Title
Computational Genetic Approaches for the Dissection of Complex Traits

Author
Furlotte, Nicholas A.

Publication Date
2013

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Computational Genetic Approaches for the Dissection of Complex Traits

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Computer Science

by

Nicholas A. Furlotte

2013

ABSTRACT OF THE DISSERTATION

Computational Genetic Approaches for the Dissection of Complex Traits

by

Nicholas A. Furlotte

Doctor of Philosophy in Computer Science

University of California, Los Angeles, 2013

Professor Eleazar Eskin, Chair

Over the past two decades, major technological innovations have transformed the field of genetics allowing researchers to examine the relationship between genetic and phenotypic variation at an unprecedented level of granularity. As a result, genetics has increasingly become a data-driven science, demanding effective statistical procedures and efficient computational methods and necessitating a new interface that some refer to as computational genetics. In this dissertation, I focus on a few problems existing within this interface. First, I introduce a method for calculating gene coexpression in a way that is robust to statistical confounding introduced through expression heterogeneity. Heterogeneity in experimental conditions causes separate microarrays to be more correlated than expected by chance. This additional correlation between arrays induces correlation between gene expression measurements, in effect causing spurious gene coexpression. By formulating the problem of calculating coexpression in a linear mixed-model framework, I show how it is possible to account for the correlation between microarrays and produce coexpression values that are robust to expression heterogeneity. Second, I introduce a meta-analysis technique that allows for genome-wide association studies to be combined across populations that are known to

contain population structure. This development was motivated by a specific problem in mouse genetics, the aim of which is to utilize multiple mouse association studies jointly. I show that by combining the studies using meta-analysis, while accounting for population structure, the proposed method achieves increased statistical power and increased association resolution. Next, I will introduce a computational and statistical procedure for performing genome-wide association using longitudinal measurements. I show that by accounting for the genetic and environmental correlation between measurements originating from the same individual, it is possible to increase association power. Finally, I will introduce a statistical and computational construct called the matrix-variate linear mixed-model (mvLMM), which is used for multiple phenotype genome-wide association. I show how the application of this method results in increased association power over single trait mapping and leads to a dramatic reduction in computational time over classical multiple phenotype optimization procedures. For example, where a classically-based approach takes hours to perform parameter optimization for moderate sample sizes mvLMM takes minutes. This technique is both a generalization and improvement on the previously proposed longitudinal analysis technique and its innovation has the potential to impact many current problems in the field of computational genetics.

The dissertation of Nicholas A. Furlotte is approved.

Amit Sahai

Christopher J. Lee

David Earl Heckerman

Aldons J. Lusis

Eleazar Eskin, Committee Chair

University of California, Los Angeles

2013

This dissertation is dedicated to my grandparents. Their love, kindness and dedication has been invaluable to me

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to first give thanks to my family. They have given me an amazing amount of support without which I would have never completed this degree and this dissertation. I would like to thank my adviser Eleazar Eskin for his guidance over the last five years. My committee has also been very helpful and I am grateful to have had their guidance and advice. In particular, I would like to thank Jake Lusis and David Heckerman for their mentorship. Lastly, I want to thank my amazing friends who have supported me through my PhD and other important stages of my life. Thank you Yuki, Ples, Peter, Doug and my fiance Luz.

VITA

1981	Born in Memphis Tennessee
1996-2000	Attended White Station High School in Memphis Tennessee
2001-2005	B.Sc., Computer Science, University of Memphis, Memphis, Tennessee.
2007-2008	M.Sc., Bioinformatics, University of Memphis, Memphis, Tennessee.
2008	Research Assistant and Software Developer, University of Tennessee Health Science Center, Memphis, Tennessee
2009	Teaching Assistant, University of California Los Angeles
2010-2012	Computational Biology and Mathematics Tutor, University of California Los Angeles
2012	Research Intern, Microsoft Research, Los Angeles, California

PUBLICATIONS

Anatole Ghazalpour, Christoph D Rau, Charles R Farber, Brian J Bennett, Luz D Orozco, Atila van Nas, Calvin Pan, Hooman Allayee, Simon W Beaven, Mete Civelek, Richard C Davis, Thomas A Drake, Rick A Friedman, Nick Furlotte, Simon T Hui, J David Jentsch, Emrah Kostem, Hyun Min Kang, Eun Yong Kang, Jong Wha Joo,

Vyacheslav A Korshunov, Rick E Laughlin, Lisa J Martin, Jeffrey D Ohmen, Brian W Parks, Matteo Pellegrini, Karen Reue, Desmond J Smith, Sotirios Tetradis, Jessica Wang, Yibin Wang, James N Weiss, Todd Kirchgessner, Peter S Gargalovic, Eleazar Eskin, Aldons J Lusis, Rene C Leboeuf, "Hybrid mouse diversity panel: a panel of inbred mouse strains suitable for analysis of complex genetic traits," in *Mammalian Genome*, October 2012.

Nicholas A. Furlotte, Eleazar Eskin and Susana Eyheramendy, "Genome-wide association mapping with longitudinal data," in *Genetic Epidemiology* , May 2012.

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.

Thomas M. Keane, Leo Goodstadt, Petr Danecek, Michael A. White, Kim Wong, Binnaz Yalcin, Andreas Heger, Avigail Agam, Guy Slater, Martin Goodson, Nicholas A. Furlotte, Eleazar Eskin, Christoffer Nellaker, Helen Whitley, James Cleak, Deborah Janowitz, Polinka Hernandez-Pliego, Andrew Edwards, T. Grant Belgard, Peter L. Oliver, Rebecca E. McIntyre, Amarjit Bhomra, Jerome Nicod, Xiangchao Gan, Wei Yuan, Louise van der Weyden, Charles A. Steward, Sendu Bala, Jim Stalker, Richard Mott, Richard Durbin, Ian J. Jackson, Anne Czechanski, Jose fonso Guerra-Assunao, Leah Rae Donahue, Laura G. Reinholdt, Bret A. Payseur, Chris P. Ponting, Ewan Birney, Jonathan Flint and David J. Adams, "Mouse genomic variation and its effect on phenotypes and gene regulation," in *Nature*, September 2011.

Nicholas A. Furlotte, Hyun Min Kang, Chun Ye and Eleazar Eskin, "Mixed Model Coexpression: calculating gene coexpression while accounting for expression heterogeneity," in *Bioinformatics*, July 2011.

Anatole Ghazalpour, Brian Bennett, Vladislav A. Petyuk, Luz Orozco, Raffi Hagopian, Imran N. Mungrue, Charles R. Farber,Janet Sinsheimer, Hyun M. Kang, Nicholas Furlotte, Christopher C. Park, Ping-Zi Wen, Heather Brewer, Karl Weitz, David G. Camp II, Calvin Pan,Roumyana Yordanova, Isaac Neuhaus, Charles Tilford, Nathan Siemers, Peter Gargalovic, Eleazar Eskin, Todd Kirchgessner, Desmond J. Smith,Richard D. Smith, Aldons J. Lusis, "Comparative analysis of proteome and transcriptome variation in mouse," in *PLoS Genetics*, June 2011.

Dan He, Farhad Hormozdiari, Nick Furlotte and Eleazar Eskin, "Efficient Algorithms for Tandem Copy Number Variation Reconstruction in Repeat-rich Regions," in *Bioinformatics*, June 2011.

Daria Van Tyne, Daniel J. Park, Stephen F. Schaffner, Daniel E. Neafsey, Elaine Angelino, Joseph F. Cortese, Kayla G. Barnes, David M. Rosen, Amanda K. Lukens, Rachel F. Daniels, Danny A. Milner, Charles A. Johnson, Ilya Shlyakhter, Sharon R. Grossman, Justin S. Becker, Daniel Yamins, Elinor K. Karlsson, Daouda Ndiaye, Ousmane Sarr, Souleymane Mboup, Christian Happi, Nicholas A. Furlotte, Eleazar Eskin, Hyun Min Kang, Daniel L. Hartl, Bruce W. Birren, Roger C. Wiegand, Eric S. Lander, Dyann F. Wirth, Sarah K. Volkman and Pardis C. Sabeti, "Identification and Functional Validation of the Novel Antimalarial Resistance Locus PF10-0355 in Plasmodium falciparum," in *PLoS Genetics*, April 2011.

Charles R. Farber, Brian J. Bennett, Luz Orozco, Wei Zou, Ana Lira, Emrah Kostem, Hyun Min Kang, Nicholas Furlotte, Ani Berberyan, Anatole Ghazalpour, Jaijam Suwanwela, Thomas A. Drake, Eleazar Eskin, Q. Tian Wang, Steven L. Teitelbaum, Aldons J. Lusis, "Mouse Genome- Wide Association and Systems Genetics Identify Asxl2 As a Regulator of Bone Mineral Density and Osteoclastogenesis," in *PLoS Genetics*, April 2011.

Lijing Xu, Nicholas Furlotte, Yunyue Lin, Kevin Heinrich, Michael W. Berry, Ebenezer O. George, Ramin Homayouni, "Functional Cohesion of Gene Sets Determined by Latent Semantic Indexing of Pubmed Abstracts," in *PLoS One*, April 2011.

Dan He, Nick Furlotte and Eleazar Eskin, "Detection and Reconstruction of Tandemly Organized de novo Copy Number Variations", in *BMC Bioinformatics*, December 2010.

Andrew Kirby, Hyun Min Kang, Claire M. Wade, Chris J. Cotsapas, Emrah Kostem, Buhm Han, Nick Furlotte, Eun Yong Kang, Manuel Rivas, Molly A. Bogue, Kelly A. Frazer, Frank M. Johnson, Erica J. Beliharz, David R. Cox, Eleazar Eskin, Mark J. Daly, "Fine Mapping in 94 Inbred Mouse Strains Using a High-density Haplotype Resource', in *Genetics*, July 2010.

Brian J. Bennett, Charles R. Farber, Luz Orozco, Hyun Min Kang, Anatole Ghazalpour, Nathan Siemers, Michael Neubauer, Isaac Neuhaus, Roumyana Yordanova, Bo Guan, Amy Truong, Wen-pin Yang, Aiqing He, Paul Kayne, Peter Gargalovic, Todd Kirchgessner, Calvin Pan, Lawrence W. Castellani, Emrah Kostem, Nicholas Furlotte, Thomas A. Drake, Eleazar Eskin, and Aldons J. Lusis, "A high-resolution association

mapping panel for the dissection of complex traits in mice", in *Genome Research*, January 2010.

Teeradache Viangteeravat, Ian M. Brooks, Joseph Ketcherside, Ramin Houmayouni, Nicholas Furlotte, Somchan Vuthipadadon, Chanchai S. McDonald, "C. Biomedical Informatics Unit (BMIU): Slim-Prim system bridges the gap between laboratory discovery and practice", in *Clinical and Translational Science*, June 2009.

Teeradache Viangteeravat, Ian M Brooks, Ebony J Smith, Nicholas Furlotte, Somchan Vuthipadadon, Rebecca Reynolds, and Chanchai Singhanayok McDonald, "Slim-Prim: A biomedical informatics database to promote translational research", in *Perspectives in Health Information Management*, May 2009.

Vinthuy Phan and Nicholas A. Furlotte, "Motif Tool Manager: a web-based framework for motif discovery", in *Bioinformatics*, December 2008.

CHAPTER 1

Introduction

Over the past two decades, major technological innovations have transformed the field of genetics allowing researchers to examine the relationship between genetic and phenotypic variation at an unprecedented level of granularity. As a result, genetics has increasingly become a data-driven science, demanding effective statistical procedures and efficient computational methods and necessitating a new interface that some refer to as computational genetics. In this dissertation, I focus on a few problems existing within this interface each related to the dissection and detailed understanding of the relationship between genetic variation and complex traits.

The origin of genetics research is often attributed to the discoveries of Gregor Mendel, due to his insightful analysis of inheritance patterns in pea plants and the formulation of his three laws of inheritance. Although, his discoveries were published in 1866 they were not truly appreciated until 1900 [Stu65]. During the period from 1866 to 1900 there were many theories floating within the academic community explaining how traits were passed from generation to generation. These theories were motivated by many great experimental accomplishments, including the development of an understanding of mitosis and meiosis and the supposition that chromosomes were the bearer of the material of heredity. However, during this period of time there was a sore lack of quantitative research and many theories of heredity were based on the Darwinian principles of natural selection and supported by unfounded assumptions [Stu65]. Disappointed by this state, scientists such as Francis Galton and William Bateson began

to promote what we might now call a Mendelian approach based on a more systematic statistical analysis of trait variation [Stu65]. The promotion of this approach along with the re-discovery of Mendel's 1866 paper, led to a dramatic shift in the way that researchers thought about heredity and set genetics on the path to become what we know of today.

The fundamental shift in the study of heredity from ad-hoc theories based on what some might call anecdotal evidence to a more rigorous statistical framework was the turning point that ushered in a completely different way of thinking about trait variation between individuals. Similarly, the next fundamental shift in genetics is happening now. Enabled by innovative technologies such as the microarray and Sanger DNA and RNA sequencing and the children of these more developed technologies, current researchers are able to examine the nature of natural variation with an amazing level of granularity. These technologies have brought about a new way of conducting genetics research based on high-throughput experimental assays and the collection of extremely large datasets. Whereas early genetics researchers attempted to understand natural variation using what we think of as basic statistics using tens of individuals, current researchers are examining data sets with hundreds of thousands and soon even millions of individuals using advanced statistical and computational procedures. This more recent shift can be seen as a shift from small data to big data science.

Growing genetics from a low-throughput small data science to a high-throughput big data science requires innovative statistical and computational thinking to deal with problems driven by both the volume of experimental data as well as the complex inter-relatedness of data derived from high-throughput experimental assays. My thesis work focuses on four problems that are aligned with this theme. In what follows, I will give a brief background of each problem and provide an intuitive explanation of the contribution. These explanations assume a general knowledge of the field of genetics

including a familiarity with the concepts of genome-wide association studies and the idea of measuring the expression of genes.

Chapter 2: Mixed-model Coexpression

The coexpression between two genes is quantified in order to determine how similarly the genes are expressed. For example, if we measure gene expression using a microarray, we may find that two genes have highly correlated expression patterns, so that when one gene is highly expressed the other gene is highly expressed. In this case, these two genes are said to be coexpressed. Measures of coexpression are often used as a way to gauge how related two genes might be and to infer the presence of a functional relationship between genes and the analysis of gene coexpression is a fairly well established form of genetic analysis. However, the frequent use of the correlation coefficient to quantify coexpression has the potential to result in a large number of spurious or false coexpressions due to a problem known as expression heterogeneity, a phenomenon that arises when expression values between separate microarray experiments are more correlated than expected by chance due to the presence of shared confounding factors. In Chapter 2, I introduce a method for calculating gene coexpression in such a way that is robust to EH. The basic intuition behind the method is that EH may be quantified through the observed correlation between arrays and this estimate of EH can be used to adjust coexpression values.

Chapter 3: Meta-analysis for Structured Populations

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 associ-

ation 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.

Chapter 4: Genome-wide Association Mapping with Longitudinal Data.

To date, most genome-wide association methodologies aim to identify genetic variations that are associated with a single measurement of a complex trait. However, most quantitative complex traits change over time as a result of natural and environmental variation. Therefore, it is reasonable to assume that a single measurement may not best represent the state of a complex trait but that this state may be better represented by considering multiple measurements taken over time. In Chapter 4, I introduce a method for evaluating the association between genetic variations and a single complex trait with multiple measurements taken over time. Considering that the correlation between separate time points is due to both the shared genetic component contributing to each time point and the shared environmental components, I introduce a statistical model for representing longitudinal data and show how it can be used for association analysis. By considering multiple time points it is possible to increase statistical power substantially.

Chapter 5: Efficient Multiple Trait Association Mapping with the Matrix-variate Linear Mixed-model.

Most methods for analyzing the relationship between genetic and phenotypic vari-

ation assume a very simplistic model for genetic systems in which one genetic variation is assumed to have an effect on a single trait. However, we know that genetic systems are extremely complex and that a more realistic model would incorporate multiple related phenotypes and multiple genetic variations. Therefore, conducting association analysis under a multiple phenotype multiple genetic variation model could intuitively increase information content and potentially statistical power to discover associated variations. Motivated by a classic paper by Henderson [HQ76], [KVS12] show how classical multiple phenotype artificial selection models incorporating both multiple phenotypes and multiple genetic variations can be utilized to perform association analysis. They show that by considering pairs of correlated phenotypes when performing association it is possible to increase statistical power. These results although encouraging are shadowed by the fact that such approaches utilizing traditional computational techniques for performing maximum-likelihood inference have a high computational complexity and do not scale well when association is performed over a large number of individuals. In Chapter 5, I introduce a mathematical and computational construct that I call the matrix-variate linear mixed-model (mvLMM) that is used to perform efficient multiple trait association mapping and results in a dramatic reduction in computational time when compared with traditional approaches. The method works by applying a simple linear transformation to the phenotype data so that a maximum-likelihood search procedure may be performed in time linear in the size of the data.

CHAPTER 2

Mixed-Model Coexpression

2.1 Background

The analysis of gene coexpression patterns has been of great interest in recent years due to the widespread availability of microarray datasets measuring thousands of genes. Gene coexpressions, evaluated by comparing the expression patterns of pairs of genes, have been utilized in order to identify loci responsible for regulating genes [LPD06, GDZ06], to evaluate the significance of known pathways [STM05] and to identify functionally related genes whose relationships have been conserved through evolution [SSK03]. Unfortunately, gene expression data can be largely affected by technical bias such as a batch effects or plate effects [JRL07]. Such non-biological effects have been shown to induce correlations between genes. For example, [BKB03] showed that spacial placement of microarray probes affected the correlation between gene expression patterns, causing genes to be more or less correlated depending on the proximity of their respective probes on the array. More generally, unobserved factors affecting gene expression have the potential to cause correlation between genes. When these factors are shared between gene expressions, they cause genes to have similar patterns of overall variation. Since these effects are not directly observed they are not incorporated into statistical models. The shared variation between genes is attributed to biological causes. This issue is referred to as expression heterogeneity and has been acknowledged as a general problem when analyzing expression datasets [LS08].

The detrimental effects of technical confounding on the results obtained from microarray analysis are well known. [QBK05], while examining the effects of stochastic dependence between arrays on the correlation of test statistics used in determining differential expression, noted that the correlation structure of arrays induced through non-biological effects can lead to spurious correlation between genes. They note that microarray normalization procedures mitigate such phenomenon, but are unable to completely negate them. In fact, the presence of spurious correlations is a general problem that arises when analyzing many types of noisy high dimensional biological datasets, and has been examined in many different contexts [CRW08]. In the context of gene coexpression, the cause of spurious correlations can be conceptualized by viewing a set of n microarrays measuring m genes as a $m \times n$ matrix. In this matrix, we expect that the microarrays represented by the columns are independent and that some of the rows, representing the genes will be correlated, indicating biological relationships. In the presence of technical confounding effects, such as batch effects, the columns will share characteristics that will cause the overall patterns of expression to be similar and thus the arrays will be statistically correlated. This increased correlation between columns induces correlation between rows, as it becomes more likely that two randomly selected rows will be correlated, given that the overall patterns of expression for each array are similar. In this way, the correlation between arrays, or inter-sample correlation, has the potential to induce correlations between genes.

Many methods have been developed that aim to remove confounding effects from gene expression data. For example, in the case of known batch effects, a method such as ComBat [JRL07] may be employed. ComBat [JRL07] uses an empirical Bayes approach to estimate parameters associated with batch and produces corrected gene expression data. This corrected expression data can then be used in subsequent analysis. Unfortunately, technical confounding such as a batch effect may not be easily observable. In this case, a method that is able to identify possible confounding effects

without prior information is of interest. For example, surrogate variable analysis (SVA) [LS07] is a method for correcting gene expression data in the absence of known confounding effects. In SVA, a set of surrogate variables are estimated and regressed out of the expression data. These surrogate variables represent the unknown confounding effects which cause expression heterogeneity. Expression heterogeneity is expected to be encoded by the inter-sample correlation matrix, which is the matrix representing the correlation between all pairs of arrays. Surrogate variables are estimated by iteratively weighting a subset of the principal components of this matrix. The SVA method is aimed at the general problem of correcting gene expression data and does not specifically target the problem of calculating pairwise gene correlations. Furthermore, SVA only utilizes the principal axes of the inter-sample correlation matrix in order to correct expression. We can reason that the full inter-sample correlation matrix contains more information than its principal components and therefore SVA is only utilizing a subset of the available correlation information in its correction procedure. When the patterns of confounding are complex, the estimated surrogate variables may not capture all of the structure encoded in the inter-sample correlation matrix and as a result the corrected expression data may contain residual correlation.

In this chapter, we propose a method for calculating pairwise gene correlations that utilizes the full inter-sample correlations matrix in order to correct for expression heterogeneity. Our proposed method, Mixed Model Coexpression (MMC), uses a linear mixed model framework in order to adjust gene expression values and calculate pairwise gene correlations. Linear mixed model frameworks have been successfully used in previous studies to remove confounding effects when performing eQTL analysis [KYE08]. The MMC procedure represents confounding as a random effect in a statistical model for coexpression. This approach allows us to more accurately calculate coexpression while removing the effects due to confounding. Unlike ComBat, our method does not require previous knowledge about the batch effects. MMC is also

able to calculate coexpression without assuming and estimating some finite number of confounding effects, such as with SVA. These two properties give our method the advantage of being able to represent a wide range of unknown effects.

A caveat to our approach, as well as other approaches that utilize the inter-sample correlation matrix as a surrogate for confounding, is the potential to remove true biological signal, as expression heterogeneity may be caused by true biological effects. Consider one transcription factor whose activity marks the beginning of many possibly unrelated pathways. When this transcription factor exhibits high activity the genes involved in the downstream pathways will appear to be highly differentially expressed. This high level of differential expression will cause the downstream genes to be statistically correlated. When this master regulator affects hundreds or even thousands of downstream genes, the global patterns of array variation become similar and thus arrays appear to be correlated. This correlation is represented in the inter-sample correlation matrix and is utilized in the correction procedure. Correlations between genes induced by such large scale biological effects are not differentiable from correlations induced through large scale technical confounding effects and therefore our method will "correct" for both types of induced correlations. In this way, it is possible for our method or a method such as SVA that utilizes the inter-sample correlation matrix to remove a true biological signal. However, this caveat can also be a useful side effect, as the goal of many coexpression analyses is to find groups of genes that are tightly functionally related. Large scale effects, whether true biological effects or technical confounding, may hinder the ability to find smaller gene modules. In this sense our method can be seen as a complementary to current coexpression methods that identify large modules.

In order to evaluate MMC, we take advantage of the fact that microarrays contain many more probes than measured genes and that expression patterns for probes mea-

suring the same gene should be among the most highly correlated within the set of all probes. We compare methods for computing coexpression by comparing their ability to highly rank these probe pairs in terms of correlation. Our results show that MMC is able to rank these pairs more highly when compared to SVA and a traditional Pearson correlation. We evaluate our method further by utilizing replicate gene expression datasets. We utilized two yeast gene expression datasets [BYC02, SK08] produced by the same lab, covering the same strains and same genes but produced 5 years apart using different microarray platforms. We applied our method to both datasets and show that it is able to produce coexpression results which are more concordant when compared to both traditional Pearson and SVA corrected coexpressions. Finally, we consider how coexpressions may be used in order to identify biologically meaningful groups of genes. Under the assumption that genes working together in the same complex or pathway should be highly coexpressed, we examined coexpression values for sets of genes belonging to known functional categories. Given a set of known gene functional modules, we evaluated the ability of MMC coexpressions to identify these modules as biologically significant. Compared to both the traditional Pearson correlation and with SVA corrected coexpressions, we show that MMC has higher power to detect biologically meaningful gene sets.

2.2 Methods

We first highlight the relationship between a traditional Pearson's correlation coefficient and a basic linear model. We demonstrate the mathematical connection between the Pearson correlation and hypothesis testing under a linear model, and use this intuition when developing the mixed-model coexpression (MMC).

2.2.1 Pearson correlation as a linear model

The coexpression between two genes is often estimated by using the traditional Pearson correlation coefficient. The Pearson correlation gives an absolute value ranging from 0 to 1. If the absolute value of the correlation is close to 1, then we say that the pair of genes is significantly coexpressed. The threshold for significance is usually domain dependent and set on a case by case basis. The Pearson correlation can be calculated for any two genes, $\mathbf{y}_1$ and $\mathbf{y}_2$, by using Equation (2.1).

$$\mathbf{r}_P = \frac{\sum_{i=1}^{n}(y_{1i} - \overline{y}_1)(y_{2i} - \overline{y}_2)}{\sqrt{\sum_{i=1}^{n}(y_{1i} - \overline{y}_1)^2}\sqrt{\sum_{i=1}^{n}(y_{2i} - \overline{y}_2)^2}} \tag{2.1}$$

In this case, $\mathbf{y}_1$ and $\mathbf{y}_2$ are both gene expression vectors of size n and the Pearson correlation is the ratio of the their sample covariance to the product of their sample standard deviations. Here $\overline{y}_1$ and $\overline{y}_2$ are the sample means for each gene. The Pearson correlation can be represented concisely using matrix notation.

$$\mathbf{r}_P = \frac{(\mathbf{y}_1 - \mathbf{1}_n\overline{y}_1)^T(\mathbf{y}_2 - \mathbf{1}_n\overline{y}_2)}{\sqrt{(\mathbf{y}_1 - \mathbf{1}_n\overline{y}_1)^T(\mathbf{y}_1 - \mathbf{1}_n\overline{y}_1)}\sqrt{(\mathbf{y}_2 - \mathbf{1}_n\overline{y}_2)^T(\mathbf{y}_2 - \mathbf{1}_n\overline{y}_2)}} \tag{2.2}$$

We use $\mathbf{1}_n$ to represent a $n \times 1$ vector of 1s.

When the elements of $\mathbf{y}_1$ and $\mathbf{y}_2$ are sampled IID from a bi-variate normal distribution and are truly uncorrelated, the following relation holds [Wea49].

$$t = r_P\sqrt{\frac{n-2}{1-r_P^2}} \tag{2.3}$$

where t has a student t-distribution with $n-2$ degrees of freedom. In order to test the hypothesis that $\mathbf{r}_P = 0$, we test the equivalent hypothesis that $t = 0$, while evaluating t using the observed $\mathbf{r}_P$.

To understand how this relationship arises, let us consider the general purpose of the correlation coefficient. The correlation coefficient gives a measure of how linearly related one variable is to another. Another way to evaluate the linear dependence between two variables is by adopting a linear regression framework. Within this framework, the linear dependence between two variables is tested by first defining a linear model, in which one variable is used as a predictor of the other variable, which is called the response. We evaluate the magnitude of the linear dependence, by testing the hypothesis that the predictor variable has no effect on the response variable. With this in mind we define the two following linear models, in which we assume that each gene is a function of its mean, some random error and the observed expression value of another gene. We use $\hat{\mathbf{y}}_i$ to represent the observed gene expression vector for $\mathbf{y}_i$.

$$\mathbf{y}_1 = \hat{\mathbf{y}}_2\beta_1 + \mu_1 + \mathbf{e}_1 \tag{2.4}$$

$$\mathbf{y}_2 = \hat{\mathbf{y}}_1\beta_2 + \mu_2 + \mathbf{e}_2 \tag{2.5}$$

In order to evaluate the significance of the effect that gene 1 has on gene 2, we test the hypothesis that $\beta_2 = 0$ in the model in equation (2.5). Under the null hypothesis that $\beta_2 = 0$, we have that the computed t-statistic follows a central student t-distribution with $n - 2$ degrees of freedom [MS01]. The computed t-statistic is a function of both the estimate $\hat{\beta}_2$ and the sample variance for $\mathbf{y}_1$. Through a series of algebraic manipulations we can show that the computed t-statistic has the relationship observed in equation 2.3 with the Pearson's correlation [Rao73]. We briefly summarize this relationship here as follows.

$$t_1 = t_2 = r_P\sqrt{\frac{n-2}{1-r_P^2}} \tag{2.6}$$

$$r_P^2 = \frac{t_1^2}{t_1^2 + (n-2)} = \frac{t_2^2}{t_2^2 + (n-2)} \tag{2.7}$$

t_1 and t_2 correspond to the t-statistics computed for the estimates of β_1 and β_2, respectively. Equation (2.7) shows that there is a direct relationship between the Pearson correlation and a linear model of the type in equations (2.4) and (2.5). Under the null hypothesis, we assume that both t_1 and t_2 asymptotically follow the t-distribution. Implicit in this assumption is the assumption that the variance of the residuals e_1 and e_2 is of the form $\sigma_e^2\mathbf{I}$. More specifically, we assume that both $\mathbf{y}_1$ and $\mathbf{y}_2$ are normally distributed with means μ_1 and μ_2, respectively, and variances $\sigma_1^2\mathbf{I}$ and $\sigma_2^2\mathbf{I}$, respectively. When these assumptions do not hold, such as when the residuals are not independent, we might experience overdispersion of the test statistics [MS01]. In other words, the variance of the test statistics will be greater than expected and thus our assumed null distribution will be incorrect. This phenomenon has been observed in many cases, for example, when the effects of population structure are not accounted for when computing association statistics [DRW01]. Since overdispersion leads to inflation of test statistics and the Pearson correlation coefficient is directly proportional to the t-statistics for the models in equations (2.4) and (2.5), this implies that overdispersion may lead to inflation of the Pearson correlation.

2.2.2 Coexpression as a linear mixed model

In the previous section, we illustrated the relationship between a traditional Pearson's correlation and a linear model. We concluded from this relationship that when the variance of the residuals is misspecified, we have the potential to observe overdispersion, which leads to inflation of the test statistics and subsequently the Pearson's correlation. In the presence of expression heterogeneity, we expect that shared confounding between arrays will make them correlated. When this is the case, we no longer ex-

pect that the residuals for the models in equations (2.4) and (2.5) will be independent. Therefore, the assumption of independent residuals is incorrect and this misspecification might lead to overdispersion and subsequently to inflation of the Pearson's correlation.

One way to deal with overdispersion is to account for the source of overdispersion with a random variable [MS01]. Therefore, we propose the following two linear models that have an additional random variable, which accounts for confounding effects.

$$\mathbf{y}_1 = \hat{\mathbf{y}}_2\beta_1 + \mu_1 + \mathbf{u}_1 + \mathbf{e}_1 \tag{2.8}$$

$$\mathbf{y}_2 = \hat{\mathbf{y}_1}\beta_2 + \mu_2 + \mathbf{u}_2 + \mathbf{e}_2 \tag{2.9}$$

In these models, we assume that $var(\mathbf{e}_1) = var(\mathbf{e}_2) = \sigma_e^2\mathbf{I}$, $var(\mathbf{u}_1) = var(\mathbf{u}_2) = \sigma_u^2\mathbf{K}$ and that $cov(\mathbf{e}_i, \mathbf{u}_j) = 0 \quad \forall \quad i, j$, where $\mathbf{K}$ represents the inter-sample correlation matrix. Given a set of n arrays each measuring m genes, we define the inter-sample correlation matrix as the $n \times n$ sample covariance matrix for the $m \times n$ matrix of the complete array data. In other words, the matrix $\mathbf{K}$ is a matrix containing all pairwise covariances for all pairs of arrays. The key assumption here is that the additional variance due to systematic confounding effects is proportional to the correlation between arrays.

When gene 1 and gene 2 are truly uncorrelated ($\beta_1 = \beta_2 = 0$), the Pearson's correlation should be zero. However, when the models in equations (2.8) and (2.9) hold, the observed Pearson's correlation will be inflated due to correlation between the elements of $\mathbf{u}_1$ and $\mathbf{u}_2$. Subtracting the true values of $\mathbf{u}_1$ and $\mathbf{u}_2$ from $\mathbf{y}_1$ and $\mathbf{y}_2$, will produce corrected vectors for which the observed Pearson's correlation will not be inflated. However, the true values of these variables are unknown and in order to

obtain estimates of them, we would have to make further assumptions and restrictions on the model. Instead, we only make an assumption about the distributions of both $\mathbf{u}_1$ and $\mathbf{u}_2$. With knowledge of these distributions, we estimate the total variance of $\mathbf{y}_1$ and $\mathbf{y}_2$.

Under the null hypothesis, $\beta_1 = \beta_2 = 0$, we have the following.

$$\mathbf{y}_1 \sim N(\mu_1, \mathbf{\Sigma}) \tag{2.10}$$

$$\mathbf{y}_2 \sim N(\mu_2, \mathbf{\Sigma}) \tag{2.11}$$

where

$$\begin{aligned} \mathbf{\Sigma} &= var(\mathbf{u}_1) + var(\mathbf{e}_1) \\ &= var(\mathbf{u}_2) + var(\mathbf{e}_2) \\ &= \sigma_u^2 \mathbf{K} + \sigma_e^2 \mathbf{I} \end{aligned}$$

When the gene expression vectors follow the distributions in equations (2.10) and (2.11), the traditional Pearson's correlation will be inflated due to overdispersion. That is, when computing the Pearson's correlation, we assume that $\mathbf{\Sigma} = \sigma_c^2 \mathbf{I}$, for some σ_c^2. In order to remove the effects of overdispersion in each gene expression vector, we need to transform the gene expression vectors so that they have the same variance-covariance structure assumed when computing the Pearson's correlation. Then using these transformed vectors we apply the definition for a traditional correlation coefficient. This is accomplished by utilizing the following rule, which is applicable to random variables with a multivariate normal distribution with mean μ and positive semi-definite variance-covariance matrix Σ [KK04].

$$\mathbf{y} \sim \quad N(\mu, \Sigma) \tag{2.12}$$

$$A\mathbf{y} + b \sim \quad N(A\mu + b, A\Sigma A') \tag{2.13}$$

Using this rule we obtain the distribution for $\mathbf{y}_1^*$ and $\mathbf{y}_2^*$ defined as follows.

$$\mathbf{y}_1^* = \mathbf{\Sigma}^{-1/2}(\mathbf{y}_1 - \mu_1) \sim N(0, \mathbf{I}) \tag{2.14}$$

$$\mathbf{y}_2^* = \mathbf{\Sigma}^{-1/2}(\mathbf{y}_2 - \mu_2) \sim N(0, \mathbf{I}) \tag{2.15}$$

When the Pearson's correlation is calculated in this transformed space (i.e.. using the transformed vectors), we expect that the assumptions of independent residuals will hold and thus the correlation will not be subject to inflation. Given the true Σ and the observed gene expression vectors $\mathbf{y}_1$ and $\mathbf{y}_2$, we transform the observed vectors and calculate a corrected Pearson's correlation.

$$r_{MMC} = \frac{\mathbf{y}_1^{*T}\mathbf{y}_2^*}{\sqrt{\mathbf{y}_1^{*T}\mathbf{y}_1^*}\sqrt{\mathbf{y}_2^{*T}\mathbf{y}_2^*}} \tag{2.16}$$

We expect that this adjusted correlation coefficient will have a mean of zero when gene 1 and gene 2 are uncorrelated. Simplifying we obtain the following.

$$= \frac{(\mathbf{y}_1 - \mu_1)^T\mathbf{\Sigma}^{-1}(\mathbf{y}_2 - \mu_2)}{\sqrt{(\mathbf{y}_1 - \mu_1)^T\mathbf{\Sigma}^{-1}(\mathbf{y}_1 - \mu_1)}\sqrt{(\mathbf{y}_2 - \mu_2)^T\mathbf{\Sigma}^{-1}(\mathbf{y}_2 - \mu_2)}} \tag{2.17}$$

We are not given the true values of the means μ_1 and μ_2 or the true value of $\mathbf{\Sigma}$. In order to calculate r_{MMC} between two given gene expression vectors, we must estimate these parameters from the data. Substituting the estimates for μ_1, μ_2 and $\mathbf{\Sigma}$, we arrive at the final formula.

$$r_{MMC} = \frac{(\mathbf{y}_1-\overline{y}_1)^T\hat{\mathbf{\Sigma}}^{-1}(\mathbf{y}_2-\overline{y}_2)}{\sqrt{(\mathbf{y}_1-\overline{y}_1)^T\hat{\mathbf{\Sigma}}^{-1}(\mathbf{y}_1-\overline{y}_1)}\sqrt{(\mathbf{y}_2-\overline{y}_2)^T\hat{\mathbf{\Sigma}}^{-1}(\mathbf{y}_2-\overline{y}_2)}} \tag{2.18}$$

The t-statistics corresponding to the βs from the models in equations (2.8) and (2.9) maintain the relationship illustrated in equation (2.6), while r_{MMC} has been substituted for r_P.

In order to determine the value of r_{MMC}, we must first determine the value of $\hat{\mathbf{\Sigma}} = \hat{\sigma_u}^2\mathbf{K} + \hat{\sigma}_e^2\mathbf{I}$. This means that we need to estimate the two variance components, $\hat{\sigma}_e^2$ and $\hat{\sigma}_u^2$. When estimating only one variance component, the estimates are obtained analytically through a maximum likelihood (ML) or restricted maximum likelihood (REML) approach. However, there does not exist a general analytical method for estimated more than one variance component. Therefore, we must incorporate a numerical search strategy in order to obtain optimal estimates. Such solutions are computationally intensive. In order to estimate these variance components, we employ a method described by [KYE08]. This method reduces the computational complexity at each search step from $O(n^3)$, using the basic Newton-Raphson algorithm, to $O(n)$ by reformulating the problem so that the singular value decomposition of $\mathbf{K}$ can be reused. The method combines grid search with the Newton-Raphson algorithm and can be applied, in order to find the optimal variance components.

For each pair of genes, i and j, we use the numerical search method to find the optimal estimates for the variance components ${}_i\sigma_e^2$, ${}_i\sigma_u^2$, ${}_j\sigma_e^2$ and ${}_j\sigma_u^2$. We use the left sub-script to identify the gene for which the component belongs to. For example, ${}_i\sigma_e^2$ and ${}_i\sigma_u^2$ are estimated using the model for gene i (refer to equations (2.8) and (2.9)). Using the estimated variance components, we obtain ${}_i\hat{\Sigma} = {}_i\sigma_u^2\mathbf{K} + {}_i\sigma_e^2\mathbf{I}$ and ${}_j\hat{\Sigma} = {}_j\sigma_u^2\mathbf{K} + {}_j\sigma_e^2\mathbf{I}$, the variance-covariance matrices for the models corresponding to $\mathbf{y}_i$ and $\mathbf{y}_j$. These variance-covariance matrices are used to obtain the observed MMC coexpression values corresponding to gene i and gene j, ${}_ir_{MMC}$ and ${}_jr_{MMC}$. Ideally

these MMC coexpressions are equal. However, in practice this is not the case, so we average them to define the final corrected correlation r_{MMC}, in order to ensure the symmetry of coexpression. It is also possible to calculate r_{MMC} by using the corrected vectors $y_i^* = {}_i\hat{\mathbf{\Sigma}}(\mathbf{y}_i - \bar{y}_i)$ and $y_j^* = {}_j\hat{\mathbf{\Sigma}}(\mathbf{y}_j - \bar{y}_j)$ and then applying the definition of the Pearson's correlation from equation (2.16). When ${}_i\hat{\mathbf{\Sigma}} \neq {}_j\hat{\mathbf{\Sigma}}$, we found the solution to be very concordant with that obtained by averaging ${}_ir_{MMC}$ and ${}_jr_{MMC}$.

2.3 Results

2.3.1 Prioritizing probe pairs targeting the same gene

In order to evaluate the ability of MMC to prioritize true coexpressions, we leverage the fact that microarrays typically contain many more probes than there are genes to measure, meaning that most genes are targeted by more than one probe. We assume that the expression levels of any two probes targeting the same gene should be highly correlated, and thus when ranked against all other pairwise coexpressions, these probe pairs should be among the most highly ranked. We compare the relative ranking of coexpressions for probes targeting the same gene between different methods, in order to determine which method is better able to prioritize strong coexpressions. It may be noted that when certain forms of alternative splicing occur or when some genes are simply not expressed, the results of this evaluation strategy may fail to differentiate the methods for calculating coexpression.

We utilized a set of 732 probe pairs obtained from the Human HapMap gene expression arrays [Int03]. The gene expression data represents 60 unrelated individuals of European descent (http://www.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi?acc=GSE6536). The probe set corresponds to those probes that target known RefSeq genes and that could be coupled with at least one other probe targeting the same gene. For each probe

pair, we calculate the MMC coexpression, the traditional Pearson correlation and an SVA corrected Pearson correlation. Each expression value is ranked with respect to all pairwise coexpressions for each method. Smaller ranks indicate higher coexpression. We expect that when examining the coexpression ranks for the set of 732 probes, the method that performs best should have an abundance of low ranks.

Figure 2.1 shows the distribution of the coexpression ranks obtained with each method. The total number of genes considered was over 26,000, meaning that there were over 300 million pairwise coexpressions (26,000 choose 2) to consider. Subsequently, there are over 300 million possible rankings for each coexpression. Each method places about 96% of the 732 probe pairs within the top 1 million ranks. The figure shows that the MMC method consistently ranks these probe pairs higher than either of the other methods. For example, MMC places 79 of the 732 probe pairs within the top 100 ranks, while SVA and Pearson place only 63 and 76, respectively. In the top 10,000 ranks, MMC places 216 probe pairs, while Pearson and SVA place only 177 and 191. If we assume that each of the 732 probe pairs should have a correlation of 1, then their ranks should be in the top 732 choose 2. MMC places 415 of the 732 probe pairs within this range, while Pearson and SVA place only 370 and 366, respectively. These results suggest that MMC is more accurately calculating the coexpressions of these probe pairs, which we assumed to be truly coexpressed.

2.3.2 Concordance between replicated data sets

Replicate datasets are great resources to use in order to validate experimental findings. When considering coexpression, we expect that genes found to be highly coexpressed in replicate dataset 1 would also be highly coexpressed in replicate dataset 2. However, due to confounding effects, we may observe a high level of discordance between coexpressions found using two separate replicate datasets. Methods that re-

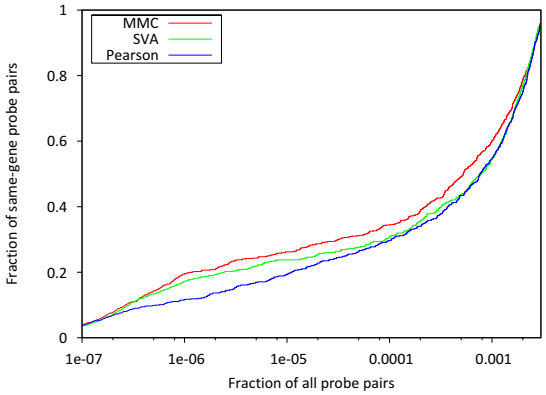


Figure 2.1: **The distributions of coexpression ranks for a set of 732 same probe pairs.** The coexpression values for each probe pair are ranked with respect to all other pairwise coexpression values. Smaller ranks indicate higher coexpression. We expect that probes targeting the same gene should be highly coexpressed and therefore should have very low rank. The MMC method consistently ranks these coexpressions lower when compared to the other two methods.

move confounding may alleviate this problem and cause coexpressions to be more concordant between replicate datasets. In order to evaluate the performance of MMC in this respect, we obtain two yeast gene expression datasets produced by the same lab and measuring the same genes over the same strains of yeast but conducted 5 years apart [BYC02, SK08]. For both datasets, we calculate coexpressions using MMC, traditional Pearson and SVA corrected Pearson. We then compare the concordance of coexpression values between the two datasets.

In order to compare coexpression values between two replicate datasets, we compare their relative rankings and compute the proportion that are common. We are considering a total of 6,143 genes, so there are over 18 million gene pairs and thus over 18 million coexpressions. We expect that the most highly coexpressed genes will be the same within both datasets. Given this, we define a measure of concordance

between two datasets in which we calculate the proportion of genes that are common within the top n most highly ranked coexpressions. For example, consider the top 100 coexpressions from dataset one, we might see that of these coexpressions only half appear in the top 100 when considering dataset two. In this case, we determine that the proportion in common is 50% for $n = 100$. By calculating the proportion in common for every n, we obtain a concordance at the top (CAT) plot, as shown in figure 2.2.

The CAT plot in Figure 2.2, illustrates the differences between concordance for each of the methods considered. Ideally, at each point on the x-axis the y-value would be 1, meaning that 100% of the coexpressions would be in common. Although, this is not the case, we do see that both MMC and SVA are concordant about 30 to 40% of the time when considering the top 200 ranks. However, when considering the ranks ranging from 300 to 50,000, our method out performs both methods by estimating coexpressions which are concordant 20 to 40% of the time. This result strongly suggests that MMC is more effective in removing confounding effects which may cause coexpressions to be discordant across datasets.

2.3.3 Gene module significance

One intention behind the calculation of coexpression is to quantify the strength of the biological relatedness between genes. For example, if two genes code for signaling proteins that act together in one particular pathway, we expect that these genes will be expressed together and that their coexpression value will be quite high. If we assume that the coexpression between two genes reflects the strength of their biological relationship, it is possible to utilize coexpressions in order to predict how biologically relevant a group of genes may be. Consider a group of genes that all code for proteins that work together in a complex. It is reasonable to assume that each pair of these genes will be coexpressed. In this case, by comparing each of the pairwise coexpressions for

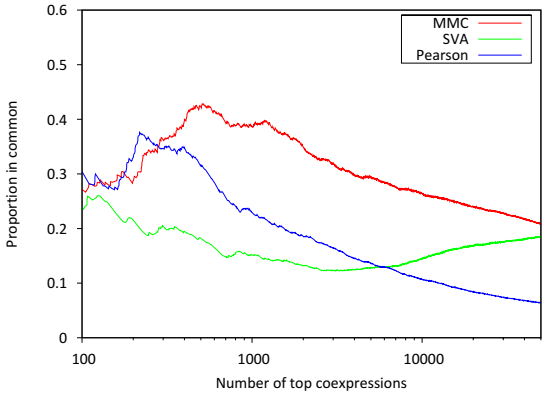


Figure 2.2: **Comparison of the concordance between two yeast datasets for both methods.** Concordance between two sets of coexpressions is compared by looking at the proportion of coexpressions in common for the top ranking coexpressions. The x-axis represents the number of top ranked coexpressions considered, while the y-axis represents the proportion of those coexpressions that are common between the new and old dataset.

the genes within this group we should see an abundance of significant coexpressions. In general, we can assume that a group of genes that are functionally related should all be significantly coexpressed. We can then use this assumption to test the significance of a group of genes in order to determine if it is biologically relevant. In practice, by using such an approach we will likely find many groups of genes which will appear to be biologically relevant, but in fact their high level of inter-coexpression is due to confounding.

We tested our ability to detect biologically significant groups of genes using MMC coexpressions. We define the statistic found in equation (2.19), which is simply the sum of the logged coexpression ranks. $rank_{ij}$ represents the relative ranking of the coexpression between gene i and gene j, with respect to all other pairwise coexpressions. When genes within a group are highly coexpressed, the value of this statistic

will be high and when they are not it will be very low. To obtain a set of gene groups which are known to be functionally related, we chose to use yeast, as it has some of the most well characterized genes. The MIPS comprehensive yeast genome database contains detailed functional data for all yeast genes [MHK99]. We used this resource in order to construct 233 gene modules ranging in size from 2 to 20. Modules were chosen such that the number of modules was maximized while the modules in each size category did not overlap. We chose sizes of 2 to 20, assuming that smaller modules would represent more closely functionally related gene sets and thus the overall coexpressions within modules would be higher. For each of the 233 modules, we calculated the statistic T using coexpressions estimated with MMC, traditional Pearson and SVA corrected Pearson. We estimate the null distribution for T under each method and each module size n, by repeatedly selecting n random coexpressions and calculating the statistic T. Each null distribution was approximated with 1 million values. Using this null approximation we calculated p-values for each known module.

$$T = \sum_{i=0}^{n-1} \sum_{j=i+1}^{n-1} log(rank_{ij}) \tag{2.19}$$

Figure 2.3 shows the distribution of the p-values for all modules. Module p-values obtained when using our method tend to be smaller than the Pearson and SVA module p-values. For example, about 40% of the tested gene modules were significant at a level of .05 when using MMC, while about 25% and 30% were significant when using Pearson and SVA. This result suggests that MMC is able to produce coexpression values which were better able to predict real biological relationships.

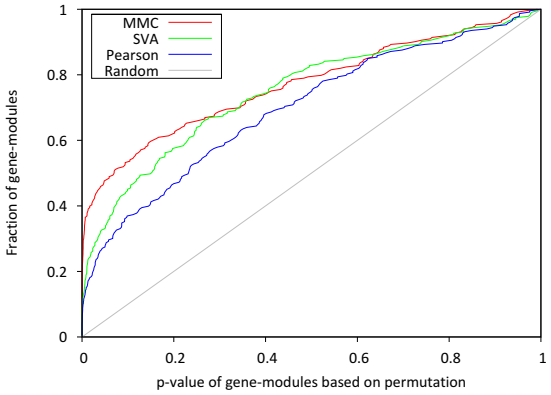


Figure 2.3: **Distribution of gene-module p-values for Pearson, SVA and MMC.** We used a set of 233 known functional modules consisting of sets of genes of size 2 to 20. For each of these modules, a p-value representing the biological significance is calculated. This figure plots the distributions of these p-values. Since the p-values were calculated for gene sets known to be functionally related, we expect that there should be an inflation of significant p-values. It can be seen that the MMC method produces a larger number of significant p-values when compared to both the traditional Pearson and SVA-corrected coexpressions.

2.4 Discussion

In this chapter, we present a statistical model for the calculation of gene coexpression called Mixed Model Coexpression (MMC). Our method calculates gene coexpressions that are robust to confounding effects. We calculate the coexpression between two genes by utilizing a mixed model framework. Unknown confounding effects are represented as a random variable in a mixed model formulation of coexpression. We use the inter-sample correlation to estimate the variance of the random variable representing unknown confounding and incorporate this variance into the model of coexpression.

We compare the coexpressions obtained with our method with those obtained using the traditional Pearson correlation and those obtained using SVA corrected expression data. Although, rank based correlation methods, such as the Spearman correlation, have been used to reduce the prevalence of spurious correlations due to deviations from assumptions of normality in expression data, we have observed in practice that the Spearman correlation coefficient performs similarly to the Pearson when comparing with MMC (data not shown). When probe pairs target the same gene, we expect that their coexpressions will be highly ranked when compared with all other pairwise coexpressions. For probe pairs of this type, MMC is shown to produce coexpressions that are more highly ranked when compared with the other two methods. We also show that MMC produces coexpressions that are more concordant across replicate datasets generated by the same lab using the same strains but generated at different times. Operating under the assumption that biologically and functionally meaningful groups of genes will be highly coexpressed, we create a simple statistic which is used to assess the functional significance of groups of genes. Our method shows increased power to discover sets of genes which are known to be biologically significant.

Although our method is able to calculate coexpression while removing the effects of confounding, it might also remove effects which are biologically meaningful. Tech-

nical confounding effects, such as a batch effect, typically have a global effect on the data. That is, these effects will increase the expression variation in a large number of genes. This shared variation within genes causes them to appear to be significantly co-expressed. Our method estimates global patterns of shared variation through the inter-sample correlation and effectively removes the effects causing the variation from the calculation of coexpression. A problem arises when we consider the case in which one gene has a large biological effect on hundreds of other genes. The effect that master regulators have on expression data as a whole is indistinguishable from the unwanted global confounding effects. That is, the variation in gene expression caused by master regulators quite closely resembles patterns of variation caused by confounding effects and will therefore be removed by our method. In this case, MMC may over-correct true biological signal and cause true coexpressions to be lost.

The drawback to our method may also be seen as a beneficial side effect. When master regulators target many genes, traditional coexpression analyses employing clustering will yield many large sized gene modules. By removing the effects of master regulators, MMC essentially enables coexpression clustering analysis to produce smaller gene modules conditional on the large modules. Large gene modules discovered through the use of standard coexpression analysis may be seen as representative of large scale cellular functionality. Small modules discovered through clustering with MMC will be subsets of these large modules. By intersecting results, it may be possible to more fully understand the detailed circuitry of the cell.

Reference to published article

Furlotte, Nicholas A, Hyun Min Kang, Chun Ye, and Eleazar Eskin. 2011.
Mixed-model coexpression: Calculating gene coexpression while accounting for expression heterogeneity. Bioinformatics 27 (13): i288-i294.

http://bioinformatics.oxfordjournals.org/content/27/13/i288.abstract.

CHAPTER 3

Meta-analysis for Structured Populations

3.1 Background

Model organisms continue to play a pivotal role in the research of human diseases. The use of mouse models in particular has been extremely effective for the the identification of genes underlying Mendelian disorders. The traditional mode of discovery used to identify loci underlying such 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 studies [FM01, BFO10b], 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 [KYE08, 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. (2010) 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 [BFO10b]. 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-statistic, which combines information across studies. The most common methods for performing meta-analysis are based on the fixed effect weighted sum of Z-scores (WSoZ) [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 [AVF11], and the use of heterogeneous stocks [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 when compared to 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 [FNG09, FBO11]. In our results, two previously implicated loci are recovered with increased significance. We also 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.

3.2 Methods

3.2.1 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 $\mathbf{y}_i = [y_{i1} \quad y_{i2} \ldots y_{in_i}]'$. 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 (3.1), where μ is the global phenotype mean and $\mathbf{x}_i$ is a vector of minor allele counts of SNP r for individuals in population i.

$$\mathbf{y}_i = \mu + \mathbf{x}_i\beta + \epsilon \tag{3.1}$$

A test statistic for testing $\beta = 0$ is derived by noting the distribution of the estimate of β under the assumption of normality. We denote the estimate of β 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 (3.2) 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.

$$Z_i = \frac{\hat{\beta}_i}{s_i} \tag{3.2}$$

3.2.2 Traditional Meta-Analysis

Most traditional methods for meta-analysis employ the weighted sum of z-scores (WSoZ) approach [SRC09, WSL08, ZSS08]. In this method, a meta statistic for each SNP is calculated using equation (3.3), where w_i denotes a weight given to each Z-score for a population i. We note that our meta statistic formulation in equation (3.3) uses a different notation, with respect to standard meta-analysis literature, which represents the meta statistic as a sum over effect sizes. However, both formulations are equivalent.

$$Z_{meta} = \frac{\sum_i w_i Z_i}{\sqrt{\sum_i w_i^2}} \tag{3.3}$$

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 β.

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

($\sum_i w_i \frac{\beta}{s_i} \leq \sqrt{\sum_i w_i^2}\sqrt{\sum(\frac{\beta}{s_i})^2}$). Under the assumption that β is the same across all populations, equality holds when $w_i = 1/s_i$

3.2.3 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 identify true associations as well as cause the appearance of a large number of false or spurious associations. Mixed effects models are often used in order to correct this problem [YPB06, KYE08, 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 (3.1).

$$\mathbf{y}_i = \mu + \beta \mathbf{x}_i + \mathbf{u}_i + \epsilon \tag{3.4}$$

In the model in equation (3.4), the random variable $\mathbf{u}_i$ 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 $\mathbf{u}_i \sim N(0, \sigma_g^2 \mathbf{K}_i)$, where $\mathbf{K}_i$ is the $n_i \times n_i$ kinship coefficient matrix for population i. With this assumption, the total variance of $\mathbf{y}_i$ is given by $\mathbf{\Sigma}_i = \sigma_g^2 \mathbf{K}_i + \sigma_e^2 \mathbf{I}$. 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 (3.4), shown in equation (3.5).

$$\mathbf{y}_i = \mathbf{X}_i \Gamma + \mathbf{u}_i + \epsilon \tag{3.5}$$

In equation (3.5), $\mathbf{X}_i$ is a $n_i \times 2$ matrix encoding the global mean and SNP vectors and Γ is a 2×1 coefficient vector. We note that this form also easily extends to models with multiple covariates. The maximum likelihood estimate for Γ in population i is given by $\hat{\Gamma}_i = (\mathbf{X}_i'\mathbf{\Sigma}_i^{-1}\mathbf{X}_i)^{-1}\mathbf{X}_i'\mathbf{\Sigma}_i^{-1}\mathbf{y}_i$ which follows a normal distribution with a mean equal to the true Γ and variance $(\mathbf{X}_i'\mathbf{\Sigma}_i^{-1}\mathbf{X}_i)^{-1}$. The z-score statistic for testing $\beta = 0$ is then given in equation (3.7), where $\mathbf{R} = [0 \quad 1]$ is a vector used to select the appropriate entry in the vector $\hat{\Gamma}_i$.

$$Z_i = [\mathbf{R}(\mathbf{X}_i'\mathbf{\Sigma}_i^{-1}\mathbf{X}_i)^{-1}\mathbf{R}']^{-1/2}\mathbf{R}\hat{\Gamma}_i \tag{3.6}$$

$$= Q_i^{-1/2}\mathbf{R}\hat{\Gamma}_i \tag{3.7}$$

$$Q_i = [\mathbf{R}(\mathbf{X}_i'\mathbf{\Sigma}_i^{-1}\mathbf{X}_i)^{-1}\mathbf{R}'] \tag{3.8}$$

3.2.4 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 (3.3), where the z-score for population i is given in equation (3.7). When $\beta \neq 0$, Z_{meta} will have a normal distribution with variance 1 and mean $\sum w_i Q_i^{-1/2}\mathbf{R}\mathbf{\Gamma} \,/\sqrt{\sum w_i^2}$. Again we employ the use of the Cauchy-Schwarz inequality, shown in equation (3.9), to show that the optimal weights are given by $w_i = Q_i^{-1/2}$. We may also arrive at this result by noting that $Q_i^{-1/2}$ from equation (3.7) 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 Γ may be evaluated.

$$\sum w_i Q_i^{-1/2}\mathbf{R}\Gamma \leq \sqrt{\sum w_i^2}\sqrt{\sum (Q_i^{-1/2}\mathbf{R}\Gamma)^2} \tag{3.9}$$

By substituting the optimal weights we arrive at the final meta statistic given in equation (3.10) with its distribution under the alternative hypothesis given in equation (3.11).

$$Z_{meta} = \frac{\sum Q_i^{-1}\mathbf{R}\hat{\Gamma}_i}{\sqrt{\sum Q_i^{-1}}} \tag{3.10}$$

$$\sim N(\beta\sqrt{\sum Q_i^{-1}}, 1) \tag{3.11}$$

It should be noted that when $\mathbf{\Sigma}_i$ is unknown, it must be estimated from the data. In this case, Z_{meta} may not follow a standard normal distribution under the null, due to the unaccounted uncertainty in the estimation of $\mathbf{\Sigma}_i$. However, we are able to side step this issue by using a global search technique [KYE08, KSS10], in order to find an optimal estimate of $\mathbf{\Sigma}_i$ for each population.

3.2.5 Simulations

Simulations were performed using a previously designed framework [KKW10, BFO10b]. For both power and resolution, phenotypes were generated by sampling a phenotype for each strain while assuming the model from equation (3.4). The genetic variance σ_g^2 was determined for a given genetic background (g^2) by using equation (3.12), where $\mathbf{S} = \mathbf{I}_n - 1/n\mathbf{J}_n$ ($\mathbf{J}_n$ is an $n \times n$ matrix of ones).

$$\sigma_g^2 = \frac{g^2\sigma_e^2(n-1)}{(1-g^2)Tr(\mathbf{S}\mathbf{K}_i\mathbf{S})} \tag{3.12}$$

The power and resolution for each effect size (β) 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.

3.2.6 Significance Threshold Estimation

Significance thresholds were estimated for each method using a technique utilized previously [KKW10, BFO10b]. 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: 3.715×10^{-6}, 2.4637×10^{-4} and 2.7×10^{-6}.

3.2.7 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 work, 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. This process assumes that the probability of a double recombination occurring between genotyped markers is close to zero.

The genotypes and phenotypes utilized in this work have been made available online at `http://genetics.cs.ucla.edu/mousemeta/`.

3.3 Results

3.3.1 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 [KKW10, BFO10b].

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 [KYE08] 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 (see Methods).

Figure 3.1 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 β from equation (3.4) 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 3.1d).

3.3.2 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 3.2 compares the distribution of these distances under each mapping method. Simulations were performed 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 association 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 resolution over the F2 panel.

3.3.3 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 [KYE08], and applied the meta-analysis approach as well. Manhattan plots summarizing these results are shown in Figure (3.3). 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 (3×10^{-7}) than either the HMDP (3.1×10^{-6}) or the F2 (1.6×10^{-3}) peaks. The original QTL on Chr. 7 (*Bmd41*) had a 1.5 LOD support

interval of 80 Mb (24.9 to 104.9 Mb) [FNG09]. We approximate the associated region obtained via meta-analysis by employing a simple approach. We define the associated region as that surrounding the peak SNP and containing SNPs with p-values $\leq 10^{-6}$. Thus defined, the Chr. 7 meta-analysis interval extending from 17.2 to 25.2 Mb is much smaller than the previously obtained support interval. This result indicates an increase in resolution for *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 (1.3×10^{-3}) and highly significant in the meta-analysis (2.8×10^{-6}). *Bmd7* 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 $\leq 10^{-6}$ spanned 10 Mb (from 129 to 139 Mb).

3.3.4 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 [KYE08] and then applied our meta-analysis approach. Figure 3.4 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 *Apoa2* [DLW90] gene exists. The peak association is 25kb upstream of the start site of the *Apoa2* gene with a p-value of 7.06×10^{-8}, which is significant at the 1×10^{-7} level estimated from a parametric bootstrap procedure. The mapping results obtained from the F2 panel(Figure 3.4a) 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 *Apoa2* locus with a p-value of 2.67×10^{-09}.

Figure 3.4b 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 1×10^{-15}.

3.4 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 increased 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.

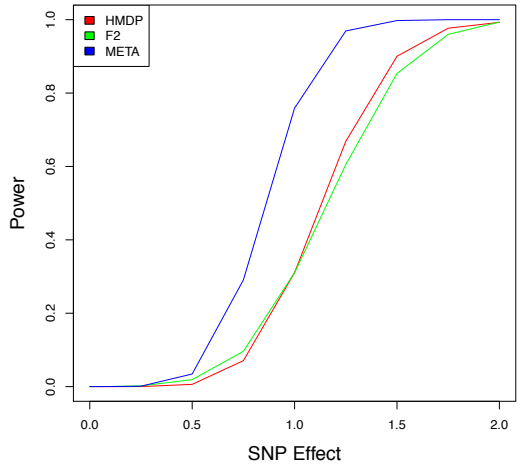
Our results have focused on the case when the HMDP panel is combined with one F2 cross. However, by using the methodology we present any number of panels can be combined. One obvious potential for this is that by adding additional F2 panels, we may increase power much further. A significant amount of 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

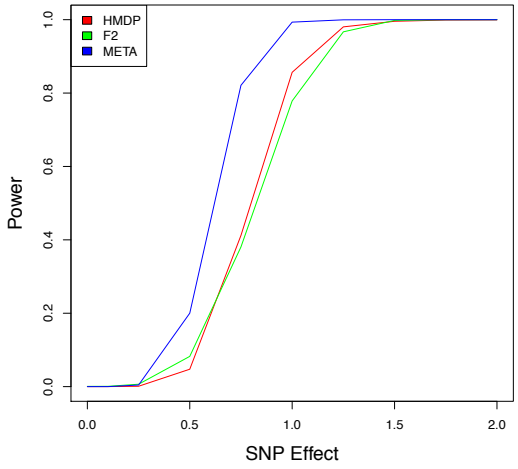
arise when combining the HMDP with such panels is that of heterogeneity of effect size. That is, the magnitude of main effects may vary between different mapping panels due to the difference in the overall genetic structure. In this case, our method may be easily extended to utilize approaches which account for such heterogeneity between effects [HE11]. 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.

Reference to published article

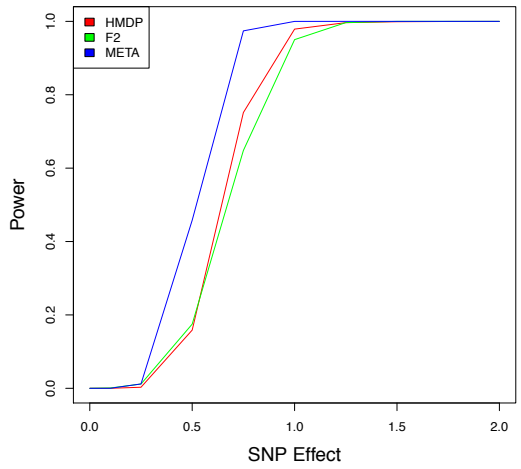
Furlotte, Nicholas A, Eun Yong Kang, Atila Van Nas, Charles R Farber, Aldons J Lusis, and Eleazar Eskin. 2012. Increasing association mapping power and resolution in mouse genetic studies through the use of meta-analysis for structured populations. Genetics doi:10.1534/genetics.112.140277.



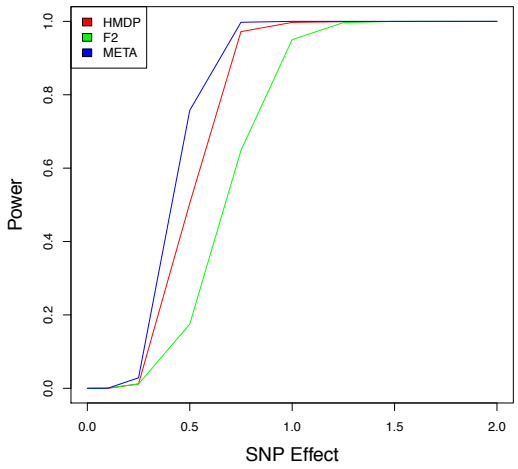
(a) 100 HMDP, 100 F2s



(b) 200 HMDP, 200 F2s



(c) 300 HMDP, 300 F2s



(d) 500 HMDP, 300 F2s

Figure 3.1: **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.

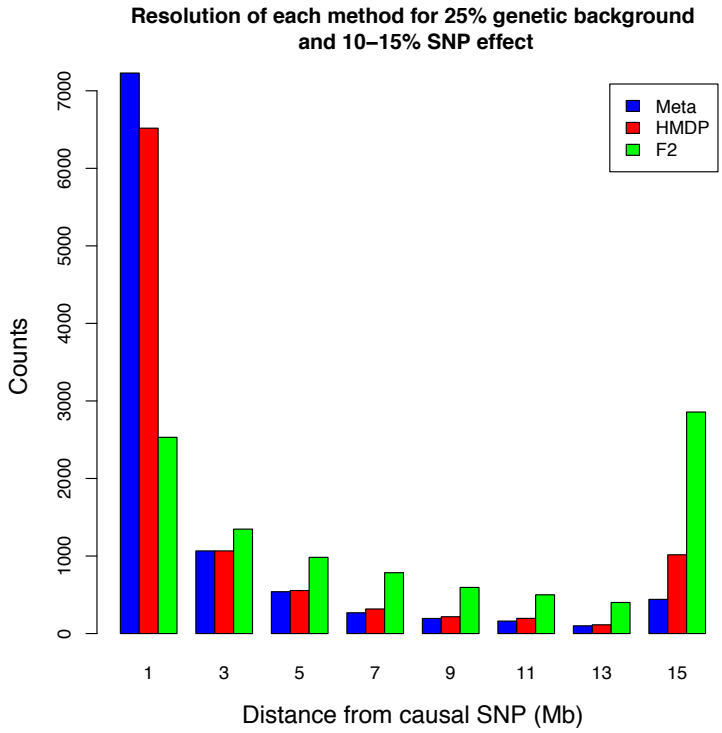


Figure 3.2: **The combined mapping result has higher resolution than that of the HMDP or F2 mapping.** The distribution of distances from the true causal SNP to the most significant association are shown in units of megabases. We considered peak associations which are within 15mb of the true causal SNP. As expected, the HMDP has much higher resolution than the F2 cross. The combined result achieves an even higher resolution.

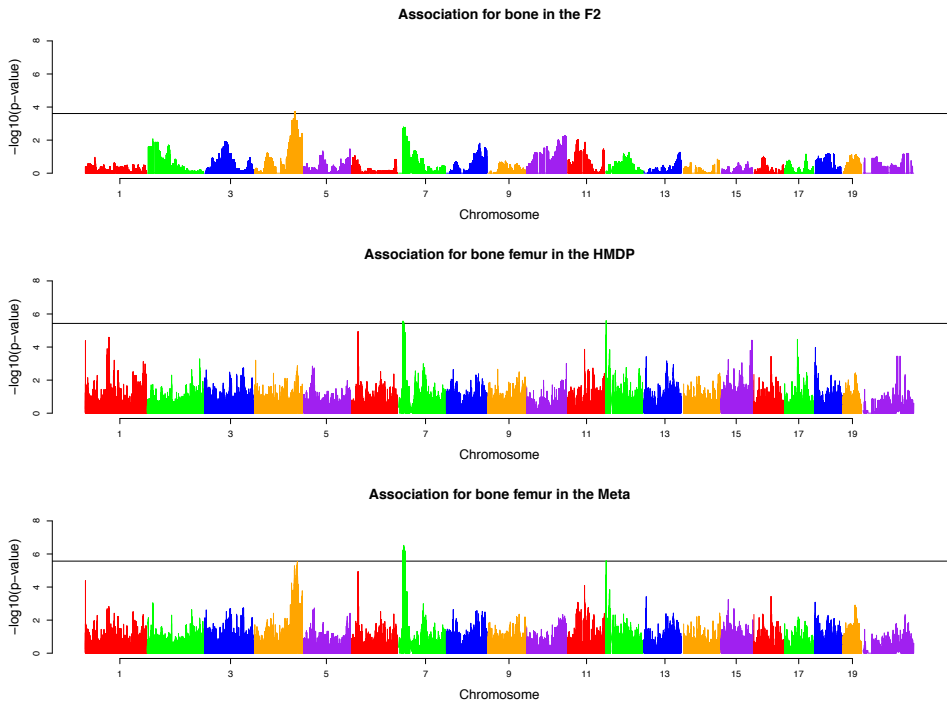
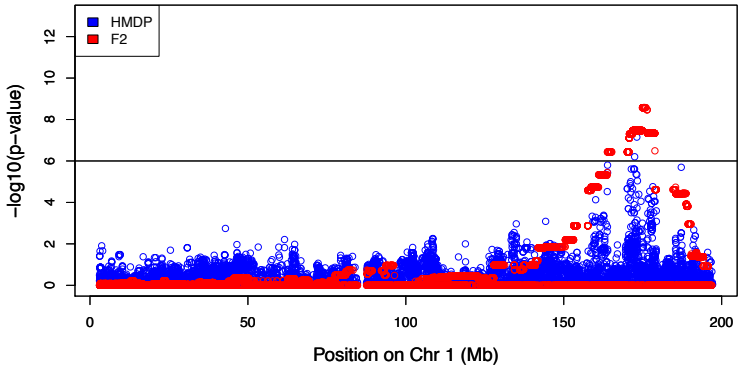
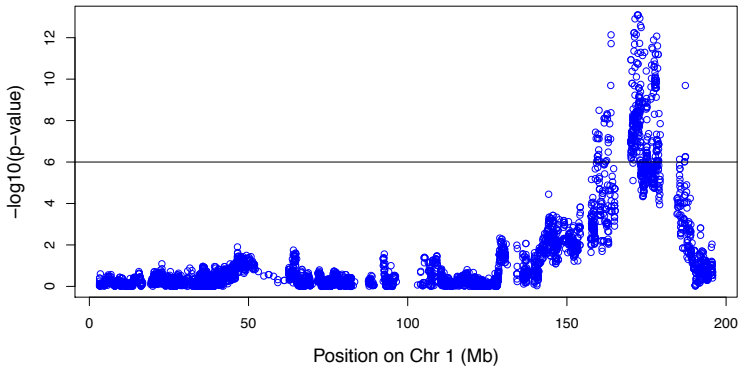


Figure 3.3: **Meta-analysis results in increased significance and increased resolution for two loci known to be associated with BMD.** Two loci, one on chr 4 and one on chr 7, were previously found to be associated with BMD (*Bmd7* and *Bmd41* respectively) [FNG09]. After applying meta-analysis, we found that the peak associated SNPs for both of these loci had increased significance with respect to the F2 and HMDP mapping panels. Thresholds for significance are indicated by the horizontal black bars in each plot.



(a) Association for HDL in the HMDP and F2 cross



(b) Meta Association for HDL

Figure 3.4: **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. The peak association in the HMDP result is 25kb away from the start site of *Apoa2* with a p-value of 7.06×10^{-08}, while the peak association in the F2 is 2Mb downstream of the start site (p-value 2.67×10^{-09}). After applying the meta-analysis method, we recover the same association identified in the HMDP with a p-value 2.36×10^{-15}. The horizontal black bar is placed at $-log10(p) = 6$ for reference.

CHAPTER 4

Genome-wide association mapping with longitudinal data

4.1 Background

The use of genome-wide association study (GWAS) methodologies has become common practice in the analysis of complex traits. However, it is well-known that the genetic variants identified for most common traits only account for a small portion of the heritability [MCC09, EFG10], implying that there are a large number of genetic associations yet to be identified. Many GWAS have been performed in cohorts, in which multiple time points are available for each individual [KMD07, Sab08, ARL08, KMO10]. However, current association methods only utilize one time point for each individual. This is accomplished by either selecting a single measurement [Sab08] or by computing the average over all time points [KMD07, IML07, KMO10]. It is reasonable to assume that a method jointly considering all time points when performing association may have increased power over single time point approaches.

In this chapter, we present a method for performing association mapping with longitudinal phenotypes and show that this method has increased power over single time point mapping approaches. Recently, there has been an interest in the prediction of the genetic contribution to traits using cohort data [YBM10]. These methods utilize a single phenotype measurement for each individual and predict the genetic contribution to

the phenotype by taking into account the relationships between individuals. We show that when utilizing multiple measurements, this genetic contribution can be differentiated from the environmental and error contribution and we show how each of these contributing factors can be accurately predicted. Our method utilizes a mixed effects approach to model phenotype measurements. Mixed effects models have been used extensively for modeling correlated data and are an important tool in the analysis of longitudinal data [Har77, LW82]. This class of models has been used in the analysis of longitudinal data for twin studies [WGH11] as well as pedigree-based family studies [AGV02]. We propose a model that partitions each individual's trait measurements into both a genetic and environmental component. We refer to the genetic component as the "genetic influence" to the trait. Similarly, we refer to the part of the environmental contribution as the "lifestyle value".

In order to evaluate our method, we first compare power with a traditional mapping procedure utilizing only one time point. Power is evaluated through an analytical approach similar to that introduced by [WB99]. Using a set of individuals obtained from the Wellcome Trust Case Control Consortium (WTCCC), we show that our method has increased power over traditional approaches. Second, we evaluate the accuracy in calculating genetic influence and lifestyle values for individuals while varying the number of available time points. We show that for phenotypes heavily influenced by the environment, the accuracy in prediction of the proportion of the phenotype due to genetics and environment has large variation. However, when ranking individuals based on their predicted lifestyle values, we find that this ranking is highly concordant with the ranking obtained using the true lifestyle values. This implies that individuals may be effectively categorized by lifestyle based on these predicted values.

4.2 Methods

4.2.1 Longitudinal Phenotypes

In experiments adopting longitudinal designs, phenotype measurements are collected for each of n individuals at m time points. We expect that measurements acquired from the same individual will tend to be more correlated than those obtained from different individuals. This correlation is due to both genetic and environmental effects shared between measurements. In order to conceptualize this, we present a generative model for phenotype measurements, which is a model specifying the mathematical process by which measurements may be systematically generated.

$$y_{ij} = \mu + G_i + E_{ij} + \epsilon_{ij} \tag{4.1}$$

The generative model in equation (4.1) states that a phenotype measurement j from individual i is a function of the global phenotype mean μ, an individual-specific genetic effect G_i, a measurement and individual-specific environmental effect E_{ij} and an error term ϵ_{ij}, accounting for other unknown factors such as measurement error. The value of G_i is a function of the genetic variation for individual i, and is expected to remain constant over time as an individual's genetic make up does not change. This assumption may not hold if there exist, for example, gene-by-environment interactions. On the other hand, the value of E_{ij} may vary across measurements due to changing environmental conditions. The correlation between each pair of E_{ij}s will depend on the magnitude of environmental change between time points as well as the degree of influence environment has over the phenotype in question. The residual terms ϵ_{ij} are expected to be independent between measurements.

4.2.2 Traditional Approach to Association Mapping

The traditional approach to association mapping considers one measurement for each of n individuals and interrogates each genetic locus individually. The traditional model is given as follows

$$y_i = \mu + \beta_r x_{ir} + \epsilon_i \tag{4.2}$$

x_{ir} represents the state of SNP r for individual i and β_r its coefficient [Bal06]. By testing the hypothesis $\beta_r = 0$, it is determined whether SNP r influences the trait or not. We note that, with respect to the model in equation (4.1), the model in equation (4.2) has folded many terms into the residual term ϵ_i. This model is represented using standard vector notation as follows

$$\mathbf{y} = \mathbf{X}_r \beta + \epsilon \tag{4.3}$$

$\mathbf{y}$ is a vector of all phenotype measurements and $\mathbf{X}_r = [\mathbf{1}_n \quad \mathbf{x}_r]$, where $\mathbf{x}_r$ is a vector representing the n SNP values for SNP r, β is a vector of coefficients and $\mathbf{1}_n$ is a column vector of ones. Other fixed effects may be added to $\mathbf{X}$ in order to account for additional confounding.

In order to apply this model to longitudinal data, the total set of mn measurements must be pre-processed into a set of n independent measurements. There are two common approaches taken and we refer to these as the single approach and the average approach. In the single approach, a single measurement from the set of m measurements is chosen for each individual i. In the average approach, the m measurements for each individual are averaged and the average value is used as the single phenotype measurement for that individual. Under the assumption that individuals are unrelated,

both of these procedures result in a set of n independent measurements and the standard model may be applied.

4.2.3 Mixed Effects Model for Association Mapping

The traditional method for association mapping interrogates each genetic locus individually while using single time points. However, it is known that traits are often influenced by many loci and ignoring this fact may have a negative impact on association mapping results. In particular, global genetic similarities between individuals may be correlated with trait similarities and this global correlation may cause many genetic loci to appear to be associated with the trait, a problem often referred to as population structure or cryptic relatedness [DRB01, VP05]. One way to account for this structure is through the use of a variance component model, in which the global genetic relatedness, referred to as polygenic background, of individuals is accounted for by the introduction of a random variable into the simple model from equation (4.2) [Lan02, YPB06, KSS10]. This model is summarized as follows.

$$y_i = \mu + \beta_r x_{ir} + u_i + \epsilon_i \tag{4.4}$$

The model is equivalently described in matrix notation using

$$\mathbf{y} = \mathbf{X}_r\beta + \mathbf{Zu} + \epsilon \tag{4.5}$$

The random variable u_i is assumed to be normally distributed with mean zero and variance σ_g^2 and the $cov(u_i, u_j) = \sigma_g^2 K_{ij}$, where K_{ij} is the kinship coefficient for individual i and j, which is a value representing their genetic relatedness. The incidence matrix $\mathbf{Z}$ maps measurements from each individual to the phenotype vector $\mathbf{y}$, and in the case when there is only one measurement for each individual $\mathbf{Z} = \mathbf{I}_n$. This

form is standard in the mixed model literature. With this the $var(\mathbf{Zu}) = \sigma_g^2 \mathbf{ZKZ}'$ and $var(\epsilon) = \sigma_\epsilon^2 \mathbf{I}$. The total variance of $\mathbf{y}$ is then given by

$$\mathbf{\Sigma} = \sigma_g^2 \mathbf{ZKZ}' + \sigma_\epsilon^2 \mathbf{I} \tag{4.6}$$

In order to test the hypothesis $\beta_r = 0$ using the model in equation (4.5), the two variance components σ_g^2 and σ_ϵ^2 must be estimated. Since there is no analytical solution, this is accomplished using a numerical search algorithm implemented in the program EMMAX [KSS10]. EMMAX combines grid search with the Newton-Raphson algorithm, in order to find the optimal variance components σ_g^2 and σ_ϵ^2 in time linear in the number of measurements, given the singular value decomposition of $\mathbf{K}$. Furthermore, by assuming that each SNP only has a small to moderate effect on the phenotype, it is reasonable to assume that variance component estimates will be the same for each SNP. With this assumption it is only necessary to perform the variance component search once and thus feasible to perform the hypothesis test for each SNP within the genome.

There are many methods to compute the kinship matrix $\mathbf{K}$. For a review of many standard relatedness estimators see [OWA06]. More recently, [YBM10] proposed a method for adjusting the relatedness matrix, to account for the fact that the true causal SNPs may not be strongly correlated with the genotyped SNPs. Such issues are beyond the scope of this work and thus we will only use the IBS allele sharing matrix [KYE08]. Furthermore, these issues and the choice of kinship matrix do not affect our simulation results. In general, the methodology introduced in this chapter may be utilized as long as the kinship matrix is positive semi-definite.

4.2.4 Association Mapping with Longitudinal Data

The models from equations (4.2) and (4.4) do not take advantage of the availability of multiple time points and do not directly account for both genetic and environmental factors. We suggest a model that directly accounts for each term using all time points by extending the model in equation (4.4) .

$$y_{ij} = \mu + \beta_r x_{ir} + u_i + v_{ij} + \epsilon_{ij} \tag{4.7}$$

The random variable v_{ij} is introduced to represent the contribution of the environment to the phenotype measurement (E_{ij} from equation (4.1)). The matrix version is as follows.

$$\mathbf{y} = \mathbf{X}_r\beta + \mathbf{Zu} + \mathbf{v} + \epsilon \tag{4.8}$$

We assume, without loss of generality, that the mean of the random components $\mathbf{u}$ and $\mathbf{v}$ are equal to zero and that the variance structure is as follows, where $\mathbf{D}$ is a known matrix representing the covariance between environmental components.

$$var\begin{bmatrix}\mathbf{u}\\ \mathbf{v}\\ \epsilon\end{bmatrix} = \begin{bmatrix}\sigma_g^2\mathbf{K} & 0 & 0\\ 0 & \sigma_v^2\mathbf{D} & 0\\ 0 & 0 & \sigma_\epsilon^2\mathbf{I}\end{bmatrix} \tag{4.9}$$

With this we define the variance of $\mathbf{y}$.

$$var(\mathbf{y}) = \mathbf{\Sigma} = \sigma_g^2\mathbf{ZKZ}' + \sigma_v^2\mathbf{D} + \sigma_\epsilon^2\mathbf{I} \tag{4.10}$$

In general, the matrix $\mathbf{D}$ will depend on the level of correlation between individual time points and can be determined through estimation techniques or by fitting

parametric models, such as models of the autoregressive class [JS86]. Most commonly $\mathbf{D}$ will take the form of a block diagonal matrix, so that environmental components between individuals will be independent. The variance of $\mathbf{v}$ is then given as $var(\mathbf{v}) = \mathbf{D} = \mathbf{E} \otimes \mathbf{I}$, where $\otimes$ represents the Kronecker product of two matrices, and $\mathbf{E}$ is an $m \times m$ matrix representing the covariance between the set of m time points for each individual.

4.2.5 Missing Data

One complication that often arises when dealing with longitudinal data is that of unbalanced or missing data [MS01]. When a study is unbalanced, meaning that all individuals do not have the same number of measurements, the model notation becomes slightly more complicated. Let us consider that individual i has m_i measurements and define $\mathbf{m} = [m_1 \quad m_2 \quad \dots m_n]$. The incidence matrix $\mathbf{Z}$ will still map genetic components to measurements and its structure will be dictated by the vector $\mathbf{m}$. For example, the first m_1 rows of $\mathbf{Z}$ will have a 1 in the first column and the second m_2 rows of $\mathbf{Z}$ will have a 1 in the second column and so on.

In order to avoid complicated notation, we suggest a simple scheme for defining the model. Define the model using $m = max(\mathbf{m})$, so that the assumed number of measurements is equal to nm and the vector $\mathbf{y}$ has missing values. Select the measurements that are missing at each individual and remove the rows and columns of the full covariance matrices of each component (genetic, environmental and residual error) that correspond to the indices of these entries in the vector $\mathbf{y}$ of size nm. The resulting covariance matrices will then correspond to a new vector $\tilde{\mathbf{y}}$, defined as the vector $\mathbf{y}$ with the missing values removed. Modeling fitting procedures are then readily adaptable to such matrices and hypothesis testing procedures can be easily applied.

4.2.6 Estimating Variance Components

In order to fit the models in equations (4.5) and (4.8), we must estimate a set of variance components. For the model in equation (4.5), a linear time search algorithm based on maximum likelihood exists, which is able to identify the optimal variance components. However, no linear time method exists to find the three variance components required for the model in equation (4.8). Therefore, we utilize an approach suggested by [LKS10], in which we use the EMMAX algorithm inside of a linear time search.

First we rewrite the variance of $\mathbf{y}$ as shown in equation (4.11), letting $\tau^2 = \sigma_v^2 - \sigma_g^2$ and $w = \sigma_g^2/(\sigma_v^2 - \sigma_g^2)$ Then, given a value w between zero and one, we apply the EMMAX search algorithm to find optimal variance components τ^2 and σ_ϵ^2. If we search q different values of w, then our approach will be q times slower than EMMAX. More specifically, EMMAX has a one time computational cost of $O(N^3)$ followed by a cost of $O(rN)$ for r search iterations, where N is the total number of measurements or the size of the vector $\mathbf{y}$. In comparison, our approach will have a one time cost of $O(qN^3)$ followed by a cost of $O(qrN)$. This is compared to the basic Newton-Raphson technique, which has a total computation cost of $O(qrN^3)$.

$$\begin{aligned} var(\mathbf{y}) = &\quad \tau^2(w\mathbf{ZKZ}' + (1-w)\mathbf{D}) + \sigma_\epsilon^2\mathbf{I} \\ = &\quad \tau^2\mathbf{K}^* + \sigma_\epsilon^2\mathbf{I} \end{aligned} \tag{4.11}$$

Additional cost will be incurred if $\mathbf{D}$ has to be estimated, such as in the case when an auto-regressive model is utilized. For example, in the case where $\mathbf{D}$ is determined through an auto-regressive model, the total computational time would be multiplied by the size of the search space for the additional auto-regressive parameter. If the number of iterations required to optimize this parameter were $O(p)$, then the total computational cost will be $O(pqN^3 + qrN)$.

4.2.7 Predicting Lifestyle Values and Genetic Influence

The realization of $\mathbf{u}$ is a vector of values representing the genetic contribution to the phenotype measurement. That is, $\mathbf{u}$ is a random variable and at the time the phenotype was measured a value for $\mathbf{u}$ was sampled from a multivariate distribution. This is the realized value. Just as fixed effects are estimated, the value of a random variable can be predicted using the best linear unbiased predictors (BLUPs) introduced by Henderson [Hen50]. The BLUP for $\mathbf{u}$, denoted by $\tilde{\mathbf{u}}$, is given by

$$\tilde{\mathbf{u}} = \sigma_g^2 \mathbf{K}\mathbf{Z}'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}_r\hat{\beta}) \tag{4.12}$$

where $\boldsymbol{\Sigma}$ is given by equation (4.10). The realized value of $\mathbf{u}$ represents the overall genetic contribution to the phenotype for each individual. By determining the value of u_i for each individual, we are able to determine what proportion of each phenotypic measurement is due to genetics and what proportion is due to other factors, specifically fixed effects and error. This enables us to compare individuals based on the magnitude of their genetic contributions. Certain individuals may have a stronger genetic effect than others. When this large genetic effect causes phenotypes to become harmful, such as in high cholesterol, we may see this difference as indicator of increased risk. When large genetic effects lead to beneficial phenotypes, this may indicate a sort of genetic robustness, a phenomenon often referred to colloquially as having "good genes". Therefore, we refer to the realized value of $\mathbf{u}$ as the genetic influence.

Just as it is possible to predict $\mathbf{u}$, it is also possible to predict the realized value of $\mathbf{v}$ from equation (4.8). The realized value of $\mathbf{v}$ for each individual represents the environmental contribution to the phenotype. By comparing realized values of $\mathbf{v}$ it is possible to uncover differences in individual environment, which may be an indicator of an individual's lifestyle. For this reason, we refer to the realized value of $\mathbf{v}$ as the

vector of lifestyle values. When calculating the realized values for $\mathbf{u}$ and $\mathbf{v}$, we are able to partition each phenotype measurement into genetics, environment, fixed effects and error, and give the proportion that each factor contributes.

$$\begin{bmatrix} \hat{\beta} \\ \tilde{\mathbf{u}} \\ \tilde{\mathbf{v}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} & \mathbf{X}' \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{K}^{-1}\alpha_1 & \mathbf{Z}' \\ \mathbf{X} & \mathbf{Z} & \mathbf{I} + \mathbf{D}^{-1}\alpha_2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \\ \mathbf{y} \end{bmatrix} \tag{4.13}$$

The BLUPs for $\mathbf{u}$ and $\mathbf{v}$ are obtained by solving the system of equations given in equation (4.13) [Hen73, MT05], where $\alpha_1 = \sigma_\epsilon^2/\sigma_g^2$ and $\alpha_2 = \sigma_\epsilon^2/\sigma_v^2$. These solutions are obtained by maximizing the joint likelihood of $\mathbf{y}$, $\mathbf{u}$ and $\mathbf{v}$ under the assumption of normality. After solving the so-called mixed model equations (MME), we obtain a prediction of the random variable v_{ij} for each individual i and time point j, as well as predictions for u_i for each individual i.

Theoretical accuracy of the random effects may be analyzed by evaluating the variance in the difference between the true and predicted effects, which are calculated by

$$var(\tilde{\mathbf{u}} - \mathbf{u}) = \sigma_g^2\mathbf{K} - \sigma_g^2\mathbf{K}\mathbf{Z}'\mathbf{P}\mathbf{Z}\sigma_g^2\mathbf{K} \tag{4.14}$$

$$var(\tilde{\mathbf{v}} - \mathbf{v}) = \sigma_v^2\mathbf{D} - \sigma_v^2\mathbf{D}\mathbf{P}\sigma_v^2\mathbf{D} \tag{4.15}$$

where $\mathbf{P} = \mathbf{\Sigma}^{-1} - \mathbf{\Sigma}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Sigma}^{-1}$.

4.2.8 Analytical Power for Mixed Effects Models

One common approach to evaluate methods for performing association is through the analysis of statistical power [BYP05]. Although power is easily calculated analytically when assuming the model from equation (4.2), it is not well-known how to calculate power when using a mixed effects model. For this reason, time consuming simula-

tions are often employed in order to estimate statistical power [BFO10a]. [WB99] introduced a likelihood-ratio based technique to compute power in variance component models used for linkage analysis. We introduce a similar derivation based on the F-test.

Let $\mathbf{y}$ be a vector of size n and assume that it has a normal distribution with mean $\mathbf{X}\beta$ and variance $\mathbf{\Sigma}$, where $\mathbf{X}$ is an $n \times q$ matrix of fixed effects, β is a $q \times 1$ vector of coefficients and $\mathbf{\Sigma}$ is an $n \times n$ covariance matrix. In order to test a hypothesis about β, we define a $q \times 1$ matrix $\mathbf{R}$, which defines a linear combination of the elements of β. For example, if $\mathbf{X}$ only encodes global mean and SNP, then we define $\mathbf{R} = [0 \quad 1]$, so that $\mathbf{R}\beta$ results in the single SNP coefficient. Given $\mathbf{R}$ we define the hypothesis test $\mathbf{R}\beta = \mathbf{r}$. The generalized least squares (GLS) F-statistic is then given by

$$\phi_F = \frac{(\mathbf{R}\hat{\beta} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\beta} - \mathbf{r})}{q} \tag{4.16}$$

We show that under the alternative hypothesis $\mathbf{R}\beta = \mathbf{r} + \delta$, ϕ_F follows an F-distribution with $n - q$ numerator and q denominator degrees of freedom and non-centrality parameter λ (see Appendix), given by

$$\lambda = \quad \delta'[\mathbf{R}(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{R}']^{-1}\delta \tag{4.17}$$

It is important to note that we have assumed an optimal estimate for the true covariance matrix $\mathbf{\Sigma}$. Given the non-centrality parameter in equation (4.17), power is calculated as the area under the curve of the distribution defined by the non-centrality parameter that is beyond the null rejection region. We expound upon the details of these calculations in what follows.

Consider that $\hat{\beta} \sim N(\beta, (\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X})^{-1})$, from equation (4.16). Now consider that the hypothesis test $\mathbf{R}\beta = \mathbf{r}$, while in truth $\mathbf{R}\beta = \mathbf{r} + \delta$ ($\mathbf{r} = \mathbf{R}\beta - \delta$). In order to

derive, a chi-square statistic, we first derive a Z-score statistic for the test $\mathbf{R}\beta = \mathbf{r}$.

$$\begin{aligned} Z = & \quad [\mathbf{R}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{R}']^{-1/2}(\mathbf{R}\hat{\beta} - \mathbf{r}) \\ = & \quad [\mathbf{R}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{R}']^{-1/2}(\mathbf{R}\hat{\beta} - \mathbf{R}\beta + \delta) \\ = & \quad [\mathbf{R}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{R}']^{-1/2}(\mathbf{R}(\hat{\beta} - \beta) + \delta) \end{aligned}$$

We know the distribution of $\hat{\beta}$ and therefore $\mathbf{R}\hat{\beta}$, thus

$$[\mathbf{R}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{R}']^{-1/2}(\mathbf{R}(\hat{\beta} - \beta)) \sim N(0, I)$$

$$[\mathbf{R}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{R}']^{-1/2}(\mathbf{R}(\hat{\beta} - \beta) + \delta) \sim N([\mathbf{R}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{R}']^{-1/2}\delta, I)$$

Squaring Z, we obtain a χ^2 statistic. Let $\mathbf{W} = [\mathbf{R}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{R}']$.

$$\begin{aligned} \phi_c = & \quad (\mathbf{R}(\hat{\beta} - \beta) + \delta)'\mathbf{W}^{-1}(\mathbf{R}(\hat{\beta} - \beta) + \delta) \\ = & \quad (\mathbf{R}\hat{\beta} - \mathbf{r})'\mathbf{W}^{-1}(\mathbf{R}\hat{\beta} - \mathbf{r}) \sim \chi^2(q; \delta'\mathbf{W}^{-1}\delta) \end{aligned} \tag{4.18}$$

Thus ϕ_c is a χ^2 statistic with q degrees of freedom and a non-centrality parameter of $\delta'\mathbf{W}^{-1}\delta$. Now consider that $\boldsymbol{\Sigma}$ is actually unknown and that we will use an estimate $\hat{\boldsymbol{\Sigma}}$, such that $\boldsymbol{\Sigma} = \sigma_c^2\hat{\boldsymbol{\Sigma}}$, where σ_c^2 is an unknown scalar. Given this, we know that $(n-q)\hat{\sigma}_c^2/\sigma_c^2 \sim \chi^2(n-q)$ and may obtain the following statistic by dividing ϕ_c by this quantity.

$$\begin{aligned} \phi_F = & \quad \frac{(\mathbf{R}\hat{\beta} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\beta} - \mathbf{r})}{\hat{\sigma}_c^2 q} \\ \phi_F \sim & \quad \mathcal{F}(q, n-q, \delta'\mathbf{W}^{-1}\delta), \end{aligned} \tag{4.19}$$

We use $\mathcal{F}(df_1, df_2, ncp)$ to represent the non-central $\mathcal{F}$-distribution with numerator degrees of freedom df_1 and denominator degrees of freedom df_2 and non-centrality parameter ncp. With optimal variance component estimates, we expect that $\hat{\sigma}_c^2 = 1$.

4.2.9 Simulations

The power gain was estimated by generating random covariance matrices for the environmental components and then by averaging analytical power over 1000 randomly selected SNPs with minor allele frequencies in the range of 1% to 5%. The expected power gain was calculated as the average power gain over 1000 such randomly generated covariance matrices. The covariance matrices were generated by randomly selecting a vector of size $m-1$ from a uniform(0,1) distribution, where m is the number of time points. The covariance between time points i and j ($i > j$) is then given as the $(i-j-1)$th entry in this vector. We then define an $m \times m$ matrix $\mathbf{E}$ using this scheme and define the full environmental covariance matrix $\mathbf{D} = \mathbf{E} \otimes \mathbf{I}$.

Phenotypes with multiple time points were generated by sampling both genetic and environmental components from their respective multivariate normal distributions, having a mean of zero and with variance as specified in equation (4.9). This results in both a genetic and environmental contribution value for each individual. These values are used as the individual's mean phenotype value to which random noise is added to generate each time point.

Random effects were predicted by fitting the model from equation (4.8), with only a mean effect, and then by obtaining the solution to equation (4.13). The environmental covariance matrix was estimated by calculating the correlation between time points using all individuals. This procedure works well, when either the genetic effect is small or the sample has little population structure. In the case where the population has a large amount of structure, we employ a simple iterative scheme. Starting with an estimate of the environmental covariance matrix calculated on the raw data, we predict the random genetic effect. This effect is regressed from the phenotype values and a new environmental covariance matrix is computed with these new phenotype values. This procedure is repeated until the environmental covariance matrix converges.

4.3 Results

4.3.1 Multiple measurements provide increased power over traditional approaches

We evaluated the gain in power achieved when using the proposed method (the full method) over using an averaging approach or single approach. In the single approach, a single time point is selected for each individual, while in the average approach time point values are averaged for each individual. Power gain is evaluated by comparing the ratios of the power achieved with one method to that of the power achieved with the single approach. Figure (1) summarizes these results. Power gain was calculated for each effect size by averaging the power gain over 1000 iterations, in which a different randomly selected environmental covariance matrix was used in the analytical calculation of power. This power was averaged over 1000 SNPs with minor allele frequency in the range of 1% to 5%. All calculations assume that the environment accounts for 80% of the phenotypic variance while both the genetic background and residual error account for 10%.

Figure (1) compares both the power curves (figures (1a) and (1c)) and power gain (figures (1b) and (1d)) for 1000 and 2000 individuals randomly selected from the Wellcome Trust Case Control Consortium (WTCCC). We see that on average the full method has increased power when compared to the average and single methods. This increased power is seen more clearly in the power gain plots, which show that the full method has as much as an 8-fold gain in power over the single approach, compared to a roughly 4.5-fold power gain achieved by the average approach.

4.3.2 Multiple measurements allow for the prediction of individual lifestyle

Another benefit of using the full method over approaches utilizing only one time point is the ability to predict the phenotypic contribution due to environment. We evalu-

ate this ability through simulation. We simulated phenotypes for 1000 individuals in which the environment accounts for 80% of the variance, while genetic and error each contribute 10%. These phenotypes represent those that are largely influenced by environment. For each phenotype, we predict the environmental contribution and the proportion of the phenotype that this value accounts for. More specifically, each phenotype can be seen as a linear combination of mean, genetic contribution, environmental contribution and error. Using random effect predictions, as summarized in the methods, we obtain predicted values for the genetic and environmental contributions at each time point. The prediction of random effects for each time point scales with the cube of the number of individuals times the number of time points. In practice, 1000 individuals and 2 time points requires a running time under 5 minutes for one phenotype, whereas the running time for 1000 individuals with 5 time points is just under 50 minutes. However, we note that recent advances in the computational theory behind linear mixed models can serve to decrease these running times [LLL11].

The results of this simulation are summarized in figure (2). First, we compare the accuracy of the predictions by summarizing the difference in the true proportion of the phenotype contributed by the environmental component with that of the predicted proportion (figure (2a)) for 1000 randomly generated environmental covariance matrices. From this plot, we see that the difference between the true and predicted proportion hovers around 25%, while increasing the number of time points available shifts this mean towards zero. Despite the high variation in accuracy, figure (2b), showing the distribution of correlations between the true lifestyle values with the predicted, shows that on average the predictions have a rank correlation of 0.94. This indicates that although accuracy is not always high, the relative ranking of individuals based on their predicted lifestyle values is highly concordant with their true ranking. This implies that individuals may be effectively ranked based on their predicted lifestyle values.

Figure (3) shows the accuracy and correlations for predicted genetic values. We find that the predicted proportions behave very similarly to that of the lifestyle values, except that the availability of additional measurements does not increase accuracy in this case. However, the correlation between the true and predicted genetic effects is on average very small and has a strange pattern. Although, we find that as the number of time points increases, the average correlation increases.

4.4 Discussion

In this chapter, we introduce a mixed model based approach to perform association mapping in GWAS, when multiple measurements are available. We show that by utilizing multiple measurements, our method achieves increased power over methods that either select a single measurement or average measurements for individuals. Furthermore, we show that when multiple measurements for each individual are available, it is possible to differentiate the genetic contribution from the environmental contribution. We call these quantities the genetic influence and lifestyle values, respectively.

The ability to partition a phenotype into its constituents may be useful for future phenotype prediction and treatment selection. For example, some individuals might gain substantially from a decrease in dietary cholesterol when the largest part of their cholesterol is due to their intake. On the other hand, some individuals who are genetically predisposed to high cholesterol, might stand to gain little from decreasing their dietary cholesterol, but instead might require medication in order to alter their overall cholesterol levels. With knowledge of the individual contributions to total cholesterol, the appropriate treatment options may be put in place in order to alter the future of the phenotype.

The previous cholesterol example extends very naturally to explain how prediction

of genetic influence and lifestyle may be useful for risk prediction. For example, it may be discovered through these methods that an individual has a very large lifestyle component for cholesterol. In this case, their risk for developing cardiovascular disease may be predicted based on the magnitude and direction of this value, such that the resulting prediction may be different from that obtained by using the total level of cholesterol.

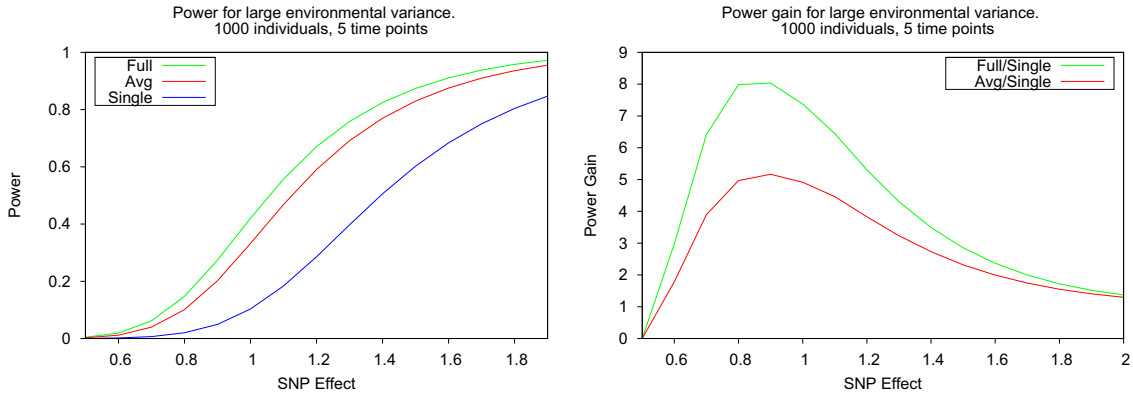
Another interesting aspect of the ability to predict lifestyle values is the subsequent ability to categorize individuals. For example, when evaluating a trait such as lung capacity, certain individuals will have decreased lung capacity due to long term smoking, while others will have relatively normal capacity given their age and genetic makeup. This might be easily discerned by using a series of questions, but it is well-known that the truth is not always told when answering such questions. In this case, the lifestyle value may help to categorize individuals based only on their phenotype measurements and genotypes.

The model we propose is tested under a certain set of assumptions, however the structure is very general and may work for a larger class of problems. For example, there are cases when it is not reasonable to assume that a phenotype follows a normal distribution, and a simple transformation such as log is not sufficient to obtain a normally distributed measure. For example, binary or categorical outcomes cannot be expected to follow a normal distribution. In this case, the phenotype may be modeled using a link function, such is done in logistic regression [MS01]. The models presented here may then be utilized in this space. Furthermore, there may be other factors to include in the model, such as gene-by-environment interactions, or even more complicated treatment-genetic-enviromental interactions. Perhaps some individuals have increased variance when subjected to certain environments and certain treatments but not with others. The model presented in this work can be used as a base to explore such

conditions, which may require more complex models with additional random effects.

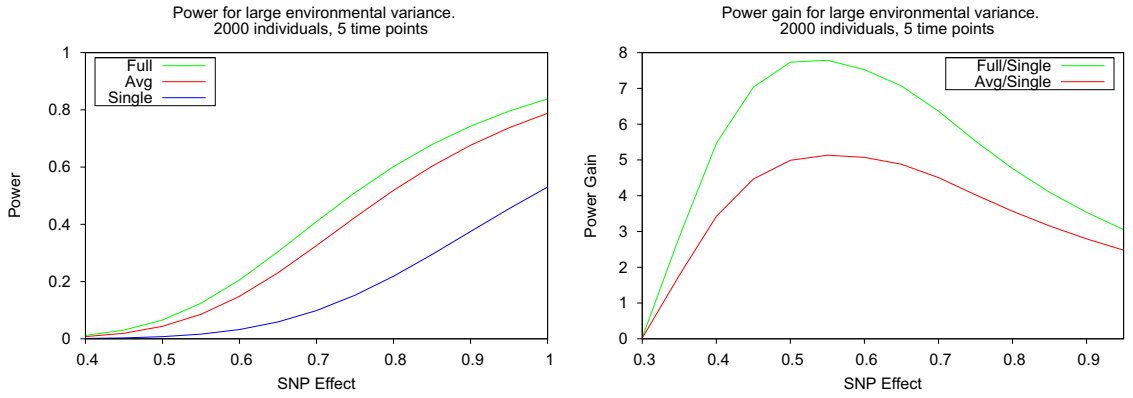
Reference to published article

Furlotte, Nicholas A, Eleazar Eskin, and Susana Eyheramendy. 2012. Genome-Wide association mapping with longitudinal data. Genetic Epidemiology. doi:10.1002/gepi.21640.



(a) Average Power - 1000 Individuals

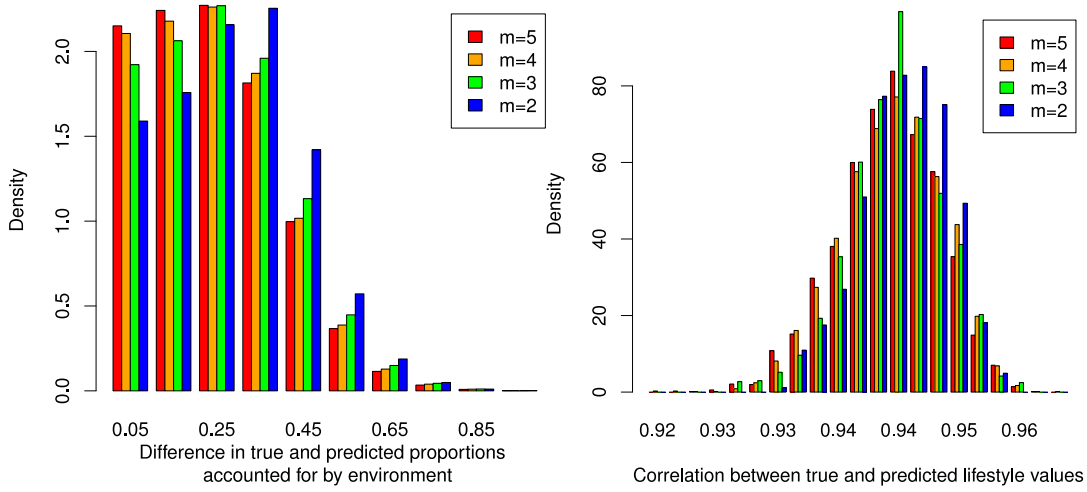
(b) Average PowerGain - 1000 Individuals



(c) Average Power - 2000 Individuals

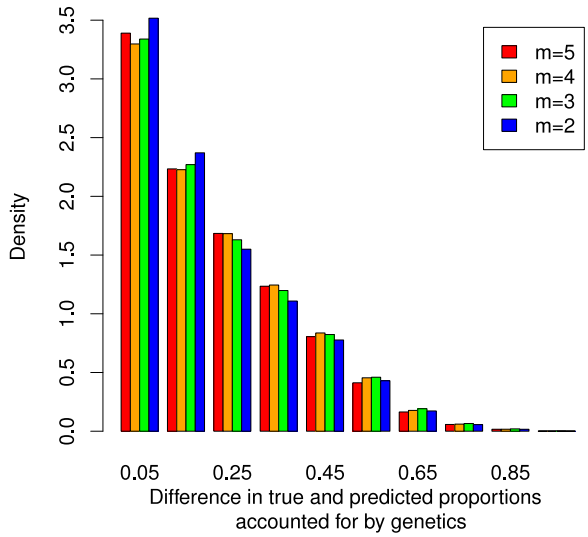
(d) Average PowerGain - 2000 Individuals

Figure 4.1: **Association mapping utilizing multiple measurements leads to an increase in power over traditional approaches.** We compare the average power gain for the proposed full model with that of the average model (using averaged measurements for each individual). Power gain is defined as the ratio of the power of a given method to that achieved with the single approach (i.e.. mapping with only one measurement for each individual) and was calculated by averaging power gain over 1000 randomly selected SNPs with MAF in the range of 1% to 5% and over 1000 randomly selected covariance structures for the multiple measurements ($m = 5$). Simulations were performed with the environmental effect accounting for 80% of the variance while the genetic background and residual error accounted for the remaining 20%.

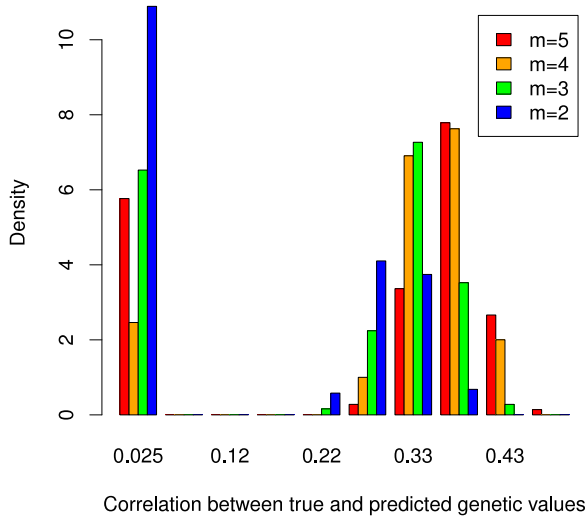


(a) (b)

Figure 4.2: **The accuracy in prediction of lifestyle values varies, while the ranking remains consistent.** For each of 1000 iterations, lifestyle values were predicted and compared with their known true values, through simulation. Figure (4.2a) evaluates the difference between the proportion accounted for by the environment as determined by the true lifestyle effect with that of the predicted lifestyle effect. This result indicates that the accuracy of these predictions has a high variation, but that by increasing the number of time points it is possible to obtain more accurate predictions . Figure (4.2b) shows the distributions of Spearman rank correlations between the true lifestyle and predicted lifestyle values. This result indicates that the ranking of individuals based on their predicted lifestyle is highly concordant with the true lifestyle ranking.



(a)



(b)

Figure 4.3: **The accuracy in prediction of genetic values is similar to that of lifestyle.** For each of 1000 iterations, genetic values were predicted and compared with their known true values, through simulation. Figure (4.3a) shows a very similar result to that found in lifestyle values, where the accuracy of these predictions has a high variation and has a relatively uniform distribution across different numbers of time points. However, the result of figure (4.3b) is much different than that found in the lifestyle value prediction. There is not a clear pattern, although the average correlation does increase as the number of time points increases.

CHAPTER 5

Efficient Multiple Trait Association with the Matrix-variate Linear Mixed-model

5.1 Background

Classically, genome-wide association studies have been carried out using single traits. However, it is well-known that genes often affect multiple traits, a phenomenon known as pleiotropy, and more recently, it has been shown that performing association mapping with multiple traits simultaneously may increase statistical power [KRI01, FP09, LPL09, AHN11, KVS12]. Analysis of multiple pleiotropic phenotypes increases power because intuitively, multiple phenotype measurements increase sample size relative to a single phenotype. However, utilizing the additional data is not straightforward as measurements from the same individual are not independent. This issue is analogous to that of association analysis in cohorts of related individuals, where phenotype measurements between related individuals are not independent. Variance component methods model this correlation structure by assuming that the covariance due to genetics between related individuals is proportional to their kinship coefficient [KYE08]. This constant of proportionality normalized by the total trait variance is related to narrow-sense heritability of the trait (the variance accounted for by additive genetic effects) [YBM10].

When the same genetic variants affect multiple traits, phenotype values for an indi-

vidual will tend to be correlated. Similarly, shared environmental effects also introduce some level of correlation between traits. A fundamental problem in understanding the relationship between the traits is determining the proportion of the total correlation due to genetics and the proportion due to environment. Classical approaches originating from animal breeding and agricultural research solve this problem by modeling the statistical relationship between traits using a linear mixed-model (LMM) [Fal81, MT05]. These approaches decompose the between trait correlation into both a genetic component and an environmental component and then use the LMM framework to obtain estimates for these quantities. The LMMs used in these classical approaches can be adapted for use in GWAS by utilizing them to test the association between genetic variants and multiple traits. Multiple phenotype variance component methods closely follow the approach utilizing kinship values to model the covariance between different phenotypes among different individuals, such that the genetic covariance between two individual's phenotypes is proportional to their kinship coefficient [HQ76]. In this case, the constant of proportionality is a function of the two trait heritabilities as well as the genetic correlation. Similarly, multiple trait models represent the covariance between phenotypes within an individual as a function of both genetics and shared environment.

In order to utilize LMMs for association analysis, an iterative procedure must be employed to identify the maximum-likelihood parameters of the statistical model used for association. The use of LMMs for single traits has been limited by the computational complexity of traditional maximum-likelihood procedures: $O(n^3 \cdot t)$, where n is the number of individuals in the study and t is the number of iterations necessary for the maximum-likelihood algorithm to converge. However, recently developed estimation algorithms have made LMMs computationally efficient and feasible for large population cohorts [KYE08, KSS10, LLL11, ZSZ12], reducing the computational complexity of from $O(n^3 \cdot t)$ to $O(n^3 + n \cdot t)$. These approaches have had a major

affect in enabling association mapping for single traits using LMMs. Unfortunately, the previous approaches [KYE08, KSS10, LLL11, ZSZ12] cannot be directly applied to multiple trait LMMs, meaning that the same computational inefficiencies that limited the widespread use of LMMs for single trait GWAS, now hinder the scale at which researchers can perform multiple trait GWAS. More specifically, with p traits measured over n individuals the running time for classical multivariate LMMs is $O(n^3p^3 \cdot t)$. In other words, even when p is small (eg. $p = 2$), the running time scales as the cube of the number of individuals in the sample, meaning that the use of multiple trait LMMs is not feasible for large sample sizes.

In this chapter, we introduce a formulation of the multiple trait linear mixed-model for use in association mapping and show that it provides a significant speed up over the classical approach. We define a statistical model relating multiple correlated traits to genetic variations based on the matrix-variate normal distribution [GN00]. Using this formulation, we show how a simple data transformation leads to a model equivalent to the classical model while allowing maximum-likelihood inference to be performed in computational time essentially linear in the size of the data set, given a one time cost of $O(n^3)$ and $O(n^2)$. In a simple case, let us assume that p is much less than n (eg. 2 vs. $10,000$) and that we only have a global mean for each phenotype; this leads to a total computational complexity of $O(n^3 + n^2p + (p^3(n + 1)) \cdot t)$. The iterative part of the algorithm is then essentially linear in the size of the dataset. We call our method the matrix-variate linear mixed-model (mvLMM) and show its efficacy by analyzing correlated phenotypes in the Northern Finland Birth Cohort [Sab08]. Comparing to a standard approach [LYG12], we show that our method results in more than a 10-fold time reduction for a pair of correlated traits, taking the analysis time from about 35 minutes to about 2.5 minutes for the cubic operations plus another 12 seconds for the iterative part of the algorithm. In addition, the cubic operation can be saved so that it does not have to be re-calculated when analyzing other traits in the same cohort.

Finally, we demonstrate how this method can be used to analyze gene expression data. Using a well-studied yeast dataset [SK08], we show how estimation of the genetic and environmental components of correlation between pairs of genes allows us for to understand the relative contribution of genetics and environment to coexpression.

5.2 Results

5.2.1 Association and genetic correlation in the Northern Finland Birth Cohort

5.2.1.1 Association

We apply our method to the Northern Finland Birth Cohort, a founder cohort consisting of 5,043 individuals each of which has multiple phenotype measurements for four different metabolic phenotypes. We analyze a total of six pairs of traits or all combinations of four traits: HDL and LDL cholesterol, C-reactive protein (CRP) and triglycerides (TG). Association between each SNP and each pair of phenotypes is evaluated by assuming that under the null hypothesis the SNP does not effect either phenotype. This same data set was analyzed by Korte et al. (2012) using a classically-based multiple trait LMM. We compare our results with their multi-trait mixed model (MTMM) method and find that the results are highly concordant, indicating that our method is consistent with classical approaches.

Over 99% of associations identified in marginal analysis are also identified when respective pairs of traits are mapped (significance threshold of 1.5e-7). However, the joint mapping uncovers more significant associations; 19 new associations are identified across all trait pairs. For example, in the analysis of TG with CRP, we identify a SNP (rs2000571) with a p-value of 8.58e-7 and with MTMM p-value of 1.7e-6. This SNP was not significant in the marginal analysis of TG (1.7e-5) or CRP (0.03), but

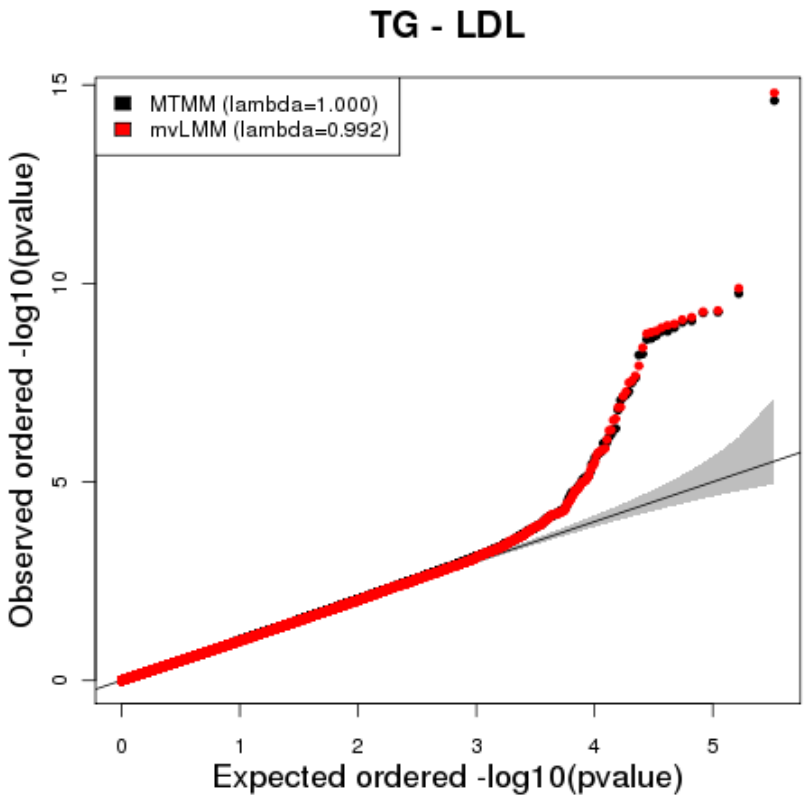


Figure 5.1: **QQ Plot comparing MTMM and mvLMM p-values obtained when performing analysis with LDL and TG.**

belongs to a region on chromosome 11 that has been shown to harbor variants contributing to triglycerides [BBS12]. For all pairs of traits, we find that the genome-wide p-values obtained using our method are highly correlated with those obtained by Korte et al. (r =0.96 - 0.99), as shown in figure (5.1) comparing QQ-plots. In addition, we identify 5 associations deemed as significant in the marginal analysis but that were no longer significant in the MTMM. These results indicate that our approach is consistent with a classical approach.

5.2.1.2 Genetic Correlations

In multiple trait models, the total trait correlation is partitioned into a genetic and an environmental component. The genetic component of the correlation (the genetic correlation) represents the part of the total trait covariance that is attributed to genetics normalized by the genetic variances. This quantity provides insight into the genetic architecture of the relationships between traits. We estimate the genetic correlations for each pair of traits analyzed in the Finland Birth Cohort and compare these estimates with those obtained using a standard implementation of a bi-variate LMM as implemented in GCTA [LYG12].

Table 5.1 compares estimates of genetic correlation obtained with GCTA and mvLMM. The utility of the genetic correlation as a parameter providing insight into the genetic architecture of trait correlation is illustrated in these results. For example, the phenotype pair HDL and TG have a total correlation of -0.19, while the estimated genetic correlation is fairly strongly positive (0.28). This result implies that HDL and TG are under the control of many of the same genetic loci, which is consistent with previous studies [SOC00]. However, since HDL serves as a protective mechanism to regulate TG, an environmental perturbation causing HDL to go down would cause an increase in TG and thus a negative environmental correlation. When the environmental correlation contributes a larger proportion to the total, this yields an overall negative correlation. When we compare our results to those of GCTA we find that the two methods yield similar results, with genetic correlation estimates falling less than one standard deviation from one another. In addition, the running time for the classical approach was around 35 minutes, while the running time for mvLMM was on average roughly 12 seconds, given a one time cost of 2.5 minutes shared across pairs of traits.

Phenotype Pair	Phenotypic Correlation	mvLMM Genetic Correlation	GCTA Genetic Correlation
HDL/CRP	-0.19	0.28 ± 0.19	0.26 ± 0.22
HDL/LDL	-0.13	-0.16 ± 0.11	-0.18 ± 0.11
HDL/TG	-0.37	-0.37 ± 0.17	-0.32 ± 0.16
LDL/CRP	0.09	0.03 ± 0.17	-0.02 ± 0.17
TG/CRP	0.21	-0.62 ± 0.26	-0.75 ± 0.41
TG/LDL	0.32	0.33 ± 0.16	0.29 ± 0.14

Table 5.1: **Genetic correlation estimates in the Finland Birth Cohort.** We compare the maximum likelihood estimates obtained with mvLMM with those obtained with GCTA and find that the results are very similar.

5.2.2 Bi-variate analysis in yeast data

Gene coexpression, defined as the correlation between expression levels of a pair of genes estimated in a set of individuals, is a fundamental quantity that has been utilized for a variety of applications [LPD06, GDZ06, STM05, SSK03]. There are two prevalent views about the meaning of significant coexpression. The first is that coexpression stems from similar environmental conditions such as disease status [HSC97]. The second comes from the systems genetics literature where it is thought that coexpressed genes have a similar genetic regulatory program and that specific genetic variants drive modules of coexpressed genes [LPD06, GDZ06]. However, correlation estimates from gene expression levels measures the combined effect of both the genetic and environmental components. Our methodology allows for the first time to decompose the coexpression into a genetic and environmental component.

We utilize the major gain in efficiency of our approach to perform an analysis that is not feasible with current methods. Using a well-studied yeast dataset [SK08] consisting of 109 yeast strains each with 5793 gene expression measurements, we perform a bi-variate analysis, estimating genetic correlations for all 5793 choose 2 gene expression pairs. Within this dataset several regions of the genome have been implicated to harbor genetic variation that affects many gene expression levels.

Using a set of hotspot locations derived from [SK08], we define a set of 13,508 hotspot gene pairs by extracting all pairs of genes that lie in each known hotspot. We then compare the phenotypic correlation to the total proportion of covariation accounted for by genetics for each of these pairs. Assuming that hotspot pairs are under the same genetic regulation, we expect that the phenotypic correlation for any given pair should reflect this by having a high value. However, this might not be the case if the environmental correlation between the pair contributes in such a way to lower the overall phenotypic correlation. Therefore, an estimation of the total phenotypic co-

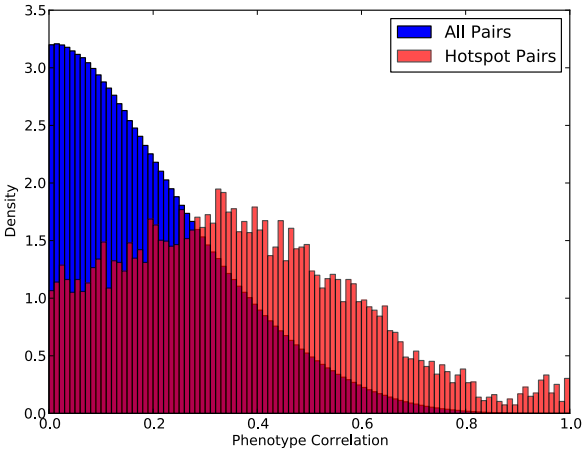
variation attributed to genetics may better reflect the fact that the two genes are under the same genetic program.

In Figure (5.2a), we plot the histogram of the absolute value of the total phenotypic correlation for all gene pairs and for hotspot gene pairs. We see that the distribution of phenotypic correlations for hotspot pairs is shifted towards higher correlations with respect to all pairs, giving an indication of co-regulation. However, most of the pairs have correlation less than 0.5. Figure (5.2b) shows the same plot generated using the total proportion of the phenotypic covariation attributed to genetics. In the figure, we observe that the estimated genetic covariation for hotspot pairs is dramatically skewed towards one. In fact, most of the pairs have a genetic covariance above 0.7. This result suggests that the estimated genetic correlations on average give a stronger indication of co-regulation compared to the phenotypic correlation.

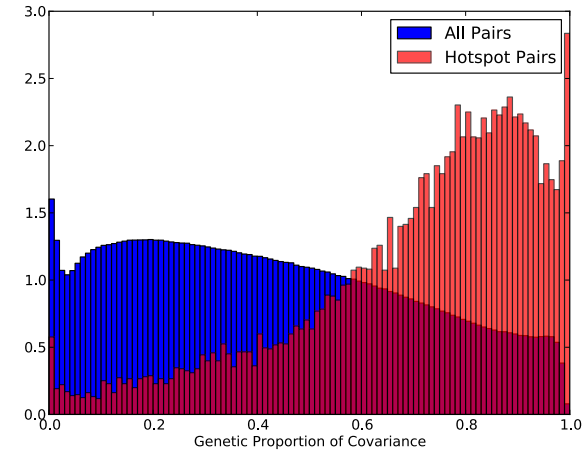
5.3 Discussion

In this chapter, we introduced a method for performing multi-trait genome-wide association analysis and for the estimation of the genetic correlation. Our method is based in classical theory, but introduces a computational advance that makes it much faster, reducing running time over 10-fold when compared with the classic approach. We have shown that our method achieves similar results to that of the classical approach. In addition, we have shown that the ability to quickly estimate genetic correlation may be of great benefit to researchers, leading to fundamental insights into the architecture of complex traits.

The ability to quickly optimize multiple trait linear mixed-models will have a large impact on the the ability to dissect complex traits. For example, multiple expression quantitative trait loci (multi-eQTL) may be discovered by mapping multiple traits to



(a) Phenotypic



(b) Genetic

Figure 5.2: **Comparison of the phenotypic correlation with the total proportion of the correlation accounted for by genetics for all gene pairs and for gene pairs from regulatory hotspots.** We compare the phenotypic correlation with the total proportion of correlation accounted for by genetics in order to assess the ability of the genetic correlation to differentiate gene pairs that are co-regulated. Utilizing a set of known hotspots, we derive a set of hotspot gene pairs, where a hotspot pair is defined as a gene pair in which both genes lie in a given hotspot. We find that the genetic correlation differentiates these co-regulated pairs better than the overall phenotype correlation.

genetic variants across the genome. The ability to perform this type of research is infeasible with current methodologies. In addition, we have shown that the genetic correlation between gene expression measurements may be a better indicator of co-regulation. It stands to reason that these genetic correlations may be used in coexpression analysis and lead to the discovery of gene modules that are truly co-regulated and not in part due to environmental correlations.

5.4 Methods

5.4.1 Modeling multiple phenotypes with the matrix-variate linear mixed-model

Given a set of p phenotypes for n individuals, let y_{ij} represent the value of the ith phenotype for the jth individual. A standard statistical model for the ith phenotype vector, denoted by $\mathbf{y}_i$, is given by the following linear mixed model (LMM), were $\mathbf{X}\beta_i$ represents the mean term for the ith phenotype such that $\mathbf{X}$ is an $n \times q$ matrix encoding covariates including SNP, $\mathbf{g}_i$ represents the population structure or genetic background component and $\mathbf{e}_i$ represents the effect due to environment and error. We have assumed that the covariates determining the mean will be shared among phenotypes, but this is not a requirement.

$$\mathbf{y}_i = \mathbf{X}\beta_i + \mathbf{g}_i + \mathbf{e}_i \tag{5.1}$$

The variance of $\mathbf{y}_i$ is given by the following, assuming that $cov(\mathbf{g}_i, \mathbf{e}_i) = 0$.

$$var(\mathbf{y}_i) = var(\mathbf{g}_i) + var(\mathbf{e}_i) \tag{5.2}$$

$$= \sigma^2_{g(i)}\mathbf{K} + \sigma^2_{e(i)}\mathbf{I} \tag{5.3}$$

where $\sigma^2_{g(i)}$ represents the genetic variance component for phenotype i, $\mathbf{K}$ represents the $n \times n$ kinship matrix calculated using a set of m known variants and $\sigma^2_{e(i)}$ represents the environmental and error variance. Assuming the models proposed by Henderson [HQ76, MT05], it then follows that the covariance between measurements for individuals j and k for phenotype i is given by $\sigma^2_{g(i)} K_{jk}$. By letting ρ_{im} represent the correlation between phenotypes i and m due to genetic effect and letting λ_{im} represent the correlation due to environment, we know that the covariance between the phenotype measurements i and m for individual j is given by the following.

$$cov(y_{ij}, y_{mj}) = cov(g_{ij}, g_{mj}) + cov(e_{ij}, e_{mj}) \tag{5.4}$$

$$= \rho_{im}\sigma_{g(i)}\sigma_{g(m)} + \lambda_{im}\sigma_{e(i)}\sigma_{e(m)} \tag{5.5}$$

Assuming that environmental effects are independent between individuals, let the covariance between phenotypes i and m for individuals j and k be $cov(y_{ij}, y_{mk}) = K_{jk}\rho_{im}\sigma_{g(i)}\sigma_{g(m)}$. We then model the full set of phenotype measurements using a matrix-variate normal distribution.

The matrix-variate normal distribution is a generalization of the multivariate normal distribution to matrices [GN00]. The main idea is that a multivariate normal distribution has three elements: a vector of data of size n, a mean vector of size n and a covariance matrix of size $n \times n$, denoted by $\mathbf{y} \sim N(\mathbf{m}, \mathbf{R})$, where $\mathbf{y}$ is the data, $\mathbf{m}$ is the mean vector and $\mathbf{R}$ is the covariance matrix. The mean vector simple determines where in n-dimensional space the mean of data sampled from this distribution will lie, while the covariance matrix encodes the correlation between individual elements of the data vector. A matrix-variate normal can be thought of as multiple multivariate normal vectors, in which elements between vectors are correlated. This means that when p multivariate data vectors are stored on the columns of an $n \times p$ matrix, the correlation

between elements on the rows of the matrix, will be determined by $\mathbf{R}$, while the correlation between the columns of the matrix will be determined by another parameter, which we will call $\mathbf{C}$. The matrix-variate normal distribution encodes these concepts and we denote such a distribution as $\mathbf{Y} \sim N_{n\times p}(\mathbf{M}, \mathbf{C}, \mathbf{R})$, where $\mathbf{Y}$ is an $n \times p$ data matrix, $\mathbf{M}$ is an $n \times p$ mean matrix, $\mathbf{C}$ is an $p \times p$ column covariance matrix and $\mathbf{R}$ is the $n \times n$ row covariance matrix.

Let $\mathbf{Y}$ represent the $n \times p$ matrix of phenotypes such that

$$\mathbf{Y} = \mathbf{Z} + \mathbf{R} \tag{5.6}$$

where $\mathbf{Z}$ follows a matrix-variate normal distribution with mean $\mathbf{X}\beta = \mathbf{X}[\beta_1 \dots \beta_p]$ and covariance matrices $\mathbf{\Psi}$ and $\mathbf{K}$, where $\mathbf{\Psi}$ is a $p \times p$ matrix representing the correlation between phenotypes due to genetics and $\mathbf{K}$ is the kinship matrix. $\mathbf{R}$ follows a matrix variate normal distribution with mean zero and covariance matrices $\mathbf{\Phi}$ and $\mathbf{I}_n$, where $\mathbf{\Phi}$ is a $p \times p$ matrix representing the covariance between phenotypes due to environment and error. The ith diagonal component of $\mathbf{\Psi}$ is given by $\sigma^2_{g(i)}$ and the i, jth component by $\rho_{ij}\sigma_{g(i)}\sigma_{g(j)}$, and similarly $\Phi_{ij} = \lambda_{ij}\sigma_{e(i)}\sigma_{e(j)}$. The distribution for $\mathbf{Y}$ is then summarized as follows, where $N_{n\times p}(\mathbf{M}, \mathbf{A}, \mathbf{B})$ denotes the matrix variate normal distribution with mean matrix $\mathbf{M}$ and columns and row covariance matrices $\mathbf{A}$ and $\mathbf{B}$.

$$\mathbf{Y} \sim N_{n\times p}(\mathbf{X}\beta, \mathbf{\Psi}, \mathbf{K}) + N_{n\times p}(0, \mathbf{\Phi}, \mathbf{I}_n) \tag{5.7}$$

5.4.2 mvLMM and Bayesian Linear Regression

The standard LMM used in GWAS has been shown to be equivalent to a Bayesian linear regression in which a number of SNPs m are assumed to each have an effect on the phenotype, such that each effect is sampled IID from a normal distribution

[HVG09, LLK12]. By integrating out these effects, one may arrive at a standard LMM using the realized relationship matrix (RRM) as the kinship matrix [GWV09, YBM10]. Here we briefly summarize this result and show how it extends to multiple trait LMMs.

Let us assume that a set of m SNPs each contribute to the background phenotypic variation for phenotype k. Let $\mathbf{W}$ be a $n \times m$ matrix allocating SNP effects to individuals, such that $E[W_{ij}] = 0$ and $var(W_{ij}) = 1$ and assume that the phenotypic effect attributed to SNP j for phenotype k is b_{jk}, so that individual i will have a total effect due to SNP j of $W_{ij}b_{jk}$. We treat the SNP effect as random and assume that each b_{jk} is sampled IID from distribution $N(0, \frac{1}{m}\sigma^2_{g(k)})$. Let $\mathbf{g}_k = \mathbf{W}\mathbf{b}_k$, where $\mathbf{b}_k = [b_{1k} \quad b_{2k} \ldots b_{mk}]'$. Therefore, the variance of $\mathbf{g}_k$ is given by equation (5.9). Thus, LMM-based population structure correction may be viewed as a basic linear model, while treating the SNP effects as random effects.

$$var(g_k) = \frac{\mathbf{W}\mathbf{W}'}{m}\sigma^2_{g(k)} \tag{5.8}$$

$$= \mathbf{K}\sigma^2_{g(k)} \tag{5.9}$$

This framework may be extended to multiple phenotypes by assuming that the correlation between SNP effect vectors has the following form, where $cor(g_{ki}, g_{ji}) = \rho_{ij}$.

$$\begin{bmatrix} \mathbf{b}_i \\ \mathbf{b}_j \end{bmatrix} \sim N\left(\mathbf{0}, \frac{1}{m}\begin{bmatrix} \sigma^2_{g(i)}\mathbf{I} & \rho_{ij}\sigma_{g(i)}\sigma_{g(j)}\mathbf{I} \\ \rho_{ij}\sigma_{g(i)}\sigma_{g(j)}\mathbf{I} & \sigma^2_{g(j)}\mathbf{I} \end{bmatrix}\right) \tag{5.10}$$

To obtain the joint distribution of $\mathbf{g}_i$ and $\mathbf{g}_j$, we apply the following linear transformation.

$$\begin{bmatrix} \mathbf{g}_i \\ \mathbf{g}_j \end{bmatrix} = \begin{bmatrix} \mathbf{W}\mathbf{b}_i \\ \mathbf{W}\mathbf{b}_j \end{bmatrix} = \begin{bmatrix} \mathbf{W} & \mathbf{0} \\ \mathbf{0} & \mathbf{W} \end{bmatrix} \begin{bmatrix} \mathbf{b}_i \\ \mathbf{b}_j \end{bmatrix} \sim \tag{5.11}$$

$$N\left(\mathbf{0}, \frac{1}{m} \begin{bmatrix} \sigma^2_{g(i)}\mathbf{W}\mathbf{W}' & \rho_{ij}\sigma_{g(i)}\sigma_{g(j)}\mathbf{W}\mathbf{W}' \\ \rho_{ij}\sigma_{g(i)}\sigma_{g(j)}\mathbf{W}\mathbf{W}' & \sigma^2_{g(j)}\mathbf{W}\mathbf{W}' \end{bmatrix}\right) \tag{5.12}$$

$$= N\left(\mathbf{0}, \begin{bmatrix} \sigma^2_{g(i)}\mathbf{K} & \rho_{ij}\sigma_{g(i)}\sigma_{g(j)}\mathbf{K} \\ \rho_{ij}\sigma_{g(i)}\sigma_{g(j)}\mathbf{K} & \sigma^2_{g(j)}\mathbf{K} \end{bmatrix}\right) \tag{5.13}$$

This result is consistent with the proposed model in the previous section.

The same basic logic is easily applied to derive the $cov(\mathbf{e}_i, \mathbf{e}_j)$. By substituting $\mathbf{W}$ for $\mathbf{I}$ as well as the appropriate variance and correlation parameters we arrive at the equivalent result for the correlation between residuals, given in the equation below.

$$\begin{bmatrix} \mathbf{e}_i \\ \mathbf{e}_j \end{bmatrix} \sim \quad N\left(\mathbf{0}, \begin{bmatrix} \sigma^2_{e(i)}\mathbf{I} & \lambda_{ij}\sigma_{e(i)}\sigma_{e(j)}\mathbf{I} \\ \lambda_{ij}\sigma_{e(i)}\sigma_{e(j)}\mathbf{I} & \sigma^2_{e(j)}\mathbf{I} \end{bmatrix}\right) \tag{5.14}$$

We note that a similar analysis may be applied when the two sets of causal SNPs are different for each phenotype. In this case, the between phenotype genetic covariance will be proportional to the $\mathbf{W}_c\mathbf{W}_c'$, where $\mathbf{W}_c$ represents the $n \times t$ SNP incidence matrix for causal SNPs that are common between the two phenotypes. If this matrix deviates significantly from the full kinship matrix $\mathbf{K}$, then it is possible that the estimated genetic correlation may be biased.

5.4.3 Efficient Maximum Likelihood Computation

In this section, we explain how to efficiently compute the log-likelihood of a matrix of phenotypes using a straightforward transformation. This section is split into two subsections: the first, giving a higher-level explanation that illustrates the basic idea

while minimizing technical details and the second, giving a more detailed overview including much technical detail. There is overlap between the two sections as a result.

A simple explanation

Likelihood evaluation for the matrix-variate distribution given by equation (5.7) is accomplished by evaluating the equivalent multivariate normal distribution. By using the $vec()$ operator, which creates a vector from a matrix input by concatenating the columns of the matrix, we are able to represent the distribution given in equation (5.7) in the following way, where $\otimes$ represents the Kronecker product of two matrices.

$$vec(\mathbf{Y}) \sim N_{np}(vec(\mathbf{X}\beta), \mathbf{\Psi} \otimes \mathbf{K} + \mathbf{\Phi} \otimes \mathbf{I}_n) \tag{5.15}$$

The likelihood computation for this model takes time on the order of $(np)^3$. This computational time becomes prohibitive when maximizing the likelihood function while considering a large cohort with multiple phenotypes. Previous work has shown how similar multivariate models with kronecker product matrices can be utilized efficiently when residual errors are independent [SIL11]. However, it is not known how these models may be used efficiently when residual errors are correlated, which is the case for our model. To remedy this problem, we introduce a transformation that results in a reduced computational time.

Let the eigendecomposition of $\mathbf{K} = \mathbf{H}_K \mathbf{S}_K \mathbf{H}'_K$. This decomposition is calculated with a computational complexity of $O(n^3)$. Let $\mathbf{L}$ be a $p \times p$ matrix that diagonalizes both $\mathbf{\Psi}$ and $\mathbf{\Phi}$, such that $\mathbf{L}\mathbf{\Psi}\mathbf{L}' = \mathbf{I}$ and $\mathbf{L}\mathbf{\Phi}\mathbf{L}' = \mathbf{D}$, a diagonal matrix. This bi-diagonalization can be accomplished in $O(p^3)$ (details are found in a later section). We then define the matrix $\mathbf{M} = (\mathbf{L} \otimes \mathbf{H}'_k)$. The transformed data vector $\mathbf{Y}_T$ is defined as $\mathbf{Y}_T = \mathbf{M}vec(\mathbf{Y})$. This transformed vector has the following distribution.

$$\mathbf{Y}_T \sim N(\mathbf{M}vec(\mathbf{X}\beta), \mathbf{I} \otimes \mathbf{S}_k + \mathbf{D} \otimes \mathbf{I}) \tag{5.16}$$

The likelihood of $\mathbf{Y}_T$ is then given as follows.

$$L(\mathbf{Y}_T|\mathbf{X}\beta, \mathbf{\Psi}, \mathbf{K}, \mathbf{\Phi}) = -\frac{np}{2}ln(2\pi) - \frac{1}{2}ln|\mathbf{I} \otimes \mathbf{S}_k + \mathbf{D} \otimes \mathbf{I}|$$
$$-\frac{1}{2}(\mathbf{M}vec(\mathbf{Y}_T - \mathbf{X}\beta))'(\mathbf{I} \otimes \mathbf{S}_k + \mathbf{D} \otimes \mathbf{I})^{-1}(\mathbf{M}vec(\mathbf{Y}_T - \mathbf{X}\beta)) + log(|\mathbf{M}|) \tag{5.17}$$

In order to calculate the likelihood given $\mathbf{\Psi}$ and $\mathbf{\Phi}$, we first obtain the transformation matrix $\mathbf{M}$, which is accomplished in $O(n^3 + p^3)$. Next, we compute the transformed data vector $\mathbf{Y}_T$ in $O(n^2p + p^2n)$. Given $\mathbf{Y}_T$, we obtain an estimate of β, denoted by $\hat{\beta}$, which we show may be accomplished in $O(np^3q^2 + p^3q^3 + np^2q)$ and given this we calculate the residual vector $\mathbf{Y}_T - \mathbf{M}vec(\mathbf{X}\hat{\beta})$ in $O(np^2q + np)$. Finally, the likelihood is computed in $O(np)$. Therefore, negating the one time $O(n^3)$ cost of the decomposition of $\mathbf{K}$ and noting that the computation contributing the n^2 term can be cached for a given experiment, the total running time to evaluate the likelihood of the data under a setting for $\mathbf{\Psi}$ and $\mathbf{\Phi}$ is given by $O(p^3 + np^3q^2 + np)$. Considering that p and q are very small (ie. 2 and 1 in our experiments), this gives a final running time of $O(np^3)$, which is essentially linear in the total size of the data. Additionally, in human data we will often determine the ML parameters while assuming that $\beta = 0$. In this case, the running time to perform ML scales like $O(np)$, given the pre-cached higher order terms.

A more detailed explanation

We are given that $vec(\mathbf{Y})$ follows a multivariate distribution given by

$$N_{np}(vec(\mathbf{X}\beta), \mathbf{\Psi} \otimes \mathbf{K} + \mathbf{\Phi} \otimes \mathbf{I}_n) \tag{5.18}$$

We wish to diagonalize the covariance matrix of $vec(\mathbf{Y})$, in order to achieve a $O(np)$ likelihood computation time, given $\mathbf{\Psi}$, $\mathbf{\Phi}$ and the kinship matrix $\mathbf{K}$. First, we define the eigendecomposition of $\mathbf{K}$ as equal to $\mathbf{H}_K\mathbf{S}_K\mathbf{H}'_K$. Next, we identify a matrix $\mathbf{L}$ that diagonalizes $\mathbf{\Psi}$ and $\mathbf{\Phi}$, as showing in the previous section. The matrix $\mathbf{M} = (\mathbf{L} \otimes \mathbf{H}'_k)$ then diagonalizes the covariance matrix as shown in the following and we can obtain it in $O(n^3 + p^3)$ time.

$$\begin{aligned}
cov(vec(\mathbf{Y})) = & \quad \mathbf{\Psi} \otimes \mathbf{K} + \mathbf{\Phi} \otimes \mathbf{I}_n & (5.19)\\
cov(\mathbf{M}vec(\mathbf{Y})) = & \quad \mathbf{M}(\mathbf{\Psi} \otimes \mathbf{K} + \mathbf{\Phi} \otimes \mathbf{I}_n)\mathbf{M}' & (5.20)\\
= & \quad (\mathbf{L} \otimes \mathbf{H}'_k)(\mathbf{\Psi} \otimes \mathbf{K})(\mathbf{L}' \otimes \mathbf{H}_k) + (\mathbf{L} \otimes \mathbf{H}'_k)(\mathbf{\Phi} \otimes \mathbf{I}_n)(\mathbf{L}' \otimes \mathbf{H}_k) & (5.21)\\
= & \quad (\mathbf{L}\mathbf{\Psi}\mathbf{L}' \otimes \mathbf{H}'_k\mathbf{K}\mathbf{H}_k) + (\mathbf{L}\mathbf{\Phi}\mathbf{L}' \otimes \mathbf{H}'_k\mathbf{I}\mathbf{H}_k) & (5.22)\\
= & \quad (\mathbf{I} \otimes \mathbf{S}_k) + (\mathbf{D} \otimes \mathbf{I}) & (5.23)
\end{aligned}$$

To apply the transformation to $\mathbf{Y}$ we avoid calculating the Kronecker products directly by using the following rule.

$$\begin{aligned}
\mathbf{M}vec(\mathbf{Y}) = & \quad (\mathbf{L} \otimes \mathbf{H}'_k)vec(\mathbf{Y}) & (5.24)\\
= & \quad vec(\mathbf{H}'_k\mathbf{Y}\mathbf{L}') & (5.25)
\end{aligned}$$

The final log-likelihood is as follows.

$$L(\mathbf{Y}_T|\mathbf{X}\beta,\mathbf{\Psi},\mathbf{K},\mathbf{\Phi}) = -\frac{np}{2}ln(2\pi) - \frac{1}{2}ln|\mathbf{I}\otimes\mathbf{S}_k + \mathbf{D}\otimes\mathbf{I}|$$
$$-\frac{1}{2}(vec(\mathbf{H}_k'\mathbf{Y}\mathbf{L}') - vec(\mathbf{H}_k'\mathbf{X}\beta\mathbf{L}'))'(\mathbf{I}\otimes\mathbf{S}_k + \mathbf{D}\otimes\mathbf{I})^{-1}$$
$$(vec(\mathbf{H}_k'\mathbf{Y}\mathbf{L}') - vec(\mathbf{H}_k'\mathbf{X}\beta\mathbf{L}')) + log(|\mathbf{M}|)$$

Since the transformation matrix $\mathbf{M}$ does not result in an orthogonal transformation the term $log(|\mathbf{M}|)$ must be used as a normalization constant. The determinant of the transformation matrix is given by $|\mathbf{M}| = |\mathbf{L}\otimes\mathbf{H}_k'| = |\mathbf{L}|^n|\mathbf{H}_k'|^p = |\mathbf{L}|^n = (|\mathbf{Q}'||\mathbf{R}|)^n = |\mathbf{R}|^n = |\mathbf{S}_k^{-1/2}|^n = (\Pi_i S_{k(i)}^{-1/2})^n$, where $S_{k(i)}^{-1/2}$ represents the inverse square root of the ith diagonal element of $\mathbf{S}_k$. Thus taking the log of $|\mathbf{M}|$, we obtain a sum, which can be computed in $O(p)$.

The final time complexity for calculating this log-likelihood is given by $O(n^3 + p^3 + n^2p + p^2n + n^2q + np^3q^2 + p^3q^3 + np^2q + p + np)$, given that it takes $O(n^3)$ time to compute the eigendecomposition of $\mathbf{K}$ and $O(p^3)$ time to obtain the matrix $\mathbf{L}$ and $O(np^3q^2 + p^3q^3 + np^2q)$, to compute the fixed effect estimates, given the $O(n^2q)$ time to compute the transformed fixed effect matrix and $O(n^2p + p^2n)$ time to compute the transformed phenotype vector, finally taking $O(np)$ time to compute the log-likelihood. We may consider that this time is reduced by the fact that the decomposition of $\mathbf{K}$ remains constant for a given experimental setting so that it may be cached only one time. Furthermore, the $O(n^2)$ part of the transformation can also be cached for a set of phenotypes. With this logic, the final running time has complexity of $O(p^3 + p^2n + np^3q^2 + p^3q^3 + np^2q + np)$. Given that n is much larger than p or q, which in our experiments have been 2 and 1, respectively, this gives a running time of $O(np^3)$. Additionally, if we only care to fit the model under the assumption that the mean is zero, we no longer need to worry about the fixed effect estimate and can compute the likelihood in $O(np)$.

5.4.4 Restricted Maximum Likelihood Computation

The restricted maximum likelihood (REML) and the maximum likelihood (ML) solutions should be similar when the model contains no covariates, or only a bias term. However, when this is not the case, parameter estimates obtained in REML analysis may deviate significantly from those of ML. We obtain the REML version of the mvLMM by extending the ML solution [WT97]. By denoting the log-likelihood obtained by ML as L_{ML} and similar for REML, we define the following log-likelihood function. For a standard multivariate normal vector $\mathbf{y}$ with distribution $N(\mathbf{T}\alpha, \mathbf{\Theta})$, where $\mathbf{T}$ is $n \times q$, the REML is $LL_{REML} = LL_{ML} + \frac{1}{2}[qln(2\pi) + ln(|\mathbf{T}'\mathbf{T}|) - ln(|\mathbf{T}'\mathbf{\Theta}^{-1}\mathbf{T}|)]$ [KYE08]. Given this standard result, we define the REML log-likelihood for the mvLMM in the following.

$$LL_{REML} = LL_{ML} + \tag{5.26}$$

$$\frac{1}{2}[qln(2\pi) + ln(|(\mathbf{L}' \otimes (\mathbf{H}_k'\mathbf{X})')(\mathbf{L} \otimes \mathbf{H}_k'\mathbf{X})|) - \tag{5.27}$$

$$ln(|(\mathbf{L}' \otimes (\mathbf{H}_k'\mathbf{X})')(\mathbf{I} \otimes \mathbf{S}_k + \mathbf{D} \otimes \mathbf{I})^{-1}(\mathbf{L} \otimes \mathbf{H}_k'\mathbf{X})|)] \tag{5.28}$$

The computational cost of the operations required to define LL_{REML} do not change the order of the computational complexity.

5.4.5 Estimating Genetic Correlation

In order to evaluate the likelihood function in equation (5.26), we obtain estimates for the parameters $\mathbf{\Psi}$ and $\mathbf{\Phi}$. We estimate these parameters under the null model, where SNPs are not included as covariates. This assumption has been used previously and is valid for cases when the effect due to each SNP is small [KSS10, LLL11]. First, for each phenotype i, we fit the basic LMM from equation (5.1), in order to

identify the optimal variance parameters $\sigma^2_{g(i)}$ and $\sigma^2_{e(i)}$. Holding these parameters constant, we perform a two dimensional global grid search in order to identify the optimal genetic and environmental correlation parameters. Given that with caching the likelihood calculation takes time on the order of $O(np^3)$, this time will be multiplied by a constant k^2 when searching over a grid of size k for each correlation parameter. That is, if we evaluate the likelihood for each genetic and environmental correlation combination for a grid size of k, then we need to evaluate the likelihood k^2 times. In practice, we have found that a course grid can be used to identify the general region where the maximum-likelihood solution lies and then a more dense grid can be used within that region to clarify the solution. In this way, the time complexity can be reduced dynamically. In our experiments, the later approach can be used to achieve the same solution as the global search but reduce the running time from roughly 10 minutes to roughly 150 seconds ($N = 5000$).

5.4.6 Calculating sampling variance for parameter estimates

We calculate the sampling variance of the variance parameters and the correlation parameters using standard multivariate theory. Generally, the sampling variance of a maximum likelihood (ML) parameter is given by the inverse of the Fisher's information (or average information) matrix evaluated at the ML parameters [SCM92]. Using the search technique we describe, we identify the ML parameters for a given set of phenotypes and then use these parameters to estimate the sampling variance using the Fisher's information matrix.

5.4.7 Assessing Association

In order to identify genetic variations that have an effect on our traits of interest, we employ a hypothesis testing framework. We first estimate the effect that a particular

SNP x has on each of the phenotypes using the mvLMM model, then we jointly test m hypotheses, each testing the effect of the SNP on a given phenotype. Our null hypothesis for this test is that the SNP has no effect on any of our phenotypes and the alternative hypothesis is that it has an effect on one or more of the phenotypes.

To obtain estimates for the SNP effect sizes, we estimate the full β matrix from equation (5.15). First, we obtain the maximum likelihood parameters for $\mathbf{\Psi}$ and $\mathbf{\Phi}$ under the null model in which the SNP has no effect, as described in the previous section. Then, given these two parameters, we compute an estimate of the coefficient matrix $\hat{\beta}$ using the following result.

In the previous section, we defined a transformation $\mathbf{M} = (\mathbf{L} \otimes \mathbf{H}_k')$ and used it to define a transformed data vector $\mathbf{Y}_T$. The mean of the transformed data is given by $\mathbf{M}vec(\mathbf{X}\beta) = (\mathbf{L} \otimes \mathbf{H}_k')vec(\mathbf{X}\beta)$, which can be reduced as follows.

$$
\begin{aligned}
& (\mathbf{L} \otimes \mathbf{H}_k')vec(\mathbf{X}\beta) && (5.29)\\
= \quad & vec(\mathbf{H}_k'\mathbf{X}\beta\mathbf{L}') && (5.30)\\
= \quad & vec(\mathbf{X}^*\beta\mathbf{L}') && (5.31)\\
= \quad & (\mathbf{L} \otimes \mathbf{X}^*)vec(\beta) && (5.32)
\end{aligned}
$$

Here we have let $\mathbf{X}^* = \mathbf{H}_k'\mathbf{X}$. By denoting $vec(\beta)$ as β_T, we obtain an estimate $\hat{\beta}$ using the following result, where $unvec()$ represents the reversal of the $vec()$ operation and we have let $\mathbf{P} = (\mathbf{I} \otimes \mathbf{S}_k + \mathbf{D} \otimes \mathbf{I})$, the transformed data covariance matrix.

$$
\begin{aligned}
\hat{\beta}_T = \quad & [(\mathbf{L}' \otimes \mathbf{X}^{*\prime})\mathbf{P}^{-1}(\mathbf{L} \otimes \mathbf{X}^*)]^{-1}(\mathbf{L}' \otimes \mathbf{X}^{*\prime})\mathbf{P}^{-1}\mathbf{M}vec(\mathbf{Y}) && (5.33)\\
\hat{\beta} = \quad & unvec(\hat{\beta}_T) && (5.34)
\end{aligned}
$$

Since $\mathbf{P}$ is a diagonal matrix, $\hat{\beta}_T$ can be computed in $O(np^3q^2 + p^3q^3 + np^2q)$ given

the one time cost of $O(n^2q)$ for computing $\mathbf{X}^*$.

The statistic for testing the proposed hypothesis is obtained by defining a transformation matrix $\mathbf{R}$ so that $\mathbf{R}\hat{\beta}_T = [\hat{\beta}_{1,x} \quad \hat{\beta}_{2x} \dots \hat{\beta}_{px}]'$, where $\hat{\beta}_{ix}$ is the coefficient estimate for the effect of SNP x on phenotype i. Therefore, given this matrix, we define the F-statistic for testing association in equation (5.35), which under the null follows an F-distribution with p numerator degrees of freedom and $np - pq$ denominator degrees of freedom, where $\hat{\sigma}^2 = \hat{var}(\mathbf{P}^{-1/2}\mathbf{Y}_T)$ and $\hat{var}(\dots)$ represents the sample variance. Details on this test can be found in [MN].

$$f = (\mathbf{R}\hat{\beta}_T)'(\mathbf{R}[(\mathbf{L}' \otimes \mathbf{X}^{*\prime})\mathbf{P}^{-1}(\mathbf{L} \otimes \mathbf{X}^*)]^{-1}\mathbf{R}')^{-1}(\mathbf{R}\hat{\beta}_T) \cdot \frac{1}{p\hat{\sigma}^2} \tag{5.35}$$

5.4.8 Diagonalizing two matrices

We are given two positive semi-definite matrices $\mathbf{\Phi}$ and $\mathbf{\Psi}$ and we wish to identify a matrix $\mathbf{L}$ that diagonalizes both of these matrices. This is accomplished in the following way. First, we obtain the eigendecomposition of $\mathbf{\Psi} = \mathbf{H}_\Psi \mathbf{S}_\Psi \mathbf{H}_\Psi'$ and then define a matrix $\mathbf{R} = \mathbf{S}_\Psi^{-1/2}\mathbf{H}_\Psi'$, so that $\mathbf{R}'\mathbf{R} = \mathbf{\Psi}^{-1}$. Next, we obtain an eigendecomposition $\mathbf{R}\mathbf{\Phi}\mathbf{R}' = \mathbf{Q}\mathbf{D}\mathbf{Q}'$ and then define a matrix $\mathbf{L} = \mathbf{Q}'\mathbf{R}$. With this we see that $\mathbf{L}\mathbf{\Psi}\mathbf{L}' = \mathbf{I}$ and that $\mathbf{L}\mathbf{\Phi}\mathbf{L}' = \mathbf{D}$. The entire procedure has complexity $O(p^3)$.

5.4.9 Genotype and phenotype data

We apply our method to the Northern Finland Birth Cohort data [Sab08] which was used in [KSS10] and [KVS12]. This data set consisted of 5326 individuals which had been filtered to reduce the presence of family structure. Missing genotypes are replaced with the MAF. Missing phenotypes are replaced with the phenotypic mean.

We use a well-studied yeast dataset [SK08] consisting of 109 yeast strains each

with 5793 gene expression measurements. Bi-variate association mapping is performed on all 2956 available SNPs. Gene expression values were normalized and subjected to quality control by [SK08] and we utilized the same data as they.

CHAPTER 6

Conclusions

The field of genetics has experienced revolutionary change in the past 20 or 30 years, transforming it from a low-throughput small data science to a high-throughput big data science. One result of this shift is the formulation of a new discipline: computational genetics, a discipline requiring both computational expertise and statistical and biological thinking. The problems lining this field are often concerned with large scale computations utilizing 100's of thousands of measurements and non-standard statistical analyses. In each of the projects discussed in this thesis, I have introduced one such problem and given my view and solution of it. However, the work I have done here is only the beginning of something much bigger. In what follows, I would like to explain how the work presented herein fits together in a much bigger picture with many more implications on the future of genetics research and of human health.

There are currently two higher level trends operating simultaneously in genetics research: discovery and prediction. Problems related to discovery, which have dominated most of the past 10 years, are aimed at identifying the causal mechanisms behind natural phenotypic variation. The most obvious example of this is embodied in the genome-wide association study (GWAS), where the idea in the simplest case is to identify genetic variations that are the cause of a particular disease or that increase the risk of developing a disease. The results of these discovery phase studies can then be utilized to develop a deeper understanding of the disease etiology and to potentially develop therapeutic interventions. On the other hand, problems related to prediction

do not necessarily have the goal of identifying causal mechanisms. Instead, these problems are focused on the prediction of a phenotypic state given all of the genetic information for an individual. For example, a phenotype prediction algorithm may be used to assess the risk for a given disease without actually knowing the causal mechanisms. Prediction problems have played a lesser role in the main stream computational genetics literature over the past 10 years when compared to discovery-based problems. However, over the past 2 or 3 years, there has been an increase in the number of publications dedicated to phenotype prediction, particularly those using so called whole genome approaches – those considering all genetic variation information jointly to perform prediction.

The work presented in this thesis has been primarily focused on specific problems related to discovery. For example, the meta-analysis, longitudinal GWAS and mvLMM chapters all focus on particular problems in the context of GWAS. However, taking a step back, the common thread between each of these projects is the use of a linear-mixed model (LMM) to account for the complex relationships between many phenotypic measurements and many genetic variations. In particular, the work presented herein has asked first how to deal with a particular form of complex relationship using a LMM (multiple measurement of a phenotype over time, measurements of multiple phenotypes, phenotype measurements from separate but related populations, etc.) and then how to do so in a computationally efficient manner. LMMs are a very powerful and well established statistical framework that are often used in a statistical hypothesis testing framework, as in my case. However these same statistical models can be used to perform phenotype prediction, a direction that has been explored very recently in a number of publications.

Moving into the future, I believe there will be a much stronger focus on phenotype prediction, especially with respect to its relation to genetic risk, drug interaction, gene

by environment interactions and generally many of the concepts behind personalized medicine. The work that I have presented in this thesis when viewed in a slightly different light, also provides a contribution to problems in this area. That is, given that standard mixed models used in GWAS can be used to predict phenotypes, the same statistical models I propose can also be utilized for this reason. This means that it may be possible to predict phenotypes over time or to predict a given phenotype based on observations of many other phenotypes. In addition, similar predictive models can be utilized to perform phenotype prediction under different environmental settings by incorporating gene-by-environment interactions. If such prediction algorithms have high accuracy, the predictive ability will have a dramatic effect on healthcare. Therefore, the continuation of the work I have presented in this thesis may take the shape of predictive algorithms that can be used in a personalized medicine setting to inform individuals about their risk for a particular disease conditioned on available observable information

REFERENCES

[AGV02] M. de Andrade, R. Guéguen, S. Visvikis, C. Sass, G. Siest, and C.I. Amos. "Extension of variance components approach to incorporate temporal trends and longitudinal pedigree data analysis." *Genetic epidemiology*, **22**(3):221–232, 2002.

[AHN11] Christy L. Avery, Qianchuan He, Kari E. North, Jose L. Ambite, Eric Boerwinkle, Myriam Fornage, Lucia A. Hindorff, Charles Kooperberg, James B. Meigs, James S. Pankow, Sarah A. Pendergrass, Bruce M. Psaty, Marylyn D. Ritchie, Jerome I. Rotter, Kent D. Taylor, Lynne R. Wilkens, Gerardo Heiss, and Dan Yu Lin. "A Phenomics-Based Strategy Identifies Loci on APOC1, BRAP, and PLCG1 Associated with Metabolic Syndrome Phenotype Domains." *PLoS Genet*, **7**(10):e1002322, 10 2011.

[ARL08] Y.S. Aulchenko, S. Ripatti, I. Lindqvist, D. Boomsma, I.M. Heid, P.P. Pramstaller, B.W.J.H. Penninx, A.C.J.W. Janssens, J.F. Wilson, T. Spector, et al. "Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts." *Nature genetics*, **41**(1):47–55, 2008.

[AVF11] David L. Aylor, William Valdar, Wendy Foulds-Mathes, Ryan J. Buus, Ricardo A. Verdugo, Ralph S. Baric, Martin T. Ferris, Jeff A. Frelinger, Mark Heise, Matt B. Frieman, Lisa E. Gralinski, Timothy A. Bell, John D. Didion, Kunjie Hua, Derrick L. Nehrenberg, Christine L. Powell, Jill Steigerwalt, Yuying Xie, Samir Np Kelada, Francis S. Collins, Ivana V. Yang, David A. Schwartz, Lisa A. Branstetter, Elissa J. Chesler, Darla R. Miller, Jason Spence, Eric Yi Liu, Leonard McMillan, Abhishek Sarkar, Jeremy Wang, Wei Wang, Qi Zhang, Karl W. Broman, Ron Korstanje, Caroline Durrant, Richard Mott, Fuad A. Iraqi, Daniel Pomp, David Threadgill, Fernando Pardo-Manuel de Villena, and Gary A. Churchill. "Genetic analysis of complex traits in the emerging collaborative cross." *Genome Res*, 3 2011.

[Bal06] D.J. Balding. "A tutorial on statistical methods for population association studies." *Nature Reviews Genetics*, **7**(10):781–791, 2006.

[BBK11] J.A. Blake, C.J. Bult, J.A. Kadin, J.E. Richardson, and J.T. Eppig. "The Mouse Genome Database (MGD): premier model organism resource for mammalian genomics and genetics." *Nucleic acids research*, **39**(suppl 1):D842, 2011.

[BBS12] T.R. Braun, L.F. Been, A. Singhal, J. Worsham, S. Ralhan, G.S. Wander, J.C. Chambers, J.S. Kooner, C.E. Aston, and D.K. Sanghera. "A Replica-

tion Study of GWAS-Derived Lipid Genes in Asian Indians: The Chromosomal Region 11q23. 3 Harbors Loci Contributing to Triglycerides." *PloS one*, **7**(5):e37056, 2012.

[BFJ08] P.I.W. de Bakker, M.A.R. Ferreira, X. Jia, B.M. Neale, S. Raychaudhuri, and B.F. Voight. "Practical aspects of imputation-driven meta-analysis of genome-wide association studies." *Human molecular genetics*, **17**(R2):R122, 2008.

[BFO10a] B.J. Bennett, C.R. Farber, L. Orozco, H. Min Kang, A. Ghazalpour, N. Siemers, M. Neubauer, I. Neuhaus, R. Yordanova, B. Guan, et al. "A high-resolution association mapping panel for the dissection of complex traits in mice." *Genome Research*, **20**(2):281, 2010.

[BFO10b] Brian J. Bennett, Charles R. Farber, Luz Orozco, Hyun Min Kang, Anatole Ghazalpour, Nathan Siemers, Michael Neubauer, Isaac Neuhaus, Roumyana Yordanova, Bo Guan, Amy Truong, Wen-pin P. Yang, Aiqing He, Paul Kayne, Peter Gargalovic, Todd Kirchgessner, Calvin Pan, Lawrence W. Castellani, Emrah Kostem, Nicholas Furlotte, Thomas A. Drake, Eleazar Eskin, and Aldons J. Lusis. "A high-resolution association mapping panel for the dissection of complex traits in mice." *Genome Res*, **20**(2):281–90, 2 2010.

[BKB03] Gábor Balázsi, Krin A. Kay, Albert-László L. Barabási, and Zoltán N. Oltvai. "Spurious spatial periodicity of co-expression in microarray data due to printing design." *Nucleic Acids Res*, **31**(15):4425–33, 8 2003.

[BYC02] R B Brem, G Yvert, R Clinton, and L Kruglyak. "Genetic dissection of transcriptional regulation in budding yeast." *Science*, **296**(5568):752–5, 2002.

[BYP05] P.I.W. de Bakker, R. Yelensky, I. Pe'er, S.B. Gabriel, M.J. Daly, and D. Altshuler. "Efficiency and power in genetic association studies." *Nature genetics*, **37**(11):1217–1223, 2005.

[CRW08] Robert Clarke, Habtom W. Ressom, Antai Wang, Jianhua Xuan, Minetta C. Liu, Edmund A. Gehan, and Yue Wang. "The properties of high-dimensional data spaces: implications for exploring gene and protein expression data." *Nat Rev Cancer*, **8**(1):37–49, 1 2008.

[DLW90] MH Doolittle, RC LeBoeuf, CH Warden, LM Bee, and AJ Lusis. "A polymorphism affecting apolipoprotein A-II translational efficiency determines high density lipoprotein size and composition." *Journal of Biological Chemistry*, **265**(27):16380, 1990.

[DRB01] B. Devlin, K. Roeder, and S.A. Bacanu. "Unbiased methods for population-based association studies." *Genetic epidemiology*, **21**(4):273–284. doi:10.1002/gepi.1034, 2001.

[DRW01] B. Devlin, K. Roeder, and L. Wasserman. "Genomic control, a new approach to genetic-based association studies." *Theor Popul Biol*, **60**(3):155–66, 11 2001.

[ECW04] D. Estrada-Smith, L.W. Castellani, H. Wong, P.Z. Wen, A. Chui, A.J. Lusis, and R.C. Davis. "Dissection of multigenic obesity traits in congenic mouse strains." *Mammalian Genome*, **15**(1):14–22, 2004.

[EFG10] Evan E. Eichler, Jonathan Flint, Greg Gibson, Augustine Kong, Suzanne M. Leal, Jason H. Moore, and Joseph H. Nadeau. "Missing heritability and strategies for finding the underlying causes of complex disease." *Nat Rev Genet*, **11**(6):446–50, 6 2010.

[Fal81] D.S. Falconer. *Introduction to quantitative genetics.* Longman., New York, 2 edition, 1981.

[FBO11] Charles R. Farber, Brian J. Bennett, Luz Orozco, Wei Zou, Ana Lira, Emrah Kostem, Hyun Min Kang, Nicholas Furlotte, Ani Berberyan, Anatole Ghazalpour, Jaijam Suwanwela, Thomas A. Drake, Eleazar Eskin, Q. Tian Wang, Steven L. Teitelbaum, and Aldons J. Lusis. "Mouse genome-wide association and systems genetics identify asxl2 as a regulator of bone mineral density and osteoclastogenesis." *PLoS Genet*, **7**(4):e1002038, 4 2011.

[FM01] J. Flint and R. Mott. "Finding the molecular basis of quantitative traits: successes and pitfalls." *Nat Rev Genet*, **2**(6):437–45, 6 2001.

[FNG09] Charles R. Farber, Atila van Nas, Anatole Ghazalpour, Jason E. Aten, Sudheer Doss, Brandon Sos, Eric E. Schadt, Leslie Ingram-Drake, Richard C. Davis, Steve Horvath, Desmond J. Smith, Thomas A. Drake, and Aldons J. Lusis. "An integrative genetics approach to identify candidate genes regulating BMD: combining linkage, gene expression, and association." *J Bone Miner Res*, **24**(1):105–16, 1 2009.

[FP09] Manuel A. R. Ferreira and Shaun M. Purcell. "A multivariate test of association." *Bioinformatics*, **25**(1):132–133, 2009.

[GDZ06] A Ghanzalpour, S Doss, B Zhang, S Wang, C Plaisier, R Castellanos, A Brozell, E E Schadt, T A Drake, A J Lusis, and S Horvath. "Integrating genetic and network analysis to characterize gene related to mouse weight." *PLoS Genetics*, **2**(8):e130, 2006.

[GN00] A.K. Gupta and D.K. Nagar. *Matrix variate distributions*, volume 104. Chapman & Hall/CRC, 2000.

[GWV09] M. E. Goddard, N. R. Wray, K. Verbyla, and P. M. Visscher. "Estimating effects and making predictions from genome-wide marker data." *Statistical Science*, **24**(4):517–529, 2009.

[Har77] D.A. Harville. "Maximum likelihood approaches to variance component estimation and to related problems." *Journal of the American Statistical Association*, **72**(358):320–338, 1977.

[HDK00] R. Hitzemann, K. Demarest, J. Koyner, L. Cipp, N. Patel, E. Rasmussen, and J. McCaughran. "Effect of genetic cross on the detection of quantitative trait loci and a novel approach to mapping QTLs." *Pharmacol Biochem Behav*, **67**(4):767–72, 12 2000.

[HE11] B. Han and E. Eskin. "Random-Effects Model Aimed at Discovering Associations in Meta-Analysis of Genome-wide Association Studies." *The American Journal of Human Genetics*, **88**(5):586–598, 2011.

[Hen50] CR Henderson. "Estimation of genetic parameters." *Ann. Math. Stat*, **21**:309, 1950.

[Hen73] C. R. Henderson. "Sire evaluation and genetic trends." *Journal of Animal Science*, **1973**(no. Symposium):10–41, 1973.

[HMC02] R. Hitzemann, B. Malmanger, S. Cooper, S. Coulombe, C. Reed, K. Demarest, J. Koyner, L. Cipp, J. Flint, C. Talbot, et al. "Multiple Cross Mapping (MCM) markedly improves the localization of a QTL for ethanol-induced activation." *Genes, Brain and Behavior*, **1**(4):214–222, 2002.

[HQ76] C. R. Henderson and R. L. Quaas. "Multiple trait evaluation using relatives' records." *Journal of Animal Science*, **43**(6):1188, 1976.

[HSC97] R.A. Heller, M. Schena, A. Chai, D. Shalon, T. Bedilion, J. Gilmore, D.E. Woolley, and R.W. Davis. "Discovery and analysis of inflammatory disease-related genes using cDNA microarrays." *Proceedings of the National Academy of Sciences*, **94**(6):2150–2155, 1997.

[HSV09] Guo-Jen J. Huang, Sagiv Shifman, William Valdar, Martina Johannesson, Binnaz Yalcin, Martin S. Taylor, Jennifer M. Taylor, Richard Mott, and Jonathan Flint. "High resolution mapping of expression QTLs in heterogeneous stock mice in multiple tissues." *Genome Res*, **19**(6):1133–40, 6 2009.

[HVG09] B. J. Hayes, P. M. Visscher, and M. E. Goddard. "Increased accuracy of artificial selection by using the realized relationship matrix." *Genet Res*, **91**(1):47–60, 2 2009.

[IML07] I. Ionita-Laza, M.B. McQueen, N.M. Laird, and C. Lange. "Genomewide weighted hypothesis testing in family-based association studies, with an application to a 100K scan." *The American Journal of Human Genetics*, **81**(3):607–614, 2007.

[Int03] International HapMap Consortium. "The International HapMap Project." *Nature*, **426**:789–796, 2003.

[JRL07] W E Johnson, A Rabinovic, and C Li. "Adjusting batch effects in microarray expression using empirical Bayes methods." *Biostatistics*, **8**(1):118–27, 2007.

[JS86] R. I. Jennrich and M. D. Schluchter. "Unbalanced repeated-measures models with structured covariance matrices." *Biometrics*, **42**(4):805–20, 12 1986.

[KK04] T. Kariya and H. Kurata. *Generalized least squares*. John Wiley & Sons Inc, 2004.

[KKW10] Andrew Kirby, Hyun Min Kang, Claire M. Wade, Chris J. Cotsapas, Emrah Kostem, Buhm Han, Nick Furlotte, Eun Yong Kang, Manuel Rivas, Molly A. Bogue, Kelly A. Frazer, Frank M. Johnson, Erica J. Beilharz, David R. Cox, Eleazar Eskin, and Mark J. Daly. "Fine Mapping in 94 Inbred Mouse Strains Using a High-density Haplotype Resource." *Genetics*, 5 2010.

[KMD07] S. Kathiresan, A. Manning, S. Demissie, R. D'Agostino, A. Surti, C. Guiducci, L. Gianniny, N. Burtt, O. Melander, M. Orho-Melander, et al. "A genome-wide association study for blood lipid phenotypes in the Framingham Heart Study." *BMC medical genetics*, **8**(Suppl 1):S17, 2007.

[KMO10] Y. Kamatani, K. Matsuda, Y. Okada, M. Kubo, N. Hosono, Y. Daigo, Y. Nakamura, and N. Kamatani. "Genome-wide association study of hematological and biochemical traits in a Japanese population." *Nature genetics*, **42**(3):210–215, 2010.

[KRI01] A.B. Korol, Y.I. Ronin, A.M. Itskovich, J. Peng, and E. Nevo. "Enhanced efficiency of quantitative trait loci mapping analysis based on multivariate complexes of quantitative traits." *Genetics*, **157**(4):1789, 2001.

[KSS10] Hyun Min Kang, Jae Hoon Sul, Susan K. Service, Noah A. Zaitlen, Sit-Yee Y. Kong, Nelson B. Freimer, Chiara Sabatti, and Eleazar Eskin. "Variance component model to account for sample structure in genome-wide association studies." *Nat Genet*, **42**(4):348–54, 4 2010.

[KVS12] Arthur Korte, Bjarni J. Vilhjálmsson, Vincent Segura, Alexander Platt, Quan Long, Magnus Nordborg, Arthur Korte, Bjarni J. Vilhjálmsson, Vincent Segura, Alexander Platt, Quan Long, and Magnus Nordborg. "A mixed-model approach for genome-wide association studies of correlated traits in structured populations." *Nature Genetics*, **44**(9):1066, 8 2012.

[KYE08] H M Kang, C Ye, and E Eskin. "Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots." *Genetics*, **180**(4):1909, 2008.

[KZW08] Hyun Min Kang, Noah A Zaitlen, Claire M Wade, Andrew Kirby, David Heckerman, Mark J Daly, and Eleazar Eskin. "Efficient Control of Population Structure in Model Organism Association Mapping." *Genetics*, **178**:1709–1723, 2008.

[Lan02] K. Lange. *Mathematical and statistical methods for genetic analysis*. Springer Verlag, New York, 2nd edition, 2002.

[LKS10] Jennifer Listgarten, Carl Kadie, Eric E. Schadt, and David Heckerman. "Correction for hidden confounders in the genetic analysis of gene expression." *Proc Natl Acad Sci U S A*, **107**(38):16465–70, 9 2010.

[LLK12] Jennifer Listgarten, Christoph Lippert, Carl M. Kadie, Robert I. Davidson, Eleazar Eskin, David Heckerman, Jennifer Listgarten, Christoph Lippert, Carl M. Kadie, Robert I. Davidson, Eleazar Eskin, and David Heckerman. "Improved linear mixed models for genome-wide association studies." *Nature Methods*, **9**(6):525, 5 2012.

[LLL11] Christoph Lippert, Jennifer Listgarten, Ying Liu, Carl M. Kadie, Robert I. Davidson, David Heckerman, Christoph Lippert, Jennifer Listgarten, Ying Liu, Carl M. Kadie, Robert I. Davidson, and David Heckerman. "FaST linear mixed models for genome-wide association studies." *Nature Methods*, **8**(10):833, 9 2011.

[LLW05] Renhua Li, Malcolm A. Lyons, Henning Wittenburg, Beverly Paigen, and Gary A. Churchill. "Combining Data From Multiple Inbred Line Crosses Improves the Power and Resolution of Quantitative Trait Loci Mapping." *Genetics*, **169**(3):1699–1709, 3 2005.

[LPD06] S I Lee, D Pe'er, A M Dudlet, G M Church, and D Koller. "Identifying regulatory mechanisms using individual variation reveals key role for chromatin modification." *Proc Nat'l Acad Sci USA*, **103**(38):14062–7, 2006.

[LPL09] Y.Z. Liu, Y.F. Pei, J.F. Liu, F. Yang, Y. Guo, L. Zhang, X.G. Liu, H. Yan, L. Wang, Y.P. Zhang, et al. "Powerful bivariate genome-wide association analyses suggest the SOX6 gene influencing both obesity and osteoporosis phenotypes in males." *PLoS One*, **4**(8):e6827, 2009.

[LS07] J T Leek and J D Storey. "Capturing heterogeneity in gene expression studies by Surrogate Variable Analysis." *PLoS Genetics*, **3**(9):e161, 2007.

[LS08] Jeffrey T. Leek and John D. Storey. "A general framework for multiple testing dependence." *Proc Natl Acad Sci U S A*, **105**(48):18718–23, 12 2008.

[LW82] N.M. Laird and J.H. Ware. "Random-effects models for longitudinal data." *Biometrics*, **38**(4):963–974, 1982.

[LYG12] S. H. Lee, J. Yang, M. E. Goddard, P. M. Visscher, and N. R. Wray. "Estimation of pleiotropy between complex diseases using SNP-derived genomic relationships and restricted maximum likelihood." *Bioinformatics*, 7 2012.

[MCC09] Teri A. Manolio, Francis S. Collins, Nancy J. Cox, David B. Goldstein, Lucia A. Hindorff, David J. Hunter, Mark I. McCarthy, Erin M. Ramos, Lon R. Cardon, Aravinda Chakravarti, Judy H. Cho, Alan E. Guttmacher, Augustine Kong, Leonid Kruglyak, Elaine Mardis, Charles N. Rotimi, Montgomery Slatkin, David Valle, Alice S. Whittemore, Michael Boehnke, Andrew G. Clark, Evan E. Eichler, Greg Gibson, Jonathan L. Haines, Trudy F. C. Mackay, Steven A. McCarroll, and Peter M. Visscher. "Finding the missing heritability of complex diseases." *Nature*, **461**(7265):747–53, 10 2009.

[MGP09] G. Manenti, A. Galvan, A. Pettinicchio, G. Trincucci, E. Spada, A. Zolin, S. Milani, A. Gonzalez-Neira, and T.A. Dragani. "Mouse genome-wide association mapping needs linkage analysis to avoid false-positive loci." *PLoS Genetics*, **5**(1):e1000331, 2009.

[MHK99] H.W. Mewes, K. Heumann, A. Kaps, K. Mayer, F. Pfeiffer, S. Stocker, and D. Frishman. "MIPS: a database for genomes and protein sequences." *Nucleic acids research*, **27**(1):44, 1999.

[MN] C.E. McCulloch and J.M. Neuhaus. "Generalized linear mixed models.".

[MS01] C.E. McCulloch and S.R. Searle. *Generalized, linear, and mixed models*. Wiley-Interscience, 2001.

[MT05] RA Mrode and R. Thompson. *Linear models for the prediction of animal breeding values*. Cabi, Cambridge, MA, 2nd edition, 2005.

[NGW09] Atila van Nas, Debraj Guhathakurta, Susanna S. Wang, Nadir Yehya, Steve Horvath, Bin Zhang, Leslie Ingram-Drake, Gautam Chaudhuri, Eric E. Schadt, Thomas A. Drake, Arthur P. Arnold, and Aldons J. Lusis. "Elucidating the role of gonadal hormones in sexually dimorphic gene coexpression networks." *Endocrinology*, **150**(3):1235–49, 3 2009.

[OWA06] Pieter A. Oliehoek, Jack J. Windig, Johan A. M. van Arendonk, and Piter Bijma. "Estimating relatedness between individuals in general populations with a focus on their use in conservation programs." *Genetics*, **173**(1):483–96, 5 2006.

[Pay07] B.A. Payseur et al. "Prospects for association mapping in classical inbred mouse strains." *Genetics*, **175**(4):1999, 2007.

[PBL07] J.L. Peirce, K.W. Broman, L. Lu, and R.W. Williams. "A simple method for combining genetic mapping data from multiple crosses and experimental designs." *PLoS One*, **2**(10):e1036, 2007.

[PMB04] Mathew T. Pletcher, Philip McClurg, Serge Batalov, Andrew I. Su, S. Whitney Barnes, Erica Lagler, Ron Korstanje, Xiaosong Wang, Deborah Nusskern, Molly A. Bogue, Richard J. Mural, Beverly Paigen, and Tim Wiltshire. "Use of a dense single nucleotide polymorphism map for in silico mapping in the mouse." *PLoS Biol*, **2**(12):e393, 12 2004.

[QBK05] Xing Qiu, Andrew I. Brooks, Lev Klebanov, and Ndrei Yakovlev. "The effects of normalization on the correlation structure of microarray data." *BMC Bioinformatics*, **6**:120, 2005.

[Rao73] C.R. Rao. "Linear statistical inference and applications." *NY: Wiley*, 1973.

[Sab08] C. Sabatti et al. "Genome-wide association analysis of metabolic traits in a birth cohort from a founder population." *Nature genetics*, **41**(1):35–46, 2008.

[SCM92] S.R. Searle, G. Casella, C.E. McCulloch, et al. *Variance components*. Wiley New York, 1992.

[SIL11] O. Stegle, M.P. Institutes, C. Lippert, J. Mooij, N. Lawrence, and K. Borgwardt. "Efficient inference in matrix-variate Gaussian models with iid observation noise." *Advances in Neural Information Processing Systems 2011*, 2011.

[SK08] E N Smith and L Kruglyak. "Gene-environment interaction in yeast gene expression." *PLoS Biology*, **6**(4):e83, 2008.

[SOC00] A.M. Shearman, J.M. Ordovas, L.A. Cupples, E.J. Schaefer, M.D. Harmon, Y. Shao, J.D. Keen, A.L. DeStefano, O. Joost, P.W.F. Wilson, et al. "Evidence for a gene influencing the TG/HDL-C ratio on chromosome 7q32. 3–qter: a genome-wide scan in the Framingham Study." *Human molecular genetics*, **9**(9):1315–1320, 2000.

[SRC09] N. Soranzo, F. Rivadeneira, U. Chinappen-Horsley, I. Malkina, J.B. Richards, N. Hammond, L. Stolk, A. Nica, M. Inouye, A. Hofman, et al. "Meta-analysis of genome-wide scans for human adult stature identifies novel Loci and associations with measures of skeletal frame size." *PLoS Genetics*, **5**(4):e1000445, 2009.

[SSK03] Joshua M. Stuart, Eran Segal, Daphne Koller, and Stuart K. Kim. "Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules." *Science*, **302**(5634):249–255, 2003.

[STM05] A Subramanian, P Tamayo, V K Mootha, S Mukherjee, B L Ebert, M A Gillette, A Paulovich, S L Pomeroy, T R Golub, E S Lander, and J P Mesirov. "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." *Proc Nat'l Acad Sci U S A*, **102**(43):15545–50, 2005.

[Stu65] AH Sturtevant. "AHistory OF GENETICS." 1965.

[VP05] B.F. Voight and J.K. Pritchard. "Confounding from cryptic relatedness in case-control association studies." *PLoS Genet*, **1**(3):32. doi:10.1371/journal.pgen.0010032, 2005.

[WB99] J. T. Williams and J. Blangero. "Power of variance component linkage analysis to detect quantitative trait loci." *Annals of Human Genetics*, **63**(6):545–563, 1999.

[Wea49] CE Weatherburn. *A First Course Mathematical Statistics*. Cambridge Univ Pr, 1949.

[WGH11] X. Wang, X. Guo, M. He, and H. Zhang. "Statistical Inference in Mixed Models and Analysis of Twin and Family Data." *Biometrics*, 2011.

[WHQ93] CH Warden, CC Hedrick, JH Qiao, LW Castellani, and AJ Lusis. "Science.", 7 1993.

[WSL08] C.J. Willer, E.K. Speliotes, R.J.F. Loos, S. Li, C.M. Lindgren, I.M. Heid, S.I. Berndt, A.L. Elliott, A.U. Jackson, C. Lamina, et al. "Six new loci associated with body mass index highlight a neuronal influence on body weight regulation." *Nature Genetics*, **41**(1):25–34, 2008.

[WT97] S. J. Welham and R. Thompson. "Likelihood Ratio Tests for Fixed Model Terms Using Residual Maximum Likelihood." *Journal of the Royal Statistical Society. Series B (Methodological)*, **59**(3):701–714, 1997.

[WYS06] S. Wang, N. Yehya, E.E. Schadt, H. Wang, T.A. Drake, and A.J. Lusis. "Genetic and genomic analysis of a fat mass trait with complex inheritance reveals marked sex specificity." *PLoS genetics*, **2**(2):e15, 2006.

[YBM10] Jian Yang, Beben Benyamin, Brian P. McEvoy, Scott Gordon, Anjali K. Henders, Dale R. Nyholt, Pamela A. Madden, Andrew C. Heath, Nicholas G. Martin, Grant W. Montgomery, Michael E. Goddard, and Peter M. Visscher. "Common SNPs explain a large proportion of the heritability for human height." *Nature Genet*, **42**(7):565–569. doi:10.1038/ng.608, 7 2010.

[YPB06] J. Yu, G. Pressoir, W.H. Briggs, I. Vroh Bi, M. Yamasaki, J.F. Doebley, M.D. McMullen, B.S. Gaut, D.M. Nielsen, J.B. Holland, et al. "A unified mixed-model method for association mapping that accounts for multiple levels of relatedness." *Nature genetics*, **38**(2):203–208, 2006.

[ZE10] N. Zaitlen and E. Eskin. "Imputation aware meta-analysis of genome-wide association studies." *Genetic Epidemiology*, 2010.

[ZSS08] E. Zeggini, L.J. Scott, R. Saxena, B.F. Voight, J.L. Marchini, T. Hu, P.I.W. de Bakker, G.R. Abecasis, P. Almgren, G. Andersen, et al. "Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes." *Nature Genetics*, **40**(5):638–645, 2008.

[ZSZ12] Xiang Zhou, Matthew Stephens, Xiang Zhou, and Matthew Stephens. "Genome-wide efficient mixed-model analysis for association studies." *Nature Genetics*, 6 2012.