) = \beta x_{mi}$ where x_{mi} is the expectation of the minor allele count (ie, the allele dosage) implied by $\mathbf{g}_{im}$, and β is a fixed effect; comparing the maximum likelihood (ML) fit of this model to that of the null model gave a likelihood ratio test and nominal p-value, reported as its negative base 10 logarithm, or logP. (Note that initially we used models testing the association between phenotype and haplotype descent, ie, $\mathbf{p}_{im}$, directly, as in the region-wide mapping of (Solberg Woods et al., 2012), but instead used the less complicated SNP modeling due to a combination of uncertainty in haplotype descent and strong imbalances in the estimated haplotype frequencies.)

Genome-wide significance thresholds for logP scores were estimated by parametric bootstrap samples from the fitted null (Valdar et al., 2009; Solberg Woods et al., 2010), with Bonferroni thresholds, which would be highly conservative due to the serial LD structure, calculated for comparison.

LD intervals for the detected QTL were defined by including neighboring markers that met a set level of LD, measured with the squared correlation coefficient r^2; we used $r^2 = 0.5$ to define intervals based on the plots of the SNP associations overlaid with LD information.

5.7.2.4 Fine-mapping through Group-LASSO with fractional resample model averaging

To prioritize SNP variants within the RetroFat chromosome 6 QTL interval, we used the multi-SNP modeling method LLARRMA-dawg (Sabourin et al., 2015), which we applied to the imputed SNP genotypes and a population structure-corrected version of the phenotype, namely the phenotypic residuals of the null model. LLARRMA-dawg uses a combination of variable selection

and resampling to identify SNPs that have stable, independent associations with the phenotype. Each SNP receives a resample model inclusion probability (RMIP), an estimate of the probability it would be included in a parsimonious multi-SNP model applied to a resampling of the individuals. SNPs with high RMIPs thus represent stronger candidates, and the existence of multiple SNPs with a high RMIP is consistent with the presence of multiple independent signals.

5.7.2.5 Estimating haplotype substitution effects at detected QTL

For detected QTL, the effect of substituting alternate haplotypes and diplotypes was estimated using the Diploffect model (Zhang et al., 2014), which can help identify interesting alleles of the candidate variants near the mapping signal. Although stability and power, along with the computational demands of a genome-wide analysis, led us to use SNP association for genetic mapping, these were no longer constraints for haplotype effect estimation at an identified QTL. Diploffect is a Bayesian hierarchical approach designed to work with probabilistically inferred haplotype descent, providing shrinkage that mitigates instability from low haplotype frequencies. In addition to the population structure effect in Eq 5.1, it models two genetic components at the QTL: additive (haplotype) effects, ie the effect of each dose of haplotype (eg WKY); and dominance deviations, those from the additive model for specific combinations of haplotype, (eg, WKY-ACI). Dominance deviations are typically less informed, but their inclusion stabilizes additive effect estimation. Both have their own variance parameters, τ^2_{add} and τ^2_{dom}, with QTL effect size recorded as the intraclass correlation coefficient

$$\rho_{\text{QTL}} = \frac{\tau^2_{\text{QTL}}}{\tau^2_{\text{QTL}} + \tau^2 + \sigma^2} \ ,$$

where $\tau^2_{\text{QTL}} = \tau^2_{\text{add}} + \tau^2_{\text{dom}}$. The model was fitted using 200 importance samples from INLA (Rue et al., 2009; Holand et al., 2013), with phenotype transformations and variance component priors set as for heritability estimation above.

5.7.2.6 Analysis of RNA-Seq data

Total RNA was extracted from the livers of 398 of the HS rats using Trizol, followed by library preparation using Illumina TruSeq Stranded mRNA library kit and sequencing on an Illu-

mina HiSeq2500 (Illumina, Inc., San Diego, CA). BN reference genome sequence (genome build Rn6) and GTF files were obtained from Ensembl. RSEM (v1.3.0) `rsem-prepare-reference` function was used to extract the transcript sequences from the genome (Li and Dewey, 2011) and to build Bowtie2 indices (Bowtie2 v2.2.8) (Langmead and Salzberg, 2012). RSEM `rsem-calculate-expression` function was then used to execute Bowtie2 to align reads of each sample to the transcriptome prepared above and to compute transcript level and gene level expression abundance. Trim Galore (`http://www.bioinformatics.babraham.ac.uk/projects/trim\_galore/`) was used to perform quality-based trimming with a cutoff at Q=20. Seven animals were removed due to low number of input reads.

5.7.3 Mediation analysis of phenotype, expression, and QTL

Mediation analysis was used to identify genes with expression levels that mediate the relationship between QTL and physiological phenotype. Expression levels of genes contained within the LD-based QTL intervals were assessed as potential candidates as full mediators (intermediates that completely explain the association between SNP and phenotype) and partial mediators (intermediates that explain some of the association between SNP and phenotype). Similar to (Baron and Kenny, 1986) and adapted for genetic data as in (Battle et al., 2014), evidence of mediation was assessed by a series of association tests, presented as a series of steps below, evaluating the relationships between previously mapped phenotype QTL (X), some transformation of the expression of level of a candidate mediator gene $j = 1, \ldots, J$ (M), and some transformation of the phenotype (Y).

1. **Potential mediators:** The relationship, represented as an arrow, with directionality encoding causality, $X \rightarrow M$ is evaluated for all J candidate genes in the physiological QTL interval with non-zero expression in greater than 0.25 of the n rats by testing for the association between QTL and expression of gene j via the regression model

$$f(\text{gene.expression}_{ij}) = \text{mapped.QTL}_i + u_i + \text{residual}_i, \tag{5.2}$$

where briefly $f(\text{gene.expression}_{ij})$ is the expression level for gene j of rat i subject to some normalizing transformation, often a rank inverse normal transformation, mapped.QTL_i is the effect of the mapped QTL for rat i, and u_i and residual_i are respectively the polygenic and

individual error terms as described in Eq 5.1. The maximum likelihood fit of the model in Eq 5.2 is compared with the null model (same as Eq 5.2 with mapped.QTL$_i$ omitted) to produce a likelihood ratio statistic and corresponding p-value. The p-values are converted to q-values using the Benjamini-Hocheberg false discovery rate (FDR) method (Benjamini et al., 1995). $X \rightarrow M$ for gene j is considered satisfied if q-value$_j < 0.1$. A lenient FDR controlling approach to multiple testing is used because the candidate set of genes is constrained to those local to the QTL interval, as well as the mediation analysis including further tests to satisfy mediator status. The set K ($K \leq J$) genes represent candidate mediators, and are also likely co-localizing eQTL to the QTL.

2. **Full mediators:** The relationship $X \perp\!\!\!\perp Y|M$ is representative of M being a full mediator of X on Y, suggesting that $X \rightarrow M \rightarrow Y$, specifically that X does not affect Y outside of through M. The support for this relationship in the data is evaluated by comparing the following regression models:

$$f(y_i) = \text{mapped.QTL}_i + f(\text{gene.expression}_{ij}) + u_i + \text{residual}_i, \tag{5.3}$$

and

$$f(y_i) = f(\text{gene.expression}_{ij}) + u_i + \text{residual}_i, \tag{5.4}$$

where Eq 5.3 is the alternative model and Eq 5.4 is the null model for a likelihood ratio test. The expression level of gene k is called a full mediator if p-value$_k > 0.05$, representing the situation in which the effect of QTL on the phenotype is fully explained by expression of gene k. After testing for all K candidate mediators, S ($0 \leq S \leq K$) full mediators are called.

3. **Partial mediators:** The relationship $M \rightarrow Y|X$ is representative of M being a partial mediator of X onto Y. To test the support for this relationship, Eq 5.3 for each candidate partial mediator t ($T = K - S$) is compared to

$$f(y_i) = \text{mapped.QTL}_i + u_i + \text{residual}_i, \tag{5.5}$$

producing a likelihood ratio statistic and p-value. The FDR controlling approach is used again to obtain corresponding q-values. If q-value$_t < 0.1$, expression of gene t is called a partial mediator of the relationship between the QTL and the phenotype. Gene t could also represent an independent effect on the phenotype from the QTL.

4. **Consistency of effects:** The consistency of the signs of the effect of the relationships of X through the mediator M onto Y ($X \rightarrow M \rightarrow Y$) with X on Y ($X \rightarrow Y$) was checked for all called mediators. $X \overset{+}{\rightarrow} Y$ means that X causally increases Y, whereas $X \overset{-}{\rightarrow} Y$ means that X causally decreases Y. Consistent signs for $X \overset{+}{\rightarrow} Y$ would be $X \overset{+}{\rightarrow} M \overset{+}{\rightarrow} Y$ or $X \overset{-}{\rightarrow} M \overset{-}{\rightarrow} Y$. Similarly, for the $X \overset{-}{\rightarrow} Y$ relationship, consistent mediation relationships would be $X \overset{+}{\rightarrow} M \overset{-}{\rightarrow} Y$ or $X \overset{-}{\rightarrow} M \overset{+}{\rightarrow} Y$. Inconsistent signs, also referred to as paradoxical effects, occur when signs of the relationships to and from the mediator are not consistent with the sign of the relationship from X to Y, suggesting that M potentially acts as a suppressive mediator of the relationship $X \rightarrow Y$.

The validity of the causal inference from the mediation analysis depends on the underlying relationships following a directed acyclic graph (DAG). If cycles are present in the graph, the causal inference will likely not be valid. Cycles cannot exist with $X \rightarrow Y$ and $X \rightarrow M$ because the QTL genotype is essentially fixed and cannot be modulated by other quantities. Notably the assumption is made that $M \rightarrow Y$, and that $M \leftarrow Y$ does not occur, though it is plausible that a QTL could modulate a phenotype ($X \rightarrow Y$), which leads the phenotype to modulate expression of certain genes ($Y \rightarrow M$). These types of relationships would produce significant associations whose causal directionality would be misinterpreted by the mediation analysis, thus their inference is dependent on the assumption.

5.7.4 Mediation analysis results

Gene expression data from the liver was measured on 398 of the 989 HS rats (all in the cohort with tissue harvested at 17 weeks of life). The three QTL intervals (RetroFat chromosome 1 and chromosome 6 loci and body weight chromosome 4 locus) were evaluated with mediation analysis in an attempt to identify and prioritize possible candidates that could affect the phenotypes through their expression level variation.

5.7.4.1 Body weight chromosome 4 locus

The QTL interval for this locus contained 11 genes (**Table 5.8**). Three of these had liver expression measured. The main candidate *Grid2* was not sufficiently expressed (non-zero expression proportion < 0.25) in liver tissue. The expression levels of the other two genes (*Ccser1* and *LOC108350839*) were not significantly associated with the QTL ($X \rightarrow M$ was not satisfied).

5.7.4.2 RetroFat chromosome 1 locus

The QTL interval contained 15 genes (**Table 5.6** and **5.7**), of which 5 were contained in the expression data (*Emx2*, *Rab11fip2*, *Fam204a*, *Prlhr*, and *Cacul1*). *Emx2* and the primary candidate *Prlhr* were not sufficiently expressed in the liver. Similar to as in body weight, the expression levels of the remaining three genes were not significantly associated with the QTL.

5.7.4.3 RetroFat chromosome 6 locus

The interval for this QTL is much wider than the previous intervals, and contains 130 genes (**Table 5.2**-**5.5**), of which 114 were measured in the liver expression data. Of the 114, 36 genes had non-zero expression below 0.25, leaving 78 genes for which to evaluate $X \rightarrow M$. 14 genes (**Table 5.9**) had a significant association (q-value < 0.1) between expression levels and the QTL. These 14 candidate mediators were then tested for evidence of being full mediators. *Krtcap3* was called a full mediator (p-value = 0.15). The remaining 13 were evaluated as partial mediators, resulting in 5 genes being selected (q-value < 0.1) (**Table 5.10**). As *Krtcap3* was a strong candidate as a full mediator, we replaced the QTL in the model of RetroFat with it. Each partial mediator was then individually included in a regression model of RetroFat with *Krtcap3* and compared to the null model with only *Krtcap3* (**Table 5.10**). Only *Slc30a3* remained significant, suggesting that it is the best candidate as an additional regulator of RetroFat, potentially separately from the QTL/*Krtcap3* signal.

5.8 Additional Figures

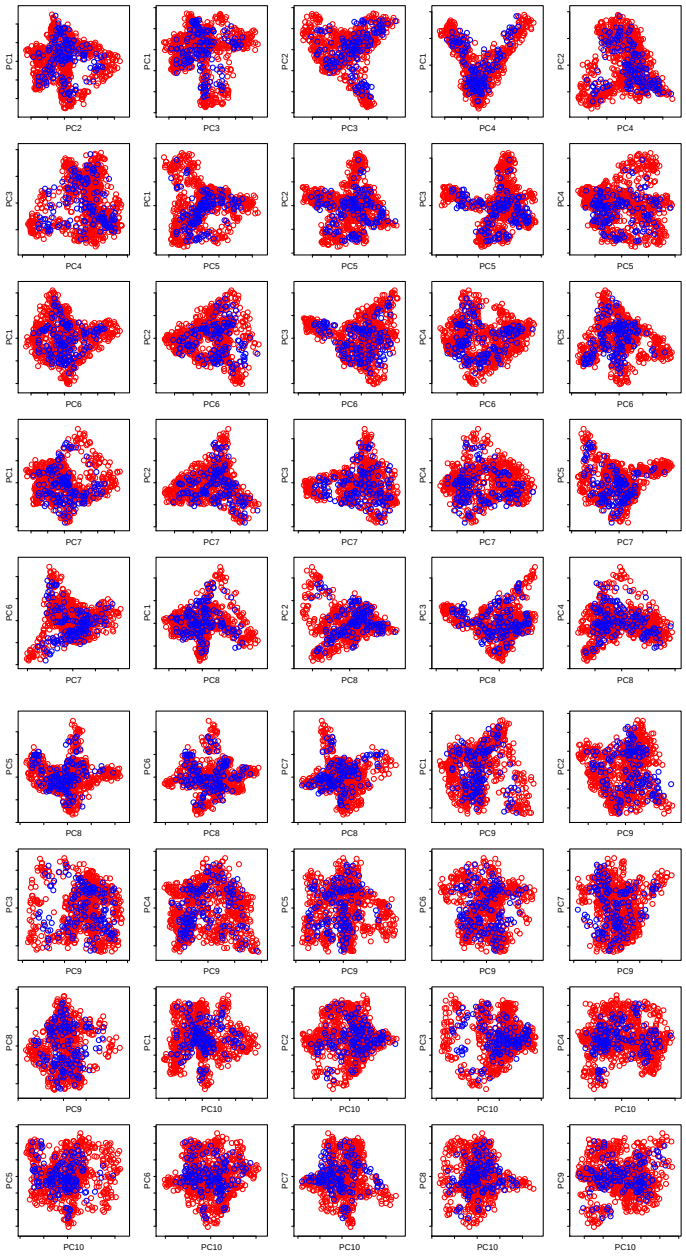

Figure 1.8: Principal component analysis for the first ten principal components of genotypes between two genotyping centers. Those genotyped at Hudson Alpha are plotted in red and those genotyped at Vanderbilt are plotted in blue. For all plots, red and blue points fall within the same general region indicating that there are no systematic differences in genotype between the two centers.

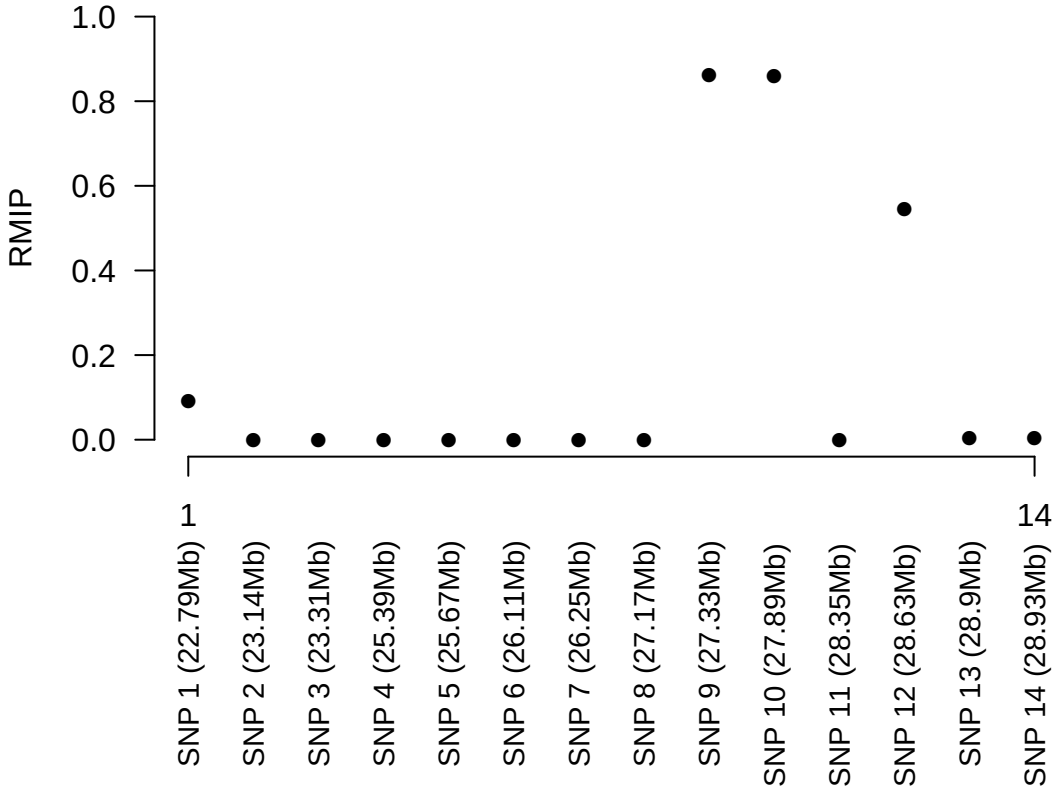

Figure 1.9: Fine-mapping of the chromosome 6 locus using LLARRMA-dawg reduced the LD support interval from 6.14 Mb to 1.46 Mb. LLARRMA-dawg jointly models and selects SNPs in a region, and returns probabilities corresponding to how often a SNP was included over many re-samples of the data (RMIP). Multiple SNPs with high RMIP suggests the potential for multiple independent signals beneath the QTL peak.

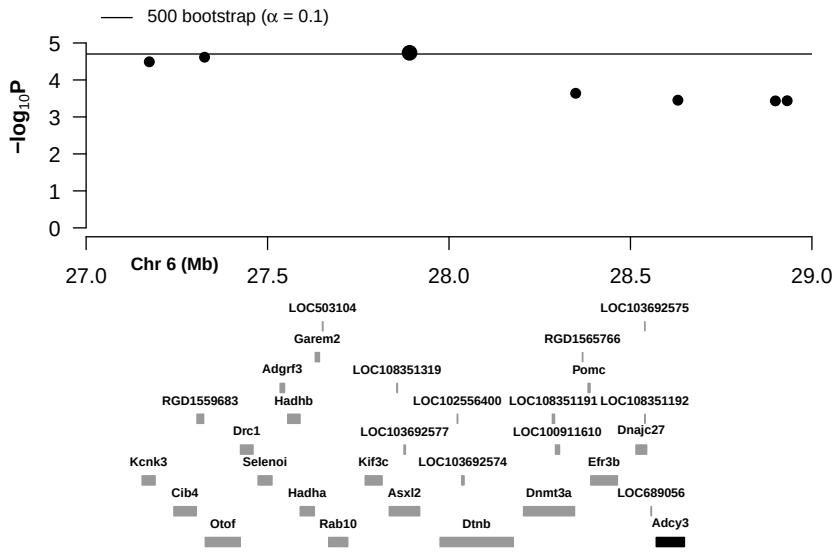

Figure 1.10: The SNP association (7 markers) present in the LLARRMA-dawg fine-mapping interval (**Figure 1.9**), including the annotations of the 30 genes local to the region. The candidate gene *Adcy3* is in bold, and possesses a non-synonymous WKY variant that is predicted to alter protein function (**Figure 1.3E**).

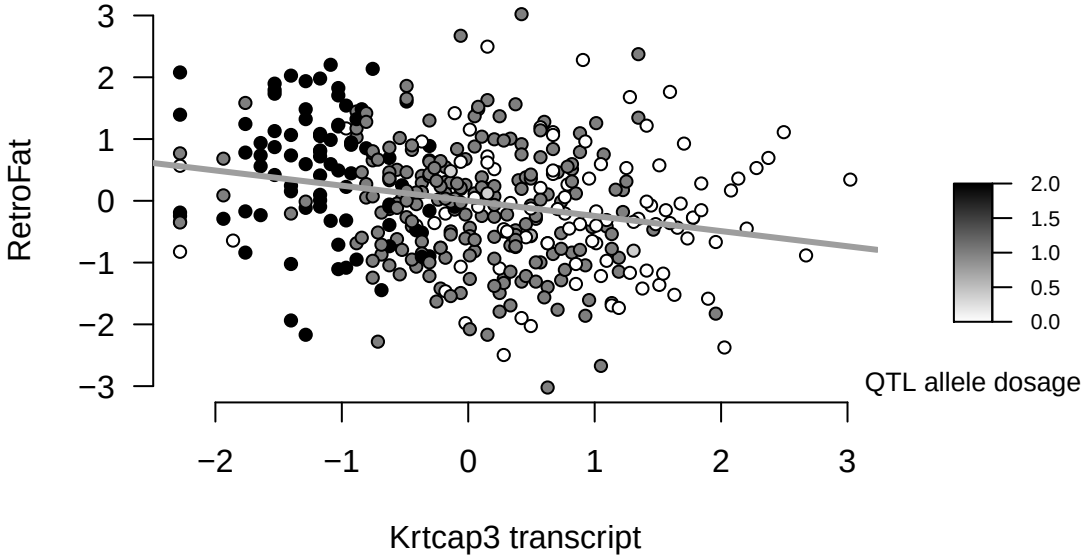

Figure 1.11: Scatterplot of RetroFat and *Krtcap3* expression levels, with data points colored by the peak SNP minor allele dosages at the QTL. RetroFat and expression levels are rank-inverse normal transformed. *Krtcap3* expression is negatively correlated with RetroFatg (negative trend line). Genotype dosage of the QTL peak SNP is positively associated with RetroFat, and negatively associated with *Krtcap3* expression, which matches **Figure 1.5**.

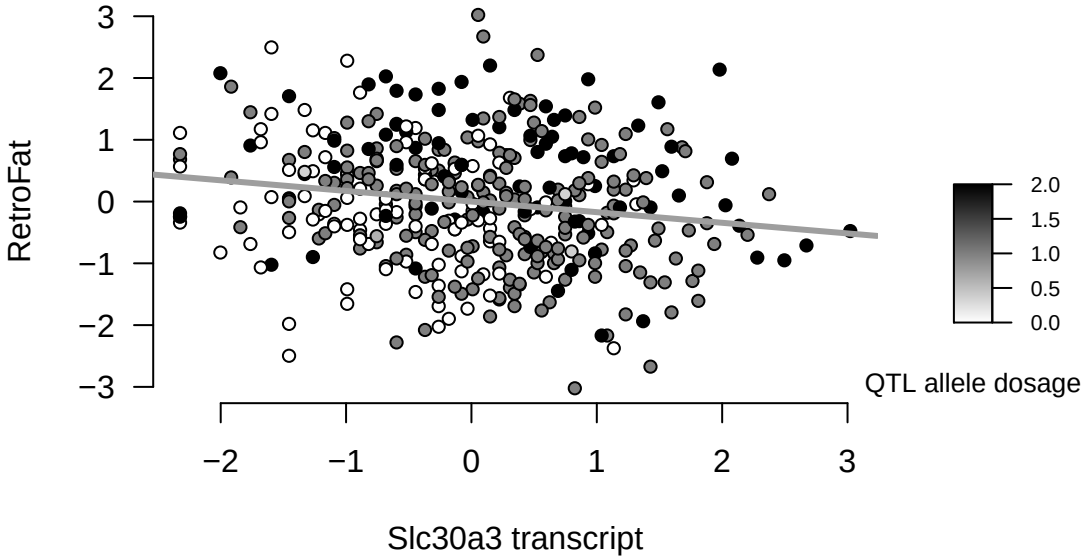

Figure 1.12: Scatterplot of RetroFat and *Slc30a3* expression levels, with data points colored by the peak SNP minor allele dosages at the QTL. RetroFat and expression levels are rank-inverse normal transformed.In contrast to *Krtcap3*, the peak SNP minor allele dosage is positive associated with *Slc30a3*, although its expression is negatively correlated with RetroFatg (negative trend line). The mediation path through *Slc30a3* is inconsistent with the QTL relationship with RetroFat, suggesting that *Slc30a3* may actually act in a suppressive manner with respect to the QTL effect.

Gene Symbol	Gene Name	Start Location	Non-synonymous variants in WKY founder†	Polyphen prediction
Alk	ALK receptor tyrosine kinase	22696415		
LOC108351180	uncharacterized	22988727		
LOC108351181	uncharacterized	23205628		
Clip4	CAP-GLY domain containing linker protein family, member 4	23222020		
LOC103692578	basic proline-rich protein-like	23298261		
RGD1304963	similar to hypothetical protein MGC38716	23337507		
Togaram2	TOG array regulator of axonemal microtubules 2	23358762		
Wdr43*	WD repeat domain 43	23433532		
Trnac-gca30	transfer RNA cysteine (anticodon GCA) 30	23487063		
LOC102551341	tRNA (adenine(58)-N(1))-methyltransferase, mitochondrial-like	23487545		
Spdya	speedy/RINGO cell cycle regulator family member A	23493686	23495595	unknown
LOC102548558	protein tyrosine phosphatase type IVA 1-like	23493704		
Ppp1cb*	protein phosphatase 1 catalytic subunit beta	23548507		
LOC108351182	ALK tyrosine kinase receptor-like	23725713		
LOC298795	similar to 14-3-3 protein sigma	23757225		
LOC108351183	uncharacterized	23771355		
LOC103692567	uncharacterized	23885316		
LOC108351327	glyceraldehyde-3-phosphate dehydrogenase pseudogene	23936327		
LOC103692568	uncharacterized	23986197		
LOC102553396	uncharacterized	24064737		
Ypel5	yippee-like 5	24069351		
Lbh	limb bud and heart development	24154207		
LOC108351184	uncharacterized	24192828		
LOC108351185	uncharacterized	24256909		
LOC102547591	uncharacterized	24336223		
LOC100912066	uncharacterized	24342924		
Lclat1	lysocardiolipin acyltransferase 1	24377398		
LOC102547438	uncharacterized	24527464		
LOC685881	hypothetical protein	24562761		
Capn13	calpain 13	24579590		
LOC102554046	uncharacterized	24623564		
LOC102553955	uncharacterized	24657682		
Galnt14	polypeptide N-acetylgalactosaminyltransferase 14	24770308		
Ehd3	EH-domain containing 3	25076012		
LOC102554201	uncharacterized	25101552		
Xdh	xanthine dehydrogenase	25149570		
LOC100363233	splicing factor 3b, subunit 4-like	25226245		
Srd5a2	steroid 5 alpha-reductase 2	25279635		

Table 5.2: Genes in RetroFat chromosome 6 QTL interval

Gene Symbol	Gene Name	Start Location	Non-synonymous variants in WKY founder†	Polyphen prediction
Plb1	phospholipase B1	25375699		
LOC683819	hypothetical protein	25565221		
Fosl2	FOS like 2, AP-1 transcription factor subunit	25598936		
Babam2	BRISC and BRCA1 A complex member 2	25666654		
LOC103692569	uncharacterized	25885973		
Rbks	ribokinase	26051568	26072561 T to A	benign
Mrpl33	mitochondrial ribosomal protein L33	26130278		
LOC102548914	uncharacterized	26201017		
Slc4a1ap	solute carrier family 4 member 1 adaptor protein	26214083		
Supt7l	SPT7-like STAGA complex gamma subunit	26241672		
Gpn1*	GPN-loop GTPase 1	26255081		
RGD1560110	similar to RIKEN cDNA 4930548H24	26278440		
Zfp512	zinc finger protein 512	26284749		
LOC102556504	titin-like	26322470		
Gckr	glucokinase regulator	26355296		
LOC100910821	uncharacterized	26387284		
Ift172	intraflagellar transport 172	26390686		
LOC108351187	uncharacterized	26407404		
LOC108351186	60S ribosomal protein L37 pseudogene	26415619		
LOC103692570	dihydropyrimidinase-related protein 5-like	26423841		
Krtcap3*	keratinocyte associated protein 3	26485126		
Nrbp1	nuclear receptor binding protein 1	26486823		
Ppm1g	protein phosphatase, Mg2+/Mn2+ dependent, 1G	26517840		
Zfp513	zinc finger protein 513	26537707		
Snx17	sorting nexin 17	26541137		
Eif2b4	eukaryotic translation initiation factor 2B subunit delta	26546917		
Gtf3c2	general transcription factor IIIC subunit 2	26560601	26581578 T to C	unknown
Mpv17	MpV17 mitochondrial inner membrane protein	26585713		
Ucn	urocortin	26602144		
Trim54	tripartite motif-containing 54	26603364		
Dnajc5g	DnaJ heat shock protein family (Hsp40) member C5 gamma	26625526		
Slc30a3*	solute carrier family 30 member 3	26629752		
Cad	carbamoyl-phosphate synthetase 2, aspartate transcarbamylase, and dihydroorotase	26657507		
Atraid*	all-trans retinoic acid-induced differentiation factor	26680628		
Slc5a6	solute carrier family 5 member 6	26685823		
Tcf23	transcription factor 23	26763159		

Table 5.3: Genes in RetroFat chromosome 6 QTL interval (continued)

Gene Symbol	Gene Name	Start Location	Non-synonymous variants in WKY founder†	Polyphen prediction
Prr30	proline rich 30	26780352		
Preb	prolactin regulatory element binding	26784088	26786379 A to G	benign
Abhd1	abhydrolase domain containing 1	26787807		
Cgref1	cell growth regulator with EF hand domain 1	26797126		
Khk	ketohexokinase	26810577		
Emilin1	elastin microfibril interfacer 1	26821249		
LOC103692571	uncharacterized	26833107		
Ost4	oligosaccharyltransferase complex subunit 4, non-catalytic	26836216		
Agbl5	ATP/GTP binding protein-like 5	26837299		
Trnaa-agc6	transfer RNA alanine (anticodon AGC) 6	26856068		
Trnay-gua	transfer RNA tyrosine (anticodon GUA)	26856459		
Trnay-gua3	transfer RNA tyrosine (anticodon GUA) 3	26856459		
Tmem214	transmembrane protein 214	26867638		
Mapre3	microtubule-associated protein, RP/EB family, member 3	26878738		
LOC108351190	uncharacterized	26890051		
LOC108351189	uncharacterized	26918219		
LOC108351188	60S ribosomal protein L37 pseudogene	26931127		
Dpysl5	dihydropyrimidinase-like 5	26939696		
LOC103692572	uncharacterized	27069013		
Cenpa	centromere protein A	27072259		
Slc35f6	solute carrier family 35, member F6	27095144		
LOC103692573	uncharacterized	27139210		
Kcnk3	potassium two pore domain channel subfamily K member 3	27154274		
Cib4	calcium and integrin binding family member 4	27241804		
RGD1559683	similar to RIKEN cDNA 1700001C02	27305402		
Otof	otoferlin	27328343		
Drc1	dynein regulatory complex subunit 1	27425237	27428501 G to A	benign
Selenoi	selenoprotein I	27473748		
Adgrf3	adhesion G protein-coupled receptor F3	27534525		
Hadhb	hydroxyacyl-CoA dehydrogenase/3-ketoacyl-CoA thiolase/enoyl-CoA hydratase (trifunctional protein), beta subunit	27555408		
Hadha	hydroxyacyl-CoA dehydrogenase/3-ketoacyl-CoA thiolase/enoyl-CoA hydratase (trifunctional protein), alpha subunit	27589840		
Garem2	GRB2 associated regulator of MAPK1 subtype 2	27631364		
LOC503104	similar to retinoblastoma binding protein 4	27651115		
Rab10	RAB10, member RAS oncogene family	27668387		

Table 5.4: Genes in RetroFat chromosome 6 QTL interval (continued)

Gene Symbol	Gene Name	Start Location	Non-synonymous variants in WKY founder†	Polyphen prediction
Kif3c	kinesin family member 3C	27768943		
Asxl2	additional sex combs like 2, transcriptional regulator	27835346		
LOC108351319	28S ribosomal protein S21, mitochondrial pseudogene	27856408		
LOC103692577	RNA pseudouridylate synthase domain-containing protein 4 pseudogene	27875868		
Dtnb	dystrobrevin, beta	27975302	28004664 A to G	benign
LOC102556400	transcription factor BTF3-like	28022498		
LOC103692574	uncharacterized	28034953		
Dnmt3a	DNA methyltransferase 3 alpha	28205375		
LOC108351191	60S ribosomal protein L37 pseudogene	28284681		
LOC100911610	dihydropyrimidinase-related protein 5-like	28293250		
RGD1565766	hypothetical gene supported by BC088468; NM_001009712	28367389		
Pomc	proopiomelanocortin	28382937		
Efr3b	EFR3 homolog B	28390541		
Dnajc27	DnaJ heat shock protein family (Hsp40) member C27	28515054		
LOC108351192	cytochrome c oxidase subunit 7B, mitochondrial pseudogene	28539158		
LOC103692575	cytochrome c oxidase subunit 7B, mitochondrial pseudogene	28539172		
LOC689056	similar to general transcription factor IIH, polypeptide 5	28556618		
Adcy3	adenylate cyclase 3	28570941	28572363 A to C	damaging
Cenpo	centromere protein O	28648804		
Ptrhd1	peptidyl-tRNA hydrolase domain containing 1	28663602		
Ncoa1	nuclear receptor coactivator 1	28677563		
LOC103692576	uncharacterized	28812571		

Genes in bold are found within the most likely region of the QTL based on multi-SNP fine-mapping analysis.
*Full or partial mediators of RetroFat called by mediation analysis.
†RetroFat chromosome 6 haplotype effects: WKY has decreased fat pad weight (Figure 3D).

Table 5.5: Genes in RetroFat chromosome 6 QTL interval (continued)

Gene Symbol	Gene Name	Start location	Gene Function (UniProt)	Non-synonymous variants in founders with the haplotype effect†
LOC103691392	uncharacterized	280573647	Unknown	
Emx2	empty spiracles homeobox 2	280633938	Transcription factor which acts to generate the boundary between the roof and archipallium in the developing brain.	
LOC108349711	uncharacterized	280653842	Unknown	
LOC502394	hypothetical	280753676	Unknown	
LOC102555781	uncharacterized	280796426	Unknown	
LOC108349712	uncharacterized	280934585	Unknown	
Rab11fip2	RAB11 family interacting protein 2	281065346	A Rab11 effector binding preferentially phosphatidylinositol 3,4,5-trisphosphate (PtdInsP3) and phosphatidic acid (PA) and acting in the regulation of the transport of vesicles from the endosomal recycling compartment (ERC) to the plasma membrane. Involved in insulin granule exocytosis. Also involved in receptor-mediated endocytosis and membrane trafficking of recycling endosomes, probably originating from clathrin-coated vesicles.	
LOC102556164	uncharacterized	281227923	Unknown	
LOC102556108	uncharacterized	281289720	Unknown	
LOC102556023	acyl carrier protein, mitochondrial-like	281304776	Unknown	
Fam204a	family with sequence similarity 204, member A	281343692	Unknown	

Table 5.6: Genes in RetroFat chromosome 1 QTL interval

Gene Symbol	Gene Name	Start location	Gene Function (UniProt)	Non-synonymous variants in founders with the haplotype effect†
LOC103691393	uncharacterized	281395030	Unknown	
LOC108349713	uncharacterized	281397476	Unknown	
Prlhr	prolactin releasing hormone receptor	281754472	Receptor for prolactin-releasing peptide (PrRP). Implicated in lactation, **regulation of food intake** and pain-signal processing.	281755911 C to T translation start site in BUF and WKY
Cacul1	CDK2-associated, cullin domain 1	281814226	Cell cycle associated protein capable of promoting cell proliferation through the activation of CDK2 at the G1/S phase transition.	

The gene in bold (Prlhr) is the most likely candidate in the region.

†RetroFat chromosome 1 haplotype effects: BUF, MR, WKY haplotypes lead to increased fat pad weight (Figure 6D).

Table 5.7: Genes in RetroFat chromosome 1 QTL interval (continued)

Gene Symbol	Gene Name	Start location	Gene Function (UniProt)	Non-synonymous variants in founders with the haplotype effect†
Ccser1	coiled-coil serine-rich protein 1	91235885	Unknown, has been associated with cocaine	None
LOC103692146	uncharacterized	91601766	Unknown	None
LOC108350840	uncharacterized	91959690	Unknown	None
LOC108350839	high mobility group protein B1-like	92431517	Unknown	None
LOC103692148	developmental pluripotency-associated protein 2 pseudogene	92443293	Unknown	None
LOC103692149	axoneme-associated protein mst101(2)-like	92501663	Unknown	None
LOC103692147	glutamate receptor ionotropic, delta-2-like	93012791	Unknown	None
Hint1-ps1	histidine triad nucleotide binding protein 1, pseudogene 1	93405665	Pseudogene, likely not functional	None
LOC103692150	thyrotropin receptor pseudogene	93447412	Unknown	None
LOC108350813	Ig kappa chain V-II region 26-10-like	93857773	Unknown	None
Grid2	glutamate ionotropic receptor delta type subunit 2	94068112	Receptor for glutamate. L-glutamate acts as an excitatory neurotransmitter at many synapses in the central nervous system.	None

The gene in bold (Grid2) is the most likely candidate in the region.

†Body Weight chromosome 4 haplotype effects: ACI, BUF, F344 and MR haplotypes lead to decreased body weight while BN haplotype leads to increased body weight (Figure 7D).

Table 5.8: Genes in body weight chromosome 4 QTL interval

Gene Symbol	Gene Name	Start location	q-value
RGD1304963	similar to hypothetical protein MGC38716	23337507	4.79E-05
Wdr43	WD repeat domain 43	23433532	7.82E-02
Ppp1cb	protein phosphatase 1 catalytic subunit beta	23548507	7.46E-03
Galnt14	polypeptide N-acetylgalactosaminyltransferase 14	24770308	3.09E-02
Rbks	ribokinase	26051568	3.13E-11
Gpn1	GPN-loop GTPase 1	26255081	3.38E-04
Krtcap3	keratinocyte associated protein 3	26485126	3.30E-41
Slc30a3	solute carrier family 30 member 3	26629752	1.19E-07
Atraid	all-trans retinoic acid-induced differentiation factor	26680628	7.46E-03
Dpysl5	dihydropyrimidinase-like 5	26939696	9.66E-14
Hadha	hydroxyacyl-CoA dehydrogenase/3-ketoacyl-CoA thiolase/enoyl-CoA hydratase (trifunctional protein), alpha subunit	27589840	3.96E-03

Table 5.9: Potential mediators in RetroFat chromosome 6 QTL interval

Gene Symbol	Gene Name	Start location	Full mediation p-value	Joint with *Krtcap3* p-value	Consistency with QTL effect
Wdr43	WD repeat domain 43	23433532	1.05E-04	0.47	Inconsistent
Ppp1cb	protein phosphatase 1 catalytic subunit beta	23548507	7.36E-05	0.28	Inconsistent
Gpn1	GPN-loop GTPase 1	26255081	1.10E-03	0.48	Consistent
Krtcap3	keratinocyte associated protein 3	26485126	0.15	.	Consistent
Slc30a3	solute carrier family 30 member 3	26629752	8.40E-07	2.36E-03	Inconsistent
Atraid	all-trans retinoic acid-induced differentiation factor	26680628	6.09E-05	0.89	Inconsistent

Genes in bold are called as mediators.

Table 5.10: Candidate mediators of the RetroFat chromosome 6 QTL

CHAPTER 6

Detecting chromatin accessibility as a mediator of gene expression in Collaborative Cross mice [1]

6.1 Introduction

Advancements in sequencing technologies over the last decade have made multi-omic experiments feasible, and further advancements in technology and interest are only increasing. Initially these data sets were observational, certainly due to the very real constraints of the populations (humans), the developing technology, and the challenge of coordinating very large experiments with multiple levels of data. With progress in these areas, powerful large-scale experiments with multiple dimensions of data per individual can be paired with modern statistical mediation analysis to draw inferences on the relationships that lie hidden between the levels of these data. These results will be more likely to identify causal rather than correlational elements, and thus provide more meaningful and actionable targets in terms of downstream applications in areas such as medicine and agriculture.

One area of strong interest has been the use of integrative analyses to better understand the regulation of fundamental biological processes that make up the processing of information from genomic DNA to phenotype, which has resulted in a variety of statistical approaches. (Degner et al., 2012) noticed the co-localization of chromatin accessibility QTL (cQTL), assessed through DNAS I sequencing, and expression QTL (eQTL) in human lymphoblastoid cell lines, detecting correlations in their positions. (Battle et al., 2014) investigated the regulation of gene expression in 922 humans, use eQTL and allele specific expression QTL. They did not measure chromatin accessibility, but assessed the evidence whether proximal genes to distal-eQTL behaved as mediators to the gene. (Pai et al., 2015) did not use mediation, rather characterizing the location of eQTL in genomic regulatory elements in human lymphoblastoid cell line data. (Alasoo et al., 2018) similarly did not

[1] This chapter was adapted from a portion of an early draft of a collaborative manuscript that is in preparation. Current author line and title are: Keele, G. R*., Quach, B*., Israel, J. W., Zhou, Y., Chappell, G. A., Lewis, L., Safi, A., Oreper D, Simon, J. M., Crawford, G. E., Valdar, W., Wright, F. A., Rusyn, I., Furey, T. S. Tissue-specific QTL analyses of gene expression and chromatin accessibility in the Collaborative Cross mouse population. Co-first authors*.

use mediation with human cell lines, but rather further elucidated the roles of eQTL within regulatory elements through the use of bacterial infections to modify the enhancer primers. (Roytman et al., 2018) used causal mediation models of histone modifications (hQTL) and expression (eQTL) to better detect signals in data from 112 humans. (Wu et al., 2018) used mediation to tease apart the relationship of DNA methylation sites, gene expression, and complex traits. (Battle et al., 2015) did not use mediation, but characterized the co-localization of QTL underlying gene expression (eQTL), ribosome occupancy (rQTL), and protein abundance (pQTL), detecting significant overlap as well as a buffering of QTL effect from ribosome occupancy up to protein abundance, again in human lymphoblastoid cell lines. (Chick et al., 2016) used a genome-wide mediation approach to characterize the transcriptional and post-translational regulation of proteins in 192 Diversity Outbred (DO) mice (Churchill et al., 2012). These studies demonstrate the span and flexibility of integrative approaches.

One appealing feature of the DO is the potential to replicate results in its related inbred sister population, the Collaborative Cross (CC) (Churchill et al., 2004; Collaborative Cross Consortium, 2012; Srivastava et al., 2017), a panel of recombinant inbred strains descended from the same founder inbred strains as the DO. (Chick et al., 2016) take advantage of this and use CC mice to confirm results in the DO by showing that estimates of founder allele effects from each of the related populations corresponded. Similar to the DO, the CC represent a powerful untapped tool for these integrative analyses of multi-omic data. Though the CC possesses certain limitations in comparison to the DO, such as a restricted number of strains and thus unique genomes, and comparatively reduced mapping resolution, it also has strengths, mainly the potential for replicate observations, which are useful to reduce noise as well as valuable for potential downstream experiments. If the assumed additive model is true for the QTL, mapping is also more powerful in inbred animals in comparison to outbred ones.

In this work, we use a small sample of 47 male CC mice with single observation per strain to investigate the dynamics between chromatin accessibility and gene expression, as done in (Degner et al., 2012) though here chromatin accessibility is measured through Assay for Transposase Accessible Chromatin sequencing (ATAC-Seq). To our knowledge, this is the first study of this kind in the CC. We aim to detect QTL underlying gene expression and chromatin accessibility separately in three tissues: lung, liver, and kidney. We will then assess the support for mediation of the effect

of eQTL on gene expression through chromatin accessibility, using methods inspired by (Chick et al., 2016). We identify and characterize examples of strong mediation, as well as co-localizing but independent eQTL and cQTL. This study is an example of the experimental power of the CC for integrative analysis of multi-omic data, particularly given the limited sample size, and provides support for its continued use in larger, more complex experiments going forward.

6.2 Materials and Methods

6.2.1 Animals

Adult male mice (8-12 weeks old) from 47 CC strains were acquired from the University of North Carolina Systems Genetics Core (Chapel Hill, NC). Animals were maintained on an NTP 2000 wafer diet (Zeigler Brothers, Inc., Gardners, PA) and water *ad libitum*. The housing room was maintained on a 12-h light-dark cycle. The experimental design sought to maximize the number of strains relative to within-strain replications based on the power analysis for QTL mapping in mouse populations (Kaeppler, 1997); therefore, one mouse was used per strain. Mice were euthanized between 8 and 10 a.m. and lungs, liver, and kidney tissues were collected, flash frozen in liquid nitrogen, and stored at -80°C until analysis. These studies were approved by the Institutional Animal Care and Use Committees at Texas A&M University and the University of North Carolina.

6.2.2 Collaborative Cross reference genomes and transcriptomes

Sequencing read mapping required CC strain-specific reference genomes and transcriptomes denoted as "pseudo-genomes" and "pseudo-transcriptomes" respectively. Pseudo-genomes in FASTA file format and corresponding MOD files were downloaded from the CC resource website (`http://csbio.unc.edu/CCstatus/index.py?run=Pseudo`) for Build 37. The Build 37 MOD files map corresponding genomic positions between the pseudo-genomes and the mm9 (C57BL/6J) genomic coordinate space. To construct pseudo-transcriptomes, the RSEM v1.2.31 command `rsem-prepare-reference` was used with default parameter specifications in conjunction with CC strain-specific gene annotations, derived from MOD files and the mm9 RefSeq gene annotations, and the pseudo-genome FASTA files.

6.2.3 mRNA sequencing and processing

Total RNA was isolated from flash-frozen tissue samples using a Qiagen miRNeasy Kit (Valencia, CA) according to the protocol of the manufacturer. RNA purity and integrity were evaluated using a Thermo Scientific Nanodrop 2000 (Waltham, MA) and an Agilent 2100 Bioanalyzer (Santa Clara, CA), respectively. A minimum RNA integrity value of 7.0 was required for RNA samples to be used for library preparation and sequencing. Libraries for samples with a sufficient RNA integrity value were prepared using the Illumina TruSeq Total RNA Sample Prep Kit (Illumina, Inc., San Diego, USA) with ribosomal depletion. Single-end (50bp) sequencing was performed using an Illumina HiSeq 2500.

Following sequencing, reads were filtered to retain only those with a quality score of 20 or greater for at least 90 percent of read positions. Reads with adapter contamination were removed using TagDust. For each sequenced RNA sample, reads were mapped to the appropriate pseudo-transcriptome using the RSEM command `rsem-calculate-expression` with STAR (v2.5.3a) as the specified aligner (parameter set: –star). RSEM utilizes STAR with alignment options that follow ENCODE3 RNA-Seq read mapping guidelines (`https://www.encodeproject.org/pipelines/ENCPL002LSE/`). Gene expression was quantified using RSEM to produce estimated read counts and transcripts per million (TPM) values.

6.2.4 ATAC-Seq data processing

Flash frozen tissue samples were pulverized in liquid nitrogen using the BioPulverizer (Biospec) to break open cells and allow even exposure of intact chromatin to Tn5 transposase. Pulverized material was thawed in glycerol containing nuclear isolation buffer to stabilize nuclear structure and then filtered through Miracloth (Calbiochem) to remove large tissue debris. Nuclei were washed and directly used for treatment with Tn5 transposase. Single-end (50bp) sequencing was performed using an Illumina HiSeq 2500.

Following sequencing, reads were filtered to retain only those with a quality score of 20 or greater for at least 90 percent of read positions, and reads with adapter contamination were removed using TagDust. A maximum of 5 read duplicates were allowed. Prior to read mapping, a GSNAP database for each pseudo-genome was built using GMAP and

the pseudo-genome FASTA file (parameter set: `-k 15, -q 1`). For each sample, reads that passed filtering were aligned to the appropriate pseudo-genome using GSNAP (parameter set: `-k 15, -m 1, -i 5, --sampling=1, --trim-mismatch-score=0, --genome-unk-mismatch=1, --query-unk-mismatch=1`). Multi-mapped reads with more than four genomic locations were removed. Satellite repetitive elements, regions with high sequence homology to mitochondrial DNA, rRNA, and regions on chromosome X with high sequence homology to chromosome Y are prone to producing artifactual signals caused by experimental or technical biases. An mm9 blacklist was constructed containing these problematic regions. Additionally, pseudo-genome specific blacklists were created by combining RepeatMasker annotations, BLAT derived chromosome X/Y homologous segments, and genomic regions in strong sequence homology to mitochondrial DNA. Regions in these blacklists were removed from consideration in subsequent analyses.

Using the CC strain MOD files, mapped reads for each ATAC-Seq sample were converted to mm9 genomic coordinates to enable direct comparison of data between samples. To account for any differences between the pseudo-genome blacklists and the mm9 blacklist, converted reads that mapped to mm9 blacklist regions were removed. Following conversion, all reads aligning to the positive strand were offset +5 bp, and all reads aligning to the negative strand were offset by -5 bp. These read shifts account for a previously characterized behavior in the integration of adapters by Tn5 transposase upon DNA binding.

6.2.5 Chromatin accessibility quantification and windowing

For each sample, genomic regions representing high chromatin accessibility, i.e. peaks, were determined using the peak-calling software F-seq with default parameters. To define an initial common set of chromatin regions, across all tissues the union set of the top 50,000 peaks (ranked by F-seq score) from each sample was derived and overlapping peaks were merged. These peaks were subsequently divided into overlapping 300 bp windows as previously described. Briefly, peaks smaller than 300 bp were expanded to 300 bp, and for any peak larger than 300 bp, the number of 300 bp windows to segment the peak and not exceed its boundaries was determined using an initial overlap constraint of 100 bp. If the windows spanned less than 90% of bases within the peak, an additional window was added and the overlap was adjusted to produce uniformly spaced windows

that exactly spanned the peak region. Per sample read coverage of each window was calculated using BEDTools coverageBed.

6.2.6 Outcome filtering for eQTL and cQTL mapping

To reduce the computational complexity and multiple testing burden of eQTL and cQTL mapping, the sets of genes and chromatin windows treated as outcome variables were reduced prior to performing association tests. For a given tissue, relative log expression (RLE) normalization, as implemented in the R package DESeq2, was applied to TPM values and read counts of genes and chromatin windows respectively. Genes with RLE-normalized TPM values less than 1 and chromatin windows with normalized counts less than 10 for greater than 50% of samples were excluded from further analysis. For each gene and chromatin window, we applied K-means clustering with $K = 2$ to identify outcomes containing outlier observations that could cause spurious, outlier-driven QTL calls. Any gene or chromatin window where the smaller K-means cluster had a cardinality of 1 was removed. For cQTL mapping, the top 15,000 chromatin windows ranked by standard deviation were selected for analysis.

6.2.7 Founder haplotype data reduction

We reduced the founder haplotype probability data by merging adjacent regions of the genome that are similar in regards to their haplotype pair, also known as diplotype, probability profile. This reduces the computational expense in QTL analysis by reducing the number of genomic loci being tested. The reduction procedure compares adjacent markers and merges their diplotype probabilities (by computing the mean) if the L2 distance between the probability vectors for the adjacent markers is less than 10% of the theoretical maximum L2 distance ($\sqrt{2}$).

Diplotype probabilities for each CC strain are available on the CC resource website (`http://csbio.unc.edu/CCstatus/index.py?run=FounderProbs`). The diplotype probabilities were constructed using an hidden Markov model (HMM) for haplotype inference as previously described (Fu et al., 2012). The genotype calls for which these probabilities are based were obtained using the MegaMUGA SNP array which contains 77,800 genotype markers.

6.2.8 Differential expression and chromatin accessibility analysis

To make the values more comparable across samples, read counts for each sample across all three tissues were converted to counts per million (CPM) and normalized using TMM normalization as implemented in the R package edgeR. To exclude windows with sparse read counts across samples, windows were removed if less than 30% of samples had a CPM value of at least 1. As a final filtering step, regions on chromosome Y and the mitochondria were excluded. For all pairwise tissue comparisons, differentially expressed genes and chromatin windows were determined using the R package limma and the following linear model:

$$\text{CPM}_i = \text{intercept} + \text{strain}_i + \text{batch}_i + \text{tissue}_i + \varepsilon_i, \tag{6.1}$$

where CPM_i represents the TMM-normalized CPM value of either expression of a gene or chromatin accessibility within a chromatin window for lung, liver, or kidney tissue, denoted as tissue_i, from individual i. The effect of sequencing center for individual i is modeled with batch_i, and the CC strain of individual i is represented by strain_i. ε_i is the error term for individual i, with $\boldsymbol{\varepsilon} \sim \text{N}(\mathbf{0}, \mathbf{I}\sigma^2)$.

To account for mean-variance relationships in gene expression and chromatin accessibility data, precision weights were calculated using the limma function voom and incorporated into the linear modeling procedure. Significantly differentially expressed genes and differentially accessible chromatin windows were called based on a Benjamini-Hochberg (BH) FDR of 0.01 and required a minimum $\log_2$ fold-change of at least 1.

In some instances, adjacent chromatin windows may exhibit significant differential chromatin accessibility. We treat these windows as a single region by merging adjacent chromatin windows that have significant differential signal in the same direction. A representative p-value is computed for the merged region using Simes' method (Sarkar and Chang, 1997). The resulting chromatin regions are then re-evaluated for significance using an FDR of 0.01 on the Simes p-values.

6.2.9 Gene set association analysis

We use the software GSAASeqSP with Reactome Pathway Database annotations (July 24, 2015 release) to identify biological pathways associated with differentially expressed genes. For a given

differential expression analysis between two tissues, the resulting gene list was provided as input to GSAASeqSP along with a corresponding weight for each gene, calculated as:

$$\text{weight}_g = sgn(\text{fc}_g) * (1 - p_g), \tag{6.2}$$

where weight_g represents the weight for gene g, $sgn(\text{fc}_g)$ is the sign of the gene expression fold change for gene g between the two tissues, and p_g is the BH adjusted p-value for gene g derived from the differential expression analysis. Pathways with gene sets of cardinality less than 15 and greater than 500 were excluded from analysis.

Gene set association analysis was applied to differentially accessible chromatin regions using a similar approach. Chromatin regions were annotated using GREAT v3.0.0 in `basal plus extension` mode with the parameters `5 kb upstream`, `1 kb downstream`, and no distal extension. GREAT associates genes to chromatin regions that can then be used for pathway enrichment analysis. For each gene output by GREAT, the associated chromatin region with the most significant BH adjusted p-value was selected to represent the gene. Gene weights were calculated as described in Eq 6.2; but $sgn(\text{fc}_g)$ corresponds to the sign of chromatin accessibility fold-changes for gene g, and p_g is the BH adjusted p-value of the chromatin region representing gene g, derived from the differential chromatin accessibility analysis.

6.2.10 QTL mapping

We use a single locus approach to QTL mapping, both when the outcome variable is gene expression and chromatin accessibility. The CC mice have well-characterized founder haplotypes, which allows the use of interval mapping (Lander and Botstein, 1989), in which the association between phenotype and haplotype descent at an interval is assessed instead of at a genotyped marker, implicitly modeling local epistasis. Because haplotype state is not directly observed but rather probabilistically inferred (Lander and Green, 1987; Mott et al., 2000; Liu et al., 2010; Fu et al., 2012; Gatti et al., 2014; Zheng et al., 2015), formal interval mapping requires an computationally inefficient expectation-maximization (EM) algorithm (Dempster et al., 1977). Instead we use a regression approximation (Haley and Knott, 1992; Martínez and Curnow, 1992) that has been commonly used in MPP (Valdar et al., 2006b, 2009; Svenson et al., 2012; Baud et al., 2013, 2014),

including the CC (Aylor et al., 2011; Kelada, 2016; Mosedale et al., 2017; Donoghue et al., 2017), and is computationally efficient. This efficiency is particularly important in the context of a study of genome-wide outcomes, such as gene expression or chromatin accessibility.

A single locus QTL genome scan involves comparing an alternative model with a locus effect to the null model with no locus effect. The alternative model is fit at loci across the genome. The general alternative model for gene expression and chromatin accessibility is the same:

$$f(y_i) = \text{intercept} + \text{QTL}_i + \text{batch}_i + \varepsilon_i, \tag{6.3}$$

where y_i represents the outcome, either levels of the expression of a gene or chromatin accessibility at a genomic site, for individual i. QTL_i is the locus effect. We fit it as seven fixed effects, each representing one of the founder haplotypes, with one founder falling into the intercept term. The effect of sequencing center for individual i is modeled with batch_i. Finally, ε_i is the error term for individual i, with $\boldsymbol{\varepsilon} \sim \text{N}(\mathbf{0}, \mathbf{I}\sigma^2)$. $f(.)$ is a normalizing function that better satisfies the regression assumption that the residuals are normally distributed. We use the rank-based inverse normal transformation, in order to be conservative towards potential extreme observations, particularly because the data are comprised of only 47 individuals. The null model is the same for all tested loci and is equivalent to Eq 6.3 with $\text{QTL}_i = 0$ for all i. The two models are compared statistically at each locus, for which we estimate an F-test p-value. eQTL or cQTL are called based on the locus effect significantly improving the fit of the alternative model compared to the null.

6.2.11 QTL mapping family-wide error rate (FWER) control

For a given outcome, expression or chromatin accessibility, we seek to control the FWER, such that the probability of a false positive result across all genome-wide tests is controlled at some nominal level ($\alpha = 0.05$), rather than at that level of a single test. A stringent approach to multiple testing is used because there are not expected to be an excess number of QTL per outcome. The CC panel, as expected through simulation (Valdar et al., 2006a) and realized to large extent (Srivastava et al., 2017), are balanced in terms of founder haplotype frequency, and as such, are relatively exchangeable, allowing for a permutation procedure to characterize a null distribution for which to compare our results (Doerge and Churchill, 1996). Specifically, we sampled 1000

permutations of the sample identities, and then performed genome scans on each permutation with each outcome. For each outcome, we collected the minimum p-value from each permutation genome scan, transform to the -$\log_{10}$ scale (logP), and then fit a null extreme value distribution (EVD) (Dudbridge and Koeleman, 2004; Valdar et al., 2006a). Genome-wide permutation p-values (permP) are then obtained by calculating the probability of a more extreme logP than the one observed from the cumulative distribution function of the EVD.

6.2.12 eQTL and cQTL false discovery rate (FDR) control

The multiple testing burden is more extreme in studies with genome-wide outcomes, such as with eQTL and cQTL. Not only is an association between a phenotype being tested with loci across the genome, but the whole process is repeated for many outcomes across the genome. Additionally, our expectations in terms of results change. Whereas we do not expect a predominance of QTL per outcome, we do expect many QTL across all the outcomes. Thus we seek to acknowledge the additional testing due to genome-wide outcomes while being more lenient by controlling FDR rather than FWER across outcomes. As in (Chick et al., 2016) we accomplish this by applying an FDR procedure (Benjamini et al., 1995; Storey and Tibshirani, 2003) to the permP, which produces q-values. We then call eQTL and cQTL based on α_{FDR}, such that q-value $\leq \alpha_{\text{FDR}}$. We used $\alpha_{\text{FDR}} = 0.1$.

6.2.13 Detection of multiple QTL per outcome

The per outcome FWER control and across outcome FDR control results in a inability to detect multiple QTL per outcome. The EVD is fit from the maximum statistical score of each outcome, and to include additional strong statistical scores in a biased fashion, such as when there are multiple signals above some FWER genome-wide threshold, would bias the FDR procedure towards more significant results. To avoid this problem, we use a multi-stage conditional fitting approach (Jansen et al., 2017). The procedure is described in the following steps:

1. For a given outcome, conduct genome scan according to the model described in Eq 6.3.

2. Conduct permutation genome scans for the outcome to characterize EVD. Calculate FWER permP from the observed max logP of the genome scan of the outcome. permP is stored to be used as input to FDR method.

3. Specify a genome-wide α_{step} for determining a whether subsequent conditional scans should be conducted for the outcome. We set $\alpha_{\text{step}} = 0.1$. If the permP $> \alpha_{\text{step}}$, no further conditional scans are conducted.

4. If permP $\leq \alpha_{\text{step}}$, the steps 1-3 are repeated for additional conditional scans of the outcome. For $j > 1$, j^{th} conditional scan use the same form of alternative and null model as described in Eq 6.3, except for the inclusion of locus effects for the the peak loci from previous stages. Generally, the alternative model for conditional stage scan J will follow as:

$$f(y_i) = \text{QTL}_i + \sum_{j=1}^{J-1} \text{QTL}_i^{\text{locus}[j]} + \text{batch}_i + \varepsilon_i, \tag{6.4}$$

with $\text{QTL}_i^{\text{locus}[j]}$ representing the locus effect of the peak locus for the j^{th} stage scan of the outcome for individual i, and is also included in the null model of conditional scans. Now repeat steps 2-4.

Initially we were concerned that a multi-stage conditional scan approach could be problematic due to over-fitting because there are only 47 data points per outcome, and each QTL effect actually represents the estimation of seven fixed effects. However, we found that this is appropriately compensated for in the recalculation of the EVD based on permutations of the conditional scans in step 2.

6.2.14 Genome-wide and local chromosome-wide significance

Given that the data represent 47 individual mice, were were concerned that there may be poor power to detect genome-wide eQTL and cQTL. Because there is a strong prior belief in the presence of local eQTL and cQTL, we also evaluated associations for gene expression and chromatin accessibility at the level of local chromosome-wide significance, meaning the chromosome on which the outcome is located. We accomplished this by fitting a local chromosome-wide EVD, producing a local permP for each outcome. We did not use a multi-stage conditional fitting approach for local

chromosome-wide significance, allowing only a single local permP per outcome. We then use the same FDR procedure on these local permP, resulting in local q-values.

6.2.15 Formal mediation analysis

Mediation, particularly causal mediation, is dependent on a number of strong assumptions, such as the underlying variables, their relationships, and the directionality of the relationships, many of which cannot be satisfied in systems far less complex than the relationship between chromatin state and gene expression in mice. However, we believe that consistent evidence of chromatin state acting as a mediator of gene expression could be supportive of the hypothesis that chromatin state has a role in the regulation of gene transcription.

(Baron and Kenny, 1986) establishes the relationships that need to be tested to declare mediation. For our study, the simplified model consists of three variables or nodes: QTL, chromatin accessibility at site k, and expression of gene j. When we call an eQTL for gene j, we detect the relationship:

$$\text{QTL} \quad \rightarrow \quad \text{gene.expression}_j \tag{6.5}$$

If such a relationship exists, the next step is to test whether the chromatin accessibility at site k is also associated with the QTL:

$$\text{QTL} \quad \rightarrow \quad \text{chromatin}_k \tag{6.6}$$

This would be consistent with the expression of gene j and chromatin accessibility at site k possessing co-localizing eQTL and cQTL. If these relationships are detected, full mediation can be tested:

$$\text{gene.expression}_j \quad \perp\!\!\!\perp \quad \text{QTL} \quad | \quad \text{chromatin}_k \tag{6.7}$$

where $\perp\!\!\!\perp$ denotes that two variables are independent. Alternatively, there may be evidence for partial mediations, also referred to as suppressors with the following relationship:

$$\text{chromatin}_k \quad \rightarrow \quad \text{gene.expression}_j \quad | \quad \text{QTL} \tag{6.8}$$

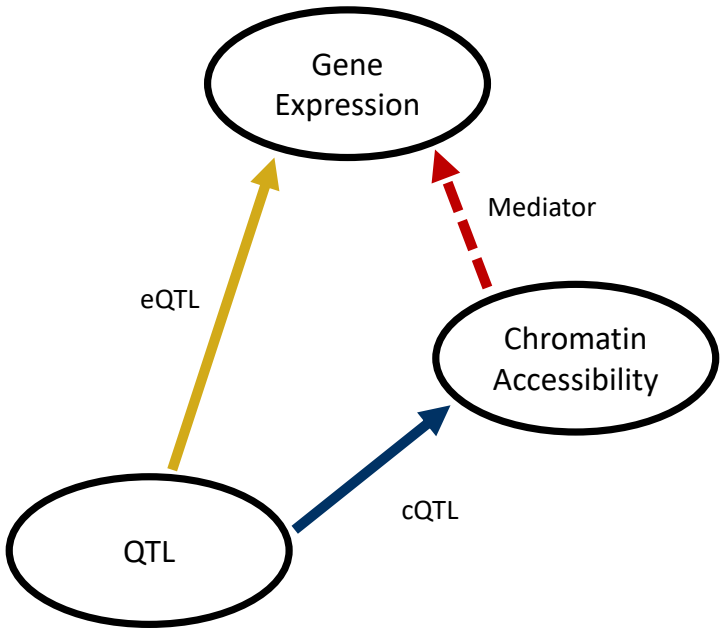

Figure 1.1: The simple relationships being assessed between QTL, chromatin accessibility, and gene expression. The directionality of the relationships between QTL and gene expression and chromatin accessibility is strongly supported in biology, as the haplotype is fixed prior to gene expression and chromatin accessibility dynamics. The assumption of the directionality of the relationship between chromatin accessibility and gene expression is less likely be true in all cases. In reality the relationship may reflect complex equilibriums or be multifactorial and complex. Mediation may also be present and the data are under-powered to detect it. Alternatively, mediation may be present and undetected, because of important additional un-modeled factors, such as the effects of transcription factors and enhancers.

The genome-wide breadth of the data poses a challenge to this framework, in terms of calling co-localizing entities amongst eQTL, cQTL, and chromatin outcomes. Instead we use a simplified, but genome-wide approach. Our simple model of these relationships are depicted in **Figure 1.1**.

6.2.16 Genome-wide mediation analysis

Mediation analysis has previously been used with genomic data (Battle et al., 2014; Roytman et al., 2018; Wu et al., 2018). Our approach to assessing the statistical support of chromatin state mediating gene expression is similar to the method used by (Chick et al., 2016) for detecting mediation of protein abundance through gene expression. We assess evidence for chromatin mediation only in detected genome-wide eQTL because the presence of mediation depends on relationship 6.5 existing.

We simplify the formal mediation analysis to a genome-wide mediation scan which allows us to detect multiple testing corrected significant mediators following relationships 6.7 and 6.8. Similar to the QTL mapping genome scan described in Eq 6.3 and Eq 6.4, the mediation scan involves a comparison of an alternative and a null model at loci across the genome. The alternative model is

$$\begin{aligned} f(\text{gene.expression}_{ij}) = \text{intercept} \\ + \text{eQTL}_i^{\text{gene}j} + \text{chromatin}_{ik} \\ + \text{batch}_i + \varepsilon_i, \end{aligned} \tag{6.9}$$

and the null model is

$$\begin{aligned} f(\text{gene.expression}_{ij}) = \text{intercept} \\ + \text{chromatin}_{ik} \\ + \text{batch}_i + \varepsilon_i, \end{aligned} \tag{6.10}$$

where gene.expression$_{ij}$ is the transcript levels of gene j that has a genome-wide significant eQTL for individual i, eQTL$_i^{\text{gene}j}$ is the locus effect for the eQTL of gene j, chromatin$_{ik}$ is the effect of chromatin accessibility at site k for individual i, batch$_i$ is the effect of the sequencing center used to sequence the gene expression of individual i, and ε_i is the random noise for individual i. Whereas with the QTL genome scan the locus effect was changed at each position, for the mediation scan, it is fixed at the eQTL locus, and the chromatin site effect is changed.

Because the eQTL is always included in the alternative model but not the null model, the average mediation logP of the mediation scan should fluctuate around the observed eQTL logP. At chromatin sites where the chromatin accessibility variable contains some or all of the information present in the eQTL founder haplotype states, the mediation logP will drop due to the competition between the eQTL and chromatin terms. Significant drops represent potential sites of full mediation, as in relationship 6.7. Alternatively, jointly fitting the eQTL and chromatin could increase the mediation logP significantly beyond the original eQTL signal. This signal does not correspond exactly to the partial mediation relationship in 6.8, but similarly represents a site where accounting for chromatin accessibility significantly improves the eQTL signal, which we will refer to as an indirect mediator.

6.2.17 Mediation scan significance thresholds

Our mediation scan approach allows us to define significance through permutations, similarly to as is done with the QTL scans. EVD can be fit for potential full mediators based on minimum logP from permutation scans, and indirect mediators based on maximum logP. We then use these EVD to produce mediation permP, which can be calculated in terms genome-wide and local chromosome-wide significance, with local in reference to the eQTL, not the gene. Finally we use an FDR procedure to obtain mediation q-values.

6.3 Preliminary Results and Discussion

6.3.1 Summaries of the number of associations

This project is ongoing, and thus the results are preliminary. In fact, results are being re-run as a result of an issue detected from the QTL mapping, which will be discussed further. A breakdown of the numbers of detected eQTL in terms of tissue, local/distal status, and genome-wide/local chromosome-wide significance are in **Table 6.1**. Similar summaries for cQTL and mediation are present in **Tables 6.2** and **6.3**, respectively.

6.3.1.1 eQTL

We detect more associations in kidney compared to lung and liver, which have similar numbers of associations. These patterns hold true for chromatin accessibility and mediation. For eQTL, we detect genome-wide eQTL for 5-8% of tested genes across the tissues. In terms of local chromosome-wide significance, the range increases to 17-28%. The eQTL signal is predominantly local signal, ranging from 74-80% local, whether genome-wide or local chromosome-wide.

6.3.1.2 cQTL

We detect more genome-wide associations in chromatin accessibility compared to gene expression, from 13-17% across the tissues, and similar for local chromosome-wide, 16-28%. The relative magnitudes for local to distal cQTL are similar to to those in gene expression.

Table 6.1: Number of genes with eQTL detected (q-value < 0.1) in lung, liver, and kidney tissues

		Tissue (%)		
		Lung	Liver	Kidney
eQTL	genome-wide	772 (5.3[a])	772 (7.2[a])	1092 (8.4[a])
	local chromosome-wide	2573 (17.8[a])	2461 (22.9[a])	3680 (28.4[a])
local-eQTL[b]	genome-wide	578 (74.9[c])	597 (77.3[c])	881 (80.7[c])
	local chromosome-wide	1935 (75.2[d])	1880 (76.4[d])	2769 (75.2[d])
distal-eQTL[e]	genome-wide	203 (26.3[c])	183 (23.7[c])	223 (20.4[c])
	local chromosome-wide	638 (24.8[d])	581 (23.6[d])	911 (24.8[d])

[a]Percentage of all tested genes.

[b]eQTL defined as local if within 10 Mb upstream or downstream of gene TSS.

[c]Percentage of genes with genome-wide eQTL.

[d]Percentage of genes with local chromosome-wide eQTL.

[e]eQTL defined as distal if not within 10 Mb upstream or downstream of gene TSS or on non-local chromosome.

6.3.1.3 Mediation

We test for mediation in only detected eQTL, and find significant genome-wide evidence ranging from 16-18% across all tissues for these eQTL. In terms of local chromosome-wide evidence, we see evidence of mediation in 30-43% of eQTL across all tissues. As with eQTL and cQTL, mediation is appears to be primarily local ranging from 72-82% and 73-74% for genome-wide and local chromosome-wide, respectively.

We expect the distal signals to be reduced upon re-processing of the sequence alignments, particularly for cQTL and mediation. Initially reads were included that could align with up to four positions, which resulted in some distal signals, particularly noticeable in the cQTL results. We are currently re-processing to restrict to the reads that align uniquely.

These results show that it is possible to detect eQTL, cQTL, as well as mediation at a genome-wide level in a relatively small sample (47 mice). Considering curves produced by the SPARCC R package from **Chapter 3** show that this sample size is not sufficiently powered to detect QTL with effects that explain 50% of the outcome variation, which suggests that many of these detected QTL have large effects ($> 50\%$). With more CC strains and replicate observations, more eQTL, cQTL, and mediators would be detected.

Table 6.2: Number of chromatin accessibility sites with cQTL detected (q-value < 0.1) in lung, liver, and kidney tissues

		Tissue (%)		
		Lung	Liver	Kidney
cQTL	genome-wide	2150 (14.3[a])	2034 (13.6[a])	2589 (17.3[a])
	local chromosome-wide	2524 (16.8[a])	4323 (28.8[a])	3351 (22.3[a])
local-cQTL[b]	genome-wide	1802 (83.8[c])	1681 (82.6[c])	1982 (76.6[c])
	local chromosome-wide	2173 (86.1[d])	3376 (78.1[d])	2672 (79.7[d])
distal-cQTL[e]	genome-wide	409 (19.0[c])	388 (19.1[c])	688 (26.6[c])
	local chromosome-wide	351 (13.9[d])	947 (21.9[d])	679 (20.3[d])

[a]Percentage of all tested chromatin site.

[b]cQTL defined as local if within 10 Mb upstream or downstream of chromatin accessibility site.

[c]Percentage of chromatin accessibility sites with genome-wide cQTL.

[d]Percentage of chromatin accessibility sites with local chromosome-wide cQTL.

[e]cQTL defined as distal if not within 10 Mb upstream or downstream of chromatin accessibility site or on non-local chromosome.

6.3.2 eQTL and cQTL mapping results

As shown in **Tables 6.1** and **6.2** most QTL for expression and chromatin accessibility are local, as can be seen in **Figure 1.2** as the band along the diagonal of the grids. There are QTL on the off-diagonal, representing distal-QTL. There is some evidence of vertical bands in the cQTL, which would represent a region of the genome that regulates the chromatin accessibility at many sites. It is likely that many of these bands will disappear once reads are restricted to those that uniquely align.

6.3.3 Mediation results

As shown in **Tables 6.3**, we do see instances of strong evidence of chromatin accessibility at genomic positions mediating the effect of an eQTL on gene expression. Although it is not possible to prove causality with these data, it is consistent with biological expectation were the eQTL to be functionally active through chromatin accessibility. A strong example of mediation of the expression of the gene *Dynltb1* in lung tissue through local chromatin accessibility is provided in **Figure 1.3**. This example also highlights an observation that cQTL tend to have larger effects than eQTL, suggesting that there is some buffering of the effect on expression. This is consistent or parallel with a similar dynamic seen in comparison of eQTL effects being larger than pQTL in (Battle et al.,

Table 6.3: Number of chromatin mediators of gene expression detected (q-value < 0.1) in lung, liver, and kidney tissues

		Tissue (%)		
		Lung	Liver	Kidney
mediators	genome-wide	101 (16.5[a])	117 (18.7[a])	170 (18.4[a])
	local chromosome-wide	188 (30.1[a])	273 (43.5[a])	380 (41.1[a])
local-mediators[b]	genome-wide	73 (72.3[c])	96 (82.1[c])	140 (82.4[c])
	local chromosome-wide	139 (73.9[d])	204 (74.7[d])	282 (74.2[d])
distal-mediators[e]	genome-wide	28 (27.7[c])	21 (17.9[c])	30 (17.6[c])
	local chromosome-wide	49 (26.1[d])	69 (25.3[d])	98 (25.8[d])
local-mediators[b] of local-eQTL[f]	genome-wide	70 (69.3[c])	92 (78.6[c])	138 (81.2[c])
	local chromosome-wide	129 (68.6[d])	185 (67.8[d])	270 (71.1[d])
local-mediators[b] of distal-eQTL[g]	genome-wide	3 (3.0[c])	4 (3.4[c])	2 (1.2[c])
	local chromosome-wide	10 (5.3[d])	19 (7.0[d])	12 (3.2[d])

[a]Percentage of genome-wide eQTL.

[b]Mediators defined as local if within 10 Mb upstream or downstream of eQTL.

[c]Percentage of genome-wide mediators.

[d]Percentage of local chromosome-wide mediators.

[e]Mediator defined as distal if not within 10 Mb upstream or downstream of eQTL or on non-local chromosome.

[f]eQTL defined as local if within 10 Mb upstream or downstream of gene TSS.

[g]eQTL defined as distal if not within 10 Mb upstream or downstream of gene TSS or on non-local chromosome.

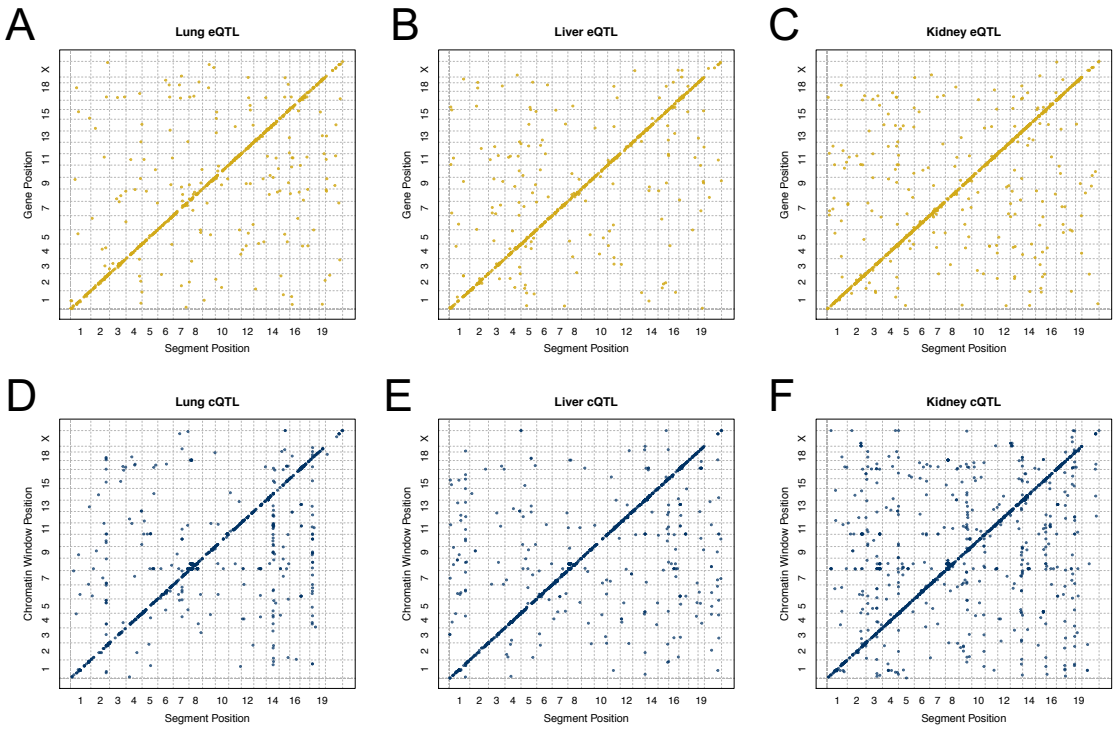

Figure 1.2: Grid plots of eQTL (yellow) and cQTL (blue) in lung (A,D), liver (B,E) , and kidney (C,F), significant at q-value ≤ 0.1. There is a predominance of local-eQTL and local-cQTL relative to distal signals, matching the biological expectation.

2015). We will discuss a systematic approach to checking this QTL effect buffering later on in the Discussion.

6.3.3.1 Identifying co-localizing eQTL and cQTL with mediation and without

Statistical detection of mediation does not simply reflect that eQTL and cQTL are located physically nearby; in fact, eQTL and cQTL can have the same position and not provide any evidence of mediation. The formal mediation involves a statistical test in which the mediator, chromatin accessibility, must absorb much of the effect of the detected eQTL for mediation to be detected. eQTL and cQTL could co-localize, but have highly different founder haplotypes driving the effects. We find that using the regressions coefficients as founder allele effects can visually distinguish co-localizing eQTL and cQTL with mediation and co-localizing eQTL and cQTL without mediation, and even quantified with Spearman's correlation, as in **Figure 1.4**. In the case of mediation, the allele effects of eQTL for *Gm14403* are highly correlated with the allele effects of the co-localizing cQTL. In the case of no mediation, as in *Ear2*, the correlation is much lower between effect vectors.

6.3.3.2 eQTL, cQTL, and mediation are highly local

As presented in **Tables 6.1**, **6.2**, and **6.3**, the QTL and mediators are largely local. We present all three levels of signals simultaneously through radial plots for each tissue in **Figure 1.5**. The plots contain a lot of information, but we emphasize that the inner circles have colored lines connecting eQTL to gene TSS (yellow), cQTL to chromosome accessibility region (blue), and mediator to eQTL (red). Local signals present as a line segment or stick, whereas distal signals are curved lines that connect positions on the circle. Overwhelmingly, the inner circle is covered in line segments representing local signals. In particular, the presence of red local mediators predominantly occur where both local eQTL and cQTL are present. We do not formally require this in our genome-wide mediation test, though the co-occurrence supports that we are detecting true signals.

6.3.3.3 Detection of alignment issue in chromatin accessibility data

In **Figure 1.5**, a strong set of distal cQTL are present in **Figures 1.5A** and **1.5C**, representing cQTL on chromosome 8 for chromatin accessibility regions on chromosome 18. Further investigation revealed that the chromosome 18 chromatin outcomes had a strong WSB signal, matching the

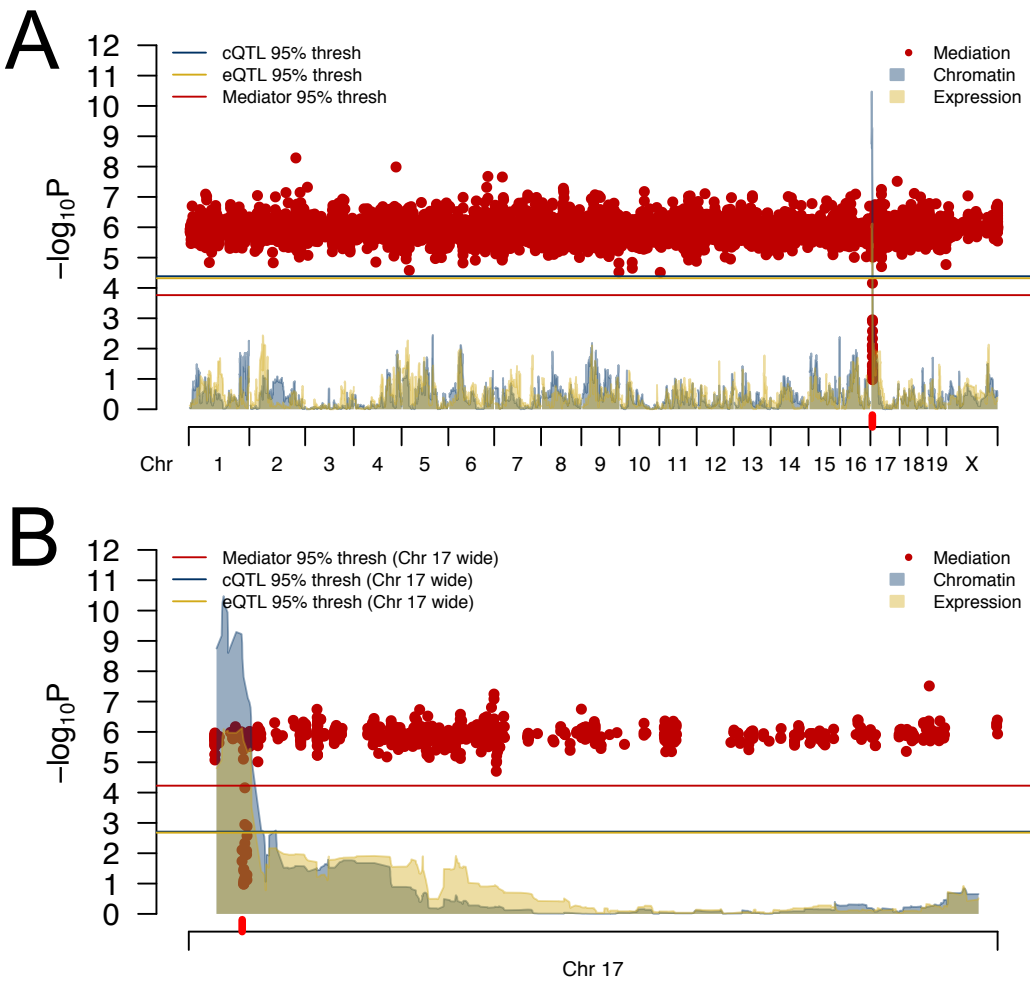

Figure 1.3: Genome scans for expression levels (yellow), chromatin accessibility (blue) and chromatin mediation (red) of *Dynltb1*, a gene located on chromosome 17, in 47 CC lines in lung tissue. There is both a strong local-eQTL and a strong local-cQTL present near the transcription start site of *Dynltb1* (red tick). The steep logP drop in the mediation scan at or near the co-localizing QTL is supportive of mediation of *Dynltb1* expression through local chromatin accessibility. Genome-wide scan with corresponding significance thresholds (A). Scan of chromosome 17, the local chromosome, with corresponding local chromosome-wide significance thresholds (B).

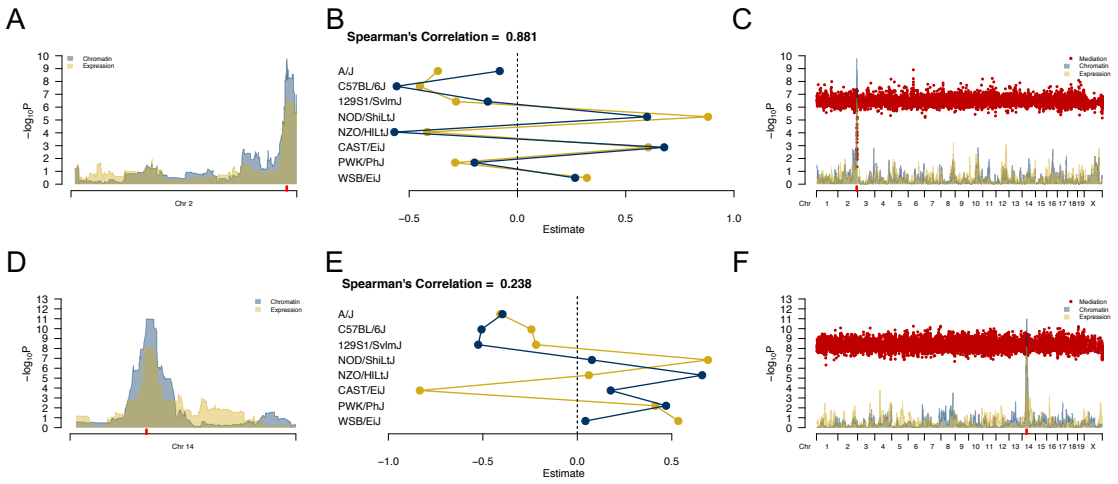

Figure 1.4: Co-localizing eQTL and cQTL are not sufficient for mediation. Co-localizing QTL are observed for which there is evidence of mediation for the gene *Gm14403* in lung tissue (A). The haplotype effects at the eQTL and cQTL position (red tick) are highly correlated (B). The resulting mediation scan shows strong evidence of mediation (C). For the gene *Ear2* in lung, co-localizing eQTL and cQTL are also observed (D). The eQTL and cQTL haplotype effects do not correlate closely, with a particularly strong CAST effect in expression but not in chromatin (E). A mediation signal is not detected for *Ear2*.

haplotype pattern present in the region on chromosome 8. This pattern of effects would correspond to true distal-cQTL, though it also raised the possibility that the sequence similarity in the regions are resulting in reads from chromosome 8 aligning to chromosome 18. To reduce the risk of false distal QTL and mediators, we are re-processing the data to only use reads that uniquely align within the genome.

6.3.4 Distance from QTL or mediator to outcome

We investigated the relationship between statistical association and physical distance from gene TSS for putative eQTL, chromatin accessibility region for putative cQTL, and eQTL for putative mediator, for signals on the local chromosome. We see that putative QTL or putative mediators nearby their outcome tend to have more statistically significant associations, which corresponds to the predominance of local signal. We present genome-wide significant results in **Figure 1.6** and local chromosome-wide significant in **Figure 1.7**. We also observe an odd pile up in cQTL with small p-values around 50 Mb away from the chromatin outcome in lung and kidney, likely representing

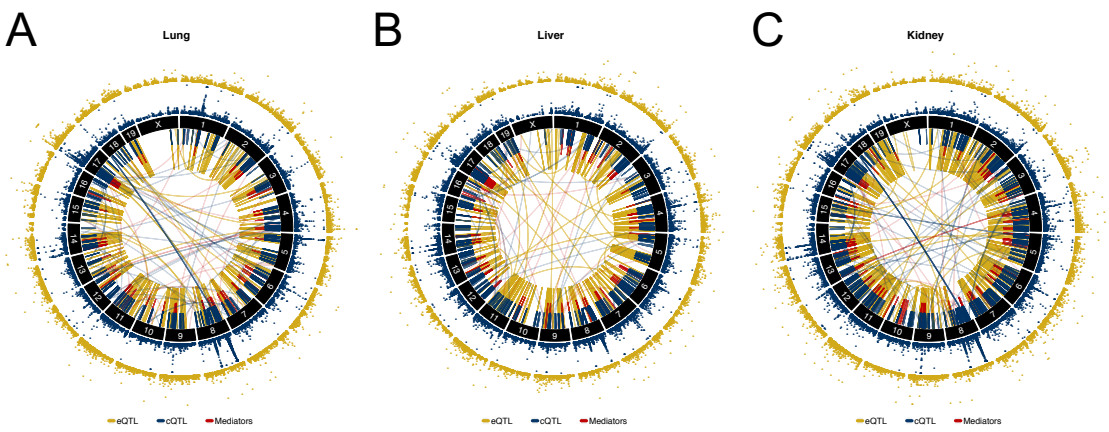

Figure 1.5: Radial plots of eQTL (yellow), cQTL (blue), and mediation (red) in (A) lung, (B) liver, and (C) kidney. The outer-most yellow ring depicts the local-eQTL permP and the middle blue ring is the local-cQTL permP. The inner circle has lines connecting gene TSS to eQTL (yellow), chromatin accessibility sites to cQTL (blue), and eQTL to mediating chromatin accessibility sites (red), each representing a significant signal of q-value ≤ 0.01. There is a predominance of local-eQTL, local-cQTL, and local-mediators compared to distal signals. Mediation signals primarily occur where both eQTL and cQTL are detected.

distal bands occurring due to the alignment issue. We expect these to disappear after the data are re-processed.

6.3.5 Buffering of eQTL effect from cQTL effect

Visually we see evidence that the cQTL effects are more extreme than the eQTL effects based on the magnitudes of the associations at the QTL in **Figures 1.3** and **1.4**. This dynamic is consistent with biology in that gene expression is further down the regulatory pathway of transcription, and thus more steps for noise to be introduced into the system. It is also similar to findings in (Battle et al., 2015) on the effect of eQTL on gene expression in comparison to the effect of pQTL on protein abundance. Prior to publication, we plan to systematically assess this dynamic. (Battle et al., 2015) uses human data and thus models QTL based on SNP genotypes, thus the QTL effect likely represents a scalar estimate of the effect of a dose of the minor allele of the SNP. With haplotype-based association in the CC, we instead estimate an eight element vector as an effect, making it non-trivial to estimate a similar quantity. We plan to use a regression approach in which the regression coefficients, call them $\boldsymbol{\beta}$ from the QTL_i term in Eq 6.3 is constrained through the imposition of a variance component: $\boldsymbol{\beta} \sim \text{N}(\mathbf{0}, \mathbf{I}\tau^2)$ (Wei and Xu, 2016). τ^2, the variance component, is a scalar summary of the QTL

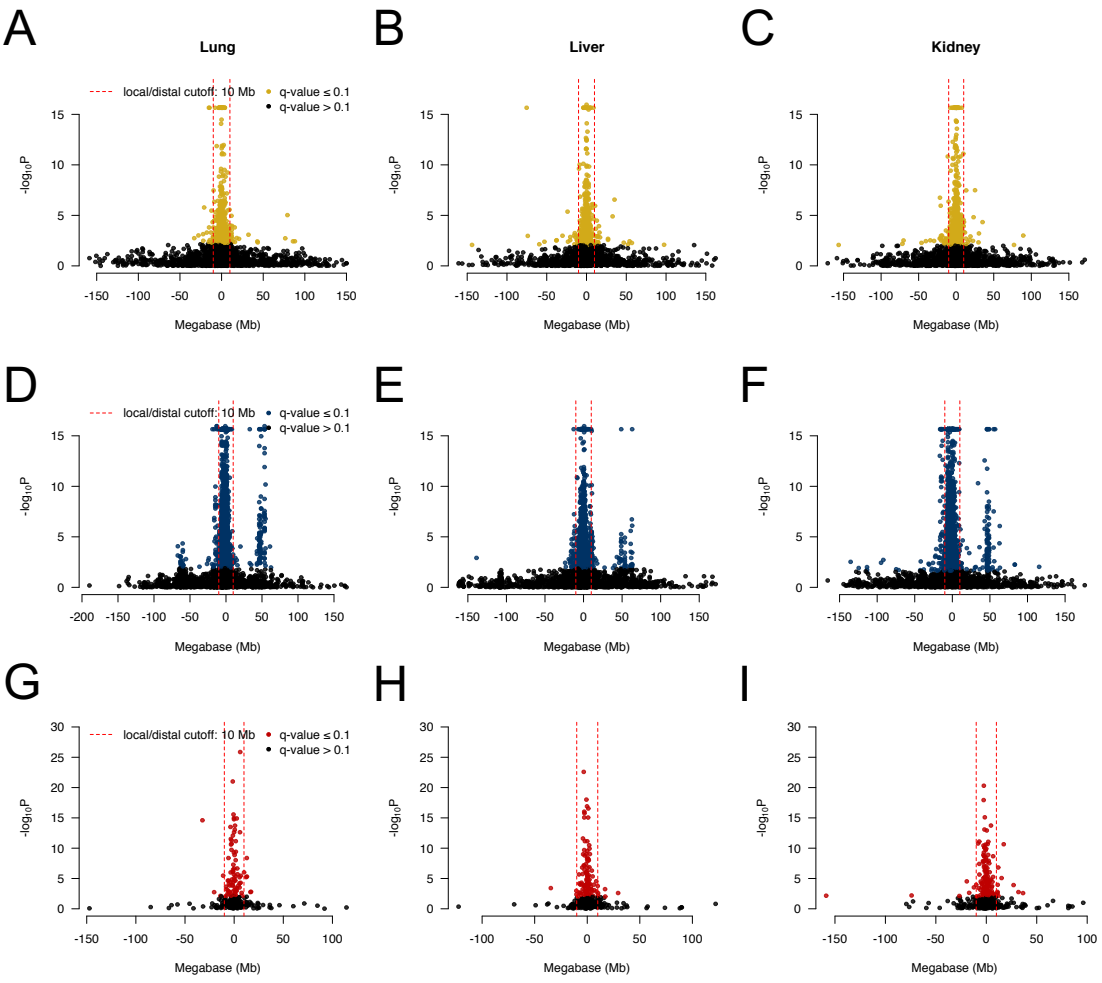

Figure 1.6: The genome-wide permP for eQTL (A-C), cQTL (D-F), and mediators (G-I) by distance (Mb) from gene TSS, chromatin site, and eQTL, respectively. Associations that do not present on the local-chromosome are not shown. The red dashed lines represent 10 Mb upstream and downstream of gene TSS, chromatin site, or eQTL for classifying an association as local or distal. Significant signals (colored symbols), based on q-value ≤ 0.1, are largely local. cQTL exhibit an interesting pattern of non-syntenic association on the local chromosome, clustering around 50 Mb from the chromatin site. This pattern is observed in all tissues, but is more pronounced in lung and kidney (**Figures 1.6D** and **1.6E**). The figures for the chromosome-wide results are present in **Figure 1.7**.

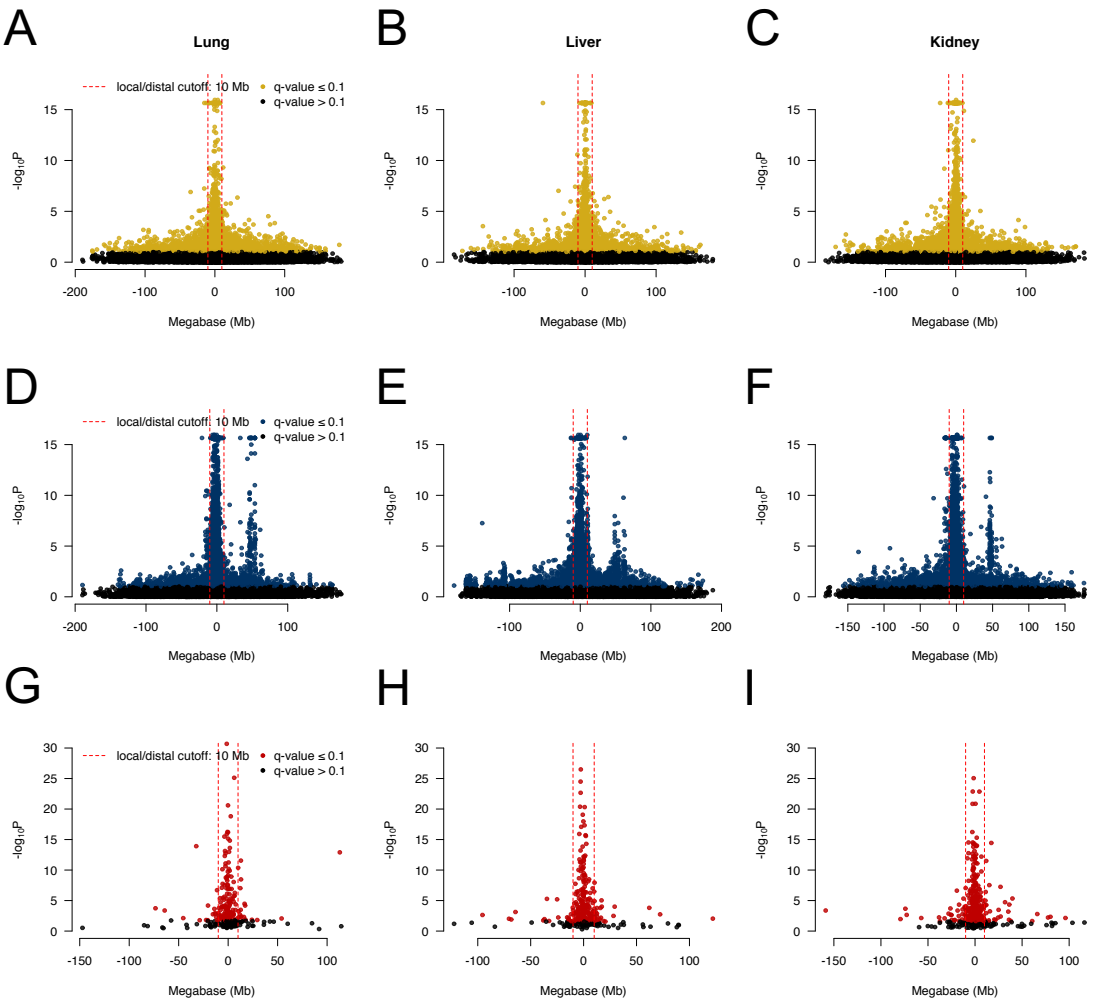

Figure 1.7: The local chromosome-wide permP for eQTL (A-C), cQTL (D-F), and mediators (G-I) by distance (Mb) from gene TSS, chromatin site, and eQTL, respectively. Associations that do not present on the local-chromosome are not shown. The red dashed lines represent 10 Mb upstream and downstream of gene TSS, chromatin site, or eQTL for classifying an association as local or distal. Significant signals (colored symbols), based on q-value ≤ 0.1, are largely local. cQTL exhibit the same interesting pattern of non-syntenic association on the local chromosome as seen in the genome-wide results, clustering around 50 Mb from the chromatin site. We see more local signals are identified through a local chromosome test, though the number of non-local associations (based on 10 Mb) on the local chromosome also increase. The figures for the genome-wide results are present in **Figure 1.6**.

effect, with larger variance components corresponding to larger effects. At all eQTL and cQTL, we can estimate $\widehat{\tau^2}_{\text{cQTL}}$ and $\widehat{\tau^2}_{\text{eQTL}}$ and summarize the overall distributions in comparison to each other. We can also look specifically at cQTL and eQTL that likely represent mediation pairs. Finally, we can look at effects sizes based on local and distal status. Random effect fits of the locus effect are computationally expensive, therefore not appealing for mapping scans with cQTL and eQTL. However, with the set of tests constrained to only the detected cQTL and eQTL, it is feasible.

6.3.6 Frequency of distal-QTL signal in comparison to local-QTL

(Battle et al., 2014) identified local-eQTL in 78.8% and distal-eQTL in 2.9% of all the genes tested. In kidney, for which we detected the most QTL, we detect local eQTL in 6.8% and distal eQTL in 1.7% of tested genes. The disparity is almost certainly the result of the data comprising 47 individuals in comparison to 922. More CC strains and replicate observations would increase power to detect the QTL, and likely mediation as well. The relative proportions between local and distal QTL are also closer in the CC mice than human data. Some of the distal-QTL will likely disappear once the data are re-processed, particularly in the cQTL. There is also the potential that a small sample of 47 CC mice are slightly prone to false positive distal QTL in comparison to 922 humans. Haplotype-based association fits a comparatively complex model in comparison to SNP genotypes, and though the CC have fairly good balance in founder haplotype contributions, there are deviations from it at certain loci. This can result in situations in which a single or a few individuals has a rare founder allele at a locus and an extreme phenotype, resulting in a sudden association (discussed in **Chapter 4**). Shrinkage approaches could account for this, and would certainly reduce distal-QTL in comparison to local-QTL, but would prove computationally challenging. More CC strains and replicate observations would respectively result in more balanced founder contributions within the sample and reduction in outliers that power false associations.

There is value in the amount of local-QTL signal that we are able to detect given the sample size, and particularly that there is evidence of mediation at some of the eQTL. Our use of local chromosome-wide significance also supports our ability to detect local-QTL, as the proportion of local-eQTL out of all genes increases to 21.4% from 6.8% compared to 7% from 1.7% for distal-eQTL in kidney tissue, suggesting that predominantly more local-eQTL, and thus likely real signal is being detected.

6.3.7 Complexity of the underlying mediation

We use a simplistic model for the mediation of the eQTL effect (**Figure 1.1**). It is likely that the true relationship between gene expression and chromatin accessibility is complex and multifactorial for many genes. A simplistic mediation model as well as potential heterogeneity in the data with respect to un-modeled factors would greatly reduce power to detect mediation. We present these results not as definitive catalogue of genes with expression that is modulated by chromatin accessibility, but rather as proof of principle that simplistic mediation models can be used to detect strong signals, even in small samples. Another approach to consider would be to extract the variants contained within the genomic intervals from the ISVdb resource (Oreper et al., 2017), and more complex mediation models could be explored.

6.3.8 Summary

In this study, we map eQTL and cQTL in lung, liver, and kidney tissues in 47 male CC mice, using a multi-stage conditional fitting approach, which allows for the detection of multiple QTL per outcome. We detect mediation of the eQTL effect on gene expression through chromatin accessibility. We find that signals for QTL and mediation are predominantly local. We note that this is a small sample of CC mice with only a single observation per strain, and is thus only powered to detect QTL with very large effects. cQTL effects appear to be larger than eQTL effects, and we describe a novel approach for quantifying this for a multiparental population like the CC. The ISVdb resource could also be used to further investigate the relationships underlying eQTL and potentially mediating cQTL. This study demonstrates that the CC can be a powerful resource for integrative experiments, which largely remains untapped.

6.4 Additional Figures

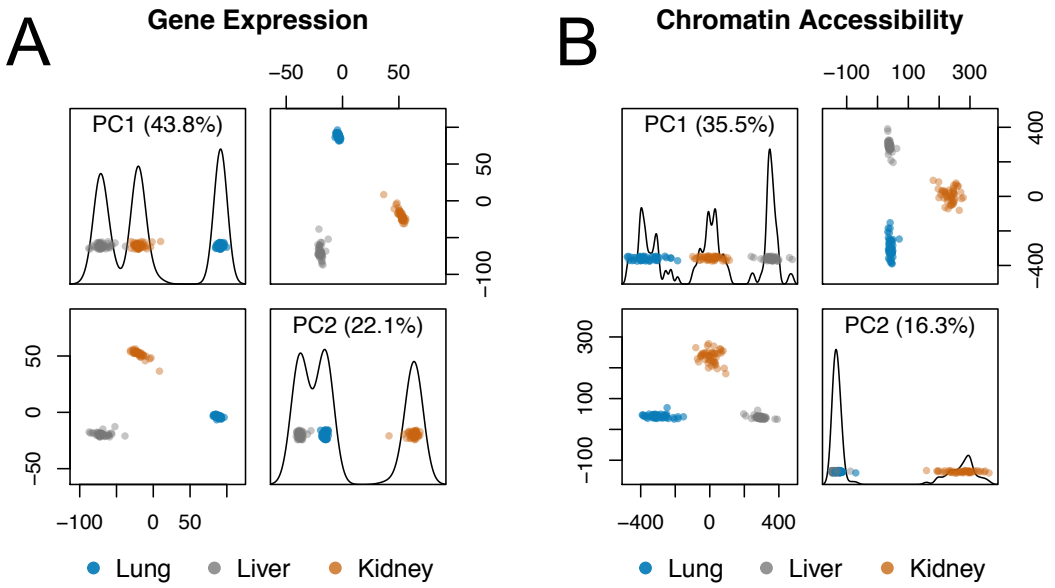

Figure 1.8: Principal components (PC) analysis of gene expression (A) and chromatin accessibility (B) for lung (blue), liver (gray), and kidney (orange) tissue samples derived from RNA-Seq and ATAC-Seq data, respectively. PC 1 and 2 capture a majority of the variance and show a greater amount of between tissue variability than within tissue variability.

CHAPTER 7
Concluding remarks

In this dissertation, statistical methods are developed and utilized that support the complete arc of an experiment involving genetic association within a multiparental population (MPP):

1. Experimental design

2. Association analysis and related fine-mapping and follow-up analyses

MPP are experimental populations composed of individuals descended from more than two inbred founders. They have been developed in a number of animal systems, including but certainly not limited to mice (Churchill et al., 2004, 2012), rats (Hansen and Spuhler, 1984), fly (King et al., 2012a; Long et al., 2014; King and Long, 2017; Najarro et al., 2017; Stanley et al., 2017), and roundworm (Noble et al., 2017), as well as numerous plant systems (Kover et al., 2009; Bandillo et al., 2013; Buckler et al., 2009; Bouchet et al., 2017; Mangandi et al., 2017; Tisné et al., 2017). These populations are valuable resources for genetic experiments due to the greater extent of genetic and likely phenotypic variation they possess in comparison to traditional inbred strains and bi-parental crosses. Though analytical procedures developed in bi-parental populations have been successfully extended to MPP, such as Haley-Knott (HK) regression (Haley and Knott, 1992; Martínez and Curnow, 1992), there is potential for modeling approaches that better accommodate these resources and can thus more efficiently and powerfully extract the underly biological inferences. The procedures and analyses presented were performed with data from MPP rodent populations (mice and rats), however, the fundamental concepts and ideas generalize to other organisms and their MPP. Herein is a summary of the findings and conclusions from the chapters of this dissertation.

7.1 Experimental design

7.1.1 Using the diallel to select optimal bi-parental crosses to map QTL

Chapter 2 dealt with, in general terms, a Bayesian decision theoretic approach for the evaluation of a utility function based on pilot data across potential downstream experiments, with the intent to improve selection of experimental designs. In application, a Bayesian hierarchical model used the diallel, a unique MPP, as pilot data to characterize strain-level effects (Lenarcic et al., 2012), which were then the inputs into the utility functions for possible bi-parental crosses (F2 intercrosses, backcrosses (BC), and parent-of-origin effect reciprocal BC), implemented as the R package DIDACT (Diallel Informed Decision theoretic Approach for Crosses Tool). The utility function used in DIDACT was the power to detect a putative QTL underlying the strain-level effects, though simpler functions could be implemented.

In practice, DIDACT was found to perform well in both Mendelian and complex phenotypes. For a largely Mendelian trait, body weight loss percentage after Influenza A infection, which is known to be largely driven by the gene *Mx1* (Maurizio et al., 2018), DIDACT correctly favors bi-parental crosses that match a strain with null *Mx1* allele with a strain with a functional allele, thus resulting in mapping populations with segregating alleles at the gene. For complex phenotypes, in which the strain-level effects are likely highly polygenic in nature, though the assumptions underlying the utility function are false and the nominal power biased upward, DIDACT favors crosses that match strains that are disparate in phenotype. Thus DIDACT provides a quantitative and principled approach to selecting bi-parental crosses that in practice will not deviate from good and standard practice of matching phenotypically distinct strains, while also allowing interesting strain-level effect combinations to inform the utility function.

7.1.2 Simulated power to map QTL in the realized Collaborative Cross

(Valdar et al., 2006a) estimated power to map QTL in the Collaborative Cross (CC) through simulation, in which the recombinant inbred (RI) strain genomes and phenotype were simulated, based on the stated expectation of 1000 RI strains. Due to allelic incompatibilities many CC lines went extinct (Shorter et al., 2017), resulting in approximately 75 final strains. **Chapter 3** describes

the R package SPARCC (Simulated Power Analysis in the Realized Collaborative Cross), which is designed to provide power calculations that are highly tailored to specific experiments in the actual finalized CC genomes, and thus assist researchers in designing their CC experiments. Additionally SPARCC can be used to explore the effect on power of various aspects of the experimental design: the number of CC strains and the number of replicate observations, and the underlying biology: QTL effect size, background strain effect size, QTL position, and allelic series (Yalcin et al., 2005), representing the number of function alleles and their distribution amongst the founder strains.

Based on large-scale simulations, SPARCC finds that increasing the number of CC strains is more important than increasing replicate observations, though both will improve mapping power. With respect to the allelic series, as the number of functional alleles increases, the power increases. An increase in background strain variance reduces mapping power, which replicate observations cannot improve. For QTL with fewer functional alleles than the number of founders (less than eight), the balance in how the alleles are distributed amongst the founders strongly affects power, with greater balance resulting in increased power. Summary power curves from SPARCC are also presented for general reference for researchers designing CC mapping experiments.

7.2 Genetic association and related analyses

7.2.1 Accounting for haplotype uncertainty in QTL mapping of multiparental populations using multiple imputation

In **Chapter 4** a haplotype-based QTL mapping approach is proposed that takes a multiple imputation (MI) approach to conservatively and stably test for QTL associations in comparison to the unstable associations observed using the standard approach for MPP, HK regression, also referred to as regression-on-probabilities (ROP) (Haley and Knott, 1992; Martínez and Curnow, 1992). Simulations show that MI is conservative in comparison to ROP when the founder haplotype contributions are roughly balanced at loci across the genome. Problems arise for ROP when imbalanced founder haplotype contributions and haplotype uncertainty combine to produce strong correlations between the phenotype and near-zero probabilities for a founder haplotype that has been lost through genetic drift, as is observed frequently in a heterogenous stock (HS) rat data set as well

as at some loci in the CC. Though more computationally intensive than ROP, MI is still feasible, as well as providing the ability to observe the fragility of association over imputations.

Also discussed is the problematic situation when founder haplotype contributions are imbalanced, but now with great certainty, resulting in haplotype parameters that are fit based on only a few individuals, representing highly leveraged data points. MI will not reduce strong associations that result from leveraged data points that have extreme outcomes. This situation actually represents the biased associations that result from fitting a fixed effect parameter to too few data points, and as such, shrinkage procedures should be used, either through variance components (Wei and Xu, 2016) or through pseudo-observations. Though computationally less efficient than ROP, MI and shrinkage methods provide QTL mapping approaches in MPP when ROP fails.

7.2.2 QTL mapping in outbred rat population with imbalanced founder allele frequencies

Chapter 5 describes a QTL mapping analysis in the the HS population first described in **Chapter 4** that detected QTL for adiposity traits, specifically two QTL for retroperitoneal fat pads (RetroFat) and one QTL for body weight, and the subsequent fine-mapping analyses to identify candidate genes and variants. As this dissertation is largely focused on statistical methods for MPP, the summary of results will focus on the development and use of methods, rather than the specific genes and variants that were indentified.

An imputed SNP association approach to QTL mapping was used rather than haplotype-based association because the HS had highly imbalanced founder haplotype contributions at most loci, as well as high levels of uncertainty in terms of distinguishing certain founder haplotypes. In such a context, imputed SNP association was found to be stable, and potentially more powerful for this population than the MI approach previously described.

Various models were used to fine-map the QTL regions. The LLARRMA-dawg method (Sabourin et al., 2015) reduced the RetroFat chromosome 6 QTL interval from 6.14 Mb to 1.46 Mb. Founder haplotype effects were estimated with the Diploffect model (Zhang et al., 2014), which were used to prioritize variants in the region that matched the effect pattern. In particular, a strong effect from the WKY founder was detected for the RetroFat chromosome 6 QTL, which identified a cluster of genes with unique WKY alleles in the region. Protein modeling (Prokop et al., 2017) was used to

predict the effect of candidate variant alleles with respect to protein function, which supported the gene *Adcy3* for the RetroFat chromosome 6 QTL, and is also supported in the literature (Speliotes et al., 2010; Nordman et al., 2008; Stergiakouli et al., 2014; Wen et al., 2012), and *Prlhr* for the RetroFat chromosome 1 QTL. *Grid2* was the primary candidate for the body weight chromosome 4 QTL based on previous literature (Dietrich and Horvath, 2013; Locke et al., 2015) and there only being two genes present in the region.

Finally, an integrative mediation analysis was used to fine-map the QTL regions as well, testing for the potential that the phenotypes are modulated partly through gene expression. The RetroFat chromosome 6 locus contained many co-localizing expression QTL (eQTL), many also driven by variants present in the WKY founder. Expression of the gene *Krtcap3* was found to be a full mediator of the QTL effect on RetroFat. There was also evidence for the expression of the gene *Slc30a* as partial mediator or suppressor. Essentially nothing is known about *Krtcap3* based on the literature and bioinformatic resources. It is possible that the WKY effect in the region is multifactorial, the result of changes to the expression levels of multiple genes as well as protein function of *Adcy3*.

7.2.3 Detecting chromatin accessibility as a mediator of gene expression in Collaborative Cross mice

Chapter 6 further develops genetic association methods in MPP, as well as the integrative mediation methodology used in **Chapter 5**, though now used to test mediation at a genome-wide level. The data consist of gene expression and chromatin accessibility sequence in 47 male CC mice from lung, liver, and kidney tissues. eQTL and chromatin accessibility QTL (cQTL) were mapped using a multi-stage conditional fitting approach (Jansen et al., 2017), which allows for potentially multiple QTL to be detected per outcome given sufficient support. After QTL mapping, support for mediation of the eQTL effect on gene expression through chromatin accessibility was assessed through a genome-wide scan, similar to the approach used in (Chick et al., 2016). Given the small sample size and strong prior expectation that QTL and mediation signals will be local (arbitrarily set to within 5 Mb of outcome position), local chromosome-wide significance (based on chromosome the outcome is located on) was assessed in addition to genome-wide significance.

Though the genetic association and integrative mediation methods are largely finalized, the results are preliminary, as the investigation is ongoing and sequence data are currently being re-

processed as a result of multiply-aligning sequences resulting in distal-cQTL. Large numbers of eQTL and cQTL are detected, though in reduced levels compared to humans (Battle et al., 2014), likely due to the small sample size (47 compared to 922). Within detected QTL, signals are largely local, with approximately ratios (local:distal) of 3:1 in eQTL and 4:1 in cQTL across the three tissues. This actually represents an excess in comparison to distal signal in humans (very roughly 30:1). It is likely that many distal QTL will be removed with the re-processed sequence data. Additionally, false distal signals may occur in a small sample of CC at sites with imbalanced founder haplotype contributions (described in **Chapter 4**). Classification as local is relatively strict, which strongly supports those signals as legitimate.

Strong signatures of full mediation of eQTL through chromatin accessibility are detected, which are largely local as expected. Mediation status is not equivalent to co-localization of eQTL and cQTL, which can be visualized through founder haplotype effects, further supporting the statistical mediation procedure. A similar dynamic is observed as in (Battle et al., 2015) in which protein abundance QTL (pQTL) effects are less extreme than the corresponding eQTL effect, but here cQTL have larger effects than eQTL, suggesting there is some buffering of the effect of chromatin accessibility on gene expression.

7.3 Final conclusion

This dissertation represents a collection of related projects, connected by their use of MPP and focus on tailoring the statistical methodology to the unique features of such data. The underlying concepts and ideas can be re-used and extended to further expand the efficiency and efficacy of these powerful genetic resources, particularly with respect to the design of experiments, as well as genetic association and related integrative analyses.

BIBLIOGRAPHY

Acar, E. F. and Sun, L. (2013). A generalized kruskal-wallis test incorporating group uncertainty with application to genetic association studies. *Biometrics*, 69(2):427–435.

Alasoo, K., Rodrigues, J., Mukhopadhyay, S., Knights, A. J., Mann, A. L., Kundu, K., HIPSCI Consortium, Hale, C., Dougan, G., and Gaffney, D. J. (2018). Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. *Nature genetics*, page 102392.

Allard, R. W. (1999). History of Plant Population Genetics. *Annual Review of Genetics*, 33(1):1–27.

Aulchenko, Y. S., Struchalin, M. V., and van Duijn, C. M. (2010). ProbABEL package for genome-wide association analysis of imputed data. *BMC bioinformatics*, 11:134.

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., Villena, F. P.-m. D., and Churchill, G. A. (2011). Genetic analysis of complex traits in the emerging Collaborative Cross. *Genome research*, 21:1213–22.

Bahari, M., Rafii, M. Y., Saleh, G. B., and Latif, M. A. (2012). Combining Ability Analysis in Complete Diallel Cross of Watermelon (Citrullus lanatus (Thunb.) Matsum. & Nakai). *The Scientific World Journal*, 2012:1–6.

Bandillo, N., Raghavan, C., Muyco, P. A., Sevilla, M. A. L., Lobina, I. T., Dilla-Ermita, C. J., Tung, C.-W., McCouch, S., Thomson, M., Mauleon, R., Singh, R. K., Gregorio, G., Redoña, E., and Leung, H. (2013). Multi-parent advanced generation inter-cross (MAGIC) populations in rice: progress and potential for genetics research and breeding. *Rice (New York, N.Y.)*, 6(1):11.

Baron, R. M. and Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, 51(6):1173–82.

Battle, A., Khan, Z., Wang, S. H., Mitrano, A., Ford, M. J., Pritchard, J. K., and Gilad, Y. (2015). Genomic variation. Impact of regulatory variation from RNA to protein. *Science (New York, N.Y.)*, 347(6222):664–7.

Battle, A., Mostafavi, S., Zhu, X., Potash, J. B., Weissman, M. M., McCormick, C., Haudenschild, C. D., Beckman, K. B., Shi, J., Mei, R., Urban, A. E., Montgomery, S. B., Levinson, D. F., and Koller, D. (2014). Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome research*, 24(1):14–24.

Baud, A., Guryev, V., Hummel, O., Johannesson, M., Rat Genome Sequencing and Mapping Consortium, and Flint, J. (2014). Genomes and phenomes of a population of outbred rats and its progenitors. *Scientific data*, 1:140011.

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):767–75.

Belknap, J. K. and Atkins, A. L. (2001). The replicability of QTLs for murine alcohol preference drinking behavior across eight independent studies. *Mammalian genome : official journal of the International Mammalian Genome Society*, 12(12):893–9.

Benjamini, Y., Hochberg, Y., Society, R. S., Benajmini, Y., Hochberg, Y., Society, R. S., Benjamini, y. Y., Hochberg, Y., Benajmini, Y., Hochberg, Y., Society, R. S., Benajmini, Y., Hochberg, Y., Society, R. S., Benjamini, y. Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society, Series B*, 57(1):289–300.

Bérénos, C., Ellis, P. A., Pilkington, J. G., and Pemberton, J. M. (2014). Estimating quantitative genetic parameters in wild populations: a comparison of pedigree and genomic approaches. *Molecular ecology*, 23(14):3434–51.

Birchler, J. a., Yao, H., and Chudalayandi, S. (2006). Unraveling the genetic basis of hybrid vigor. *Proceedings of the National Academy of Sciences*, 103(35):12957–12958.

Bogue, M. A., Churchill, G. A., and Chesler, E. J. (2015). Collaborative Cross and Diversity Outbred data resources in the Mouse Phenome Database. *Mammalian Genome*, 26(9-10):511–520.

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.

Broman, K. W. (2001). Review of statistical methods for QTL mapping in experimental crosses. *Lab animal*, 30(7):44–52.

Broman, K. W. (2016). *qtl2geno: Treatment of Marker Genotypes for QTL Experiments*. R package version 0.4-23.

Broman, K. W. (2017). *qtl2scan: Genome Scans for QTL Experiments*. R package version 0.5-13.

Broman, K. W. and Sen, S. (2009). *A Guide to QTL Mapping with R/qtl*. Statistics for Biology and Health. Springer New York, New York, NY.

Buckler, E. S., Holland, J. B., Bradbury, P. J., Acharya, C. B., Brown, P. J., Browne, C., Ersoz, E., Flint-Garcia, S., Garcia, A., Glaubitz, J. C., Goodman, M. M., Harjes, C., Guill, K., Kroon, D. E., Larsson, S., Lepak, N. K., Li, H., Mitchell, S. E., Pressoir, G., Peiffer, J. a., Rosas, M. O., Rocheford, T. R., Romay, M. C., Romero, S., Salvo, S., Sanchez Villeda, H., da Silva, H. S.,

Sun, Q., Tian, F., Upadyayula, N., Ware, D., Yates, H., Yu, J., Zhang, Z., Kresovich, S., and McMullen, M. D. (2009). The genetic architecture of maize flowering time. *Science (New York, N.Y.)*, 325(5941):714–718.

Carbonetto, P., Cheng, R., Gyekis, J. P., Parker, C. C., Blizard, D. A., Palmer, A. A., and Lionikas, A. (2014). Discovery and refinement of muscle weight QTLs in B6 D2 advanced intercross mice. *Physiological genomics*, 46(16):571–82.

Casares, P., Carracedo, M. C., Piñeiro, R., Miguel, E. S., and Garcia-Florez, L. (1992). Genetic basis for female receptivity in Drosophila melanogaster: a diallel study. *Heredity*, 69(5):400–405.

Casellas, J. (2011). Inbred mouse strains and genetic stability: a review. *Animal : an international journal of animal bioscience*, 5(1):1–7.

Chick, J. M., Munger, S. C., Simecek, P., Huttlin, E. L., Choi, K., Gatti, D. M., Raghupathy, N., Svenson, K. L., Churchill, G. A., and Gygi, S. P. (2016). Defining the consequences of genetic variation on a proteome-wide scale. *Nature*, 534(7608):500–5.

Churchill, G., Gatti, D., Munger, S., and Svenson, K. (2012). The Diversity outbred mouse population. *Mammalian Genome*, 23:713–8.

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., de Villena, F. P.-M., 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 (2012). The genome architecture of the Collaborative Cross mouse genetic reference population. *Genetics*, 190(2):389–401.

Connolly, S. and Heron, E. a. (2015). Review of statistical methodologies for the detection of parent-of-origin effects in family trio genome-wide association data with binary disease traits. *Briefings in Bioinformatics*, 16(3):429–448.

Crawford, L., Wood, K. C., Zhou, X., and Mukherjee, S. (2017). Bayesian Approximate Kernel Regression with Variable Selection. *Journal of the American Statistical Association*, 1459(November):0–0.

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.

Cubillos, F. A., Brice, C., Molinet, J., Tisné, S., Abarca, V., Tapia, S. M., Oporto, C., García, V., Liti, G., and Martínez, C. (2017). Identification of Nitrogen Consumption Genetic Variants in Yeast Through QTL Mapping and Bulk Segregant RNA-Seq Analyses. *G3*, 7(6):1693–1705.

Darvasi, A. and Soller, M. (1995). Advanced intercross lines, an experimental population for fine genetic mapping. *Genetics*, 141(3):1199–207.

Degner, J. F., Pai, A. A., Pique-Regi, R., Veyrieras, J.-B., Gaffney, D. J., Pickrell, J. K., De Leon, S., Michelini, K., Lewellen, N., Crawford, G. E., Stephens, M., Gilad, Y., and Pritchard, J. K. (2012). DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature*, 482(7385):390–4.

Dempster, A., Laird, N., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B Methodological*, 39(1):1–38.

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

Dietrich, M. O. and Horvath, T. L. (2013). Hypothalamic control of energy balance: insights into the role of synaptic plasticity. *Trends in neurosciences*, 36(2):65–73.

Doerge, R. and Churchill, G. (1996). Permutation tests for multiple loci affecting a quantitative character. *Genetics*, 142(1977):285–94.

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.

Dos Santos, E. A., de Almeida, A.-A. F., Ahnert, D., Branco, M. C. d. S., Valle, R. R., and Baligar, V. C. (2016). Diallel Analysis and Growth Parameters as Selection Tools for Drought Tolerance in Young Theobroma cacao Plants. *PloS one*, 11(8):e0160647.

Dudbridge, F. and Koeleman, B. P. (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.

Dupuis, J. and Siegmund, D. (1999). Statistical methods for mapping quantitative trait loci from a dense set of markers. *Genetics*, 151:373–86.

Durrant, C. and Mott, R. (2010). Bayesian quantitative trait locus mapping using inferred haplotypes. *Genetics*, 184(3):839–52.

Ellacott, K. L. J., Donald, E. L., Clarkson, P., Morten, J., Masters, D., Brennand, J., and Luckman, S. M. (2005). Characterization of a naturally-occurring polymorphism in the UHR-1 gene encoding the putative rat prolactin-releasing peptide receptor. *Peptides*, 26(4):675–81.

Emdin, C. A., Khera, A. V., Natarajan, P., Klarin, D., Zekavat, S. M., Hsiao, A. J., and Kathiresan, S. (2017). Genetic Association of Waist-to-Hip Ratio With Cardiometabolic Traits, Type 2 Diabetes, and Coronary Heart Disease. *JAMA*, 317(6):626–634.

Fairfull, R. W., Gowe, R. S., and Emsley, J. A. B. (1983). Diallel cross of six longterm selected leghorn strains with emphasis on heterosis and reciprocal effects. *British Poultry Science*, 24(2):133–158.

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.

Flegal, K. M., Kruszon-Moran, D., Carroll, M. D., Fryar, C. D., and Ogden, C. L. (2016). Trends in Obesity Among Adults in the United States, 2005 to 2014. *JAMA*, 315(21):2284–91.

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*, 4(9):1623–1633.

Ghareeb Zeinab, E. and Helal, A. (2014). Diallel analysis and separation of genetic variance components in eight faba bean genotypes. *Annals of Agricultural Sciences*, 59(1):147–154.

Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H., Ch'ang, L.-Y., Huang, W., Liu, B., and Shen, Y. (2003). The international HapMap project. *Nature*, 426(6968):789–796.

Gilbert, N. E. G. (1958). Diallel cross in plant breeding. *Heredity*, 12(4):477–492.

Gonzalez, P. N., Pavlicev, M., Mitteroecker, P., Pardo-Manuel de Villena, F., Spritz, R. A., Marcucio, R. S., and Hallgrímsson, B. (2016). Genetic structure of phenotypic robustness in the collaborative cross mouse diallel panel. *Journal of Evolutionary Biology*, 29(9):1737–1751.

Gonzalo, M., Vyn, T. J., Holland, J. B., and McIntyre, L. M. (2007). Mapping reciprocal effects and interactions with plant density stress in Zea mays L. *Heredity*, 99(1):14–30.

Graham, J. B., Thomas, S., Swarts, J., McMillan, A. A., Ferris, M. T., Suthar, M. S., Treuting, P. M., Ireton, R., Gale, M., and Lund, J. M. (2015). Genetic diversity in the collaborative cross model recapitulates human West Nile virus disease outcomes. *mBio*, 6(3):e00493–15.

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.

Greenberg, A. J., Hackett, S. R., Harshman, L. G., and Clark, A. G. (2010). A Hierarchical Bayesian Model for a Novel Sparse Partial Diallel Crossing Design. *Genetics*, 185(1):361–373.

Griffing, B. (1956). Concept of General and Specific Combining Ability in Relation to Diallel Crossing Systems. *Australian Journal of Biological Sciences*, 9(4):463.

Gu, W., Geddes, B. J., Zhang, C., Foley, K. P., and Stricker-Krongrad, A. (2004). The prolactin-releasing peptide receptor (GPR10) regulates body weight homeostasis in mice. *Journal of molecular neuroscience : MN*, 22(1-2):93–103.

Halekoh, U. and Højsgaard, S. (2014). A Kenward-Roger Approximation and Parametric Bootstrap Methods for Tests in Linear Mixed Models - The R Package pbkrtest. *Journal of Statistical Software*, 59(9):1–32.

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

Hansen, C. and Spuhler, K. (1984). Development of the National Institutes of Health genetically heterogeneous rat stock. *Alcoholism, clinical and experimental research*, 8(5):477–9.

Harper, K. M., Tunc-Ozcan, E., Graf, E. N., Herzing, L. B. K., and Redei, E. E. (2014). Intergenerational and parent of origin effects of maternal calorie restriction on Igf2 expression in the adult rat hippocampus. *Psychoneuroendocrinology*, 45:187–91.

Hartmann, J., Garland, T., Hannon, R. M., Kelly, S. a., Muñoz, G., and Pomp, D. (2008). Fine mapping of "mini-muscle," a recessive mutation causing reduced hindlimb muscle mass in mice. *The Journal of heredity*, 99(6):679–87.

Hirai, H., Launey, T., Mikawa, S., Torashima, T., Yanagihara, D., Kasaura, T., Miyamoto, A., and Yuzaki, M. (2003). New role of delta2-glutamate receptors in AMPA receptor trafficking and cerebellar function. *Nature neuroscience*, 6(8):869–76.

Hoffman, G. E. (2013). Correcting for population structure and kinship using the linear mixed model: theory and extensions. *PloS one*, 8(10):e75707.

Holand, A. M., Steinsland, I., Martino, S., and Jensen, H. (2013). Animal models and integrated nested Laplace approximations. *G3 (Bethesda, Md.)*, 3(8):1241–51.

Huang, X., Paulo, M.-J., Boer, M., Effgen, S., Keizer, P., Koornneef, M., and van Eeuwijk, F. a. (2011). Analysis of natural allelic variation in Arabidopsis using a multiparent recombinant inbred line population. *Proceedings of the National Academy of Sciences of the United States of America*, 108:4488–4493.

Ivy, T. M. (2007). Good genes, genetic compatibility and the evolution of polyandry: use of the diallel cross to address competing hypotheses. *Journal of evolutionary biology*, 20(2):479–87.

Jansen, R., Hottenga, J.-J., Nivard, M. G., Abdellaoui, A., Laport, B., de Geus, E. J., Wright, F. A., Penninx, B. W. J. H., and Boomsma, D. I. (2017). Conditional eQTL analysis reveals allelic heterogeneity of gene expression. *Human molecular genetics*, 26(8):1444–1451.

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.

Katzmarzyk, P. T., Malina, R. M., Pérusse, L., Rice, T., Province, M. A., Rao, D., and Bouchard, C. (2000). Familial resemblance in fatness and fat distribution. *American journal of human biology : the official journal of the Human Biology Council*, 12(3):395–404.

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-Assunção, J. A., 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):289–94.

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

Kelly, S. a., Bell, T. a., Selitsky, S. R., Buus, R. J., Hua, K., Weinstock, G. M., Garland, T., Pardo-Manuel de Villena, F., and Pomp, D. (2013). A novel intronic single nucleotide polymorphism in the myosin heavy polypeptide 4 gene is responsible for the mini-muscle phenotype characterized by major reduction in hind-limb muscle mass in mice. *Genetics*, 195(4):1385–95.

Khisti, R. T., Wolstenholme, J., Shelton, K. L., and Miles, M. F. (2006). Characterization of the ethanol-deprivation effect in substrains of C57BL/6 mice. *Alcohol (Fayetteville, N.Y.)*, 40(2):119–26.

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

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

King, E. G., Merkes, C. M., McNeil, C. L., Hoofer, S. R., Sen, S., Broman, K. W., Long, A. D., and Macdonald, S. J. (2012b). Genetic dissection of a model complex trait using the Drosophila Synthetic Population Resource. *Genome Research*, 22(8):1558–1566.

Kirkpatrick, S. L. and Bryant, C. D. (2014). Behavioral architecture of opioid reward and aversion in C57BL/6 substrains. *Frontiers in behavioral neuroscience*, 8:450.

Klein, R. J. (2007). Power analysis for genome-wide association studies. *BMC Genetics*, 8(1):58.

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).

Krieger, E., Joo, K., Lee, J., Lee, J., Raman, S., Thompson, J., Tyka, M., Baker, D., and Karplus, K. (2009). Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: Four approaches that performed well in CASP8. *Proteins*, 77 Suppl 9:114–22.

Kumar, V., Kim, K., Joseph, C., Kourrich, S., Yoo, S.-H., Huang, H. C., Vitaterna, M. H., de Villena, F. P.-M., Churchill, G., Bonci, A., and Takahashi, J. S. (2013). C57BL/6N mutation in cytoplasmic

FMRP interacting protein 2 regulates cocaine response. *Science (New York, N.Y.)*, 342(6165):1508–12.

Kutalik, Z., Johnson, T., Bochud, M., Mooser, V., Vollenweider, P., Waeber, G., Waterworth, D., Beckmann, J. S., and Bergmann, S. (2011). Methods for testing association between uncertain genotypes and quantitative traits. *Biostatistics*, 12(1):1–17.

Lander, E. S. and Botstein, D. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, 121(1):185–99.

Lander, E. S. and Green, P. (1987). Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences*, 84(8):2363–2367.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4):357–9.

Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N., and Davey Smith, G. (2008). Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in medicine*, 27(8):1133–63.

Lawrence, C. B., Celsi, F., Brennand, J., and Luckman, S. M. (2000). Alternative role for prolactin-releasing peptide in the regulation of food intake. *Nature neuroscience*, 3(7):645–6.

Lawson, H. A., Cheverud, J. M., and Wolf, J. B. (2013). Genomic imprinting and parent-of-origin effects on complex traits. *Nature reviews. Genetics*, 14(9):609–17.

Lee, J. J., Vattikuti, S., and Chow, C. C. (2016). Uncovering the Genetic Architectures of Quantitative Traits. *Computational and Structural Biotechnology Journal*, 14:28–34.

Lenarcic, A. B., Svenson, K. L., Churchill, G. A., and Valdar, W. (2012). A general Bayesian approach to analyzing diallel crosses of inbred strains. *Genetics*, 190(2):413–35.

Li, B. and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics*, 12(1):323.

Li, K. H., Raghunathan, T. E., and Rubin, D. B. (1991). Large-Sample Significance Levels from Multiply Imputed Data Using Moment-Based Statistics and an F Reference Distribution. *Journal of the American Statistical Association*, 86(416):1065–1073.

Li, Y., Willer, C., Sanna, S., and Abecasis, G. (2009). Genotype Imputation. *Annual Review of Genomics and Human Genetics*, 10(1):387–406.

Lister, C. and Dean, C. (1993). Recombinant inbred lines for mapping RFLP and phenotypic markers in Arabidopsis thaliana. *The Plant Journal*, 4(4):745–750.

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.

Liu, Y. and Zeng, Z.-B. B. (2000). A general mixture model approach for mapping quantitative trait loci from diverse cross designs involving multiple inbred lines. *Genetical Research*, 75(3):345–355.

Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., Day, F. R., Powell, C., Vedantam, S., Buchkovich, M. L., Yang, J., Croteau-Chonka, D. C., Esko, T., Fall, T., Ferreira, T., Gustafsson, S., Kutalik, Z., Luan, J., Mägi, R., Randall, J. C., Winkler, T. W., Wood, A. R., Workalemahu, T., Faul, J. D., Smith, J. A., Zhao, J. H., Zhao, W., Chen, J., Fehrmann, R., Hedman, Å. K., Karjalainen, J., Schmidt, E. M., Absher, D., Amin, N., Anderson, D., Beekman, M., Bolton, J. L., Bragg-Gresham, J. L., Buyske, S., Demirkan, A., Deng, G., Ehret, G. B., Feenstra, B., Feitosa, M. F., Fischer, K., Goel, A., Gong, J., Jackson, A. U., Kanoni, S., Kleber, M. E., Kristiansson, K., Lim, U., Lotay, V., Mangino, M., Leach, I. M., Medina-Gomez, C., Medland, S. E., Nalls, M. A., Palmer, C. D., Pasko, D., Pechlivanis, S., Peters, M. J., Prokopenko, I., Shungin, D., Stančáková, A., Strawbridge, R. J., Sung, Y. J., Tanaka, T., Teumer, A., Trompet, S., van der Laan, S. W., van Setten, J., Van Vliet-Ostaptchouk, J. V., Wang, Z., Yengo, L., Zhang, W., Isaacs, A., Albrecht, E., Ärnlöv, J., Arscott, G. M., Attwood, A. P., Bandinelli, S., Barrett, A., Bas, I. N., Bellis, C., Bennett, A. J., Berne, C., Blagieva, R., Blüher, M., Böhringer, S., Bonnycastle, L. L., Böttcher, Y., Boyd, H. A., Bruinenberg, M., Caspersen, I. H., Chen, Y.-D. I., Clarke, R., Daw, E. W., de Craen, A. J. M., Delgado, G., Dimitriou, M., Doney, A. S. F., Eklund, N., Estrada, K., Eury, E., Folkersen, L., Fraser, R. M., Garcia, M. E., Geller, F., Giedraitis, V., Gigante, B., Go, A. S., Golay, A., Goodall, A. H., Gordon, S. D., Gorski, M., Grabe, H.-J., Grallert, H., Grammer, T. B., Gräßler, J., Grönberg, H., Groves, C. J., Gusto, G., Haessler, J., Hall, P., Haller, T., Hallmans, G., Hartman, C. A., Hassinen, M., Hayward, C., Heard-Costa, N. L., Helmer, Q., Hengstenberg, C., Holmen, O., Hottenga, J.-J., James, A. L., Jeff, J. M., Johansson, Å., Jolley, J., Juliusdottir, T., Kinnunen, L., Koenig, W., Koskenvuo, M., Kratzer, W., Laitinen, J., Lamina, C., Leander, K., Lee, N. R., Lichtner, P., Lind, L., Lindström, J., Lo, K. S., Lobbens, S., Lorbeer, R., Lu, Y., Mach, F., Magnusson, P. K. E., Mahajan, A., McArdle, W. L., McLachlan, S., Menni, C., Merger, S., Mihailov, E., Milani, L., Moayyeri, A., Monda, K. L., Morken, M. A., Mulas, A., Müller, G., Müller-Nurasyid, M., Musk, A. W., Nagaraja, R., Nöthen, M. M., Nolte, I. M., Pilz, S., Rayner, N. W., Renstrom, F., Rettig, R., Ried, J. S., Ripke, S., Robertson, N. R., Rose, L. M., Sanna, S., Scharnagl, H., Scholtens, S., Schumacher, F. R., Scott, W. R., Seufferlein, T., Shi, J., Smith, A. V., Smolonska, J., Stanton, A. V., Steinthorsdottir, V., Stirrups, K., Stringham, H. M., Sundström, J., Swertz, M. A., Swift, A. J., Syvänen, A.-C., Tan, S.-T., Tayo, B. O., Thorand, B., Thorleifsson, G., Tyrer, J. P., Uh, H.-W., Vandenput, L., Verhulst, F. C., Vermeulen, S. H., Verweij, N., Vonk, J. M., Waite, L. L., Warren, H. R., Waterworth, D., Weedon, M. N., Wilkens, L. R., Willenborg, C., Wilsgaard, T., Wojczynski, M. K., Wong, A., Wright, A. F., Zhang, Q., LifeLines Cohort Study, Brennan, E. P., Choi, M., Dastani, Z., Drong, A. W., Eriksson, P., Franco-Cereceda, A., Gådin, J. R., Gharavi, A. G., Goddard, M. E., Handsaker, R. E., Huang, J., Karpe, F., Kathiresan, S., Keildson, S., Kiryluk, K., Kubo, M., Lee, J.-Y., Liang, L., Lifton, R. P., Ma, B., McCarroll, S. A., McKnight, A. J., Min, J. L., Moffatt, M. F., Montgomery, G. W., Murabito, J. M., Nicholson, G., Nyholt, D. R., Okada, Y., Perry, J. R. B., Dorajoo, R., Reinmaa, E., Salem, R. M., Sandholm, N., Scott, R. A., Stolk, L., Takahashi, A., Tanaka, T., van 't Hooft, F. M., Vinkhuyzen, A. A. E., Westra, H.-J., Zheng, W., Zondervan, K. T., ADIPOGen Consortium, AGEN-BMI Working Group, CARDIOGRAMplusC4D Consortium, CKDGen Consortium, GLGC, ICBP, MAGIC Investigators, MuTHER Consortium, MIGen Consortium, PAGE Consortium, ReproGen Consortium, GENIE Consortium, International Endogene Consortium, Heath, A. C., Arveiler, D., Bakker, S. J. L., Beilby, J., Bergman, R. N., Blangero, J., Bovet, P., Campbell, H., Caulfield, M. J., Cesana, G., Chakravarti, A., Chasman, D. I., Chines, P. S., Collins, F. S., Crawford, D. C., Cupples, L. A., Cusi, D., Danesh, J., de Faire, U., den Ruijter, H. M., Dominiczak, A. F., Erbel, R., Erdmann, J., Eriksson, J. G., Farrall, M., Felix, S. B., Ferrannini, E., Ferrières, J., Ford, I., Forouhi, N. G., Forrester, T., Franco, O. H., Gansevoort, R. T., Gejman, P. V., Gieger, C., Gottesman, O., Gudnason, V., Gyllensten, U., Hall, A. S., Harris, T. B., Hattersley, A. T., Hicks, A. A., Hindorff,

L. A., Hingorani, A. D., Hofman, A., Homuth, G., Hovingh, G. K., Humphries, S. E., Hunt, S. C., Hyppönen, E., Illig, T., Jacobs, K. B., Jarvelin, M.-R., Jöckel, K.-H., Johansen, B., Jousilahti, P., Jukema, J. W., Jula, A. M., Kaprio, J., Kastelein, J. J. P., Keinanen-Kiukaanniemi, S. M., Kiemeney, L. A., Knekt, P., Kooner, J. S., Kooperberg, C., Kovacs, P., Kraja, A. T., Kumari, M., Kuusisto, J., Lakka, T. A., Langenberg, C., Marchand, L. L., Lehtimäki, T., Lyssenko, V., Männistö, S., Marette, A., Matise, T. C., McKenzie, C. A., McKnight, B., Moll, F. L., Morris, A. D., Morris, A. P., Murray, J. C., Nelis, M., Ohlsson, C., Oldehinkel, A. J., Ong, K. K., Madden, P. A. F., Pasterkamp, G., Peden, J. F., Peters, A., Postma, D. S., Pramstaller, P. P., Price, J. F., Qi, L., Raitakari, O. T., Rankinen, T., Rao, D. C., Rice, T. K., Ridker, P. M., Rioux, J. D., Ritchie, M. D., Rudan, I., Salomaa, V., Samani, N. J., Saramies, J., Sarzynski, M. A., Schunkert, H., Schwarz, P. E. H., Sever, P., Shuldiner, A. R., Sinisalo, J., Stolk, R. P., Strauch, K., Tönjes, A., Trégouët, D.-A., Tremblay, A., Tremoli, E., Virtamo, J., Vohl, M.-C., Völker, U., Waeber, G., Willemsen, G., Witteman, J. C., Zillikens, M. C., Adair, L. S., Amouyel, P., Asselbergs, F. W., Assimes, T. L., Bochud, M., Boehm, B. O., Boerwinkle, E., Bornstein, S. R., Bottinger, E. P., Bouchard, C., Cauchi, S., Chambers, J. C., Chanock, S. J., Cooper, R. S., de Bakker, P. I. W., Dedoussis, G., Ferrucci, L., Franks, P. W., Froguel, P., Groop, L. C., Haiman, C. A., Hamsten, A., Hui, J., Hunter, D. J., Hveem, K., Kaplan, R. C., Kivimaki, M., Kuh, D., Laakso, M., Liu, Y., Martin, N. G., März, W., Melbye, M., Metspalu, A., Moebus, S., Munroe, P. B., Njølstad, I., Oostra, B. A., Palmer, C. N. A., Pedersen, N. L., Perola, M., Pérusse, L., Peters, U., Power, C., Quertermous, T., Rauramaa, R., Rivadeneira, F., Saaristo, T. E., Saleheen, D., Sattar, N., Schadt, E. E., Schlessinger, D., Slagboom, P. E., Snieder, H., Spector, T. D., Thorsteinsdottir, U., Stumvoll, M., Tuomilehto, J., Uitterlinden, A. G., Uusitupa, M., van der Harst, P., Walker, M., Wallaschofski, H., Wareham, N. J., Watkins, H., Weir, D. R., Wichmann, H.-E., Wilson, J. F., Zanen, P., Borecki, I. B., Deloukas, P., Fox, C. S., Heid, I. M., O'Connell, J. R., Strachan, D. P., Stefansson, K., van Duijn, C. M., Abecasis, G. R., Franke, L., Frayling, T. M., McCarthy, M. I., Visscher, P. M., Scherag, A., Willer, C. J., Boehnke, M., Mohlke, K. L., Lindgren, C. M., Beckmann, J. S., Barroso, I., North, K. E., Ingelsson, E., Hirschhorn, J. N., Loos, R. J. F., and Speliotes, E. K. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518(7538):197–206.

Long, A. D., Macdonald, S. J., and King, E. G. (2014). Dissecting complex traits using the Drosophila Synthetic Population Resource. *Trends in Genetics*, 30(11):488–495.

Lu, Y., Day, F. R., Gustafsson, S., Buchkovich, M. L., Na, J., Bataille, V., Cousminer, D. L., Dastani, Z., Drong, A. W., Esko, T., Evans, D. M., Falchi, M., Feitosa, M. F., Ferreira, T., Hedman, Å. K., Haring, R., Hysi, P. G., Iles, M. M., Justice, A. E., Kanoni, S., Lagou, V., Li, R., Li, X., Locke, A., Lu, C., Mägi, R., Perry, J. R. B., Pers, T. H., Qi, Q., Sanna, M., Schmidt, E. M., Scott, W. R., Shungin, D., Teumer, A., Vinkhuyzen, A. A. E., Walker, R. W., Westra, H.-J., Zhang, M., Zhang, W., Zhao, J. H., Zhu, Z., Afzal, U., Ahluwalia, T. S., Bakker, S. J. L., Bellis, C., Bonnefond, A., Borodulin, K., Buchman, A. S., Cederholm, T., Choh, A. C., Choi, H. J., Curran, J. E., de Groot, L. C. P. G. M., De Jager, P. L., Dhonukshe-Rutten, R. A. M., Enneman, A. W., Eury, E., Evans, D. S., Forsen, T., Friedrich, N., Fumeron, F., Garcia, M. E., Gärtner, S., Han, B.-G., Havulinna, A. S., Hayward, C., Hernandez, D., Hillege, H., Ittermann, T., Kent, J. W., Kolcic, I., Laatikainen, T., Lahti, J., Mateo Leach, I., Lee, C. G., Lee, J.-Y., Liu, T., Liu, Y., Lobbens, S., Loh, M., Lyytikäinen, L.-P., Medina-Gomez, C., Michaëlsson, K., Nalls, M. A., Nielson, C. M., Oozageer, L., Pascoe, L., Paternoster, L., Polašek, O., Ripatti, S., Sarzynski, M. A., Shin, C. S., Narančić, N. S., Spira, D., Srikanth, P., Steinhagen-Thiessen, E., Sung, Y. J., Swart, K. M. A., Taittonen, L., Tanaka, T., Tikkanen, E., van der Velde, N., van Schoor, N. M., Verweij, N., Wright, A. F., Yu, L., Zmuda, J. M., Eklund, N., Forrester, T., Grarup, N., Jackson, A. U., Kristiansson, K., Kuulasmaa, T.,

Kuusisto, J., Lichtner, P., Luan, J., Mahajan, A., Männistö, S., Palmer, C. D., Ried, J. S., Scott, R. A., Stancáková, A., Wagner, P. J., Demirkan, A., Döring, A., Gudnason, V., Kiel, D. P., Kühnel, B., Mangino, M., Mcknight, B., Menni, C., O'Connell, J. R., Oostra, B. A., Shuldiner, A. R., Song, K., Vandenput, L., van Duijn, C. M., Vollenweider, P., White, C. C., Boehnke, M., Boettcher, Y., Cooper, R. S., Forouhi, N. G., Gieger, C., Grallert, H., Hingorani, A., Jørgensen, T., Jousilahti, P., Kivimaki, M., Kumari, M., Laakso, M., Langenberg, C., Linneberg, A., Luke, A., Mckenzie, C. A., Palotie, A., Pedersen, O., Peters, A., Strauch, K., Tayo, B. O., Wareham, N. J., Bennett, D. A., Bertram, L., Blangero, J., Blüher, M., Bouchard, C., Campbell, H., Cho, N. H., Cummings, S. R., Czerwinski, S. A., Demuth, I., Eckardt, R., Eriksson, J. G., Ferrucci, L., Franco, O. H., Froguel, P., Gansevoort, R. T., Hansen, T., Harris, T. B., Hastie, N., Heliövaara, M., Hofman, A., Jordan, J. M., Jula, A., Kähönen, M., Kajantie, E., Knekt, P. B., Koskinen, S., Kovacs, P., Lehtimäki, T., Lind, L., Liu, Y., Orwoll, E. S., Osmond, C., Perola, M., Pérusse, L., Raitakari, O. T., Rankinen, T., Rao, D. C., Rice, T. K., Rivadeneira, F., Rudan, I., Salomaa, V., Sørensen, T. I. A., Stumvoll, M., Tönjes, A., Towne, B., Tranah, G. J., Tremblay, A., Uitterlinden, A. G., van der Harst, P., Vartiainen, E., Viikari, J. S., Vitart, V., Vohl, M.-C., Völzke, H., Walker, M., Wallaschofski, H., Wild, S., Wilson, J. F., Yengo, L., Bishop, D. T., Borecki, I. B., Chambers, J. C., Cupples, L. A., Dehghan, A., Deloukas, P., Fatemifar, G., Fox, C., Furey, T. S., Franke, L., Han, J., Hunter, D. J., Karjalainen, J., Karpe, F., Kaplan, R. C., Kooner, J. S., McCarthy, M. I., Murabito, J. M., Morris, A. P., Bishop, J. A. N., North, K. E., Ohlsson, C., Ong, K. K., Prokopenko, I., Richards, J. B., Schadt, E. E., Spector, T. D., Widén, E., Willer, C. J., Yang, J., Ingelsson, E., Mohlke, K. L., Hirschhorn, J. N., Pospisilik, J. A., Zillikens, M. C., Lindgren, C., Kilpeläinen, T. O., and Loos, R. J. F. (2016). New loci for body fat percentage reveal link between adiposity and cardiometabolic disease risk. *Nature communications*, 7:10495.

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

Mangandi, J., Verma, S., Osorio, L., Peres, N. A., van de Weg, E., and Whitaker, V. M. (2017). Pedigree-Based Analysis in a Multiparental Population of Octoploid Strawberry Reveals QTL Alleles Conferring Resistance to Phytophthora cactorum. *G3*, 7(6):1707–1719.

Mansur, L. M., Orf, J. H., Chase, K., Jarvik, T., Cregan, P. B., and Lark, K. G. (1996). Genetic Mapping of Agronomic Traits Using Recombinant Inbred Lines of Soybean. *Crop Science*, 36(5):1327.

Marchini, J. and Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews. Genetics*, 11(7):499–511.

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.

Maurizio, P. L., Ferris, M. T., Keele, G. R., Miller, D. R., Shaw, G. D., Whitmore, A. C., West, A., Morrison, C. R., Noll, K. E., Plante, K. S., Cockrell, A. S., Threadgill, D. W., Pardo-Manuel de Villena, F., Baric, R. S., Heise, M. T., and Valdar, W. (2018). Bayesian Diallel Analysis Reveals Mx1 -Dependent and Mx1 -Independent Effects on Response to Influenza A Virus in Mice. *G3 (Bethesda, Md.)*, 8(2):427–445.

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

McMullan, R. C., Kelly, S. A., Hua, K., Buckley, B. K., Faber, J. E., PardoManuel de Villena, F., and Pomp, D. (2016). Longterm exercise in mice has sexdependent benefits on body composition and metabolism during aging. *Physiological Reports*, 4(21):e13011.

Meng, X.-L. and Rubin, D. B. (1992). Performing Likelihood Ratio Tests with Multiply-Imputed Data Sets. *Biometrika*, 79(1):103.

Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4):1819–1829.

Monforte, A. J. and Tanksley, S. D. (2000). Development of a set of near isogenic and backcross recombinant inbred lines containing most of the Lycopersicon hirsutum genome in a L. esculentum genetic background: A tool for gene mapping and gene discovery. *Genome*, 43(5):803–813.

Mosedale, M., Kim, Y., Brock, W. J., Roth, S. E., Wiltshire, T., Scott Eaddy, J., 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):kfw269.

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):12649–54.

Mulligan, M. K., Ponomarev, I., Boehm, S. L., Owen, J. A., Levin, P. S., Berman, A. E., Blednov, Y. A., Crabbe, J. C., Williams, R. W., Miles, M. F., and Bergeson, S. E. (2008). Alcohol trait and transcriptional genomic analysis of C57BL/6 substrains. *Genes, brain, and behavior*, 7(6):677–89.

Najarro, M. A., Hackett, J. L., and Macdonald, S. J. (2017). Loci Contributing to Boric Acid Toxicity in Two Reference Populations of Drosophila melanogaster. *G3*, 7(June):1631–1641.

Ng, M. C. Y., Graff, M., Lu, Y., Justice, A. E., Mudgal, P., Liu, C.-T., Young, K., Yanek, L. R., Feitosa, M. F., Wojczynski, M. K., Rand, K., Brody, J. A., Cade, B. E., Dimitrov, L., Duan, Q., Guo, X., Lange, L. A., Nalls, M. A., Okut, H., Tajuddin, S. M., Tayo, B. O., Vedantam, S., Bradfield, J. P., Chen, G., Chen, W.-M., Chesi, A., Irvin, M. R., Padhukasahasram, B., Smith, J. A., Zheng, W., Allison, M. A., Ambrosone, C. B., Bandera, E. V., Bartz, T. M., Berndt, S. I., Bernstein, L., Blot, W. J., Bottinger, E. P., Carpten, J., Chanock, S. J., Chen, Y.-D. I., Conti, D. V., Cooper, R. S., Fornage, M., Freedman, B. I., Garcia, M., Goodman, P. J., Hsu, Y.-H. H., Hu, J., Huff, C. D., Ingles, S. A., John, E. M., Kittles, R., Klein, E., Li, J., McKnight, B., Nayak, U., Nemesure, B., Ogunniyi, A., Olshan, A., Press, M. F., Rohde, R., Rybicki, B. A., Salako, B., Sanderson, M., Shao, Y., Siscovick, D. S., Stanford, J. L., Stevens, V. L., Stram, A., Strom, S. S., Vaidya, D., Witte, J. S., Yao, J., Zhu, X., Ziegler, R. G., Zonderman, A. B., Adeyemo, A., Ambs, S., Cushman, M., Faul, J. D., Hakonarson, H., Levin, A. M., Nathanson, K. L., Ware, E. B., Weir, D. R., Zhao, W., Zhi, D., Bone Mineral Density in Childhood Study (BMDCS) Group, Arnett, D. K., Grant, S. F. A., Kardia, S. L. R., Oloapde, O. I., Rao, D. C., Rotimi, C. N., Sale, M. M., Williams, L. K., Zemel, B. S., Becker, D. M., Borecki, I. B., Evans, M. K., Harris, T. B., Hirschhorn, J. N., Li, Y., Patel, S. R., Psaty, B. M., Rotter, J. I., Wilson, J. G., Bowden, D. W., Cupples, L. A., Haiman, C. A., Loos, R. J. F., and North, K. E. (2017). Discovery and fine-mapping of adiposity loci using high density imputation of genome-wide association studies in individuals of African ancestry: African Ancestry Anthropometry Genetics Consortium. *PLoS genetics*, 13(4):e1006719.

Nikpay, M., Šeda, O., Tremblay, J., Petrovich, M., Gaudet, D., Kotchen, T. A., Cowley, A. W., and Hamet, P. (2012). Genetic mapping of habitual substance use, obesity-related traits, responses to

mental and physical stress, and heart rate and blood pressure measurements reveals shared genes that are overrepresented in the neural synapse. *Hypertension research : official journal of the Japanese Society of Hypertension*, 35(6):585–91.

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. *bioRxiv*, 207(December):1663–1685.

Nordman, S., Abulaiti, A., Hilding, A., Långberg, E.-C., Humphreys, K., Ostenson, C.-G., Efendic, S., and Gu, H. F. (2008). Genetic variation of the adenylyl cyclase 3 (AC3) locus and its influence on type 2 diabetes and obesity susceptibility in Swedish men. *International journal of obesity (2005)*, 32(3):407–12.

Nystrand, M., Dowling, D. K., and Simmons, L. W. (2011). Complex Genotype by Environment interactions and changing genetic architectures across thermal environments in the Australian field cricket, Teleogryllus oceanicus. *BMC Evolutionary Biology*, 11(1):222.

Oreper, D., Cai, Y., Tarantino, L. M., de Villena, F. P.-M., and Valdar, W. (2017). Inbred Strain Variant Database (ISVdb): A Repository for Probabilistically Informed Sequence Differences Among the Collaborative Cross Strains and Their Founders. *G3 (Bethesda, Md.)*, 7(6):1623–1630.

Ott, R. L. and Longnecker, M. T. (2006). *Introduction to Statistical Methods and Data Analysis (with CD-ROM)*. Duxbury Press.

Pai, A. A., Pritchard, J. K., and Gilad, Y. (2015). The genetic and mechanistic basis for variation in gene regulation. *PLoS genetics*, 11(1):e1004857.

Parker, C. C., Carbonetto, P., Sokoloff, G., Park, Y. J., Abney, M., and Palmer, A. A. (2014). High-resolution genetic mapping of complex traits from a combined analysis of F2 and advanced intercross mice. *Genetics*, 198(1):103–16.

Parker, C. C., Cheng, R., Sokoloff, G., Lim, J. E., Skol, A. D., Abney, M., and Palmer, A. A. (2011). Fine-mapping alleles for body weight in LG/J SM/J F and F(34) advanced intercross lines. *Mammalian genome : official journal of the International Mammalian Genome Society*, 22(9-10):563–71.

Parker, C. C., Sokoloff, G., Cheng, R., and Palmer, A. A. (2012). Genome-wide association for fear conditioning in an advanced intercross mouse line. *Behavior genetics*, 42(3):437–48.

Parsons, P. A. (1964). A diallel cross for mating speeds in Drosophila melanogaster. *Genetica*, 35(1):141–151.

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.

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., and Frelinger, J. A. (2014). Using the emerging Collaborative Cross to probe the immune system. *Genes & Immunity*, 15(1):38–46.

Pitman, J. L., Wheeler, M. C., Lloyd, D. J., Walker, J. R., Glynne, R. J., and Gekakis, N. (2014). A gain-of-function mutation in adenylate cyclase 3 protects mice from diet-induced obesity. *PloS one*, 9(10):e110226.

Prokop, J. W., Lazar, J., Crapitto, G., Smith, D. C., Worthey, E. A., and Jacob, H. J. (2017). Molecular modeling in the age of clinical genomics, the enterprise of the next generation. *Journal of molecular modeling*, 23(3):75.

Purcell, S., Cherny, S. S., and Sham, P. C. (2003). Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics*, 19(1):149–150.

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

Raghavan, C., Mauleon, R., Lacorte, V., Jubay, M., Zaw, H., and Bonifacio, J. (2017). Approaches in Characterizing Genetic Structure and Mapping in a Rice Multiparental Population. *G3*, 7(June):1721–1730.

Raiffa, H. and Schlaifer, R. (2000). *Applied Statistical Decision Theory*. John Wiley & Sons, INC., New York, NY, wiley clas edition.

Ramstein, G. P., Lipka, A. E., Lu, F., Costich, D. E., Cherney, J. H., Buckler, E., and Casler, M. D. (2015). Genome-Wide Association Study Based on Multiple Imputation with Low-Depth Sequencing Data : Application to Biofuel Traits in Reed Canarygrass. *G3*, 5(July):891–909.

Rantala, M. J. and Roff, D. A. (2006). Analysis of the importance of genotypic variation, metabolic rate, morphology, sex and development time on immune function in the cricket, Gryllus firmus. *Journal of Evolutionary Biology*, 19(3):834–843.

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. (2014). Host genetic diversity enables Ebola hemorrhagic fever pathogenesis and resistance. *Science (New York, N.Y.)*, 346(6212):987–91.

Rebai, A. and Goffinet, B. (1993). Power of tests for QTL detection using replicated progenies derived from a diallel cross. *Theoretical and Applied Genetics*, 86(8):1014–1022.

Roberts, A. J., Casal, L., Huitron-Resendiz, S., Thompson, T., and Tarantino, L. M. (2018). Intravenous cocaine self-administration in a panel of inbred mouse strains differing in acute locomotor sensitivity to cocaine. *Psychopharmacology*.

Robsa Shuro, A. (2017). Review Paper on Approaches in Developing Inbred Lines in Cross-Pollinated Crops. *Biochemistry and Molecular Biology*, 2(4):40.

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):435–47.

Roy, A., Kucukural, A., and Zhang, Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction. *Nature protocols*, 5(4):725–38.

Roytman, M., Kichaev, G., Gusev, A., and Pasaniuc, B. (2018). Methods for fine-mapping with chromatin and expression data. *PLoS genetics*, 14(2):e1007240.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392.

Sabourin, J., Nobel, A. B., and Valdar, W. (2015). Fine-mapping additive and dominant SNP effects using group-LASSO and fractional resample model averaging. *Genetic epidemiology*, 39(2):77–88.

Sarkar, S. K. and Chang, C.-K. (1997). The Simes Method for Multiple Hypothesis Testing With Positively Dependent Test Statistics. *Journal of the American Statistical Association*, 92(440):1601.

Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511(7510):421–7.

Sen, S. and Churchill, G. (2001). A statistical framework for quantitative trait mapping. *Genetics*, 159:371–87.

Sen, S., Satagopan, J. M., Broman, K. W., and Churchill, G. A. (2007). R/qtlDesign: inbred line cross experimental design. *Mammalian genome*, 18(2):87–93.

Sen, S., Satagopan, J. M., and Churchill, G. A. (2005). Quantitative trait locus study design from an information perspective. *Genetics*, 170(1):447–64.

Servin, B. and Stephens, M. (2007). Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS genetics*, 3(7):e114.

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.

Simon, M. M., Greenaway, S., White, J. K., Fuchs, H., Gailus-Durner, V., Wells, S., Sorg, T., Wong, K., Bedu, E., Cartwright, E. J., Dacquin, R., Djebali, S., Estabel, J., Graw, J., Ingham, N. J., Jackson, I. J., Lengeling, A., Mandillo, S., Marvel, J., Meziane, H., Preitner, F., Puk, O., Roux, M., Adams, D. J., Atkins, S., Ayadi, A., Becker, L., Blake, A., Brooker, D., Cater, H., Champy, M.-F., Combe, R., Danecek, P., di Fenza, A., Gates, H., Gerdin, A.-K., Golini, E., Hancock, J. M., Hans, W., Hölter, S. M., Hough, T., Jurdic, P., Keane, T. M., Morgan, H., Müller, W., Neff, F., Nicholson, G., Pasche, B., Roberson, L.-A., Rozman, J., Sanderson, M., Santos, L., Selloum, M., Shannon, C., Southwell, A., Tocchini-Valentini, G. P., Vancollie, V. E., Westerberg, H., Wurst, W., Zi, M., Yalcin, B., Ramirez-Solis, R., Steel, K. P., Mallon, A.-M., de Angelis, M. H., Herault, Y., and Brown, S. D. M. (2013). A comparative phenotypic and genomic analysis of C57BL/6J and C57BL/6N mouse strains. *Genome biology*, 14(7):R82.

Smidt, K., Jessen, N., Petersen, A. B., Larsen, A., Magnusson, N., Jeppesen, J. B., Stoltenberg, M., Culvenor, J. G., Tsatsanis, A., Brock, B., Schmitz, O., Wogensen, L., Bush, A. I., and Rungby, J.

(2009). SLC30A3 responds to glucose- and zinc variations in beta-cells and is critical for insulin production and in vivo glucose-metabolism during beta-cell stress. *PloS one*, 4(5):e5684.

Smith, G. D. and Ebrahim, S. (2003). 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International journal of epidemiology*, 32(1):1–22.

Solberg Woods, L. C. (2014). QTL mapping in outbred populations: successes and challenges. *Physiological genomics*, 46(3):81–90.

Solberg Woods, L. C., Holl, K., Tschannen, M., and Valdar, W. (2010). Fine-mapping a locus for glucose tolerance using heterogeneous stock rats. *Physiological genomics*, 41(1):102–8.

Solberg Woods, L. C., Holl, K. L., Oreper, D., Xie, Y., Tsaih, S.-W., and Valdar, W. (2012). Fine-mapping diabetes-related traits, including insulin resistance, in heterogeneous stock rats. *Physiological genomics*, 44(21):1013–26.

Speliotes, E. K., Willer, C. J., Berndt, S. I., Monda, K. L., Thorleifsson, G., Jackson, A. U., Lango Allen, H., Lindgren, C. M., Luan, J., Mägi, R., Randall, J. C., Vedantam, S., Winkler, T. W., Qi, L., Workalemahu, T., Heid, I. M., Steinthorsdottir, V., Stringham, H. M., Weedon, M. N., Wheeler, E., Wood, A. R., Ferreira, T., Weyant, R. J., Segrè, A. V., Estrada, K., Liang, L., Nemesh, J., Park, J.-H., Gustafsson, S., Kilpeläinen, T. O., Yang, J., Bouatia-Naji, N., Esko, T., Feitosa, M. F., Kutalik, Z., Mangino, M., Raychaudhuri, S., Scherag, A., Smith, A. V., Welch, R., Zhao, J. H., Aben, K. K., Absher, D. M., Amin, N., Dixon, A. L., Fisher, E., Glazer, N. L., Goddard, M. E., Heard-Costa, N. L., Hoesel, V., Hottenga, J.-J., Johansson, A., Johnson, T., Ketkar, S., Lamina, C., Li, S., Moffatt, M. F., Myers, R. H., Narisu, N., Perry, J. R. B., Peters, M. J., Preuss, M., Ripatti, S., Rivadeneira, F., Sandholt, C., Scott, L. J., Timpson, N. J., Tyrer, J. P., van Wingerden, S., Watanabe, R. M., White, C. C., Wiklund, F., Barlassina, C., Chasman, D. I., Cooper, M. N., Jansson, J.-O., Lawrence, R. W., Pellikka, N., Prokopenko, I., Shi, J., Thiering, E., Alavere, H., Alibrandi, M. T. S., Almgren, P., Arnold, A. M., Aspelund, T., Atwood, L. D., Balkau, B., Balmforth, A. J., Bennett, A. J., Ben-Shlomo, Y., Bergman, R. N., Bergmann, S., Biebermann, H., Blakemore, A. I. F., Boes, T., Bonnycastle, L. L., Bornstein, S. R., Brown, M. J., Buchanan, T. A., Busonero, F., Campbell, H., Cappuccio, F. P., Cavalcanti-Proença, C., Chen, Y.-D. I., Chen, C.-M., Chines, P. S., Clarke, R., Coin, L., Connell, J., Day, I. N. M., den Heijer, M., Duan, J., Ebrahim, S., Elliott, P., Elosua, R., Eiriksdottir, G., Erdos, M. R., Eriksson, J. G., Facheris, M. F., Felix, S. B., Fischer-Posovszky, P., Folsom, A. R., Friedrich, N., Freimer, N. B., Fu, M., Gaget, S., Gejman, P. V., Geus, E. J. C., Gieger, C., Gjesing, A. P., Goel, A., Goyette, P., Grallert, H., Grässler, J., Greenawalt, D. M., Groves, C. J., Gudnason, V., Guiducci, C., Hartikainen, A.-L., Hassanali, N., Hall, A. S., Havulinna, A. S., Hayward, C., Heath, A. C., Hengstenberg, C., Hicks, A. A., Hinney, A., Hofman, A., Homuth, G., Hui, J., Igl, W., Iribarren, C., Isomaa, B., Jacobs, K. B., Jarick, I., Jewell, E., John, U., Jørgensen, T., Jousilahti, P., Jula, A., Kaakinen, M., Kajantie, E., Kaplan, L. M., Kathiresan, S., Kettunen, J., Kinnunen, L., Knowles, J. W., Kolcic, I., König, I. R., Koskinen, S., Kovacs, P., Kuusisto, J., Kraft, P., Kvaløy, K., Laitinen, J., Lantieri, O., Lanzani, C., Launer, L. J., Lecoeur, C., Lehtimäki, T., Lettre, G., Liu, J., Lokki, M.-L., Lorentzon, M., Luben, R. N., Ludwig, B., MAGIC, Manunta, P., Marek, D., Marre, M., Martin, N. G., McArdle, W. L., McCarthy, A., McKnight, B., Meitinger, T., Melander, O., Meyre, D., Midthjell, K., Montgomery, G. W., Morken, M. A., Morris, A. P., Mulic, R., Ngwa, J. S., Nelis, M., Neville, M. J., Nyholt, D. R., O'Donnell, C. J., O'Rahilly, S., Ong, K. K., Oostra, B., Paré, G., Parker, A. N., Perola, M., Pichler, I., Pietiläinen, K. H., Platou, C. G. P., Polasek, O., Pouta, A., Rafelt, S., Raitakari, O., Rayner, N. W., Ridderstråle, M., Rief, W., Ruokonen, A., Robertson, N. R., Rzehak, P., Salomaa,

V., Sanders, A. R., Sandhu, M. S., Sanna, S., Saramies, J., Savolainen, M. J., Scherag, S., Schipf, S., Schreiber, S., Schunkert, H., Silander, K., Sinisalo, J., Siscovick, D. S., Smit, J. H., Soranzo, N., Sovio, U., Stephens, J., Surakka, I., Swift, A. J., Tammesoo, M.-L., Tardif, J.-C., Teder-Laving, M., Teslovich, T. M., Thompson, J. R., Thomson, B., Tönjes, A., Tuomi, T., van Meurs, J. B. J., van Ommen, G.-J., Vatin, V., Viikari, J., Visvikis-Siest, S., Vitart, V., Vogel, C. I. G., Voight, B. F., Waite, L. L., Wallaschofski, H., Walters, G. B., Widen, E., Wiegand, S., Wild, S. H., Willemsen, G., Witte, D. R., Witteman, J. C., Xu, J., Zhang, Q., Zgaga, L., Ziegler, A., Zitting, P., Beilby, J. P., Farooqi, I. S., Hebebrand, J., Huikuri, H. V., James, A. L., Kähönen, M., Levinson, D. F., Macciardi, F., Nieminen, M. S., Ohlsson, C., Palmer, L. J., Ridker, P. M., Stumvoll, M., Beckmann, J. S., Boeing, H., Boerwinkle, E., Boomsma, D. I., Caulfield, M. J., Chanock, S. J., Collins, F. S., Cupples, L. A., Smith, G. D., Erdmann, J., Froguel, P., Grönberg, H., Gyllensten, U., Hall, P., Hansen, T., Harris, T. B., Hattersley, A. T., Hayes, R. B., Heinrich, J., Hu, F. B., Hveem, K., Illig, T., Jarvelin, M.-R., Kaprio, J., Karpe, F., Khaw, K.-T., Kiemeney, L. A., Krude, H., Laakso, M., Lawlor, D. A., Metspalu, A., Munroe, P. B., Ouwehand, W. H., Pedersen, O., Penninx, B. W., Peters, A., Pramstaller, P. P., Quertermous, T., Reinehr, T., Rissanen, A., Rudan, I., Samani, N. J., Schwarz, P. E. H., Shuldiner, A. R., Spector, T. D., Tuomilehto, J., Uda, M., Uitterlinden, A., Valle, T. T., Wabitsch, M., Waeber, G., Wareham, N. J., Watkins, H., Procardis Consortium, Wilson, J. F., Wright, A. F., Zillikens, M. C., Chatterjee, N., McCarroll, S. A., Purcell, S., Schadt, E. E., Visscher, P. M., Assimes, T. L., Borecki, I. B., Deloukas, P., Fox, C. S., Groop, L. C., Haritunians, T., Hunter, D. J., Kaplan, R. C., Mohlke, K. L., O'Connell, J. R., Peltonen, L., Schlessinger, D., Strachan, D. P., van Duijn, C. M., Wichmann, H.-E., Frayling, T. M., Thorsteinsdottir, U., Abecasis, G. R., Barroso, I., Boehnke, M., Stefansson, K., North, K. E., McCarthy, M. I., Hirschhorn, J. N., Ingelsson, E., and Loos, R. J. F. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature genetics*, 42(11):937–48.

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.

Stanley, P. D., Ng'oma, E., O'Day, S., and King, E. G. (2017). Genetic dissection of nutrition-induced plasticity in insulin/insulin-like growth factor signaling and median life span in a Drosophila multiparent population. *Genetics*, 206(2):587–602.

STAR Consortium, Saar, K., Beck, A., Bihoreau, M.-T., Birney, E., Brocklebank, D., Chen, Y., Cuppen, E., Demonchy, S., Dopazo, J., Flicek, P., Foglio, M., Fujiyama, A., Gut, I. G., Gauguier, D., Guigo, R., Guryev, V., Heinig, M., Hummel, O., Jahn, N., Klages, S., Kren, V., Kube, M., Kuhl, H., Kuramoto, T., Kuroki, Y., Lechner, D., Lee, Y.-A., Lopez-Bigas, N., Lathrop, G. M., Mashimo, T., Medina, I., Mott, R., Patone, G., Perrier-Cornet, J.-A., Platzer, M., Pravenec, M., Reinhardt, R., Sakaki, Y., Schilhabel, M., Schulz, H., Serikawa, T., Shikhagaie, M., Tatsumoto, S., Taudien, S., Toyoda, A., Voigt, B., Zelenika, D., Zimdahl, H., and Hubner, N. (2008). SNP and haplotype mapping for genetic analysis in the rat. *Nature genetics*, 40(5):560–6.

Stergiakouli, E., Gaillard, R., Tavaré, J. M., Balthasar, N., Loos, R. J., Taal, H. R., Evans, D. M., Rivadeneira, F., St Pourcain, B., Uitterlinden, A. G., Kemp, J. P., Hofman, A., Ring, S. M., Cole, T. J., Jaddoe, V. W. V., Davey Smith, G., and Timpson, N. J. (2014). Genome-wide association study of height-adjusted BMI in childhood identifies functional variant in ADCY3. *Obesity (Silver Spring, Md.)*, 22(10):2252–9.

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, 100(16):9440–5.

Stunkard, A. J., Foch, T. T., and Hrubec, Z. (1986). A twin study of human obesity. *JAMA*, 256(1):51–4.

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):437–47.

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

Teslovich, T. M., Musunuru, K., Smith, A. V., Edmondson, A. C., Stylianou, I. M., Koseki, M., Pirruccello, J. P., Ripatti, S., Chasman, D. I., Willer, C. J., Johansen, C. T., Fouchier, S. W., Isaacs, A., Peloso, G. M., Barbalic, M., Ricketts, S. L., Bis, J. C., Aulchenko, Y. S., Thorleifsson, G., Feitosa, M. F., Chambers, J., Orho-Melander, M., Melander, O., Johnson, T., Li, X., Guo, X., Li, M., Shin Cho, Y., Jin Go, M., Jin Kim, Y., Lee, J.-Y., Park, T., Kim, K., Sim, X., Twee-Hee Ong, R., Croteau-Chonka, D. C., Lange, L. A., Smith, J. D., Song, K., Hua Zhao, J., Yuan, X., Luan, J., Lamina, C., Ziegler, A., Zhang, W., Zee, R. Y. L., Wright, A. F., Witteman, J. C. M., Wilson, J. F., Willemsen, G., Wichmann, H.-E., Whitfield, J. B., Waterworth, D. M., Wareham, N. J., Waeber, G., Vollenweider, P., Voight, B. F., Vitart, V., Uitterlinden, A. G., Uda, M., Tuomilehto, J., Thompson, J. R., Tanaka, T., Surakka, I., Stringham, H. M., Spector, T. D., Soranzo, N., Smit, J. H., Sinisalo, J., Silander, K., Sijbrands, E. J. G., Scuteri, A., Scott, J., Schlessinger, D., Sanna, S., Salomaa, V., Saharinen, J., Sabatti, C., Ruokonen, A., Rudan, I., Rose, L. M., Roberts, R., Rieder, M., Psaty, B. M., Pramstaller, P. P., Pichler, I., Perola, M., Penninx, B. W. J. H., Pedersen, N. L., Pattaro, C., Parker, A. N., Pare, G., Oostra, B. A., O'Donnell, C. J., Nieminen, M. S., Nickerson, D. A., Montgomery, G. W., Meitinger, T., McPherson, R., McCarthy, M. I., McArdle, W., Masson, D., Martin, N. G., Marroni, F., Mangino, M., Magnusson, P. K. E., Lucas, G., Luben, R., Loos, R. J. F., Lokki, M.-L., Lettre, G., Langenberg, C., Launer, L. J., Lakatta, E. G., Laaksonen, R., Kyvik, K. O., Kronenberg, F., König, I. R., Khaw, K.-T., Kaprio, J., Kaplan, L. M., Johansson, A., Jarvelin, M.-R., Janssens, A. C. J. W., Ingelsson, E., Igl, W., Kees Hovingh, G., Hottenga, J.-J., Hofman, A., Hicks, A. A., Hengstenberg, C., Heid, I. M., Hayward, C., Havulinna, A. S., Hastie, N. D., Harris, T. B., Haritunians, T., Hall, A. S., Gyllensten, U., Guiducci, C., Groop, L. C., Gonzalez, E., Gieger, C., Freimer, N. B., Ferrucci, L., Erdmann, J., Elliott, P., Ejebe, K. G., Döring, A., Dominiczak, A. F., Demissie, S., Deloukas, P., de Geus, E. J. C., de Faire, U., Crawford, G., Collins, F. S., Chen, Y.-d. I., Caulfield, M. J., Campbell, H., Burtt, N. P., Bonnycastle, L. L., Boomsma, D. I., Boekholdt, S. M., Bergman, R. N., Barroso, I., Bandinelli, S., Ballantyne, C. M., Assimes, T. L., Quertermous, T., Altshuler, D., Seielstad, M., Wong, T. Y., Tai, E.-S., Feranil, A. B., Kuzawa, C. W., Adair, L. S., Taylor, H. A., Borecki, I. B., Gabriel, S. B., Wilson, J. G., Holm, H., Thorsteinsdottir, U., Gudnason, V., Krauss, R. M., Mohlke, K. L., Ordovas, J. M., Munroe, P. B., Kooner, J. S., Tall, A. R., Hegele, R. A., Kastelein, J. J. P., Schadt, E. E., Rotter, J. I., Boerwinkle, E., Strachan, D. P., Mooser, V., Stefansson, K., Reilly, M. P., Samani, N. J., Schunkert, H., Cupples, L. A., Sandhu, M. S., Ridker, P. M., Rader, D. J., van Duijn, C. M., Peltonen, L., Abecasis, G. R., Boehnke, M., and Kathiresan, S. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, 466(7307):707–13.

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 : official journal of the International Mammalian Genome Society*, 13(4):175–8.

Tisné, S., Pomiès, V., Riou, V., Syahputra, I., and Cochard, B. (2017). Identification of Ganoderma Disease Resistance Loci Using Natural Field Infection of an Oil Palm Multiparental Population. *G3*, 7(June):1683–1692.

Tong, T., Shen, Y., Lee, H.-W., Yu, R., and Park, T. (2016). Adenylyl cyclase 3 haploinsufficiency confers susceptibility to diet-induced obesity and insulin resistance in mice. *Scientific reports*, 6:34179.

Tsaih, S.-W., Holl, K., Jia, S., Kaldunski, M., Tschannen, M., He, H., Andrae, J. W., Li, S.-H., Stoddard, A., Wiederhold, A., Parrington, J., Ruas da Silva, M., Galione, A., Meigs, J., Meta-Analyses of Glucose and Insulin-Related Traits Consortium (MAGIC) Investigators, Hoffmann, R. G., Simpson, P., Jacob, H., Hessner, M., and Solberg Woods, L. C. (2014). Identification of a novel gene for diabetic traits in rats, mice, and humans. *Genetics*, 198(1):17–29.

Tsaih, S.-W., Lu, L., Airey, D. C., Williams, R. W., and Churchill, G. A. (2005). Quantitative trait mapping in a diallel cross of recombinant inbred lines. *Mammalian Genome*, 16(5):344–355.

Turner, S. D., Maurizio, P. L., Valdar, W., Yandell, B. S., and Simon, P. W. (2018). Dissecting the Genetic Architecture of Shoot Growth in Carrot (Daucus carota L.) Using a Diallel Mating Design. *G3 (Bethesda, Md.)*, 8(2):411–426.

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):1783–97.

Valdar, W., Holmes, C. C., Mott, R., and Flint, J. (2009). Mapping in structured populations by resample model averaging. *Genetics*, 182(4):1263–77.

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.

Valdar, W., Solberg, L. C., Gauguier, D., Cookson, W. O., Rawlins, J. N. P., Mott, R., and Flint, J. (2006c). Genetic and environmental effects on complex traits in mice. *Genetics*, 174(2):959–84.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

Verbyla, A. P., George, A. W., Cavanagh, C. R., and Verbyla, K. L. (2014). Whole-genome QTL analysis for MAGIC. *Theoretical and Applied Genetics*, 127(8):1753–1770.

Verhoeven, K. J. F., Jannink, J.-L., and McIntyre, L. M. (2006). Using mating designs to uncover QTL and the genetic architecture of complex traits. *Heredity*, 96(2):139–49.

Wang, Y. C., McPherson, K., Marsh, T., Gortmaker, S. L., and Brown, M. (2011). Health and economic burden of the projected obesity trends in the USA and the UK. *Lancet (London, England)*, 378(9793):815–25.

Wang, Z., Li, V., Chan, G. C. K., Phan, T., Nudelman, A. S., Xia, Z., and Storm, D. R. (2009). Adult type 3 adenylyl cyclase-deficient mice are obese. *PloS one*, 4(9):e6979.

Wei, J. and Xu, S. (2016). A random-model approach to QTL mapping in multiparent advanced generation intercross (MAGIC) populations. *Genetics*, 202(2):471–486.

Wen, W., Cho, Y.-S., Zheng, W., Dorajoo, R., Kato, N., Qi, L., Chen, C.-H., Delahanty, R. J., Okada, Y., Tabara, Y., Gu, D., Zhu, D., Haiman, C. A., Mo, Z., Gao, Y.-T., Saw, S.-M., Go, M.-J., Takeuchi, F., Chang, L.-C., Kokubo, Y., Liang, J., Hao, M., Le Marchand, L., Zhang, Y., Hu, Y., Wong, T.-Y., Long, J., Han, B.-G., Kubo, M., Yamamoto, K., Su, M.-H., Miki, T., Henderson, B. E., Song, H., Tan, A., He, J., Ng, D. P.-K., Cai, Q., Tsunoda, T., Tsai, F.-J., Iwai, N., Chen, G. K., Shi, J., Xu, J., Sim, X., Xiang, Y.-B., Maeda, S., Ong, R. T. H., Li, C., Nakamura, Y., Aung, T., Kamatani, N., Liu, J.-J., Lu, W., Yokota, M., Seielstad, M., Fann, C. S. J., Genetic Investigation of ANthropometric Traits (GIANT) Consortium, Wu, J.-Y., Lee, J.-Y., Hu, F. B., Tanaka, T., Tai, E. S., and Shu, X.-O. (2012). Meta-analysis identifies common variants associated with body mass index in east Asians. *Nature genetics*, 44(3):307–11.

Williams, R. W. and Williams, E. G. (2017). Resources for systems genetics. In Schughart, K. and Williams, R. W., editors, *Systems Genetics: Methods and Protocols*, pages 3–29. Springer New York, New York, NY.

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

Zhang, Q., Arveiler, D., Bakker, S. J. L., Beilby, J., Bergman, R. N., Bergmann, S., Biffar, R., Blangero, J., Boomsma, D. I., Bornstein, S. R., Bovet, P., Brambilla, P., Brown, M. J., Campbell, H., Caulfield, M. J., Chakravarti, A., Collins, R., Collins, F. S., Crawford, D. C., Cupples, L. A., Danesh, J., de Faire, U., den Ruijter, H. M., Erbel, R., Erdmann, J., Eriksson, J. G., Farrall, M., Ferrannini, E., Ferrières, J., Ford, I., Forouhi, N. G., Forrester, T., Gansevoort, R. T., Gejman, P. V., Gieger, C., Golay, A., Gottesman, O., Gudnason, V., Gyllensten, U., Haas, D. W., Hall, A. S., Harris, T. B., Hattersley, A. T., Heath, A. C., Hengstenberg, C., Hicks, A. A., Hindorff, L. A., Hingorani, A. D., Hofman, A., Hovingh, G. K., Humphries, S. E., Hunt, S. C., Hypponen, E., Jacobs, K. B., Jarvelin, M.-R., Jousilahti, P., Jula, A. M., Kaprio, J., Kastelein, J. J. P., Kayser, M., Kee, F., Keinanen-Kiukaanniemi, S. M., Kiemeney, L. A., Kooner, J. S., Kooperberg, C., Koskinen, S., Kovacs, P., Kraja, A. T., Kumari, M., Kuusisto, J., Lakka, T. A., Langenberg, C., Le Marchand, L., Lehtimäki, T., Lupoli, S., Madden, P. A. F., Männistö, S., Manunta, P., Marette, A., Matise, T. C., McKnight, B., Meitinger, T., Moll, F. L., Montgomery, G. W., Morris, A. D., Morris, A. P., Murray, J. C., Nelis, M., Ohlsson, C., Oldehinkel, A. J., Ong, K. K., Ouwehand, W. H., Pasterkamp, G., Peters, A., Pramstaller, P. P., Price, J. F., Qi, L., Raitakari, O. T., Rankinen, T., Rao, D. C., Rice, T. K., Ritchie, M., Rudan, I., Salomaa, V., Samani, N. J., Saramies, J., Sarzynski, M. A., Schwarz, P. E. H., Sebert, S., Sever, P., Shuldiner, A. R., Sinisalo, J., Steinthorsdottir, V., Stolk, R. P., Tardif, J.-C., Tönjes, A., Tremblay, A., Tremoli, E., Virtamo, J., Vohl, M.-C., Electronic Medical Records and Genomics (eMEMERGEGE) Consortium, MIGen Consortium, PAGEGE Consortium, LifeLines Cohort Study, Amouyel, P., Asselbergs, F. W., Assimes, T. L., Bochud, M., Boehm, B. O., Boerwinkle, E., Bottinger, E. P., Bouchard, C., Cauchi, S., Chambers, J. C., Chanock, S. J., Cooper, R. S., de Bakker, P. I. W., Dedoussis, G., Ferrucci, L., Franks, P. W., Froguel, P., Groop, L. C., Haiman, C. A., Hamsten, A., Hayes, M. G., Hui, J., Hunter, D. J., Hveem, K., Jukema, J. W., Kaplan, R. C., Kivimaki, M., Kuh, D., Laakso, M., Liu, Y., Martin, N. G., März, W., Melbye, M., Moebus, S., Munroe, P. B., Njølstad, I., Oostra, B. A., Palmer, C. N. A., Pedersen, N. L., Perola, M., Pérusse, L., Peters, U., Powell, J. E., Power, C., Quertermous, T., Rauramaa, R., Reinmaa, E., Ridker, P. M., Rivadeneira, F., Rotter, J. I., Saaristo, T. E., Saleheen, D., Schlessinger, D., Slagboom, P. E., Snieder, H., Spector, T. D., Strauch, K., Stumvoll, M., Tuomilehto, J., Uusitupa, M., van der Harst, P., Völzke, H., Walker, M., Wareham, N. J., Watkins, H., Wichmann, H.-E., Wilson, J. F., Zanen, P., Deloukas, P., Heid, I. M., Lindgren, C. M., Mohlke, K. L., Speliotes, E. K., Thorsteinsdottir, U., Barroso, I., Fox, C. S., North, K. E., Strachan, D. P., Beckmann, J. S., Berndt, S. I., Boehnke, M., Borecki, I. B., McCarthy, M. I., Metspalu, A., Stefansson, K., Uitterlinden, A. G., van Duijn, C. M., Franke, L., Willer, C. J., Price, A. L., Lettre, G., Loos, R. J. F., Weedon, M. N., Ingelsson, E., O'Connell, J. R., Abecasis, G. R., Chasman, D. I., Goddard, M. E., Visscher, P. M., Hirschhorn, J. N., and Frayling, T. M. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature genetics*, 46(11):1173–86.

Wu, L., Shen, C., Seed Ahmed, M., Östenson, C.-G., and Gu, H. F. (2016). Adenylate cyclase 3: a new target for anti-obesity drug development. *Obesity reviews : an official journal of the International Association for the Study of Obesity*, 17(9):907–14.

Wu, Y., Zeng, J., Zhang, F., Zhu, Z., Qi, T., Zheng, Z., Lloyd-Jones, L. R., Marioni, R. E., Martin, N. G., Montgomery, G. W., Deary, I. J., Wray, N. R., Visscher, P. M., McRae, A. F., and Yang, J. (2018). Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits. *Nature Communications*, 9(1):918.

Xu, S. (2003). Estimating polygenic effects using markers of the entire genome. *Genetics*, 163(2):789–801.

Yalcin, B., Flint, J., and Mott, R. (2005). Using progenitor strain information to identify quantitative trait nucleotides in outbred mice. *Genetics*, 171(2):673–81.

Yamamoto, A. H. (1994). Diallel analysis of temperature preference in Drosophila immigrans. *Idengaku zasshi*, 69(1):77–86.

Yang, H., Bell, T. A., Churchill, G. A., and Pardo-Manuel de Villena, F. (2007). On the subspecific origin of the laboratory mouse. *Nature genetics*, 39(9):1100–7.

Yang, H., Wang, J. R., Didion, J. P., Buus, R. J., Bell, T. A., Welsh, C. E., Bonhomme, F., Yu, A. H.-T., Nachman, M. W., Pialek, J., Tucker, P., Boursot, P., McMillan, L., Churchill, G. A., and de Villena, F. P.-M. (2011). Subspecific origin and haplotype diversity in the laboratory mouse. *Nature genetics*, 43(7):648–55.

Yi, N. and Xu, S. (2008). Bayesian LASSO for quantitative trait loci mapping. *Genetics*, 179(2):1045–1055.

Zhang, Z., Wang, W., and Valdar, W. (2014). Bayesian modeling of haplotype effects in multiparent populations. *Genetics*, 198(1):139–56.

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

Zheng, J., Li, Y., Abecasis, G. R., and Scheet, P. (2011). A Comparison of Approaches to Account for Uncertainty in Analysis of Imputed Genotypes. *Genetic Epidemiology*, 35(2):102–110.

Zhu, J. and Weir, B. S. (1996). Mixed model approaches for diallel analysis based on a bio-model. *Genetical research*, 68(3):233–40.

Zou, F., Gelfond, J. A. L., Airey, D. C., Lu, L., Manly, K. F., Williams, R. W., and Threadgill, D. W. (2005). Quantitative trait locus analysis using recombinant inbred intercrosses: theoretical and empirical considerations. *Genetics*, 170(3):1299–311.

Zou, F., Sun, W., Crowley, J. J., Zhabotynsky, V., Sullivan, P. F., and Pardo-Manuel de Villena, F. (2014). A Novel Statistical Approach for Jointly Analyzing RNA-Seq Data from F1 Reciprocal Crosses and Inbred Lines. *Genetics*, 197(1):389–99.

Zou, F., Xu, Z., and Vision, T. (2006). Assessing the significance of quantitative trait loci in replicable mapping populations. *Genetics*, 174(2):1063–8.