Title
Computational Genetic Approaches for Understanding the Genetic Basis of Complex Traits

Permalink
https://escholarship.org/uc/item/50j7w041

Author
Kang, Eun Yong

Publication Date
2013

Peer reviewed|Thesis/dissertation
Computational Genetic Approaches for Understanding the Genetic Basis of Complex Traits

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Computer Science

by

Eun Yong Kang

2013
ABSTRACT OF THE DISSERTATION

Computational Genetic Approaches for Understanding the Genetic Basis of Complex Traits

by

Eun Yong Kang

Doctor of Philosophy in Computer Science

University of California, Los Angeles, 2013

Professor Eleazar Eskin, Chair

Recent advances in genotyping and sequencing technology have enabled researchers to collect an enormous amount of high-dimensional genotype data. These large scale genomic data provide unprecedented opportunity for researchers to study and analyze the genetic factors of human complex traits. One of the major challenges in analyzing these high-throughput genomic data is requirements for effective and efficient computational methodologies. In this thesis, I introduce several methodologies for analyzing these genomic data which facilitates our understanding of the genetic basis of complex human traits. First, I introduce a method for inferring biological networks from high-throughput data containing both genetic variation information and gene expression profiles from genetically distinct strains of an organism. For this problem, I use causal inference techniques to infer the presence or absence of causal relationships between yeast gene expressions in the framework of graphical causal models. In particular, I utilize prior biological knowledge that genetic variations affect gene expressions, but not vice versa, which allow us to direct the subsequent edges between two gene expression levels. The prediction of a presence of causal relationship as well as the absence of causal relationship between gene expressions can facilitate distinguishing between
direct and indirect effects of variation on gene expression levels. I demonstrate the utility of our approach by applying it to data set containing 112 yeast strains and the proposed method identifies the known “regulatory hotspot” in yeast. Second, I introduce efficient pairwise identity by descent (IBD) association mapping method, which utilizes importance sampling to improve efficiency and enables approximation of extremely small p-values. Two individuals are IBD at a locus if they have identical alleles inherited from a common ancestor. One popular approach to find the association between IBD status and disease phenotype is the pairwise method where one compares the IBD rate of case/case pairs to the background IBD rate to detect excessive IBD sharing between cases. One challenge of the pairwise method is computational efficiency. In the pairwise method, one uses permutation to approximate p-values because it is difficult to analytically obtain the asymptotic distribution of the statistic. Since the p-value threshold for genome-wide association studies (GWAS) is necessarily low due to multiple testing, one must perform a large number of permutations which can be computationally demanding. I present Fast-Pairwise to overcome the computational challenges of the traditional pairwise method by utilizing importance sampling to improve efficiency and enable approximation of extremely small p-values. Using the WTCCC type 1 diabetes data, I show that Fast-Pairwise can successfully pinpoint a gene known to be associated to the disease within the MHC region. Finally, I introduce a novel meta analytic approach to identify gene-by-environment interactions by aggregating the multiple studies with varying environmental conditions. Identifying environmentally specific genetic effects is a key challenge in understanding the structure of complex traits. Model organisms play a crucial role in the identification of such gene-by-environment interactions, as a result of the unique ability to observe genetically similar individuals across multiple distinct environments. Many model organism studies examine the same traits but, under varying environmental conditions. These studies when examined in aggregate provide an opportunity to identify genomic loci
exhibiting environmentally-dependent effects. In this project, I jointly analyze multiple studies with varying environmental conditions using a meta-analytic approach based on a random effects model to identify loci involved in gene-by-environment interactions. Our approach is motivated by the observation that methods for discovering gene-by-environment interactions are closely related to random effects models for meta-analysis. We show that interactions can be interpreted as heterogeneity and can be detected without utilizing the traditional uni- or multi-variate approaches for discovery of gene-by-environment interactions. I apply our new method to combine 17 mouse studies containing in aggregate 4,965 distinct animals. We identify 26 significant loci involved in High-density lipoprotein (HDL) cholesterol, many of which show significant evidence of involvement in gene-by-environment interactions.
The dissertation of Eun Yong Kang is approved.

Adnan Darwiche
Christopher J. Lee
David Earl Heckerman
Aldons J. Lusis
Eleazar Eskin, Committee Chair

University of California, Los Angeles
2013
To my wife Yun Jin, my parents Jai Soo Kang and Su Hee Kim and my two sons
Aaron and Luke
# Table of Contents

1 Introduction ................................................................. 1

2 Detecting the Presence and Absence of Causal Relationships Between Expression of Yeast Genes with Very Few Samples ........................................ 10
   2.1 Background .......................................................... 10
   2.2 Methodology ......................................................... 16
      2.2.1 Causal Graphs for Genetical Genomics ................. 16
      2.2.2 Inference Algorithm Overview ......................... 20
      2.2.3 Finding potential causal SNPs ......................... 20
      2.2.4 Finding causal relationships between genes .......... 21
      2.2.5 Distinguishing between direct and indirect effects of variation .......... 24
      2.2.6 Excluding causal relationships between genes ......... 25
   2.3 Results ............................................................ 27
   2.4 Discussion .......................................................... 33
   2.5 Figures ............................................................ 35
   2.6 Tables ............................................................. 39

3 Respecting Markov Equivalence in Computing Posterior Probabilities of Causal Graphical Features ................................................................. 43
   3.1 Introduction ........................................................ 43
      3.1.1 Background .................................................... 43
      3.1.2 Previous Approaches and Their Limitations ............ 44
3.1.3 Proposed Approach .............................................. 47
3.2 Posterior Computation of Graphical Features ..................... 48
3.3 Previously Proposed Method ....................................... 50
3.4 Novel Proposed Method ............................................. 53
3.4.1 Novel DP Algorithm for Computing Posteriors of All Adjacency Features ............................................. 54
3.4.2 Novel DP Algorithm for Computing Posteriors of All V-Structure Features ............................................. 54
3.5 Experiments .......................................................... 56
3.6 Conclusion ........................................................... 59

4 Increasing Association Mapping Power and Resolution in Mouse Genetic Studies through the Use of Meta-analysis for Structured Populations .... 61
4.1 INTRODUCTION ..................................................... 61
4.2 MATERIALS AND METHODS ...................................... 64
4.2.1 Association Studies ............................................. 64
4.2.2 Association Studies in Structured Populations ................. 66
4.2.3 Meta-Analysis in Structured Populations ....................... 67
4.2.4 Mouse Association Data ....................................... 69
4.3 RESULTS ............................................................ 70
4.3.1 Combining the HMDP with an F2 cross increases power ... 70
4.3.2 Meta-analysis leads to an increase in resolution over HMDP and F2 mapping ............................................. 71
4.3.3 Application to Bone Mineral Density ........................... 72
5 Fast Pairwise IBD Association Testing in Genome-wide Association Studies

5.2 Methods

5.2.1 IBD graph

5.2.2 Pairwise methods for IBD association mapping

5.2.3 Permutation test

5.2.4 Fast-Pairwise

5.2.5 IBD-degreetype

5.2.6 Substitution strategy

5.2.7 Sampling with replacement

5.2.8 Sampling without replacement

5.2.9 P-value calculation

5.2.10 Adjusting for population structure

5.3 Results

5.3.1 Efficiency

5.3.2 Accuracy

5.3.3 Application to WTCCC type 1 diabetes data

5.4 Discussion
## 6 Meta-analysis identifies Gene-by-environment Interactions as Demonstrated in a Study of 4,965 Mice

<table>
<thead>
<tr>
<th>Sections</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1 Background</td>
</tr>
<tr>
<td>6.2 Results</td>
</tr>
<tr>
<td>6.2.1 Discovering environmentally-specific loci using meta-analysis</td>
</tr>
<tr>
<td>6.2.2 The connection between random effects meta-analysis and gene-by-environment interactions</td>
</tr>
<tr>
<td>6.2.3 Gene-by-environment interactions are prevalent in mouse association studies</td>
</tr>
<tr>
<td>6.2.4 Power of meta-analysis for detecting gene-by-environment interactions</td>
</tr>
<tr>
<td>6.2.5 Application to 17 mouse HDL studies</td>
</tr>
<tr>
<td>6.3 Discussion</td>
</tr>
<tr>
<td>6.4 Methods and Materials</td>
</tr>
<tr>
<td>6.4.1 Standard study design for testing gene-by-environment interactions</td>
</tr>
<tr>
<td>6.4.2 Multivariate interactions model</td>
</tr>
<tr>
<td>6.4.3 Standard meta-analysis approach</td>
</tr>
<tr>
<td>6.4.4 Relation between gene-by-environment interactions and meta-analysis</td>
</tr>
<tr>
<td>6.4.5 Controlling for population structure within studies</td>
</tr>
<tr>
<td>6.4.6 Meta-analysis of studies with population structure</td>
</tr>
<tr>
<td>Section</td>
</tr>
<tr>
<td>---------</td>
</tr>
<tr>
<td>6.4.7</td>
</tr>
<tr>
<td>6.5</td>
</tr>
<tr>
<td>6.6</td>
</tr>
<tr>
<td>6.7</td>
</tr>
<tr>
<td>6.8</td>
</tr>
<tr>
<td>6.8.1</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>6.8.2</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>6.8.3</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>7.1</td>
</tr>
<tr>
<td>7.2</td>
</tr>
<tr>
<td>7.2.1</td>
</tr>
<tr>
<td>7.2.2</td>
</tr>
<tr>
<td>7.2.3</td>
</tr>
<tr>
<td>7.3</td>
</tr>
<tr>
<td>7.4</td>
</tr>
<tr>
<td>7.4.1</td>
</tr>
<tr>
<td>7.5</td>
</tr>
<tr>
<td>7.5.1</td>
</tr>
<tr>
<td>7.5.2</td>
</tr>
</tbody>
</table>
### 8 A Novel Meta-analysis Approach for Genome-wide Association Studies with Sex-specific Effects

#### 8.1 Background

- Page 206

#### 8.2 Results

1. Applying meta-analysis to handle sex-specific effects... Page 208
2. The presence of background GxS interactions explains the advantage of the sex-differentiated meta-analysis approach... Page 211
3. Statistical power comparison between FE, UFE, GWAMA, WG-WAMA, SST and RE2 approaches for Sex-differentiated meta-analysis... Page 214

#### 8.3 Discussion

- Page 219

#### 8.4 Methods

1. Fixed effects model meta analysis for sex-differentiated analysis... Page 220
2. Unweighted fixed effects model meta analysis for sex-differentiated analysis... Page 220
3. SST approach for sex-differentiated analysis... Page 221
4. GWAMA approach for sex-differentiated analysis... Page 221
5. Weighted GWAMA approach for sex-differentiated analysis... Page 222
6. Meta-sex approach for sex-differentiated analysis... Page 222
7. Controlling for population structure within studies... Page 223
8. Computation of statistical power of the methods... Page 225
9. Simulating GxS interactions... Page 226
10. Computing heritability of GxS interactions... Page 227
LIST OF FIGURES

2.1 Complete causal network in yeast with the nine regulatory hotspots colored. Circles designate regulators, squares designate targets and diamonds designate genes that are both regulators and targets. The spring-embedded view of the causal network shows that some hotspots, hotspots 2 (red), 3 (bright green), and 9 (light blue), overlap well with the hub like structures of the network where regulators are positioned in the middle and targets surround the causal hub. 36

2.2 Complete causal network in yeast with the nine regulatory hotspots colored. Circles designate regulators, squares designate targets and diamonds designate genes that are both regulators and targets. Causal network grouped by hotspot shows that some regulators and targets (indicated by gray) are not part of known regulatory hotspots. 37

2.3 Possible causal graphs relating a triplet considering a SNP $S$ with the level of gene expression for genes $E_i$ and $E_j$. Bidirected edges denote hidden common causes. (a) Nine possible causal models consistent with $S$ being a causal ancestor of $E_i$ and $E_j$ (models $H_4$ through $H_9$ are indistinguishable from observations of the triplet). (b) Four possible causal models consistent with $S$ being a causal ancestor of $E_i$ while being uncorrelated with $E_j$. 38

3.1 Three graphical models from a Markov equivalent class 46

3.2 (a) Directed edge approach (b) Adjacency + v-structure approach (c) Maximum likelihood approach 58
4.1 **Mapping power is increased by combining populations using meta-analysis** We performed simulations assuming a background genetic effect of 25%. Power was calculated as the percentage of associations detected at a given level of significance for SNPs simulated to be causal. The meta-analysis method is shown to provide increased power at all effect sizes. 77

4.2 **The combined mapping result has higher resolution than that of the HMDP or F2 mapping.** The distribution of distances from the true causal SNP to the most significant association are shown in units of megabases. We considered peak associations which are within 15mb of the true causal SNP. As expected, the HMDP has much higher resolution than the F2 cross. The combined result achieves an even higher resolution. 78

4.3 **Meta-analysis results in increased significance and increased resolution for two loci known to be associated with BMD.** Two loci, one on chr 4 and one on chr 7, were previously found to be associated with BMD (Bmd7 and Bmd41 respectively) [FNG09]. After applying meta-analysis, we found that the peak associated SNPs for both of these loci had increased significance with respect to the F2 and HMDP mapping panels. We also found that using significant SNPs to define intervals within these loci, resulted in tighter confidence intervals, indicating increased resolution. 79
4.4 **Meta-analysis increases significance of known association.** We compare the association mapping results obtained from the HMDP, F2 cross and meta-analysis for HDL cholesterol on chromosome 1. We have previously shown that the *Apoa2* gene underlies the chromosome 1 HDL locus [WHQ93].

6.1 Application of Meta-GxE to *Apoa2* locus. The forest plot (A) shows heterogeneity in the effect sizes across different studies. The PM-plot (B) predicts that 7 studies have an effect at this locus, even though only 1 study (HMDP-chow(M)) is genome-wide significant with *P*-value.

6.2 The prevalence of heterogeneity in effect size of significant loci. Each dot represents association between SNPs and HDL phenotype from applying random effects based meta-analysis approach. Dots with larger $I^2$ value represents the existence of more heterogeneity at the locus between studies. The distribution of the heterogeneity statistic for significant SNPs (red dots) in the meta analysis is skewed toward higher heterogeneity while the non-significant SNPs are much less skewed.

6.3 Power of mouse meta-analysis to identify gene-by-environment interactions in 4,965 animals from 17 studies under varying mean effect sizes and the per study variance of the effect size which corresponds to gene-by-environment effects.

6.4 Power of (a) random-effect, (b) fixed-effect meta-analysis and (c) heterogeneity meta-analysis methods as a function of the effect size and the strength of the interaction effect (heterogeneity). (d) shows a comparison of the three methods with the color corresponding to the method with the highest power.
<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.5</td>
<td>Manhattan plots showing the results of Meta-GxE applied to (a) 17 HDL studies, (b) 9 HDL studies consisting only of male animals and (c) 8 studies consisting only of female animals.</td>
</tr>
<tr>
<td>6.6</td>
<td>Peak SNP in chromosome 8 shows interesting gene-by-environment interactions between sex × mutation-driven LDL levels.</td>
</tr>
<tr>
<td>6.7</td>
<td>Power comparison between random-effect, fixed-effect meta-analysis and heterogeneity testing</td>
</tr>
<tr>
<td>6.8</td>
<td>The association result of the fixed effects meta analysis.</td>
</tr>
<tr>
<td>6.9</td>
<td>Power comparison between random-effect meta-analysis and traditional wald test based approach</td>
</tr>
<tr>
<td>6.10</td>
<td>Power comparison between heterogeneity testing approach and traditional wald test based approach</td>
</tr>
<tr>
<td>6.11</td>
<td>Forest plot and PM-plot for Chr1:64752822 locus</td>
</tr>
<tr>
<td>6.12</td>
<td>Forest plot and PM-plot for Chr1:107271282 locus</td>
</tr>
<tr>
<td>6.13</td>
<td>Forest plot and PM-plot for Chr1:171199523 locus</td>
</tr>
<tr>
<td>6.14</td>
<td>Forest plot and PM-plot for Chr2:77837584 locus</td>
</tr>
<tr>
<td>6.15</td>
<td>Forest plot and PM-plot for Chr2:134421733 locus</td>
</tr>
<tr>
<td>6.16</td>
<td>Forest plot and PM-plot for Chr3:32944259 locus</td>
</tr>
<tr>
<td>6.17</td>
<td>Forest plot and PM-plot for Chr3:76066632 locus</td>
</tr>
<tr>
<td>6.18</td>
<td>Forest plot and PM-plot for Chr3:107430396 locus</td>
</tr>
<tr>
<td>6.19</td>
<td>Forest plot and PM-plot for Chr3:143466942 locus</td>
</tr>
<tr>
<td>6.20</td>
<td>Forest plot and PM-plot for Chr4:131925523 locus</td>
</tr>
<tr>
<td>6.21</td>
<td>Forest plot and PM-plot for Chr5:119034507 locus</td>
</tr>
<tr>
<td>6.22</td>
<td>Forest plot and PM-plot for Chr8:46903188 locus</td>
</tr>
</tbody>
</table>
6.23 Forest plot and PM-plot for Chr8:64150094 locus . . . . . . . . . . . 159
6.24 Forest plot and PM-plot for Chr8:84073148 locus . . . . . . . . . . . 160
6.25 Forest plot and PM-plot for Chr9:101972687 locus . . . . . . . . . . 161
6.26 Forest plot and PM-plot for Chr10:21399819 locus . . . . . . . . . . 162
6.27 Forest plot and PM-plot forChr10:90146088 locus . . . . . . . . . . 163
6.28 Forest plot and PM-plot for Chr11:69906552 locus . . . . . . . . . . 164
6.29 Forest plot and PM-plot for Chr11:114083173 locus . . . . . . . . . . 165
6.30 Forest plot and PM-plot for Chr14:33632464 locus . . . . . . . . . . 166
6.31 Forest plot and PM-plot for Chr15:21194226 locus . . . . . . . . . . 167
6.32 Forest plot and PM-plot for Chr15:59860191 locus . . . . . . . . . . 168
6.33 Forest plot and PM-plot for Chr17:46530712 locus . . . . . . . . . . 169
6.34 Forest plot and PM-plot for Chr18:82240606 locus . . . . . . . . . . 170
6.35 Forest plot and PM-plot for Chr19:3319089 locus . . . . . . . . . . . 171
6.36 Forest plot and PM-plot for ChrX:151384614 locus . . . . . . . . . . 172

7.1 Mouse hearing association results from random effect meta analysis . 182
7.2 (A) Forest plot and (B) PM-plot for Chr11:120818214 locus (8khz) . 185
7.3 (A) Forest plot and (B) PM-plot for Chr11:120818214 locus (16khz) 186
7.4 (A) Forest plot and (B) PM-plot for Chr11:120818214 locus (32khz) 186
7.5 The association result of three studies for 8khz mouse hearing. . . . . 196
7.6 The association result of three studies for 16khz mouse hearing. . . . . 197
7.7 The association result of three studies for 32khz mouse hearing. . . . . 198
7.8 (A) Forest plot and (B) PM-plot for Chr1:196270162 locus (8khz) . . 200
7.9 (A) Forest plot and (B) PM-plot for Chr8:130247700 locus (8khz) . . 200
7.10 (A) Forest plot and (B) PM-plot for Chr11:120818214 locus (8khz) . 201
7.11 (A) Forest plot and (B) PM-plot for Chr1:196270162 locus (16khz) . 201
7.12 (A) Forest plot and (B) PM-plot for Chr8:130247700 locus (16khz) . 202
7.13 (A) Forest plot and (B) PM-plot for Chr9:30287464 locus (16khz) . . 202
7.14 (A) Forest plot and (B) PM-plot for Chr11:120818214 locus (16khz) 203
7.15 (A) Forest plot and (B) PM-plot for Chr1:196345862 locus (32khz) . 203
7.16 (A) Forest plot and (B) PM-plot for Chr8:131366159 locus (32khz) . 204
7.17 (A) Forest plot and (B) PM-plot for Chr11:120818214 locus (32khz) 204
7.18 (A) Forest plot and (B) PM-plot for Chr18:24320393 locus (32khz) . 205

8.1 Power of (a) random-effect meta-analysis (RE2), (b) fixed-effect meta-
analysis (FE) and (c) sex-as-covariate methods (CV) as a function of
the GxS heritability and the heterogeneity between the male and the
female study effect size. (d) shows a comparison of the three methods
with the color corresponding to the method with the highest power. . . 214

8.2 Statistical power comparison between FE, GWAMA, SST and RE2 ap-
proaches while varying the effect sizes of males and females study. (a)
The power comparison under the assumption of balanced number of
individuals (b) and when there are twice as many females as males.
The diagonal line shows the points where the effect sizes between
males and females are equal. . . . . . . . . . . . . . . . . . . . . . . . . . . . . 217
8.3 Statistical power comparison between FE, UFE, WGWAMA, GWAMA, and SST approaches while varying the effect sizes of male and female study. (a) The power comparison under the assumption of balanced number of individuals (b) and when there are twice as many females as males. The diagonal line shows the points where the effect sizes between males and females are equal. 

S1 QQplots for BMI phenotype 
S2 QQplots for Crp phenotype 
S3 QQplots for Dia phenotype 
S4 QQplots for Glu phenotype 
S5 QQplots for Hdl phenotype 
S6 QQplots for height phenotype 
S7 QQplots for ins phenotype 
S8 QQplots for ldl phenotype 
S9 QQplots for sys phenotype 
S10 QQplots for tg phenotype
LIST OF TABLES

2.1 Summary Statistics For Different Likelihood Thresholds .......... 40
2.2 Regulatory Hotspots and Corresponding Regulators ............... 41
2.3 Significantly Enriched Processes For Causal and Not Causal Genes . . 42
5.1 Running time for pairwise IBD association testing for the WTCCC whole genome data .................................................. 97
6.1 17 HDL studies for meta analysis ........................................... 140
6.2 26 significant loci identified by applying Meta-GxE analysis .......... 141
6.3 False Positive Rate of RE versus Traditional Wald Test Based Approaches at Tresholds of Increasing Significance .................... 142
6.4 False Positive Rate of HE versus Traditional Wald Test Based Approaches at Tresholds of Increasing Significance .................... 142
7.1 (a) Three significant loci were identified in the meta-analysis for the 8kHz tone burst. (b) Four significant loci were identified in the meta-analysis for the 16kHz tone burst. (c) Three significant loci were identified in the meta-analysis for the 8kHz tone burst. E denotes the number of studies with an effect on the 8 kHz phenotype. N denotes the number of studies without an effect. A denotes the number of studies with an ambiguous effect. .................................................. 183
7.2 Genetic factors that contribute to age-related hearing loss in inbred mouse strains. CS=chromosome substitution; RI=recombinant inbred. 199
8.1  The comparison of genomic control $\lambda$ for 5 approaches across 10 NFBC phenotypes. .................................................. 210

8.2  Comparison of 4 association mapping approaches: Fixed effect meta-analysis (FE), random effect meta-analysis (RE2), sex-as-covariate (Covariate), combined approach (Combined) ........................................... 212
ACKNOWLEDGMENTS

First and foremost, I thank my advisor, Eleazar Eskin, for his support and guidance in my research. He always trusted me and encouraged me, even when I was depressed, so that I could continue to focus on my research with enthusiasm. It is amazing that he was always available when I needed his help. In the initial stages of my research, I found it difficult to familiarize myself with genetics because I did not have a strong background in the subject. Through experiencing a wide variety of genetics projects, I learned valuable lessons and now feel that there are so many interesting and exciting areas to study and discover in the field of genetics. I also had lot of fun with Eleazar both when carrying out research and outside of the laboratory.

I would like to thank rest of my committee; Adnan Darwiche, Christopher Lee, David Heckerman, and Aldons Jake Lusis. I especially thank David Heckerman for his mentorship and allowing me the opportunity to work on FastLmm projects. I also had lots of fun when I was working at Microsoft Research. I especially thank Jake Lusis for his support, great guidance and mentorship during my research. I had a great opportunity to work with Jake and his lab members on several very interesting projects; this is one of the greatest benefits to me for the development of my research career.

I would like to thank all of my current and previous colleagues in Zarlab. I especially thank Hyun Min Kang for his mentorship and guidance in my research. Most of all, Hyun introduced me to Eleazar, giving me the opportunity to join the Zarlab. I met Hyun on the bus on the way home from UCLA. Although I barely knew him, he shared his research and suggested that I join Zarlab. I thank Chun (Jimmie) Ye for teaching me statistics and several useful R commands during the first year of my PhD in Zarlab. I especially thank Ilya Shpitser for teaching me graphical modeling theory and causality.
I can always feel his enthusiasm to the research, while I was working with him. I especially thank Buhm Han for his insightful suggestions, for sharing scientific insights and for being a good mentor for my research. I especially thank Richard Davis for his efforts in the Meta-GxE project. We spent lots of time together interpreting the Meta-GxE results and he taught me many interesting biological knowledge behind 17 mouse HDL studies. I also thank the rest of my current and previous lab colleagues Noah Zaitlen, Sean O’Rourke, Emrah Kostem, Chris Jones, Olivera Grujic, Nick Furlotte, Jae-hoon Sul, Dan He, Warin Isvilanonda, Serghei Mangul, Farhad Hormozidiari, Jong Wha (Joanne) Joo, Wen-Yun Yang, and Zhanyong Wang, for sharing scientific insights and our enjoyable time together. I especially thank Christoph Lippert from Microsoft Research for teaching me mixed models and useful linear algebra. I also thank all the rest of my colleagues in Microsoft Research; Jennifer Listgarten, Carl M Kadie, and Robert I. Davidson. I had a lot of fun when I was working at Microsoft Research. I also thank all of my great collaborators; Sagiv Shiffman, Eyal Ben-David, Rick A. Friedman, Jeffry Ohmen, Lisa Martin, Soumya Raychaudhuri, Paul I. W. de Bakker, Atila Van Nas, Elin Org, Thomas Drake, and Charles Farber.

My church family in Olympic ward always encourages and inspires me, both temporally and spiritually. I especially thank my parents, Jai Soo Kang and Su Hee Kim, my sister, Eunchai Kang, and my in-laws for their endless love and support. I would like to thank my wife, Yun Jin Kang, and two sons, Aaron and Luke, for their presence, love, encouragement, comfort and support. I especially thank my wife, Yun Jin Kang for waiting so long for my graduation. Without her help I would never be in this state; happy, healthy, full of gratitude, and filled with love. Finally and most importantly, I would like to thank my Heavenly Father and his son Jesus Christ, my Lord, Savior and Redeemer, for his everlasting love, grace, mercy and guidance in my life.
VITA

2005  Bachelor of Science in Computer Science, University of Utah, Salt Lake City, Utah.

2002-2004  Teaching Assistant, University of Utah, Salt Lake City, Utah

2003-2004  Software Engineer, Global Knowledge Management Center-University of Utah, Salt Lake City, Utah

2004-2007  Research Assistant, University of Utah, Salt Lake City, Utah

2007  Master of Science in Computer Science, University of Utah, Salt Lake City, Utah.

2008-2009  Teaching Assistant, University of California Los Angeles, Los Angeles, California

2012-2012  Research Intern, Microsoft Research, Los Angeles, California

2008-2013  Research Assistant, University of California Los Angeles, Los Angeles, California

PUBLICATIONS


Christoph Lippert, Gerald Quon, Eun Yong Kang, Carl M Kadie, Jennifer Listgarten and David Heckerman, “The benefits of selecting phenotype-specific variants for applications of mixed models in genomics”, in *Scientific Reports*, April 2013.


Eun Yong Kang, Chun Ye, Ilya Shpitser and Eleazar Eskin, “Detecting the presence and absence of causal relationships between expression of yeast genes with very few samples”, in *Journal of Computational Biology*, March, 2010.


Eun Yong Kang, Ilya Shpitser, Chun Ye and Eleazar Eskin, “Detecting the presence and absence of causal relationships between expression of yeast genes with very few


CHAPTER 1

Introduction

The human genome contains the complete set of genetic information for humans. This information is encoded by DNA sequences within 23 chromosomes comprised of 3 billion nucleotides. Of the 3 billion nucleotides that make up the human genome, only a small fraction of nucleotide sequences differ between individuals. The most common type of individual DNA variation is the single nucleotide polymorphism (SNP), which represents a variation in a single nucleotide at a particular genomic position, mainly due to a single point mutation. If a particular genetic variant has a functional property that affects a phenotype, such as susceptibility to a disease, it is referred to as causal with respect to that phenotype. There has been a huge effort to discover the causal genetic variants of many phenotypes by the human genetics research community. The search for variants related to human disease are particularly important and unfortunately remain amongst the most elusive to discover.

One major driver for the elucidation of the genetic basis of human complex diseases is the advancement in genotyping and sequencing technology. These technological advances have enabled researchers to collect an enormous amount of high-throughput genotype data. Large scale international research efforts, such as Hapmap or 1000 genome project, have characterized a detailed catalogue of genetic variation between individuals from a number of different ethnic groups. Due to the completion of these human genome projects, it is possible to gain significant insights into the nature of human genetic diversity. These include changes to individual nucleotides
SNPs), changes in the number of copies of a segment of DNA (copy number variations; CNVs), and other structural changes such as inversions and translocations.

The challenges of discovering the genetic basis of human complex traits including disease is that this process is inherently computationally demanding due to the high-dimensionality of the human genetic data. To tackle these computational challenges in analyzing these large scale human genetic data, efficient and effective computational genetic approaches are necessary. Computational genetics refers to the field of study that utilizes computational and statistical analyses to understand causal biological mechanisms by analyzing large scale genetic data such as whole genome sequencing data, exome sequencing data, genotyping data, gene expression arrays, or transcriptome sequencing data. To analyze each of these data and uncover important biological mechanisms requires an understanding of the fundamental characteristics of human genetic data through human genome projects, and effective and efficient computational methodologies, each of which are tailored to a particular goal of interest.

One popular computational genetic approach to analyze large scale human genetic data is a genome-wide association study (GWAS). Genome-wide association studies systematically search for disease susceptibility loci by computing the correlation between common genetic variants and phenotypes such as disease status or various quantitative traits. The method searches the genome for genetic variations such as SNPs that occur more frequently in people with a particular disease than in people without the disease. The fundamental assumption of genome-wide association studies is that the SNP changes typically show a local correlation structure within a genomic segment. The tendency of a high correlation between physically close SNPs is due to linkage disequilibrium (LD) induced by the high likelihood of linkage between the two loci during recombination. In past decades, genome-wide association studies have revealed thousands of disease-associated loci and have provided insights into the allelic
architecture of complex traits. High-density genome-wide association (GWA) studies have identified a large number of disease susceptible loci that contribute to complex human diseases. These findings are providing valuable clues to the allelic architecture of complex traits.

Although individual GWAS has identified many disease-associated loci, all of these identified loci can only explain some of the variance in the phenotype of interest. This lack of the ability to explain the full variance of the phenotype is termed the “missing heritability” problem. One possible reason for this phenomenon is that performing GWAS with a single cohort is lacks the power to identify all of the disease-associated loci, particularly genetic variation with a small effect size or low minor allele frequency. The popular approach to address the lack of statistical power of a single GWA study is meta-analysis. In the meta analysis approach, the summary statistics of many independent studies are often combined, such as effect size and standard error of effect size estimates from each of the studies. This also helps to overcome the limitation of sharing individual level data. Additionally, the combined information from a large number of independent studies leads to increasing statistical power and better control of false positives. Thus meta-analysis has become the most popular approach for association studies and has led to the discovery of new disease-associated genetic loci.

In this thesis, I introduce several statistical and computational approaches for the efficient and effective genetic studies. Here I will give brief background of each problem and provide an intuitive explanation of the contribution.

In Chapter 2, I present the method to infer biological networks from high-throughput data, which particularly contains both genetic variation information and gene expression profiles from a set of genetically distinct group of individuals. Inference of biological networks from high-throughput genomic data is an important yet very challenging problem in genetical genetics and systems biology. Gene expression and genetic
variation data from high-throughput technologies provides unprecedented opportunity, which allow us to infer the underlying biological networks that govern how individual genetic variation mediates gene expression and how genes regulate and interact with each other. The method presented in this chapter try to infer much richer information on network between gene expressions compared to coexpression network, which provide only correlation between gene expressions. In particular, infer the presence or absence of causal relationships between yeast gene expressions in the framework of graphical causal models. In addition, our method can detect the absence of causal relationships and can distinguish between direct and indirect effects of variation on a gene expression level. I evaluate our method using a well studied dataset consisting of both genetic variations and gene expressions collected over randomly segregated yeast strains. Our predictions of causal regulators, genes that control the expression of a large number of target genes, are consistent with previously known experimental evidence.

In Chapter 3, I present efficient dynamic programming algorithm for computing posterior probability of causal graphical feature such as directed edges and v-structures. There have been many efforts to identify causal graphical features such as directed edges between random variables from observational data. Tian et al. [TH09] proposed a dynamic programming algorithm which computes marginalized posterior probabilities of directed edge features over all the possible structures in $O(n3^n)$ time when the number of parents per node is bounded by a constant, where $n$ is the number of variables of interest. However the main drawback of this approach is that deciding a single appropriate threshold for the existence of the directed edge feature is difficult due to the scale difference of the posterior probabilities between the directed edges forming v-structures and the directed edges not forming v-structures. I show that computing posterior probabilities of both adjacencies and v-structures is necessary and more effective for discovering causal graphical features, since it allows us to find a single appropriate decision threshold for the existence of the feature that is tested.
For efficient computation, I devise a novel dynamic programming algorithm which computes the posterior probabilities of all of $\frac{n(n-1)}{2}$ adjacency and $n\binom{n-1}{2}$ v-structure features in $O(n^3 3^n)$ time.

In Chapter 4, I introduce the meta analysis approach to combine multiple studies, where each study contains population structure. In meta-analysis, the results of separate studies are combined to obtain an aggregate result. This type of analysis has been popular in human genome-wide association studies due to issues of data privacy and the potential for reduced computational burden and increased statistical power [BFJ08]. However, one problem that has not been effectively addressed is how to combine multiple studies using meta-analysis when each study has some degree of population structure, a well-known problem in association studies in single populations [KZW08]. For single populations, there are effective algorithms and procedures for dealing with issues related to population structure [KZW08, LLL11, ZSZ12], but it is not clear how these methods may be adapted to meta-analysis. Motivated by a specific problem in mouse genetics, I will introduce a method for combing separate study populations when each study contains population structure. The method works by correcting each study separately and then combining the studies while considering the degree of population structure in each. I show that this method results in increased power and increased association resolution when combining two separate mouse populations.

In Chapter 5, I introduce efficient pairwise identity by descent (IBD) association mapping method, which utilizes importance sampling to improve efficiency and enable approximation of extremely small p-values. Two individuals are IBD at a locus if they have identical alleles inherited from a common ancestor. Recently investigators have proposed state-of-the-art identity-by-descent (IBD) detection methods to detect IBD segments between purportedly unrelated individuals. The IBD information can then
be used for association testing in genetic association studies. One popular approach to
find the association between IBD status and disease phenotype is the pairwise method
where one compares the IBD rate of case/case pairs to the background IBD rate to
detect excessive IBD sharing between cases. One challenge of the pairwise method
is computational efficiency. In the pairwise method, one uses permutation to approxi-
mate p-values because it is difficult to analytically obtain the asymptotic distribution of
the statistic. Since the p-value threshold for genome-wide association studies (GWAS)
is necessarily low due to multiple testing, one must perform a large number of permuta-
tions which can be computationally demanding. I present Fast-Pairwise to overcome
the computational challenges of the traditional pairwise method by utilizing impor-
tance sampling to improve efficiency and enable approximation of extremely small
p-values. Fast-pairwise takes only days to complete a genome-wide scan. In the ap-
application to the WTCCC type 1 diabetes data, Fast-Pairwise successfully fine-maps a
HLA gene which is known to cause the disease.

In Chapter 6, I introduce a novel meta analytic approach to identify gene-by-
environment interactions by aggregating the multiple studies with varying environ-
mental conditions. Identifying environmentally specific genetic effects is a key chal-
lenge in understanding the structure of complex traits. Model organisms play a crucial
role in the identification of such gene-by-environment interactions, as a result of the
unique ability to observe genetically similar individuals across multiple distinct envi-
ronments. Many model organism studies examine the same traits but, under varying
environmental conditions. For example, in many mouse genetic studies the same traits
are examined under different environmental conditions. Specifically, knock-out or
diet-controlled mice are often utilized in the study of cholesterol levels. The availabil-
ity of these studies presents a unique opportunity to identify genomic loci involved in
gene-by-environment interactions as well as those loci involved in the trait independent
of the environment. These studies when examined in aggregate provide an opportu-
nity to identify genomic loci exhibiting environmentally-dependent effects. In this project, I jointly analyze multiple studies with varying environmental conditions using a meta-analytic approach based on a random effects model to identify loci involved in gene-by-environment interactions. Our approach is motivated by the observation that methods for discovering gene-by-environment interactions are closely related to random effects models for meta-analysis. I show that interactions can be interpreted as heterogeneity and can be detected without utilizing the traditional uni- or multi-variate approaches for discovery of gene-by-environment interactions. I also show that our approach is more powerful than the traditional uni- or multi-variate gene-by-environment association approach which assumes knowledge of the co-variates involved in gene-by-environment interactions. The traditional uni- or multi-variate approach requires prior knowledge about environmental factors including kinds of variable (e.g. sex, age, gene knockouts) and encoding of the variables (e.g. binary values or continuous values) due to the requirement of explicit modeling of environmental variables in the model. However, the proposed approach does not require such knowledge to identify gene-by-environment interaction. I apply our new method to combine 17 mouse studies containing in aggregate 4,965 distinct animals. I identify 26 significant loci involved in High-density lipoprotein (HDL) cholesterol, many of which show significant evidence of involvement in gene-by-environment interactions.

In Chapter 7, I analyze the mice age-related hearing loss data with a meta-analytic approach. Age-related hearing impairment (ARHI) is characterized by a symmetric sensorineural hearing loss primarily in the high frequencies and individuals have different levels of susceptibility to ARHI. Heritability studies have shown that the sources of this variance are both genetic and environmental, with approximately half of the variance attributable to hereditary factors [HT10]. Only a limited number of large-scale association studies for ARHI have been undertaken in humans, to date. An alternate and complementary approach to these human studies is through the use of mouse
models. Advantages of mouse models include that the environment can be more carefully controlled, measurements can be replicated in genetically identical animals, and the proportion of the variability explained by genetic variation is increased. Complex traits in mouse strains have been shown to have higher heritability and genetic loci often have stronger effects on the trait compared to humans. Motivated by these advantages, we have performed the first genome-wide association study of its kind in the mouse by combining several data sets in a meta-analysis to identify loci associated with age-related hearing loss. I identified five genome-wide significant loci ($< 10^{-8}$). One of these loci generated a narrow peak in the region of Ahl8. These data confirm the utility of this approach and provide new high-resolution mapping information about variation within the mouse genome associated with hearing loss.

In Chapter 8, I introduce the powerful random effect based meta-analytic approach for genome-wide association studies with sex-specific effects. The prevalence of sex-specific effects complicates the analysis of association studies consisting of both males and females. The traditional approach to address this issue in genome wide association is to include sex as a covariate in the statistical model when performing association analysis. Unfortunately, this approach ignores the potential for a difference in effect sizes between the two sexes and thus may lead to a loss in power. Furthermore, genome-wide gene-by-sex effects introduce additional background phenotypic variance, which can induce additional power loss. I present a novel meta-analytic approach for the analysis of genome-wide association studies consisting of both males and females. In this approach, males and females are analyzed separately and the results are combined using a random effects meta-analysis approach allowing for a difference in effect sizes between sexes. I show that by analyzing males and females separately, our method reduces the overall variance in each study leading to an increase in statistical power when applying meta-analysis. I apply our method to the Northern Finland Birth Cohort data and I show that our method has increased power over the
traditional approach while controlling for false positives.
CHAPTER 2

Detecting the Presence and Absence of Causal Relationships Between Expression of Yeast Genes with Very Few Samples

2.1 Background

Inference of biological networks from high-throughput genomic data is a central problem in bioinformatics where many different types of methods have been proposed and applied to a wide diversity of datasets [MS07]. Several recent studies have collected data in model organisms such as yeast and mouse which contain both genetic variations as well as gene expressions from a set of genetically distinct group of individuals. Originally, these “genetical genomics” datasets were used to identify genetic variations located at specific genomic locations that affect expression levels in the form of linkages or associations [BYC02, BK05]. These studies treated expression levels as quantitative traits and each associated genomic location is called an expression quantitative trait locus (eQTL). More recently, various statistical approaches have been applied to these datasets demonstrating them as being particularly powerful for teasing out the underlying biological networks that govern how genetic variations mediate differential gene expression and how genes regulate and interact with each other [LPD06, RGM02, GDZ06, STM05]. Some of these methods build on pioneering work in using graphical models to model gene regulatory networks
Extracting meaningful causal relationships from these networks has been a challenging but important area of genetical genomics. What differs in genetical genomics studies compared to traditional microarray analysis and what makes causal inference possible is the idea to model genetic variations as random perturbations to the underlying regulatory network. A principled way of representing the causal relationships in a biological network is using graphical causal models [Pea88, Pea00]. Such models represent causal relationships between random variables by means of a directed acyclic graph called a causal graph, where a directed edge between two variables represents direct causal influence. The data-generating process represented by a causal graph imposes a variety of constraints, such as conditional independence constraints, on the observed data. A rich theory of causal inference has been developed [Pea00, SGS00] which attempts to reconstruct aspects of the graph from the pattern of constraints in the observations. Causal relationships can then be read off directly from the reconstructed graph.

The advantage of the causal inference paradigm is that predictions made are in fact causal, and so can be directly verified with knockout, siRNA or allele swap experiments. Compared to other methods such as co-expression networks which aim to capture the global structure in the regulatory network, causal inference methods attempt to identify the actual biological mechanisms regulating gene expression. Furthermore, for many applications where the final goal is to perturb the biological system in some way, causal networks are advantageous because they naturally predict the effect of possible interventions. The resulting models can be perturbed in silico to help guide which experimental perturbations to apply.

The disadvantage of these methods is that existing causal inference theory is a large sample theory, and is only guaranteed to work asymptotically. Unfortunately, in the
case of inferring biological networks from gene expression data, there are far fewer
samples than genes, which means practical applications must be a successful synthesis
of ideas from both causal inference and small sample statistics.

There are two main approaches to learning causal graphs in biological networks.
Score-based methods assign scores to models which both produce high likelihood of
the observed data, and have limited complexity, and search for the highest scoring
model [Suz93, LB94]. These methods have been used in identifying causal regula-
tors in yeast [FLN00, PRE01, SSR03, ZZS08, BH05, KJ06] and causal mediators of
disease in mice [SLY05]. Constraint-based methods rule out those causal graphs in-
consistent with patterns of conditional independence constraints in the observations.
These methods have been applied to discovering causal relationship between pairs of
genes [CES07].

In this chapter, we discover the presence and absence of causal relationships be-
tween genes in yeast by examining their expression levels over a set of individuals
with random genetic variations. Causal discovery is challenging in our case because
there are several thousand genes, while the number of samples is very limited. In
particular, most conventional conditional independence tests or model selection algo-
rithms are not reliable in the small sample case, since conditioning severely reduces
the number of samples available, and as a result we cannot infer independence with
high confidence, limiting our ability to induce features of the causal graph.

Our approach is to rely on basic properties of graphical models to infer or exclude
edge directionality based on either simple unconditional independence tests which are
possible to perform even in the small sample case, and on results of simple model
selection amongst small causal sub-graphs of the overall causal model which have
particularly strong signals. Our philosophy is that due to the small number of samples,
it is impossible to accurately recover the complete causal graph. We opt to predict only
the subset of the network where our predictions are likely to be correct.

We take advantage of prior biological knowledge that genetic variations affect gene expressions, but not vice versa. This knowledge can be expressed graphically as forbidding directed paths from gene expressions to genetic variations. While in general it is not possible to recover most causal structures based on unconditional independence tests, the availability of prior knowledge allows us to “bootstrap” certain edge orientations, which in turn allows us to orient more paths as causal using basic properties of $d$-separation (described below). Moreover, we can also rule out certain edge orientations using the same principals, thus identifying the absence of certain causal relationships.

Our method is inherently conservative, only predicting the existence and orientation of edges in the causal graph if there is strong support from the sample data. As expected, our approach predicts only a small fraction of the complete causal regulatory network of yeast. However, the actual predictions made by the method are surprisingly consistent with previous experimentally validated knowledge of yeast gene regulation.

We demonstrate the utility of our method by analyzing the Brem et al. yeast strains [BYC02]. The 112 yeast strains in this dataset was created by crossing a laboratory strain with a wild strain of $S.\, cerevisiae$. Both genetic variations and gene expressions from each offspring have been collected. We focus our analysis on an interesting feature of this dataset known as “regulatory hotspots” or regions in the genome in which a genetic variation is correlated with the expression levels of many genes. Compared to traditional eQTL mapping techniques that first identified these “regulatory hotspots”, our method provides much richer causal information that simple correlation can not capture. First, our method allows us to infer causal relationships between pairs of genes to identify global regulators that control the gene expression of target genes correlated with a “regulatory hotspot”. Second, our method can exclude causal
relationships between genes. Third, even when considering only variation-expression pairs, our method can distinguish whether a variation has a direct or an indirect effect on expression. While several other methods attempt to infer causal relationships between genes [ZZS08, CES07], our method is the first to be able to exclude causal relationships and distinguish between direct and indirect effects of a variation.

We evaluate our method’s ability to infer regulatory relationships by comparing it directly to two other competing methods as well as verifying our results with previous experimental validations. Using our method, of the 12 genes for which there is some experimental evidence that they behave as master regulators [YBW03, ZZS08], we recover 9 of them. Furthermore, for one of our predictions, ILV6, a competing method by Zhu et al., 2008 [ZZS08] was not able to identify the gene as a causal regulator based on expression data alone. The gene was only identified when additional transcription factor binding data was incorporated. Combined with our method’s ability to exclude specific causal relationships, we used gene set enrichment analysis to find that gene targets not causally affected by a regulator to be enriched for different pathways and biological processes than gene targets affected by the same regulator.

To evaluate our ability to distinguish between direct and indirect effects of a genetic variation on gene expression, we take advantage of the fact that most expression transcripts are affected directly by a few variations close to the gene through a mechanism called cis-regulation. Since our method does not rely on information about the relative positions of a genetic variation and its effected gene, an enrichment of cis-effects in our predictions for direct causal effects validates our method.

A shorter version of this article has previously been published as part of a conference proceeding [KSY09]. In this article, we provide more details on our causal inference procedure by providing the exact likelihoods for each gene in a triplet. We also updated our results by systematically identifying “regulatory hotspots” using a
previously published method based on dividing the genome into discrete sections and approximating the appearance of a linkage as a Poisson process [BYC02]. Using this method, we identified 9 “regulatory hotspots” and 38 regulator genes which mediate the genetic variations. Finally, we provide two additional visualizations for the causal relationships we discover. We use a spring embedded algorithm to construct the yeast causal network and show that the “regulatory hotspots” overlap well with the inherent hub structures. We also use a representation grouped by the “regulatory hotspots” to show that there is significant cross talk between hotspots.
2.2 Methodology

2.2.1 Causal Graphs for Genetical Genomics

We first introduce the machinery of causal inference needed to formalize our approach to inferring causal relationships between a genetic variation (a SNP) and the expression of a pair of genes. Our primary object of study is the probabilistic causal model [Pea00].

**Definition 1.** A probabilistic causal model (PCM) is a tuple \( M = \langle U, V, F, P(u) \rangle \), where

- \( U \) is a set of background or exogenous variables, which cannot be observed or experimented on, but which can influence the rest of the model.
- \( V \) is a set \( \{V_1, \ldots, V_n\} \) of observable or endogenous variables. These variables are considered to be functionally dependent on some subset of \( U \cup V \).
- \( F \) is a set of functions \( \{f_1, \ldots, f_n\} \) such that each \( f_i \) is a mapping from a subset of \( U \cup V \setminus \{V_i\} \) to \( V_i \), and such that \( \bigcup F \) is a function from \( U \) to \( V \).
- \( P(u) \) is a joint probability distribution over the variables in \( U \).

PCMs represent causal relationships between observable variables in \( V \) by means of the functions in \( F \): a given variable \( V_i \) is causally determined by \( f_i \) using the values of the variables in the domain of \( f_i \). Causal relationships entailed by a given PCM have an intuitive visual representation using a graph called a causal diagram. In this graph, each node is represented by a vertex, and a directed edge is drawn from a variable \( X \) to a variable \( V_i \) if \( X \) appears in the domain of \( f_i \). A graph obtained in this way from a model is said to be induced by said model.
A node $Y$ is an *ancestor* of node $Z$ in a causal diagram $G$ if there is a directed path from $Y$ to $Z$. Causal diagrams are generally assumed to be acyclic. While we expect the full yeast regulatory network to have causal cycles (they serve as common regulatory mechanisms), in this chapter we concentrate our efforts on the fragments of the overall network where acyclicity holds.

One advantage of causal graphs, and graphical models in general [Pea88, JW02] is their ability to represent conditional independence relations between variables in a qualitative and intuitive way using the notion of path blocking known as d-separation [Pea88]. Two variables $X, Y$ are d-separated if all causal and confounding paths from $X$ to $Y$ contain at least one variable whose value is known, and the value of no common effect of both $X$ and $Y$ is known. Every d-separation statement involving two nodes (or sets of nodes) in the graph corresponds to a conditional independence among the corresponding sets of variables. That is, if every path from $X$ to $Y$ is blocked or d-separated by $Z$ in a causal diagram $G$, then $X$ and $Y$ are conditionally independent given $Z$ in every probability distribution compatible with $G$ [Pea88]. Furthermore in stable [PV91] or faithful [SGS93] models the converse is also true: conditional independencies in the observations imply the corresponding d-separation statement holds in the underlying causal graph. The faithfulness assumption thus allows us to infer aspects of the generating causal graph from conditional independence constraints apparent in the data, and is crucial for inductive causal inference. Faithfulness holds in “most” causal models, and can thus be justified on Occam’s Razor grounds [Pea00].

Constraint-based inference of correct edge orientations in a causal diagram has two fundamental limits in practical applications. The first is that it can be difficult to collect sufficient samples to perform reliable conditional independence tests, and the second is that some causal diagrams may disagree on orientations of particular edges while entailing the same set of conditional independence constraints (such causal diagrams
are called Markov-equivalent [VP90]).

In this chapter, we will use causal graphs to represent causal interactions between genetic variations and gene expression levels in yeast. In our case the genetic variations is the set of single nucleotide polymorphisms (SNPs). In this chapter, we limit our focus to inferring the presence or absence of a causal relationship between gene expression levels based on independence tests and model selection we can actually perform. We will be relying on the following three (elementary) theorems in graphical models.

**Theorem 1.** Let $G$ be a causal graph where $X$ is d-connected to $Y$ via a path ending in an arrow pointing to $Y$, $X$ is d-connected to $Z$, and $X$ and $Z$ are d-separated by $Y$. Then $Y$ is an ancestor of $Z$.

If we assume faithfulness, this theorem implies we can infer causal directionality based on the result of two unconditional independence tests, and one conditional independence test. In our case, $X$ is a SNP, $Y$ is the expression level of a gene and $Z$ is the expression level of a second gene. We are using our prior knowledge that expression levels do not affect SNP values to satisfy one of the preconditions of the theorem, namely that the d-connected path must end in an arrow pointing to $Y$. In particular, if $Y$ is a gene expression value, and $X$ is a SNP value correlated with on $Y$, then $Y$ cannot cause $X$. Using this theorem, we are able to infer a causal relationship between the expression levels of genes $Y$ and $Z$.

Unfortunately, testing whether $X$ is conditionally independent of $Z$ given $Y$ in the small sample case is not feasible. An alternative approach which is more appropriate in our case is to use a model selection method, that is rather than performing the independence test, find the causal model over the local variables of interest, and read off causal directionality from its graph. In general, if we restrict ourselves to a small part of a large causal model which contains three variables $X,Y,Z$, the causal diagram
which captures conditional independencies in the corresponding marginal distribution, that is \( P(x, y, z) \), will be a mixed graph containing both directed and bidirected arcs, called a latent projection [VP90]. In latent projections, a directed arc from \( X \) to \( Y \) corresponds to a \( d \)-connected path which starts with an arrow pointing away from \( X \) and ends with an arrow pointing towards \( Y \) in the original, larger graph such that every node on the path other than \( X \) and \( Y \) is marginalized out or latent. Similarly, a bidirected arc from \( X \) to \( Y \) corresponds to a \( d \)-connected path in the larger graph which starts with an arrow pointing to \( X \), ends with an arrow pointing to \( Y \), and every node on the path other than \( X \) and \( Y \) is marginalized out or latent.

If we restrict ourselves to local models of three variable marginal distributions, where certain causal relationships are excluded due to prior knowledge (e.g. genes cannot cause SNPs), the complete set of causal hypotheses is captured by a small set of latent projections.

**Theorem 2.** Let \( G \) be a causal graph where \( X \) is an ancestor of \( Y \) and \( Z \). Then the latent projection which represents conditional independencies of \( P(x, y, z) \) is one of the graphs in Figure 2.3(a).

Theorem 2 allows us to select the graph in Figure 2.3(a) which best fits the available data (we use a version of the likelihood ratio test), and use this graph to conclude causal directionality. The next theorem allows us to conclude the opposite, that a variable cannot be a causal ancestor of another.

**Theorem 3.** Let \( G \) be a causal graph where \( X \) is \( d \)-connected to \( Y \), and \( X \) and \( Z \) are \( d \)-separated. Then \( Y \) cannot be an ancestor of \( Z \).

As before, faithfulness allows us to apply this theorem to conclude the absence of causal directionality based on the results of two unconditional independence tests. In our case SNP \( X \) is associated with the expression level of gene \( Y \), and SNP \( X \) is...
independent of the expression level of gene $Z$. In this case we can rule out a direct
causal relation between expression levels of genes $Y$ and $Z$. In our case, the possible
models are shown in Figure 2.3(b). In the small sample case we again use a maximum
likelihood method to perform such tests.

In the next section, we describe our statistical methodology in more detail.

### 2.2.2 Inference Algorithm Overview

Our algorithm for inferring the presence or absence of causal relationships of gene
expression proceeds in four steps. First, we find for every gene expression, the set
of potential causal SNPs using the standard $F$-test. Second, we infer the presence of
causal relationships between pairs of genes correlated with the same SNP by com-
paring the likelihoods of possible models. Third, we distinguish between direct and
indirect effects of genetic variation on gene expression. Fourth, we infer the absence
of causal relationships based on the results of step one and Theorem 3.

### 2.2.3 Finding potential causal SNPs

In the first step, we attempt to find, for every gene expression level, the set of potential
causal SNPs, in other words the set of SNPs which are either causal or which are
confounded with causal SNPs.

To examine the (potential) causal relationship between SNP $S_i$ and expression level
$E_j$ in our small sample case, we assume the following linear relationship between the
two: $E_j = \alpha S_i + \epsilon$. We use an arrow notation to signify potential causality ($\rightarrow$) and the
negation ($\not\rightarrow$) as no potential causality. Under the null hypothesis of no potential causal
relationship between the SNP and expression levels ($S_i \not\rightarrow E_j$), we expect $\alpha = 0$ ($H_0$).
Under the alternate hypothesis of a potential causal relationship ($S_i \rightarrow E_j$), we expect
\( \alpha \neq 0 (H_1) \). To decide between these hypotheses, one could calculate the likelihood ratio statistic \( x_{ij} = -2 \log \frac{L(H_0)}{L(H_1)} \) or use the standard \( F \)-test which is related to the likelihood ratio statistic \( F_{ij} = (N - 2)e^{x_{ij}} - 1 \) and follows asymptotically the \( F \) distribution with \( 1, k - 2 \) degrees of freedom where \( k \) is the number of samples. We calculate the \( F \) statistic \( F_{ij} \) for every SNP/expression pair \((S_i, E_j)\). To assign significance, we shuffle the labels of the individuals \( B \) times to obtain the null statistics \( F_{ijb}^0, b = 1, 2, ..., B \). Then the p-value of each SNP and expression pair can be calculated by looking at the ranking of the statistic of the pair in the permuted null statistic distribution.

We can easily estimate the false discovery rate (FDR) for our statistic using previous approaches [ST03]. To limit the number of potential causal networks to evaluate in subsequent steps, we filter the SNP/expression pairs for those with a FDR of \( q < 0.01 \).

Due to linkage disequilibrium or local correlation of variation, the SNPs which are correlated with expressions are not likely to be actually causal, but instead correlated with causal SNPs in the same genomic region. Since all of the SNPs are correlated in a region, this does not affect our ability to make inferences about the causal regulatory network, but we must keep in mind that the SNPs which we predict to have direct effects are likely proxies for the true causal variants.

### 2.2.4 Finding causal relationships between genes

The next stage of our algorithm consists of inferring causal directionality between gene expressions by using Theorem 1 and Theorem 2. Since the two unconditional independence tests have already been performed in the first step, all that remains is to test conditional independence. Unfortunately, conditional tests present a problem in the small sample case since conditioning further limits the number of samples we have to test. An alternative approach is to consider multiple models consistent with the results of the unconditional independence tests where in some models the conditional...
independence holds, and in others it does not. If a model where the conditional independence test holds is the best fit for the data, and moreover accounts for more of the fit compared to a “default” model making no conditional independence assumptions, then we assume the conditional independence is likely true.

In our case, we are considering fragments of the causal graph consisting of a single SNP $S$ and two expression levels $E_i, E_j$ dependent on $S$ (due to step 1). Figure 2.3(a) shows the nine possible causal models in the case that all of the elements are pairwise correlated. In $H_1$ the SNP affects both expression levels independently. In $H_2$ and $H_3$ there is a direct causal relationship between the two expression levels. The “default” models $H_4$ through $H_9$ impose no constraints on the data and are indistinguishable based on conditional independence tests. Since they are all equivalent, for simplicity, we only consider $H_4$ below.

We obtain information about the network whenever we predict a triplet to have a model $H_1$, $H_2$ or $H_3$. To distinguish between the three hypothesis $H_1$, $H_2$ and $H_3$, we compute likelihood ratio statistics for each hypothesis against the alternative $H_4$, and conclude that a hypothesis is likely true if the corresponding ratio exceeds the other ratio (e.g. fits better than the other simple hypothesis) and is close to unity (e.g. a simpler hypothesis accounts for the observations). The fact that the likelihood ratio is close to unity means that the missing edge in the triplet does not hurt the likelihood of the model compared to “default” model ($H_4$). This is equivalent to the standard approach of performing a likelihood ratio test for model selection taking into account a complexity penalty. In this case, the complexity penalty would be applied to $H_4$ since the model has an additional degree of freedom. We also pairwise compare the likelihoods between $H_1$, $H_2$ and $H_3$ against each other and only consider triplets where the most likely hypothesis is more likely than the others using a threshold.

We compute the likelihood for each model by computing the likelihoods at each
target node. Since we are interested only in the causal effects on individual genes, we can represent the causal effects on an individual gene using a linear model assuming Gaussian noise. For every triplet, we can write the following linear model for genes $g_1$ and $g_2$ and the common associated SNP $s$.

\[
g_1 = \mu_1 + \beta_{g_2} g_2 + \beta_{s_1} s + e_1 \tag{2.1}
\]

\[
g_2 = \mu_2 + \beta_{g_1} g_1 + \beta_{s_2} s + e_2 \tag{2.2}
\]

where $\mu_1$ and $\mu_2$ are the means for $g_1$ and $g_2$ respectively, and $\beta_{g_2}, \beta_{s_1}, \beta_{g_1},$ and $\beta_{s_2}$ are causal coefficients for $g_2, s, g_1,$ and $s$ to their causal target nodes respectively, and $e_1$ and $e_2$ represent noise terms which follow Gaussian distribution. In this model, all coefficients are estimated by maximum likelihood estimation. The regression coefficients are interpreted as the Wright’s rule [Wri21], sum of path products of coefficients in the underlying (and unknown) true causal graph.

Since we assume that each gene expression is independently sampled from a underlying generative model, computing the likelihood of the model given data is done by multiplying all the Gaussian density of errors calculated by least square method. We can represent this mathematically as follows:

\[
\mathcal{L}(M|D) = \prod_{i=1}^{k} \frac{1}{\hat{\sigma} \sqrt{2\pi}} \exp\left(-\frac{(X_i - \hat{\mu})^2}{2\hat{\sigma}^2}\right) \tag{2.3}
\]

where, $X_i$ is the data sample, and $k$ is the number of samples, and $\hat{\sigma}$ and $\hat{\mu}$ is computed by maximum likelihood estimation from given data.
2.2.5 Distinguishing between direct and indirect effects of variation

If a SNP \( S \) is associated with two genes \( E_i \) and \( E_j \), the nine possible models are \( H_1 \), \( H_2 \), \( H_3 \) and the “default” models \( H_4 \) through \( H_9 \). The models \( H_2 \) and \( H_3 \) explain the associations as a direct effect of the SNP on one gene and an indirect effect on the other. Model \( H_1 \) suggests that the SNP directly affects the expression levels of both genes. Since our statistical methodology uses \( H_4 \) as the default model, we are unable to distinguish between direct and indirect effects if we cannot classify a triplet as one of either \( H_1 \), \( H_2 \), or \( H_3 \).

Establishing direct and indirect effects in causal analysis is always done with respect to particular model granularity. This is because it is generally always possible to observe intermediate variables between any direct cause and its effect – finer granularity removes directness of causation. In our case, when distinguishing direct versus indirect effects, the specific triplet that we are observing determines whether or not an effect is direct or indirect. Consider the following motivating example of a SNP \( S \) and three genes with the underlying network \( S \rightarrow E_1 \rightarrow E_2 \rightarrow E_3 \). If we consider the triplet \((S,E_2,E_3)\), the correct structure of the subgraph is \( H_2 \) and \( S \) will have a direct effect on \( E_2 \) and an indirect effect on \( E_3 \). Now if we consider the triplet \((S,E_1,E_2)\), the correct structure is again \( H_2 \) and \( S \) will have a direct effect on \( E_1 \) and an indirect effect on \( E_2 \). Intuitively, this is because when we consider the triplet \((S,E_2,E_3)\), \( E_1 \) is unobserved. Thus each prediction of a triplet as \( H_2 \) or \( H_3 \) induces a partial order on the causal relationships between gene pairs. After examining all pairs, we return the minimum set of causal relationships which are consistent with all of the triplet predictions.

More complicated networks introduce ambiguity into our ability to distinguish between direct and indirect effects. For example, if we add the edge \( S \rightarrow E_3 \) in our example, we can still identify \( E_1 \rightarrow E_2 \) as a direct effect from the triplet \((S,E_1,E_2)\),
but are unable to identify $E_2 \rightarrow E_3$ as a direct effect. This is because we will predict the structure of each triplet containing $E_3$ and either $E_1$ or $E_2$ as $H_4$ where the effects are ambiguous. However, if there is an additional edge in the graph $E_3 \rightarrow E_4$, the triplet $(S, E_3, E_4)$ would identify $E_3 \rightarrow E_4$ as a direct effect.

### 2.2.6 Excluding causal relationships between genes

The ability to exclude certain causal relationships between genes, an inherent advantage of causal analysis, is important to obtaining a more complete understanding of genetic regulation. For example, a gene might be causal to a number of genes enriched for a biological process but not causal to a number genes enriched for a different biological process even though it is correlated with both sets of genes. We attempt to determine the absence of causal relationships by looking at a SNP and a pair of genes where the SNP is the potential cause of one gene, but not the other. In this case, basic properties of d-separation (Theorem 3) guarantee that there are only four possibilities $H_{10}$ through $H_{13}$ (see Figure 2.3(b)).

In $H_{10}$, the SNP affects gene expression $E_i$, but gene expression $E_j$ is completely independent from both the SNP and gene expression $E_i$. In $H_{11}$, both the SNP and gene expression $E_j$ affect gene expression $E_i$ simultaneously. In $H_{12}$, the SNP affects gene expression $E_i$, and gene expression $E_i$ and gene expression $E_j$ has a hidden common causal parent. In $H_{13}$, both the SNP and gene expression $E_j$ affect gene expression $E_i$ and at the same time, gene expression $E_i$ and gene expression $E_j$ has a hidden common causal parent. In none of these models, gene expression $E_i$ affects gene expression $E_j$.

We model the association between a SNP and a gene expression using a linear Gaussian model as in Section 3.2. We correct multiple hypothesis testing problem by computing the false discovery rate (FDR) [ST03]. We identify pairs of genes $E_i$ and $E_j$ where we can exclude causal relationships using the following criterion: a SNP is
significantly associated with $E_i$ (FDR of $q < 0.01$) and not associated with $E_j$ (FDR of $q > 0.9$).
2.3 Results

We applied our method to an expression dataset of 5534 genes and a genotyping dataset of 2956 SNPs collected over 112 genetic segregants of yeast. After step 1, we found 42331 (SNP, expression) pairs where the SNP is causal to the expression at a FDR of \( q < 0.01 \). We constructed triplets from these causal pairs to significantly reduce the number of possible causal models to evaluate for causal relationships between the genes in step 2. For each triplet, we considered the four possible models \( H_1, H_2, H_3 \) and \( H_4 \) and identified the most likely as described above. We find the most likely of \( H_1, H_2 \) and \( H_3 \) and required that the log likelihood difference of the best model be within 2 of \( H_4 \). This is equivalent to penalizing the likelihood of \( H_4 \) and applying using the likelihood ratio for model selection. Inferring causal relationships with few samples can result in directional and causal conflicts. A directional conflict occurs when the direction of causation predicted between two genes is inconsistent using different SNPs. A causal conflict occurs when the presence and absence of a causal relationship predicted between two genes is inconsistent using different SNPs.

We examined the robustness of our method by quantifying the number of directional and causal conflicts. Directional conflicts result when triples containing the same pair of genes and different SNPs predict different directional causal relations between the genes. Causal conflicts result when different triples both predict and exclude the same causal edge. As Table 2.1 shows, consistent across complexity penalties, fewer than 3% of predicted causal relationships are in conflict. These prediction conflicts are due to the limited number of samples available. We exclude all conflict predictions from our final result.

The genetic variations inherent in the individuals we study can be seeing as naturally occurring random perturbations to the underlying regulatory networks that ultimately give rise to subtle differences in gene expression. We present our results in
the context of these regulatory networks by identifying genes that are directly effected by the SNPs, regulators and those genes that are controlled by the regulators, targets. Formally, we call a gene a regulator if there exists a directed edge from a \textit{cis} SNP to the gene and a gene a target if there exists a directed edge from a regulator to the gene. Intuitively, the requirement for a causal \textit{cis} SNP ensures a high probability that the SNP directly perturbs the gene expression of the regulator. In our data, we found 3370 causal relationships consisting of 212 causal regulator genes and 1396 affected target genes. Table 2.1 shows the number of causal relationships, causal regulators and affected target genes discovered using various model complexity penalties for $H_4$.

One way to make sense of the large number of causal relationships detected is to look for causal regulators that affect a number of genes or “causal hubs”. Of particular interest is identifying causal regulators that are associated with “regulatory hotspots”, defined as regions of the yeast genome linked to the expression of a large number of genes. Presumably, these “causal hubs” are important regulatory elements that lead to subtle changes in expression of genes belonging to a number of different biological processes and functions. Previous analyses have identified several “regulatory hotspots” in the yeast genome but very little is known about the corresponding “causal hubs” because of the limited resolution of genotyping studies. In a few isolated cases, several groups have performed experimental knock out studies to confirm the existence of causal regulators and allele swap studies to further show that these regulators are perturbed by the corresponding “regulatory hotspot” [YBW03, ZZS08].

We first identified 9 “regulatory hotspots” similar to previous methods [BYC02] by dividing the genome into 611 bins and approximating the number of linkages expected in each bin as a Poisson process. Figure 2.1, 2.2, shows the complete causal network inferred by our method with regulators and targets colored by the “regulatory hotspots” they belong to. Gray nodes indicate that a gene does not belong to
any identified “regulatory hotspot”. Figure 2.1 shows the spring embedded network where the position of the nodes are determined so that the Euclidean distance is approximately proportional to the geodesic distance between two nodes [KK89]. Several regulatory hotspots overlap remarkably well with the inherent hub structures that are present in this representation including hotspot 2 (bright red), hotspot 3 (bright green) and hotspot 9 (light blue). Figure 2.2 shows the same causal network but with the nodes grouped in a circle by the “regulatory hotspot” they belong to. This representation shows that there is significant cross talk between the regulatory hotspots and there is a significant number of genes, indicated by the gray nodes, that are not part of any regulatory hotspot in our causal network.

We further summarize our results by examining each regulatory hotspot in detail. The 38 genes which are involved in the 9 regulatory hotspots among the top 45 genes which have the largest number of targets are summarized in Table 2.2 (The 7 genes which doesn’t belong to 9 hotspots include $SDS24(60), LYS2(17), URA3^+(17), GAS2(19), NMA111(20), NAM9^+(14), YML133C(30)$). Both Chen et al. [CES07] and Zhu et al. [ZZS08] applied causal inference methods to the same data allowing us to perform a direct comparison of the results. Among the genes suspected to be global regulators in the hotspots, there are a total of 12 causal regulators with some experimental evidence. Nine were proposed by the original group that collected the data: $AMN1, MAK5, LEU2, MATALPHA1, URA3, GPA1, HAP1, SIR3$ and $CAT5$ [YBW03]. Three additional were validated in Zhu et al.: $ILV6, SAL1$ and $PHM7$ [ZZS08]. Our method discovers all but 3 of these ($MAK5, SIR3$ and $CAT5$). We note that $SIR3$ and $CAT5$ have much weaker experimental evidence than the others and none of the comparison methods (neither Chen et al. [CES07] or Zhu et al. [ZZS08]) were able to find these three. The best validation of our method is that we were able to find $ILV6$ which was experimentally validated in Zhu et al. [ZZS08]. However, Zhu et al. [ZZS08] used additional types of data (incorporating TFBS data from ChIP-chip experiments, phy-
logenetic conservation, and protein protein interaction data (PPI)) in order to discover \(ILV6\) and they claim that they would not have been able to discover \(ILV6\) if they used only the data that we used. We note that \(ILV6\) was also suggested as a regulator for this hotspot by Kulp et al.[KJ06]. We recover the highlighted genes from Chen et al.[CES07] including \(NAM9\) which was not found by Zhu et al. [ZZS08] and is supported by “bioinformatics type evidence” (GO analysis, etc). A direct comparison to Chen et al.[CES07] is difficult because their results are organized in a different way, yet our results are consistent with Chen et al.[CES07] in that they highlight their discovery of 6 of the experimentally validated regulators which we also discover.

Table 2.2 summarizes our results. Experimentally validated predictions are shown in bold. Regulators with an asterisk (*) were found by Zhu et al. [ZZS08]. Regulators marked with a plus (+) were found in the Chen et al. [CES07] study and unlabeled regulators are novel predictions. In parentheses after the name of the regulator is the number of targets that we found. We note that in most cases the experimentally validated regulator is at the top of the list. We also observed that with various model complexity cut off, the ranking of predicted genes is maintained, if the model complexity cut off is less than a certain threshold. Of particular interest are a group of regulators linked to chromosome 14 which is enriched for mitochondrial genes. Previous published studies in yeast did not identify any putative regulators in this region [YBW03]. We found a number of genes in this region including three previously identified genes \(SAL1\) and \(TOP2\) and several proteins of unknown function including \(NMA111\) and \(YNL035C\).

We validate our ability to distinguish between direct and indirect effects of variation by considering the genomic positions of SNPs and the locations of genes that they are associated with. Variation that affects expression can be classified into two broad categories: \(cis\)-regulation which is an effect of a variation near a gene that affects expression of the gene and \(trans\)-regulation which is an effect of variation located in one
region of the genome affecting expression of genes in other regions. It is suspected that most cis-regulation is direct while trans-regulation may be either direct or indirect. Of the 42,331 SNP gene pairs where the SNP is associated with the expression of the gene, 11,328 are predicted as cis-regulated gene while 31,003 are trans-regulated gene. Using our approach, out of the 11,328 cis-regulated genes, we predict 9,385 of them to have a direct effect on expression. Out of 31,003 the trans-regulated genes, 20,509 of the SNP gene pairs have indirect effects. Thus cis-regulated genes are enriched in our predicted set of directly affected genes, while trans-regulated genes are enriched in indirectly affected genes.

We speculate that the identified causal regulators are likely to either directly control or perturb biological processes. However, step 3 of our analysis also identifies a collection of genes that are causally irrelevant to other genes. Combining results from these two steps can help us identify specific biological processes that are either regulated or not regulated by these causal regulators. We examined those eight significant causal regulators from our results with previous experimental validation. For each regulator, we construct two sets of genes, those that are causal targets and those that are causally irrelevant. We then use the hypergeometric distribution to assess the statistical significance of overlap of each gene set to known gene sets. Table 2.3 shows the different GO pathways that are enriched when we performed this analysis. The eight regulators appear to be involved in very different biological processes. For example, AMN1 is a causal regulator for ribosome biogenesis and assembly while four other regulators LEU2, MATA, URA3 and ILV6 are causally irrelevant for the process. Similarly, SAL1 is a causal regulator for the process of translation while HAP1 and PHM7 are causally irrelevant for the process. We notice that all significant processes are crucial for cell growth and survival but are controlled by different global regulators. The causal analysis shows that most of these global regulators participate specifically in certain biological processes. The one example of multiple regulators
from the same regulatory hotspot includes LEU2 and ILV6. In this case, these two regulators participate in similar biological processes of organic acid metabolic process and amine biosynthetic process respectively. We further confirmed the specificity of these global regulators by enrichment analysis for localization of the causal and causally irrelevant targets. For example, SALI’s causal targets are enriched for localization to the ribosome while HAPI’s targets are enriched for localization to the mitochondrial membrane. Furthermore, both PHM7 and HAPI’s causally irrelevant targets localized to cytosolic region of the cell where translation takes place. Similarly, although LEU2 and ILV6’s causal targets are not enriched for a specific cellular compartment, their causally irrelevant targets are both enriched for the nucleolus where ribosome biogenesis and assembly takes place.
2.4 Discussion

In this chapter we combined a principled representation of causality using graphical causal models with small sample statistical methods to infer the presence and absence of causal relationships between yeast genes. Working with a dataset of genetically identical yeast strains allowed us to make strong causal assumptions about edge directionality in the underlying causal model. These assumptions, in turn, allowed us to take maximum advantage of the limited samples we had available by employing either unconditional independence tests, and simple model selection to discover or exclude causal directionality between gene expressions. This work motivates theoretical questions about the limits of causal inference based on either restricting or eliminating conditional independence tests, and relying strictly on unconditional tests. In addition, our method does not explicitly account for hidden confounding effects and could potentially make erroneous predictions. Detecting causal relationships with latent variables is a challenging and active area of both theoretical and applied research. Promising new techniques have been suggested and can potentially be incorporated into our method.

We demonstrated the usefulness of our method by examining yeast expressions collected over a segregated population derived from two parental strains to identifying many experimentally validated causal regulators. In addition, our approach is able to distinguish between direct and indirect variations and exclude causal relationships between genes. These results provide a rich description of the yeast gene regulation network beyond any previous results from mapping studies, coexpression analysis and competing causal methods.

Several interesting extensions can be applied to our method. One can either empirically or theoretically characterize the strength of effects recoverable by our method to hypothesize about the strength of regulation between genes. Many biological networks
are in fact cyclical in nature and the assumption of certain type of noise structures has been shown to be useful in identifying cycles in causal graphs. Finally, incorporating additional phenotype information can potentially help us understand the genetic basis of complex phenotypes.
2.5 Figures
Figure 2.1: Complete causal network in yeast with the nine regulatory hotspots colored. Circles designate regulators, squares designate targets and diamonds designate genes that are both regulators and targets. The spring-embedded view of the causal network shows that some hotspots, hotspots 2 (red), 3 (bright green), and 9 (light blue), overlap well with the hub like structures of the network where regulators are positioned in the middle and targets surround the causal hub.
Figure 2.2: Complete causal network in yeast with the nine regulatory hotspots colored. Circles designate regulators, squares designate targets and diamonds designate genes that are both regulators and targets. Causal network grouped by hotspot shows that some regulators and targets (indicated by gray) are not part of known regulatory hotspots.
Figure 2.3: Possible causal graphs relating a triplet considering a SNP $S$ with the level of gene expression for genes $E_i$ and $E_j$. Bidirected edges denote hidden common causes. (a) Nine possible causal models consistent with $S$ being a causal ancestor of $E_i$ and $E_j$ (models $H_4$ through $H_9$ are indistinguishable from observations of the triplet). (b) Four possible causal models consistent with $S$ being a causal ancestor of $E_i$ while being uncorrelated with $E_j$. 
2.6 Tables
<table>
<thead>
<tr>
<th>Complexity Penalty</th>
<th># Causal Regulators</th>
<th># Affected Genes</th>
<th># Causal Relationships</th>
<th># Direction Conflicts</th>
<th># Causal Conflicts</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>146</td>
<td>1106</td>
<td>2135</td>
<td>34 (1.6%)</td>
<td>7 (.3%)</td>
</tr>
<tr>
<td>1.5</td>
<td>183</td>
<td>1272</td>
<td>2794</td>
<td>30 (1.1%)</td>
<td>11 (.4%)</td>
</tr>
<tr>
<td>2</td>
<td>212</td>
<td>1396</td>
<td>3370</td>
<td>44 (1.3%)</td>
<td>13 (.4%)</td>
</tr>
<tr>
<td>2.5</td>
<td>240</td>
<td>1524</td>
<td>3983</td>
<td>59 (1.5%)</td>
<td>18 (.5%)</td>
</tr>
<tr>
<td>3</td>
<td>266</td>
<td>1615</td>
<td>4571</td>
<td>81 (1.8%)</td>
<td>20 (.4%)</td>
</tr>
</tbody>
</table>

Table 2.1: Summary Statistics For Different Likelihood Thresholds
<table>
<thead>
<tr>
<th>Hot Spot</th>
<th>SNP Chr</th>
<th>SNP Loc</th>
<th>Regulators</th>
<th>Num Targets</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>370000</td>
<td>TAT1(49)</td>
<td>49</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>530000</td>
<td><strong>AMN1</strong>(226), <strong>YSWI</strong>(190), <strong>TBSI</strong>(177), CNS1**(166), <strong>ARA1</strong>(162), <strong>SUP45</strong>(31), AGP2(17), <strong>TOSI</strong>(16), YBR137W(16)</td>
<td>303</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>90000</td>
<td><strong>NFSI</strong>(106), <strong>CIT2</strong>(100), <strong>LEU2</strong>(277), HIS4(66), <strong>ILV6</strong>(29)</td>
<td>169</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>200000</td>
<td>MATALPHA1**(40)**, MATALPHA2 (24)</td>
<td>41</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>110000</td>
<td>GPA1**(15)**</td>
<td>15</td>
</tr>
<tr>
<td>6</td>
<td>12</td>
<td>640000</td>
<td><strong>HAP1</strong>(22), MAP1**(22)**</td>
<td>40</td>
</tr>
<tr>
<td>7</td>
<td>12</td>
<td>1060000</td>
<td>YLR464W**(32)**, <strong>YRF1-4</strong>(30), <strong>YRF1-4</strong>(22)</td>
<td>33</td>
</tr>
<tr>
<td>8</td>
<td>14</td>
<td>503000</td>
<td><strong>SAL1</strong>(138), LAT1(77), COG6(69), <strong>TOP2</strong>(62), MSK1(38), YNL035C(38), SWS2(17)</td>
<td>320</td>
</tr>
<tr>
<td>9</td>
<td>15</td>
<td>150000</td>
<td><strong>PHM7</strong>(227), RFC4(96), <strong>NDJ1</strong>(69), <strong>HAL9</strong>(66), ZEO1(55), WRS1(38), SKM1(28), YOL092W(18)</td>
<td>263</td>
</tr>
</tbody>
</table>

Table 2.2: Regulatory Hotspots and Corresponding Regulators
<table>
<thead>
<tr>
<th>Gene</th>
<th>Regulated Targets</th>
<th>Unregulated Targets</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GO Pathway</td>
<td>p value</td>
</tr>
<tr>
<td>AMN1</td>
<td>Ribosome biogenesis and assembly</td>
<td>$1.7 \times 10^{-34}$</td>
</tr>
<tr>
<td>LEU2</td>
<td>Organic acid metabolic process</td>
<td>$3.0 \times 10^{-7}$</td>
</tr>
<tr>
<td>MATα1</td>
<td>Biological regulation</td>
<td>$5.9 \times 10^{-6}$</td>
</tr>
<tr>
<td>URA3</td>
<td>De novo pyrimidine base biosynthetic process</td>
<td>$4.6 \times 10^{-6}$</td>
</tr>
<tr>
<td>HAP1</td>
<td>Mitochondrial electron transport chain</td>
<td>$2.4 \times 10^{-10}$</td>
</tr>
<tr>
<td>ILV6</td>
<td>Amine biosynthetic process</td>
<td>$2.4 \times 10^{-17}$</td>
</tr>
<tr>
<td>SAL1</td>
<td>Translation</td>
<td>$7.9 \times 10^{-30}$</td>
</tr>
<tr>
<td>PHM7</td>
<td>Carbohydrate metabolic process</td>
<td>$3.7 \times 10^{-9}$</td>
</tr>
</tbody>
</table>

Table 2.3: Significantly Enriched Processes For Causal and Not Causal Genes

Reference to published article

Eun Yong Kang, Chun Ye, Ilya Shpitser and Eleazar Eskin, “Detecting the presence and absence of causal relationships between expression of yeast genes with very few samples”, in Journal of Computational Biology, March, 2010.
CHAPTER 3

Respecting Markov Equivalence in Computing

Posterior Probabilities of Causal Graphical Features

3.1 Introduction

3.1.1 Background

In many scientific problems, identifying causal relationships is an important part of the solution. The gold standard for identifying causal relationships is a randomized experiment. However in many real world situations, the randomized experiment cannot be performed due to various reasons such as ethical, practical or financial issues. Therefore identifying causal relationships from observational data is an unavoidable and important step in understanding and solving many scientific problems.

A popular tool to represent causal relationships is a graphical model called a causal diagram [Pea88], [Pea00]. A causal diagram is a directed acyclic graph (DAG), which consists of nodes and directed edges. Nodes represent variables of interest while directed edges between nodes represent directed causal influence between two end nodes of those edges. A causal diagram is a data generating model, with a bundle of directed arrows pointing to each node representing the causal mechanism which determines values of that node in terms of values of that node’s direct causes.

The ideal goal of inferring causal relationships from observational data is to iden-
tify the exact data generating model. However inferring causal relationships has a fundamental limitation, if we only use observational data. That is, the best we can infer with observational data is the Markov equivalence class [VP90] of the data generating model.

The Markov equivalence class is the set of graphical models which all represent the same set of conditional independence assertions among observable variables. All graphical models in a Markov equivalence class share the same set of adjacencies and v-structures. Here a v-structure represents a node triple where two nodes are non-adjacent parents of another node. If we find the correct Markov equivalence class of the data generating model, then the Markov equivalence class contains the true data generating model as its member. The true generating model and other models in its Markov equivalence class must agree on directions of certain edges, while possibly disagreeing on others. Since all models in the Markov equivalence class share v-structures, all edges taking part in v-structures must have the same direction in all models in the class. Furthermore, certain other edges must have the same direction in the entire class as well. These edges have the property that reversing their direction would entail the creation of a v-structure which is not present in the class.

3.1.2 Previous Approaches and Their Limitations

Computing marginal posterior probabilities of graphical features considering all the possible causal network structures is computationally infeasible due to the exponential number of possible network structures, unless the number of variables is small. Due to this problem, approximation methods for computing marginal posterior probabilities of graphical features such as directed or undirected edges have been proposed. Madigan and York [MY95] applied a Markov chain Monte Carlo (MCMC) method in the space of all the network structures. Friedman and Koller [FK03] developed a
more efficient MCMC procedure which applies to the space of all orderings of variables. The problem with these approximation methods is that inference accuracy is not guaranteed in the finite runs of MCMC.

Several methods have been developed to compute the exact posterior probability of graphical features using dynamic programming algorithms. Koivisto and Sood [KS04] proposed a DP algorithm to compute the marginal posterior probability of a single undirected edge in $O(n^{2^n})$ time, where $n$ is the number of variables. Koivisto [Koi06] also developed a DP algorithm for computing all $n(n-1)$ undirected edges in $O(n^{2^n})$ time and space. In the above two DP algorithms, the number of adjacent node is bounded by a constant. These methods compute the posterior probability by marginalizing over all the possible orderings of variables in the graphical models. These DP approaches and order MCMC approaches require a special prior for the graphical structure when summing over all the possible orders of the variables. The drawback of these structure priors is that due to averaging DAGs over variable ordering space instead of over all possible structures, this structure prior results in biased posterior probability of features and leads to incorrect inferences [EW08]. To fix this bias problem, new MCMC algorithms were proposed recently [EM07], [EW08]. But these sampling-based algorithms still cannot compute the exact posterior probability.

Tian and He [TH09] proposed a new dynamic programming algorithm which computes marginalized posterior probabilities of directed edges over all the possible structures in $O(n^{3^n})$ total time when the number of parents per node is bounded by a constant. This algorithm requires longer running time than Koivisto’s approach [Koi06] but it can compute exact or unbiased posterior probabilities since it marginalizes over all the possible structures instead of all the possible orders of variables.

However there is a problem with the approach of predicting only directed edges for the two following reasons. First, if we predict a directed edge for two fixed nodes
Figure 3.1: Three graphical models from a Markov equivalent class

There are only three possible predictions: $a \rightarrow b$, $a \leftarrow b$, or $a \leftrightarrow b$, where $a \leftrightarrow b$ means no edge between $a$ and $b$. If the presence of an edge is predicted, then there are two possibilities. If the edge is constrained to always point in the same direction in every model in the Markov equivalence class, then, in the limit, its posterior should approach 1. On the other hand, if an edge can point in different directions in different models in the Markov equivalence class, then, in the limit, the posterior probability of a particular orientation should approach the fraction of the models in the class which agree on this orientation. For example, in the class showed in Figure 3.1, in the limit, $P(c \rightarrow d) = \frac{1}{3}$ while, $P(c \leftarrow d) = \frac{2}{3}$. Since the scale of posterior probabilities of these two kinds of edge features is different, this makes it difficult to decide a single appropriate posterior threshold. However one can say that scale difference problem in the posterior computation can be addressed by direct scale adjustment. But direct scale adjustment is not possible, since we do not know in advance which edges are compelled to point in the same direction across the equivalence class, and which are not. Second, we usually have only limited samples available in real world situations. This can lead to inaccurate posterior probability computation. This small sample error can make the scale difference problem worse when considering only directed edges for posterior computation, which leads to determining an inappropriate decision threshold.
3.1.3 Proposed Approach

The scale difference of the posterior probabilities of directed edge features between different kinds of edges makes it difficult to determine a single appropriate threshold. Therefore to correctly recover the model, inferring only directed edge features is not sufficient.

The following well-known theorem [VP90] will guide us in the choice of features.

**Theorem 4.** Any two Markov-equivalent graphs have the same set of adjacencies and the same set of v-structures, where a v-structure is a node triple $\langle A, B, C \rangle$, where $A, C$ are non-adjacent parents of $B$.

Theorem 4 gives us a simple set of features to test in order to completely determine the Markov equivalence class of the data generating model: adjacencies and v-structures. If the posterior probabilities of adjacency features and v-structure features are computed by marginalizing over all possible DAGs, the scale of the posterior probabilities of these two graphical features is the same unlike the posterior probabilities of the directed edge features. Therefore we propose that posteriors of both adjacency and v-structure features must be computed, if one wants to recover the Markov equivalence class of true data generating model by computing posteriors of the graphical features. For a graph of size $n$, our feature vector consists of $\frac{n(n-1)}{2}$ binary features for adjacencies and $n\binom{n-1}{2}$ binary features for v-structures. For the efficient posterior computation, we provide a novel dynamic programming algorithm which computes the posterior probabilities of all of $\frac{n(n-1)}{2}$ adjacency and $n\binom{n-1}{2}$ v-structure features in $O(n^3 3^n)$ time.
3.2 Posterior Computation of Graphical Features

To learn the causal structure, we must compute the posteriors of causal graphical features such as directed edges. The posterior probability of a graphical feature can be computed by averaging over all the possible DAG structures.

The posterior probability of a certain DAG $G$ given observational data $D$ can be represented as follows:

$$ P(G|D) \propto \mathcal{L}(G)\pi(G), $$  \hfill (3.1)

where $\mathcal{L}(G)$ is the likelihood score of the structure given the observation, and $\pi(G)$ is the prior of this DAG structure. The likelihood score computation can be performed based on the underlying model representing the causal relationships. We will talk more about the model that we used for our experiments in the experiment section.

In this chapter, we take widely accepted assumptions including parameter independence, parameter modularity [GH94], and prior modularity which simplifies our learning task. Parameter independence means that each parameter in our network model is independent. Parameter modularity means that if a node has the same set of parents in two distinct network structures, the probability distribution for the parameters associated with this node is identical in both networks. Prior modularity means that the structural prior can be factorized into the product of the local structure priors. Now if we assume global and local parameter independence, parameter modularity, and prior modularity, $\mathcal{L}(G)$ and $\pi(G)$ can be factorized into the product of local marginal likelihoods and local structural prior scores respectively.

$$ P(G|D) \propto \prod_{i=1}^{n} \mathcal{L}(V_i|Pa_{V_i})\pi_i(Pa_i), $$  \hfill (3.2)

where $n$ is the number of variables, $V_i$ is the $i^{th}$ node, $Pa_{V_i}$ is the parent set of node
$V_i$, and $\pi_i(Pa_i)$ is the structure prior for the local structure which consists of node $i$ and its parent set $Pa_i$. This factorization implies that conditional independencies in the distribution of observable variables are mirrored by a graphical notion of “path blocking” known as d-separation [Pea88], in a sense that any time two variable sets $X, Y$ are d-separated by $Z$ in a graph $G$ induced by the underlying causal model, then $X$ is conditionally independent from $Y$ given $Z$ in the distribution of variables in underlying causal model.

In stable [PV91] or faithful [SGS93] models, the converse is also true: conditional independence implies d-separation. Informally, in faithful models all probabilistic structure embodied by conditional independencies is precisely characterized by graphical structure, there are no extraneous independencies due to “numerical coincidences” in the observable joint distribution. Faithfulness is typically assumed by causal discovery algorithms since it allows us to make conclusions about the graph based on the outcome of conditional independence tests. In addition, we assume that unobserved variables which are involved in the data generating process are jointly independent. Finally, since we are interested in learning causal, not just statistical models, we make the causal Markov assumption; that is, each variable $V_i$ in our model is independent of all its non-effects given its direct causes [PV91].

Now the posterior probability for a feature $f$ of interest can be computed by averaging over all the possible DAG structures which can be expressed as:

$$P(f|D) = \frac{P(f, D)}{P(D)} = \frac{\sum_{G_i \in G_{f+}} \prod_{i=1}^{n} L(V_i|Pa_V) \pi_i(Pa_i)}{\sum_{G_i \in G} \prod_{i=1}^{n} L(V_i|Pa_V) \pi_i(Pa_i)} \quad (3.3)$$

where $G_{f+}$ is the set of DAG structures where $f$ occurs, $G$ is the set of all the possible DAG structures, from $i$ to $n$ is nodes in $G_j$, and $\pi_i(Pa_i)$ is a prior over local structures. If we introduce an indicator function to equation (3.3), $P(f, D)$ can be expressed as follows:
\[ P(f, D) = \sum_{G_j \in G} \prod_{i \in G_j} f_i(Pa_i) \mathcal{L}(V_i|Pa_{V_i}) \pi_i(Pa_i), \tag{3.4} \]

where \( f_i(Pa_i) \) is an indicator function with the value of 1 or 0. The function \( f_i(Pa_i) \) can have the value 1, if the \( G_j \) has the feature \( f \) of interest (directed edge, adjacency or v-structure), or 0 otherwise. For the directed edge feature \( j \rightarrow i \), if \( j \in Pa_i \) then \( f_i(Pa_i) \) will equal 1. For the v-structure feature \( a \rightarrow i \leftarrow b \), if \( a \in Pa_i \), \( b \in Pa_i \), \( a \notin Pa_b \), and \( b \notin Pa_a \) then \( f_i(Pa_i) \) will have 1. We can compute (3.3), if we know \( P(f, D) = \sum_{G_j \in G} \prod_{i \in G_j} f_i(Pa_i) \mathcal{L}(V_i|Pa_{V_i}) \pi_i(Pa_i) \), since \( P(f, D) \) computation with \( f = 1 \) will gives us \( P(D) \). Since the number of possible DAGs is super-exponential \( O(n!2^{\frac{n(n-1)}{2}}) \), the brute-force approach of posterior computation of all the possible graphical features takes \( O(n(n-1)n!2^{\frac{n(n-1)}{2}}) \) for all \( n(n-1) \) directed edge features and \( O(n\binom{n-1}{2}n!2^{\frac{n(n-1)}{2}}) \) for all \( n\binom{n-1}{2} \) v-structure features, where \( n \) is the number of nodes in DAG. In the following sections, we will explain how to compute \( \sum_{G_j \in G} \prod_{i \in G_j} f_i(Pa_i) \mathcal{L}(V_i|Pa_{V_i}) \pi_i(Pa_i) \) by dynamic programming more efficiently. For simplicity, we let \( B_i(Pa_i) = f_i(Pa_i) \mathcal{L}(V_i|Pa_{V_i}) \pi_i(Pa_i) \).

### 3.3 Previously Proposed Method

#### 3.3.0.1 Posterior Computation of Directed Edge Features by Exploiting Root Nodes and Sink Nodes

In this section, we describe how to compute \( P(f, D) = \sum_{G_j \in G} \prod_{i \in G_j} B_i(Pa_i) \) for all directed edge features using DP algorithm which was introduced in [TH09]. For the summation of \( \prod_{i \in G_j} B_i(Pa_i) \) over all possible \( G \) by dynamic programming, we use the set inclusion-exclusion principle for computing unions of the overlapping sets. We first split DAGs into sets whose DAG elements have common root nodes or sink nodes. Then \( P(f, D) \) can be computed by union of those sets. Now we describe the
computation of $P(f, D)$ by exploiting sets with common root nodes. Let $V$ be the set of all variables of interest and let $\zeta^+(S)$ be the set of DAGs over $V$ such that all variables in $V - S$ are root nodes. Then for any $S \subseteq V$, we define $RR(S)$ as follows:

$$RR(S) = \sum_{G \in \zeta^+(S)} \prod_{i \in S} B_i(Pa_i). \quad (3.5)$$

If we apply the set inclusion-exclusion principle to $RR(V)$, then we can derive a recursive formula for the dynamic programming in terms of $RR(V)$ as follows:

First, we define the $A_i(S)$ as follows:

$$A_i(S) = \sum_{Pa_i \subseteq S} B_i(Pa_i) = \sum_{Pa_i \subseteq S} f_i(Pa_i) \mathcal{L}(V_i|Pa_{V_i}) \pi_i(Pa_i). \quad (3.6)$$

Then Equation (3.5) can be rewritten as follows:

$$RR(S) = \sum_{k=1}^{|S|} (-1)^{k+1} \sum_{T \subseteq S, |T| = k} RR(S - T) \prod_{j \in T} A_j(V - S). \quad (3.7)$$

Now if we see Equation (3.7)$^1$, then $RR(V)$ can be recursively computed by dynamic programming with base case $RR(\emptyset) = 1$ and $RR(j) = A_j(V - j)$. For each $A_j(S)$, with the assumption of fixed number of maximum parents $k$, we can compute it using the truncated Möbius transform algorithm in time $O(k2^n)$ [KS04]. Since $RR(S)$ for all $S \subseteq V$ can be computed in $\sum_{k=0}^n \binom{n}{k} 2^k = 3^n$, $RR(V) = P(f\rightarrow v, D)$ can be computed in $O(3^n)$ time for the fixed maximum number of parents. Therefore all $n(n-1)$ directed edge features can be computed in $O(n^2 3^n)$ time.

Now we describe how to compute $P(f, D)$ by computing union of DAG sets with common sink nodes. For all $S \subset V$, let $\zeta(S)$ denote the set of all possible DAGs over $S$. Then for any $S \subseteq V$, we define $H(S)$ as follows:
\[
H(S) = \sum_{G \in \zeta(S)} \prod_{i \in S} B_i(P_{a_i}). \tag{3.8}
\]

If we apply the set inclusion-exclusion principle to \(H(S)\), then we can derive a recursive formula for the dynamic programming in terms of \(H(S)\) as follows:

\[
H(S) = \sum_{k=1}^{|S|} (-1)^{k+1} \sum_{T \subseteq S, |T| = k} H(S - T) \prod_{j \in T} A_j(S - T). \tag{3.9}
\]

Equation (3.9) \(^1\) can be efficiently computed by dynamic programming. Each \(H(S)\) can be computed in time \(\sum_{k=1}^{|S|} \binom{|S|}{k} k = |S|2^{|S|-1}\). All \(H(S)\) for \(S \subseteq V\) can be computed in time \(\sum_{k=1}^n \binom{n}{k} 2k^{k-1} = n3^{n-1}\). In the next section, we will explain how to combine \(H(S)\) and \(RR(S)\) to compute posteriors of all directed edge features which is a factor of \(n\) faster than DP algorithm using only \(RR(S)\).

**Combining \(H(S)\) and \(RR(S)\) to Compute Posteriors of All Directed Edge Features**

To compute the posteriors of all features efficiently, we extract the terms which are feature specific from the posterior computation. After we identify the terms which are common for all posterior computations, we can precompute those terms, then reuse them for all posterior computation. Let \(V\) be all the nodes in the DAG. For a fixed node \(v\), summation over all DAGs can be decomposed into the set of nodes which are ancestors(\(U\)) of \(v\), non-ancestors \(V - U - \{v\}\) and feature specific components. Now the computation of posterior amounts to the summation over DAGs over \(U\), which corresponds to \(H(U)\) and the summation over DAGs over \(V - U - \{v\}\) with \(U \cup \{v\}\) as root nodes, which corresponds to \(RR(V - \{v\} - U)\) and feature specific components which corresponds to \(A_v(U)\).

\(^1\) The proof of Eq. (3.7) and (3.9) can be found in [TH09].
We define $K_v(U)$ which corresponds to $RR(V - \{v\} - U)$ for any $v \in V$ and $U \subseteq V - \{v\}$ as follows:

$$
K_v(U) = \sum_{T \subseteq V - \{v\} - U} (-1)^{|T|} RR(V - \{v\} - U - T) \prod_{j \in T} A_j(U). 
$$

(3.10)

Then the posteriors of $u \rightarrow v$ can be computed by summation over all the subset of $V - \{v\}$ of three terms as we mentioned above. We can express this as follows:

$$
P(f_{u \rightarrow v}, D) = \sum_{U \subseteq V - \{v\}} A_v(U) H(U) K_v(U).
$$

(3.11)

In the Eq. (3.11), $A_v(U)$ is a feature specific term and $H(U)$ and $K_v(U)$ are feature independent terms. To compute the posteriors for all directed edge features, first we have to compute the feature independent terms which include $B_i$, $A_i$, $RR$, $H$, and $K_i$ under the condition of $f = 1$. Since for each directed edge feature, $f_i(Pa_i)$ will have different values, we need to recompute $A_v(S)$, where $S \subseteq V - \{v\}$. With the assumption of a fixed maximum indegree $k$, the computation of feature independent terms takes $O(n3^n)$ in time. And since each feature specific term $A_v(U)$ can be computed in $O(k2^n)$ time for all $U \in V - \{v\}$, it take $O(kn^22^n)$ for all $n(n - 1)$ directed edge features. Therefore total computational complexity is $O(n3^n + kn^22^n) = O(n3^n)$ for all $n(n - 1)$ directed edge features.

### 3.4 Novel Proposed Method

In this section, we describe a novel dynamic programming algorithm which computes the posterior probabilities of all $\frac{n(n-1)}{2}$ adjacency and $n\binom{n-1}{2}$ v-structure features in $O(n^33^n)$ time.
3.4.1 Novel DP Algorithm for Computing Posteriors of All Adjacency Features

The adjacency feature includes both direction of directed edge features. Therefore we can compute $P(f_{u\rightarrow v}, D)$ by simply adding joint probabilities for both directions. 

$$P(f_{u\rightarrow v}, D) = P(f_{u\leftarrow v}, D) + P(f_{u\rightarrow v}, D)$$

Since the computational complexity for all adjacency features is same as that of all directed edge features, we can compute all $\frac{n(n-1)}{2}$ adjacency features in $O(n^3)$ time.

3.4.2 Novel DP Algorithm for Computing Posteriors of All V-Structure Features

Now we describe a novel DP algorithm to compute posteriors for all $n\binom{n-1}{2}$ v-structure features. A directed edge feature can be represented by one edge feature, while a v-structure feature requires two non-adjacent parent nodes and their common child. Therefore the computation of $P(f, D)$ needs to be modified accordingly.

Among three terms in the equation (3.11), since $K_v(U)$ term for v-structure feature computation is same as a directed edge feature computation, we need to modify only $A_v(U)$ and $H(U)$ terms from the equation (3.11).

For the computation of posteriors of v-structure, in the equation (3.6) of definition $A_v(U)$, only the indicator function $f_v(Pa_v)$ needs to be modified from the equation (3.6). For example, if the feature $f$ represents the v-structure $a \rightarrow v \leftarrow b$, since v-structure requires two parents $a$ and $b$ pointing to node $v$, the indicator function $f_v(Pa_v)$ will have 1, if $a \in Pa_v$, and $b \in Pa_v$, or, 0 otherwise.

Now $H(U)$ needs to be modified for the posterior computation of v-structures. $H(U)$ represents the summation over the parents ($U$) of $v$. Let the v-structure feature we are testing be $a \rightarrow v \leftarrow b$. Since this v-structure requires that two parents $a$ and $b$ should not be adjacent, DAGs with edges between $a$ and $b$ should be excluded from
the computation of $H(U)$. For $(n - 2)$ v-structures with fixed two parents $a$ and $b$, there are significant overlaps in the posterior computation. If we compute the posterior probabilities of v-structures with same two parents nodes at once, we can exploit these overlaps to reduce the time and space complexity.

To this end, for all pairs of $(a, b)$, and any $S$, where $a \in V$, $b \in V$ and $S \subset V$, we define $HV(S, a, b)$ as follows:

$$HV(S, a, b) = \sum_{k=1}^{\vert S \vert} (-1)^{k+1} \sum_{T \subseteq S, |T| = k} HV(S - T, a, b) \prod_{j \in T} AT_j(S - T, a, b),$$ (3.12)

where $AT_j(S - T, a, b)$ is (1)$A_j(S - T - \{a, b\})$ when $j = a$ or $j = b$, (2) otherwise $A_j(S - T)$. Equation (3.12) computes the summation over all DAGs over $S$ except the DAGs which contain $a \rightarrow b$ or $a \leftarrow b$.

Then to apply our two modifications to equation (3.11), $H(U)$ needs to be replaced by $HV(U, a, b)$ for the v-structure $a \rightarrow v \leftarrow b$. Now we can express $P(f, D)$ for v-structure feature $a \rightarrow v \leftarrow b$ as follows:

$$P(f_{a\rightarrow v\leftarrow b}, D) = \sum_{U \subseteq V - \{v\}} A_v(U) HV(U, a, b) K_v(U).$$ (3.13)
computed in $O(n^{3n-1})$ time. Therefore, all $\frac{n(n-1)}{2}$ pairs of $(a, b)$ can be computed in $O(n^3n^{n-1})$ time. The computational complexity of $A_v(U)$ of v-structure is same as that of the directed edge case. So for $n\left(\begin{array}{c} n-1 \\ 2 \end{array}\right)$ v-structure features, $A_v(U)$ computation can be done in $O(kn^3n^r)$ time. Lastly computation complexity for $K_v(U)$ computation is same as that of directed edge case, which is $O(n^3)$ time. Therefore we can compute $P(f, D)$ for all $n\left(\begin{array}{c} n-1 \\ 2 \end{array}\right)$ v-structures in $O(n^3n^r)$ time.

3.5 Experiments

In this section we empirically compare the effectiveness of the three approaches (Inference with adjacencies + v-structures, inference with only directed edges, and the maximum likelihood approach) for discovering causal graphical features. In particular, we focus on determining which of the three approaches allows us to choose an appropriate threshold for the existence of features, which is the key step for the discovery procedure. For this comparison, we have applied these three approaches to a synthetic data set generated in the following way.

The continuous-valued synthetic data sets were generated under the assumption that the generative causal model contains linear functions while unobserved variables are Gaussian. The samples of each variable $v_i$ which has $k$ parents were thus generated using the following equation.

$$v_i = \frac{\sum_{j=1}^{k} \alpha_j P\alpha_j(v_i) + N(0, 1)}{\sqrt{\sum_{j=1}^{k} \alpha_j^2 + 2\sum_{a<b} \alpha_a \alpha_b \text{cov}(P\alpha_a(v_i), P\alpha_b(v_i)) + 1}},$$

(3.14)

where $P\alpha_j(v_i)$ is the $j^{th}$ parent of $v_i$, $\alpha_j$ is the causal coefficient of $j^{th}$ parent to the $v_i$, and $N(0, 1)$ is the Gaussian noise with 0 mean and 1 standard deviation. We assume that these unobserved Gaussian noise variables are independent of each other. After adding all the effects from parents, we divide this variable by
\[
\sqrt{\sum_{j=1}^{k} \alpha_j^2 + 2 \sum_{a<b} \alpha_a \alpha_b \text{cov}(P_{a}(v_i), P_{ab}(v_i))} + 1
\]
to normalize the variance of each variable. This synthetic data set does not contain any missing values. For this continuous-valued synthetic data set, we model it as linear model with Gaussian noise and each parameter (\(\alpha_j\)) is estimated by maximum likelihood fashion. We give the following as the structural prior for our model.

\[
\pi(G) = \frac{1}{N^{df}}, \quad (3.15)
\]

where \(N\) is the number of samples and \(df\) is the degree of freedom of given model \(G\). The above structural prior has Markov equivalent and modularity property which meets our assumption, since our algorithm computes posterior probabilities by summing over all possible DAG structures, and Eq. (3.15) only depends on the number of samples and the degree of freedom, which amounts to the number of edges in \(G\).

Now we want to compare three different prediction approaches for the task of identifying graphical features. As we mentioned above, our hypothesis is that just predicting directed edge features under limited sample makes it difficult to determine a single appropriate threshold for existence of graphical features. To verify this hypothesis, we evaluated these approaches using data sets with various sample sizes. We randomly generated 100 causal diagrams with 5 variables. For each causal graphical model, we simulated samples under linear model with Gaussian noise as we explained above. Three different data sets were generated, containing 50, 100, and 200 samples.

We computed the posterior probabilities for all possible graphical features by three different approaches. For the directed edge approach, posterior probabilities for all \(n(n-1)\) directed edge features were computed. For adjacency + v-structure approach, posterior probabilities for all \(\frac{n(n-1)}{2}\) adjacency features and all \(n\binom{n-1}{2}\) v-structure features were computed. For the ML-approach, we chose one model which had highest likelihood score among all the possible models. If the feature exists in the maximum
Figure 3.2: (a) Directed edge approach (b) Adjacency + v-structure approach (c) Maximum likelihood approach
likelihood model, its posterior is 1, otherwise 0.

Figure (3.2) shows the posterior probability distribution of positive features (features which exist in the data generating model) and negative features (features which don’t exist in the data generating model) with three approaches when 200 samples are generated. Due to space limitation, we only show the 200 sample case, but we observed similar results in the 50 and 100 sample cases. Figure (3.2) (a) represents the posterior distribution when applying the directed edge only approach. As can be seen in the upper and lower plots of Fig. 1 (a), there is no clean separation for posterior probabilities of edge features, which makes it difficult to choose a threshold. Figure (3.2) (b) represents posterior distribution, when adjacency + v-structure approach applied. As Figure (3.2) (b) shows, the posterior of positive features are mostly greater than 0.9 and the posterior of negative features are mostly less than 0.5. We can clearly see that adjacency + v-structure approach has better distinguishing power than other approaches.

3.6 Conclusion

In this chapter we propose a more effective way of identifying the Markov equivalence class of the data generating model from observational data by computing posteriors of graphical features. The previous approach of computing posteriors of only directed edge features has the problem of deciding a single appropriate threshold due to the scale difference between directed edges forming v-structures and directed edges not forming v-structures. We claim that computing posteriors of both adjacencies and v-structures is necessary and more effective for discovering graphical features, since it allows us to find a single appropriate decision threshold for the existence of the features that we are testing. Empirical validation supports that adjacency + v-structure approach is more effective than traditional directed edge only approach. For the efficient
computation, we provide a novel dynamic programming algorithm which computes all $\frac{n(n-1)}{2}$ adjacency and all $n^{(n-1)}$ v-structure features in $O(n^33^n)$ time based on DP suggested in [TH09].

Reference to published article

CHAPTER 4

Increasing Association Mapping Power and Resolution in Mouse Genetic Studies through the Use of Meta-analysis for Structured Populations

4.1 INTRODUCTION

Model organisms have long played a pivotal role in the research of human diseases. The use of mouse models in particular has been extremely effective for the identification of genes underlying Mendelian disorders. The traditional mode of discovery used to identify loci underlying these disorders has been the F2 cross. In an F2 cross, two inbred mice are used to produce F1 progeny and then these progeny are crossed to obtain F2 mice, each of which have a genetic structure that is a mix of the two original inbred strains. By applying linkage analysis to F2 populations, regions harboring causal variants are identified with high statistical power. Unfortunately, these approaches have had limited success in identifying genetic variations underlying complex, polygenic traits due to the low resolution of the study [FM01, BFO10a], meaning that the regions found to harbor causal variants are very large.

As an alternative to F2 mapping, a number of groups have proposed the use of GWAS methodologies to map traits in inbred populations [PMB04, Pay07]. Such approaches result in increased resolution, as inbred strains have a more diverse genetic structure, in which only small portions of the genome are shared between any two
strains. The initial results were promising, but it was later found that the significant population structure within inbred strains causes a large number of spurious associations and inflates the significance of true associations. Upon correction for population structure, most of the associations identified as significant were found to be spurious [KZW08, MGP09]. Also, when corrected for population structure, existing panels of classical inbred strains were under powered to detect genetic variants explaining less than 10% of the phenotypic variation. In order to address these issues, Bennett et al. (2008) [BFO10a] utilized a panel of mice called the Hybrid Mouse Diversity Panel (HMDP), which combines inbred strains with recombinant inbred (RI) strains, which resemble an inbred version of an F2 cross. The idea is that inbred strains provide high resolution, while RI strains provide increased power. They showed that when performing association mapping within this panel they achieved higher resolution than when performing mapping only using RI strains and showed that they achieved higher power than when performing mapping with only inbred strains. However, the power to detect small effects remains quite low, a problem that is due to an inherit limitation in the design of the HMDP: the limit on the availability of inbred strains.

Limited power and resolution are noted problems in many mapping panels and in order to combat these issues, a number of groups have suggested methods to combine the results from multiple studies [HDK00, HMC02, PBL07, LLW05]. The core concepts behind these methods, all of which are formed on linkage-based methodologies, may be adapted to work in association analysis. However, such approaches may not be well-suited for studies in structured populations. For example, a shared feature of these linkage-based methods is the attribution of equal informativeness to each study. Such an assumption may not hold in studies with population structure, as the informativeness of a given panel will be locus-dependent. In this case, methods attributing equal weight to each population may result in sub-optimal power.
In this chapter, we propose a method to combine studies in a locus-specific manner, weighting each study relative to its level of informativeness, and show that our method achieves optimal power within the proposed framework. Our method is based on the concept of meta-analysis. In a meta-analysis, the statistics obtained for each SNP in two separate studies are used to obtain a meta-statistics that combines information from these studies. The most common methods for performing meta-analysis are based on the fixed effect weighted sum of Z-scores (WSoZ) \cite{BFJ08}, in which Z-scores from each study are combined using a pre-defined weighting scheme. Typically, weights are set as proportional to the number of individuals in the study. Using this basic idea, we propose a meta-analysis method for combining the results obtained from mapping in the HMDP with those obtained from mapping within an F2 population. Since the best way to combine results from these two populations at a given SNP is dependent on the strain distribution pattern in each population at that SNP, current meta-analysis methodologies are not well suited. We introduce a method that accounts for the genetic structure within each population when combining results. Using a mixed-model-based approach to correct for population structure, we derive a meta-statistic based on the WSoZ. By applying an optimal weighting scheme, our method achieves both higher power and increased resolution over mapping performed only within one population. We note that the HMDP is only one of several recently proposed strategies for increasing the resolution of mouse genetic studies over traditional crosses. Other strategies include the collaborative cross \cite{AVF11}, and the use of heterogeneous stocks \cite{HSV09}. The meta-analysis method we introduce is flexible and may be used to combine studies conducted within these panels as well.

We evaluate our method through simulation and by applying it to real phenotype data for which previous discoveries have been made. First, we evaluate both power and resolution through a simulation framework. We find under many different settings that the meta-analysis approach results in higher power than either single panel. We also
find that when applying the meta-analysis approach, resolution is increased 1.5-fold with respect to the HMDP and 3.5-fold with respect to an F2 panel. Next, we apply the meta-analysis approach to map bone mineral density (BMD), which was measured from the femurs of 865 HMDP mice and 161 F2 mice, a cross between C57BL/6 and C3H [FBO11]. In our results, two previously implicated loci are recovered with increased significance. Deriving support intervals using the span of significant SNPs, we find that our method results in increased resolution over results obtained through linkage analysis. Finally, we apply our method to map HDL cholesterol in 687 HMDP mice and 164 F2 mice [NGW09] and find that a gene (Apoa2 [WHQ93]) known to be associated is identified with increased significance.

4.2 MATERIALS AND METHODS

4.2.1 Association Studies

Let us assume that we have measured a phenotype within a population $i$ that contains $n_i$ individuals. We denote the $n_i \times 1$ column vector of phenotype measurements as $y_i = [y_{i1}, y_{i2}, \ldots, y_{in_i}]'$. In order to test the association between the phenotype and a given SNP $r$, we test the hypothesis $\beta = 0$ under the model in equation (4.1), where $\mu$ is the global phenotype mean and $x_i$ is a vector of minor allele counts of SNP $r$ for individuals in population $i$.

$$y_i = \mu + x_i \beta + \epsilon$$ (4.1)

A test statistic for testing $\beta = 0$ is derived by noting the distribution of the estimate of $\beta$ under the assumption of normality. We denote the estimate of $\beta$ in population $i$ as $\hat{\beta}_i$, where $\hat{\beta}_i \sim N(\beta, s_i^2)$ and $s_i^2$ denotes the squared standard error of the estimate in population $i$. The z-score statistic for SNP $r$ in population $i$, $Z_i$, is given in equation
(4.2) and may be used to test the hypothesis $\beta = 0$ or may be used in order to derive other statistics, such as a chi-square or F-statistic.

$$Z_i = \frac{\hat{\beta}_i}{s_i}$$

(4.2)

**Traditional Meta-Analysis**

Most of traditional methods for meta-analysis employ the weighted sum of z-scores (WSoZ) [SRC09] [WSL09] [ZSS08]. In this method, a meta statistic for each SNP is calculated using equation (4.3), where $w_i$ denotes a weight given to each Z-score for a population $i$.

$$Z_{meta} = \frac{\sum_i w_i Z_i}{\sqrt{\sum_i w_i^2}}$$

(4.3)

The weights, $w_i$, are often a function of the sample size of their respective population, so that larger population samples obtain a higher weight [BFJ08]. This weighting scheme make sense intuitively as we may want to attribute greater confidence to studies with more individuals. Alternatively, weights are set as the inverse of the standard error of the estimate of the beta coefficient, so that $w_i = 1/s_i$. The resulting meta statistic is the so-called pooled inverse variance-weighted beta coefficient [BFJ08]. As has been done for case-control studies [ZE10], it is possible to show that this particular weighting scheme is optimal in the sense that these weights maximize the power of detecting an effect of size $\beta$.

Given the distribution of $\hat{\beta}_i$, we have that when $\beta \neq 0$, $Z_{meta} \sim (\lambda, 1)$, where $\lambda$ is a non-centrality parameter with $\lambda = \sum_i w_i \frac{\beta}{s_i} / \sqrt{\sum_i w_i^2}$. $\lambda$ is maximized, when $w_i = 1/s_i$, thus meta-statistic has optimal power to detect an effect of size of $\beta$. The optimality of the weight ($w_i = 1/s_i$) is shown by using the Cauchy-Schwarz inequality.
\[ \sum_i w_i \frac{\beta}{s_i} \leq \sqrt{\sum_i w_i^2} \sqrt{\sum_i (\frac{\beta}{s_i})^2} . \] Under the assumption that \( \beta \) is the same across all populations, equality holds when \( w_i = 1/s_i \).

### 4.2.2 Association Studies in Structured Populations

Although the traditional approach to association mapping is often used, there are a number of issues that arise when performing this basic analysis. One problem is that of population structure or cryptic relatedness [DRB01, VP05], in which genetic similarities between individuals both inhibit the ability to find true associations and cause the appearance of a large number of false or spurious associations. Mixed effects models are often used in order to correct this problem [Lan02, YPB06, KZW08, KSS10]. Methods employing a mixed effects correction account for the genetic similarity between individuals with the introduction of a random variable into the traditional model from equation (4.1).

\[ y_i = \mu + \beta x_i + u_i + \epsilon \]  

In the model in equation (8.8), the random variable \( u_i \) represents the vector of genetic contributions to the phenotype for individuals in population \( i \). This random variable is assumed to follow a normal distribution with \( u_i \sim N(0, \sigma^2_g K_i) \), where \( K_i \) is the \( n_i \times n_i \) kinship coefficient matrix for population \( i \). With this assumption, the total variance of \( y_i \) is given by \( \Sigma_i = \sigma^2_g K_i + \sigma^2_e I \). A z-score statistic is derived for the test \( \beta = 0 \) by noting the distribution of the estimate of \( \hat{\beta}_i \). In order to avoid complicated notation, we introduce a more basic matrix form of the model in equation (8.8), shown in equation (8.9).

\[ y_i = X_i \Gamma + u_i + \epsilon \]
In equation (8.9), $X_i$ is a $n_i \times 2$ matrix encoding the global mean and SNP vectors and $\Gamma$ is a $2 \times 1$ coefficient vector. We note that this form also easily extends to models with multiple covariates. The maximum likelihood estimate for $\Gamma$ in population $i$ is given by $\hat{\Gamma}_i = (X_i'\Sigma^{-1}_i X_i)^{-1}X_i'\Sigma^{-1}_i y_i$ which follows a normal distribution with a mean equal to the true $\Gamma$ and variance $(X_i'\Sigma^{-1}_i X_i)^{-1}$. The z-score statistic for testing $\beta = 0$ is then given in equation (4.7), where $R = [0 \ 1]$ is a vector used to select the appropriate entry in the vector $\hat{\Gamma}_i$.

$$Z_i = [R(X_i'\Sigma^{-1}_i X_i)^{-1}R']^{-1/2}R\hat{\Gamma}_i$$ (4.6)

$$Q_i = [R(X_i'\Sigma^{-1}_i X_i)^{-1}R']$$ (4.7)

$$\sum w_i Q_i^{-1/2} R \Gamma \leq \sqrt{\sum w_i^2} \sqrt{\sum (Q_i^{-1/2} R \Gamma)^2}$$ (4.9)

### 4.2.3 Meta-Analysis in Structured Populations

In order to perform meta-analysis using multiple structured populations, we adopt the weighted sum of z-scores approach shown in equation (4.3), where the z-score for population $i$ is given in equation (4.7). When $\beta \neq 0$, $Z_{meta}$ will have a normal distribution with variance 1 and mean $\sum w_i Q_i^{-1/2} R \sqrt{\sum w_i^2}$. Again we employ the use of the Cauchy-Schwarz inequality, shown in equation (4.9), to show that the optimal weights are given by $w_i = Q_i^{-1/2}$. This result may also be obtained by noting that $Q_i^{-1/2}$ from equation (4.7) is the mixed-model equivalent to $s$ from section (Traditional Meta-Analysis). However, this result is more general, allowing for a more flexible hypothesis testing framework in which any linear combinations of the elements of $\Gamma$ may be evaluated.
Substituting the optimal weights we arrive at the final meta statistic given in equation (4.10) with its distribution under the alternative hypothesis given in equation (4.11).

\[
Z_{\text{meta}} = \frac{\sum Q_i^{-1} R_i}{\sqrt{\sum Q_i^{-1}}} \sim N(\beta \sqrt{\sum Q_i^{-1}}, 1)
\]

It should be noted that when \(\Sigma_i\) is unknown, it must be estimated from the data. In this case, \(Z_{\text{meta}}\) may not follow a standard normal distribution under the null, due to the unaccounted uncertainty in the estimation of \(\Sigma_i\). However, we are able to sidestep this issue by using a global search technique [KZW08, KSS10], in order to find an optimal estimate of \(\Sigma_i\) for each population.

**Simulations**

Simulations were performed using a previously designed framework [KKW10b, BFO10a]. For both power and resolution, phenotypes were generated by sampling a phenotype for each strain while assuming the model from equation (8.8). The genetic variance \(\sigma_g^2\) was determined for a given genetic background \((g^2)\) by using equation (4.12), where \(S = I_n - 1/nJ_n\) (\(J_n\) is an \(n \times n\) matrix of ones).

\[
\sigma_g^2 = \frac{g^2\sigma_e^2(n-1)}{(1 - g^2)Tr(SK_i)}
\]

The power and resolution for each effect size \((\beta)\) was determined by first applying the association mapping procedures to each simulated phenotype. Power was calculated as the percentage of associations for the known causal SNP that reached significance. For resolution, association was applied to each SNP on the same chromosome as the causal SNP. The distance between the true causal SNP and the peak associations...
were recorded. If the peak association is greater than 15Mb away from the causal SNP, then the value is recorded as 15Mb. This procedure helps to reduce the mean shift that occurs because of low power within a region.

**Significance Threshold Estimation**

Significance thresholds were estimated for each method using a technique utilized previously [KKW10b, BFO10a]. Ten-thousand null phenotypes were generated and association statistics were calculated for each phenotype over all SNPs. We selected the minimum p-value for each phenotype, resulting in a set of 10,000 minimum null p-values. The threshold was chosen by selecting the p-value for which only 5% of the minimum p-values were smaller. This p-value then represents our threshold controlling for 5% FDR. Thresholds for the HMDP, F2 and Meta-analysis approach were found as follows: $3.715 \times 10^{-6}$, $2.4637 \times 10^{-4}$ and $2.7 \times 10^{-6}$.

**4.2.4 Mouse Association Data**

Genotypes for the F2 cross were obtained from a previous study [ECW04, WYS06, NGW09, FBO11]. The original cross contained 311 mice, but we randomly sampled only 300 for our simulation studies. Each mouse was genotyped at about 1200 markers spread across the genome and it was this set of markers which was used previously to perform linkage analysis. In order to apply the meta-analysis approach outlined in this chapter, we require that the F2 mice be typed at the same markers as the HMDP. Fortunately, since the parental strains for the F2s are part of the HMDP, genotyping is not necessary. Instead we perform imputation in order to determine the state of each marker which is typed in the HMDP but is not part of the markers typed in the F2 cross. By applying the imputation algorithm described below, we obtained a set of 113,650 SNPs which were polymorphic in both the HMDP and the F2 cross. This is compared
to the total set of markers available for the HMDP, which is of size 132,285.

We utilize a straightforward approach to imputation by noting the simple structure of the F2 genomes. For any two adjacent markers in a given F2 mouse, the state of the intervening markers will be determined by the state of the two adjacent markers. Let two adjacent markers be $x_i$ and $x_{i+k}$, where $k$ is the number of intervening markers. If both $x_i$ and $x_{i+k}$ are in the same state as parent one, then the markers from $x_{i+1}$ to $x_{i+k-1}$ will be set to be the same as the corresponding markers in parent one. Likewise, if both $x_i$ and $x_{i+k}$ share the same state as parent two, the intervening markers will be set to those from parent two. If there is a switch in state between the two adjacent markers, this indicates a recombination. In this case, we are not able to determine the state of the intervening markers and these will be labeled as unknown.

4.3 RESULTS

4.3.1 Combining the HMDP with an F2 cross increases power

We show that by combining the mapping results obtained in the HMDP with those obtained in an F2 cross through meta-analysis, we achieve higher power than when mapping within only one panel. Simulations are performed with genotypes for 300 F2 mice, which were obtained from a previously generated cross [ECW04, WYS06, NGW09]. The F2s were genotyped at about 1200 markers and imputation was performed (see Methods) to obtain genotypes at all markers typed in the HMDP strains.

Power simulations were performed as described in previous studies [KKW10b, BFO10a]. We randomly selected a set of 10,000 SNPs that are polymorphic in both the F2 cross and the HMDP. For each SNP we generated a phenotype with a 25% genetic background effect and a SNP effect of a given size. The genetic background effect can be thought of as the heritability of the trait. Association between each SNP
and its corresponding set of generated phenotypes was tested using EMMA [KZW08] for the F2 and HMDP panels alone. Power for each SNP effect size was calculated as the percentage of tests that resulted in a significant p-value. Significance thresholds for each panel were obtained through a parametric bootstrap procedure.

Figure 4.1 shows the comparison of power between the meta-analysis approach and mapping within the individual panels. In these simulations, we varied both the number of F2 mice as well as the number of HMDP replicates. Power is reported on the y-axis and the magnitude of the SNP effect is reported on the x-axis. The SNP effect is reported in terms of $\beta$ from equation (8.8) and the actual variance explained for a given value will depend on the SNP as well as the genetic background. Therefore, we determine the variance explained by a given effect size under a given genetic background by taking the average variance explained in the HMDP across all SNPs. The meta-analysis method has higher power than mapping within the single populations in all simulations. As power within each of the single populations increases, so does the power of the meta-analysis method. For a large number of F2 mice and HMDP mice, the power to detect small effects increases dramatically by applying meta-analysis. For example, for a SNP effect accounting for 5% of the phenotypic variance ($\beta = 0.5$), we find that mapping within only the HMDP with 5 replicates results in a 50% power, while mapping within only the F2 cross results in a power of 17%. When combining the results through meta-analysis, the power increases to 75% (Figure 4.1D).

### 4.3.2 Meta-analysis leads to an increase in resolution over HMDP and F2 mapping

We evaluate the mapping resolution when using the HMDP, F2 and the meta-analysis approaches through simulation. Resolution was evaluated by calculating the genetic distance between a SNP simulated to be causal and the peak associated SNP, while
only considering the region within 15Mb of the causal SNP. Figure 4.2 compares the
distribution of these distances under each mapping method. Simulations were per-
formed assuming a 25% genetic background effect and a SNP effect accounting for
10-15% of the phenotypic variance.

Using one replicate for the HMDP, we find that the mean distance of the peak as-
sociation to the true causal SNP is 3.17Mb. This compares with a mean of 7.5Mb
obtained when mapping within the F2 panel. When combining results through the
meta-analysis approach, the mean distance is decreased to 2.21Mb. This is an almost
1.5-fold increase in resolution over the HMDP and an almost 3.5-fold increase in res-
olution over the F2 panel.

4.3.3 Application to Bone Mineral Density

We obtained a set of bone mineral density (BMD) measurements from the femurs of
865 HMDP mice and 161 male F2 mice. We applied association mapping in each panel
separately using EMMA [KZW08], and applied the meta-analysis approach as well.
Manhattan plots summarizing these results are shown in figure (4.3). Two loci (Chr 4
and 7) showed an increase in significance relative to the associations in either the F2 or
HMDP. The significance of the Chr. 7 meta-analysis peak was an order of magnitude
more significant \(3 \times 10^{-7}\) than either the HMDP \(3.1 \times 10^{-6}\) or the F2 \(1.6 \times 10^{-3}\)
peaks. The original QTL on Chr. 7 (\textit{Bmd41}) had a 1.5 LOD support interval of 80 Mb
(24.9 to 104.9 Mb) [FNG09]. The Chr. 7 meta-analysis SNPs with P-values of \(\leq1e-6\)
extended from 17.2 to 25.2 Mb, representing a significant improvement in resolution
for \textit{Bmd41}. The QTL on Chr. 4, previously referred to as Bmd7 [FNG09], was the
strongest locus affecting femoral BMD in the F2 \(p = 7.8 \times 10^{-4}\). The peak F2 SNP
was moderately significant in the HMDP \(1.3 \times 10^{-3}\) and highly significant in the
meta-analysis \(2.8 \times 10^{-6}\). \textit{Bmd41} was previously found to have a 1.5 LOD support
interval of 11.0 Mb (126.2 to 137.2 Mb). In the meta-analysis SNPs with P-values of \( \leq 1 \times 10^{-6} \) spanned 10 Mb (from 129 to 139 Mb).

### 4.3.4 Application to HDL cholesterol

We obtained a set of HDL measurements for 687 male mice each a member of the HMDP and a set of 164 male F2s [NGW09]. We applied association mapping in the HMDP and F2 panels separately using EMMA [KZW08] and then applied our meta-analysis approach. Figure 4.4 shows the results of this experiment. As shown in the original paper introducing the HMDP, the peak association for HDL is found on distal chromosome one, in which a well-known association with the \( \text{Apoa2} \) [DLW90] gene exists. The peak association is 25kb upstream of the start site of the \( \text{Apoa2} \) gene with a p-value of \( 7.06 \times 10^{-8} \), which is significant at the \( 1 \times 10^{-7} \) level estimated from a parametric bootstrap procedure. The mapping results obtained from the F2 panel (Figure 4.4A) resembles a linkage peak, due to the large amount of linkage disequilibrium within the F2 genomes. The peak association identified in the F2 population is over 2Mb downstream of the end site for the \( \text{Apoa2} \) locus with a p-value of \( 2.67 \times 10^{-9} \).

Figure 4.4B shows the mapping result obtained with the meta-analysis procedure. Using the meta-analysis result, we again obtain the association which is 25kb from the start site of the gene, however the p-value is greatly reduced to less than \( 1 \times 10^{-15} \).

### 4.4 DISCUSSION

In this chapter, we introduce a study design in which the Hybrid Mouse Diversity (HMDP) inbred panel is combined with an F2 cross in order to perform association mapping. We show that by utilizing a meta-analysis approach which accounts for the genetic structure of the populations, both association power and resolution are in-
creased when compared with mapping within either of the individual panels. The reason for increased power can be understood intuitively as, in general, increased sample sizes lead to increases in power. However, an increase in resolution when combining a high resolution panel with a low resolution panel is somewhat counter intuitive. One way to understand why we achieve higher resolution is by considering that by combining panels we are increasing the number of overall unique genomic break points.

Our results have focused on the case when the HMDP panel is combined with one F2 cross. However, by using the methodology we present any number of panels can be combined. One obvious potential for this is that by adding additional F2 panels, we may increase power much further. A significant amount of cross data exists in publicly accessible databases such as MGI [BBK11]. By utilizing existing cross data researchers will be able to use our technique, in order to increase the power of their studies without spending money to generate F2s of their own.

Another advantage of our method is that it is general enough to be used in order to combine the HMDP with other types of study designs such as the Collaborative Cross [AVF11] and heterogeneous stock [HSV09]. However, one potential issue that may arise when combining the HMDP with such panels is that of heterogeneity of effect size. That is, the magnitude of main effects may vary between different mapping panels due to the difference in the overall genetic structure. In this case, our method may be easily extended to utilize approaches which account for such heterogeneity between effects [HE11a]. Heterogeneity between effect sizes is also known to be a problem between sexes within the same population. Therefore, a similar approach may be utilized in order to combine results across sexes within the same mapping panel.
LIST OF FIGURES
4.5 FIGURES

Reference to published article

Nicholas A. Furlotte*, Eun Yong Kang*, Atila Van Nas, Charles R. Farber, Aldons J. Lusis and Eleazar Eskin, “Increasing association power and resolution in mouse genetic studies through the use of meta-analysis for structured populations” in Genetics, April 2012.
Figure 4.1: **Mapping power is increased by combining populations using meta-analysis** We performed simulations assuming a background genetic effect of 25%. Power was calculated as the percentage of associations detected at a given level of significance for SNPs simulated to be causal. The meta-analysis method is shown to provide increased power at all effect sizes.
Figure 4.2: **The combined mapping result has higher resolution than that of the HMDP or F2 mapping.** The distribution of distances from the true causal SNP to the most significant association are shown in units of megabases. We considered peak associations which are within 15mb of the true causal SNP. As expected, the HMDP has much higher resolution than the F2 cross. The combined result achieves an even higher resolution.
Figure 4.3: Meta-analysis results in increased significance and increased resolution for two loci known to be associated with BMD. Two loci, one on chr 4 and one on chr 7, were previously found to be associated with BMD (Bmd7 and Bmd41 respectively) [FNG09]. After applying meta-analysis, we found that the peak associated SNPs for both of these loci had increased significance with respect to the F2 and HMDP mapping panels. We also found that using significant SNPs to define intervals within these loci, resulted in tighter confidence intervals, indicating increased resolution.
Figure 4.4: Meta-analysis increases significance of known association. We compare the association mapping results obtained from the HMDP, F2 cross and meta-analysis for HDL cholesterol on chromosome 1. We have previously shown that the Apoa2 gene underlies the chromosome 1 HDL locus [WHQ93].

80
CHAPTER 5

Fast Pairwise IBD Association Testing in Genome-wide Association Studies

5.1 Background

Identity by descent (IBD) is a fundamental concept in genetics. Two individuals are IBD at a locus if they have identical alleles inherited from a common ancestor. Investigators have put tremendous efforts to map the IBD segments between purportedly unrelated individuals [PNT07, GLS09, BB10]. The current state-of-the-art methods such as GERMLINE [GLS09] and Beagle [BB10, BB11] can detect even small (several megabases) IBD segments between individuals.

One promising application of IBD mapping is to use discovered IBD segments in the association testing [PNT07]. Investigators usually test single nucleotide polymorphisms (SNPs) for association, but SNPs may not “tag” low frequency causal variations well [BYP05]. Imputation also performs poorly on rare variations [MHM07, BB09]. Association testing based on the IBD information, or IBD association testing, can complement standard association testing methods [BB11].

There are two categories of methods in IBD association testing. The first method is the pairwise method [PNT07] where one compares the IBD rate of case/case pairs to the background IBD rate to detect excessive IBD between cases. The rationale is that if a rare causal variations has occurred at a relatively recent ancestor, cases will
likely share an IBD segment containing the causal variation. The second method is the clustering method [GKL11] where one divides individuals into clusters based on the IBD information and then test each cluster for association assuming the cluster “tags” a rare causal variation. In this chapter, we focus on the pairwise method.

The pairwise method have two computational challenges. The first challenge is computational efficiency. In the pairwise method, one uses permutation to approximate p-values because it is difficult to analytically obtain the asymptotic distribution of the statistic. Since the p-value threshold for genome-wide association studies (GWAS) is necessarily low due to multiple testing [BT12], one must perform a large number of permutations which can be computationally demanding. The second challenge is fine-mapping. After one identifies a significant loci, it is important to pinpoint the most significant peak within the loci to further follow-up candidate genes. The permutation is limited for this purpose because the smallest p-value it can approximate is constrained by the number of permutations, often resulting in many SNPs having the same minimal p-values in the region.

In this chapter, we present a new method, Fast-Pairwise, to overcome the computational challenges of the traditional pairwise method. Fast-Pairwise utilizes importance sampling [Was04] to improve efficiency and enable approximation of extremely small p-values. To devise an importance sampling procedure, we introduce a new statistic that has two properties; it can be conveniently sampled and it can approximate the pairwise method statistic. We show that the new statistic has close relationship to the pairwise method statistic through the properties of the graph representation of IBD.

Fast-Pairwise is efficient and takes only days to complete a genome-wide scan. To demonstrate the utility in fine-mapping, we apply our method to the type 1 diabetes dataset of the [Wel07]. In this dataset, the traditional pairwise method can identify a significant region in chromosome 6 [BT12], but it gives the same minimal p-value
Figure 5.1: An example of IBD graph. IBD detection method provides IBD information (Table). Then we build a graph where vertices are individuals and edges are IBD relationships.

For a very wide region (26.7Mb-35.5Mb) including all eight classical human leukocyte antigen (HLA) genes. Among these, Fast-Pairwise pinpoints *HLA-DQB1* which is known to cause the disease [TBM87].


### 5.2 Methods

#### 5.2.1 IBD graph

Given *N* individuals, the IBD information at a genomic locus can be represented as a graph with *N* vertices (Figure 5.1). An edge exists between a pair of vertices if the individuals are IBD.
5.2.2 Pairwise methods for IBD association mapping

We refer to a class of IBD association mapping methods as “pairwise methods” if they examine the relative numbers of edges in the IBD graph at each locus. There are three different types of edges: edges that connect two case individuals, edges that connect two control individuals and edges that connect a case and a control individual. Pairwise methods can be performed in two different ways. One way is to compare the number of case/case pairs to control/control pairs. A second way is to compare the number of case/case pairs to non-case/case pairs (union of control/control pairs and case/control pairs). We will call the first method CC and the second method CN. In this chapter, we mainly focus on CN consistent with previous studies [BT12, PNT07]. If we simply refer “pairwise method”, it refers to CN.

Suppose that we have \( N^+ \) cases and \( N^- \) controls \( (N^+ + N^- = N) \). Let \( V^+ \) be the set of case vertices and \( V^- \) be the set of control vertices. Let \( E^{++} \) be the set of all possible case/case vertex pairs, \( E^{--} \) be the set of all possible control/control vertex pairs, and \( E^{+-} \) be the set of all possible case/control vertex pairs. Let \( e_{ij} \) be 1 if there exists an edge between vertices \( i \) and \( j \) and 0 otherwise. Then, the CC and CN statistics are defined

\[
S_{CC} = \hat{IBD}_{case/case} - \hat{IBD}_{control/control} = \sum_{(i,j) \in E^{++}} e_{ij} \left( \binom{N^+}{2} \right) - \sum_{(i,j) \in E^{--}} e_{ij} \left( \binom{N^-}{2} \right)
\]

\[
S_{CN} = \hat{IBD}_{case/case} - \hat{IBD}_{non-case/case} = \sum_{(i,j) \in E^{++}} e_{ij} \left( \binom{N^+}{2} \right) - \sum_{(i,j) \in E^{--} \cup E^{+-}} e_{ij} \left( \binom{N^-}{2} + N^+N^- \right)
\]

The asymptotic distributions of these statistics are difficult to obtain analytically. This is because the statistics are based on the edges that depend on each other. For this
Figure 5.2: High correlation between CC and CN statistics. We simulated 1,000 studies under the alternative hypothesis (See Section 5.3.2). We then permuted phenotypes to simulate the null hypothesis. Spearman $\rho$ is 0.89 and 0.99 under the null and the alternative respectively.

reason, statistical significance is assessed by permutation. We assume a one sided test where IBD segments carry variants that are involved in disease [BT12].

The relationship between CC and CN is worth noting. Under the condition that the background IBD rates of control/control pairs and the non-case/case pairs are equivalent ($IBD_{\text{control/control}} = IBD_{\text{non-case/case}}$), CN will be more powerful than CC due to the additional $N^+N^-$ pairs it considers. We expect that, however, the relative ordering of the two statistics are similar to each other (Figure 2), since most of the pairs for both CC and CN are the same. As we will show below, we will utilize this similarity as the basis of increasing the computational efficiency of computing $S_{CN}$. 
5.2.3 Permutation test

Permutation is the standard approach for assessing significance in the pairwise method. A single permutation can be thought of as randomly sampling a vector of case/control disease status. Let \( \mathbf{v} = (v_1, ..., v_N), \forall v_i \in \{0, 1\} \) be the vector of disease status of \( N \) individuals where 0 denotes control and 1 denotes case. The test statistic of pairwise method, \( S_{CN} \), is a function of \( \mathbf{v} \). Let \( \hat{\mathbf{v}} \) be the case/control status that was originally observed in the data. The standard permutation test is equivalent to sampling new \( \mathbf{v} \) from all possible permutations of \( \hat{\mathbf{v}} \) assuming a uniform distribution. Let \( \mathcal{B} \) be the set of sampled \( \mathbf{v} \). The estimated p-value is

\[
\hat{p} = \frac{1}{|\mathcal{B}|} \sum_{\mathbf{v} \in \mathcal{B}} \delta(S_{CN}(\mathbf{v}) \geq S_{CN}(\hat{\mathbf{v}})) \tag{5.1}
\]

where \( \delta \) is the indicator function.

The drawback of the permutation test is that it is computationally inefficient. If the true p-value is very small, which is required in genome-wide studies, we will need a large number of permutations. For example, to assess a p-value \( p \) with standard error \( p/10 \), one needs approximately \( 100/p \) samples. For the genome-wide threshold of IBD association testing \((6 \times 10^{-6}, [BT12])\), more than 10 million permutations are required.

5.2.4 Fast-Pairwise

We develop a new method Fast-Pairwise that uses importance sampling technique to speed up CN method [Was04]. Unlike the standard permutation test that samples case/control status \( \mathbf{v} \) from the uniform distribution, we sample \( \mathbf{v} \) non-uniformly. Specifically, we aim to sample \( \mathbf{v} \) from all permutations of \( \hat{\mathbf{v}} \) such that \( S_{CN}(\mathbf{v}) \) will be similar to \( S_{CN}(\hat{\mathbf{v}}) \) on average. The intuition is that by intentionally sampling \( \mathbf{v} \) that
gives large value of $S_{CN}$, we can reduce the variance of the p-value estimate. Thus, our goal is to design a sampling procedure that satisfies

$$
\mathbb{E}_f(S_{CN}(v)) = S_{CN}(\tilde{v})
$$

(5.2)

where the expectation is with respect to $f$, our sampling distribution for $v$. However, designing such a sampling procedure is not straightforward. To this end, we leverage the fact that we can construct a simpler statistic that approximates $S_{CN}$ which we use for the sampling.

### 5.2.5 IBD-degreetype

In order to apply importance sampling, we must identify a statistic which roughly approximates $S_{CN}$ but can be conveniently used for designing a sampling procedure. Since we empirically have observed that $S_{CC}$ approximates $S_{CN}$ (Figure 2), we want to find a statistic that approximates $S_{CC}$. Our proposed statistic $S_{SUM}$ is related to $S_{CC}$ through a concept that we introduce called the IBD-degreetype which is simply the degree of each individual in the IBD graph. Obtaining the degrees of vertices is equivalent to splitting all edges and counting how many split edges are adjacent to each vertex (Figure 5.3). Then we assign these numbers to the vertices. Given this, we define the IBD-degreetype as conceptually similar to a genotype where the allele of each individual equals to the degree of the corresponding vertex in the IBD graph.

The IBD-degreetypes can be used for statistical testing for IBD association testing. The intuition is that if case/case pairs have an excessive number of IBDs, then case vertices will have higher degrees than control vertices. The test based on IBD-degreetype will be comparing the average degrees between cases and controls,

$$
S_{ID} = \sum_{i \in V^+} \frac{w_i}{N^+} - \sum_{i \in V^-} \frac{w_i}{N^-}
$$

(5.3)
where $w_i$ is the IBD-degreetype of individual $i$, or equivalently the degree of vertex $i$ in the graph. We note that this statistic is conceptually similar to the traditional association statistic which compares the frequency of the genotypes between the cases and controls. Here we instead compute the difference between the case and controls of the IBD-degreetypes (hence the name).

We are interested in the monotonic relationship between statistics within the permutation procedure. Let $v_1$ and $v_2$ be the permuted case/control status. We define *monotonic increasing relationship* (MIR) as follows,

**Definition 1.** Two statistics $S$ and $T$ are in a monotonic increasing relationship if,
\[
\forall v_1 \neq v_2, \quad S(v_1) \geq S(v_2) \iff T(v_1) \geq T(v_2).
\]

It is clear that if two statistics are in MIR, they will give the same p-value under permutation. It also follows that MIR is transitive (if $S$ and $T$ are in MIR and $T$ and $U$ are in MIR, then $S$ and $U$ are in MIR).

The IBD-degreetype test has the following relationship to the pairwise CC method. Figure 5.3 illustrates this relationship with a toy example.

**Lemma 1.** In a balanced study design ($N^+ = N^-$), CC and the IBD-degreetype test are in MIR.
Figure 5.3: Equivalence of pairwise (CC) method and IBD-degreetype test in a balanced study. Note that in pairwise CC method, the edge between (D,F) pair is ignored, which would not be ignored in CN method.
Proof. 

\[ S_{ID} = \sum_{i \in V^+} \frac{w_i}{N^+} - \sum_{i \in V^-} \frac{w_i}{N^-} = \frac{1}{N^+} \left\{ \sum_{i \in V^+} w_i - \sum_{i \in V^-} w_i \right\} \]

\[ = \frac{1}{N^+} \left\{ \left( \sum_{(i,j) \in E^{++}} e_{ij} + \sum_{(i,j) \in E^{+-}} e_{ij} \right) \right. \]

\[ - \left. \left( \sum_{(i,j) \in E^{--}} e_{ij} + \sum_{(i,j) \in E^{+-}} e_{ij} \right) \right\} \]

\[ = \frac{2}{N^+} \left\{ \sum_{(i,j) \in E^{++}} e_{ij} - \sum_{(i,j) \in E^{--}} e_{ij} \right\} \]

\[ = \frac{2}{N^+} \left( \frac{N^+}{2} \right) \left\{ \sum_{(i,j) \in E^{++}} \frac{e_{ij}}{\binom{N^+}{2}} - \sum_{(i,j) \in E^{--}} \frac{e_{ij}}{\binom{N^-}{2}} \right\} \]

\[ = \frac{2}{N^+} \left( \frac{N^+}{2} \right) S_{CC} = (N^+ - 1) S_{CC} \]

Since \( S_{ID} \) and \( S_{CC} \) differ by a constant multiplication factor \((N^+ - 1)\), they are in MIR.

We introduce another simple form of test statistic based on IBD-degreetype. This sum statistic is the sum of IBD-degreetype alleles or the degrees of the vertices in cases,

\[ S_{SUM} = \sum_{i \in V^+} w_i \]

Lemma 2. \( S_{ID} \) and \( S_{SUM} \) are in MIR.

Proof. Note that

\[ \sum_{(i,j) \in E^{++}} e_{ij} + \sum_{(i,j) \in E^{+-}} e_{ij} + \sum_{(i,j) \in E^{--}} e_{ij} = |e| \]

where \(|e|\) denotes the total count of edges. In addition, the sum of the IBD-degreetypes over all vertices is equal to \(2|e|\) because each edge is counted twice.

\[ \sum_{i \in V^+} w_i + \sum_{i \in V^-} w_i = 2|e| \]
Pairwise method

\[
\begin{array}{c|c|c|c}
S_{CN} & S_{CC} & S_{ID} & S_{SUM} \\
\hline
\text{Correlated} & \text{MIR in balanced design (Lemma 1)} & \text{MIR (Lemma 2)} & \text{Approximate}
\end{array}
\]

Figure 5.4: Relationship between different statistics.

Therefore,

\[
S_{ID} = \sum_{i \in V^+} \frac{w_i}{N^+} - \sum_{i \in V^-} \frac{w_i}{N^-} = \left( \frac{1}{N^+} + \frac{1}{N^-} \right) \sum_{i \in V^+} w_i - \frac{2|e|}{N^-}
\]

Since \( \left( \frac{1}{N^+} + \frac{1}{N^-} \right) > 0 \) and \( \frac{2|e|}{N^-} \) is a constant, \( S_{SUM} \) is a monotonic increasing linear transformation of \( S_{ID} \). Thus, they are in MIR.

\[\square\]

5.2.6 Substitution strategy

Here we propose to use \( S_{SUM} \) in sampling as a substitution to the pairwise method statistic, \( S_{CN} \). The logical ground for this strategy comes from the relationship between \( S_{CN} \) and \( S_{SUM} \). We have empirically shown that \( S_{CN} \) and \( S_{CC} \) are highly correlated (Figure 5.2). Since \( S_{CC} \) and \( S_{ID} \) are in MIR in a balanced study (Lemma 1), we expect that they will be correlated in general even in an unbalanced study and we show this property through a simulation experiment (Supplementary Figure 1). \( S_{ID} \) is in MIR with \( S_{SUM} \) (Lemma 2). Thus, \( S_{SUM} \) can be an approximation to \( S_{CN} \) (Figure 5.4).

Given this relationship between \( S_{CN} \) and \( S_{SUM} \), our strategy is to sample \( v \) such that
$S_{SUM}(v)$ will be similar to $S_{SUM}(\hat{v})$ on average. Our new goal can be described

$$\mathbb{E}_f(S_{SUM}(v)) = S_{SUM}(\hat{v})$$

(5.4)

It turns out that this new goal is much easier to achieve. Note that $S_{SUM}$ is used only for sampling. After the sampling is done, the sampled $v$ are used to approximate the p-value of CN method.

This substitution approach works because in importance sampling, the sampling distribution needs not guarantee optimality (the smallest variance of p-value estimate). Instead, a reasonably similar distribution to the optimal distribution suffices. It is clear that this strategy will perform the best if the balanced study condition is met. However, even if the condition is not met, only the variance of p-value estimate is affected and not the mean. The p-value estimate will still be unbiased and it only means that we will need a larger number of samples to obtain the same accuracy.

5.2.7 Sampling with replacement

In this section, we devise a sampling procedure satisfying equation (5.4). Such a sampling procedure will be the core part in our importance sampling framework for speeding up CN method.

Sampling a random $v$ from all permutations of $\hat{v}$ can be thought of as sampling $N^+$ out of $N$ individuals that will be assigned case status, or equivalently, sampling $N^+$ case indices among 1, ..., $N$. Let $a_1, ..., a_{N^+}$ be the sampled case indices. These are the indices in $v$ that will be assigned “1”. Sampling case indices is without-replacement sampling procedure; we cannot sample the same index twice, so that exactly $N^+$ distinct indices will be sampled at the end ($\forall i \neq j, a_i \neq a_j$). This way, we can restrict sample space of $v$ to the set of all permutations of $\hat{v}$.

However, the design of sampling procedure satisfying equation (5.4) is consider-
ably easier if we assume sampling case indices with replacement. That is, we allow the same index can be sampled multiple times. Although this assumption is not valid for our purpose, our strategy is to devise a sampling approach satisfying equation (5.4) assuming sampling with replacement first, and then extend the approach to the sampling without replacement.

Suppose that we pick \( a_1 \) among \( 1, \ldots, N \) with probability \( P(a_1 = k) = g(k) \), \( \sum_{k=1}^{N} g(k) = 1 \). Since we assume sampling with replacement, sampling \( a_2 \) is no different from sampling \( a_1 \); in fact, for any \( 1 \leq i \leq N^+ \), \( a_i \) is independent and identically distributed (IID) with distribution \( g \). Now consider \( w_{a_1} \), the IBD-degreetype allele of \( a_1 \). Let \( \mathbb{E}_g(w_{a_1}) \) be the expected value of \( w_{a_1} \) with respect to \( g \). Again, since we assume sampling with replacement, \( \mathbb{E}_g(w_{a_1}) = \ldots = \mathbb{E}_g(w_{a_{N^+}}) \).

Then, by the definition of \( S_{SUM} \), we can easily see that

\[
\mathbb{E}_g(S_{SUM}) = N^+ \mathbb{E}_g(w_{a_1})
\]

Thus, equation (5.4) can be described

\[
N^+ \mathbb{E}_g(w_{a_1}) = S_{SUM}(\hat{v})
\]

or

\[
\mathbb{E}_g(w_{a_1}) = \frac{S_{SUM}(\hat{v})}{N^+} \quad (5.5)
\]

where the left hand side is the expected case IBD-degreetype allele in our distribution \( g \) and the right hand side is the average case IBD-degreetype allele in the observation \( \hat{v} \). This shows that, if the p-value is highly significant (e.g. the right hand side is large), we should pick \( a_1 \) (and all \( a_i \)) such that the expected value of IBD-degreetype allele can be large.

Here we propose a new sampling procedure that satisfies condition (5.5). We define distribution \( g \) as follows;

\[
P(a_1 = k) = g(k) \propto t_k \quad \text{where} \quad t_k = 1 + \rho w_k
\]
and
\[ \rho = \frac{N \frac{S_{SUM}(\hat{v})}{N^+} - \sum_{k=1}^{N} w_k}{\sum_{k=1}^{N} w_k^2 - \frac{S_{SUM}(\hat{v})}{N^+} \sum_{k=1}^{N} w_k} \] (5.6)

It is easy to show that this sampling procedure meets condition (5.5), since condition (5.5) can be described
\[ \mathbb{E}_g(w_{a_1}) = \frac{\sum_{k=1}^{N} t_k w_k}{\sum_{k=1}^{N} t_k} = \frac{\sum_{k=1}^{N} (1 + \rho w_k) w_k}{\sum_{k=1}^{N} (1 + \rho w_k)} = \frac{S_{SUM}(\hat{v})}{N^+} \]
and by solving this for \( \rho \) we exactly have equation (5.6). If \( \rho < 0 \), then we set \( \rho \) to zero to prevent negative \( t_k \). Such a case is not of our interest at any rate, since we focus on the one-sided test for detecting excess of case/case IBD. We choose the most simple linear function for \( t_k \), which enables us to calculate \( \rho \) easily. As a result, for any \( \hat{v} \), we can calculate \( \rho \) and completely define the distribution \( g \). So far, we have successfully defined a sampling procedure satisfying equation (5.4) assuming sampling with replacement.

### 5.2.8 Sampling without replacement

In this section, we extend the sampling procedure from the previous section to the sampling without replacement. Here we propose to heuristically apply the same sampling scheme based on \( t_1, ..., t_N \) to the without-replacement context. When we pick \( a_i \) (ith case index), we pick index \( k \) among \( \{1, ..., N\} \setminus \{a_1, ..., a_{i-1}\} \) with probability \( \frac{t_k}{\left( \sum_{j=1}^{N} t_j - \sum_{l=1}^{i-1} t_{a_l} \right)} \). That is, we assume the same sampling probability proportional to \( t_k \), but we exclude indices previously picked as cases from our consideration.

However, this sampling procedure does not exactly satisfy equation (5.4) in the without-replacement sampling context. The indices with larger IBD-degreetype alleles are likely to be picked as cases earlier in the procedure and removed. Thus, if we use \( \rho \) calculated assuming sampling with replacement, the expected case IBD-degreetype allele (the left hand side of (5.5)) will be smaller than what we would obtain in the
with-replacement context. To compensate for this difference, we use the following heuristic. In the middle of sampling, we empirically assess the left hand side of (5.4) by examining the currently obtained samples. Then, we dynamically increment $\rho$ until the left hand side of (5.4) is close to the right hand side. This simple heuristic is sufficient because, again, in importance sampling, we only need to approximately satisfy equation (5.4).

### 5.2.9 P-value calculation

By using the sampling procedure developed in the previous section, we can obtain many sample $v$ that approximately satisfy equation (5.2). The final step is to use these samples to assess the p-value of pairwise (CN) method. In importance sampling, we must account for the fact that we used a sampling distribution that is different from the original distribution. Our original distribution is the uniform distribution defined by the standard permutation approach. The sampling distribution is defined by the sampling procedure that we developed in the previous section.

For a given $v$, the probability of sampling $v$ differs in the two distributions as follows. In order to sample a $v$, we sample the case indices $a_1, ..., a_{N+}$. The probability of sampling case indices in the standard uniform distribution is

$$P_{\text{Uniform}}(v) = \frac{1}{N \cdots (N - N^+ + 1)}$$

On the other hand, the probability of sampling case indices in our sampling procedure described in the previous section is

$$P_{\text{New}}(v) = \prod_{k=1}^{N^+} \left\{ t_{a_k} / \left( \sum_{j=1}^{N} t_j - \sum_{l=1}^{k-1} t_{a_l} \right) \right\}$$

This is because at each step we pick $i$th case index $a_i$, we sample the index with probability proportional to $t_{a_i}$, where the previously picked indices $a_1, ..., a_{N^+}$ are excluded.
from consideration. Using the standard formula of importance sampling, we approximate the p-value

\[ \hat{p} = \frac{1}{|B|} \sum_{v \in B} \frac{P_{\text{Uniform}}(v)}{P_{\text{New}}(v)} \delta(S_{CN}(v) \geq S_{CN}(\hat{v})) \]

Note that we use the pairwise CN statistic in this formula. We use \( S_{SUM} \) only to facilitate the sampling of \( v \), and then use the obtained samples for CN method at the final step.

We can approximate the variance of \( \hat{p} \) with the following formula

\[ \left\{ \frac{1}{|B|} \sum_{v \in B} \left( \frac{P_{\text{Uniform}}(v)}{P_{\text{New}}(v)} \right)^2 \delta(S_{CN}(v) \geq S_{CN}(\hat{v})) - \hat{p}^2 \right\} / |B| \]

5.2.10 Adjusting for population structure

A simple correction for population structure has been previously proposed [PNT07, BT12] for the pairwise method. In this simple approach, the genomic average is subtracted from each of the two contrasting terms of the statistic. For example, in CN method, the genomic average of case/case IBD rate is subtracted from the observed case/case IBD rate and the genomic average of non-case/case IBD rate is subtracted from the observed non-case/case IBD rate before calculating the statistic. The same approach can be applied to our Fast-Pairwise method.

5.3 Results

5.3.1 Efficiency

To assess the efficiency gain of our Fast-Pairwise method, we use the Wellcome Trust Case Control Consortium (WTCCC) data [Wel07]. We first run Beagle FastIBD to detect IBD between individuals [BB11]. Then we test individual IBD segments for
Table 5.1: Running time for pairwise IBD association testing for the WTCCC whole genome data

<table>
<thead>
<tr>
<th>Method</th>
<th>10^7 permutations</th>
<th>Adaptive perm.</th>
<th>10^4 samples</th>
<th>10^3 samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traditional Pairwise</td>
<td>35 years</td>
<td>474 days</td>
<td>40 days</td>
<td>4 days</td>
</tr>
<tr>
<td>Fast-Pairwise</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

We perform IBD association testing using both the traditional pairwise method based on permutation and our Fast-Pairwise. We perform 10 million permutations for the traditional pairwise method. For our Fast-Pairwise method, we perform importance sampling with 1,000 samples and 10,000 samples. We implemented both methods in the Java programming language.

Table 5.1 shows the estimated running time of both methods for analyzing the whole genome data (500,000 SNPs) of a single disease. The time is extrapolated from the estimated time for chromosome 22. Our Fast-Pairwise method is several orders of magnitude faster than the traditional pairwise method. It only takes 4 days for the whole genome while the traditional method can take 13,000 days or 35 years of CPU time.

We can reduce the computation time for the traditional pairwise method by employing an adaptive permutation approach. We can terminate the permutation earlier if the p-value approximates to a non-significant value. Given a p-value $p$, we need $100/p$ permutations to obtain the standard error of $\sim p/10$. Suppose that we sample $100/p$ permutations for each p-value with upper limit of 10 millions. When we apply this adaptive approach to the WTCCC type 1 diabetes data, the estimated computation time is 474 days. Thus, Fast-Pairwise is still an order of magnitude faster than the traditional pairwise method with an adaptive permutation approach.
5.3.2 Accuracy

To assess the accuracy of our importance sampling, we use the simulation framework similar to [BT12]. Using the HapMap ENCODE regions [Int05], we run HapGen2 to simulate 10,000 individuals [SMD11]. These individuals define our founder population. Then we simulate the first generation by sampling 100,000 individuals from the founders. Then we simulate the second generation by sampling 100,000 individuals from the first generation. We repeat until we obtain the 25th generation. Finally, we use the 25th generation to simulate a case/control study. Within the ENCODE region, we randomly select 5 causal variants among all rare variants (minor allele frequency \(< 1\%\)). If a haplotype contains one or more causal variants, it confers risk with relative risk selected from uniform(3,10). We assume the disease prevalence of 0.1. Given this disease model, using the standard formula of the case and control minor allele frequencies [HKE09a], we sample 1,500 cases and 1,500 controls from the 25th generation. The IBD information between a pair of individuals is determined by tracking if they are descendants of the same founder. We repeat this simulation 100 times per each of the ten ENCODE regions to generate 1,000 sets of case/control studies.

Given these case/control study sets, we assess the p-values of pairwise (CN) method using both the standard permutation and the importance sampling of Fast-Pairwise. We use 10,000 samples for importance sampling and compare to $10^4$, $10^5$, and $10^6$ permutations. Figure 5.5 shows that the p-values of two methods track very well within the p-value range that permutation can approximate (up to p-values of $10^{-4}$, $10^{-5}$, and $10^{-6}$ respectively). Within this range, the Pearson correlation $r^2$ of two log p-values are 0.98, 0.94, and 0.99 respectively. This shows that our importance sampling procedure obtains accurate p-values.

Moreover, Figure 5.5 emphasizes a fundamental difference between the two methods. In permutation, the range of p-values one can obtain is limited by the number
Figure 5.5: Accuracy of importance sampling. In simulations using the HapMap ENCODE region, we assess the p-values of pairwise (CN) method using both the standard permutation and Fast-Pairwise (importance sampling). The vertical dashed line denotes the lower bound of p-value that permutation can approximate given the number of permutations. The red dots denote the simulations where none of the permutations exceeds the observed statistic and therefore the lower bound of p-value is reported by the permutation test. The green triangles in (C) are the p-values that we performed extra permutations (> $10^6$).

of permutations. Given $|B|$ permutations, if none of the permutations exceeds the observed statistic, a conservative approximation of p-value is $1/|B|$. By contrast, in Fast-Pairwise, the p-value range is not bounded by the number of samples. With a relatively small number of samples (10,000), Fast-Pairwise can obtain accurate p-values comparable to the permutation test for a wide range of p-values. We also performed extra permutations (> $10^6$) to estimate p-values between $10^{-6}$ and $10^{-8}$. The p-values of the two methods are still consistent within this p-value range (green triangles in Figure 5.5C).
Figure 5.6: IBD association testing results of the WTCCC type 1 diabetes data. (A) The results that we would obtain if we use the traditional pairwise method. Given 5 million permutations, the smallest p-value is bounded to $2 \times 10^{-7}$. The top peak is stretched over >8Mb region complicating the fine-mapping. (B) Fast-Pairwise results. The top peak is at TAP2 whose closest HLA gene is HLA-DQB1. In both plots, the dashed green horizontal line denotes genome-wide threshold $6 \times 10^{-6}$. 
5.3.3 Application to WTCCC type 1 diabetes data

[BT12] applied the pairwise (CN) method to the WTCCC type 1 diabetes (T1D) data based on the Beagle FastIBD IBD mapping results [BB11]. Using 5 million permutations, they found that the major histocompatibility complex (MHC) region in the chromosome 6 is statistically significant given the genome-wide threshold $6 \times 10^{-6}$. Since the MHC association to T1D has been historically known [TBM87], this result was a validation that IBD association testing can detect the true association signal.

The limitation of [BT12]'s permutation approach is that although it is possible to determine whether each test is significant ($p < 6 \times 10^{-6}$) using 5 million permutations, it is not possible to approximate much smaller p-values. Given 5 million permutations, the smallest p-value one can approximate is bounded to $2 \times 10^{-7}$. In the MHC region, due to the strong signal and the long linkage disequilibrium, the location of the top peak of p-value is important for interpreting and fine-mapping the results. Figure 5.6A shows that the top peak of p-value is stretched over a wide region (>8Mb) including all eight classical HLA genes. Therefore, it is difficult to interpret which HLA gene is likely to be involved in the association.

We applied our Fast-Pairwise to the same dataset. Since Fast-Pairwise is the same pairwise method with increased efficiency, we expected to see the similar results as [BT12]. Indeed, we discovered significant associations within the MHC region. However, the difference is that since our method can approximate very small p-values well beyond the genome-wide threshold, it is possible to localize the statistical signal to a single marker. (Figure 5.6B). The top hit is at SNP rs241432 ($p = 7 \times 10^{-45}$) at the intron of TAP2 gene. Among all major class I and II HLA genes, the closet gene to this peak is HLA-DQB1 (150kb upstream from the peak). It is historically known that the main contributing genes for the MHC association to T1D is HLA-DQB1 [TBM87]. Thus, this result demonstrates that our Fast-Pairwise method can pinpoint the causal
gene among many HLA genes within the MHC region, while the traditional pairwise method cannot.

One interesting observation is that the peak association of our IBD association test is on the \textit{TAP2} gene that encodes antigen peptide transporter 2 and has been shown to confer independent risk to the T1D when conditioned on the DQ haplotypes [QLM07]. Thus, the peak revealed by our Fast-Pairwise method may imply the added effect of \textit{TAP2} in addition to the primary effect of \textit{HLA-DQB1} which is in linkage disequilibrium.

\subsection*{5.4 Discussion}

We have developed a new efficient method for pairwise IBD association testing called \textit{Fast-Pairwise}. Fast-Pairwise employs importance sampling and can perform the pairwise method more efficiently than the traditional method based on permutation. Moreover, unlike permutation, Fast-Pairwise can approximate extremely small p-values beyond the genome-wide threshold. Using the WTCCC type 1 diabetes data, we show that Fast-Pairwise can successfully pinpoint a gene known to be associated to the disease within the MHC region.

The true utility of the IBD association testing is on finding novel loci where there are potentially multiple rare variants which can not be found using single SNP tests [BT12]. An important advantage of IBD association testing is its wide applicability. The analysis can be performed using the same genotype data collected for single SNP tests without incurring additional cost. For this reason, we feel that many investigators will apply our method to search for these additional loci bearing rare causal variants. What is preventing researchers from applying this approach is an efficient method for IBD association testing which we provide in this chapter. We expect that our new
method will promote the wide use of IBD association testing and facilitate further research on the power and utilities of IBD association testing.

Reference to published article

5.5 Supplementary Section
Figure 5.7: Supplementary Figure 1. Correlation between different statistics. We simulated 1,000 studies under a disease model assuming balanced study design (See Section 3.2). Each study consists of 1,500 cases and 1,500 controls. (A) We measured the correlation between different statistics under the balanced study design. Then we sub-sampled 750 cases to simulate unbalanced study design (#case:#control=1:2). (B) We measured the correlation between statistics under the unbalanced study design. The numbers on the upper pane are the Spearman correlation between statistics.
### Figure 5.8

<table>
<thead>
<tr>
<th></th>
<th>CN</th>
<th>1.00</th>
<th>0.98</th>
<th>0.97</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CC</td>
<td>0.99</td>
<td></td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td>ID</td>
<td></td>
<td></td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td>SUM</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
CHAPTER 6

Meta-analysis identifies Gene-by-environment Interactions as Demonstrated in a Study of 4,965 Mice

6.1 Background

Identifying environmentally specific genetic effects is a key challenge in understanding the structure of complex traits. In humans, gene-by-environment (GxE) interactions have been widely discussed [GLR10, MMF12, SK08, Tal07, FET12, OBF12, DWH13, PCK13, WKZ12, MXX12, GNS12, WWM12] yet only a few have been replicated. One reason for this discrepancy is the inability to accurately control for environmental conditions in humans as well as the inability to observe the same individuals in multiple distinct environments. Model organisms do not share such difficulties and for this reason can play a crucial role in the identification of gene-by-environment interactions. For example, in many mouse genetic studies the same traits are examined under different environmental conditions. Specifically, knock-out or diet-controlled mice are often utilized in the study of cholesterol levels. The availability of these studies presents a unique opportunity to identify genomic loci involved in gene-by-environment interactions as well as those loci involved in the trait independent of the environment.

In order to utilize genetic studies in model organisms to identify gene-by-environment interactions, one needs to directly compare the effects of genetic variations in studies conducted under different conditions. This practice is complicated for a number of
reasons, when combining more than two studies. First, environmental conditions are often variable across studies and do not fit to the standard univariate model for interactions. For example, in one study, cholesterol may be examined under different diet conditions (e.g. low fat and high fat) and then in another study cholesterol is examined using gene knockouts. In this case, it is not straightforward to analyze these studies in aggregate using a single variable to represent the environmental condition. Applying a multivariate model, one in which the environment is represented using multiple environmental variables, results in increased degrees of freedom and low statistical power. Second, model organisms such as the mouse exhibit a large degree of population structure. Population structure is well-known for causing false positives and spurious associations [PPP06, DRW01] in association analysis and can be expected to complicate the ability to combine separate studies.

In this paper, we propose a random-effects based meta-analytic approach to combine multiple studies conducted under varying environmental conditions and show that this approach can be used to identify both genomic loci involved in gene-by-environment interactions as well as those loci involved in the trait independent of the environment. By making the connection between gene-by-environment interactions and random effects model meta-analysis, we show that interactions can be interpreted as heterogeneity and detected without requiring uni- or multi-variate models. We also define an approach for correcting population structure in the random effects model meta-analysis, extending the methods developed for fixed effects model meta-analysis [FKV12]. We show that this method enables the analyses of large scale meta-analyses with dozens of heterogeneous studies and leads to dramatic increases in power. We demonstrate that insights regarding gene-by-environment interactions are obtained by examining the differences in effect sizes among studies facilitated by the recently developed m-value statistic [HE12], which allows us to distinguish between studies having an effect and studies not having an effect at a given locus.
We applied our approach, which we refer to as Meta-GxE, to combine 17 mouse High-density lipoprotein (HDL) studies containing 4,965 distinct animals. To our knowledge, this is the largest mouse genome-wide association study conducted to date. The environmental factors of the 17 studies vary greatly and include various diet conditions, knock-outs, different ages and mutant animals. By applying our method, we have identified 26 significant loci. Consistent with the experience of meta-analysis in human studies, our combined study finds many loci which were not discovered in any of the individual studies. Among the 26, 24 loci have been previously implicated in having an effect on HDL cholesterol or closely related lipid levels in the blood, while 2 loci are novel findings. In addition, our study provides insights into genetic effects on several disease loci and their relationship between environment and sex. For example, we identified 3 loci (Chr10:21399819, Chr19:3319089, ChrX:151384614), where female mice show a more significant effect on HDL phenotypes than male mice. We also identified 7 loci (Chr1:171199523, Chr8:46903188, Chr8:64150094, Chr8:84073148, Chr10:90146088, Chr11:69906552, Chr15:21194226) where male mice show a more significant effect on HDL than female mice. In addition, many of the loci show strong gene-by-environment interactions. Using additional information describing the studies and our predictions of which studies do and do not contain an effect, we gain insights into the interaction. For example, locus on chromosome 8 (Chr8:84073148) shows a strong sex by mutation-driven LDL level interaction, which affects HDL cholesterol levels.

Part of the reason for our success in identifying a large number of loci is that our study combined multiple mouse genetic studies many of which use very different mapping strategies. Over the past few years, many new strategies have been proposed beyond the traditional F2 cross [FE12] which include the hybrid mouse diversity panel (HMDP) [BFO10b, GRF12], heterogeneous outbred stocks [VSG06], commercially available outbred mice [YNB10], and the collaborative cross [AVF11]. In our current
study, we are combining several HMDP studies with several F2 cross studies and benefit from the statistical power and resolution advantages of this combination [FKV12]. The methodology presented here can serve as a roadmap for both performing and planning large scale meta-analysis combining the advantages of many different mapping strategies.

6.2 Results

6.2.1 Discovering environmentally-specific loci using meta-analysis

The Meta-GxE strategy uses a meta-analytic approach to identify gene-by-environment interactions by combining studies that collect the same phenotype under different conditions. Our method consists of four steps. First, we apply a random effects model meta-analysis (RE) to identify loci associated with a trait considering all of the studies together. The RE method explicitly models the fact that loci may have different effects in different studies due to gene-by-environment interactions. Second, we apply a heterogeneity test to identify loci with significant gene-by-environment interactions. Third, we compute the \( m \)-value of each study to identify in which studies a given variant has an effect and in which it does not. Forth, we visualize the result through a forest plot and PM-plot to understand the underlying nature of gene-by-environment interactions.

We illustrate our methodology by examining a well-known region on mouse chromosome 1 harboring the \( Apoa2 \) gene, which is known to be strongly associated with HDL cholesterol [WHQ93]. Figure 6.1 shows the results of applying our method to this locus. We first compute the effect size and its standard deviation for each of the 17 studies. These results are shown as a forest plot in Figure 6.1 (a). Second we compute the \( P \)-value for each individual study also shown in Figure 6.1 (a). If we were to fol-
low traditional methodology and evaluate each study separately, we would declare an
effect present in a study if the $P$-value exceeds a predefined genome-wide significance
threshold ($P < 1.0 \times 10^{-6}$). In this case, we would only identify the locus as associated
in a single study, HMDP-chow(M) ($P = 6.84 \times 10^{-9}$). On the other hand, in our ap-
proach, we combine all studies to compute a single $P$-value for each locus taking into
account heterogeneity between studies. This approach leads to increased power over
the simple approach considering each study separately. The combined meta $P$-value
for the Apoa2 locus is very significant ($4.41 \times 10^{-22}$), which is consistent with the
fact that the largest individual study only has 749 animals compared to 4,965 in our
combined study.

In order to evaluate how significantly different the effect sizes of the locus are be-
tween studies, we apply a heterogeneity test. The statistical test is based on Cochran’s
Q test [DL86a, Coc09], which is a non-parametric test for testing if studies have the
same effect or not. In this locus, the effect sizes are clearly different and not surpris-
ingly the $P$-value of the heterogeneity test is significant ($5.80 \times 10^{-5}$). This provides
strong statistical evidence of a gene-by-environment interaction at the locus. Below
we more formally describe how heterogeneity in effect size at a given locus can be
interpreted as gene-by-environment interaction.

If a variant is significant in the meta-analytic testing procedure, then this implies
that the variant has an effect on the phenotype in one or more studies. Examining in
which subset of the studies an effect is present and comparing to the environmental
conditions of the studies can provide clues to the nature of gene-by-environment in-
teractions at the locus. However, the presence of the effect may not be reflected in
the study-specific $P$-value due to a lack of statistical power. Therefore, it is difficult
to distinguish only by a $P$-value if an effect is absent in a particular study due to a
gene-by-environment interaction at the locus or a lack of power. In order to identify
which studies have effects, we utilize a statistic called the m-value [HKE09b], which estimates the posterior probability of an effect being present in a study given the observations from all other studies. We visualize the results through a PM-plot, in which

\( P \)-values are simultaneously visualized with the m-values at each tested locus. These plots allow us to identify in which studies a given variant has an effect and in which it does not. M-values for a given variant have the following interpretation: a study with a small m-value (\( \leq 0.1 \)) is predicted not to be affected by the variant, while a study with a large m-value (\( \geq 0.9 \)) is predicted to be affected by the variant.

The PM-plot for the \textit{Apoa2} locus is shown in Figure 6.1 (b). If we only look at the separate study \( P \)-values (y-axis), we can conclude that this locus only has an effect in HMDP-chow(M). However, if we look at m-value (x-axis), then we find 8 studies (HMDPxB-ath(M), HMDPxB-ath(F), HMDP-chow(M), HMDP-fat(M), HMDP-fat(F), BxD-db-5(M), BxH-apoe(M), BxH-apoe(F)), where we predict that the variation has an effect, while in 3 studies (BxD-db-12(F), BxD-db-5(F), BxH-wt(F)) we predict there is no effect. The predictions for the remaining 6 studies are ambiguous.

From Figure 1, we observe that differences in effect sizes among the studies are remarkably consistent when considering the environmental factors of each study as described in Table 1. For example, when comparing study 1 – 4, the effect size of the locus decreases in both the male and female HMDPxB studies in the chow diet (chow study) relative to the fat diet (ath study). Thus we can see that when the mice have Leiden/CETP transgene, which cause high total cholesterol level and high LDL cholesterol level, effect size of this locus on HDL cholesterol level in blood is affected by the fat level of diet. Similarly, when comparing study 12 – 15, the knockout of the \textit{Apoe} gene affects the effect sizes for both male and female BxH crosses. However, in the BxD cross (study 8 – 11), where each animal is homozygous for a mutation causing a deficiency of the leptin receptor, the effect of the locus is very strong in the young
male animals, while as animals get older and become fatter, the effect becomes weaker. However in the case of female mice, the effect of the locus is nearly absent at both 5 and 12 weeks of age. Thus we can see that sex plays an important role in affecting HDL when the leptin receptor activity is deficient. We note that there are many genes in this locus and the genetic mechanism of interactions may involve genes other than Apoa2. Despite this caveat, the results of Meta-GxE at this locus provides insights into the nature of GxE and can provide a starting point for further investigation.

We note that an alternate explanation for differences in effect sizes between studies is the presence of gene-by-gene interactions and differences in the genetic backgrounds of the studies. While this is a possible explanation for differences in effect sizes between the different crosses and the HMDP studies, in Figure 6.1, we see many differences in effect sizes among studies with the same genetic background. Thus gene-by-gene interactions can only partially explain the differences in observed effect sizes.

6.2.2 The connection between random effects meta-analysis and gene-by-environment interactions

Gene-by-environment interactions, random effects meta-analysis and heterogeneity testing are closely related. Suppose we have \( k \) studies each conducted under different environmental conditions. We define the following linear model, where \( y_i \) is the observed phenotype for study \( i \), \( \alpha_i \) is the phenotype mean for study \( i \), \( \delta_i \) is the genetic effect on the phenotype for study \( i \), \( X \) is the genotype, and \( e \) is the residual error.

\[
y_i = \alpha_i + \delta_i X + e
\]  

(6.1)

Since each environment is different, the effect size \( \delta_i \) is partially determined by
environmentally-specific factors and partially determined by factors common to all studies. Given that we can decompose the effect $\delta_i$ into environment-independent and environment-dependent factors. Then we define the following linear model, where $\beta$ is the environment-independent genetic effect and $\gamma_i$ is the environment-dependent genetic effect for study $i$.

$$y_i = \alpha_i + \beta X + \gamma_i X + e$$  \hspace{1cm} (6.2)

In order to test for the presence of an effect shared across environments, we test the null hypothesis $\beta = 0$ and to test for the presence of a gene-by-environment interaction, we test the hypothesis that $\gamma_i = 0$.

In the random effects meta-analysis, we assume that the effect size $\delta_i$ is sampled from a normal distribution with mean $\mu$ and variance $\tau^2$, denoted $\delta_i \sim N(\mu, \tau^2)$. Under this assumption, we test the null hypothesis $\mu = 0$ and $\tau^2 = 0$, in order to obtain a study-wide $P$-value. Additionally, we perform a heterogeneity test to test the null hypothesis $\delta_1 = \ldots = \delta_k$ versus the alternative hypothesis $NOT(\delta_1 = \ldots = \delta_k)$. We posit that by conducting hypotheses tests in the meta-analysis framework, we are simultaneously testing for the presence of environmentally-independent and environmentally-specific effects and that by applying heterogeneity testing we are testing for only environmentally-specific effects.

Consider that in the meta-analysis framework $\mu$ is analogous to $\beta$ and the variation ($\tau^2$) around $\mu$ is analogous to variation among $\gamma_i$s. In the random effects meta analysis testing framework we are testing if $\mu = 0$ and $\tau^2 = 0$. This is equivalent to testing both environmentally-independent ($\beta = 0$) and environmentally-dependent ($\gamma_i = 0$) effects simultaneously. In heterogeneity testing, we test the null hypothesis $\delta_1 = \ldots = \delta_k$ versus the alternative hypothesis $NOT(\delta_1 = \ldots = \delta_k)$. When the environment-dependent effect ($\gamma_i$) is 0 it means that $\tau^2 = 0$ and thus $\delta_1 = \ldots = \delta_k$. When $\tau^2 \neq 0$,
we expect that $\delta_i$ will vary around $\mu$, so that we do not expect that $\delta_i = \delta_j$. Since the variation ($\tau^2$) of $\delta_i$ around $\mu$ is analogous to the variable $\gamma_i$, heterogeneity testing in the meta-analysis framework is approximately equivalent to testing for environmentally-specific effects.

### 6.2.3 Gene-by-environment interactions are prevalent in mouse association studies

The presence of heterogeneity in the effect size at causal genetic loci due to gene-by-environment interactions is naturally expected in mouse genetic studies when combining studies with varying environmental conditions. One extreme example comes from a knock-out experiment. If the knocked-out gene is causal for a particular trait, then we can expect that the gene would have no effect on a knock-out mouse, while the gene would have an effect on the wild type mouse. This is a binary form of heterogeneity. In a less extreme form of heterogeneity, the effect of a given gene may be affected by an environmental factor which varies in different mice – ranging from small effects to large effects.

To see the relationship between significance of the association and gene-by-environment interactions, we compute and compare this $P$-value for each SNP from the 17 studies using the random effects meta-analysis to a measure of heterogeneity. Heterogeneity can be assessed by $I^2$ statistic, which describes the percentage of variation across studies that is due to heterogeneity rather than chance [HT02].

Figure 6.2 compares $I^2$ statistic with the meta-analysis $P$-value for each SNP. In this figure, we see that $I^2$ is uniformly distributed for the non-significant SNPs (blue dots), while it is right skewed for significant SNPs (red dots), indicating that more significant SNPs have a greater potential for exhibiting heterogeneity in effect. Since heterogeneity in this case can be interpreted as representing gene-by-environment in-
interactions, as heterogeneity is induced by differences in the environment, we see that
the presence of a GxE interaction confers higher power to detect an association.

6.2.4 Power of meta-analysis for detecting gene-by-environment interactions

The power to identify both gene-by-environment and main effects in a meta-analysis
of mouse studies depends on both the main effect size and the amount of heterogeneity.
We performed simulations using the genotypes of the 17 mouse studies analyzed in this
paper. We simulated a range of main effect (mean effect) sizes and a range of gene-by-
environment effects. We are simulating the realistic scenario in which we do not know
exactly the set of covariates which are responsible for the gene-by-environment effects.
We simulated gene-by-environment effects by drawing the effect in each study from a
distribution with a mean given by the main effect size and a variance controlling the
magnitude of gene-by-environment interactions. If this variance is small, then all of
the studies have close to the same effect size and there are few gene-by-environment
effects. If the variance is high, then there are strong gene-by-environment effects.
Figure 6.3 shows the results of our simulations. 1000 simulated phenotypes were
generated for each mean and variance pair. Statistical power is estimated by computing
the proportion of the datasets in which a simulated effect is detected. We observe that
the power is high for a wide range of main effect sizes and gene-by-environment effect
sizes which is explained by the large sample size of the study. We also observe that
even for small main effects, if there are strong gene-by-environment effects, we can
still identify the locus. This is because in this case a subset of the studies will have
strong effect sizes due to gene-by-environment effects.

Our approach is not the only way to analyze a meta-analysis study. We compare
the power to two other meta-analytic approaches. The first is the traditional meta-
analysis strategy which uses a fixed effects model (FE) in which all of the effect sizes
across studies are assumed to be the same. We utilize an extension of the fixed effects model which corrects for population structure [FKV12]. A second alternate strategy is to simply apply the heterogeneity test (HE), which in our framework is only applied to loci first identified using random effects meta-analysis. The HE test follows the intuition that loci with high heterogeneity will harbor gene-by-environment interactions. For the purposes of the comparison we refer to Meta-GxE as the random effects (RE) model.

The level of gene-by-environment interactions can be simulated by changing both the environment-dependent and environment-independent effect simultaneously, when simulating the phenotype. Figure 8.1 (a) - (c) shows the power of the three approaches (RE, FE, HE) respectively when we vary the mean and variance of the effect size distribution we sampled from. In this simulation study, mean effect represents shared effect and variance of the effect size represents interaction effect. As expected, RE has high power in cases where the shared effect or the interaction effect is large. FE has high power when the shared effect is large and the HE test has high power when the interaction effect is large. Figure 8.1 (d) shows the heatmap which is colored with the color of highest powered approach. FE is most powerful at the top-left region, HE is most powerful at the bottom-right region, while RE is most powerful for a majority of the simulations. In the Supplementary information, we show through simulations that our methodology outperforms the alternative fixed effects and heterogeneity testing approaches when the effect is present in a subset of the studies, which is another possible interaction model we can assume. We also show in the Supplementary information that our approach is more powerful than the traditional uni- or multi-variate gene-by-environment association approach which assumes knowledge of the covariates involved in gene-by-environment interactions. For the traditional uni- or multi-variate approach, required knowledge includes kinds of variable (e.g. sex, age, gene knockouts) and encoding of the variables (e.g. binary values, continuous values). In
the Supplementary section, we also show the our proposed approach controls the false positive rate.

### 6.2.5 Application to 17 mouse HDL studies

We applied Meta-GxE to 17 mouse genetic studies conducted under various environmental conditions where each study measured HDL cholesterol. Table 6.1 summarizes each study. More details are provided in the Methods and Supplementary Information. We analyzed all 17 studies together and we also analyzed the 9 male and 8 female studies separately. Some significant associations are shared and some associations are specific to males and females.

The Manhattan plots in Figure 6.5 show the meta-GxE result when applied to the 17 studies, 9 male only studies and 8 female only studies. Table 6.2 summarizes 26 significant peaks ($P < 1.0 \times 10^{-6}$) showing the $P$-values obtained by applying meta-GxE to the male only studies (9 studies), the female only studies (8 studies) and the male + female studies (17 studies). For each significant locus, we computed m-values, interpreted as the posterior probability of having an effect on the phenotype and report the number of studies with an effect (E), the number of studies with ambiguous effect size (A) and the number of studies without an effect (N). We also report the number of individual studies where the locus was significant ($P < 1.0 \times 10^{-6}$). As seen in the table, many of the loci were not significant in any of the individual studies and would not have been discovered without combining the studies. We note that we use a more stringent genome wide threshold of $p < 1.0 \times 10^{-6}$ than was used in the original studies. The Genes in Region and Gene Refs columns contain the gene names near the locus previously known to affect HDL cholesterol level or closely related lipid level in the blood and associated literature citations.

Among the 26 loci that we identified by applying Meta-GxE, 24 loci are near the
genes (mostly genes are located within 1MB of the peak) known to affect HDL or closely related lipid level in the blood, while 3 loci are novel.

For example, we identified 3 loci (Chr10:21399819, Chr19:3319089, ChrX:151384614) female mice show a more significant effect on HDL phenotypes than male mice. We also identified 7 loci (Chr1:171199523, Chr8:46903188, Chr8:64150094, Chr8:84073148, Chr10:90146088, Chr11:69906552, Chr15:21194226) where male mice show a more significant effect on HDL than female mice.

Interestingly, we observed that in 3 loci (Chr10:21399819, Chr19:3319089, ChrX:151384614), female mice are more highly affected, while in 7 loci (Chr1:171199523, Chr8:46903188, Chr8:64150094, Chr8:84073148, Chr10:90146088, Chr11:69906552, Chr15:21194226) male mice are more highly affected. Among 26 loci, many show a significant heterogeneity in effect sizes between the 17 studies, which we interpret as gene-by-environment interactions.

One interesting example showing strong gene-by-environment interaction is a locus in Chr8:84073148. This locus is located near the gene Prkaca, which is known to affect the abnormal lipid levels in blood [NWW05]. Figure 6.6 shows the forest plot and PM-plot for this locus. If we look at the forest plot of the locus in Figure 6.6, we can easily see that there are two groups: 12 studies with an effect (red dots) and 5 studies with an ambiguous prediction of the existence of an effect (green dots). Interestingly, the log odds ratios of effect size for the 12 studies with an effect is about the same (around 0.2). The common characteristic in 4 of the 5 studies (HMDPxB-chow(F), HMDPxB-ath(F), BXH-apoe(F), CXB-ldlr(F)) is that they are female mice with high LDL levels in the blood. In addition, in all 4 cases, these high LDL levels are caused by mutant genes. Mice in HMDPxB-chow and HMDPxB-ath studies have transgenes for both Apoe Leiden [37] and for human Cholesterol Ester Transfer Protein (CETP), while mice in the BXH-apoe and CXB-ldlr studies carried knockouts
of the genes for Apoe and LDL receptor, respectively. This is a strong evidence that there is an interaction between sex × mutation-driven LDL levels through this locus (Chr8:84073148) when affecting HDL levels in mice.

Supplementary Figures S5 - S30 show the forest plots and PM-plots for each locus, which show information such as effect sizes, the direction of the effect, which study has an effect and which study does not have an effect for each of 17 studies at the given locus.

6.3 Discussion

In this paper, we present a new meta-analysis approach for discovering gene-by-environment interactions that can be applied to a large number of heterogeneous studies each conducted in different environments and with animals from different genetic backgrounds. We show the practical utility of the proposed method by applying it to 17 mouse HDL studies containing 4,965 mice, and we successfully identify many known loci involved in HDL. Consistent with the results of meta-analysis in human studies, our combined study finds many loci which were not discovered in any of the individual studies.

A point of emphasis is that in our study design, in each of the combined studies, all of the individuals in the study are subject to only a single environment. This is distinct from other approaches for discovery of gene-by-environment interactions using meta-analysis such as those described in [MLL11]. In these approaches, in each of the combined studies, the individuals in the study are subject to multiple environments and information on each individual’s environment is collected. Gene-by-environment statistics are then computed in each study and then combined in the meta-analysis. In our study design, we compute main effect sizes for each SNP and then look for variants where the effect sizes are different suggesting the presence of a gene-by-environment
interaction.

In our meta-analysis approach, we assume that we do not have any prior knowledge of the effect size in any particular study. However one might incorporate prior knowledge of the specific environmental effects. In some cases, one might know that some of the studies have similar effect sizes as compared to others. Or the prior knowledge might suggest that one specific study needs to be eliminated in the meta analysis. If we utilize such prior knowledge, we may be able to achieve even higher statistical power.

In this paper we have addressed how to perform meta-analysis when the studies have different genetic structures, building off the results of our previous study [FKV12]. While in this paper we combine 7 HMDP studies with 10 genetic crosses, the approach in principle can be used to combine any variety of study types. Recently, several strategies for mouse genome-wide association mapping have been proposed [FE12]. These include HMDP [BFO10b], collaborative cross [CAA04] and outbred-stock [YNB10] [FE12]. The approach presented here can be utilized to combine these different kinds of studies and is a roadmap for integrating the results of different strategies for mouse GWAS.

Although we have focused on explaining heterogeneity by gene-by-environment interaction, it is possible that the differences in effect sizes can be caused by gene-by-gene interactions on different genetic backgrounds, where the interacting variants differ in frequency in the different studies. While gene-by-gene interactions certainly contribute to locus heterogeneity, we predict that, in combining studies with similar genetic structures, locus heterogeneity more likely arises from gene-by-environment interactions. In any case, determining whether or not these heterogeneous loci are environment-driven or interaction-driven is an important and interesting direction for future study.
6.4 Methods and Materials

6.4.1 Standard study design for testing gene-by-environment interactions

In the model organism studies for which we can control the environment, the standard study design for testing gene-by-environment interactions is to combine multiple cohorts whose environments are known. The environmental value that we vary is typically a quantitative measure that we can model with a single random variable. Thus, the standard univariate linear model can be applied

\[ y = \mu + \alpha D + \beta X + \gamma X \cdot D + e \]

where \( y \) is \( n \times 1 \) vector of phenotype measurements from \( n \) individuals, \( \mu \) is the phenotype mean, \( \alpha \) is the main environmental effect mean, \( D \) is \( n \times 1 \) environmental status vector, \( \beta \) is the genetic effect, \( X \) is \( n \times 1 \) genotype vector, \( \gamma \) is GxE interactions effect, \( \cdot \) denotes the dot-product between two vectors, and \( e \) is the residual error, which follows normal distribution. In this model, vector \( D \) is a vector of indicators which describes the environmental status of each individual study. For example, Suppose the environmental condition of one study is wildtype and that of another is gene knockout. In this case, the environmental condition of wildtype is described as 0 and that of knockout is described as 1. In order to test if there are interactions, we test the null hypothesis \( \gamma = 0 \) versus the alternative hypothesis \( \gamma \neq 0 \). Another possible testing strategy is to test the interactions effect together with the genetic effect, that is, the null hypothesis \( \beta = 0 \) and \( \gamma = 0 \) versus the alternative hypothesis \( \beta \neq 0 \) or \( \gamma \neq 0 \). This strategy is powerful in detecting loci exhibiting both the genetic effects and the interactions effects.
6.4.2 Multivariate interactions model

For more complicated scenarios where the different environments can not be modeled with a single variable, a straightforward extension of the standard univariate interactions model is the multivariate model. Suppose that there are k different possible environments and the information on the environments of each individual are captured by a matrix $D$ which has k columns where each column corresponds to one environment. Then, the standard multivariate interactions model will be

$$y = \sum_{i=1}^{k} \alpha_i D_i + \beta X + \sum_{i=1}^{k} \gamma_i X \cdot D_i + e$$

(6.3)

$D_i$ is the $i^{th}$ column of the D matrix, $\alpha_i$ is the environment specific mean, $y$ denotes the phenotype measurements, $X$ denotes the genotypes, $\beta$ denotes the fixed genetic effect, $\gamma_i$ denotes GxE interactions effect of $i^{th}$ environmental variable and , and $e$ denotes the residual error. Then the testing will be between the null hypothesis $\gamma_1 = 0$ and ... and $\gamma_k = 0$ versus the alternative hypothesis $\gamma_1 \neq 0$ or ... or $\gamma_k \neq 0$. The test statistic will be

$$S_{Mult} = \sum_{i=1}^{k} Z_i^2$$

where $Z_i$ is the z-score corresponding to $\gamma_i$. $S_{Mult}$ follow $\chi^2_{(k)}$ under the null. Similarly to the univariate model, if we want to test the interactions effect together with genetic effect, we add the z-score corresponding to $\beta$ into the statistic, in which case the statistic will follow $\chi^2_{(k+1)}$.

6.4.3 Standard meta-analysis approach

Before we describe the relationship between gene-by-environment interactions and meta-analysis, we first describe the standard fixed effects and random effects meta-analysis in details.
6.4.3.1 Fixed effects model meta analysis

In standard meta-analysis, we have $N$ studies. In each of the $N$ studies, we estimate the effect size of interest. Suppose that we estimate the genetic effect in study $i$, 

$$y_i = \alpha_i + \delta_i X_i + e_i \quad (6.4)$$

We can obtain the estimates of $\delta_i$ and its variance $V_i$. In the fixed effects model meta-analysis, we assume that the underlying effect sizes are the same as $\delta$ ($\delta = \delta_1 = \ldots = \delta_N$). The best estimate of $\delta$ is the inverse-variance weighted effect size, 

$$\bar{\delta} = \frac{\sum W_i \delta_i}{\sum W_i}, \quad (6.5)$$

where $W_i = 1/V_i$ is the so-called inverse variance. Then we test the null hypothesis $\delta = 0$ versus the alternative hypothesis $\delta \neq 0$.

6.4.3.2 Testing heterogeneity

The phenomenon that the underlying effect sizes differ between studies is called heterogeneity. The presence of heterogeneity is tested using the Cochran’s Q test [DL86a, Coc09]. Cochran’s Q test is a non-parametric test for testing if $N$ studies have the same effect or not. Particularly it tests the null hypothesis $\delta_1 = \ldots = \delta_N$ versus the alternative hypothesis $\text{NOT}(\delta_1 = \ldots = \delta_N)$. Cochran’s Q statistic can be calculated as the weighted sum of squared differences between individual study effects and the pooled effect across studies.

$$Q = \sum_{i=1}^{N} W_i (\delta_i - \bar{\delta})^2 \quad (6.6)$$

Cochran’s Q statistic has a chi-square statistic with $N - 1$ degrees of freedom.
6.4.3.3 Random effects model meta analysis

Under the random effects model meta-analysis, we explicitly model heterogeneity by assuming a hierarchical model. We assume that the effect size of each study $\delta_i$ is a random variable sampled from a distribution with mean $\delta$ and variance $\tau^2$,

$$\delta_i \sim N(\delta, \tau^2)$$

Traditional formulations of a random effects meta-analysis method are known to be overly conservative [DL86a, IPE07a, IPE07b]. However, we recently developed a random effects model that addresses this issue [HE11b]. The method assumes that there is no heterogeneity under the null, a modification that is natural in the context of association studies because the effect size should be fixed to be zero under the null hypothesis. This random effects model tests the null hypothesis $\delta = 0$ and $\tau^2 = 0$ versus the alternative hypothesis $\delta \neq 0$ or $\tau^2 \neq 0$.

Similarly to the traditional random effects model [DL86a], we use the likelihood ratio framework considering each statistic as a single observation. Since we assume no heterogeneity under the null, $\mu = 0$ and $\tau^2 = 0$ under the null hypothesis. The likelihoods are then

$$L_0 = \prod_i \frac{1}{\sqrt{2\pi V_i}} \exp \left( -\frac{\delta_i^2}{2V_i} \right)$$

$$L_1 = \prod_i \frac{1}{\sqrt{2\pi (V_i + \tau^2)}} \exp \left( -\frac{(\delta_i - \mu)^2}{2(V_i + \tau^2)} \right).$$

The maximum likelihood estimates $\hat{\mu}$ and $\hat{\tau}^2$ can be found by an iterative procedure suggested by Hardy and Thompson [HT96]. Then the likelihood ratio test statistic can be built

$$S_{meta} = -2 \log(\lambda) = \sum \log \left( \frac{V_i}{V_i + \hat{\tau}^2} \right) + \sum \frac{\delta_i^2}{V_i} - \sum \frac{(\delta_i - \hat{\mu})^2}{V_i + \hat{\tau}^2}, \quad (6.7)$$

whose $P$-value is calculated using tabulated values [HE11b].
6.4.4 Relation between gene-by-environment interactions and meta-analysis

Here we explain more about the relationship between gene-by-environment interactions and meta-analysis based on the explanation in Results section. If we do not consider the interactions, it has been already known that the fixed effects model meta-analysis is approximately equivalent to the linear model of combined cohorts [LS09]. That is, the fixed effects model equation (8.3) gives approximately equivalent results to the combined linear model

\[ y = \sum_{i=1}^{k} \alpha_i A_i + \beta X + e \]  

where \( X \) is the combined genotype vector from all cohorts, \( A \) is a matrix that includes indicator columns which identify which individual is in each cohort, \( A_i \) is the \( i^{th} \) column of matrix \( A \), and \( \alpha_i \) is the cohort specific mean. The two methods are approximately equivalent because they both test the fixed mean effect (\( \beta \) in equation (6.8) and \( \delta \) in equation (8.3)). The subtle difference between the two models is that in equation (6.8), we assume the error \( e \) follows a single normal distribution (e.g. \( N(0, \sigma^2) \)), whereas in equation (8.3), the variance of the distributions may differ between studies (e.g. \( e_j \sim N(0, \sigma^2_j) \) for each \( j \)). In other words, under the constant error variance assumption (\( \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_N^2 \)), the two models become equivalent and \( \beta \) in equation (6.8) equals \( \delta \) in equation (8.3),

\[ \beta = \delta. \]

Similarly, by considering interactions, we extend this argument to show the relationship between gene-by-environment interactions and meta-analysis. We consider the relationship between equation (6.3) and equation (6.4). For simplicity of the notation, we consider the case where the matrix D is defined in such a way that each individual is only in one environment such that the D matrix is equivalent to the matrix A described above. If we assume the constant error variance assumption, we establish
the following relationship,

$$\delta_i = \beta + \gamma_i$$

where the left hand side is the coefficient of the genotype $X_i$ of study $i$ from the meta-analysis equation (6.4) and the right hand side is the same coefficient of $X_i$ (the study $i$’s part within the combined genotype matrix $X$) from the equation (6.3).

Suppose that there are no interactions (null hypothesis of interaction testing). Then, $\gamma_i = 0$ for each study $i$. Thus, the effect size of meta-analysis $\delta_i$ is equivalent to $\beta$, the genetic effects that are invariant across studies. Therefore, $\tau^2 = 0$ (null hypothesis of heterogeneity testing). On the other hand, suppose that $\tau^2 = 0$ (null hypothesis of heterogeneity testing). Naturally, $\gamma_i = 0$ for all studies (null hypothesis of interaction testing). This shows that the null hypothesis of the interactions test in the model (6.3) and the null hypothesis of the heterogeneity test in meta-analysis are equivalent. As a result, we can utilize meta-analytic heterogeneity testing to detect interactions.

Using reasoning, it is straightforward to show that we can utilize the random effects model meta-analysis method to detect the mean effect and the interaction effect at the same time, which can be powerful for identifying loci bearing both kinds of effects.

### 6.4.5 Controlling for population structure within studies

Model organism such as the mouse are well-known to exhibit population structure or cryptic relatedness [DRB01, VP05], where genetic similarities between individuals both inhibit the ability to find true associations and cause the appearance of a large number of false or spurious associations. Mixed effects models are often used in order to correct this problem [Lan02, YPB06, KZW08, KSS10, LQK13, LLK13]. Methods employing a mixed effects correction account for the genetic similarity between individuals with the introduction of a random variable into the traditional linear model.
\[ y_i = \mu + \delta_i X + u_i + \epsilon \] (6.9)

In the model in equation (8.8), the random variable \( u_i \) represents the vector of genetic contributions to the phenotype for individuals in population \( i \). This random variable is assumed to follow a normal distribution with \( u_i \sim N(0, \sigma^2 g K_i) \), where \( K_i \) is the \( n_i \times n_i \) kinship coefficient matrix for population \( i \). With this assumption, the total variance of \( y_i \) is given by \( \Sigma_i = \sigma^2 g K_i + \sigma^2 I \). A z-score statistic is derived for the test \( \delta_i = 0 \) by noting the distribution of the estimate of \( \hat{\delta}_i \). In order to avoid complicated notation, we introduce a more basic matrix form of the model in equation (8.8), shown in equation (8.9).

\[ y_i = S_i \Gamma + u_i + \epsilon \] (6.10)

In equation (8.9), \( S_i \) is a \( n_i \times 2 \) matrix with the first column being a vector of 1s representing the global mean and the second vector is the vector and \( \Gamma \) is a \( 2 \times 1 \) coefficient vector containing the mean \( \alpha_i \) and genotype effect \( (\delta_i) \). We note that this form also easily extends to models with multiple covariates. The maximum likelihood estimate for \( \Gamma \) in population \( i \) is given by \( \hat{\Gamma}_i = (S_i'\Sigma_i^{-1}S_i)^{-1}S_i'\Sigma_i^{-1}y_i \) which follows a normal distribution with a mean equal to the true \( \Gamma \) and variance \( (S_i'\Sigma_i^{-1}S_i)^{-1} \). The estimates of the effect size \( \delta_i \) and standard error of the \( \delta_i \) \( (SE(\delta_i)) \) are then given in equation (8.10) and equation (8.11), where \( R = [0 \ 1] \) is a vector used to select the appropriate entry in the vector \( \hat{\Gamma}_i \).

\[
\delta_i = R(S_i'\Sigma_i^{-1}S_i)^{-1}S_i'\Sigma_i^{-1}y_i \quad (6.11)
\]
\[
SE(\delta_i) = [R(S_i'\Sigma_i^{-1}S_i)^{-1}R']^{1/2} \quad (6.12)
\]
6.4.6 Meta-analysis of studies with population structure

When we test gene-by-environment interactions with meta analysis approaches, one important step is correcting for population structure. This can be achieved by correcting for population structure within each study first as described above. For example, consider the random effects model meta-analysis method that we primarily focus on. We employ population structure control, using (8.10) and (8.11). Then the likelihood ratio test statistic will be

\[ S_{\text{Pop}} = -2 \log(\lambda) = \sum \log \left( \frac{V_i}{V_i + \hat{\tau}^2} \right) + \sum \frac{\delta_i^2}{V_i} - \sum \frac{(\delta_i - \hat{\mu})^2}{V_i + \hat{\tau}^2}, \]  

(6.13)

where \( \delta_i = R(S_i^\prime \Sigma_i^{-1} S_i)^{-1} S_i^\prime \Sigma_i^{-1} y_i \) and \( V_i = [R(S_i^\prime \Sigma_i^{-1} S_i)^{-1} R'] \).

6.4.7 Identifying studies with an effect

After identifying loci exhibiting interaction effects, we employ the meta-analysis interpretation framework that we recently developed. The \( m \)-value [HE12] is the posterior probability that the effect exists in each study. Suppose we have \( n \) number of studies we want to combine. Let \( E = [\delta_1, \delta_2, ..., \delta_n] \) be the vector of estimated effect sizes and \( V = [V_1, V_2, ..., V_n] \) be the vector of estimated variance of \( n \) effect sizes. We assume that the effect size \( \delta_i \) follows the normal distribution.

\[ P(\delta_i | \text{no effect}) = N(\delta_i; 0, V_i) \]  

(6.14)

\[ P(\delta_i | \text{effect}) = N(\delta_i; \mu, V_i) \]  

(6.15)

We assume that the prior for the effect size is

\[ \mu \sim N(0, \sigma^2) \]  

(6.16)

A possible choice for \( \sigma \) in GWAS is 0.2 for small effect and 0.4 for large effect [SB09]. We also denote \( C_i \) be a random variable whose value is 1 if a study \( i \) has an effect and
0 otherwise. We also denote $C$ as a vector of $C_i$ for $n$ studies. Since $C$ has $n$ binary values, $C$ can be $2^n$ possible configurations. Let $U = [c_1, ..., c_{2^n}]$ be a vector containing all the possible these configurations. We define $m$-value $m_i$ as the probability $P(C_i = 1|E)$, which is the probability of study $i$ having an effect given the estimated effect sizes. We can compute this probability using the Bayes’ theorem in the following way.

$$m_i = P(C_i = 1|E) = \frac{\sum_{c \in U_i} P(E|C = c)P(C = c)}{\sum_{c \in U} P(E|C = c)P(C = c)} \quad (6.17)$$

where $U_i$ is a subset of $U$ whose elements’ $i^{th}$ value is 1. Now we need to compute $P(E|C = c)$ and $P(C = c)$. $P(C = c)$ can be computed as

$$P(C = c) = \frac{\text{B}(|c| + \alpha, n - |c| + \beta)}{\text{B}(\alpha, \beta)} \quad (6.18)$$

where $|c|$ denotes the number of 1’s in $c$ and $\text{B}$ denotes the beta function and we set $\alpha$ and $\beta$ as 1 [HE12]. The probability $E$ given configuration $c$, $P(E|C = c)$, can be computed as

$$P(E|C = c) = \int_{-\infty}^{\infty} \prod_{i \in c_0} N(\delta_i; 0, V_i) \prod_{i \in c_1} N(\delta_i; \mu, V_i) p(\mu) \, d\mu \quad (6.19)$$

$$= \bar{C} N(\bar{\delta}; 0, \bar{V}) \prod_{i \in c_0} N(\delta_i; 0, V_i) \quad (6.20)$$

$$\bar{\delta} = \frac{\sum_i W_i \delta_i}{\sum_i W_i} \quad \text{and} \quad \bar{V} = \frac{1}{\sum_i W_i} \quad (6.21)$$

where $c_0$ is the indices of 0 in $c$ and $c_1$ is the indices of 1 in $c$, $N(\delta; a, b)$ denotes the probability density function of the normal distribution with mean $a$ and variance $b$. $W_i = V_i^{-1}$ is the inverse variance or precision and $\bar{C}$ is a scaling factor.

$$\bar{C} = \frac{1}{(\sqrt{2\pi})^{N-1} \sqrt{\prod_i W_i^{1/2}}} \exp \left\{ -\frac{1}{2} \left( \sum_i W_i \delta_i^2 - \frac{(\sum_i W_i \delta_i)^2}{\sum_i W_i} \right) \right\} \quad (6.22)$$

All summations appeared for computing $\bar{\delta}, \bar{V}$ and $\bar{C}$ are with respect to $j \in t_1$. 

130
The *m-values* have the following interpretations: small *m-values* (0.1) represent a study that is predicted to not have an effect, large *m-values* (0.9) represent a study that is predicted to have an effect, otherwise it is ambiguous to make a prediction. It was previously reported that *m-values* can accurately distinguish studies having an effect from the studies not having an effect [HE12]. For interpreting and understanding the result of the meta-analysis, it is informative to look at the *P*-value and m-value at the same time. We propose to apply the PM-plot framework [HE12], which plots the *P*-values and m-values of each study together in two dimensions. Figure 6.1 (b) shows one example of a PM-plot. In this example, studies with an *m-value* less than 0.1 are interpreted as studies not having an effect while studies with an *m-value* greater than 0.9 are interpreted as studies having an effect. For studies with an *m-value* between 0.1 and 0.9, we cannot make a decision. One reason that studies are ambiguous (0.9 ≤ *m-value* ≤ 0.1) is that they are underpowered due to small sample size. If the sample size increases, the study can be drawn to either the left or the right side.
6.5 Tables
6.6 Figures

Figure 6.1: Application of Meta-GxE to Apoa2 locus. The forest plot (A) shows heterogeneity in the effect sizes across different studies. The PM-plot (B) predicts that 7 studies have an effect at this locus, even though only 1 study (HMDP-chow(M)) is genome-wide significant with $P$-value.
Figure 6.2: The prevalence of heterogeneity in effect size of significant loci. Each dot represents association between SNPs and HDL phenotype from applying random effects based meta-analysis approach. Dots with larger $I^2$ value represents the existence of more heterogeneity at the locus between studies. The distribution of the heterogeneity statistic for significant SNPs (red dots) in the meta analysis is skewed toward higher heterogeneity while the non-significant SNPs are much less skewed.
Figure 6.3: Power of mouse meta-analysis to identify gene-by-environment interactions in 4,965 animals from 17 studies under varying mean effect sizes and the per study variance of the effect size which corresponds to gene-by-environment effects.
Figure 6.4: Power of (a) random-effect, (b) fixed-effect meta-analysis and (c) heterogeneity meta-analysis methods as a function of the effect size and the strength of the interaction effect (heterogeneity). (d) shows a comparison of the three methods with the color corresponding to the method with the highest power.
Figure 6.5: Manhattan plots showing the results of Meta-GxE applied to (a) 17 HDL studies, (b) 9 HDL studies consisting only of male animals and (c) 8 studies consisting only of female animals.
Figure 6.6: Peak SNP in chromosome 8 shows interesting gene-by-environment interactions between sex × mutation-driven LDL levels.
6.7 Supplementary Legends
Seventeen HDL studies are combined in the meta analysis. U in the Ref column represents a
data set that is not yet published. Mice for the HMDPxB panel were created by breeding fe-
males of the various HMDP inbred strains to males carrying transgenes for both Apoe Leiden
and for human Cholesterol Ester Transfer Protein (CETP) on a C57BL/6 genetic background.
The Leiden/CETP transgenes [MHK93, JAW92] cause high total cholesterol level and high
LDL cholesterol level in the circulation, along with reduced HDL cholesterol. BxD db/db
denotes a population of F2 mice from a cross between C57BL/6 DBA/2 with homozygous de-
ficiency in leptin receptor (db/db), which results in obese mice. BxH Apoe −/− denotes a
population of F2 mice from a cross between C57BL/6 and C3H also carrying a deficiency in
apolipoprotein E. CxC LDLR −/− denotes a population of F2 mice from a cross between
C57BL/6 and BALB/cBy also carrying a deficiency in LDL receptor, which results in high
LDL cholesterol level in the circulation BXH wildtype denotes a population of F2 mice from
a cross between C57BL/6 and C3H.
Table 6.2: 26 significant loci identified by applying Meta-GxE analysis

<table>
<thead>
<tr>
<th>SNP Location</th>
<th>Meta GxE P (Male)</th>
<th>Meta GxE P (Female)</th>
<th>Meta GxE P (Male/Female)</th>
<th># of Studies w/ Significant P</th>
<th>HE Meta P (Male/Female)</th>
<th># Studies E/AN</th>
<th>Genes in Region</th>
<th>Gene Ref</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chr1:6475282 (rs31078051)</td>
<td>1.12 x 10^{-4}</td>
<td>3.05 x 10^{-5}</td>
<td>9.82 x 10^{-7}</td>
<td>0</td>
<td>3.82 x 10^{-5}</td>
<td>3/4/1</td>
<td>Pikfyve</td>
<td>[THS13]</td>
</tr>
<tr>
<td>Chr1:102771282 (rs32803839)</td>
<td>6.69 x 10^{-4}</td>
<td>2.66 x 10^{-5}</td>
<td>7.67 x 10^{-7}</td>
<td>0</td>
<td>7.55 x 10^{-2}</td>
<td>6/0/1</td>
<td>Bel2</td>
<td>[HBM12]</td>
</tr>
<tr>
<td>Chr1:171199523 (rs32075748)</td>
<td>3.45 x 10^{-16}</td>
<td>1.39 x 10^{-7}</td>
<td>4.41 x 10^{-22}</td>
<td>1</td>
<td>5.80 x 10^{-5}</td>
<td>8/6/3</td>
<td>Apoa2</td>
<td>[PAO97]</td>
</tr>
<tr>
<td>Chr2:142271733 (rs72328893)</td>
<td>1.55 x 10^{-4}</td>
<td>7.10 x 10^{-5}</td>
<td>5.74 x 10^{-7}</td>
<td>0</td>
<td>2.25 x 10^{-2}</td>
<td>3/4/0</td>
<td>Agps</td>
<td>[LCD11]</td>
</tr>
<tr>
<td>Chr3:29944129 (rs28962784)</td>
<td>5.97 x 10^{-4}</td>
<td>2.54 x 10^{-5}</td>
<td>2.36 x 10^{-7}</td>
<td>0</td>
<td>3.9 /</td>
<td>3/8/6</td>
<td>Prkci</td>
<td>[FSY07]</td>
</tr>
<tr>
<td>Chr3:76666632 (rs14870708)</td>
<td>5.89 x 10^{-3}</td>
<td>2.90 x 10^{-5}</td>
<td>5.30 x 10^{-7}</td>
<td>0</td>
<td>6.56 x 10^{-5}</td>
<td>1/1/10</td>
<td>Novel</td>
<td></td>
</tr>
<tr>
<td>Chr3:107430396 (rs30013147)</td>
<td>9.59 x 10^{-6}</td>
<td>3.48 x 10^{-5}</td>
<td>1.56 x 10^{-7}</td>
<td>0</td>
<td>7.74 x 10^{-7}</td>
<td>7/1/10</td>
<td>Csf1</td>
<td>[QFM97]</td>
</tr>
<tr>
<td>Chr3:143669492 (rs30267671)</td>
<td>1.82 x 10^{-3}</td>
<td>3.97 x 10^{-5}</td>
<td>3.34 x 10^{-7}</td>
<td>0</td>
<td>8.60 x 10^{-2}</td>
<td>7/1/00</td>
<td>H2s1t1</td>
<td>[SWC10]</td>
</tr>
<tr>
<td>Chr4:131925523 (rs2595861)</td>
<td>1.72 x 10^{-4}</td>
<td>2.84 x 10^{-5}</td>
<td>1.41 x 10^{-7}</td>
<td>0</td>
<td>8.65 x 10^{-4}</td>
<td>6/0/3</td>
<td>Fabp3</td>
<td>[MBB10]</td>
</tr>
<tr>
<td>Chr5:119034507 (rs33131194)</td>
<td>1.23 x 10^{-4}</td>
<td>2.59 x 10^{-5}</td>
<td>9.00 x 10^{-7}</td>
<td>0</td>
<td>3.94 x 10^{-1}</td>
<td>9/0/0</td>
<td>Nos1</td>
<td>[LGP12, NT508]</td>
</tr>
<tr>
<td>Chr6:4663688 (rs33272588)</td>
<td>1.47 x 10^{-7}</td>
<td>6.52 x 10^{-3}</td>
<td>1.66 x 10^{-6}</td>
<td>1</td>
<td>1.60 x 10^{-6}</td>
<td>2/1/4</td>
<td>Acd1</td>
<td>[LEP09]</td>
</tr>
<tr>
<td>Chr6:64150094 (rs31750594)</td>
<td>1.96 x 10^{-7}</td>
<td>1.89 x 10^{-5}</td>
<td>1.33 x 10^{-10}</td>
<td>0</td>
<td>8.34 x 10^{-1}</td>
<td>11/0/0</td>
<td>Cpe</td>
<td>[NIN94]</td>
</tr>
<tr>
<td>Chr8:34073148 (rs34358859)</td>
<td>1.95 x 10^{-5}</td>
<td>4.53 x 10^{-4}</td>
<td>5.49 x 10^{-8}</td>
<td>0</td>
<td>3.85 x 10^{-1}</td>
<td>2/5/0</td>
<td>Prkca</td>
<td>[NNW05]</td>
</tr>
<tr>
<td>Chr9:107976837 (rs4333010)</td>
<td>1.22 x 10^{-4}</td>
<td>2.90 x 10^{-5}</td>
<td>4.09 x 10^{-7}</td>
<td>0</td>
<td>1.98 x 10^{-8}</td>
<td>2/1/4</td>
<td>Pik3cb</td>
<td>[CMO08]</td>
</tr>
<tr>
<td>Chr10:12190919 (rs39193411)</td>
<td>9.07 x 10^{-4}</td>
<td>3.61 x 10^{-5}</td>
<td>3.36 x 10^{-9}</td>
<td>0</td>
<td>1.18 x 10^{-7}</td>
<td>3/2/2</td>
<td>Hqgr1</td>
<td>[GPW07]</td>
</tr>
<tr>
<td>Chr10:90146088 (rs29705052)</td>
<td>1.93 x 10^{-7}</td>
<td>0.786</td>
<td>1.06 x 10^{-5}</td>
<td>1</td>
<td>8.94 x 10^{-4}</td>
<td>2/1/4</td>
<td>Nrlh4</td>
<td>[KAR07]</td>
</tr>
<tr>
<td>Chr11:69906552 (rs2497701)</td>
<td>5.77 x 10^{-8}</td>
<td>1.35 x 10^{-5}</td>
<td>3.17 x 10^{-12}</td>
<td>0</td>
<td>2.37 x 10^{-9}</td>
<td>6/0/2</td>
<td>Phe3</td>
<td>[WIZ04]</td>
</tr>
<tr>
<td>Chr11:140813173 (rs29416888)</td>
<td>0.0001</td>
<td>7.83 x 10^{-1}</td>
<td>1.71 x 10^{-7}</td>
<td>0</td>
<td>5.28 x 10^{-5}</td>
<td>3/3/1</td>
<td>Acor1</td>
<td>[FPC06]</td>
</tr>
<tr>
<td>Chr14:33632464 (rs31601259)</td>
<td>1.96 x 10^{-4}</td>
<td>1.65 x 10^{-5}</td>
<td>8.90 x 10^{-7}</td>
<td>0</td>
<td>2.02 x 10^{-5}</td>
<td>3/1/0</td>
<td>Pyr1</td>
<td>[SBS03]</td>
</tr>
<tr>
<td>Chr15:21194226 (rs31670969)</td>
<td>1.96 x 10^{-8}</td>
<td>1.29 x 10^{-5}</td>
<td>9.87 x 10^{-10}</td>
<td>1</td>
<td>5.65 x 10^{-4}</td>
<td>3/2/12</td>
<td>Novel</td>
<td></td>
</tr>
<tr>
<td>Chr15:59860195 (rs37182172)</td>
<td>5.64 x 10^{-6}</td>
<td>1.45 x 10^{-5}</td>
<td>5.31 x 10^{-8}</td>
<td>0</td>
<td>9.92 x 10^{-5}</td>
<td>3/0/2</td>
<td>Trib1, Sdle</td>
<td>[ERS11, FRR13]</td>
</tr>
<tr>
<td>Chr17:65537012 (rs32593131)</td>
<td>1.09 x 10^{-5}</td>
<td>1.00 x 10^{-7}</td>
<td>1.30 x 10^{-7}</td>
<td>0</td>
<td>5.35 x 10^{-5}</td>
<td>5/0/2</td>
<td>Gnnm</td>
<td>[ELC07]</td>
</tr>
<tr>
<td>Chr18:82240606 (rs33483466)</td>
<td>2.05 x 10^{-4}</td>
<td>2.23 x 10^{-5}</td>
<td>1.05 x 10^{-7}</td>
<td>0</td>
<td>9.05 x 10^{-7}</td>
<td>5/2/0</td>
<td>Mbp</td>
<td>[BCG08]</td>
</tr>
<tr>
<td>Chr19:3310899 (rs10043232)</td>
<td>5.58 x 10^{-2}</td>
<td>8.53 x 10^{-5}</td>
<td>4.56 x 10^{-7}</td>
<td>0</td>
<td>1.08 x 10^{-3}</td>
<td>3/1/0</td>
<td>Lrp5</td>
<td>[FAK03]</td>
</tr>
<tr>
<td>ChrX:151836416 (rs31208008)</td>
<td>2.59 x 10^{-4}</td>
<td>4.72 x 10^{-5}</td>
<td>8.09 x 10^{-9}</td>
<td>0</td>
<td>1.22 x 10^{-1}</td>
<td>5/5/0</td>
<td>Htr2c</td>
<td>[KGT08]</td>
</tr>
</tbody>
</table>

Twenty-six significant loci identified by applying Meta-GxE analysis of both random effects meta-analysis and heterogeneity testing to 17 mouse HGL studies under different environments containing 4,965 total animals. # studies E denotes the number of studies with an effect on HGL phenotype. # studies N denotes the number of studies with no effect on HGL phenotype. # studies A denotes the number of studies with an ambiguous effect size. Genes in region denotes candidate genes for each locus based on close proximity to the peak SNP and previously suggested role in lipid or apolipoprotein metabolism: Pikfyve (phosphoinositide kinase), Bel2 (B cell leukemia/lymphoma 2), Apoa2 (apolipoprotein A-II), Agps (alkylglycerone phosphate synthase), Jag1 (jagged 1), Prkci (protein kinase C), Prkci (colony stimulating factor 1 (macrophage)), Hs2s1 (heparan sulfate 2-O-sulfotransferase 1), Fabp3 (fatty acid binding protein 3), Nos1 (nitric oxide synthase 1), Acs1 (acyl-CoA synthetase long-chain family member 1), Cpe (carboxypeptidase E), Prkaca (protein kinase, cAMP dependent, catalytic, alpha), Acox1 (peroxisomal acyl-coenzyme A oxidase 1), Ppyr1 (pancreatic polypeptide receptor 1), Trib1 (tribbles homolog 1), Ssq1 (squalene epoxidase), Gmnt (glucose N-methyltransferase), Mbp (myelin basic protein), Lrp5 (low density lipoprotein receptor-related protein 5), Htr2c (5-hydroxytryptamine (serotonin) receptor 2C)
Table 6.3: False Positive Rate of RE versus Traditional Wald Test Based Approaches at Thresholds of Increasing Significance

<table>
<thead>
<tr>
<th># of Knockout Studies</th>
<th>RE</th>
<th>Wald test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Threshold α</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.0 × 10^{-2}</td>
<td>5.00 × 10^{-2}, 4.98 × 10^{-2}, 4.99 × 10^{-2}, 4.99 × 10^{-2}</td>
<td>5.07 × 10^{-2}, 5.05 × 10^{-2}, 5.06 × 10^{-2}, 5.05 × 10^{-2}</td>
</tr>
<tr>
<td>5.0 × 10^{-3}</td>
<td>4.96 × 10^{-3}, 5.00 × 10^{-3}, 4.98 × 10^{-3}, 4.94 × 10^{-3}</td>
<td>5.16 × 10^{-3}, 5.18 × 10^{-3}, 5.17 × 10^{-3}, 5.12 × 10^{-3}</td>
</tr>
<tr>
<td>5.0 × 10^{-4}</td>
<td>5.03 × 10^{-4}, 4.99 × 10^{-4}, 5.04 × 10^{-4}, 4.99 × 10^{-4}</td>
<td>5.27 × 10^{-4}, 5.26 × 10^{-4}, 5.20 × 10^{-4}, 5.19 × 10^{-4}</td>
</tr>
<tr>
<td>5.0 × 10^{-5}</td>
<td>4.93 × 10^{-5}, 4.93 × 10^{-5}, 4.89 × 10^{-5}, 5.14 × 10^{-5}</td>
<td>5.37 × 10^{-5}, 5.36 × 10^{-5}, 5.21 × 10^{-5}, 5.31 × 10^{-5}</td>
</tr>
</tbody>
</table>

False positive rate is measured as we increase the number of knockout studies having interaction effects with 100 million null panel simulations.

Table 6.4: False Positive Rate of HE versus Traditional Wald Test Based Approaches at Thresholds of Increasing Significance

<table>
<thead>
<tr>
<th># of Knockout Studies</th>
<th>HE</th>
<th>Wald test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Threshold α</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.0 × 10^{-2}</td>
<td>5.01 × 10^{-2}, 4.99 × 10^{-2}, 4.98 × 10^{-2}, 4.98 × 10^{-2}</td>
<td>5.06 × 10^{-2}, 5.03 × 10^{-2}, 5.05 × 10^{-2}, 5.03 × 10^{-2}</td>
</tr>
<tr>
<td>5.0 × 10^{-3}</td>
<td>4.98 × 10^{-3}, 5.01 × 10^{-3}, 4.97 × 10^{-3}, 4.97 × 10^{-3}</td>
<td>5.17 × 10^{-3}, 5.15 × 10^{-3}, 5.11 × 10^{-3}, 5.12 × 10^{-3}</td>
</tr>
<tr>
<td>5.0 × 10^{-4}</td>
<td>4.99 × 10^{-4}, 4.98 × 10^{-4}, 4.98 × 10^{-4}, 5.15 × 10^{-4}</td>
<td>5.21 × 10^{-4}, 5.24 × 10^{-4}, 5.30 × 10^{-4}, 5.19 × 10^{-4}</td>
</tr>
<tr>
<td>5.0 × 10^{-5}</td>
<td>4.91 × 10^{-5}, 4.95 × 10^{-5}, 4.86 × 10^{-5}, 4.99 × 10^{-5}</td>
<td>5.35 × 10^{-5}, 5.32 × 10^{-5}, 5.27 × 10^{-5}, 5.15 × 10^{-5}</td>
</tr>
</tbody>
</table>

False positive rate is measured as we increase the number of knockout studies having interaction effects with 100 million null panel simulations.
Figure 6.7: Power comparison between random-effect, fixed-effect meta-analysis and heterogeneity testing
Figure 6.8: The association result of the fixed effects meta analysis.
Figure 6.9: Power comparison between random-effect meta-analysis and traditional wald test based approach
Figure 6.10: Power comparison between heterogeneity testing approach and traditional wald test based approach
Figure 6.11: Forest plot and PM-plot for Chr1:64752822 locus
**RE Summary**

<table>
<thead>
<tr>
<th>P-value</th>
<th>Study Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3187</td>
<td>1.HMDPxB–chow(M)</td>
</tr>
<tr>
<td>0.3466</td>
<td>2.HMDPxB–chow(F)</td>
</tr>
<tr>
<td>0.8350</td>
<td>3.HMDPxB–ath(M)</td>
</tr>
<tr>
<td>0.00346</td>
<td>4.HMDPxB–ath(F)</td>
</tr>
<tr>
<td>0.00016</td>
<td>5.HMDP–chow(M)</td>
</tr>
<tr>
<td>0.0234</td>
<td>6.HMDP–fat(M)</td>
</tr>
<tr>
<td>0.0020</td>
<td>7.HMDP–fat(F)</td>
</tr>
<tr>
<td>0.1256</td>
<td>8.BXD–db-12(M)</td>
</tr>
<tr>
<td>0.2428</td>
<td>9.BXD–db-12(F)</td>
</tr>
<tr>
<td>0.9941</td>
<td>10.BXD–db-5(M)</td>
</tr>
<tr>
<td>0.6261</td>
<td>11.BXD–db-5(F)</td>
</tr>
<tr>
<td>0.0137</td>
<td>12.BXH–apoe(M)</td>
</tr>
<tr>
<td>0.0314</td>
<td>13.BXH–apoe(F)</td>
</tr>
<tr>
<td>0.2863</td>
<td>14.BXH–wt(M)</td>
</tr>
<tr>
<td>0.1820</td>
<td>15.BXH–wt(F)</td>
</tr>
<tr>
<td>0.9166</td>
<td>16.CXB–ldlr(M)</td>
</tr>
<tr>
<td>0.4062</td>
<td>17.CXB–ldlr(F)</td>
</tr>
</tbody>
</table>

**Figure 6.12: Forest plot and PM-plot for Chr1:107271282 locus**
Figure 6.13: Forest plot and PM-plot for Chr1:171199523 locus
Figure 6.14: Forest plot and PM-plot for Chr2:77837584 locus
Figure 6.15: Forest plot and PM-plot for Chr2:134421733 locus
Figure 6.16: Forest plot and PM-plot for Chr3:32944259 locus
Figure 6.17: Forest plot and PM-plot for Chr3:76066632 locus
Figure 6.18: Forest plot and PM-plot for Chr3:107430396 locus
Figure 6.19: Forest plot and PM-plot for Chr3:143466942 locus
Figure 6.20: Forest plot and PM-plot for Chr4:131925523 locus
Figure 6.21: Forest plot and PM-plot for Chr5:119034507 locus
Figure 6.22: Forest plot and PM-plot for Chr8:46903188 locus
Figure 6.23: Forest plot and PM-plot for Chr8:64150094 locus
Figure 6.24: Forest plot and PM-plot for Chr8:84073148 locus
Figure 6.25: Forest plot and PM-plot for Chr9:101972687 locus
**Figure 6.26: Forest plot and PM-plot for Chr10:21399819 locus**

### Chr10:21399819 [rs29363941] (Meta $P = 3.36 \times 10^{-5}$)

**Gene:** Ifngr1

<table>
<thead>
<tr>
<th>Study Name</th>
<th>$P$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.HMDPxB−chow(M)</td>
<td>0.0036</td>
</tr>
<tr>
<td>2.HMDPxB−chow(F)</td>
<td>$5.10 \times 10^{-5}$</td>
</tr>
<tr>
<td>3.HMDPxB−ath(M)</td>
<td>0.8923</td>
</tr>
<tr>
<td>4.HMDPxB−ath(F)</td>
<td>$7.96 \times 10^{-6}$</td>
</tr>
<tr>
<td>5.HMDP−chow(M)</td>
<td>0.2037</td>
</tr>
<tr>
<td>6.HMDP−fat(M)</td>
<td>0.0434</td>
</tr>
<tr>
<td>7.HMDP−fat(F)</td>
<td>0.087</td>
</tr>
<tr>
<td>8.BXD−db−12(M)</td>
<td>0.2955</td>
</tr>
<tr>
<td>9.BXD−db−12(F)</td>
<td>0.5558</td>
</tr>
<tr>
<td>10.BXD−db−5(M)</td>
<td>0.5669</td>
</tr>
<tr>
<td>11.BXD−db−5(F)</td>
<td>0.6659</td>
</tr>
<tr>
<td>12.BXH−apoe(M)</td>
<td>0.1185</td>
</tr>
<tr>
<td>13.BXH−apoe(F)</td>
<td>0.3642</td>
</tr>
<tr>
<td>14.BXH−wt(M)</td>
<td>0.7105</td>
</tr>
<tr>
<td>15.BXH−wt(F)</td>
<td>0.1983</td>
</tr>
<tr>
<td>16.CXB−ldlr(M)</td>
<td>0.2788</td>
</tr>
<tr>
<td>17.CXB−ldlr(F)</td>
<td>0.5708</td>
</tr>
</tbody>
</table>

**RE Summary**

- **Log odds ratio**
  - $-0.6$ to $0.6$

- **PM-Plot**
  - $-10 \log_{10}(p)$
  - $m$-value

- **Legend**
  - ● Study has an effect ($m > .9$)
  - ● Study does not have an effect ($m < .1$)
  - ● Study's effect is uncertain ($0.1 < m < 0.9$)
Log odds ratio
-0.6 -0.2 0.2 0.6
1.HMDPxB −chow(M)
2.HMDPxB −chow(F)
3.HMDPxB −ath(M)
4.HMDPxB −ath(F)
5.HMDP −chow(M)
6.HMDP −fat(M)
7.HMDP −fat(F)
8.BXD −db−12(M)
9.BXD −db−12(F)
10.BXD −db−5(M)
11.BXD −db−5(F)
12.BXH−apoE(M)
13.BXH−apoE(F)
14.BXH−wt(M)
15.BXH−wt(F)
16.CXB−ldlr(M)
17.CXB−ldlr(F)

RE Summary

Figure 6.27: Forest plot and PM-plot for Chr10:90146088 locus
Figure 6.28: Forest plot and PM-plot for Chr11:69906552 locus
Figure 6.29: Forest plot and PM-plot for Chr11:114083173 locus
Figure 6.30: Forest plot and PM-plot for Chr14:33632464 locus
Figure 6.31: Forest plot and PM-plot for Chr15:21194226 locus
Figure 6.32: Forest plot and PM-plot for Chr15:59860191 locus
Figure 6.33: Forest plot and PM-plot for Chr17:46530712 locus
Figure 6.34: Forest plot and PM-plot for Chr18:82240606 locus
### RE Summary

<table>
<thead>
<tr>
<th>Study Name</th>
<th>P-value</th>
<th>Log odds ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. HMDPxB−chow(M)</td>
<td>0.0211</td>
<td>−1.5</td>
</tr>
<tr>
<td>2. HMDPxB−chow(F)</td>
<td>8.43 x 10^{-7}</td>
<td>−1.0</td>
</tr>
<tr>
<td>3. HMDPxB−ath(M)</td>
<td>0.8327</td>
<td>−0.5</td>
</tr>
<tr>
<td>4. HMDPxB−ath(F)</td>
<td>0.0024</td>
<td>0.0</td>
</tr>
<tr>
<td>5. HMDP−chow(M)</td>
<td>0.3186</td>
<td>0.5</td>
</tr>
<tr>
<td>6. HMDP−fat(M)</td>
<td>0.2221</td>
<td>1.0</td>
</tr>
<tr>
<td>7. HMDP−fat(F)</td>
<td>0.0395</td>
<td></td>
</tr>
<tr>
<td>8. BXD−db−12(M)</td>
<td>0.3282</td>
<td></td>
</tr>
<tr>
<td>9. BXD−db−12(F)</td>
<td>0.1333</td>
<td></td>
</tr>
<tr>
<td>10. BXD−db−5(M)</td>
<td>0.1813</td>
<td></td>
</tr>
<tr>
<td>11. BXD−db−5(F)</td>
<td>0.6694</td>
<td></td>
</tr>
<tr>
<td>12. BXH−apoE(M)</td>
<td>0.7650</td>
<td></td>
</tr>
<tr>
<td>13. BXH−apoE(F)</td>
<td>0.7273</td>
<td></td>
</tr>
<tr>
<td>14. BXH−wt(M)</td>
<td>0.9516</td>
<td></td>
</tr>
<tr>
<td>15. BXH−wt(F)</td>
<td>0.8735</td>
<td></td>
</tr>
<tr>
<td>16. CXB−ldlr(M)</td>
<td>0.3721</td>
<td></td>
</tr>
<tr>
<td>17. CXB−ldlr(F)</td>
<td>0.0204</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 6.35: Forest plot and PM-plot for Chr19:3319089 locus**
Figure 6.36: Forest plot and PM-plot for ChrX:151384614 locus
6.8 Supplementary Materials

6.8.1 Power comparison between Meta-GxE, fixed effects meta-analysis and Heterogeneity testing approach

We estimate power for the case in which the genetic effect exists in a subset of the studies. In our simulation study, assuming five studies of equal sample size of 1,000, we estimate power when all five studies exhibited the effect, when only 4 studies exhibited the effect, and so on. Supplementary Figure S1 shows that as the number of studies with effect decreases, the power of FE drops. In contrast, the RE method maintains high power regardless of the number of studies exhibiting and effect. The power of HE method is increasing as there exists more heterogeneity due to gene-by-environment interactions. Note that RE method is more powerful than HE method when there are 2 studies having interaction effects. In this case, two studies have interaction effects and three studies do not have an effect. Since the magnitude and direction of effects are same in 2 studies with interaction effects, the regression for estimating effect sizes will behavior as if there exists a common effect with effect size $\frac{2}{5}\beta$. For this reason, RE method is most powerful over HE and FE in this case. We applied FE analysis method to 17 HDL studies. Supplementary Figure S2 shows the Manhattan plot of the result of FE analysis. The FE method found fewer (14) significant loci compared to the 26 discovered by Meta-GxE.

6.8.2 Power and type I error comparison between Meta-GxE approach and traditional Wald test based GxE testing approach

To compare the power and type I error (false positive rate) between the Meta-GxE approach and traditional Wald test based approach to identify gene-by-environment interactions, we first used each approach to identify both environment-shared and
environment-specific effect on the phenotype in simulated data sets. For this power comparison, we generated 6 simulated genotype data sets with 1000 individuals assuming a minor allele frequency of 0.3 and we simulated the phenotype using equation (3). For all 6 simulated studies, we assume that there is no environment-shared effect ($\beta = 0$), We also assume that one study has no gene-by-environment effect ($\gamma_i = 0$) and some of the remaining 5 studies have gene-by-environment interactions ($\gamma_i \neq 0$). Supplementary Figure S1 and S2 show the power comparison of RE approach vs traditional Wald Test based approach and that of HE approach vs traditional Wald Test based approach. The proposed RE and HE approaches consistently outperform the traditional Wald test based testing framework. Supplementary Table S1 and S2 show the false positive rate comparison of RE approach vs traditional Wald Test based approach and that of HE approach vs traditional Wald Test based approach. Both the RE and HE approaches correctly control the false positive rate.

6.8.3 Details of 17 HDL mouse studies

Here we provide details of each of the mouse genetic studies involved in the meta-analysis.

6.8.3.1 BxH-Apoe(M/F)

C57BL/6J Apoe$^{-/-}$ (B6.Apoe$^{-/-}$) were purchased from Jackson Laboratory (Bar Harbor, Maine, United States). C3H/HeJ Apoe$^{-/-}$ (C3H.Apoe$^{-/-}$) were generated by backcrossing B6.Apoe$^{-/-}$ to C3H for ten generations. F1 mice were generated from reciprocal intercrossing between B6.Apoe$^{-/-}$ and C3H.Apoe$^{-/-}$, and F2 mice were subsequently bred by intercrossing F1 mice. A total of 334 F2 mice (169 female, 165 male) were produced. All mice were fed Purina Chow containing 4% fat until 8 wk of age and then transferred to a "Western" diet containing 42% fat and 0.17%
cholesterol for 16 wk. Mice were euthanized and plasma collected at 24 wk of age [WYS06].

6.8.3.2 BxD-db-5(M/F) and BxD-db-12(M/F)

We previously carried out an F2 intercross between the inbred strains DBA/2 and C57BL/6 [DNC12]. These parental mice were obtained from The Jackson Laboratory (Bar Harbor, ME). The male C57BL/6 parents carried heterozygosity deficiency in the leptin receptor (db +/-) and F1 progeny were selected for homozygosity of the mutant allele. Among F2 progeny, only those with homozygous deficiency in leptin receptor (db/db) were selected. Animals were fed a chow diet with 6% fat by weight. Plasma levels of HDL were measured in both genders at 5 weeks of age (BxD-db-5(M/F)) and at 12 weeks of age (BxD-db-12(M/F)).

6.8.3.3 HMDP-chow(M)

Male mice from the hybrid HMDP panel were purchased from The Jackson Laboratory and studied as previously described [BFO10b]. Mice were between 6 and 10 wk of age on arrival and to ensure adequate acclimatization to a common environment the mice were aged until 16 wk of age. All mice were maintained on a chow diet (Ralston-Purina Co.) until sacrifice at 16 wk of age. Following a 16-h fast, mice were bled retro-orbitally under isoflurane anesthesia.

6.8.3.4 HMDP-fat(M/F)

All mice were obtained from The Jackson Laboratory and were bred at University of California, Los Angeles to generate mice used in this study as previously described [PNO13]. Purchased mice were maintained on a chow diet (Ralston Purina Company)
until 8 weeks of age. Then they were given a high-fat, high-sucrose diet (Research DietsD12266B) with the following composition: 16.8% kcal protein, 51.4% kcal carbohydrate, and 31.8% kcal fat for 8 weeks. Plasma was collected at 16 weeks of age.

6.8.3.5 HMDPxB-chow(M/F) and HMDPxB-ath(M/F)

Mice for the HMDPxB panel were created by breeding females of the various HMDP inbred strains [BFO10b] to males carrying transgenes for both Apoe Leiden [MHK93] and for human Cholesterol Ester Transfer Protein (CETP) [JAW92] on a C57BL/6 genetic background. Male and female F1 progeny were genotyped to verify presence of both transgenes and then provided a normal chow diet until 8 weeks of age. At that time, after a 4h fast, animals were anesthetized with isoflurane vapor and a sample of blood (HMDPxB-chow) was collected via the retro orbital sinus and plasma was prepared using EDTA coated tubes (Cat.365973, Becton Dickinson, Franklin Lakes, NJ). These plasmas represent the HMDPxB-chow(M) (male) and HMDPxB-chow(F) (female) populations. The animals were then fed a high-fat, high-cholesterol diet (33 kcal% fat (mostly cocoa butter) + 1% cholesterol) (Research Diets, New Brunswick, NJ) for 16 weeks at which time a second blood sample was collected by the same procedures. These plasmas represent the HMDPxB-ath(M) (male) and HMDPxB-ath(F) (female) populations. All mice were maintained on a 12h light and 12h dark cycle and fed ad libitum.

6.8.3.6 CxB-ldlr(M/F)

Female BALB/cByJ-LDLRKO (designated as C) mice were crossed with male C57BL/6J-LDLRKO (designated as B) to generate F1 mice. Intercross of F1 was performed to generate F2 mice. A total of 124 male and 64 female F2 mice were phenotyped and genotyped. These F2 mice were fed a Western diet (Harlan Teklad) for 12 weeks be-
fore euthanasia and collection of plasma. For Genotyping, Illumina 1440 SNP Mouse medium density microarray was used with 799 informative SNPs. Two additional SNPs at 30 Mb and 67 Mb of chromosome 2 were also typed manually. Therefore a total of 801 markers were used. The average marker density was 1 SNP per 2.2 Mb (Unpublished).

6.8.3.7 BxH-wt(M/F)

BXH wild type (BXH/wt) mice were produced as previously described [NIS10]. Briefly, C57BL/6J mice were intercrossed with C3H/HeJ mice to generate 321 F2 progeny (161 females, 160 males). These mice were fed Purina Chow (Ralston-Purina) containing 4% fat until 8 wk of age, after which, they were fed a western diet (Teklad 88137, Harlan Teklad) containing 42% fat and 0.17% cholesterol. At 20 weeks of age, plasma samples were collected as described above. All mice were maintained on a 12 h light and 12 h dark cycle and fed ad libitum.

Reference to published article

CHAPTER 7

Genome Wide Association Study for Age-related Hearing Loss in the Mouse: A Meta-analysis

7.1 Introduction

Age-related hearing impairment (ARHI) is characterized by a symmetric sensorinueral hearing loss primarily in the high frequencies. Age of onset, progression and severity of ARHI show great variation, but males are generally more affected than females. Approximately 35% of people over the age of 65, and 50% of octogenarians suffer from ARHI [GM05].

Heritability studies have shown that the sources of this variance are both genetic and environmental, with approximately half of the variance attributable to hereditary factors [HT10]. Only a limited number of large-scale genome wide association studies (GWAS) for ARHI have been undertaken in humans, to date. A first pooled genome-wide association study for ARHI was carried out in a population consisting of eight subpopulations from 6 European countries. An association was found between ARHI and SNPs within the GRM7 gene. This gene encodes the metabotropic glutamate receptor type7, which is activated through L-glutamate, the primary excitatory neurotransmitter in the auditory system [FVH09]. This study was followed up in a US population and an association was found with GRM7 and ARHI on several measures of central auditory function (DINA). A GWAS in 352 samples from the Saami, an isolated
population originating from northern Finland, revealed no genome-wide significant associations for ARHI, although the authors noted one SNP immediately downstream of the GRM7 gene among the most significant association signals [VHH10].

As delineated above for ARHI, the genetic analysis of human complex traits has been revolutionized by the ability to carry out association studies on a genome-wide basis. Such GWAS have been applied to numerous complex traits [MCC09, ADL08]. Despite these successes, the fraction of the genetic component that has been explained is relatively modest for most traits. Furthermore, formal proof that a specific variant is responsible for a given trait has proven difficult. Buoyed by the prospects and successes of human association studies, several groups have proposed mouse GWAS [BFO10b, GRF12, VSG06, YNB10, CAA04, FE12, MF13, KKW10a]. For obvious reasons, mouse models have several advantages over human studies. The environment can be more carefully controlled, measurements can be replicated in genetically identical animals, and the proportion of the variability explained by genetic variation is increased. Complex traits in mouse strains have been shown to have higher heritability and genetic loci often have stronger effects on the trait compared to humans [LWD00, WPB03, YWF04]. Furthermore, several recently developed strategies for mouse genetic studies, such as use of the Hybrid Mouse Diversity Panel (HMDP), provide much higher resolution for associated loci than traditional approaches to quantitative trait loci (QTL) mapping [BFO10b]. In genome-wide association studies in humans, the use of meta-analysis is becoming more and more popular because one can virtually collect tens of thousands of individuals that will provide power to identify associated variants with small effect sizes [HMK13, AWG13, BSJ13].

Motivated by the success in human studies, in this chapter, we describe the use of a meta-analytic approach based upon a random effects model to identify loci for ARHI in the mouse by combining data sets from several studies. Utilizing a frame-
work facilitating the interpretation of the results of the meta-analysis combining inbred strains within the HMDP, a QTL study mapping ahl8, and auditory evoked potential data on common inbred strains (JAX), we are able to distinguish between the data sets predicted to have an effect, the data sets predicted not to have an effect, and the ambiguous studies that were under-powered [HE12] for each locus. Several loci that we have identified with our combined analysis were not previously evident in any of the individual studies.

7.2 Results

7.2.1 New loci discovered through random-effects meta-analysis

We combined heterogeneous phenotypic data sets (auditory evoked potential (ABR) thresholds for 8 kHz, 16 kHz, and 32 kHz) including 226 classic inbred strains (Zheng M and F), 387 F2 backcross mice (M and F) from the ahl8 mapping study [JLG08], and 324 mice from our HMDP panel. The ages of mice at the time of testing were different in each study. Among these datasets, the Zheng dataset phenotyped animals at different ages which, as we show in our analysis, leads to confounding factors likely causing spurious associations. Application of the random effect meta-analysis approach requires that we first compute the effect size and its standard deviation for each of the 5 studies using linear mixed model association mapping [HE11b, FKV12]. This strategy corrects for population structure underlying the data sets. Supplementary Figures S1 ∼ S3 show the manhattan plots of 5 data sets for three hearing phenotypes (8 kHz, 16 kHz, and 32 kHz tone bursts). As seen in these figures, the Zheng data yield many association peaks, likely false positives secondary to insufficient power. Thus we excluded the Zheng data from the meta-analysis. We applied meta-analysis on the remaining three data sets (Johnson(F), Johonson(M), and HMDP(F)). Figure
7.1 shows the manhattan plot combining these data for the three hearing phenotypes (8khz, 16khz, 32khz). Tables 7.1 summarize significant peaks from our meta-analysis for the three hearing phenotypes (8khz, 16khz, 32khz). The combined meta p-value for the ah18 locus is very significant at all frequencies tested.

The effect sizes, standard errors and p-values for the chromosome 11 locus are shown as forest plots in Figures 2 - 4 (a). The size of black square and horizontal lines for each study in forest plot represent the precision of the effect size estimate and standard error of the effect size estimate respectively. The diamond rectangle represents the summarized log odds ratio using all studies. While we excluded the Zheng data due to possible confounding, we include it in the forest plots to aid in interpretation of the significant peaks. Similar effect sizes in the Zheng data compared to the other studies for the chromosome 11 peak suggest that even in the presence of confounding factors, the effect is present in the Zheng data.

7.2.2 Interpreting meta-analysis using posterior probabilities

As seen in the tables and figures, the presence of the effect may not be reflected in the study-specific p-value due to a lack of statistical power. Therefore, it is difficult to distinguish if an effect is absent in a particular study due to a gene-by-environment interaction at the locus or a lack of power. In order to identify which studies have effects, we utilize a statistic called the m-value [HE12], which estimates the posterior probability of an effect being present in a study given the observations from all other studies. We visualize the results through a PM-plot, in which p-values (y-axis) are simultaneously visualized with the m-values (x-axis) at each tested locus. These plots allow us to identify in which studies genetic variation at the locus has an effect and in which it does not. M-values for a given variant have the following interpretation: a study with a small m-value(\(<= 0.1\)) is predicted not to be affected by the variant,
Figure 7.1: Mouse hearing association results from random effect meta analysis

while a study with a large m-value (\( \geq 0.9 \)) is predicted to be affected by the variant. Whether or not an effect is present for a study with a m-value between 0.1 and 0.9 is ambiguous. The m-values are also displayed by the study name near the forest plot where the color dot on the left-hand side of the study name provides information about the m-value. A red dot indicates that the study’s m-value is greater than 0.9, a blue dot represents that the study’s m-value is less than 0.1, and a green dot represents that the
Table 7.1: (a) Three significant loci were identified in the meta-analysis for the 8khz tone burst. (b) Four significant loci were identified in the meta-analysis for the 16khz tone burst. (c) Three significant loci were identified in the meta-analysis for the 8khz tone burst. E denotes the number of studies with an effect on the 8 kHz phenotype. N denotes the number of studies without an effect. A denotes the number of studies with an ambiguous effect.
study’s m-value is between 0.1 and 0.9.

The PM-plots for the chromosome 11 locus for each tested frequency are shown in Figures 7.2 (b) - 7.4 (b). If we only look at the separate study p-values (y-axis), we can conclude that this locus only has an effect for 8 kHz in the Johnson (F) cohort. However, if we look at m-value (x-axis), then we find 2 studies (Johnson (F) and Johnson (M)), where we predict that the variation has an effect, while in the other 3 studies we predict there is no effect. For the 16 kHz trait, we see only the Johnson (M) and (F) cohorts have an effect individually but if we consider the m-values all 5 demonstrate an effect at this locus. Lastly, for the 8 kHz trait, only the Johnson (M) cohort has an individual effect but taken in aggregate three studies, Johnson (M and F) and HMDP (F) have an effect at this locus.

In Tables 1-3, the column E/A/N summarizes the m-values for each associated peak providing the counts of the number of studies where there is a predicted effect (E), the number of studies where the effect is ambiguous (A) and the number of studies where we predict that the effect is not present (N). A comparison of this column compared to the column providing the count of the studies with a significant p-value shows that, using the m-value predicts significantly more studies that have an effect compared with the p-value alone. This is consistent with the fact that each individual study is underpowered.

Through forest plot, we can easily see which study shows a strong effect on the hearing phenotype at this locus and which study does not show an effect on the hearing phenotype at this locus. One interesting observation is that for 16kHz hearing phenotype in Figure 7.3, the log odds ratio of the effect size of all 5 studies are very close (around 0.5). This is a strong evidence that an effect of the locus Chr11:120818214 on mouse hearing ability is likely to be real biological signal. For two other hearing phenotypes (8kHz and 32kHz) shown in Figure 7.2 and 7.4, the forest plots show weak
effect size in two Johnson studies. Supplementary Figures 7.8 to 7.18 show the forest and PM plots for all of the discovered loci for each of the phenotypes.

7.2.3 Novel loci identified in meta-analysis

Our approach discovered an additional four loci for ARHI in mice. What is evident from the data in Table 7.1 is the power of the meta-analysis approach to identify additional loci in the combined data set that would have gone unrecognized in the individual studies due to insufficient power. As seen in Table 1, all but one of the cohorts in the meta-analysis detected SNP associations with the 8 kHz trait whereas in the individual cohorts only two of the five were significant. For the 16 kHz and 32 kHz traits, only one study was powered to detect an association at the Ahl8 locus by itself, whereas 3 studies showed association in the meta-analysis. The other loci would have gone unrecognized without the combined analysis.
Chr11:120818214 (Meta $P = 3.32 \times 10^{-15}$)

<table>
<thead>
<tr>
<th>Study Name</th>
<th>$P$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zheng(F)</td>
<td>0.0367</td>
</tr>
<tr>
<td>Zheng(M)</td>
<td>0.1866</td>
</tr>
<tr>
<td>Johnson(F)</td>
<td>$6.47 \times 10^{-6}$</td>
</tr>
<tr>
<td>Johnson(M)</td>
<td>$3.14 \times 10^{-6}$</td>
</tr>
<tr>
<td>HMDP(F)</td>
<td>0.0039</td>
</tr>
</tbody>
</table>

RE Summary:

![Forest plot and PM-plot for Chr11:120818214 locus (16kHz)](image)

Figure 7.3: (A) Forest plot and (B) PM-plot for Chr11:120818214 locus (16kHz)

Chr11:120818214 (Meta $P = 1.06 \times 10^{-12}$)

<table>
<thead>
<tr>
<th>Study Name</th>
<th>$P$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zheng(F)</td>
<td>0.9407</td>
</tr>
<tr>
<td>Zheng(M)</td>
<td>0.2500</td>
</tr>
<tr>
<td>Johnson(F)</td>
<td>$1.35 \times 10^{-6}$</td>
</tr>
<tr>
<td>Johnson(M)</td>
<td>$2.42 \times 10^{-6}$</td>
</tr>
<tr>
<td>HMDP(F)</td>
<td>0.0687</td>
</tr>
</tbody>
</table>

RE Summary:

![Forest plot and PM-plot for Chr11:120818214 locus (32kHz)](image)

Figure 7.4: (A) Forest plot and (B) PM-plot for Chr11:120818214 locus (32kHz)
Lastly, a look at the meta-analysis data reveals that there are three loci associated with the 8 kHz trait on chromosomes 1, 8, and 11. However, for the 16 and 32 kHz studies, there are additional loci on chromosomes 9 and 18, respectively. This underlies the complexity of the hearing phenotype, its polygenic nature, and the likely effects of genetic variation on specific regions of the cochlea. The identification of an additional locus for the 32 kHz trait is also interesting in light of the fact that the \textit{Fscn2} BAC transgenic rescue of the \textit{ahl8} associated hearing loss in DBA/2J was successful at all frequencies except 32 kHz. It is possible that a locus on chromosome 18 may contribute to DBA/2J hearing loss. Equally plausible, given the size of the BAC transgenic, there may be additional genes in the \textit{ahl8} interval contributing to the phenotype.

The resolution in mouse association studies is limited by the size of linkage disequilibrium blocks. For the loci reported here, these blocks span roughly 2 Mb (consistent with average LD blocks in previous HMDP studies [DNB13, BFO10b] and contain a number of potential candidates including membrane transporter genes, transcription factors, and neurotransmitter receptors. Many of the genes within the intervals are expressed in the cochlea (data not shown) complicating candidate selection to some degree; however, the gene lists are relatively small and serve as a starting point for gene discovery. \textit{Snhl2}, an ARHI locus identified in a backcross between \textit{ALR/LtJ} and \textit{C3HeB/FeJ} mapped to chromosome 1 (133 Mb-172 Mb) is nearby and these data may serve to refine this locus or may represent an additional locus [LNN11]. The two loci, \textit{Hfhl3} on chromosome 9 and \textit{ahl6} on chromosome 18, map closely to the regions identified in our meta-analysis [KN12, DN06]. In contrast, although \textit{Hfhl2} maps to chromosome 8, it is very proximal to the locus identified in our study and likely represents a novel locus [KNL11].
7.3 Discussion

In this chapter, we used a recently developed meta-analysis approach that can be applied to a large number of heterogeneous studies each conducted in different environments with animals from different genetic backgrounds, different genders and different ages at the time of testing. We show the practical utility of the proposed method by applying it to 5 mouse ARHI studies containing 937 samples, and we successfully identify several loci involved in ARHI in the mouse, including the known locus, ahl8. Consistent with the results of meta-analysis in human studies, our combined study recognized loci that were not discovered in any of the individual studies.

Part of the reason for our success in identifying a several loci is that our study combined multiple mouse studies using fundamentally different mapping strategies. Over the past few years, many new strategies have been proposed beyond the traditional F2 cross [FE12] which include the hybrid mouse diversity panel (HMDP) [BFO10b, GRF12], heterogeneous outbred stocks [VSG06], commercially available outbred mice [YNB10], and the collaborative cross [CAA04]. In our current study, we are combining data from two HMDP strategies with an N2 backcross. The meta-analysis benefits from the statistical power and resolution advantages of this combination [FKV12].

As shown in Table 7.2, there are currently 17 autosomal loci and one mitochondrial locus for ARHI in mice (www.hearingimpairment.jax.org). Of these 18 loci, only 6 have been characterized at the gene level illustrating the inherent difficulties in traditional gene mapping studies. Much of the progress in the genetics of hearing disorders in the mouse has come from the application of linkage analysis (i.e. QTL analysis) to identify naturally occurring single gene mutations (Mendelian traits) and the analysis of targeted gene deletions. Little attention has been directed towards the definition of the genetics of common hearing disorders. Classical genetic approaches have been
used to identify several QTLs that are associated with ARHI in mice. However one of the most significant shortcomings of QTL analysis is the use of a limited resource (i.e. a segregating F2 population). Another limitation of this approach is the genomic resolution, typically on the order of 10 Mb, or greater. Hence, further progress with these strategies depends upon execution of large-scale fine-mapping crosses where-in several thousand N2 or F2 animals are screened for recombination within the region of interest. Following this several generations of subcongenic lines must be established followed by progeny testing. This process can take years of effort and never achieve the resolution necessary to identify individual genes or causal variants. In this manuscript we have demonstrated the utility of GWAS using a meta-analysis approach for the high resolution mapping of loci associated with ARHI in the mouse. Subsequent studies will include a replication of this experiment utilizing the entire HMDP and the selection of candidate genes based upon cochlear eQTL information, cochlear gene expression, and genes known to be associated with hearing loss in either mouse or humans. Since the mouse inner ear is functionally and genetically very similar to the human ear, these studies will facilitate the identification of genes and pathways associated with ARHI in humans.

7.4 Methods and Materials

7.4.1 Controlling for population structure within studies

Model organism such as mouse is well-known for the population structure or cryptic relatedness [DRB01, VP05], in which genetic similarities between individuals both inhibit the ability to find true associations and cause the appearance of a large number of false or spurious associations. Mixed effects models are often used in order to correct this problem [Lan02, YPB06, KZW08]. Methods employing a mixed effects
correction account for the genetic similarity between individuals with the introduction of a random variable into the traditional linear model.

\[ Y_i = \mu + \delta_i X_r + u_i + \epsilon \]  

(7.1)

In the model in equation (8.8), the random variable \( u_i \) represents the vector of genetic contributions to the phenotype for individuals in population \( i \). This random variable is assumed to follow a normal distribution with \( u_i \sim N(0, \sigma_g^2 K_i) \), where \( K_i \) is the \( n_i \times n_i \) kinship coefficient matrix for population \( i \). With this assumption, the total variance of \( Y_i \) is given by \( \Sigma_i = \sigma_g^2 K_i + \sigma_e^2 I \). A z-score statistic is derived for the test \( \delta_i = 0 \) by noting the distribution of the estimate of \( \hat{\delta}_i \). In order to avoid complicated notation, we introduce a more basic matrix form of the model in equation (8.8), shown in equation (8.9).

\[ Y_i = S_i \Gamma + u_i + \epsilon \]  

(7.2)

In equation (8.9), \( S_i \) is a \( n_i \times 2 \) matrix encoding the global mean and SNP vectors and \( \Gamma \) is a \( 2 \times 1 \) coefficient vector. We note that this form also easily extends to models with multiple covariates. The maximum likelihood estimate for \( \Gamma \) in population \( i \) is given by \( \hat{\Gamma}_i = (S_i' \Sigma_i^{-1} S_i)^{-1} S_i' \Sigma_i^{-1} Y_i \), which follows a normal distribution with a mean equal to the true \( \Gamma \) and variance \( (S_i' \Sigma_i^{-1} S_i)^{-1} \). The estimates of the effect size \( X_i \) and standard error of the \( X_i \) \( (SE(X_i)) \) are then given in equation (8.10) and equation (8.11), where \( R = [0 \ 1] \) is a vector used to select the appropriate entry in the vector \( \hat{\Gamma}_i \).

\[ \delta_i = \mathbf{R}(S_i' \Sigma_i^{-1} S_i)^{-1} S_i' \Sigma_i^{-1} Y_i \]  

(7.3)

\[ SE(\delta_i) = \left[ \mathbf{R}(S_i' \Sigma_i^{-1} S_i)^{-1} \mathbf{R}' \right]^{1/2} \]  

(7.4)
This mixed model based approach, well-known for correcting population structure was applied to the studies reported here.

### 7.4.1.1 Random effects model meta analysis

Under the random effects model meta-analysis, we explicitly model heterogeneity by assuming a hierarchical model. We assume that the effect size of each study $\delta_i$ is a random variable picked by random from a distribution with the grand mean $\delta$ and the variance $\tau^2$,

$$\delta_i \sim N(\delta, \tau^2)$$

We recently developed a powerful random effects model, which addresses the problem of the conservative nature of traditional random effects model by assuming no heterogeneity under the null hypothesis [HE11b]. This modification is natural because the effect size should be fixed to be zero under the null hypothesis. This random effects model tests the null hypothesis $\delta = 0$ and $\tau^2 = 0$ versus the alternative hypothesis $\delta \neq 0$ or $\tau^2 \neq 0$.

Similarly to the traditional random effect model [DL86a], we use the likelihood ratio framework considering each statistic as a single observation. Since we assume no heterogeneity under the null, $\mu = 0$ and $\tau^2 = 0$ under the null hypothesis. The likelihoods are then

$$L_0 = \prod_i \frac{1}{\sqrt{2\pi V_i}} \exp\left(-\frac{\delta_i^2}{2V_i}\right)$$

$$L_1 = \prod_i \frac{1}{\sqrt{2\pi(V_i + \tau^2)}} \exp\left(-\frac{(\delta_i - \mu)^2}{2(V_i + \tau^2)}\right).$$

The maximum likelihood estimates $\hat{\mu}$ and $\hat{\tau^2}$ can be found by an iterative procedure suggested by Hardy and Thompson [HT96]. Then the likelihood ratio test statistic can
be built

\[ S_{\text{meta}} = -2 \log(\lambda) = \sum \log \left( \frac{V_i}{V_i + \tau^2} \right) + \sum \frac{\delta_i^2}{V_i} - \sum \frac{(\delta_i - \hat{\mu})^2}{V_i + \tau^2}, \quad (7.5) \]

where \( \delta_i = R(S'_i\Sigma^{-1}_iS_i)^{-1}S'_i\Sigma^{-1}_iY_i \) and \( V_i = [R(S'_i\Sigma^{-1}_iS_i)^{-1}R'] \). \textit{P-values} can be easily computed using precomputed tabulated values [HE11b].

**Identifying studies with an effect**

After identifying loci exhibiting interaction effects, we employ the meta-analysis interpretation framework that we recently developed. The \textit{m-value} [HE12] is the posterior probability that the effect exists in each study. Suppose we have \( N \) number of studies we want to combine. Let \( D = [\delta_1, \delta_2, ..., \delta_N] \) be the vector of estimated effect sizes and \( V = [V_1, V_2, ..., V_N] \) be the vector of estimated variance of \( N \) effect sizes. We assume that the effect size \( \delta_i \) follows the normal distribution.

\[ P(\delta_i|\text{no effect}) = N(\delta_i; 0, V_i) \quad (7.6) \]

\[ P(\delta_i|\text{effect}) = N(\delta_i; \mu, V_i) \quad (7.7) \]

Normality assumption above can be easily justified by the large number of sample size in the current GWASs. We assume that the prior for the effect size is

\[ \mu \sim N(0, \sigma^2) \quad (7.8) \]

A possible choice for \( \sigma \) in GWAS is 0.2 for small effect and 0.4 for large effect [SB09]. We also denote \( E_i \) be a random variable whose value is 1 if a study \( i \) has an effect and 0 otherwise. We also denote \( E \) as a vector of \( E_i \) for \( N \) studies. Since \( E \) has \( N \) binary values, \( E \) can be \( 2^N \) possible configurations. Let \( U = [e_1, ..., e_{2^N}] \) be a vector containing all the possible these configurations. We define \textit{m-value} \( m_i \) as the probability \( P(E_i = 1|D) \), which is the probability of study \( i \) having an effect given
the estimated effect sizes. We can compute this probability using the Bayes’ theorem in the following way.

\[ m_i = \frac{\sum_{e \in U_i} P(D|E = e)P(E = e)}{\sum_{e \in U} P(D|E = e)P(E = e)} \]  

(7.9)

where \( U_i \) is a subset of \( U \) whose elements’ \( i \)th value is 1. Now we need to compute \( P(D|E = e) \) and \( P(E = e) \). \( P(E = e) \) can be computed as

\[ P(E = e) = \frac{\text{B}(|e| + \alpha, N - |e| + \beta)}{\text{B}(\alpha, \beta)} \]  

(7.10)

where \(|e|\) denotes the number of 1’s in \( e \) and \( \text{B} \) denotes the beta function and we set \( \alpha \) and \( \beta \) as 1 [HE12]. The probability \( D \) given configuration \( e \), \( P(D|E = e) \), can be computed as

\[ P(D|E = e) = \int_{-\infty}^{\infty} \prod_{i \in t_0} N(\delta_i; 0, V_i) \prod_{i \in t_1} N(\delta_i; \mu, V_i)p(\mu) \, d\mu \]  

(7.11)

\[ = \bar{C}N(\bar{\delta}; 0, \bar{V} + \sigma^2) \prod_{i \in t_0} N(\delta_i; 0, V_i) \]  

(7.12)

where

\[ \bar{\delta} = \frac{\sum_i W_i \delta_i}{\sum_i W_i} \quad \text{and} \quad \bar{V} = \frac{1}{\sum_i W_i} \]  

(7.13)

where \( N(\delta; a, b) \) denotes the probability density function of the normal distribution with mean \( a \) and variance \( b \). \( W_i = V_i^{-1} \) is the inverse variance or precision and \( \bar{C} \) is a scaling factor.

\[ \bar{C} = \frac{1}{(\sqrt{2\pi})^{N-1}} \sqrt{\prod_i W_i} \exp \left\{ -\frac{1}{2} \left( \sum_i W_i \delta_i^2 - \frac{(\sum_i W_i \delta_i)^2}{\sum_i W_i} \right) \right\} \]  

(7.14)

All summations appeared for computing \( \bar{\delta}, \bar{V} \) and \( \bar{C} \) are with respect to \( i \in t_1 \).

The \( m \)-values have the following interpretations: small \( m \)-values (0.1) represent a study that is predicted to not have an effect, large \( m \)-values (0.9) represent a study that is predicted to have an effect, otherwise it is ambiguous to make a prediction.
It was previously reported that *m-values* can accurately distinguish studies having an effect from the studies not having an effect [HE12]. For interpreting and understanding the result of the meta-analysis, it is informative to look at the *P*-value and *m*-value at the same time. We applied the PM-plot framework [HE12], which plots the *P*-values and *m*-values of each study together in two dimensions. For studies with an *m*-value between $0.1$ and $0.9$, we cannot make a decision. One reason that studies are ambiguous ($0.9 \leq m\_value \leq 0.1$) is that they are underpowered due to small sample size. If the sample size increases, the study can be drawn to either the left or the right side.
7.5 Supplementary Materials
7.5.1 Mouse hearing association results

Figure 7.5: The association result of three studies for 8kHz mouse hearing.
Figure 7.6: The association result of three studies for 16khz mouse hearing.

7.5.2 PM-plots for significant locus
Figure 7.7: The association result of three studies for 32khz mouse hearing.
<table>
<thead>
<tr>
<th>AHL locus</th>
<th>Chr</th>
<th>Location</th>
<th>Mapping</th>
<th>Known strains</th>
<th>Known strains</th>
<th>references</th>
</tr>
</thead>
<tbody>
<tr>
<td>(gene)</td>
<td></td>
<td></td>
<td>methods</td>
<td>with</td>
<td>with</td>
<td></td>
</tr>
<tr>
<td>*</td>
<td></td>
<td></td>
<td>susceptibility</td>
<td>allele</td>
<td>resistance</td>
<td></td>
</tr>
<tr>
<td>ah1(Cdh23)</td>
<td>10</td>
<td>60Mb</td>
<td>backcross, strain</td>
<td>Many</td>
<td>Many</td>
<td>[JEC97]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>haplotypes</td>
<td></td>
<td></td>
<td>[JZE00]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>[NZJ03]</td>
</tr>
<tr>
<td>ah14(Cs)</td>
<td>10</td>
<td>120-130 Mb</td>
<td>RI, CS strains</td>
<td>A/J</td>
<td>C57BL/6J,</td>
<td>[ZDY09]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>backcross</td>
<td></td>
<td>CAST/Ei</td>
<td>[JGL12]</td>
</tr>
<tr>
<td>ah15(Gipc3)</td>
<td>10</td>
<td>81Mb</td>
<td>backcross</td>
<td>Black Swiss</td>
<td>CAST/Ei</td>
<td>[DN06]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>intercross</td>
<td></td>
<td></td>
<td>[CLS11]</td>
</tr>
<tr>
<td>ah18(Fscn2)</td>
<td>11</td>
<td>120Mb</td>
<td>RI strains</td>
<td>DBA/2J</td>
<td>C57BL/6J</td>
<td>[JLG08]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>backcross</td>
<td></td>
<td></td>
<td>[SLG10]</td>
</tr>
<tr>
<td>Gpr98, G protein</td>
<td>13</td>
<td>81Mb</td>
<td>backcross</td>
<td>BUB/BnJ</td>
<td>CAST/Ei</td>
<td>[JZW05]</td>
</tr>
<tr>
<td>coupled receptor 9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mtDNA(mt-tr)</td>
<td>13</td>
<td>mitochondria</td>
<td>backcross</td>
<td>A/J</td>
<td>CAST/Ei</td>
<td>[JZB01]</td>
</tr>
</tbody>
</table>

Table 7.2: Genetic factors that contribute to age-related hearing loss in inbred mouse strains. CS=chromosome substitution; RI=recombinant inbred.
Chr1:196270162 (Meta $P = 1.05 \times 10^{-9}$)

**PM-Plot**

<table>
<thead>
<tr>
<th>$P$-value</th>
<th>Study Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.4667</td>
<td>Zheng(F)</td>
</tr>
<tr>
<td>0.5960</td>
<td>Zheng(M)</td>
</tr>
<tr>
<td>$3.87 \times 10^{-6}$</td>
<td>Johnson(F)</td>
</tr>
<tr>
<td>$6.58 \times 10^{-6}$</td>
<td>Johnson(M)</td>
</tr>
<tr>
<td>0.1668</td>
<td>HMDP(F)</td>
</tr>
</tbody>
</table>

**RE Summary**

- Study has an effect ($m > .9$)
- Study does not have an effect ($m < .1$)
- Study's effect is uncertain ($0.1 < m < .9$)

Figure 7.8: (A) Forest plot and (B) PM-plot for Chr1:196270162 locus (8khz)

Chr8:130247700 (Meta $P = 3.92 \times 10^{-6}$)

**PM-Plot**

<table>
<thead>
<tr>
<th>$P$-value</th>
<th>Study Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8162</td>
<td>Zheng(F)</td>
</tr>
<tr>
<td>0.3230</td>
<td>Zheng(M)</td>
</tr>
<tr>
<td>0.0001</td>
<td>Johnson(F)</td>
</tr>
<tr>
<td>0.0002</td>
<td>Johnson(M)</td>
</tr>
<tr>
<td>0.9248</td>
<td>HMDP(F)</td>
</tr>
</tbody>
</table>

**RE Summary**

- Study has an effect ($m > .9$)
- Study does not have an effect ($m < .1$)
- Study's effect is uncertain ($0.1 < m < .9$)

Figure 7.9: (A) Forest plot and (B) PM-plot for Chr8:130247700 locus (8khz)
**Chr11:120818214 (Meta \( P = 3.08 \times 10^{-11} \))**

<table>
<thead>
<tr>
<th>Study Name</th>
<th>( P )-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zheng(F)</td>
<td>0.9353</td>
</tr>
<tr>
<td>Zheng(M)</td>
<td>0.2562</td>
</tr>
<tr>
<td>Johnson(F)</td>
<td>1.65 x 10^{-5}</td>
</tr>
<tr>
<td>Johnson(M)</td>
<td>6.07 x 10^{-7}</td>
</tr>
<tr>
<td>HMDP(F)</td>
<td>0.1104</td>
</tr>
</tbody>
</table>

**Log odds ratio**

-1.0 −0.5 0.0 0.5 1.0

**PM-Plot**

- Study has an effect (\( m > 0.9 \))
- Study does not have an effect (\( m < 0.1 \))
- Study's effect is uncertain (0.1 < \( m < 0.9 \))

---

**Chr1:196270162 (Meta \( P = 5.97 \times 10^{-7} \))**

<table>
<thead>
<tr>
<th>Study Name</th>
<th>( P )-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zheng(F)</td>
<td>0.7866</td>
</tr>
<tr>
<td>Zheng(M)</td>
<td>0.8271</td>
</tr>
<tr>
<td>Johnson(F)</td>
<td>0.0009</td>
</tr>
<tr>
<td>Johnson(M)</td>
<td>0.0019</td>
</tr>
<tr>
<td>HMDP(F)</td>
<td>0.0235</td>
</tr>
</tbody>
</table>

**Log odds ratio**

-1.0 −0.5 0.0 0.5 1.0

**PM-Plot**

- Study has an effect (\( m > 0.9 \))
- Study does not have an effect (\( m < 0.1 \))
- Study's effect is uncertain (0.1 < \( m < 0.9 \))

---

Figure 7.10: (A) Forest plot and (B) PM-plot for Chr11:120818214 locus (8khz)

---

Figure 7.11: (A) Forest plot and (B) PM-plot for Chr1:196270162 locus (16khz)
Figure 7.12: (A) Forest plot and (B) PM-plot for Chr8:130247700 locus (16khz)

Figure 7.13: (A) Forest plot and (B) PM-plot for Chr9:30287464 locus (16khz)
### Chr11:120818214 (Meta $P = 3.32 \times 10^{-15}$)

<table>
<thead>
<tr>
<th>Study Name</th>
<th>$P$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zheng(F)</td>
<td>0.00367</td>
</tr>
<tr>
<td>Zheng(M)</td>
<td>0.1866</td>
</tr>
<tr>
<td>Johnson(F)</td>
<td>$6.47 \times 10^{-6}$</td>
</tr>
<tr>
<td>Johnson(M)</td>
<td>$3.14 \times 10^{-6}$</td>
</tr>
<tr>
<td>HMDP(F)</td>
<td>0.0039</td>
</tr>
</tbody>
</table>

**RE Summary**
- **●**: Study has an effect ($m > 0.9$)
- **●**: Study does not have an effect ($m < 0.1$)
- **●**: Study's effect is uncertain ($0.1 < m < 0.9$)

![Forest plot and PM-plot for Chr11:120818214 locus (16khz)](image)

Figure 7.14: (A) Forest plot and (B) PM-plot for Chr11:120818214 locus (16khz)

### Chr1:196345862 (Meta $P = 3.23 \times 10^{-7}$)

<table>
<thead>
<tr>
<th>Study Name</th>
<th>$P$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zheng(F)</td>
<td>0.6626</td>
</tr>
<tr>
<td>Zheng(M)</td>
<td>0.1313</td>
</tr>
<tr>
<td>Johnson(F)</td>
<td>0.00010</td>
</tr>
<tr>
<td>Johnson(M)</td>
<td>0.0003</td>
</tr>
<tr>
<td>HMDP(F)</td>
<td>0.0061</td>
</tr>
</tbody>
</table>

**RE Summary**
- **●**: Study has an effect ($m > 0.9$)
- **●**: Study does not have an effect ($m < 0.1$)
- **●**: Study's effect is uncertain ($0.1 < m < 0.9$)

![Forest plot and PM-plot for Chr1:196345862 locus (32khz)](image)

Figure 7.15: (A) Forest plot and (B) PM-plot for Chr1:196345862 locus (32khz)
Figure 7.16: (A) Forest plot and (B) PM-plot for Chr8:131366159 locus (32khz)

Figure 7.17: (A) Forest plot and (B) PM-plot for Chr11:120818214 locus (32khz)
Chr18:24320393 (Meta \( P = 2.32 \times 10^{-6} \))

Log odds ratio

-0.6 -0.2 0.2 0.6

Study Name

<table>
<thead>
<tr>
<th>( P )-value</th>
<th>Study Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6232</td>
<td>Zheng(F)</td>
</tr>
<tr>
<td>0.5230</td>
<td>Zheng(M)</td>
</tr>
<tr>
<td>0.0294</td>
<td>Johnson(F)</td>
</tr>
<tr>
<td>0.0068</td>
<td>Johnson(M)</td>
</tr>
<tr>
<td>0.0004</td>
<td>HMDP(F)</td>
</tr>
</tbody>
</table>

RE Summary

● Study has an effect (\( m > .9 \))
● Study does not have an effect (\( m < .1 \))
● Study's effect is uncertain (\( .1 < m < .9 \))

Figure 7.18: (A) Forest plot and (B) PM-plot for Chr18:24320393 locus (32khz)
CHAPTER 8

A Novel Meta-analysis Approach for Genome-wide Association Studies with Sex-specific Effects.

8.1 Background

Genome-wide association studies (GWAS) have successfully identified numerous genetic loci associated with complex human traits. Most GWAS studies collect both male and female individuals in the studies. Recently, there has been an increasing attention to sex-specific effects in genome-wide association studies and several recent studies have observed differences in effect sizes between males and females [KCS13, PHW13, CDL13, BJC12, KCD12, MLM10, FLW12, PMP13, ML12]. In particular, a large scale study, Randall et al., (2013) applied meta-analysis over 46 studies of anthropomorphic phenotypes analyzing males and females separately to estimate effect sizes in males and females and discovered many loci with sex-specific effects. The discovery of these sexually dimorphic associations may lead to a better understanding of the role of gene-by-sex interaction in disease mechanism.

The prevalence of sex-specific effects complicates the analysis of association studies consisting of both males and females when gene-by-sex interactions exist. The most widely used traditional approach of analyzing studies consisting of both males and females involves in analyzing all of the samples together and including sex as a covariate in the statistical model when performing the analysis [KCS13, PHW13,
CDL13, BJC12, KCD12, MLM10, FLW12, PMP13, ML12]. Treating sex as a covariate in the model adjusts for differences in phenotype mean between males and females, but ignores the possibility of sex specific effects. This naturally raises the question of how does the presence of sex-specific effects affect statistical power of GWAS when sex is modeled as a covariate and what is the best analysis strategy for a single study consisting of males and females.

Unfortunately, standard association study approaches ignores the potential for differences in effect sizes between the two sexes which may lead to a loss in statistical power. Furthermore, another complication in the analysis, as we elaborate on below, is that since gene-by-sex effects are prevalent throughout the genome, when testing for association at a given locus, the gene-by-sex effects in the rest of the genome cause what we call a gene-by-sex background effect. Improperly accounting for this effect can also lead to a decrease in statistical power.

In this paper, we present a meta-analytic approach for the analysis of genome-wide association studies consisting of both males and females. In our meta-analytic approach, called Meta-Sex, males and females are analyzed separately and the results are combined using a random effects meta-analysis approach which explicitly models difference in effect sizes between sexes. Our approach naturally increases power over the traditional sex-as-covariate approach at loci where sex specific effects are present because of the advantage of the random effects meta-analysis. We also show that meta-analysis methods are more powerful even at loci which do not have sex specific effects because of the background gene-by-sex effect. We apply our method to the Northern Finland Birth Cohort data and we show that our method has increased power over a traditional approach while controlling for false positives.

Our approach is not the first meta-analysis approach proposed for analysis of sex specific effects. A previous method, GWAMA, method simply adds the two $\chi^2$ statis-
tics estimated from each sex and performs a two degree of freedom $\chi^2$ test [MLM10]. We analytically show the connections between several meta-analysis methods including our random effects meta-analysis, traditional fixed effects meta-analysis and GWAMA. From our analysis, GWAMA loses power compared to either random effect or fixed effect meta-analysis when the effect sizes for males and females are similar. GWAMA also suffers from a loss in power when the number of male and female samples are not equal. We also extend GWAMA to handle the scenario where different numbers of males and females are in the study by adding weights to each study following similar intuitions to inverse variance weighting for fixed effect meta-analysis.

8.2 Results

8.2.1 Applying meta-analysis to handle sex-specific effects

We refer to our meta-analytic approach to handling sex-specific effects in genome wide association studies as Meta-Sex. Our approach is as follows. We first separate the male and female individuals and estimate effect sizes and standard errors for each group separately. We then combine these estimates using a meta-analysis method. While we advocate the use of the recently developed random effects meta-analysis designed for association studies (RE2) [HE11b] to combine the studies, we report results using a variety of meta-analysis methods including the fixed effects (FE) meta-analysis and the traditional random effects meta-analysis (RE) [DL86b, IPE07a, IPE07b, EMI07]. The difference between RE and RE2 is that RE is overly conservative due to that it assume there is heterogeneity present in the null hypothesis of no association. While this assumption is reasonable for some problems, in the context of association studies, loci which do not affect a trait have a true effect size of zero in all studies. RE2 is a novel random effect meta analysis approach which addresses the issues with the traditional
random effects meta-analysis [HE11b] We then check to see if there is inflation of the resulting statistics because of the potential population structure between the male and female individuals in which case we apply genomic control to correct for this inflation.

To demonstrate the performance of meta-analysis approaches, we applied 5 different approaches including three meta-analysis methods (FE, RE and RE2) and two traditional approaches ("Covariate" and "Combined) to the North Finland Birth Cohort data (NFBC) data. “Covariate” is the approach combining the male and female data and including a sex as a covariate in the statistical model. “Combined” is the approach simply combining the male and female data before performing association analysis.

8.2.1.1 Meta-analysis increases statistical power in the NFBC data

We applied our meta-analysis approach to the NFBC data. We computed the association statistics and p-values with the 5 different approaches and compute the genomic control $\lambda$ to see the inflation factor of each method. Table 8.1 shows the genomic control $\lambda$ for 5 different approaches. FE shows slight inflation and RE shows significant deflation of the association statistic as expected, since it is overly conservative. For each method, we apply genomic control to correct for the inflation in order to be able to fairly compare the different approaches. Since RE shows significant deflation, we drop this method from the further analysis.

Table 8.2 shows the significant associations for NFBC phenotypes with 4 different sex-differentiated approaches (FE, RE2, Covariate and Combined) and 2 sex-specific approaches (male only and female only). After correcting with genomic control, if any p-value from those 6 approaches is significant (threshold : $5 \times 10^{-7}$), then we include this result in the table. For each SNP, the most significant p-value among four methods (FE, RE2, Covariate and Combined) is in bold font. The “beta” column shows the effect size of male and female only study and their estimated standard errors. When
there exists sex-specific effects, the effect size differences between male and female study are large.

For each significant association, we report the “best methods”. We define the best methods by the set of the methods whose p-value is less than 2 times the most significant p-value for each association. For example, for the association at Chr11:47226831 in the hdlres phenotype, the most significant p-value is $3.96 \times 10^{-8}$. Since p-value for FE meta ($4.65 \times 10^{-8}$) is less than twice that value. Thus we also consider FE as one of the best methods. However the p-value for Covariate and Combined is larger than twice that value. Thus only FE and RE2 are included in the best methods for this

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>FE</th>
<th>RE</th>
<th>RE2</th>
<th>Covariate</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>tgres</td>
<td>1.016</td>
<td>0.785</td>
<td>0.995</td>
<td>1.001</td>
<td>1.002</td>
</tr>
<tr>
<td>hdlres</td>
<td>1.024</td>
<td>0.794</td>
<td>1.008</td>
<td>1.003</td>
<td>1.003</td>
</tr>
<tr>
<td>ldlres</td>
<td>1.024</td>
<td>0.795</td>
<td>1.011</td>
<td>1.002</td>
<td>1.002</td>
</tr>
<tr>
<td>bmiresh</td>
<td>1.014</td>
<td>0.789</td>
<td>0.992</td>
<td>0.994</td>
<td>0.994</td>
</tr>
<tr>
<td>crpres</td>
<td>1.004</td>
<td>0.774</td>
<td>0.998</td>
<td>0.993</td>
<td>0.993</td>
</tr>
<tr>
<td>glures</td>
<td>1.017</td>
<td>0.784</td>
<td>1.011</td>
<td>1.008</td>
<td>1.007</td>
</tr>
<tr>
<td>insres</td>
<td>1.011</td>
<td>0.781</td>
<td>1.006</td>
<td>1.005</td>
<td>1.004</td>
</tr>
<tr>
<td>sysres</td>
<td>1.023</td>
<td>0.796</td>
<td>1.012</td>
<td>1.006</td>
<td>1.005</td>
</tr>
<tr>
<td>diures</td>
<td>1.021</td>
<td>0.789</td>
<td>1.013</td>
<td>1.006</td>
<td>1.006</td>
</tr>
<tr>
<td>height</td>
<td>1.042</td>
<td>0.812</td>
<td>1.028</td>
<td>1.003</td>
<td>1.003</td>
</tr>
</tbody>
</table>

Table 8.1: The comparison of genomic control $\lambda$ for 5 approaches across 10 NFBC phenotypes.
SNP. The interesting observation in the best methods category, when applying these four methods, is that meta analysis approaches (FE and RE2) are typically selected as the best methods together and the standard approaches (CV, CM) are likely selected as the best methods together. To get an overall comparison of the relative power of the approaches, we compare the number of times the meta-analysis approaches are selected as best methods relative to the number of times the traditional approaches are selected as the best methods. For the 6 NFBC phenotypes with significant associations, FE/RE2 approaches are the best methods for 13 associated loci whereas CV/CM approaches are the best methods for only 3 associated loci. For 11 associated loci, all four approaches provide similar p-values.

### 8.2.2 The presence of background GxS interactions explains the advantage of the sex-differentiated meta-analysis approach

To explain this phenomenon, we make the note that sex specific effects is present throughout the genome and we refer to their effect on the phenotype as the gene-by-sex background effect. We hypothesized that the presence or absence of gene-by-sex interactions throughout the genome may lend an advantage to one method over another.

To test this hypothesis, we generate simulated phenotype data using the Finland Birth Cohort as a model population. For each simulated phenotype, we assume a given heritability \( h_g \) and a given GxS heritability \( h_s \) and generate a phenotype by sampling from a multivariate normal distribution with mean zero and covariance equal to \( h_g K + h_s K_s + h_f I \), where \( K \) represents the kinship matrix for males and females and \( K_s \) represents the GxS kinship matrix (see Methods) and \( h_f \) can be computed by \( 1 - h_g - h_s \). We insert an effect for a given SNP and then attempt to identify this effect as significant using each method. Power is calculated as the proportion of the number of effects identified as significant for a large number of such simulations. We calculate
<table>
<thead>
<tr>
<th>Position</th>
<th>rsid</th>
<th>BE</th>
<th>sex-specific</th>
<th>beta ± std error</th>
<th>best</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chr1:157945440</td>
<td>rs2794520</td>
<td>1.60e-13</td>
<td>3.01e-13</td>
<td>3.01e-13</td>
<td>Chr1:157945440</td>
</tr>
<tr>
<td>Chr12:119873345</td>
<td>rs26500000</td>
<td>1.06e-12</td>
<td>1.96e-12</td>
<td>1.96e-12</td>
<td>Chr12:119873345</td>
</tr>
<tr>
<td>Chr17:68999647</td>
<td>rs720617</td>
<td>0.0000011</td>
<td>8.15e-05</td>
<td>8.15e-05</td>
<td>Chr17:68999647</td>
</tr>
</tbody>
</table>

**Table 8.2:** Comparison of 4 association mapping approaches: Fixed effect meta-analysis (FE), random effect meta-analysis (RE2), sex-as-covariate (Covariate), combined approach (Combined)
power for a range of $h_g$ and $h_s$ values and we vary the level of heterogeneity in effect size between males and females.

We compare the statistical power between RE2, FE, and CV approaches while varying gene-by-sex background effect and heterogeneity between male and female effect size. We first fix the background noise heritability ($h_I$) to be 0.5 and vary GxS heritability ($h_s$) from 0 to 0.25. For each chosen GxS heritability ($h_s$), we set the $h_g$ to be $0.5 - h_s$. We also vary the effect size of main genetic effect between male and female, which is represented as “Heterogeneity” in the Figure 8.1. Figure 8.1 (a), (b), (c) shows power of proposed random effect meta-analysis approach, fixed effect meta-analysis approach and sex-as-covariate method respectively. In these heat maps, as the color become darker, the power is higher as the legends show. In Figure 8.1 (d), the region of heat map is colored with the color representing most powerful method in the region. This resulting heat map plot indicates that in most cases, except when the effect size in the male and female studies are equal or there is no background gene-by-sex effect, the RE2 approach is the most powerful. When there is neither a difference in effect sizes between males and females nor a gene-by-sex background effect, CV is most powerful and when there is no difference in effect sizes but there is a gene-by-sex background effect, FE is most powerful. One possible explanation for this observation is that the presence of many gene-by-sex interactions may increase the variance of the trait when both genders are analyzed together compared to if they are analyzed separately. This is consistent with our observation that the meta-analysis methods perform better than the combined methods when there is a large amount of gene-by-sex interactions.
Figure 8.1: Power of (a) random-effect meta-analysis (RE2), (b) fixed-effect meta-analysis (FE) and (c) sex-as-covariate methods (CV) as a function of the GxS heritability and the heterogeneity between the male and the female study effect size. (d) shows a comparison of the three methods with the color corresponding to the method with the highest power.

8.2.3 Statistical power comparison between FE, UFE, GWAMA, WGWAMA, SST and RE2 approaches for Sex-differentiated meta-analysis

As we have shown, meta-analysis approaches have advantages in statistical power in the presence of gene-by-sex interactions compared to traditional approaches. The
question remains which meta-analysis method is the best approach to use. Unfortu-
nately, there is no definite answer since which meta-analysis approach is most power-
ful depends on the level of heterogeneity resulting from the gene-by-sex interaction. We show this by analyzing the power of FE, RE2, GWAMA (a previously proposed meta-analysis approach) and an approach which analyzes males and females separately (which we refer to as SST (Sex-Specific Test) which declares a loci associated if any sex-specific p-value is significant).

We compare the statistical power between FE, GWAMA, SST, and RE2 approaches by both analytically estimating the power and performing simulation experiments. The statistic for FE, GWAMA, and SST approach have known analytical, closed form distributions. Thus we can easily compute the statistical power using a analytically derived formula as shown in section 8.4.1, 8.4.2, 8.4.3 and 8.4.4. Because the distribution of the statistics of RE2 does not have an analytical closed form formula, we generate the simulated data to estimate the statistical power of the method.

To assess the statistical power of the four methods, we first varies true effect sizes of male and female studies under the assumption that the number of samples in male and female studies are same. Figure 8.2 (A) shows the 50% power line of four methods (FE, SST, RE2, GWAMA). The 50% power line shows the male and female effect sizes where the method achieves exactly a power of 50%. Note that the power increases as the effect size increase. Thus the closer the 50% line is to the bottom leftmost point on the graph, the more powerful the method is. As expected, when one of the effect size is zero, the SST approach is the most powerful, One interesting observation in this power comparison experiment is that when the effect size are moderately similar between male and female study, FE and RE2 approach clearly outperform SST and GWAMA approach. Based on the graph, we feel that RE2 is a good balance between losing very little in terms of power when the effect sizes are close to each other compared to FE
while gaining a significant advantage over FE when the effect sizes are different.

We also perform a similar power comparison experiment when the number of individuals are different in the female and male study. Specifically, we assume that there are twice as many females as males \(N_{female} = 2N_{male}\). Figure 8.2 (B) shows the 50% power line of four methods (FE, SST, RE2, GWAMA) in the unbalanced study. The 50% power lines shows very similar patterns to the balanced study case. However when the female effect size is small and male effect size is large, there is a region that GWAMA outperform other methods. The reason for this phenomenon is that GWAMA ignores the unbalanced number of individuals between males and females. A major difference between GWAMA and the other meta-analysis methods is that GWAMA always weighs each study equally while the other studies take into account the differences in the variance of the effect size estimates which are typically driven by differences in sample sizes. It can be easily shown using the Cauchy-Schwartz inequality that the typical inverse variance weighting of the fixed-effect meta-analysis is optimal with respect to power. We extended GWAMA to allow for weighing each study and we set the weights in a similar fashion to optimize power under the same assumptions of the other meta-analysis approaches. We demonstrate our new method Weighted GWAMA (WGWAMA) by comparing to FE, RE2 and also an unweighted version of the fixed-effect meta-analysis (UFE) to gain intuitions on the effect of weights.

Figure 8.3 (A) shows the 50% power line of four methods (FE, UFE, GWAMA, WGWAMA, SST) under the balanced study assumption between male and female studies. As shown, unweighted FE (UFE) and weighted GWAMA (WGWAMA) shows same power as FE and GWAMA respectively, when the male and female studies are balanced. Figure 8.3 (B) shows the 50% power line of four methods (FE, UFE, GWAMA, WGWAMA, SST) in an unbalanced study where there are twice as many females as males. In this case, weighted GWAMA (WGWAMA) and FE are more
Figure 8.2: Statistical power comparison between FE, GWAMA, SST and RE2 approaches while varying the effect sizes of males and females study. (a) The power comparison under the assumption of balanced number of individuals (b) and when there are twice as many females as males. The diagonal line shows the points where the effect sizes between males and females are equal.

appropriate and natural approach in that it gives more weight to the larger studies.
Figure 8.3: Statistical power comparison between FE, UFE, WGWAMA, GWAMA, and SST approaches while varying the effect sizes of male and female study. (a) The power comparison under the assumption of balanced number of individuals (b) and when there are twice as many females as males. The diagonal line shows the points where the effect sizes between males and females are equal.
8.3 Discussion

We have presented Meta-sex, a novel powerful meta-analytic approach for the analysis of genome-wide association study containing the sex-specific effects. Meta-sex utilize the random effect based meta-analysis approach, which allows for the heterogeneity in the effect size between the male and the female studies leading to statistically powerful compared to the traditional approach, which considers the sex-as-covariates. Through simulations and analysis of the NFBC data, we have shown that our proposed approach achieves higher power than traditional approaches.

One significant challenge encountered in applying meta-analysis is the potential for between study population structure, which in this case corresponds to genetic relationships between males and females in the original study resulting in correlation between the effect size estimates which can lead to inflation and spurious associations. The traditional sex-as-covariate approach would apply a mixed model to the combined sample that would model these relationships through a kinship matrix estimated from the genotypes. Unfortunately, if we are analyzing the males and females separately, we are unable to correct for this between sex population structure. One simple way to address this issue is to compute genomic control lambda values [DR99], and check to see if there is any inflation. If there is, genomic control can be applies to the combined statistics. A more sophisticated approach would be to apply the method of Sul et al., (2013), which utilizes a mixed model and the kinship matrix between individuals in each study to correct for the correlation between effect size estimates directly when performing a meta-analysis. In our experiments, we did not observe any inflation of the combined statistics.
8.4 Methods

8.4.1 Fixed effects model meta analysis for sex-differentiated analysis

In standard fixed effect model meta analysis, we first estimate the genetic effect size of male and female study separately in the linear model.

\[ y_f = \mu_f + \beta_f X_f + e_f \]  \hspace{1cm} (8.1)
\[ y_m = \mu_m + \beta_m X_m + e_m \]  \hspace{1cm} (8.2)

Given the estimates of genetic effect size \((\beta_m, \beta_f)\) and their variance \((\sigma^2_m, \sigma^2_f)\), In the fixed effects model meta-analysis, we assume that the underlying effect sizes are the same in male and female study as \(\beta (\beta = \beta_m = \beta_f)\). The best estimate of \(\beta\) is the inverse-variance weighted effect size \([BFJ08]\),

\[ \bar{\beta} = \frac{w_m \beta_m + w_f \beta_f}{w_m + w_f}, \quad SE(\bar{\beta}) = \frac{1}{\sqrt{w_m + w_f}} \]  \hspace{1cm} (8.3)

\[ Z_{FE} = \frac{\bar{\beta}}{SE(\bar{\beta})} = \frac{w_m \beta_m + w_f \beta_f}{\sqrt{w_m + w_f}} \]  \hspace{1cm} (8.4)

where \(w_m = 1/\sigma^2_m\) and \(w_f = 1/\sigma^2_f\) so-called inverse variance. Then we can test the null hypothesis \(\beta = 0\) versus the alternative hypothesis \(\beta \neq 0\).

8.4.2 Unweighted fixed effects model meta analysis for sex-differentiated analysis

Unlike the standard fixed effect model meta analysis, unweighted fixed effects meta analysis assigns same weights to both the male and the female study regardless of the variance of the effect size estimates. Given the estimates of genetic effect size \((\beta_m, \beta_f)\) and their variance \((\sigma^2_m, \sigma^2_f)\) from Equation (8.1) and (8.2), We compute the z-scores
for each male and female study,
\[ Z_m = \frac{\beta_m}{\sigma_m^2}, \quad Z_f = \frac{\beta_f}{\sigma_f^2}. \]  
(8.5)

Then unweighted fixed effect z-score can be obtained by equally weighting \( Z_m \) and \( Z_f \).
\[ Z_{UFE} = \frac{wZ_m + wZ_f}{\sqrt{w^2 + w^2}} = \frac{Z_m + Z_f}{\sqrt{2}} \]  
(8.6)

The p-value of unweighed fixed effect model meta analysis can be easily computed, as \( Z_{UFE} \) follows standard normal distribution under the null hypothesis that \( \beta_m = 0 \) and \( \beta_f = 0 \).

### 8.4.3 SST approach for sex-differentiated analysis

The SST approach is similar to sex-specific association test. We first estimate the genetic effect size \((\beta_m, \beta_f)\) and its standard error \((\sigma_m^2, \sigma_f^2)\) in male and female study separately in the linear model as in equation (8.1). Then we compute the z-score statistic of male and female study separately by \( z_m = \frac{\beta_m}{\sigma_m} \) and \( z_f = \frac{\beta_f}{\sigma_f} \). Then the SST statistic is chosen by maximum between absolute value of z-scores \( S_{\text{max}} = \max(|z_m|, |z_f|) \). \( p\text{-value} \) can be obtained with \( S_{\text{max}} \) under the null hypothesis of standard normal distribution.

### 8.4.4 GWAMA approach for sex-differentiated analysis

In the GWAMA approach [MLM10], we first estimate the genetic effect size \((\beta_m, \beta_f)\) and its standard error \((\sigma_m^2, \sigma_f^2)\) in male and female study separately in the linear model as in equation (8.1). Then we compute the \( \chi^2 \) statistic of male and female study separately by squaring z-scores \( \chi_m^2 = z_m^2 = \left(\frac{\beta_m}{\sigma_m}\right)^2 \) and \( \chi_f^2 = z_f^2 = \left(\frac{\beta_f}{\sigma_f}\right)^2 \). GWAMA statistic can be obtained by summing male \( \chi^2 \) and female \( \chi^2 \) \( \left( S_{GWAMA} = \chi_m^2 + \chi_f^2 \right) \). \( p\text{-value} \) can be obtained with \( S_{GWAMA} \) under the null hypothesis of central \( \chi^2 \) distribu-
bution with 2 degrees of freedom.

### 8.4.5 Weighted GWAMA approach for sex-differentiated analysis

Unlike the traditional GWAMA approach, weighted GWAMA (WGWAMA) seeks the optimal weight with respect to the power, when combining two $\chi^2$ statistics from the male and female study. The relationship between WGWAMA and GWAMA is analogous to the that between FE and unweighted FE. Similar to FE, WGWAMA assumes the fixed effect at the causal loci in both the male and the female study. Under this assumption, WGWAMA performs numerical grid search for optimal weights to combine two $\chi^2$ statistics. For each weights, it first generates 2 million null $\chi^2$ statistics by squaring z-scores sampled from standard normal distribution. Subsequently we can obtain 1 million null statistics by combining two null $\chi^2$ statistics with given weights. Given significant level $\alpha$, $(\alpha \times 1 \text{ mil})$-th statistic is the significant threshold for the given weights. Then we generate 10000 alternative statistics. For this, we first generate 10000 alternative $\chi^2$ statistics sampled from normal distribution with mean $\sqrt{N_m}$ and $\sqrt{N_f}$, where $N_m$ and $N_f$ are the number of samples for the male study and the female study respectively. Then the power for the specific weights can be computed by the proportion of alternative statistics more significant than the threshold for the weights. The above power estimation is performed for wide range of weights then the most powerful weights can be chosen for the WGWMA approach.

### 8.4.6 Meta-sex approach for sex-differentiated analysis

In the Meta-sex approach, random effects model is assumed. In other words, we assume that the effect size of male and female studies $\beta_i (i = m, f)$ is a random variable
followed by a distribution with the grand mean $\beta$ and the variance $\tau^2$ [HE11b, DL86a],

$$\beta_i \sim N(\beta, \tau^2).$$

This random effects model tests the null hypothesis $\beta = 0$ and $\tau^2 = 0$ versus the alternative hypothesis $\beta \neq 0$ or $\tau^2 \neq 0$. Similarly to the traditional random effect model [DL86a], we use the likelihood ratio framework considering each statistic ($\beta_i$ and $\sigma_i$ pair) as a single observation. Here $\beta_i$ and $\sigma_i$ represents the effect size estimate and its standard error respectively. However to avoid the problem of conservative nature of traditional random effects model, we assume no heterogeneity under the null hypothesis [HE11b], which is natural because the effect size should be fixed to be zero under the null hypothesis The likelihoods are then

$$L_0 = \prod_{i=m,f} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp \left( -\frac{\beta_i^2}{2\sigma_i^2} \right),$$

$$L_1 = \prod_{i=m,f} \frac{1}{\sqrt{2\pi(\sigma_i^2 + \tau^2)}} \exp \left( -\frac{(\beta_i - \mu)^2}{2(\sigma_i^2 + \tau^2)} \right).$$

The maximum likelihood estimates $\hat{\mu}$ and $\hat{\tau}^2$ can be found by an iterative procedure suggested by Hardy and Thompson [HT96]. Then the likelihood ratio test statistic can be built

$$S_{meta} = -2 \log(\lambda) = \sum \log \left( \frac{\sigma_i^2}{\sigma_i^2 + \hat{\tau}^2} \right) + \sum \frac{\beta_i^2}{\sigma_i^2} - \sum \frac{(\beta_i - \hat{\mu})^2}{\sigma_i^2 + \hat{\tau}^2}. \quad (8.7)$$

$p$-value can be calculated efficiently using pre-computed tabulated values [HE11b].

### 8.4.7 Controlling for population structure within studies

Genetic similarities between individuals, called the population structure or cryptic relatedness, is well-known for both inhibiting the ability to find true associations and causing the appearance of a large number of false or spurious associations [DRB01, VP05]. One popular method to correct this problem is employing mixed models[Lan02,
YPB06, KZW08, KSS10, LQK13, LLK13]. These mixed model approaches introduce a random variable into the traditional linear model.

\[ Y_i = \mu + \beta_i X_i + u_i + \epsilon \quad (i = m, f) \]  

(8.8)

In the model in equation (8.8), the random variable \( u_i \) represents the vector of genetic contributions to the phenotype for individuals in study \( i \). This random variable is assumed to follow a normal distribution with \( u_i \sim N(0, \sigma^2 K_i) \), where \( K_i \) is the \( n_i \times n_i \) kinship coefficient matrix for study \( i \). With this assumption, the total variance of \( Y_i \) is given by \( \Sigma_i = \sigma^2 K_i + \sigma^2_e I \). A z-score statistic is derived for the test \( \beta_i = 0 \) by noting the distribution of the estimate of \( \hat{\beta}_i \). In order to avoid complicated notation, we introduce a more basic matrix form of the model in equation (8.8), shown in equation (8.9).

\[ Y_i = S_i \Gamma + u_i + \epsilon \]  

(8.9)

In equation (8.9), \( S_i \) is a \( n_i \times 2 \) matrix encoding the global mean and SNP vectors and \( \Gamma \) is a \( 2 \times 1 \) coefficient vector. The maximum likelihood estimate for \( \Gamma \) in study \( i \) is given by \( \hat{\Gamma}_i = (S_i' \Sigma_i^{-1} S_i)^{-1} S_i' \Sigma_i^{-1} Y_i \) which follows a normal distribution with a mean equal to the true \( \Gamma \) and variance \( (S_i' \Sigma_i^{-1} S_i)^{-1} \). The estimates of the effect size \( \beta_i \) and standard error of the \( \beta_i \) (\( \sigma_i \)) are then given in equation (8.10) and equation (8.11), where \( R = [0 \ 1] \) is a vector used to select the appropriate entry in the vector \( \hat{\Gamma}_i \).

\[ \beta_i = R(S_i' \Sigma_i^{-1} S_i)^{-1} S_i' \Sigma_i^{-1} Y_i \]  

(8.10)

\[ \sigma_i = [R(S_i' \Sigma_i^{-1} S_i)^{-1} R']^{1/2} \]  

(8.11)

The above mixed model based approach can be utilized in conjunction with any of above sex-differentiated methods such as fixed effects meta analysis, SST approach,
GWAMA approach, Meta-Sex approach to control the population structure in each male and female study.

8.4.8 Computation of statistical power of the methods

8.4.8.1 Analytical power computation for FE, SST and GWAMA methods

Let $\alpha$ be the $p$-value threshold, which will control the false positive rate of sex-differentiated association testing. For the analytical power computation of fixed effect meta analysis approach, let $\Phi(x; \mu, \sigma)$ be the CDF of normal distribution for statistic $x$ given mean $\mu$ and standard deviation $\sigma$ of the distribution. Since FE test is two-sided test, the threshold statistic can be obtained by inverse CDF of $\frac{\alpha}{2}$ ($T_\alpha = \Phi^{-1}(\frac{\alpha}{2}; 0, 1)$). Given effect size estimates $(\beta_m, \beta_f)$ and their standard errors $(\sigma_m, \sigma_f)$ from male and female study, the non-centrality parameter can be easily obtained by equation (8.4) $S_{FE} = \frac{w_f \beta_f + w_m \beta_m}{\sqrt{w_f + w_m}}$. Given $T_\alpha$ and $S_{FE}$ we derive the formula for analytical power of fixed effect model meta analysis.

$$Power_{FE} = \Phi(T_\alpha; S_{FE}, 1) + 1 - \Phi(-T_\alpha; S_{FE}, 1) \quad (8.12)$$

The statistical power for SST approach is the probability that the hypothesis testings for both male and female study are not rejected given $z_m$ and $z_f$, $T_\alpha$. This quantity can be analytically computed by the following equation.

$$Power_{Max} = 1 - \left[ \left( 1 - \Phi(T_\alpha; z_m, 1) - (1 - \Phi(-T_\alpha, z_m, 1)) \right) \times \right.$$

$$\left. \left( 1 - \Phi(T_\alpha; z_f, 1) - (1 - \Phi(-T_\alpha, z_f, 1)) \right) \right]$$

225
To derive the statistical power formula for GWAMA approach, we first define $F_{\chi^2}(x; m, k)$ as CDF of $\chi^2$ distribution with non-centrality parameter $m$ and $k$ degree of freedom. Since GWAMA statistic follows central $\chi^2$ distribution with 2 degree of freedom under the null hypothesis, given the $p$-value threshold $\alpha$ for controlling false positive rate, the threshold statistic for GWAMA approach can be expressed as $T_\alpha = F^{-1}_{\chi^2}(1 - \alpha; 0, 2)$. The non-centrality parameter for GWAMA approach can be computed by $S_{GWAMA} = z_m^2 + z_f^2$, as explained in the section 8.4.4. Then the statistical power of GWAMA approach can be computed analytically with the following formula:

$$Power_{GWAMA} = F_{\chi^2}(T_\alpha; S_{GWAMA}, 2)$$

(8.13)

### 8.4.8.2 Empirical power estimation for Meta-Sex methods

Unlike FE, GWAMA, SST methods, the analytical distribution of the Meta-Sex statistic is unknown. Thus to estimate the statistical power of the Meta-Sex approach, we employ simulation approach. Given true effect sizes and standard errors of male and female, we sample $N$ number of male and female effect sizes. $N$ number of $p$-values are computed as explained in section 8.4.6. Then the statistical power is computed as the proportion of $p$-values that are more significant than predefined genome-wide significant threshold such as $5.0 \times 10^{-7}$.

### 8.4.9 Simulating GxS interactions

In order to simulate GxS interactions, we sample from the appropriate multivariate normal distribution. We are given the kinship matrices $K_m$, $K_f$ and $K_a$, for males, female and all individuals, respectively. The sex-specific kinship matrix $K_s$ is then given by the following.
Given a heritability $h$ and a GxS heritability $h_s$, a simulated phenotype is generated by sampling from a multivariate normal distribution with mean zero and variance $hK_a + h_sK_s + (1 - h - h_s)I$. The intuition is that shared genetics between individuals accounts on average for some proportion of the overall phenotypic variance ($h$), while the proportion of variance accounted for by genetics for only males or only females is $h + h_s$. That is, same sex individuals have a similar genetic contribution to their overall trait value and this contribution is different when comparing across sexes.

### 8.4.10 Computing heritability of GxS interactions

We assume an additive model for the genetic and GxS interaction effects and estimate the phenotypic variation due to these effects using a linear mixed model. Let $N$ be the number of individuals and $M$ be the number of genotypes collected in the study.

$$y = X\beta + g + gs + e,$$

where $y$ is the $N \times 1$ vector of the observed phenotypic values from the individuals. $X$ is the $N \times K$ matrix of covariates, including the intercept and the sex, and $\beta$ is the $K \times 1$ vector of coefficients for the covariates. We treat $g, gs, e$ as random effects. Let $x$ be the $N \times M$ matrix representing SNP matrix and $x_i$ denotes the $N \times 1$ normalized genotype vector of SNP $i$ and the components of $x_i$ are encoded as \(\{2 - 2p_i/\sqrt{2p_i(1-p_i)}, 1 - 2p_i/\sqrt{2p_i(1-p_i)}, -p_i/\sqrt{2p_i(1-p_i)}\}\) for minor allele homozygous, heterozygous, and major allele homozygous ($p_i$ is the observed minor allele frequency). Then we assume $g \sim N(0, \sigma_g^2K_g)$, where $K_g = xx^T$. The second random effect variable $gs$ denotes the background gene-by-sex interaction effect and we
assume $g_s \sim N(0, \sigma_{gs}^2 K_{gs})$, in which $K_{gs}$ representing covariance structure for GxS. For the pairs of individuals with same sex, the covariance will be the covariance in $K_g$ but for the pairs of individuals with different sex, the covariance will be 0. Finally, $e$ denotes the $N \times 1$ vector of environmental effects and we assume $e \sim N(0, \sigma_e^2 I)$, in which $I$ is the identity matrix. After estimating $\sigma^2_g, \sigma^2_{gs}, \sigma^2_e$ by restricted maximum likelihood using GCTA software [YBM10, LWG11] Then the gene-by-sex (GxS) heritability can be computed by $h^2_{gs} = \frac{\sigma^2_{gs}}{\sigma^2_g + \sigma^2_{gs} + \sigma^2_e}$. 
8.5 Supplementary Materials

QQplots for NFBC data
Figure S1: QQplots for BMI phenotype
Figure S2: QQplots for Crp phenotype
Figure S3: QQplots for Dia phenotype
Figure S4: QQplots for Glu phenotype
Figure S5: QQplots for Hdl phenotype
Figure S6: QQplots for height phenotype
Figure S7: QQplots for ins phenotype
Figure S8: QQplots for ldl phenotype
Figure S9: QQplots for sys phenotype
Figure S10: QQplots for tg phenotype
CHAPTER 9

Conclusion

In this thesis, I presented several computational genetic approaches for analyzing genetic data, including that from humans as well as model organisms such as mice. In Chapter 2, I introduced a method to infer biological networks from high-throughput data including both genetic variation and gene expression profiles. The key insight employed in the proposed method is utilizing the prior knowledge that genetic variation affects gene expression, but gene expression cannot affect genetic variation. This knowledge allows us to force the direction of the causal relationship between genetic variation and gene expression, which in turn allows us to determine the causal directionality between two gene expressions, using the basic properties of causality theory. Moreover, we can also rule out certain directed causal relationships between gene expressions using the same principals, thus identifying the absence of certain causal relationships. The proposed method combines a principled representation of causality using graphical causal models with small sample statistical methods to infer the presence and absence of causal relationships between yeast genes. To overcome the problem of limited sample size, we performed a likelihood ratio test instead of performing a conditional independence test directly. The advantage of our approach is the ability to distinguish between direct and indirect variations, and minimizing false positives by inferring the absence of causal relationships. These results provide a rich description of the yeast gene regulation network beyond previous results from mapping studies and co-expression network analysis. The interesting extension to this approach
is developing advanced method which can be utilized in mouse model organism. This extension requires incorporating random effects in the model to correct for batch effect and population structure which exists in the mouse model organism data.

Next I presented an efficient dynamic programming algorithm to compute the marginalized posterior probability of causal graphical features, such as adjacencies and v-structures. The goal of inferring causal relationships from observational data is to identify the exact data generating model. However inferring causal graphical features has a fundamental limitation if we only use observational data. That is, the best we can infer with observational data is the Markov equivalence class of the data generating model. Previous approaches to this problem have computed the marginalized posterior probability of directed edges or adjacencies between the variables of interest. However, the posterior probability of directed edges in the graphical model is dependent on the shape of the true data generating model, which consequently decides the Markov equivalence class. As proven in Chapter 3, the scale of the posterior probability of directed edge features will not be the same across all directed graphical edge features. One problem of this phenomenon is that it becomes difficult to decide on a single threshold for posterior probability to claim the existence of the directed edge features. The solution we proposed is computing the posterior probability of graphical features defining the Markov equivalence class, such as adjacencies and v-structures, instead of considering directed edges. Based on this theoretical justification, we devised the dynamic programming algorithm to efficiently compute the posterior probability of the existence of adjacencies and v-structures.

I also introduced meta-analysis to combine multiple genome-wide association studies, where each study contained a population structure. Meta-analysis is a popular tool to combine multiple studies to achieve a higher statistical power. In addition, we can also combine studies at the summary statistic level, making it easier for data to be
shared when there is a limitation in the data sharing process. The method of combining multiple studies at the summary statistic level for optimal statistical power when each study does not contain population structure is very well known. In Chapter 4, a solution to the problem of performing meta-analysis was proposed, while each study contains a different. Population structure is well known to cause spurious association signals and false positives in genome-wide association studies, especially those involved in model organism. In our example of two different meta-analysis studies (mouse HDL and BMD), we showed the utility of the proposed approach. Interesting observation of the application of the proposed meta-analysis approach to these mouse studies is that we demonstrated increased resolution as well as increased statistical power in the meta-analysis results compared to any single study.

Identity by descent (IBD) is the foundation for many of the important problems in genetics, including the determination of the haplotype phase, understanding familial diseases, detecting population structure and heritability estimation. Segmental IBD sharing indicates that the common ancestor is relatively recent. Two individuals share a haplotype segment identical by descent when the haplotype is inherited from a recent common ancestor without recombination. If the segmental IBD haplotype contains the disease causing mutation, then the individuals who share this particular IBD segment are likely to share the disease as well. The IBD mapping is a method searching for the IBD segments containing the disease causing mutation. The major bottleneck of the traditional approach to this problem is that it requires permutation to estimate the significance of the association to pass the stringent genome-wide threshold ($1.0 \times 10^{-7}$), which takes years to compute. In Chapter 5, I introduced the importance sampling approach to estimate the significance efficiently, which reduces computation time from years to days. Another important benefit of the proposed method is that we can perform fine mapping, which helps to pinpoint the candidate causal gene. We demonstrated the applicability of our approach to the fine mapping problem with WTCCC type 1 dia-
Identifying environmentally specific genetic effects is a key challenge in understanding the structure of complex traits. The traditional approach of identifying gene-by-environment interactions utilizes the linear model and includes the interaction effect as a covariate in the model. In this linear model, a Wald test based on a hypothesis testing framework is performed to determine whether there is a non-zero coefficient for the gene-by-environment interaction variable. To discover gene-by-environment interactions using this traditional framework requires several challenges to be overcome. First, a much larger number of samples are required to identify gene-by-environment interactions, to achieve similar statistical powers to simple GWA studies, because more degrees of freedom for environmental variables is required in the model. Second, identifying gene-by-environment interaction with the traditional approach requires prior knowledge of environmental variables. Specifically, before setting up the linear model for gene-by-environment interaction, we must know what kind of variable (e.g. sex, age, gene knockout) to be considered in the model, and how we encode these variables in the model (for example, binary or continuous values). In many cases, it is difficult to predict the gene-by-environment interaction we are expecting. Thus the kind of environments that should be incorporated in the model can be difficult to determine. In addition, the encoding of the environmental variables affects the statistical power of the study. In Chapter 6, I proposed a meta-analytic approach to identify gene-by-environment interaction, which can address these challenges. The proposed approach utilizes the random effect model of meta-analysis to combine studies with varying environmental conditions to identify gene-by-environment interactions. The major advantage of the proposed approach over the traditional approach is that it does not require explicit modeling of the environmental variables in the model, since we combine the studies with varying environments with meta analytic approach. In addition, by combining studies with a meta-analytical approach, we can easily combine
multiple studies at the summary statistic level, which allow us to perform large scale meta-analysis to achieve higher statistical powers. As demonstrated in the mouse HDL study, the method has exploratory aspects, meaning that it can discover the unexpected gene-by-environment interaction. Our proposed visualization framework allows the researcher to gain biological insights on gene-by-environment interactions.

In Chapter 7, I also analyzed age-related hearing loss data from mice, using a meta-analytic approach. Only a limited number of large-scale association studies for ARHI have been undertaken in humans, to date. In human studies it is difficult to control the environments affecting the ARHI, which might lead to noise in the genome-wide association study. An alternate and complementary approach to these human studies is the use of mouse models. Advantages of mouse models include more control over the environment, the ability to replicate measurements in genetically identical animals, and an increase in the proportion of the variability explained by genetic variation. Motivated by these advantages, the presented study in Chapter 7 is the first genome-wide association study of its kind by combining several data sets in a meta-analysis to identify loci associated with age-related hearing loss in mice. We identified 5 genome-wide significant loci, one of which has a peak located in the previously implicated region. This application study suggests that we can achieve high resolution and power in genome-wide association study by combining multiple independent studies.

Lastly, I introduced the powerful random effect-based meta-analytic approach for genome-wide association studies with sex-specific effects. Many genome-wide association studies in humans and model organisms show sex-specific effects. The prevalence of sex-specific effects complicates the analysis of association studies consisting of both males and females. The question that I addressed in Chapter 8 was how to achieve optimal power in such genome-wide association studies. In Chapter 8, I compared many different meta-analytic approaches and traditional covariate approaches.
to this problem of GWAS with sex-specific effects. Our extensive simulation experiments show the behavior of the different approaches in various scenarios, including heterogeneity between male and female effect sizes, various numbers of sample sizes between male and female studies, and the existence of various background gene-by-sex interaction cases. We proposed random effect model meta-analysis which analyzes male and female studies separately, and combine the summary statistics with the meta-analytic approach. We applied our method to the Northern Finland Birth Cohort data and demonstrated increased power over the traditional approach, while controlling for false positives.
REFERENCES


tin Dichgans, Maija Wessman, George Davey Smith, Kari Stefansson, Mark J. Daly, Dale R. Nyholt, Daniel I. Chasman, and Aarno Palotie. “Genome-wide meta-analysis identifies new susceptibility loci for mi-
graine.” Nat Genet, 6 2013.


[BFO10b] Brian J. Bennett, Charles R. Farber, Luz Orozco, Hyun Min Kang, Anatole Ghazalpour, Nathan Siemers, Michael Neubauer, Isaac Neuhaus,


Richard C. Davis, Atila van Nas, Brian Bennett, Luz Orozco, Calvin Pan, Christoph D. Rau, Eleazar Eskin, and Aldons J. Lusis. “Genome-wide association mapping of blood cell traits in mice.” *Mamm Genome, 24*(3-4):105–18, 4 2013.


Ai Kubo, Michael Blaise Cook, Nicholas J. Shaheen, Thomas L. Vaughan, David C. Whiteman, Liam Murray, and Douglas A. Corley. “Sex-specific associations between body mass index, waist circumference and the risk of Barrett’s oesophagus: a pooled analysis from the international BEACON consortium.” *Gut, 1* 2013.


association mapping needs linkage analysis to avoid false-positive loci.”


263


[SLY05] Eric E Schadt, John Lamb, Xia Yang, Jun Zhu, Steve Edwards, Debraj GuhaThakurta, Solveig K Sieberts, Stephanie Monks, Marc Reitman, Chunsheng Zhang, Pek Yee Lum, Amy Leonardson, Rolf Thieringer,


[SWC10] Kristin I. Stanford, Lianchun Wang, Jan Castagnola, Danyin Song, Joseph R. Bishop, Jillian R. Brown, Roger Lawrence, Xaiomei Bai, Hiroko Habuchi, Masakazu Tanaka, Wellington V. Cardoso, Koji Kimata,


