UNIVERSITY OF CALIFORNIA, SAN DIEGO

Effective Design and Analysis of Systems Genetics Studies

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

in

Computer Science

by

Hyun Min Kang

Committee in charge:

Professor Pavel Pevzner, Chair
Professor Eleazar Eskin, Co-Chair
Professor Vineet Bafna
Professor Sanjoy Dasgupta
Professor Trey Ideker
Professor Nicholas J. Schork

2009

The dissertation of Hyun Min Kang is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Co-Chair

Chair

University of California, San Diego

2009

DEDICATION

To Jihye and Joseph.

EPIGRAPH

Get the facts, or the facts will get you.
And when you get them, get them right, or they will get you wrong.
— Thomas Fuller

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I thank my advisor Eleazar Eskin for guiding my research, sharing insights, always being accessible, trusting me and showing endurance. I thank my fanatistic colleagues who joined in the lab with me at the same year. Noah Zaitlen, Chun (Jimmie) Ye and Sean O'Rourke, They have always kindly shared their knowledge and insight in the research area especially when I had very little idea when I first started my research, and I truly enjoyed my years of graduate studies with them. I also thank the rest of my current and previous lab colleagues Buhm Han, Emrah Kostem, Chris Jones, Olivera Grujic, Eun Yong Kang, Nick Furlotte, Jaehoon Sul, Dan He, Nils Homer, Michael Sanders, Ilya Shpitster, and Juan Lorenzo Roderiguez for sharing scientific insights and spending great time together. Joining in Zarlab at UC San Diego is one of the best decisions I have ever made in my life.

There are many more terrific people I would like to offer my gratitude. Aldons Jake Lusis and Chiara Sabatti for their extremely strong support for my research direction, which truly nourished this dissertation. My great collaborators Erica Beilharz, Brian Bennett, Chris Cotsapas, Thomas Drake, Charles Farber, Kelly Frazer, Anatole Ghanzalpour, Carl Kadie, Eran Halperin, Andrew Kirby, Susan Service, Myeong Seong Seo, Claire Wade, and Matthew Zapala. My great teachers and thesis comittee members Pavel Pevzner, Vineet Bafna, Sanjoy Dasgupta, Trey Ideker, and Nicholas Schork, from whom I learned most of knowledge and insight during my years of graduate studies. My church family for always encouraging and inspiring me, both mentally and spiritually. My parents Chi-won Kang and Soon-rye Han, my sister Hyunhee Kang and my in-laws for their enduring love and support. I especially thank my lovely wife and son, Jihye and Joseph, for their presence, love, encouragement, endurance, comfort, and support. Finally and most importantly, I would like to thank the almight God and Jesus Christ, my Lord and Savior, for his everlasting love, grace and providence.

I was supported by the Samsung Fellowship, National Science Foundation grant no. 0513612, 0713455, and 029049, National Institute of Health grant no. 1K25HL-080029, and U01-DA024417, and the UCLA subcontract of contract N01-ES-45530 from the National Toxicology Program/ National Institute of Environmen-

tal Health Sciences to Perlegen Sciences. I thank the National Biomedical Computational Resource (NBCR), the UCSD Fast Wired and Wireless Grid Project (FW-Grid), Insitutide for Digital Research and Education (IDRE) for allowing us to utilize their cluster system.

Chapter 2 is currently in submission for publication for the material. Andrew Kirby, Hyun Min Kang, Claire M. Wade, Chris J. Cotsapas, Emrah Kostem, Buhm Han, Manuel Rivas, Molly A. Bogue, Kelly A. Frazer, Frank M. Johnson, Erica J. Beilharz, David R. Cox, Eleazar Eskin, and Mark J. Daly, "A high-density haplotype resource of 94 inbred mouse strains". The dissertation author and Andrew Kirby are the primary investigators and authors of this paper.

Chapter 3 was published in Proceedings of the 13th Annual Conference on Research in Computational Biology (RECOMB-2009), Tuscon, Arizona, May 18-21, 2009. Hyun Min Kang, Noah Zaitlen, Buhm Han, and Eleazar Eskin, "An adaptive and efficient algorithm for genotype imputation". The dissertation author was the primary investigator and author of this paper.

Chapter 4 was published in Genetics, Volume 178, pp 1709-23, 2008. 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". The dissertation author was the primary investigator and author of this paper.

Chapter 5 is currently in submission for publication for the material. Hyun Min Kang, Jae-hoon Sul, Susan Service, Noah Zaitlen, Sit-yee Kong, Nelson Freimer, Chiara Sabatti, and Eleazar Eskin. "Accounting for sample structure in large scale genome-wide association studies using a variance component model". The dissertation author and Jae-hoon Sul are the primary investigators and authors of this paper.

Chapter 6 is published in Genetics, Volume 180, pp 1909-25, 2008. Hyun Min Kang, Chun Ye, and Eleazar Eskin, "Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots". The dissertation author and Chun Ye were the primary investigators and authors of this paper.

Chapter 7, is currently being prepared for submission for publication of the material. Brian J. Bennett, Charles R. Farber, Luz Orozco, Hyun Min Kang, Anatole Chanzalpour, Todd Kirchgessner, Peter Gargalovic, 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”. The dissertation author, Brian J. Bennett, Charles R. Farber, and Luz Orozco are the primary investigators and authors of this paper.

VITA

1998	Bachelor of Science in Electrical Engineering, Seoul National University, Seoul, Korea
2000	Master of Science in Electrical Engineering , Seoul National University, Seoul, Korea
2000-2002	Republic of Korea Army
2002-2003	Research Fellow, Genome Research Center for Diabetes and Endocrine Disease, Seoul National University Hospital, Seoul, Korea
2004-2005	Teaching Assistant, University of California, San Diego
2005-2006	Research Assistant, University of California, San Diego
2007-2009	Research Assistant, University of California, Los Angeles
2009	Doctor of Philosophy in Computer Science, University of California, San Diego

PUBLICATIONS

Hyun Min Kang, Noah A. Zaitlen, Buhm Han, and Eleazar Eskin, “An adaptive and memory efficient algorithm for genotype imputation”, *In Proceedings of the Thirteenth Annual Conference on Research in Computational Biology (RECOMB-2009).* Tuscon, Arizona: May 18th-21st, 2009

Buhm Han, Hyun Min Kang, and Eleazar Eskin, “Rapid and accurate multiple testing correction and power estimation for millions of correlated markers”, PLoS Genetics, 5:e1000456, 2009

Noah A. Zaitlen, Hyun Min Kang, and Eleazar Eskin, “Linkage effects and analysis of finite sample errors in the HapMap”, Human Heredity, 68:73-86, 2009

Hyun Min Kang, Chun Ye, and Eleazar Eskin “Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots”, Genetics, 180:1909-25, 2008

Eun Yong Kang, Hyun Min Kang, Chun Ye, Ilya Shipster, and Eleazar Eskin, “Detecting the presence and absence of causal relationships between expression of yeast genes with very few samples”, *In Proceedings of the NIPS 2008 Workshop of Machine Learning in Computational Biology*, Whistler, Canada: December 12th, 2008

Buhm Han, Hyun Min Kang, Myeong Seong Seo, Noah A. Zaitlen, and Eleazar Eskin, "Efficient association study via power-optimized tag SNP selection", Annals of Human Genetics, 72:834-47, 2008

Anatole Ghanzalpour, Sudheer Doss, Hyun Min Kang, Charles Farber, Ping-Zi Wen, Alec Brozell, Ruth Castellanos, Eleazar Eskin, Desmond J. Smith, Thomas A. Drake, and Aldons J. Lusis, "High-resolution mapping of gene expression using association in an outbred mouse stock", PLoS Genetics, 4:e1000149, 2008

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, 178:1709-23, 2008

Kelly A. Frazer, Eleazar Eskin, Hyun Min Kang, Molly A. Bogue, David A. Hinds, Erica J. Beilharz, Robert V. Gupta, Julie Montgomery, Matt M. Morenzoni, Geoffrey B. Nilsen, Charit L. Pethiyagoda, Laura L. Stuve, Frank M. Johnson, Mark J. Daly, Claire M. Wade, and David R. Cox, "A sequence-based variation map of 8.27 million SNPs in inbred mouse strains", Nature 448:1050-3, 2007

Noah A. Zaitlen, Hyun Min Kang, Eleazar Eskin, and Eran Halperin, "Leveraging the HapMap correlation structure in association studies", American Journal of Human Genetics, 80:683-91, 2007

Chun Ye, Matthew A. Zapala, Hyun Min Kang, Jennifer Wessel, Eleazar Eskin, and Nicholas J. Schork, "High-density QTL mapping to identify phenotypes and loci influencing gene expression patterns in entire biochemical pathways", *In Proceedings of the Second RECOMB Satellite Worshop of Systems Biology*, San Diego, CA: December 1st-2nd, 2006

Noah A. Zaitlen, Hyun Min Kang, Michael L. Feolo, Stephen T. Sherry, Eran Halperin, and Eleazar Eskin, "Inference and analysis of haplotypes from combined genotyping studies deposited in dbSNP", Genome Research, 15:1594-600, 2005

FIELDS OF STUDY

Major Field: Computer Science
Professor Eleazar Eskin

ABSTRACT OF THE DISSERTATION

Effective Design and Analysis of Systems Genetics Studies

by

Hyun Min Kang
Doctor of Philosophy in Computer Science

University of California San Diego, 2009

Professor Pavel Pevzner, Chair
Professor Eleazar Eskin, Co-Chair

Systems genetics studies for unraveling genetic basis of complex traits have been one of the most propitious research area with the advance of high-throughput biotechnologies. This thesis presents several computational and statistical challenges in effective design and analysis of systems genetics studies and present novel methodological advances and corresponding results in several specific contexts of systems genetics studies.

First, I present an extensive haplotype analysis on a recently collected catalogue of genetic variation among inbred mouse strains, which revealed the contribution from ancestral subspecies, haplotype block structure, and complex history of each genomic segments among the inbred mouse strains. In addition, I accurately imputed the uncollected genotypes in the resource by developing a novel and efficient genotype imputation method which adaptively learns parameters from data using an Expectaion-Maximuzation (EM) algorithm. Our method is demonstrated to outperform previous methods in both mouse and human data.

Statistical analyses in systems genetics studies are often confounded by unmodeled factors such as heterogeneous sample structure. Recent studies suggested that mixed models correct for the sample structure in association mapping, but the available methods suffer from substantial computational cost to be applied in genome-wide association mapping. I developed the Efficient Mixed Model Association (EMMA), which takes advantage of the invariant structure of eigenvectors in applying mixed models for association mapping, which substantially increase the computational efficiency in several orders of magintude. Our method was shown to successfully reduce inflated false positives in *in silico* genome-wide association mapping of inbred mouse strains involving hundreds of thousands of markers. I further extend EMMA to accommodate even larger scale of genome-wide association mapping in humans, typically involving several thousands or more individuals, and demonstrate that the method consistently eliminates the significant over-dispersion of test statistics across multiple human data sets. The method has been further employed in correcting for a different type of confounding effects in expression studies. I developed a novel mixed-model method that corrects for the spurious associations and trans-regulatory bands caused by systematic confounding effects using inter-sample correlation of expression measurements.

Finally, in the design of association studies using inbred strains, I propose a novel trait mapping strategy using hybrid mouse diversity panel (HMDP). By integrating classical inbreds and multiple sets of recombinant inbreds while precisely accounting for the sample structure using high-density markers with EMMA, the proposed design is shown to much more powerfully and precisely identify previously known associations than previous approaches.

Chapter 1

Introduction

The advent of high throughput biotechnology such as sequencing or expression technologies has greatly impacted on many areas in bioscience by enabling us to tackle the problems that had been extremely difficult to address with conventional methods. The availability of extensive resource of genomic, transcriptomic, proteomic, and metabolic data motivated the area of systems biology which aims to systematically understand the complex interaction in biological system, under the philosophy of reductionism[111, 90].

Among various types of high-throughput biological data, array-based genotyping platforms and gene expression platforms have been most popularly used in the past several years[135, 78, 181]. These genotyping and expression arrays provide genome-wide profiles of DNA and mRNA variations respectively, and it is important to understand the relationship between these elements first in order to further understand the complex interaction within the biological system. Using the genetic segregants of budding yeast, Brem et al. [27] carried out genome-wide linkage analysis of expression patterns to reveal the relationship between the DNA variation and mRNA variation. These approaches have been further followed up in many other organisms including human, with a name of "genetical genomics"[98].

One of the main goal of genetic analysis of high-throughput data is to understand the genetic basis of complex traits related to human disease. Direct mapping of these traits with DNA or mRNA variation has also been extensively studied. For example, genome-wide association studies (GWAS) have been popular in the past few years resulting in many identified association between DNA variants and disease phenotypes[223, 143, 13]. Differential expression analysis between different disease outcomes or environmental conditions have been actively studied for several years[208, 190].

Systems genetics is referred to as systems biology involving populations[188, 101]. This is a broader sense than the genetical genomics. Additional high throughput data, phenotypes, and external data such as protein-protein interaction network may be combined in systems genetics studies[228]. Figure 1.1 represent the goal of systems genetics studies. Because comprehensive understanding of the whole system is extremely difficult, current instances of systems genetics studies involves only a few of these layers. For example, genetical genomics studies involves the analysis of the first two layers, and the differential expression analysis involved the mRNA and the phenotype layer. GWAS aims to understand the relationship between DNA and phenotypes. But even such a simple type of analysis confronts many statistical and computational challenges to accurately identify genetic effects due to such as the complex structure of genetic variation, lack of mapping resolution in linkage analysis, heterogeneous sample structure, and technical confounding effects.

In this thesis, I addresses several specific statistical and computational challenges for effective design and analysis of systems genetics studies with proposed solutions. Here I outline specific problems and novel contribution to solve the problems.

In order to perform a systems genetics study effectively, it is important to first understand the structure of genetic variation, which causes the genetic difference between individuals. Of the 3 billion nucleotides of human genome, only a small fraction of nucleotide sequences differ between individuals. The most common type of individual DNA variation is SNP (Single Nucleotide Polymorphism), which represents a variation of a single nucleotide at a particular genomic position mainly due to a single point mutation[114, 36]. Through a substantial collaborative effort, it has been reported that there are at least 3.1 million common SNPs in human, and the total number of SNPs are estimated to be greater than 10 million[93]. These single nucleotide changes typically show local correlation structure within a genomic segment because physically close SNPs tend to be correlated due to linkage disequi-

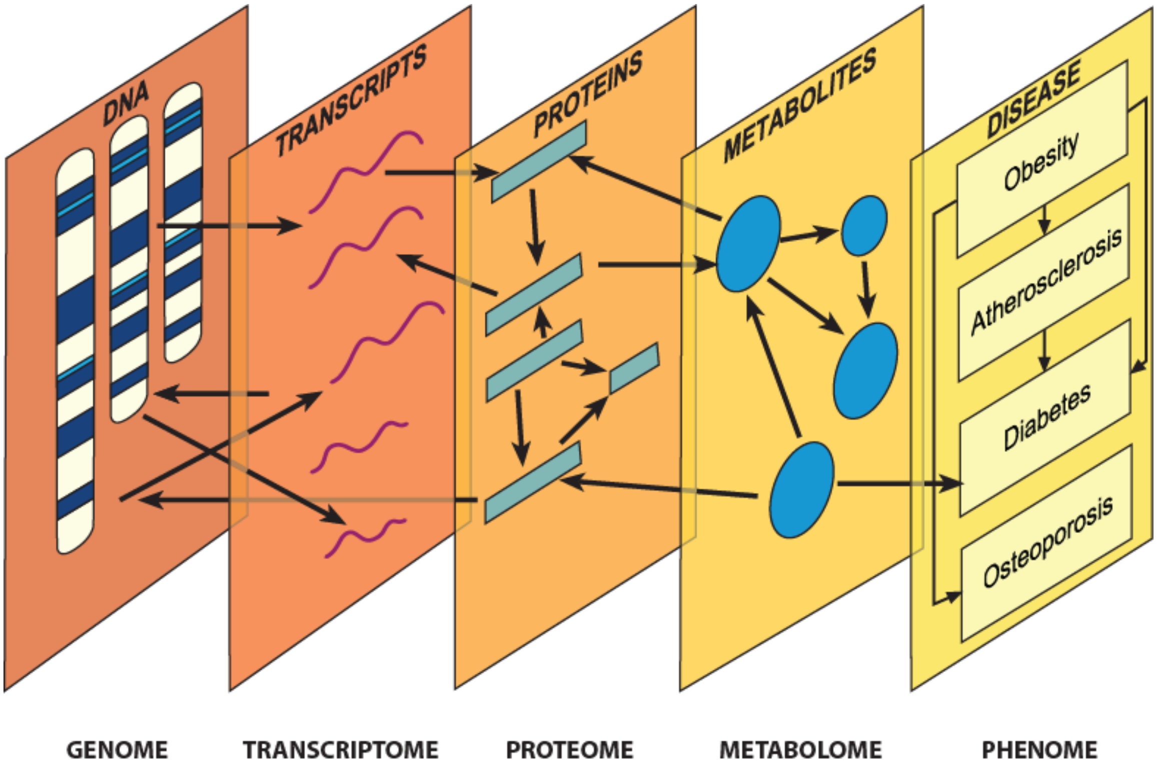


Figure 1.1: A conceptual diagram of systems genetics studies[56]

librium (LD) induced by the high likelihood of linkage between the two loci during recombination. While the structure of human genetic variation has been relatively well understood, the genetic variation in other model organisms such as laboratory mouse, which has a significant implication for human disease trait mapping, has not been comprehensively studied until recently.

Chapter 2 presents the structure of genetic variation among commonly used laboratory mouse strains. I present the analysis of mouse HapMap resource profiled genetic variations among 94 common laboratory mouse. These strains are complexly related with each other because of the heterogeneous contribution from ancestral subspecies, recent population bottleneck, and complex history of hybridization and inbreeding. Understanding the genetic variation structure among the strains at a haplotype level is a very important problem for a comprehensive understanding of genetic variation structure of inbred mouse. By combining the mouse HapMap resouce and the recently collected resequencing-based resource spanning over 8 million SNPs [67], I performed an extensive haplotype analysis which revealed the contribution from each ancestral subspecies per strain, and the patterns of shared segments between each pair of strains, the haplotype block structure among classical inbred mouse strains. These results provide a valuable resource for understanding a detailed history of each genomic segment among the strains. Our results demonstrate that classical inbred strains have a limited genetic diversity due to recent population bottleneck, enabling us to accurately impute the uncollected genotypes using a small set of tag SNPs, comparably to the accuracy of current genotyping technologies. Our analysis also provides detailed genetic relationship among the strains and provide implications for high-resolution association mapping.

Understanding the local correlation structure of genetic variation is very important in the design and analysis of systems genetics studies especially when interpreting a genetic effect associated with a DNA variation at a particular genomic locus. In family-based or segregants-based linkage studies where the local correlation structure are clearly explained by shared segments during recombination, a fraction of genome segment can be interpreted to be 'linked' to the observed phenotypes, meaning one or more of the variants within the segment are significantly associated with the phenotype traits[113, 178]. On the other hand, population-based association studies rely on the linkage disequilibrium (LD) between a causal variant and the nearby markers typically collected by genotyping array to identify 'association' between a marker and a phenotype[22]. By leveraging the local correlation structure, it is also possible to understand the haplotype structure[82, 87] or accurately impute the genotypes at uncollected markers from higher density of reference genotypes[132].

Chapter 3 describes a novel method to impute uncollected genotypes using a high-density reference panel. Standard high-throughput genotyping technologies capture only a fraction of the total genetic variation. Recent efforts have shown that it is possible to "impute" with high accuracy the genotypes of SNPs that are not collected in the study provided that they are present in a reference panel which contains both SNPs collected in the study as well as other SNPs. I introduce a novel hidden Markov Model (HMM) based technique to solve the imputation problem that addresses several shortcomings of existing methods. First, our method is adaptive in the sense that it estimates the HMM parameters from the observed genotype data using an exact EM algorithm rather than using predefined parameters as other methods do. The adaptivity of our method is especially important for the inbred mouse imputation problem described in Chapter 2 where the parameters for recombination and mutation are not well known. Compared to previous methods, our method is up to ten times more accurate on model organisms such as mouse. Second, our algorithm scales in memory usage in the number of collected markers as opposed to the number of known SNPs by utilizing silent states. This issue is very relevant due to the size of the reference data sets currently being generated. I compare our method over mouse and human data sets to existing methods and show that each has either comparable or better performance and much lower memory usage.

In genome-wide association studies to identify associations between genetic variants and disease outcomes or other complex traits, undocumented sample structure among the individuals have been demonstrated to increase false positives and false negatives in association mapping [30, 170]. This sample structure are previously examined in two extreme forms: population structure or cryptic relatedness[50, 211, 170]. Each of these method only partially captures the confounding effects from the complex sample structure, and recently it has been demonstrated that linear mixed model accounts for sample structure in model organisms association mapping where the confounding effect from sample structure is known to be complex and substantial, but the substantial amount of computational cost impeded the application of to large-scale genome-wide association mapping.

In Chapter 4, I present an Efficient Mixed Model Association (EMMA) method that accounts for sample structure with thousands of times higher computational efficiency than previous methods. Our method takes advantage of the specific nature of the optimization problem in applying mixed models for association mapping, which allows us to substantially increase computational speed and reliability of the results with guaranteed convergence properties and global optimization. In particular, the method reduces the time complexity of each iteration of maximum likelihood (ML) or restricted maximum likelihood (REML) from cubic to linear complexity, enabling us to perform genome-wide association mapping in a feasible amount of time. EMMA is shown to robustly correct for inflation of false positives in *in silico* whole genome association mapping of inbred mouse strains involving hundreds of thousands of SNPs.

In the Chapter 5, I extend the EMMA algorithm to accommodate even larger scale of genome-wide association mapping in humans, typically involving several thousands or more individuals. Our method takes advantage of the fact that each loci involved in human disease has a relatively small effect, which allows our method to scale to large samples by only estimating the variance component once during a genome wide scan. Using the Northern Finland Birth Cohort (NFBC66) and Wellcome Trust Case Control Consortium (WTCCC) data sets, we demonstrate that our method consistently eliminates the significant over-dispersion of test statistics that have not been fully resolved by principal component analysis across 17 quantitative and dichotomous phenotypes.

In the genetic analysis of expression data, additional types of systematic confounding effects induces spurious signals and confound the expression quantitative trait loci (eQTL) mapping. Many previous studies suggested that thousands of genes are *trans*-regulated by a small number of genomic regions called “regulatory hotspots”, resulting in “*trans*-regulatory bands” in an eQTL map. As several recent studies have demonstrated, technical confounding factors such as batch effects can complicate eQTL analysis by causing many spurious associations including spurious regulatory hotspots. Yet little is understood about how these technical confounding factors affect eQTL analyses and how to correct for these factors.

In Chapter 6, through an analysis of datasets with biological replicates, I demonstrate that it is this inter-sample correlation structure inherent in expression data that leads to spurious associations between genetic loci and a large number of transcripts inducing spurious regulatory hotspots. I propose a statistical method based on mixed models that corrects for the spurious associations caused by complex inter-sample correlation of expression measurements in eQTL mapping. Due the the computational advances of EMMA described in Chapter 4 and 5, we are able to efficiently perform large-scale eQTL mapping accounting for the complex inter-sample correlation structure using mixed models. Applying our Inter-sample Correlation Emended (ICE) eQTL mapping method to mouse, yeast, and human identifies many more *cis* associations while eliminating most of the spurious *trans* associations. The concordances of *cis* and *trans* associations have consistently increased between different replicates, tissues, and populations; demonstrating the higher accuracy of our method to identify real genetic effects.

These advances in statistical analysis in systems genetics studies can have a great impact in systems genetics studies by enabling us to design a more flexible and powerful study. For example, in the genetic studies of complex traits in mouse studies, due to the substantial amount of sample structure among inbred mouse strains, linkage-based methods have been preferred over association-based methods[64]. One of the major concern in the linkage-based studies have been the mapping resolution to narrow down the region of the associated loci to the gene level. The current linkage studies using F2 progeny or recombinant inbred (RI) strains have the ability to map the associated loci only by tens or several megabases[64]. The power of linkage studies using currently available RI strains is not sufficient to identify modest size of effects explaining the 10% or less variance of phenotypes[39].

In Chapter 7, I present a new design of mouse association mapping by leveraging the ability to precisely account for the complex sample structure using high density genotype markers. Using our Hybrid Mouse Diversity Panel (HMDP), it is possible to combine the classical inbred strains and multiple recombinant inbred strains and perform association mapping by precisely accounting for the pairwise relatedness between the strains obtained from the high-density marker information described in Chapter 2 and 3, and applying the efficient mixed model approach presented in Chapter 4. HMDP achieves both high power and high resolution than any of the currently available single mouse association data set. Our method increases the trait mapping resolution by orders of magnitude while increasing the power of association studies more than twice in most contexts of our power simulation studies. Our integrative approach has a great implication for a more powerful and precise trait mapping in mouse genetic studies and other model organism genetics, and also in human association mapping.

Chapter 2

A high-density haplotype resource of 94 inbred mouse strains

2.1 Motivation

Phenotypic variation among inbred mouse strains exposed to a disease causing agent (be it genetic, infectious or environmental) provides potential insight into human disease processes that often cannot be practically achieved through direct human studies. Indeed hundreds of phenotype measurements related to human diseases are available for dozens of inbred strains[74] in common use over the last 50-100 years. As with the direct study of chronic disease in humans, a key step towards determining the genetic underpinnings of this phenotypic variation is to develop a catalogue of the genetic variation among inbred mouse strains and interpreting the structure of variation patterns across the strains. Recent advances in high-throughput genotyping and DNA resequencing technologies are making it possible to rapidly uncover the genetic variation maps of many model organisms[93, 121, 23, 193, 67]. A recent whole genome resequencing study of 15 inbred mouse strains captured a significant fraction of the genetic variation among a limited number of strains allowing researchers to infer patterns of genetic variation and identify the ancestral origin of the genetic variation[67, 232]. Yet, the availability and common experimental employment of hundreds of inbred strains, including over 190 stocks available from the Jackson Laboratory, motivates the need of a high-density variation map for a larger set of strains.

2.2 Results

2.2.1 The mouse HapMap resource

We have assembled a dense set of genotypes for a total of 138,980 unique biallelic single nucleotide polymorphisms (SNPs) in 94 inbred mouse strains, at an average spacing of 20kb on chromosomes 1-19 and X. We selected the most commonly used inbred laboratory strains - especially targeting priority strains from the Mouse Phenome Database[74] - and 19 wild-derived strains both as reference out-groups and to help identify ancestry of genomic segments. Our dataset is a composite of 121,433 SNPs discovered and genotyped at the Broad Institute by comparing data from the two inbred mouse genome sequencing projects[146, 145], with additional discovery in a wild-derived strain in regions of low marker density; 7,570 SNPs covering physical gaps in the Broad Institute map revealed by examining data from the concurrent NIEHS/Perlegen effort to resequence 15 inbred strains[67] and also genotyped at the Broad Institute; and 13,094 SNPs discovered and genotyped at the Wellcome Trust Center for Human Genetics (WTCHG) that could be mapped to Build 37 of the mouse genome.

To evaluate the quality of these resources, we examined SNPs typed in common by Broad and WTCHG as well as compared each resource to the genotypes of strains produced from the NIEHS/Perlegen sequence data. SNPs overlapping between the Broad and WTCHG sets demonstrate a discordance rate of 0.00058, while SNPs overlapping WTCHG and NIEHS/Perlegen sequence-based genotypes demonstrate a discordance of 0.00688. The extremely high concordance of the Broad and WTCHG data and significantly higher accuracy than the array-based sequence genotypes are unsurprising; the Broad and WTCHG utilized established SNP genotyping techniques and need only distinguish between two homozygous genotype classes. An interesting disparity in discordance rate is observed between Perlegen and WTCHG genotypes. When the WTCHG genotype is the reference strain allele (C57BL/6J) the disparity with Perlegen genotype is 0.00335 and is 0.0106 otherwise. This is

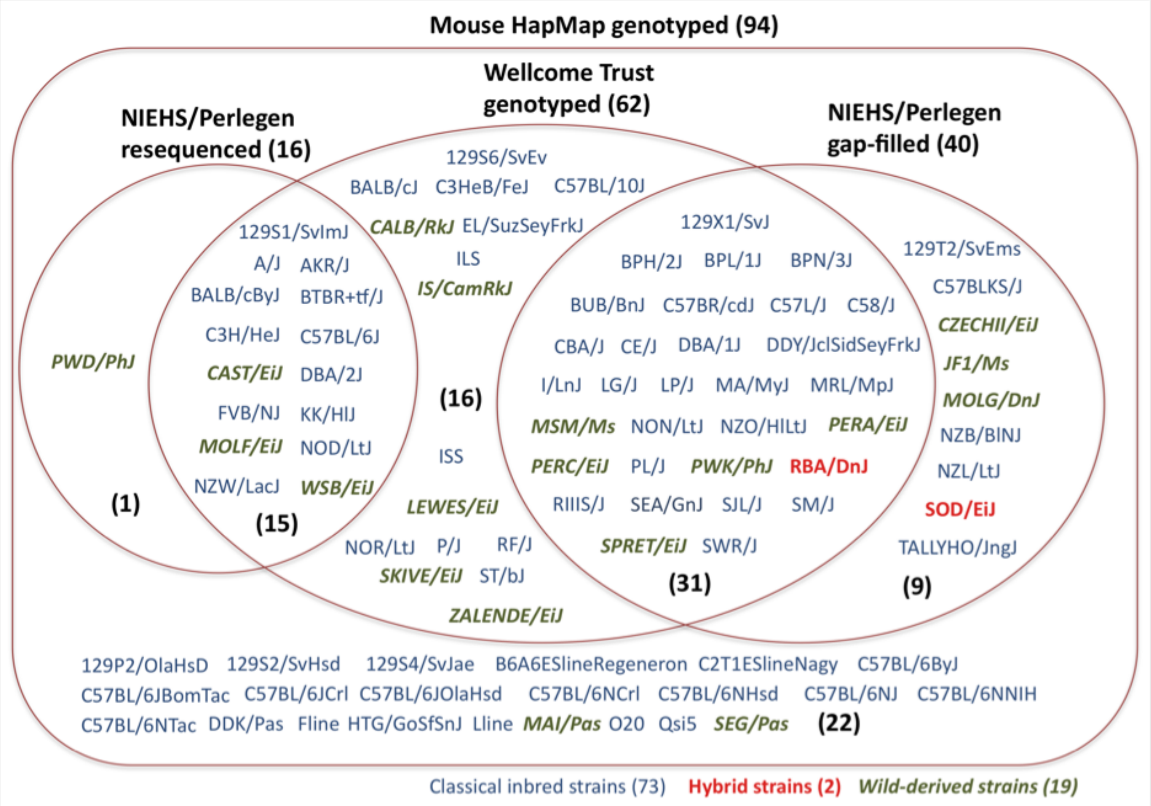


Figure 2.1: Classification of 94 strains used in the mouse HapMap projects based on the availability in other resources, including 8.27 million NIEHS/Perlegen resequencing-based SNPs, WTCHG SNPs, and additional gap-filling SNPs. (C57BL/6J is not included in the 15 resequenced strains, but it is the reference strain that has been fully resequenced)

consistent with the variant discovery strategy employed by Perlegen, which emphasized low false positive variant discovery at the expense of a higher false negative rate[93, 67]. Figure 2.1 and Table 2.1 summarizes the genotype resources for each of the 94 strains.

Table 2.1: List of strains used in mouse HapMap projects and the availability in other resources *C57BL/6J is not included in the 15 resequenced strain, but it is the reference strain that has been fully sequenced (Continued to next page)

Strain name	Perlegen resequenced	WTCHG genotyped	Additional gap-filling	Wild-derived or classical
129P2/OlaHsD	X	X	X	IN

(Contintued to next page)

Table 2.1: (Continued from previous page) List of strains used in mouse HapMap projects and the availability in other resources *C57BL/6J is not included in the 15 resequenced strain, but it is the reference strain that has been fully sequenced

Strain	Perlegen	WTCHG	Additional	Wild-derived
name	resequenced	genotyped	gap-filling	or classical
129S1/SvImJ	O	O	X	IN
129S2/SvHsd	X	X	X	IN
129S4/SvJae	X	X	X	IN
129S6/SvEv	X	O	X	IN
129T2/SvEms	X	X	O	IN
129X1/SvJ	X	O	O	IN
A/J	O	O	X	IN
AKR/J	O	O	X	IN
B6A6ESlineRegeneron	X	X	X	IN
BALB/cByJ	O	O	X	IN
BALB/cJ	X	O	X	IN
BPH/2J	X	O	O	IN
BPL/1J	X	O	O	IN
BPN/3J	X	O	O	IN
BTBRT<+>tf/J	O	O	X	IN
BUB/BnJ	X	O	O	IN
C2T1ESlineNagy	X	X	X	IN
C3H/HeJ	O	O	X	IN
C3HeB/FeJ	X	O	X	IN
C57BL/10J	X	O	X	IN
C57BL/6ByJ	X	X	X	IN
C57BL/6J	O*	O	X	IN
C57BL/6JBomTac	X	X	X	IN
C57BL/6JCrl	X	X	X	IN
C57BL/6JOlaHsd	X	X	X	IN
C57BL/6NCrl	X	X	X	IN
C57BL/6NHsd	X	X	X	IN

(Continutued to next page)

Table 2.1: (Continued from previous page) List of strains used in mouse HapMap projects and the availability in other resources *C57BL/6J is not included in the 15 resequenced strain, but it is the reference strain that has been fully sequenced

Strain	Perlegen	WTCHG	Additional	Wild-derived
name	resequenced	genotyped	gap-filling	or classical
C57BL/6NJ	X	X	X	IN
C57BL/6NNIH	X	X	X	IN
C57BL/6NTac	X	X	X	IN
C57BLKS/J	X	X	O	IN
C57BR/cdJ	X	O	O	IN
C57L/J	X	O	O	IN
C58/J	X	O	O	IN
CALB/RkJ	X	O	X	WI
CAST/EiJ	O	O	X	WI
CBA/J	X	O	O	IN
CE/J	X	O	O	IN
CZECHII/EiJ	X	X	O	WI
DBA/1J	X	O	O	IN
DBA/2J	O	O	X	IN
DDK/Pas	X	X	X	IN
DDY/JclSidSeyFrkJ	X	O	O	IN
EL/SuzSeyFrkJ	X	O	X	IN
FVB/NJ	O	O	X	IN
Fline	X	X	X	IN
HTG/GoSfSnJ	X	X	X	IN
I/LnJ	X	O	O	IN
ILS	X	O	X	IN
IS/CamRkJ	X	O	X	WI
ISS	X	O	X	IN
JF1/Ms	X	X	O	WI
KK/HlJ	O	O	X	IN
LEWES/EiJ	X	O	X	WI

(Continuted to next page)

Table 2.1: (Continued from previous page) List of strains used in mouse HapMap projects and the availability in other resources *C57BL/6J is not included in the 15 resequenced strain, but it is the reference strain that has been fully sequenced

Strain	Perlegen	WTCHG	Additional	Wild-derived
name	resequenced	genotyped	gap-filling	or classical
LG/J	X	O	O	IN
LP/J	X	O	O	IN
Lline	X	X	X	IN
MA/MyJ	X	O	O	IN
MAI/Pas	X	X	X	WI
MOLF/EiJ	O	O	X	WI
MOLG/DnJ	X	X	O	WI
MRL/MpJ	X	O	O	IN
MSM/Ms	X	O	O	WI
NOD/LtJ	O	O	X	IN
NON/LtJ	X	O	O	IN
NOR/LtJ	X	O	X	IN
NZB/B1NJ	X	X	O	IN
NZL/LtJ	X	X	O	IN
NZO/HlLtJ	X	O	O	IN
NZW/LacJ	O	O	X	IN
O20	X	X	X	IN
P/J	X	O	X	IN
PERA/EiJ	X	O	O	WI
PERC/EiJ	X	O	O	WI
PL/J	X	O	O	IN
PWD/PhJ	O	X	X	WI
PWK/PhJ	X	O	O	WI
Qsi5	X	X	X	IN
RBA/DnJ	X	O	O	HY
RF/J	X	O	X	IN
RIIIS/J	X	O	O	IN

(Continuted to next page)

Table 2.1: (Continued from previous page) List of strains used in mouse HapMap projects and the availability in other resources *C57BL/6J is not included in the 15 resequenced strain, but it is the reference strain that has been fully sequenced

Strain name	Perlegen resequenced	WTCHG genotyped	Additional gap-filling	Wild-derived or classical
SEA/GnJ	X	O	O	IN
SEG/Pas	X	X	X	WI
SJL/J	X	O	O	IN
SKIVE/EiJ	X	O	X	WI
SM/J	X	O	O	IN
SOD1/EiJ	X	X	O	HY
SPRET/EiJ	X	O	O	WI
ST/bJ	X	O	X	IN
SWR/J	X	O	O	IN
TALLYHO/JngJ	X	X	O	IN
WSB/EiJ	O	O	X	WI
ZALENDE/EiJ	X	O	X	WI

2.2.2 Haplotype structure among the strains

Using these genotype resources, we are able to examine the fine-level haplotype structure among the strains. For example, a comparison of the six 129 strains shows that they share the vast majority of their genomic segments, but that there are several noticeable differences. In particular, there is a large disparity between 129P2/OlaHsD and 129X1/SvJ from 35Mb to 100Mb on chromosome 7, and there are also differences specific to 129S6/SvEv on chromosomes 3, 5, and 12. Similarly, comparisons between the fifteen C57 strains revealed significant discrepancies between C57BL/6J and the other C57 strains. We also identified that some strains appear to result from recent hybridizations between two or more strains. We observed that HTG/GoSfSnJ shares more than 99.9% of genome with either BALB/cByJ or C57BL/6J, and that NOR/LtJ shares more than 99.9% segments with either

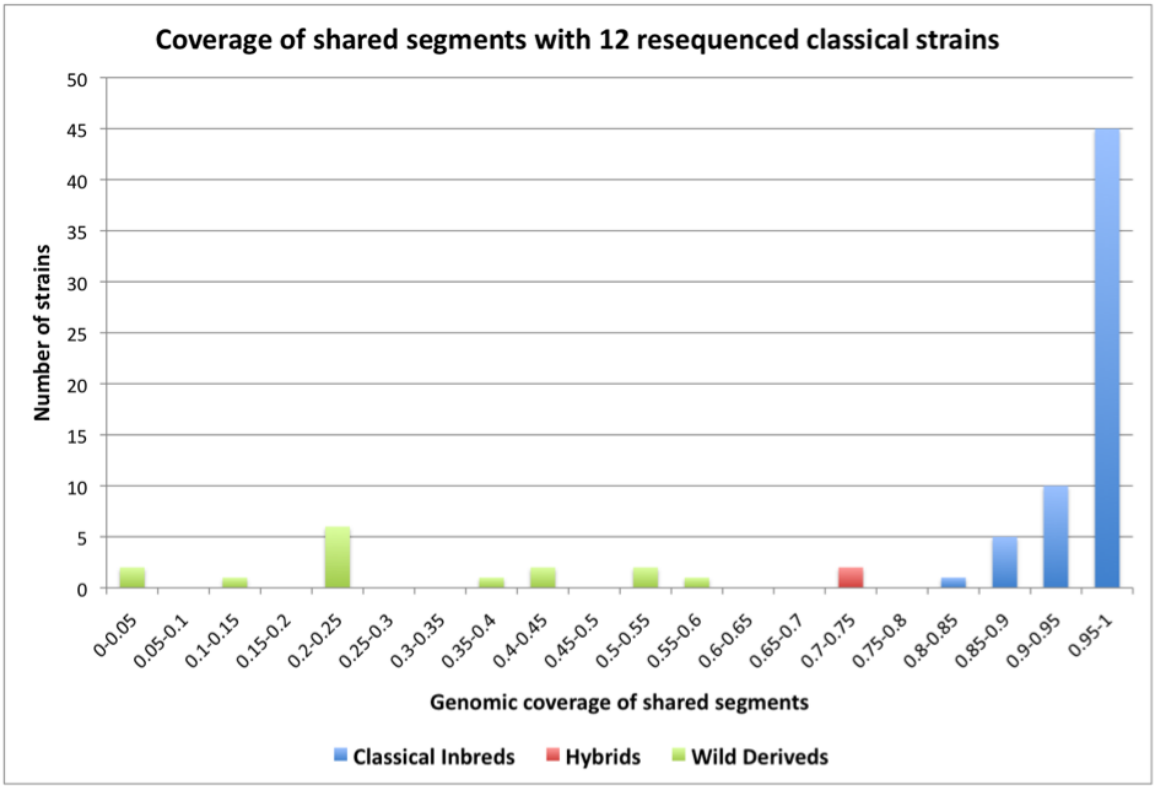


Figure 2.2: A histogram of the fractions of genome covered by shared segments with one of the 12 classical inbred strains, over 78 non-resequenced mouse HapMap strains. The classical inbred strains are colored in blue, the hybrid strains in red, and the wild-derived strains in green

NOD/LtJ or C57BLKS/J, confirming the annotated genealogical history[200]. We also observed that two strains (RBA/DnJ and SOD/EiJ) are âĂİhybridâĂİ strains with genetic content from both classical inbred and wild-derived strains. When comparing the fraction of genome shared by any of the 12 classical inbred resequenced strains, there is a clear difference between rates of sharing with the wider set of classical inbred strains (97% of the genome on average and 81% minimum) and with the wild-derived strains (28% on average, 56% maximum) (Figure 2.2).

We allocated ancestry of local genomic regions to one of the four "founder" strains using the methods described previously for resequencing data[67]. For each of the remaining 90 strains, we identified the fractions of genomic regions unequivocally close to *domesticus*, *musculus*, *castaneus*, and *molossinus* strains. On average these ancestral strains contribute 32.3%, 9.19%, 4.52%, and 11.8%, respectively. 42.2% of observed total genomic regions are ambiguous for ancestry, meaning either that the ancestry is not precisely represented by any of the four founder strains (37.3%), or else that two or more ancestral sub-species share haplotypes in these regions (4.86%). The fractions of regions identified as having domesticus or unknown ancestry differ from previous studies[67] due to the sparser resolution of the SNP map, and the SNP ascertainment bias inherent in both current and former datasets. All of the classical inbred strains and hybrid strains share predominantly domesticus ancestry, while the wild-derived strains are divided into four groups corresponding to their respective ancestral subspecies.

To investigate the average sizes of shared haplotype segments among strains, we identified common (low SNP density) and divergent (high SNP density) ancestral segments across the genome for each pair of inbred strains using a hidden Markov model[67]. Among the 4371 possible pair-wise comparisons of the 94 strains, an average of 32.5% of the genomic regions are shared between any pair of strains (Figure 2.3). The average number of shared ancestral segments genome-wide is 280 per comparison, which is about one segment per 10Mb. On average, there are 176 segments longer than 1Mb covering 28.8% of the genome, and 39 segments longer than 5Mb covering 15.6% of the genome - reflective of the tight recent co-ancestry of these strains. Given a cross between any of the two parental strains, it is possible to estimate the genomic region excluded from mapping variations associated with phenotype traits due to the shared segments between them. For example, among BXD recombinant-inbred strains, 48.6% of genomic regions are excluded for the mapping.

To ascertain whether intervening genotypes might be successfully imputed from the resequencing data, we counted how many distinct haplotypic segments exist for each genomic region and compared this with the numbers derived from the resequencing data by combining the shared segment analysis using hierarchical clustering. The average number of distinct segments within any region is estimated to be 4.70 over 73 classical inbred strains. This limited diversity likely reflects recent bottlenecks, where a limited number of chromosomes from the founder strains gave rise to the modern inbred strains[67, 213, 68]. Among the 12 resequenced classical inbred strains, an average of 3.46 ancestral segments were identified. Like the analysis

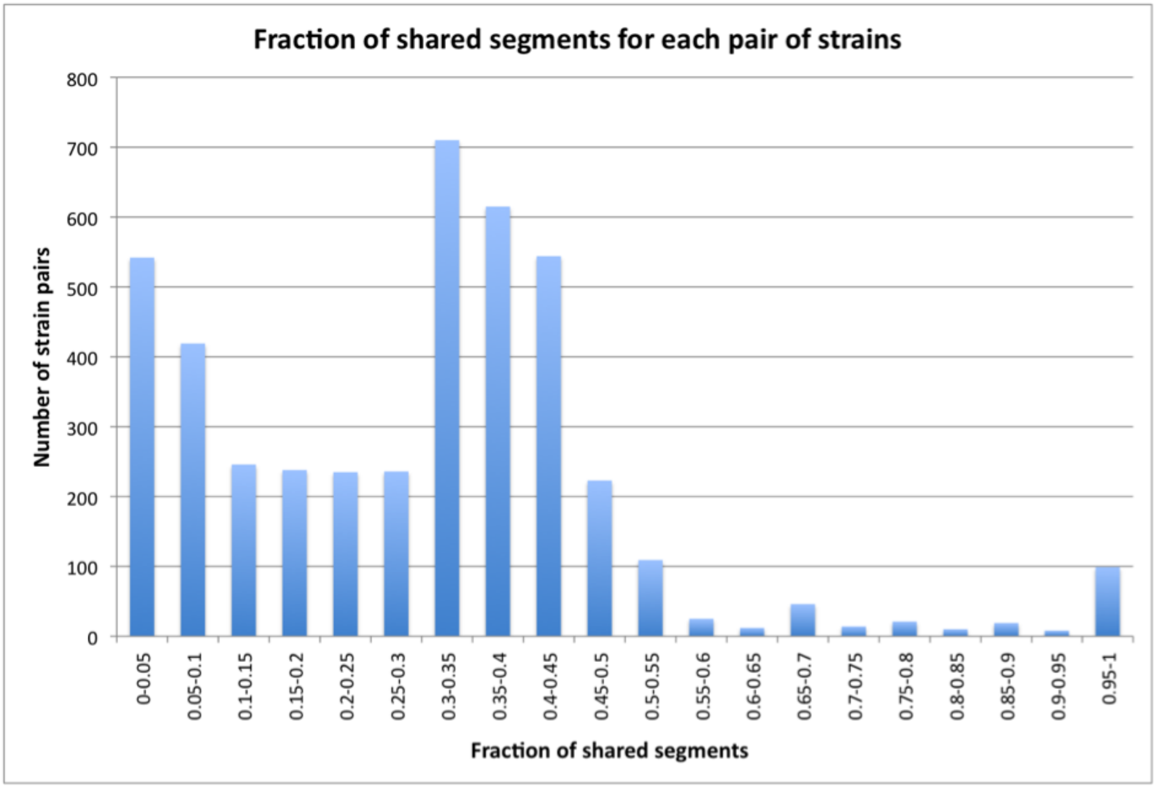


Figure 2.3: A histogram of the fractions of shared genomic segments between each of 4371 pairs between the 94 strains

of shared segments, these results suggest that most of the genetic variation existing among the classical inbred strains can be explained by the variation present in the resequenced strains.

2.2.3 Integrating NIEHS/Perlegen resequencing and HapMap data

Now confident that we could identify segment ancestry by reference to the 16 resequenced strains, we proceeded to impute genotypes for the 8.27 million NIEHS/Perlegen SNPs on the 78 genotyped strains using a hidden Markov model that learns genome wide transition and mutation parameters using Expectation-Maximization (EM) algorithm[47]. We were able to call the majority of SNPs (79.2%) with high confidence (posterior probability > 0.98), when genotypes were successfully called from the all 16 resequenced strains (see Table 2.2, top panel for details). We found that confidence scores vary greatly, with 11 wild-derived strains having no high-confidence imputed genotypes because their estimated mutation rates were very high. In contrast, all 9 strains with the C57BL/6 prefix have more than 99.7% of high-confidence call rate, due to their genetic proximity to the reference strain C57BL/6J. We were also able to impute genotypes missing in the 16 resequenced strains, but only 17.2% of these with high confidence due to poor probe quality resulting in unreliable data (Table 2.2, bottom panel).

We estimated the accuracy of our imputed genotypes in two different ways. First, we used a leave-one-out cross-validation approach to impute genotypes for each of the 16 resequenced strains using the remainder. When considering the SNPs with complete data in the resequenced strains, the average leave-one-out imputation error over the 12 classical inbred resequenced strains was 1.59%, dropping to 0.27% when only high confidence genotypes were used (Table 2.3). We found that these rates varied substantially between the 12 classical inbred strains (range 0.60% - 2.67%; high-confidence genotype error range 0.10% - 0.57%). The call rate of high confidence genotypes also varied, ranging from 84.6% to 97.0%. These errors increase when considering the four wild-derived strains, with total imputation error ranging from 10.9% to 33.4%. These error differences likely reflect the divergent ancestry of the

Table 2.2: Classification of imputed genotypes that are untyped or experimentally missing. The fraction of imputed genotypes in each category is shown within a parenthesis

Category	NIEHS/Perlegen	High	Medium	Low	All
	SNP quality	confidence	confidence	confidence	confidence
Untyped	Fully	235,728,507	48,532,073	13,431,178	297,691,758
8 million	resequenced	(36.7%)	(7.57%)	(2.09%)	(46.4%)
NIEHS/Perlegen	Mostly	137,628,908	34,464,866	21,237,494	193,331,268
genotypes	resequenced	(21.5%)	(5.37%)	(3.31%)	(30.2%)
over 78	Poorly	72,753,547	25,350,239	52,284,738	150,388,524
non-resequenced	resequenced	(11.3%)	(3.95%)	(8.15%)	(23.4%)
strains	Total	446,110,962	108,347,178	86,953,410	641,411,550
		(69.5%)	(16.9%)	(13.6%)	(100%)
Experimentally	Mostly	1,109,113	959,986	1,316,561	3,384,660
missing	resequenced	(7.58%)	(6.56%)	(9.00%)	(23.1%)
NIEHS/Perlegen	Poorly	1,407,303	1,753,637	8,077,233	11,238,223
genotypes	resequenced	(9.62%)	(12.0%)	(55.2%)	(76.9%)
over 16	Total	2,516,416	2,712,673	9,393,794	14,622,883
strains		(17.2%)	(18.6%)	(64.2%)	(100%)
Mouse HapMap	Total	744,725	263,196	257,847	1,265,768
missing genotypes		(58.8%)	(20.8%)	(20.4%)	(100%)
Grand		449,372,103	111,323,047	96,605,051	657,300,201
total		(68.4%)	(16.9%)	(14.7%)	(100%)

imputed strains, as the marker set remains biased towards the strains used for SNP discovery. Next, we estimated accuracy by comparing our imputed genotypes to data previously generated by the WTCHG on 47 of the 78 genotyped strains, and found a total error rate of 4.86% (2.26% when excluding the 11 wild-derived and hybrid strains). Restricting to the 71.7% of the imputed genotypes called at high confidence genotypes reduces this error to 0.37%, more than ten times smaller than recently published results for this marker subset using a different method[200]. As in the previous error estimate, the imputation error again differs greatly by strain, ranging from 0.065% to 20.9% (0.019% to 4.41% for high confidence imputed genotypes).

In summary, we were able to impute 657,300,201 genotypes across 8.27 million markers in 94 inbred strains, including 14,622,883 experimentally missing genotypes in the resequencing strains and 1,265,768 genotypes missing in the combined genotype sets. This creates a near-comprehensive snapshot of variation in commonly available mouse strains.

Table 2.3: Leave-one-out imputation error rates of 12 resequenced classical inbred strains using mouse HapMap SNPs, WTCHG SNPs, and gap-filling Perlegen SNPs. The fraction of imputed genotypes in each category is shown within a parenthesis.

SNP quality	High	Medium	Low	Total
	confidence	confidence	confidence	
Fully	0.27%	6.40%	19.0%	1.59%
resequenced	(46.1%)	(2.79%)	(2.73%)	(51.7%)
Mostly	0.40%	3.94%	16.1%	2.26%
resequenced	(25.3%)	(3.50%)	(2.98%)	(31.8%)
Poorly	0.76%	4.95%	15.8%	5.18%
resequenced	(9.59%)	(2.62%)	(4.29%)	(16.5%)
Total	0.38%	4.74%	16.8%	2.40%
	(81.1%)	(8.91%)	(10.0%)	(100%)

Table 2.4: Total number of imputed genotypes of 16 resequenced strains missing in NIEHS/Perlegen SNPs, mouse HapMap SNPs, WTCHG SNPs, and the gap-filling NIEHS/Perlegen SNPs, in addition to the estimation of imputation errors using leave-one-out cross-validation

Strain	# imputed	# high-confidence	Overall	High-confidence
name	genotypes	genotypes	imputation error	imputation error
129S1/SvImJ	923,959	245,792	0.02505	0.00379
A/J	797,363	164,028	0.01473	0.00195
AKR/J	899,745	213,076	0.01976	0.00288
BALB/cByJ	856,126	225,381	0.01172	0.00211
BTBRT+tf/J	905,104	270,244	0.02092	0.00430
C3H/HeJ	987,508	282,717	0.01181	0.00289
C57BL/6J	14	0	0.03854	0.00369
CAST/EiJ	1,257,795	0	0.34073	N/A
DBA/2J	956,054	272,074	0.01874	0.00341
FVB/NJ	914,948	208,107	0.02886	0.00395
KK/HlJ	908,091	211,090	0.03628	0.00674
MOLF/EiJ	1,226,418	0	0.15148	N/A
NOD/LtJ	929,392	224,838	0.02584	0.00379
NZW/LacJ	905,125	214,152	0.03395	0.00542
PWD/PhJ	1,312,167	0	0.16713	N/A
WSB/EiJ	943,362	5,665	0.13027	N/A
Total (or Avg.)	14,733,063	2,542,320	0.06156	0.00367

Table 2.5: Total number of imputed genotypes missing in NIEHS/Perlegen SNPs, mouse HapMap SNPs, WTCHG SNPs, and the gap-filling NIEHS/Perlegen SNPs, in addition to the coverage of shared segments and the imputation errors over 78 non-resequenced strains

Strain	# imputed		% Miscoverage of		Imputation error	
name	genotypes		shared segments		using WTCHG SNPs	
	overall	high-confidence	16 strains	12 strains	overall	high-confidence
129P2/OlaHsD	8,245,441	7,230,952	0.00059	0.00059	0.00335	0.00094
129S2/SvHsd	8,244,722	7,307,195	0.00000	0.00000		
129S4/SvJae	8,245,906	7,314,437	0.00000	0.00000		
129S6/SvEv	8,234,178	7,223,744	0.00000	0.00000		
129T2/SvEms	8,243,699	7,224,079	0.00000	0.00000		
129X1/SvJ	8,225,454	7,140,392	0.00059	0.00059	0.00602	0.00104
B6A6ESlineRegeneron	8,245,745	8,230,951	0.00000	0.00000		
BALB/cJ	8,235,566	7,184,094	0.00000	0.00000	0.02036	0.00019
BPH/2J	8,228,527	6,977,701	0.00000	0.00000	0.01104	0.00187
BPL/1J	8,232,603	6,906,869	0.00117	0.00117	0.01430	0.00286
BPN/3J	8,237,827	6,900,403	0.00351	0.00351	0.01548	0.00250
BUB/BnJ	8,226,291	6,305,478	0.01992	0.01992	0.02731	0.00454
C2T1ESlineNagy	8,245,903	8,229,596	0.00000	0.00000		
C3HeB/FeJ	8,233,065	7,207,943	0.00000	0.00000	0.00065	0.00037
C57BL/10J	8,233,995	7,948,029	0.00000	0.00000	0.00964	0.00597
C57BL/6ByJ	8,247,008	8,231,062	0.00000	0.00000		
C57BL/6JBomTac	8,248,037	8,239,698	0.00000	0.00000		
C57BL/6JCrl	8,243,609	8,237,257	0.00000	0.00000		
C57BL/6JOlaHsd	8,244,352	8,236,971	0.00000	0.00000		
C57BL/6NCrl	8,244,887	8,229,351	0.00000	0.00000		
C57BL/6NHsd	8,244,020	8,227,807	0.00000	0.00000		
C57BL/6NJ	8,244,963	8,230,202	0.00000	0.00000		
C57BL/6NNIH	8,244,167	8,229,485	0.00000	0.00000		
C57BL/6NTac	8,245,953	8,229,946	0.00000	0.00000		
C57BLKS/J	8,236,725	7,723,976	0.00000	0.00000		
C57BR/cdJ	8,226,716	6,965,038	0.04277	0.09080	0.02635	0.00428
C57L/J	8,227,549	6,916,320	0.04277	0.09080	0.02599	0.00459
C58/J	8,225,696	6,862,822	0.05214	0.11834	0.02523	0.00490
CALB/RkJ	8,244,873	0	0.39602	0.56298	0.13004	N/A
CBA/J	8,227,287	6,945,743	0.00000	0.00000	0.01009	0.00251
CE/J	8,228,150	5,251,341	0.13064	0.19215	0.06007	0.00749
CZECHII/EiJ	8,246,104	0	0.00293	0.78617		
DBA/1J	8,226,067	7,027,623	0.00000	0.00000	0.00787	0.00201
DDK/Pas	8,244,842	6,053,745	0.08436	0.08494		
DDY/JclSidSeyFrkJ	8,236,772	6,303,940	0.06503	0.06503	0.02930	0.00333
EL/SuzSeyFrkJ	8,235,603	6,258,752	0.03866	0.03866	0.02807	0.00482
Fline	8,250,432	5,967,688	0.04218	0.04218		
HTG/GoSfSnJ	8,245,769	7,297,844	0.00000	0.00000		
I/LnJ	8,227,162	6,046,611	0.07967	0.10486	0.03574	0.00429
ILS	8,233,633	6,890,084	0.00000	0.00000	0.01324	0.00148

(Continued to the next page)

Table 2.5: (Continued from previous page) Total number of imputed genotypes missing in NIEHS/Perlegen SNPs, mouse HapMap SNPs, WTCHG SNPs, and the gap-filling NIEHS/Perlegen SNPs, in addition to the coverage of shared segments and the imputation errors over 78 non-resequenced strains

Strain	# imputed		% Miscoverage of		Imputation error	
name	genotypes		shared segments		using WTCHG SNPs	
	overall	high-confidence	16 strains	12 strains	overall	high-confidence
IS/CamRkJ	8,236,630	0	0.26889	0.47686	0.14918	N/A
ISS	8,233,475	6,867,012	0.00996	0.00996	0.01806	0.00282
JF1/Ms	8,246,764	5,444,331	0.00000	0.78676		
LEWES/EiJ	8,235,345	977,677	0.21324	0.44757	0.16626	0.00919
LG/J	8,227,062	6,128,027	0.00996	0.00996	0.03319	0.00534
LP/J	8,226,598	7,029,170	0.00059	0.00059	0.00806	0.00182
Lline	8,249,621	5,957,733	0.08260	0.08260		
MA/MyJ	8,226,967	6,300,478	0.04511	0.06503	0.03063	0.00495
MAI/Pas	8,253,960	0	0.00059	0.79262		
MOLG/DnJ	8,244,056	6,351,851	0.00000	0.87756		
MRL/MpJ	8,230,636	6,454,050	0.01054	0.01054	0.02311	0.00397
MSM/Ms	8,238,358	5,577,646	0.00059	0.78676	0.04643	0.01150
NON/LtJ	8,229,599	6,349,568	0.05858	0.05858	0.02463	0.00334
NOR/LtJ	8,242,454	7,316,242	0.00000	0.00000	0.00241	0.00038
NZB/B1NJ	8,239,399	5,965,759	0.09315	0.09315		
NZL/LtJ	8,246,177	6,081,332	0.10076	0.10896		
NZO/HlLtJ	8,234,925	6,156,349	0.11013	0.11775	0.03378	0.00537
O20	8,247,518	5,736,937	0.09022	0.10076		
P/J	8,239,416	6,119,956	0.00000	0.00000	0.03595	0.00588
PERA/EiJ	8,231,368	0	0.42004	0.64968	0.14189	N/A
PERC/EiJ	8,229,729	0	0.40949	0.59930	0.20895	N/A
PL/J	8,225,812	6,415,946	0.05858	0.05858	0.02221	0.00411
PWK/PhJ	8,242,952	6,370,734	0.00000	0.78735	0.16018	0.04412
Qsi5	8,246,243	6,202,186	0.00879	0.00879		
RBA/DnJ	8,229,526	2,318,943	0.17047	0.28881	0.11836	0.00834
RF/J	8,239,999	6,532,284	0.00000	0.00000	0.02189	0.00373
RIIIS/J	8,226,755	5,802,294	0.02636	0.02636	0.04037	0.00483
SEA/GnJ	8,227,130	6,637,917	0.00410	0.00410	0.02021	0.00285
SEG/Pas	8,267,796	0	0.51787	1.00000		
SJL/J	8,226,610	6,344,651	0.01875	0.01875	0.02377	0.00350
SKIVE/EiJ	8,242,754	0	0.00293	0.79262	0.10861	N/A
SM/J	8,235,831	5,511,067	0.01465	0.03281	0.05459	0.00757
SOD1/EiJ	8,241,898	0	0.20152	0.27651		
SPRET/EiJ	8,254,823	0	0.40773	1.00000	0.15051	N/A
ST/bJ	8,233,413	6,021,016	0.04159	0.04159	0.03758	0.00670
SWR/J	8,226,433	6,051,249	0.02695	0.02695	0.03176	0.00454
TALLYHO/JngJ	8,238,580	5,872,209	0.02519	0.05682		
ZALENDE/EiJ	8,241,228	0	0.29525	0.48975	0.18275	N/A
Total (or Avg.)	642,567,138	446,829,783	0.06087	0.16907	0.04859	0.00374

Table 2.6: Imputation error rates of 47 inbred strains genotyped only in WTCHG SNPs, using mouse HapMap SNPs, and gap-filling Perlegen SNPs. The fraction of imputed genotypes in each category is shown within a parenthesis

SNP quality	High confidence	Medium confidence	Low confidence	Total
36 classical inbreds	0.35% (88.9%)	9.63% (6.74%)	29.7% (4.37%)	2.25% (100%)
All 47 strains	0.37% (71.7%)	8.85% (16.7%)	27.0% (11.5%)	4.86% (100%)

2.2.4 Effects of larger resources

To estimate the cost-effectiveness of expanding this resource, we evaluated the potential imputation coverage made possible by either increasing the number of resequenced strains or the number of SNPs in the HapMap.

To determine the effect of using a larger number of resequenced strains, we assumed that the 62 WTCHG strains were all resequenced, and estimated the imputation accuracy of imputing the WTCHG genotypes from the mouse HapMap SNPs and the gap-filling SNPs, using the leave-one-out cross-validation for each of the 62 strains. Because each strain targeted for imputation now has 61 instead of 16 reference strains, the imputation accuracy is expected to be high. Overall, the errors are reduced from 4.86% to 2.45%, and the errors in the 36 classical inbred strains are reduced from 2.25% to 0.96%. In contrast, the accuracy in the high-confidence genotypes of the classical inbred strains is reduced from 0.35% to 0.16%. More importantly, high-confidence call rate was increased from 88.9% to 95.6% for the 36 classical inbred strains, and from 71.7% to 84.9% for all 47 strains. Several strains such as MRL/MpJ, C57L/J, C57BR/cdJ, PERA/EiJ and PWK/EiJ showed a substantial improvement in imputation accuracy when a larger set of reference strains was used, while many other wild-derived strains still retained imputation errors of greater than 10%.

Since the resequencing of more strains is expected to increase the imputation coverage significantly, we prioritized the strains that might be targeted for resequencing to improve the coverage, based on our analysis of the shared segments. To do

Table 2.7: Leave-one-out imputation error rates of 12 resequenced classical inbred strains using mouse HapMap SNPs only, and combining mouse HapMap SNPs and gap-filling NIEHS/Perlegen SNPs. The fraction of imputed genotypes in each category is shown within a parenthesis

Category	NIEHS/Perlegen SNP quality	High confidence	Medium confidence	Low confidence	All confidence
Mouse HapMap SNPs only	Fully resequenced	0.33% (46.3%)	7.18% (3.08%)	21.3% (2.25%)	1.66% (51.7%)
	Mostly resequenced	0.47% (25.0%)	4.16% (4.09%)	16.8% (2.73%)	2.35% (31.8%)
	Poorly resequenced	0.87% (9.35%)	4.21% (2.85%)	15.7% (4.29%)	5.31% (16.5%)
	Total	0.44% (80.7%)	4.74% (8.91%)	16.8% (10.0%)	2.48% (100%)
Mouse HapMap SNPs + gap-filling SNPs	Fully resequenced	0.27% (45.6%)	5.82% (3.14%)	18.9% (2.85%)	1.63% (51.7%)
	Mostly resequenced	0.39% (25.0%)	3.79% (3.77%)	16.2% (3.04%)	2.30% (31.8%)
	Poorly resequenced	0.76% (9.35%)	3.97% (2.71%)	15.8% (4.32%)	5.23% (16.5%)
	Total	0.37% (80.2%)	4.50% (9.62%)	16.8% (10.0%)	2.44% (100%)

this, we picked the strain that maximized the additional genomic coverage of shared segments with the other strains given the coverage by the resequenced reference strains. This procedure is repeated greedily to select the next target of reference strain given the previous set of reference strains. To increase the coverage including the wild-derived strains, many wild-derived strains are prioritized for resequencing. When considering only classical inbred strains, the strains with relatively higher imputation errors tend to be prioritized (Table 2.8).

Next, we estimated the effectiveness of imputation when different numbers of mouse HapMap SNPs are collected. To do this, we selected a range of sparse subsets (10,000 to 1,000,000 markers) of the NIEHS/Perlegen SNPs with complete data in the resequenced strains and estimated the imputation errors for each of 12 resequenced classical inbred strains using leave-one-out cross-validation. As expected, accuracy increased proportionally to subset size. Selecting a 100,000 SNP subset gave an overall imputation error of 1.36% (high-confidence genotype error 0.36% with 93.8% call rate). This is comparable to the imputation accuracy using the current mouse

Table 2.8: Top 10 strains greedily targeted for resequencing in addition to 16 resequenced strains in order to improve the genomic coverage over mouse HapMap strains.

All 94 strains		73 classical inbred strains	
Strain name	Coverage increase (avg)	Strain name	Coverage increase (avg.)
PERC/EiJ	0.012039	NZL/LtJ	0.004697
ZALENDE/EiJ	0.006353	P/J	0.003432
SPRET/EiJ	0.004942	DDK/Pas	0.002914
IS/CamRkJ	0.002984	LG/J	0.002070
NZL/LtJ	0.002416	C57/L	0.001928
LEWES/EiJ	0.001931	SJL/J	0.001574
P/J	0.001624	O20	0.001359
DDK/Pas	0.001354	RIIIS/J	0.001184
C57/L	0.001013	I/LnJ	0.000985
LG/J	0.000919	SM/J	0.000789

HapMap SNPs. We note that the current size of the HapMap SNP is well powered to capture the majority of variation at low error rates and high confidence. (Figure 2.4). A several-fold increase in SNP map density to 1,000,000 markers further optimizes these rates, and as current genotyping platforms can accommodate this number of assays this would be a viable design for the next generation mouse HapMap.

2.2.5 Trait mapping with the mouse HapMap resource

This detailed picture of haplotype diversity in the mouse allows us to map traits in the inbred strains by correlating genomic ancestry to trait measurements, rather than generating de novo experimental crosses. This *in silico* association mapping has two advantages: it allows us to capture the full spectrum of diversity in the inbred strains rather than a subset used as progenitors of an experimental cross; and phenotypic noise can be minimized by performing replicates on genetically identical individuals. In particular, this approach should complement traditional QTL linkage mapping (often successful at locating large chromosomal segments) by providing a higher resolution, association-based component and indeed has already yielded several positive results[166, 124, 32].

The high degree of relatedness between strains described above introduces a

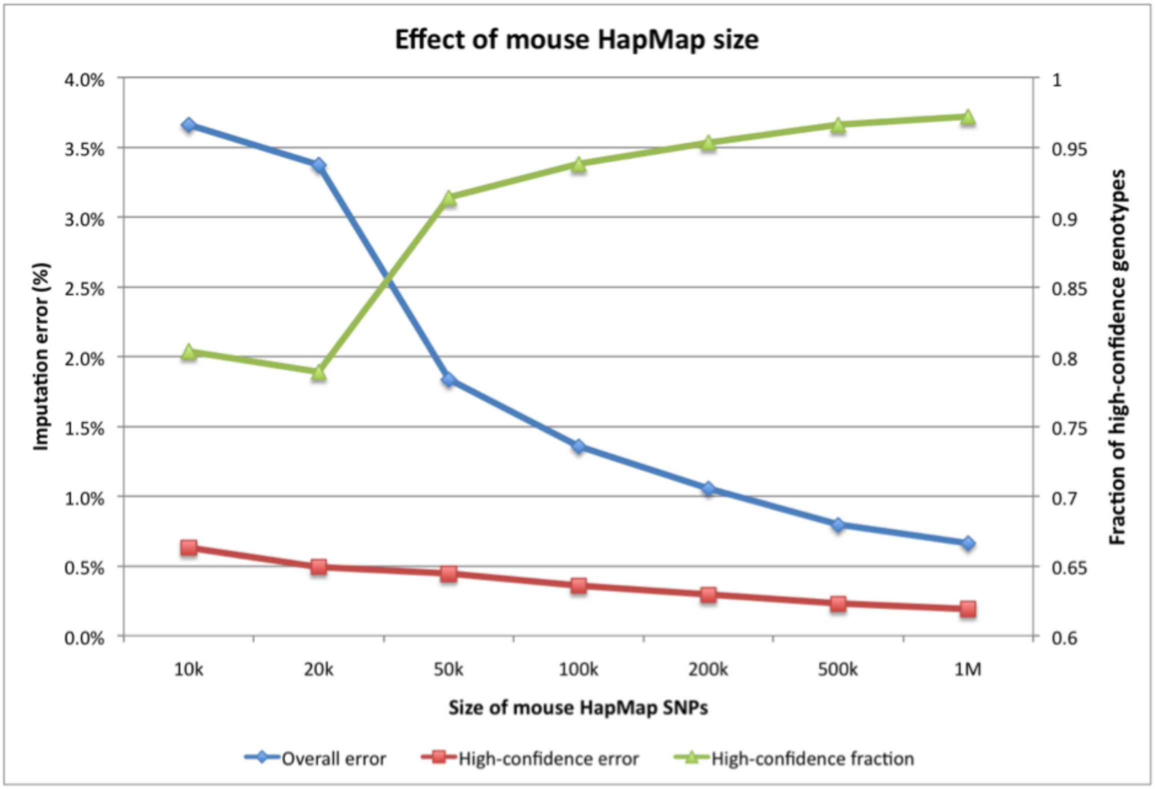


Figure 2.4: Estimated imputation accuracy and coverage over fully-resequenced NIEHS/Perlegen SNPs across 12 classical inbred strains with various sizes of randomly selected SNP sets

systematic bias in association mapping *in silico*: an inflation of test statistics leading to false positive associations, caused by population structure and genetic relatedness among the strains[103, 235, 240, 8]. For example, among the 180 phenotypes deposited in the Mouse Phenome Database at the Jackson Laboratory (MPD) with more than 30 distinct strains, 59% (106) of them have more than 50% of the inter-strain phenotypic variance explained by population structure and genetic relatedness measured using a variance component test. (Figure 2.5). At an FDR level of 0.05, 51% (91) of them are significantly associated with population structure. We and others have shown that these issues can effectively be corrected using linear mixed models[103, 235, 240]. We have therefore developed a corrected association database in conjunction with the MPD, in which we find 71/180 phenotypes collected in more than 30 strains have at least one significant association ($p < 1 \times 10^{-6}$). Among them, 11 (6.1%) phenotypes showed significant associations across more than 20 different genomic regions, which may indicate residual bias from other sources generating false positives. This may be compared to 24 (13%) phenotypes showing association without population structure correction to more than 20 different genomic regions, while the total number of phenotypes with significant associations is similar (Figure 2.6). When comparing the "inflation factor" suggested by Genomic Control between different statistical tests, t-test showed much higher overall inflation ($\lambda = 2.08 \pm 1.29$) compared to the linear mixed model ($\lambda = 1.15 \pm 0.18$) over the 180 MPD phenotypes, confirming the overly inflated false positive rates with the conventional t-test. (Figure 2.7)

2.3 Discussion

We have described the high-density genotype resource for 94 inbred mouse strains. Our genotype data is available at `http://www.mousehapmap.org`. In addition, we have established a website `http://mouse.cs.ucla.edu/` at which researchers can download genotype data, and access a genome browser which allows the visualization of the haplotype and shared segment analyses. The website also supports inbred association mapping and includes association results using the genotypes and all collected phenotype data in the Mouse Phenome Database.

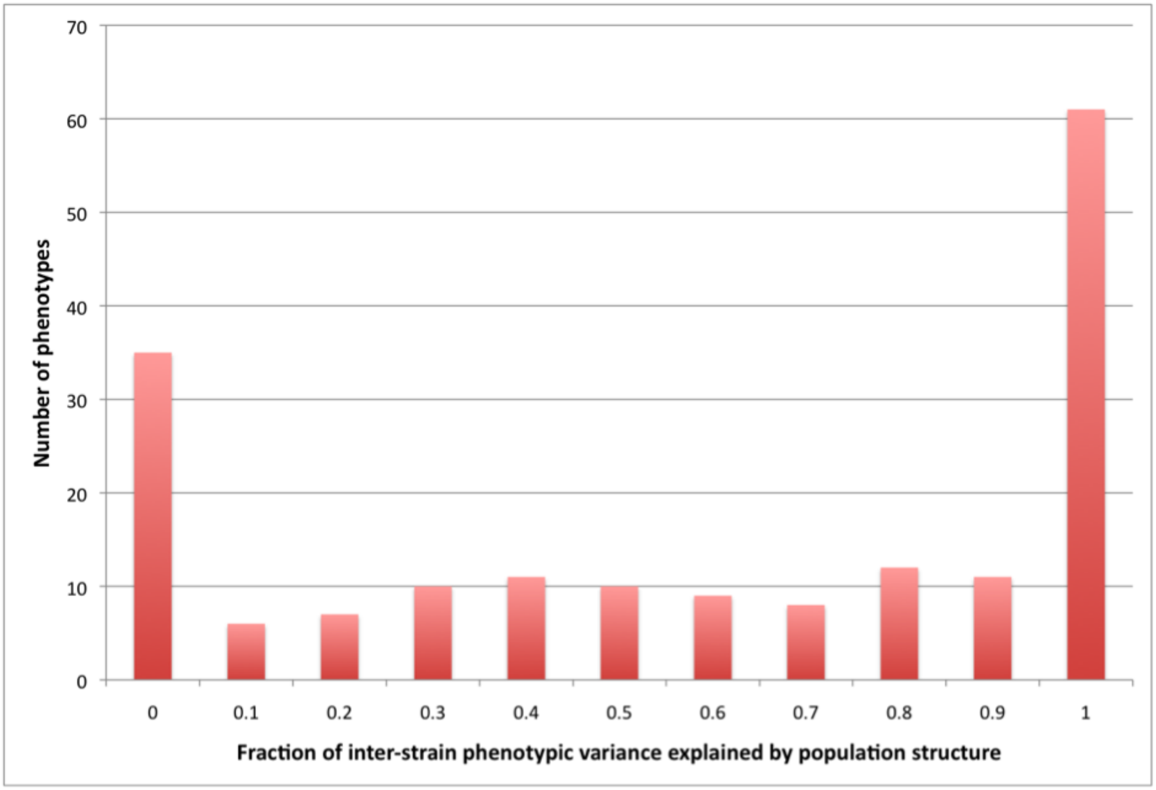


Figure 2.5: Distribution of fraction of phenotypic variation explained by population structure among the strains over 180 quantitative phenotypes deposited in the Mouse Phenome Database (MPD) with 30 or more strains

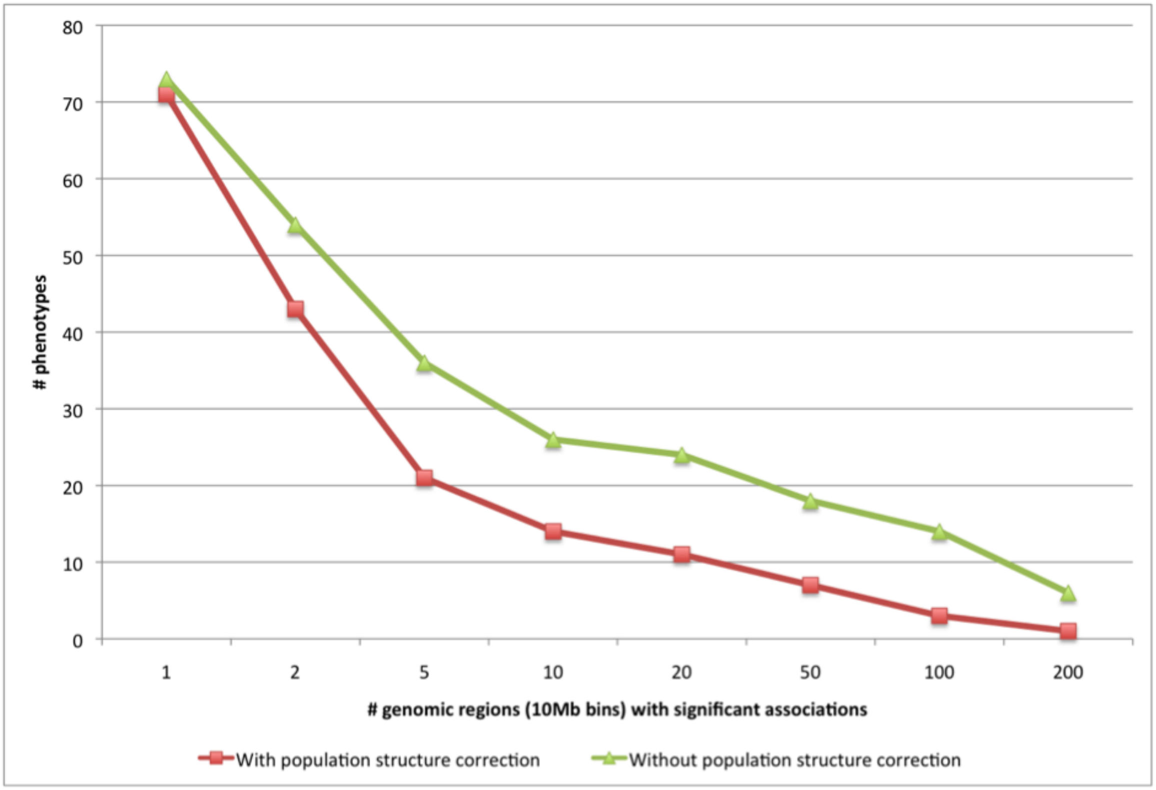


Figure 2.6: Number of phenotypes with multiple genomic regions with significant associations illustrating the degree of inflated false positives, over 180 quantitative phenotypes deposited in the Mouse Phenome Database (MPD) with 30 or more strains

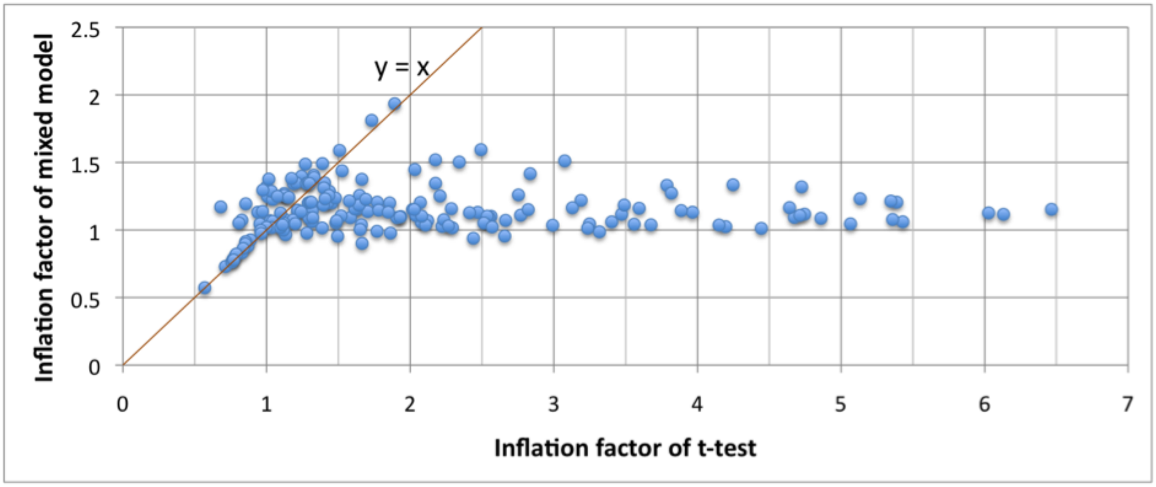


Figure 2.7: Comparison of genomic control "inflation factors" between t-test and linear mixed model across 180 MPD phenotypes

2.4 Methods

Array Design The mouse HapMap chips consist of two Affymetrix genotyping arrays with 20 or 36 PM/MM probe-pairs. SNPs were selected to as evenly spaced as possible across the NCBI build 33, and mapped to NCBI build 37. Genotypes were called with Affymetrix DM algorithm, and the genotypes with low confidence genotypes or with conflicting calls between replicated samples or any discovery strain were called as missing.

Analysis of shared segments The mapping with four founder strains was performed with a hidden Markov model with four reference strains with additional state for unknown reference, learning the parameters from the genotype data using EM algorithm as described in the imputation method. A hidden Markov model with two states representing common and divergent regions was constructed for each pairwise comparison, with recombination parameter $\theta = 10^{-8}$ and mutational parameter $\mu = 0.03$, estimated from the distribution of maximum likelihood parameters using EM algorithm among all 4371 comparisons. The fraction of genome with shared segments was computed as the fraction of genome wide SNPs with the probability of shared segments greater than 0.9. The number of distinct ancestral segments at a genomic position was computed by taking all the pairwise probabilities of shared segments, and by performing hierarchical clustering with a median agglomeration method by taking the pairwise probabilities as elements of a similarity matrix.

Imputation of missing genotypes A hidden Markov model was constructed, for each strain targeted for imputation, with 16+1 states per SNP representing each of 16 resequenced reference strains and a state representing equivocal reference strain, similar to previously suggested method. Unlike the previous methods[132, 186], the maximum-likelihood parameters of genome wide mutation and recombination parameters were learned from the data using EM algorithm and forward-backward algorithm, independently for each strain. For leave-one-out imputation for experimentally missing genotypes in the resequenced strains, 15+1 states were used excluding the target strain for imputation.

In-silico association mapping We downloaded the individual phenotype measurements of Mouse Phenome Database (MPD) from Jackson Laboratory, and selected 696 quantitative phenotypes containing phenotype measurements in at least 10 strains, and having at least 10 distinct values of phenotypes with the maximum occurrence of 20 per each value to filter out the categorical phenotypes and survival data. We applied EMMA (Efficient Mixed Model Association)[103] as an implementation of linear mixed models to correct for population structure and genetic relatedness, using the kinship matrix generated as a genotype similarity matrix. The variance component was based on REML (Restricted Maximum Likelihood) estimate, and a standard F test was performed as previously suggested[235, 240]. The FDR significance level was estimated using the q-value R package[196]. The males and females were mapped for association separately. The Genomic Control inflation factor was computed by taking the median p-value and computing the corresponding chi-square statistic divided by 0.455[50]

Chapter 2 is currently in submission for publication for the material. Andrew Kirby, Hyun Min Kang, Claire M. Wade, Chris J. Cotsapas, Emrah Kostem, Buhm Han, Manuel Rivas, Molly A. Bogue, Kelly A. Frazer, Frank M. Johnson, Erica J. Beilharz, David R. Cox, Eleazar Eskin, and Mark J. Daly, “A high-density haplotype resource of 94 inbred mouse strains”. The dissertation author and Andrew Kirby are the primary investigators and authors of this paper.

Chapter 3

An adaptive and memory efficient algorithm for genotype imputation

3.1 Motivation

Recent advances in high-throughput genotyping technologies are helping to uncover the genetic basis of complex phenotypes in human[223], mouse[67], rat[193], dog[104], arabidopsis[23], and many other model organisms. While the vast majority of positions in a genome are identical among individuals in a population, a significant portion of positions differ. Many of these positions are single nucleotide polymorphisms (SNPs). In a typical association study that attempts to identify variation involved in a trait, variation information (e.g. SNP genotypes), is collected from a set of individuals and the trait is measured in each individual. Each SNP is then correlated (or associated) with the trait. Any statistically significant associations are reported as possible causal variation with respect to the trait [179, 45].

Genotyping arrays, such as those developed by Affymetrix and Illumina simultaneously probe hundreds of thousands of marker SNPs in an individual's genome [135, 78]. While this is a significant amount of information, it is only a fraction of the millions of SNPs and other genetic variation in the population. Only complete resequencing of individual genomes will guarantee collecting all variation in a study. However, resequencing still remains prohibitively expensive. Array based genotyping is currently the most practical cost-effective method for collecting large amounts of variation information on a set of individuals. Although only a subset of individual genetic variation is collected by a genotyping array, due to the correlation structure of variation in the genome, SNPs on the array can serve as proxies for SNPs which are not collected [49, 41]. This property is called linkage disequilibrium (LD), and it greatly extends the coverage of the array since a causal SNP need not be collected, but only strongly correlated with one of the collected markers on the array[237]. However, if the causal variants are not in LD with one of the SNPs included on the array then the study will not be able to discover the association. Thus, increasing the number of collected SNPs in the study increases the study's power to identify casual variation and is of fundamental importance.

Recently, several studies have proposed methods to increase the ability of a study to identify associations at SNPs which are not collected by "imputing" or predicting the genotypes of SNPs that are not contained in the study data set. These methods work by using a reference sample, such as the HapMap [67] for humans, which has genotyped millions of SNPs at great cost and effort. These reference samples contain both SNPs which are collected in the study as well as other SNPs. An imputation method uses the correlation patterns between the collected and uncollected SNPs inferred from the reference sample to make predictions of the uncollected SNPs in the study sample. This problem is effectively a missing data problem in which partial data is observed in the study and complete data is observed in the reference sample.

Consider the example shown in Figure 3.1, there is a set of reference individuals shown on top and a study individual shown on the bottom. In the reference set all the SNPs are genotyped in all five individuals. In the study individual, some of the SNPs are uncollected and denoted by a "?". The goal of imputation is to resolve the genotypes of the uncollected SNPs by using the overlap of the typed SNPs between the reference set and the study set. Our method selects the most likely reference individual for each marker (both collected and uncollected). The path in bold shown in Figure 3.1 denotes that the sequence of SNP values in target individual is composed of pieces of the three reference individuals along the six collected and four uncollected markers.

Multiple techniques have been successfully employed to solve the imputation

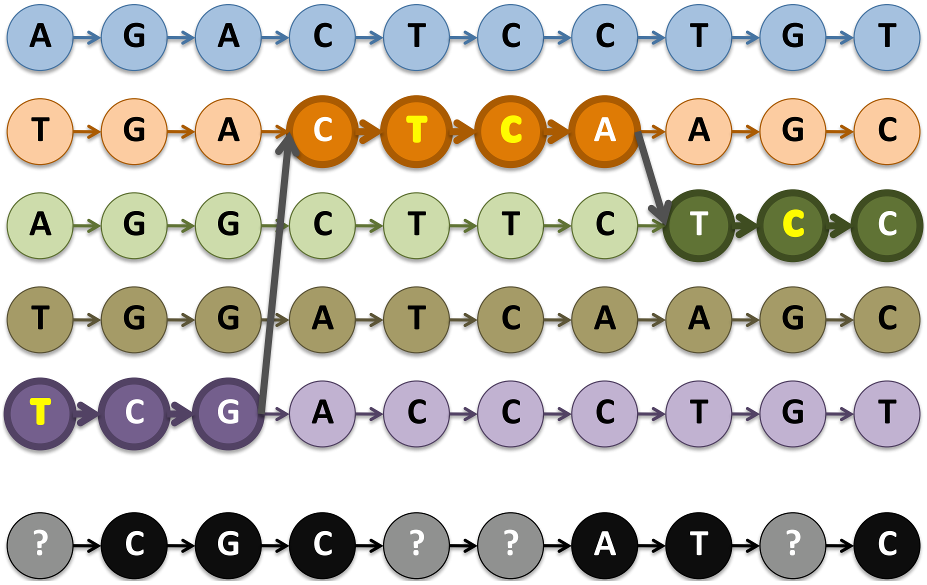


Figure 3.1: An example of the imputation problem. The five reference individuals are genotyped on all ten SNPs, while the study individual is genotyped on only six SNPs. The goal of imputation is to resolve the genotypes of the uncollected SNPs.

problem, and Hidden Markov models (HMMs) have been amongst the most popular, and have been used in several studies in both human and mouse. However, each of these existing HMM techniques fails to address at least one of several important problems. IMPUTE[132] applies a standard population genetics model of recombination and mutation using a set of reference haplotypes. The transitional and mutational parameters of the HMM are predefined for each reference population, so the method is not adaptive to different populations or different organisms. For example, using IMPUTE on mouse genotypes will result in largely inaccurate estimates of uncollected SNPs. Moreover, it has memory requirements that grow linearly in the size of the reference data set, which may become prohibitively large as more SNPs are discovered. On the other hand, MACH[88, 151] allows us to learn parameters of an HMM from the observed genotypes. However, in order to learn parameters easily, it uses a much simpler transitional model which does not utilize the continuous-time Markov chain model of recombination in population genetics[108]. Moreover, since the transitional parameters are estimated per each marker interval separately, MACH requires a large number of samples to be simultaneously imputed for an accurate parameter estimation. For inbred mouse imputation, the problem differs from human because the reference data sets have dramatically different linkage properties from human and do not have heterozygous genotypes. Recently, a fastPHASE[186] like method recently employed HMMs to solve the imputation problem in mouse heuristically using a predefined sets of clusters with Dirichlet prior distribution and Viterbi training method[200]. Although this recent method did impute many of the SNPs in the mouse, the average imputation error was 10.4% and 4.4% for high-confidence genotypes, which is higher than the error rates in most of the human imputation studies.

In this chapter, we propose EMINIM (Expectation-Maximized INtegrative IMputation), an adaptive genotype imputation method that learns HMM parameters with the standard population genetics model of recombination and mutation using Expectation-Maximization (EM) algorithm. Our method is motivated by the fact that the previous methods may not be applied to model organisms such as inbred mouse strains due to the predefined parameters being inappropriate or the small number of strains in the reference sample (in this case only 16). Our method utilizes various types of silent states in the HMM to estimate the EM parameters, to increase memory-efficiency, to impute genotypes at collected SNPs, and to obtain a leave-one-snp-out estimate of imputation accuracies. Our method is also more memory efficient allowing much larger data sets to be imputed. In addition, we provide an extensive implementation detail of our method that improves accuracy and computational efficiency in addition to the core statistical model, in order to facilitate further progress in the area of genotype imputation.

We applied our method to the imputation of 8.27 million SNPs that have been discovered from a resequencing of 15 inbred mouse strains, based on the 138,980 SNPs collected from the mouse HapMap project over 94 inbred strains. Imputation in mouse strains differs from human imputation because the reference datasets have drastically different linkage properties. Using a leave one out procedure, we measured the error rate of our method and compared to the recently published mouse imputation paper [200]. Our method's overall error rate is less than half the error rate of the previous method, and for high confidence genotypes our error rate is ten time smaller.

Next, we applied our method to the imputation of human HapMap SNPs from the Wellcome Trust Case-Control Consortium (WTCCC) genotypes. Our results show that our method consistently achieves similar or better imputation accuracy over different populations than other state-of-the-art methods without requiring predefined parameters for each population. Our method also shows a significant increase of memory-efficiency which can be an important technical issue when imputing genotypes of a dense SNP sets over a large number of reference samples. For example, on a recent study EMINIM used only 508 MB of memory to impute chromosome 22 while IMPUTE required 6.6 GB. This problem will become even more severe on larger chromosomes and data sets. Our method is publicly available at `http://genetics.cs.ucla.edu/eminim`.

3.2 Materials and methods

3.2.1 The imputation problem.

In this section we formalize the problem of imputing missing genotype data in an individual using a reference population. The terms and definitions described here will be used throughout the text. We classify the imputation problem into two categories - haploid (or inbred) imputation and diploid imputation. Suppose that we genotype m SNP *markers* on an individual (*target individual*) and wish to determine the genotypes of additional *"uncollected" SNPs*. We will employ a set of *reference haplotype* that are genotyped on the m collected markers as well as an additional set of uncollected SNPs. In diploid model, we assume that each reference individuals is already phased into two reference haplotypes. The allele of the i-th reference haplotype at the t-th marker is represented as $G_{i,t} \in \{0, 1, 2\}$, where 0 represents a missing reference genotype and 1,2 represents two alleles of biallelic SNP marker. Let n be the number of reference haplotype collected at m markers in the target individual. Let $\mathbf{d} = \{d_1, d_2, \cdots, d_{m-1}\}$ be the physical or genetic distance between consecutive markers. In the target individual, the genotype at the t-th marker is represented as $g_t \in \{1, 2\}$ in haploid model, and $g_t \in \{\{1, 1\}, \{1, 2\}, \{2, 2\}\}$ in diploid model, where 1 and 2 represents two alleles of the biallelic SNP marker.

In our model, we assume that in each region, the target individual has similar haplotypes to the reference haplotypes. The goal of imputation can then be rephrased as assigning one (haploid model) or a pair of (diploid model) the n reference haplotypes to each of the m markers. From these assignments, we will assign the individual SNP genotype values to each uncollected SNP in the target individual by using the alleles of the assigned reference haplotypes at the nearby markers. We employ a hidden Markov model (HMM) to assign the reference haplotypes to the markers of the target individual. We describe the details of the HMMs used to solve the different cases of the imputation problem.

3.2.2 Imputation algorithm for haploid model

Hidden Markov Model for imputation In the following section we describe the algorithm for performing imputation for haploid or inbred organisms. In this case, the reference haplotypes (or individuals) are called reference strains and we assume that we do not observe any genotypes where an individual has both alleles 1 and 2 of the SNP (i.e. no heterozygous genotypes).

The goal is then to assign one of the n reference strains to each of the m markers described in the previous section. To accomplish this we use an HMM like that shown in Figure 3.2. For each of the m markers there are n states corresponding to each of the reference strains. From each state there are n edges (with the exception of the states representing the final marker) directed towards the n states for the next marker. The edges corresponding to a change in the reference strain are called transitions or a recombination. Each state can also emit one of the two possible alleles for that marker. Emitting an allele that does not match the target strain is called a mutation.

Let $S_t \in \{1, 2, \cdots, n\}$ be the reference strain assigned to the target strain at marker $t \in \{0, 1, \cdots, m-1\}$. Since some of genomic segments may not be represented by any of the reference strains, we introduce another reference strain consisting of only missing genotypes to represent the unknown reference state. We let the initial probability that marker S_0 is assigned to strain i be $\Pr(S_0 = i) = \pi_i$, with $\pi = \{\pi_1, \cdots, \pi_n\}$. Similar to many other methods designed for genotype imputation and haplotype phasing[132, 186], our HMM relies on a typical Markov chain model

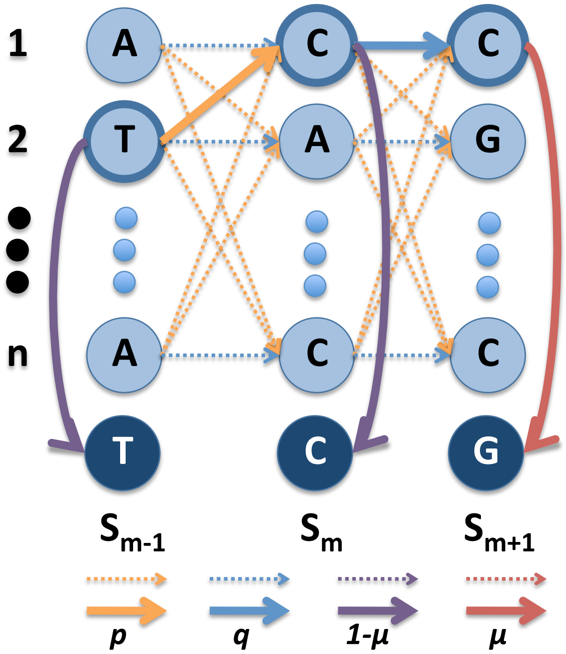


Figure 3.2: An example of the hidden Markov model for the imputation problem

of population genetics based on neutral Wright-Fisher model[108]. We use two parameters μ, and θ to be the mutation and recombination parameters, and use the standard distributions for computing the probability of transitioning between states $p_n(x) = (1 - e^{-x})/n$ and $q_n(x) = p_n(x) + e^{-x}$ based on the continuous-time Markov process. Figure 3.2 shows how the these probabilities correspond to the edges in the HMM. The transition probabilities are computed from the recombination parameter and the distance between markers as follows.

$$\Pr(S_t = j | S_{t-1} = i, \theta) = \begin{cases} q_n(-\theta d_t) & i = j \\ p_n(-\theta d_t) & i \neq j \end{cases} \tag{3.1}$$

The probability of an observed genotype given a state is computed from the mutation parameter and the allele observed at the reference strain at the state. If the reference strain has missing genotype at the marker, then the probabilities are equally assigned between the alleles. Figure 3.2 shows examples of matching and mutated genotypes in the emission states of the HMM.

$$\Pr(g_t|S_t,\mu) = \begin{cases} 1-\mu & g_t = G_{S_t,t} > 0 \\ \mu & g_t \neq G_{S_t,t} > 0 \\ 0.5 & G_{S_t,t} = 0 \end{cases} \tag{3.2}$$

We assume a uniform distribution of initial state probabilities, $\pi_1 = \cdots = \pi_n = 1/n$, and learn the mutation and transition parameters from the data using the EM algorithm presented below.

EM algorithm for learning maximum-likelihood parameters Many of the previous methods suggested for HMM-based imputation of missing genotypes use either predefined transitional parameters[132] or Viterbi training which may not converge to a local maximum[200]. Other methods estimate the transition probabilities per marker interval independently to avoid the computational complexity in constraining the transition parameters consistently across different states[186, 151].

In the genotype imputation of inbred mouse strains, independent unconstrained estimation of parameters at each marker is prone to inherent bias and inaccuracy because of two reasons. First, the total number of target strains is small, so estimating parameters per marker independently may be highly inaccurate. Second, these strains have complex genetic relationship, so the transitional and mutational parameters vary greatly across different strains. We constrain the parameters to be equal over the genome, but allows different transition and mutation parameters for each strain. Instead of the simple Viterbi training algorithm, we present an EM algorithm based on the exact conditional probabilities obtained from the forward-backward algorithm.

Let us denote $X_t^- = (X_1, \cdots, X_t)$ and $X_t^+ = (X_{t+1}, \cdots, X_m)$ as observed data, and $\lambda = (\pi, \mu, \theta)$ be the initial, mutational, and transitional parameters of the hidden Markov model. The forward-backward algorithm estimates $\alpha_t(i) = \Pr(X_t^-, S_t = i|\lambda)$ and $\beta_t(i) = \Pr(X_t^+|S_t = i, \lambda)$ using dynamic programming.

Let $X = (g, G)$. The EM algorithm starts with initial parameters (μ_0, θ_0). At the E-step of r-th iteration, $\Pr(S_t|X, \lambda_r)$ are computed from the forward-backward algorithm. Let $S = \{S_0, \cdots, S_{m-1}\}$. At the M-step, the expected likelihood function can be written as follows.

$$\begin{aligned}
Q(\mu,\theta) &= \sum_{S} \Pr(S|X,\lambda_r)\log\Pr(g,S|\lambda) & (3.3)\\
&= \sum_{t=1}^{m-1}\sum_{(S_t,S_{t-1})} \log\Pr(S_t|S_{t-1},\theta)\Pr(S_t,S_{t-1}|X,\lambda_r)\\
&\quad+\sum_{t=0}^{m-1}\sum_{S_t}\log\Pr(g_t|S_t,\mu)\Pr(S_t|X,\lambda_r)-\log n
\end{aligned}$$

The expectation-maximized parameters used for the next round of E-step can be obtained as follows.

$$\begin{aligned}
\mu_{r+1} &= \frac{\sum_{t=1}^{m}\sum_{S_t} I(g_t\neq G_{i,t})\Pr(S_t|X,\lambda_r)}{\sum_{t=1}^{m}\sum_{S_t} I(G_{i,t}>0)\Pr(S_t|X,\lambda_r)} & (3.4)\\
\theta_{r+1} &= \arg\max_{\theta}\left[\sum_{t=1}^{m-1}\sum_{(S_t,S_{t-1})}\log\Pr(S_t|S_{t-1},\theta)\Pr(S_t,S_{t-1}|X,\lambda_r)\right] & (3.5)\\
&= \arg\max F(\theta) & (3.6)
\end{aligned}$$

In order to estimate the joint probability $\Pr(S_t,S_{t-1}|X,\lambda_r)$, we introduce a silent state J_t between S_{t-1} and S_t with the following transition probabilities which keeps $\Pr(S_t|S_{t-1})$ unchanged.

$$\begin{aligned}
\Pr(J_t=(i,b)|S_{t-1}=i,\theta) &= \begin{cases} q_n(-\theta d_t) & b=0\\ (n-1)p_n(-\theta d_t) & b=1\end{cases} & (3.7)\\
\Pr(S_t=j|J_t=(i,b),\theta) &= \begin{cases} 1 & b=0,\ \ i=j\\ 1/(n-1) & b=1,\ \ i\neq j\end{cases} & (3.8)
\end{aligned}$$

The probabilities of the other transitions are set to zero. In general, the marginal probabilities of these silent states can be computed, in the following way. Let L be a silent state connecting S_{t-1} and S_t preserving the transition probability of $\Pr(S_t|S_{t-1},\lambda)=\sum_L \Pr(S_t|L,\lambda)\Pr(L|S_{t-1},\lambda)$ unchanged. The forward and backward probability of any silent state L are defined as $\alpha(L)=\sum_{S_{t-1}}\alpha_{t-1}(S_{t-1})\Pr(L|S_{t-1})$ and $\beta(L)=\sum_{S_t}\beta_t(j)\Pr(S_t|L)\Pr(X_t|S_t)$. Then the objective function of M-step transitional parameter becomes

$$F(\theta) = \sum_{t=1}^{m-1}\left[\log q_n(-\theta d_t)\sum_{i=1}^{n}\Pr(J_t = (i,0)|X,\lambda_r) + \log p_n(-\theta d_t)\sum_{i=1}^{n}\Pr(J_t = (i,1)|X,\lambda_r)\right] \tag{3.9}$$

This function can be numerically optimized using a Newton-Raphson algorithm.

Imputation of uncollected genotypes Let h_t be the number of 'uncollected SNPs' that need to be imputed between marker $(t-1)$ and t. An uncollected SNP is represented as (t,s), where $t \in \{1,2,\cdots,m-1\}$ and $s \in \{1,\cdots,h_t\}$. Let $T_{t,s} \in \{1,\cdots,n\}$ be the state at an uncollected SNP (t,s).

We again modify the HMM by adding a silent state $T_{t,s}$ to the original HMM. The transition probabilities between $S_{t-1}, T_{(t,s)}$, and S_t is defined in the same way to to Equation 3.1 based on the distance between them. Let $H_{i,t,s} \in \{0,1,2\}$ be the genotypes of i-th reference strain at the uncollected SNP (t,s). Then distribution of the imputed genotype $z_{t,s}$ at the uncollected SNP is estimated as follows.

$$\Pr(z_{t,s}|X,\lambda) = \begin{cases} \begin{array}{l}(1-\mu)\sum_{i=1}^{n} I(H_{i,t,s} = z_{t,s})p_{t,s}(i) \\ \quad +\mu\sum_{i=1}^{n} I(H_{i,t,s} \neq z_{t,s})p_{t,s}(i)\end{array} & z_{t,s} > 0 \\ \sum_{i=1}^{n} I(H_{i,t,s} = 0)p_{t,s}(i) & z_{t,s} = 0 \end{cases} \tag{3.10}$$

where $p_{t,s}(i)$ denotes the marginal probability of state i at the uncollected SNP (t,s). When estimating leave-one-snp-out imputation accuracy, the same imputation methods are applied by pretending marker t has a silent state by ignoring the observed genotype.

Improving computation time complexity A standard HMM implementation requires squared time complexity with respect to the number of individuals. It is possible to reduce the time complexity to be linear in the number of states, by leveraging the fact that the transition probabilities are uniform over different states. When $t > 1$, $\alpha_t(i)$ follows that

$$\alpha_t(S_t) \quad = \quad [\exp(-\theta d_t)\alpha_{t-1}(S_t) + p_n(-\theta d_t)\sum_{x=1}^{n}\alpha_{t-1}(x)]\Pr(X_t|S_t) \qquad (3.11)$$

This can be computed in a constant time if $\sum_{j=1}^{n}\alpha_{t-1}(x)$ are precomputed, so the computation of $\alpha_t(S_t)$ over all states can be performed in linear time. In a similar way, the computation of $\beta_t(S_t)$ and the computation over silent states are linear to the number of states.

3.2.3 Extension to unphased genotypes (diploid model)

When imputing human genotype data, the reference individuals are typically provided as phased haplotypes across a dense set of SNPs, and a number of unphased genotypes of a target individual are provided as a subset of SNPs. In this case, the state at each collected marker $Z_t = (i,j)$ represents the combined states of each haplotype. Here we assume there are no missing alleles in the reference haplotypes because they are phased. However, missing alleles can also be handled in a similar way presented as in the haploid model. Their initial state probabilities are defined as $\Pr(Z_0 = (i,j)) = \pi_{ij} = 1/n^2$, and the transition probabilities are defined as follows.

$$\Pr(Z_t = (i,j)|Z_{t-1} = (k,l),\theta) = \begin{cases} q_n(-\theta d_t)^2 & i = k \ , \ j = l \\ p_n(-\theta d_t)q_n(-\theta d_t) & i = k \ \oplus \ j = l \\ p_n(-\theta d_t)^2 & i \neq k \ , \ j \neq l \end{cases} \qquad (3.12)$$

where $\oplus$ denotes exclusive OR operator. Let $Z_t = (Z_{t,0}, Z_{t,1})$ be the individual states of each chromosome, then the imputed genotype $(g_{t,0}|Z_{t,0}, G,g_{t,1}|Z_{t,1}, G)$ independently follows the mutational distribution in the inbred imputation.

The observed genotype $g_t = \{g_{t,0}, g_{t,1}\} \in \{\{1,1\},\{1,2\},\{2,2\}\}$ represent one of homozygous base alleles, heterozygous alleles, or homozygous mutant alleles. Based on these probability models, an HMM can be constructed with n^2 states for each collected marker.

Let b_t be the number of state changes between marker $t-1$ and t. In order to estimate EM parameters, we introduce a silent state J_t connecting Z_{t-1} and Z_t, with the following transition probabilities.

$$\Pr(J_t = (k,l,b_t)|Z_{t-1} = (i,j)) = \begin{cases} q_n(-\theta d_t)^2 & k = i,\ l = j,\ b_t = 0 \\ 2(n-1)p_n(-\theta d_t)q_n(-\theta d_t) & k = i,\ l = j,\ b_t = 1 \\ (n-1)^2 p_n(-\theta d_t)^2 & k = i,\ l = j,\ b_t = 2 \end{cases} \tag{3.13}$$

$$\Pr(Z_t = (i,j)|J_t = (k,l,b)) = \begin{cases} 1 & k = i,\ l = j,\ b_t = 0 \\ 1/(2n-2) & (k = i \oplus l = j),\ b_t = 1 \\ 1/(n-1)^2 & k \neq i,\ l \neq j,\ b_t = 2 \end{cases} \tag{3.14}$$

From the expectation maximized parameters in the M-step it follows that

$$\begin{aligned} \mu_{r+1} &= \frac{1}{m}\sum_{t=0}^{m-1}\sum_{Z_t}\eta(g_t, G_{Z_{t,0},t}, G_{Z_{t,1},t})\Pr(Z_t|X,\lambda_r) && (3.15) \\ \theta_{r+1} &= \arg\max_\theta\left[\sum_{t=1}^{m-1}\sum_{(Z_t,Z_{t-1})}\log\Pr(Z_t|Z_{t-1},\theta)\Pr(Z_t,Z_{t-1}|X,\lambda_r)\right] && (3.16) \\ &= \arg\max_\theta G(\theta) && (3.17) \end{aligned}$$

where $\eta(g,h_1,h_2)$ is the number of mismatched alleles between genotype g and $\{h_1,h_2\}$. $G(\theta)$ can be rewritten to be numerically optimized using a Newton-Raphson algorithm as follows.

$$\begin{aligned} \sum_{t=1}^{m}[2\log q_n(-\theta d_t)\Pr(b_t = 0|X,\lambda_r) + \log\left(p_n(-\theta d_t)q_n(-\theta d_t)\right)\Pr(b_t = 1|X,\lambda_r) \\ +2\log(p_n(-\theta d_t))\Pr(b_t = 2|X,\lambda_r)] \end{aligned} \tag{3.18}$$

Similar to the inbred case, each uncollected genotypes is imputed without increasing memory by adding a silent state. For example, $\alpha_t(Z_t)$ can be computed in a linear time with the number of states if we precompute $\sum_{x=1}^{n}\alpha_t(Z_{t,b},x)$ for each $Z_{t,b}$ and $\sum_{x=1}^{n}\sum_{y=1}^{n}\alpha_t(x,y)$ in order to increase the computational complexity.

3.3 Results

3.3.1 Genotype imputation of 94 inbred mouse strains

A recent NIEHS/Perlegen mouse resequencing project identified 8.27 million SNPs among 16 inbred mouse strains[67]. The Broad mouse HapMap project collected genotypes over 94 strains at 138,980 SNPs, which is only 1.7% of the number of SNPs identified in the resequencing project. We can achieve high imputation accuracy even with such a small fraction of the SNPs because of the very long regions of linkage disequilibrium.

We evaluated the accuracy of our genotype imputation method through leave-one-out analysis. For each of 16 resequenced strains, we ran our EMINIM algorithm to impute the genotypes at NIEHS/Perlegen SNPs using the mouse HapMap genotypes and the NIEHS/Perlegen SNPs of the rest 15 strains. Singleton SNPs polymorphic only in the target strain were removed in the evaluation of accuracy since they are not able to be imputed using the rest of strains. The leave-one-strain-out validation provides a conservative estimate of the genome wide imputation accuracy of a unresequenced strain using 16 resequenced strains.

The overall average imputation error over 12 classical strains is 2.40%. We classified the imputed genotypes into the 'high-confidence' category if the posterior probability is greater than 0.98, and 'medium-confidence' in between 0.8 and 0.98. When considering only high-confidence imputed genotypes after discarding 18.9% of low and medium confidence genotypes, the average imputation error significantly reduces to 0.37%. When including wild-derived strains, the imputation error significantly increases. The average imputation error between four wild-derived strains was 19.2%, each of them ranging from 13.0% (WSB/EiJ) to 34.0% (CAST/EiJ). None of the wild-derived strains have high-confidence imputed genotypes due to high estimates of mutation rates.

Unlike previous imputation methods based on hidden Markov models, we introduce an additional state to account for genomic regions that are not explained by any of the reference strain. We compared this model to one without an additional state, by computing imputation accuracy using leave-one-strain-out cross-validation. The results over 12 classical inbred strains show that the overall imputation error

Table 3.1: Imputation error rates of inbred mouse strains. First two rows (LOOCV) use leave-one-strain-out estimation using 15 reference strains and 138,980 combined SNPs of the target strain across 12 classical inbred strains. The unknown reference strain is used in the first but not in the second. Last two rows uses WTCHG genotypes as validation set, and impute those genotypes in 47 WTCHG strains not included in the reference strains. The fraction of imputed genotypes in each category is shown within a parenthesis

.

Confidence cutoff	high (> 0.98)	high + medium (> 0.8)	all (≥ 0)
LOOCV with unknown reference	0.37% (81%)	0.81% (90%)	2.40% (100%)
LOOCV without unknown reference	0.52% (76%)	1.24% (93%)	2.46% (100%)
36 non-wild WTCHG strains	0.35% (89%)	1.00% (96%)	2.25% (100%)
all 47 WTCHG strains	0.37% (72%)	1.98% (89%)	4.86% (100%)

increased from 2.40% to 2.46%. More notably, the average imputation errors in high confidence category increased from 0.37% to 0.52%, and the coverage of high-confidence category reduced from 81.1% to 75.7%. This suggests that our model with additional state for unknown reference strains significantly affects the imputation accuracy probably because some genomic segments are not well characterized by any of the 16 reference strains.

Next, we evaluated the imputation accuracy by comparing the genotypes typed for 78 non-resequenced strains. We used the Wellcome Trust genotypes as a validation set and evaluated how accurately our method can impute the genotypes in the validation set using the 16 resequenced strains as reference strains. 62 strains out of 94 strains were genotyped by Wellcome Trust, and 47 of them were not included in the 16 reference strains. A total of 493,033 genotypes in the validation set were evaluated for imputation accuracy, and the overall imputation error was 4.86%. 353,704 (71.7%) genotypes fall into high-confidence genotypes, and the imputation errors on these high-confidence genotypes are 0.37%. It should be noted that our imputation errors for high-confidence category is more than ten times smaller than the recently published results which used a different imputation method at a similar level of call-rate[200]. Their imputation errors at high-confidence genotypes were reported to be 4.4% with 69.5% call rate. When excluding eleven wild-derived strains the average error reduced to 2.26%, which is slightly lower than what we observed in 12 classical inbred strains with leave-one-strain-out cross-validation. Among the rest of 36 non-wild strains, 344,747 (88.9%) genotypes out of 387,817 fall into the high-confidence category with an average imputation error of 0.35%, suggesting a high coverage of the mouse HapMap SNP sets with high imputation accuracy.

3.3.2 Imputation of HapMap SNPs in WTCCC samples

We applied EMINIM to impute the uncollected HapMap SNPs of the 1,376 WTCCC control samples. We compared the imputation accuracy and memory efficiency with other published methods to demonstrate the robustness of our method. Our evaluation of chromosome 22 can be extrapolated to the estimate the performance of each method on a genome-wide scale.

First, we evaluated the accuracy of our method by randomly choosing 25% of SNPs out of the collected SNPs and imputing them from the rest of the collected SNPs. We varied the initial HMM transitional parameters of each method and observed the changes of the imputation accuracy to compare the adaptivity of the methods against the bias of the initial parameter. While IMPUTE shows a considerable change of imputation accuracy based on the transitional parameters, EMINIM shows almost the same accuracy regardless of the initial values of HMM parameters, because the optimal parameters are learned from the genotype and haplotype data using EM algorithm. The accuracy table consistently shows that our method has a higher accuracy than the previous methods (Table 3.2). The imputation accuracy of MACH was outperformed by EMINIM and IMPUTE. Since MACH does not use genetic map as input, we ran EMINIM using physical map instead of genetic map to compare their performance in the absence of genetic map, and EMINIM still showed a higher accuracy.

Next, we compared the memory efficiency between different algorithms. Since each imputation method requires a significantly large amount of memory to impute a large genomic region, the memory efficiency is an important issue when practically using the methods. The methods requiring too large amount of memory need to partition the chromosome into smaller segments at the expense of increased error rates and redundant computation. While IMPUTE requires 6.6GB of memory space, to impute all the polymorphic HapMap SNPs in chromosome 22, EMINIM requires only 508MB of memory, and MACH used 502MB of memory. This is mainly due

Table 3.2: Estimated error rates of each imputation method with different transition parameters across different confidence cutoffs. $\theta = 3.8$ is suggested by Marchini et. al.[132], and different parameters are applied to demonstrate the effect of the initial parameters. The values in parenthesis represents the fraction of imputed genotypes with confidence above the threshold. Note that MACH does not provide the posterior probability at genotype level

confidence cutoff	> 0.9	> 0.8	> 0.7	all
EMINIM ($\theta_0 = 3.8$)	1.23% (81%)	2.09% (87%)	3.07% (91%)	6.57% (100%)
EMINIM ($\theta_0 = 0.38$)	1.23% (81%)	2.09% (87%)	3.07% (91%)	6.57% (100%)
EMINIM ($\theta_0 = 0.038$)	1.23% (81%)	2.09% (87%)	3.07% (91%)	6.57% (100%)
IMPUTE ($\theta = 3.8$)	1.35% (81%)	2.25% (87%)	3.21% (91%)	6.61% (100%)
IMPUTE ($\theta = 0.38$)	2.79% (87%)	4.00% (91%)	5.04% (94%)	7.52% (100%)
IMPUTE ($\theta = 0.038$)	3.97% (88%)	5.19% (92%)	6.16% (95%)	8.23% (100%)
EMINIM (physical map)	1.50% (79%)	2.62% (86%)	3.78% (91%)	7.29% (100%)
MACH	N/A	N/A	N/A	7.69% (100%)

to the fact that IMPUTE consumes memory space for each uncollected SNP while EMINIM requires memory space only for collected SNPs using a silent state when imputing each uncollected SNP. Such a difference may be substantial in a larger chromosome such as chromosome 1 which has more than five times as many SNPs as chromosome 22. In this case, EMINIM is expected to use 2.5GB of memory while IMPUTE may require 33GB of memory space. Such a difference may be more crucial as the number of reference samples increases. The overall CPU time of EMINIM was 4.7 hours with Intel Xeon E5320 Processor, which was faster than IMPUTE 0.5.0 (6.5 hours) and MACH 1.0 (7.2 hours with only 10 rounds), despite the fact that EMINIM runs HMM multiple times per individuals to estimate the EM parameters.

3.4 Conclusion

We have proposed an adaptive and memory efficient imputation method EMINIM. Our method adaptively learns HMM parameters using an exact EM algorithm. As a result, both in the human and inbred mouse strain imputation problems, our method is shown to outperform previous imputation methods specifically designed for each organism. In addition, the memory requirement of our method is independent of the number of uncollected SNPs by utilizing silent states, which significantly increase the scalability and computational efficiency of our method to genome-wide imputation.

Chapter 3 was published in Proceedings of the 13th Annual Conference on Research in Computational Biology (RECOMB-2009), Tuscon, Arizona, May 18-21, 2009. Hyun Min Kang, Noah Zaitlen, Buhm Han, and Eleazar Eskin, “An adaptive and efficient algorithm for genotype imputation”. The dissertation author was the primary investigator and author of this paper.

Chapter 4

Efficient control of population structure in model organism association mapping

4.1 Motivation

With the recent development of high-throughput genotyping technologies, genetic variation in many model organisms such as mice, Arabidopsis, and maize are being discovered on a genome-wide scale[166, 97, 65]. Genome-wide association mapping in model organisms has great potential to identify risk factors for complex traits related to human diseases. Although straight inference from model organisms to human traits may not always apply, model organism association mapping is potentially more powerful than human association mapping because it is possible to reduce the effect of environmental factors by replicating phenotype measurements in genetically identical organisms[17]. In addition, it is often easier and more cost effective to verify associated signals via *in vivo* and/or *in vitro* experiments in model organisms than in human subjects. Moreover, many ongoing genotyping and phenotyping projects in model organisms such as the Mouse Phenome Database (MPD)[74] and Mouse HapMap projects (`http://www.mousehapmap.org`) provide publicly available resources to perform *in-slico* mapping of complex traits in model organisms[162].

However, genetic association studies in inbred model organisms are confronted by the problem of inflated false positive rates due to population structure and genetic relatedness among inbred strains often caused by the complex genealogical history of most model organisms strains. Applying conventional statistical tests of independence between a genetic marker and a phenotype is prone to spurious associations because the marker and the phenotype are likely to be correlated via population structure which violates the independence assumption under the null hypothesis. Recent association or linkage mapping studies in model organisms attempt to avoid inflated false positive rates by designing the studies using recombinant inbred lines generated from a handful number of parental strains[243, 29]. However, these studies are limited by the variation present in the parental strains and have long regions between recombinations due to relatively few generations between the recombinant inbred strains and the parental strains. Traditional QTL mapping using F2 or backcross suffers from the same problem in the fine-resolution mapping in addition to the problem of expensive genotyping cost[17, 64].

An alternative approach to reduce the inflation of false positives is to apply a statistical test that corrects for the bias due to population structure or genetic relatedness. The most widely used methods to reduce such bias in human association mapping are Genomic Control[50], Structured Association[170], and EIGENSTRAT which uses principal component analysis for population stratification[168, 158]. However, these methods are inadequate in the case of model organism association mapping. Genomic Control suffers from weak power when the effect of population structure is large as in model organisms. Structured Association or EIGENSTRAT, which assume a handful number of ancestral populations and admixture, only partially captures the multiple levels of population structure and genetic relatedness in model organisms[8, 235, 240]. Recently, it is suggested that linear mixed model can effectively correct for population structure in the association mapping of quantitative traits[235]. Linear mixed models incorporate pairwise genetic relatedness between every pair of individuals in the statistical model directly, reflecting that the phenotypes of two genetically similar subjects are more likely to be correlated than are genetically dissimilar ones. Applications of mixed models to association mapping in maize, Arabidopsis, and potato panels demonstrate that mixed models obtain fewer false positives and higher power than previous methods including Genomic Control, Structured Association, and EIGENSTRAT[235, 240].

Although mixed models can effectively capture statistical confounding due to population structure, the currently available implementations have several limitations in the context of model organism association mapping. First, the variance component numerically estimated by Nelder-Mead simplex algorithm[147, 73, 141], EM algorithm, and Newton-Raphson algorithm[122, 72, 99]. only provide locally optimal solution, which may cause the statistical inferences based on these estimates to be inaccurate. Second, the computational cost of the numerical optimization procedure is substantial, requiring a large number of computationally expensive matrix operations at each iteration. Computational considerations are important when a large number of SNPs are to be tested. For example, the association mapping with maize panels consisting of hundreds of SNPs over hundreds of strains takes hours for a single run with currently available implementations such as TASSEL[235] or SAS[92]. If one million SNPs were available for genome-wide analysis with the same number of strains, a single run of association mapping with mixed model would take several months of CPU time. Third, when inferring the genetic variance component referred to as the kinship matrix, the importance of a mathematically correct form of kinship matrix estimation is often overlooked. For example, Yu et. al.[235] proposed to infer kinship matrix using SPAGeDi software, setting negative kinship coefficients to zero. However, such a kinship matrix may not be positive semidefinite and thus not be a valid form of variance component. Using a non-positive semidefinite kinship matrix generates noncontinuous search space for optimization and may disrupt the convergence[92].

In this chapter, we propose a new method, Efficient Mixed Model Association (EMMA), which corrects for population structure and genetic relatedness in model organism association mapping. Our method takes advantage of the specific nature of the optimization problem in applying mixed models for association mapping, which allows us to substantially increase computational speed by several orders of magnitude and improve the reliability of results by achieving near global optimization. Standard iterative methods for estimating variance components imputes unessential parameters such as individual random effects in BLUP(Best Linear Unbiased Prediction)[86]. Our method improves the efficiency of the mixed model method by enabling us to perform statistical tests without BLUP estimation, reducing the number of dimensions that need to be numerically optimized to one. Our method's efficiency is further increased by avoiding redundant computationally expensive matrix operation at each iteration in the computation of likelihood function by leveraging spectral decomposition, reducing the computational cost of each iteration from cubic to linear complexity. Our method is related to a similar technique developed in a different context of linear mixed model, for simulating null distribution of likelihood ratio test statistics efficiently when testing for the significance of a variance component[43]. Due to substantially decreased computational cost of each iteration, it is possible to converge global optimum of the likelihood in variance component estimation with high confidence by combining grid search and Newton-Raphson algorithm even though the likelihood function may not be convex.

We discuss how to design the kinship matrix accounting for genetic relatedness while guaranteeing convergence in the optimization procedure. We show that simple genetic similarity matrix with appropriate handling of missing genotypes guarantees convergence in optimization. Our results are consistent with other studies[240], suggesting that these simpler kinship matrices reduce the false positive rate as effectively as or more effectively than the kinship matrices generated by previous methods[235]. We suggest another method called *phylogenetic control* based on the assumption that a phylogenetic tree is a good approximation of the genealogical history of an inbred model organism. In such cases, the phylogenetic tree may be used as a confounding factor, correcting for complex genetic relatedness between strains. We show that phylogenetic control can be formulated into a linear mixed model, and present an algorithm for inferring the phylogenetic kinship matrix. We show that the kinship matrix is always positive semidefinite and its optimal variance components are unique regardless of the choice of root.

One of the important questions in the design of model organism association mapping studies is estimating the study power for any specific set of inbred strains. We performed a simulation study of the power of our EMMA method to identify causal SNPs both on a genome-wide scale and within a smaller region such as a QTL interval. Our results show that with a limited number of genetically diverse strains, such as the currently available panel of inbred mice, it is possible to identify causal loci with a genome-wide significance only if the locus explains a large portion of phenotypic variance. However, with more strains, the power of these association studies increases dramatically. Our analysis of statistical power in model organism association mapping demonstrates the dramatic increase in power using multiple measurements of phenotypes from multiple animals for each strain. Study designs that do not replicate phenotype measurements and analysis methods that do not take individual measurements into account suffer a significant decrease in statistical power.

We applied our EMMA method to association mappings of various inbred model organisms. First, we verified that EMMA gives almost identical results to other widely used implementations using the maize panel datasets[235]. In terms of computational time, EMMA is shown to be orders of magnitude faster than the previous methods while performing global optimization. Second, we performed a genome-wide association mapping of Arabidopsis flowering time phenotypes. Our results are consistent to the recently published results[240], reducing most of inflated false positives. Finally, we used our EMMA method to perform a whole genome association mapping study of inbred mouse strains. We analyzed nearly 140,000 mouse HapMap SNPs over 48 strains and three quantitative phenotypes, liver weight, body weight, and saccharin preference, with QTLs identified by previous studies. We identified significant associations for the three phenotypes while our results show a significant reduction in the inflation of false positives. Interestingly, many of the significantly associated SNPs fall into the known QTLs, suggesting the results are likely to be true associations. The implementation of our method via R package, as well as the mouse association results are are publicly available online at `http://mouse.cs.ucla.edu/emma`.

4.2 Materials and methods

4.2.1 Genotypes and phenotypes

Genotypes, phenotypes, SPAGeDi-based kinship matrix, and the STRUCTURE outputs from 277 maize strains across 553 SNPs as described in [235] are downloaded from Ed Buckler's lab web site(`http://www.maizegenetics.net`). The Arabidopsis genotypes, phenotypes, and the output from STRUCTURE were obtained from the published datasets[8, 150]. The 13,416 non-singleton SNPs with no more than 10% of genotype calls missing, were tested for association after imputing the missing alleles using HAP[82]. The flowering time phenotypes over 95 strains were log-transformed to fit to a normal distribution.

For inbred mouse association mapping, the BROAD mouse HapMap SNP sets were obtained from the Mouse HapMap web site. The 106,040 SNPs which have no more than 10% of genotype calls missing were tested after imputing the missing alleles. The initial body weight(MPD10305) and liver weight phenotypes(MPD2907) were downloaded from Jackson Laboratory MPD[74]. They consist of 374 and 308 phenotype measurements over 38 and 34 strains, respectively. The saccharin preference phenotypes consist of 280 phenotype measurements in 24 strains[176].

4.2.2 Efficient mixed model association (EMMA)

Suppose that n measurements of a phenotype are collected across t inbred strains. A linear mixed model in model organism association mapping is typically expressed as the following equation.

$$\mathbf{y} = X\beta + Z\mathbf{u} + \mathbf{e} \tag{4.1}$$

where $\mathbf{y}$ is a $n \times 1$ vector of observed phenotypes, X is are $n \times q$ matrix of fixed effects including mean, SNPs and other confounding variables. β is $q \times 1$ vector representing coefficients of the fixed effects. Z is $n \times t$ incidence matrix mapping each observed phenotype to one of t inbred strains. $\mathbf{u}$ is the random effect of the mixed model with $\text{Var}(\mathbf{u}) = \sigma_g^2 K$ where K is the $t \times t$ kinship matrix inferred from genotypes as described in the following section, and $\mathbf{e}$ is a $n \times n$ matrix of residual effect such that $\text{Var}(\mathbf{e}) = \sigma_e^2 I$. The overall phenotypic variance-covariance matrix can be represented as $V = \sigma_g^2 ZKZ' + \sigma_e^2 I$.

Instead of solving mixed model equations by obtaining the best linear unbiased prediction (BLUP) of random effects $\mathbf{u}$ via Henderson's iterative procedure, we directly estimate the variance components σ_g and σ_e maximizing the full likelihood or restricted likelihood which is defined as full likelihood with the fixed effects integrated out[48]. The restricted likelihood avoids a downward bias of maximum likelihood estimates of variance components by taking into account the loss in degrees of freedom associated with fixed effects. Under the null hypothesis, the full log-likelihood and restricted log-likelihood function can be formulated as follows[222].

$$\begin{aligned} l_F(\mathbf{y};\beta,\sigma,\delta) &= \frac{1}{2}\left[-n\log(2\pi\sigma^2) - \log|H| - \frac{1}{\sigma^2}(\mathbf{y}-X\beta)'H^{-1}(\mathbf{y}-X\beta)\right] \quad (4.2)\\ l_R(\mathbf{y};\sigma,\delta) &= l_F(\mathbf{y};\hat{\beta},\sigma^2,\delta) + \frac{1}{2}\left[q\log(2\pi\sigma^2) + \log|X'X| - \log|X'H^{-1}X|\right] \quad (4.3)\end{aligned}$$

where $\sigma = \sigma_g$ and $H = \sigma^{-1}V = ZKZ' + \delta I$ is a function of δ, defined as $\delta = \sigma_e^2/\sigma_g^2$.

The full likelihood function is maximized when β is $\hat{\beta} = (X'H^{-1}X)^{-1}X'H^{-1}\mathbf{y}$, and optimal σ^2 is $\hat{\sigma_F}^2 = R/n$ for l_F for l_F and $\hat{\sigma_R}^2 = R/(n-q)$ for l_R for l_R, where $R = (\mathbf{y}-X\hat{\beta})'H^{-1}(\mathbf{y}-X\hat{\beta})$ is a function of δ as well.

Using spectral decomposition, it is possible to find ξ_i and λ_s such that

$$\begin{aligned} H = ZKZ' + \delta I &= U_F\text{diag}(\xi_1+\delta,\cdots,\xi_n+\delta)U_F' \quad (4.4)\\ SHS = S(ZKZ'+\delta I)S &= [U_R\ W_R]\text{diag}(\lambda_1+\delta,\cdots,\lambda_{n-q}+\delta,0,\cdots,0)[U_R\ W_R]'\\ &= U_R\text{diag}(\lambda_1+\delta,\cdots,\lambda_{n-q}+\delta)U_R' \quad (4.5)\end{aligned}$$

where $S = I - X(X'X)^{-1}X'$, U_F is $n \times n$, U_R is $n \times (n-q)$ eigenvector matrix corresponding to the non-zero eigenvalues. W_R is $n \times q$ eigenvector matrix corresponding to zero eigenvalues[157]. Note that U_R is independent of δ. Let $U_R'\mathbf{y} = [\eta_1\ \eta_2 \cdots \eta_{n-q}]'$, then finding ML or REML estimates is equivalent to optimizing the following functions with respect to δ.

$$\begin{aligned} f_F(\delta) &= l_F(\mathbf{y};\hat{\beta},\hat{\sigma},\delta) = \frac{1}{2}\left[n\log\frac{n}{2\pi} - n - n\log\left(\sum_{s=1}^{n-q}\frac{\eta_s^2}{\lambda_s+\delta}\right) - \sum_{i=1}^{n}\log(\xi_i+\delta)\right] \quad (4.6)\\ f_R(\delta) &= l_R(\mathbf{y};\hat{\sigma},\delta) \quad (4.7)\\ &= \frac{1}{2}\left[(n-q)\log\frac{n-q}{2\pi} - (n-q) - (n-q)\log\left(\sum_{s=1}^{n-q}\frac{\eta_s^2}{\lambda_s+\delta}\right) - \sum_{s=1}^{n-q}\log(\lambda_s+\delta)\right]\end{aligned}$$

(See Section 4.2.7 for the mathematical details) The derivatives of these functions follows that

$$f'_F(\delta) = \frac{n}{2} \cdot \frac{\sum_s \eta_s^2/(\lambda_s+\delta)^2}{\sum_s \eta_s^2/(\lambda_s+\delta)} - \frac{1}{2}\sum_i \frac{1}{\xi_i+\delta} \tag{4.8}$$

$$f'_R(\delta) = \frac{n-q}{2} \cdot \frac{\sum_s \eta_s^2/(\lambda_s+\delta)^2}{\sum_s \eta_s^2/(\lambda_s+\delta)} - \frac{1}{2}\sum_s \frac{1}{\lambda_s+\delta} \tag{4.9}$$

The suggested procedure in computing likelihood and its derivatives involves only linear time vector operation at each iteration once the spectral decomposition is computed. The time complexity of the method is $O(n^3 + rn)$ where r is the number of iterations required. The time complexity of standard EM or Newton-Raphson algorithms is $O(rn^3)$, and the actual ratio of the running time is much bigger than r because the existing methods typically requires a large number of matrix multiplications and inverses at each iteration while EMMA computes spectral decomposition only once. Since the computational cost of each iteration has decreased dramatically, instead of obtaining locally optimal solution during the numerical optimization, it is now computationally feasible to perform grid search combining with Newton-Raphson algorithm in a single dimensional parameter space of δ, which is the ratio of environmental random effect to genetic background effect, to optimize the likelihood globally with high confidence.

Furthermore, when a large number of multiple measurements are phenotyped per strain, i.e. $n \gg t$, the execution time can be further reduced using the fact that the nonnegative eigenvalues of ZKZ' and $SZKZ'S$ are the same as those of $KZ'Z$ and $KZ'SZ$, respectively. Combining this fact with a simple modification of Gram-Schmidt process greatly reduces the execution time of eigenvalue decomposition, reducing the time complexity into $O(t^3 + n^2t + rn)$.

In the application of our EMMA method to the various datasets presented in this chapter, the δ ranged from 10^{-5} (almost pure population structure effect) to 10^5 (almost pure environmental or residual effect), and are divided evenly into 100 regions in logarithm scale to compute the derivatives of likelihood functions. The global ML or REML is searched for by applying the Newton-Raphson algorithm to all the intervals where the signs of derivatives change, and taking the optimal δ

amongst all of the stationary points and endpoints. Since the derivatives of both the full and restricted likelihood function are continuous with positive semidefinite kinship matrices suggested in the following sections, such an optimization technique has guaranteed convergence properties as long as the kinship matrix is positive semidefinite. In the following two sections, we describe different methods to infer a kinship matrix K, based on either genetic similarity matrix or phylogenetic tree.

4.2.3 Similarity-based kinship matrix

A number of methods for inferring kinship matrix from a large number of molecular markers have been suggested, including simple identical by state (IBS) allele sharing matrix, allele-frequency weighted IBS matrix Lynch[127], Maximum-likelihood kinship matrix[203], and Monte Carlo simulation-based matrix[215]. Comparisons of different kinship matrices for explaining genetic differentiation among populations shows similar results but small quantitative difference[149]. Recent studies on the association mapping of *Arabidopsis thaliana* in structure population shows that a simple IBS allele sharing matrix effectively corrects for confounding from population structure, even better than more sophisticated methods[240]. Although recently suggested estimator of pairwise relatedness have some desirable statistical properties than simple IBS allele sharing matrix[207], they are not guaranteed to be positive semidefinite and thus may disrupt the convergence is the estimation of variance components.

Here we show that a simple IBS allele sharing matrix based on the assumption of each SNP or haplotype inducing same level of small random changes on the phenotype guarantees positive semidefiniteness and convergence if missing alleles are handled appropriately.

Let $l_{i,j,h} \in \{0,1\}$ be a binary variable which has a value of one only when the genotype (or haplotype) allele at j-th locus in i-th strain is $h \in 1, \cdots, |\mathcal{H}_j|$ where $|\mathcal{H}_j|$ is the total number of alleles at j-th locus. Let $x_{h,j}$ be random variables independently sampled from $N(0, \sigma^2)$, then the genetic background effect u_i of strain i can be modeled as an accumulation of small random effects as follows, assuming that $x_{h,j}$ denote the random genetic effect caused by allele h at j-th locus.

$$u_i = \sum_j \sum_{h=1}^{\mathcal{H}_j} l_{i,j,h} x_{h,j} \tag{4.10}$$

Let $|\mathcal{H}| = \max(|\mathcal{H}_j|)$, and let L_h be the matrix whose element at (i, j) is $l_{i,j,h}$, then the overall genetic background effect $\mathbf{u}$ is expressed in the following form.

$$\mathbf{u} = \sum_{h=1}^{|\mathcal{H}|} L_h \mathbf{x}_h \tag{4.11}$$

Assuming that each $x_{h,j}$ follows a normal distribution with zero mean and variance of σ^2 independently, the variance-covariance matrix of $\mathbf{u}$ becomes $\mathrm{Var}(\mathbf{u}) = \sigma^2 \sum_h L_h L_h'$. Since its (i_0, i_1)-th element $\sum_h \sum_j l_{i_0,j,h} l_{i_1,j,h}$ represent the number of shared IBS alleles between the i_0-th and i_1-th strains directly, $\mathrm{Var}(\mathbf{u})$ is equivalent to the IBS allele sharing kinship matrix with a scaling factor. It is obvious from the equation 4.11, that the kinship matrix is positive semidefinite. When missing genotypes exists, we estimate $l_{i,j,h}$ to be the square root of the probability of the SNP or haplotype allele at j-th locus having the allele h. This is so that the random effect for each allele is assigned probabilistically. When haplotype similarity matrix is used, the haplotype window size resulting in the largest ML estimates is selected as optimal window size. In the Arabidopsis and mouse association mapping results of this chapter, the optimal haplotype window size is set to five in both cases.

4.2.4 Phylogenetic control

Evolutionary biologists have tried modeling inter-specific phenotype distribution using various phylogenetic comparative methods (PCMs)[133]. The correlation structure between phenotypes can be effectively captured with phylogenetic trees, and PCMs have been applied to evolutionary analysis of quantitative traits such as gene-expression[154, 76], or, very recently, to the association mapping of dichotomous phenotypes[19]. Felsentein's independent contrast (FIC) method [58] models the correlation between phenotypes under the assumption of Brownian motion of phenotypic change along the phylogeny due to random genetic drift. Since random genetic drift occurs within a species as well, in cases where the phylogenetic tree is a good approximation of genealogical history, it is reasonable to apply PCMs such as the FIC method in modeling the phenotypic variation in model organisms.

We followed the Felsenstein's assumption of Brownian phenotypic changes along the phylogeny. Under this assumption, the branch length between any two nodes is proportional to the phylogenetic covariance of phenotypes. Let T be a phylogenetic tree with t leafs and m edges, and let $\mathbf{z} \in \mathbf{R}^m$ be random variables independently sampled from $N(0, \sigma_g^2)$. At each branch i whose length is b_i, we represent the amount of random phenotypic changes along the branch as $\sqrt{b_i} z_i$. Let Ψ_i denote the set of branches connecting to a leaf node i from the root. Then the amount of phenotypic changes due to genetic drift is equivalent to $\sum_{e \in \Psi_i} \sqrt{b_e} z_e$. If $X\beta$ is the ancestral mean at an arbitrarily chosen root node, then the phenotype values at the leaf nodes are expressed in the following form,

$$\mathbf{y} = X\beta + ZE\mathbf{z} + \mathbf{e} \tag{4.12}$$

where E is an $t \times m$ matrix whose (i, j)-th element is $\sqrt{b_j}$ if branch j exists in the path from the root to the leaf node i, and zero otherwise. The kinship matrix of random effect $\mathbf{u} = E\mathbf{z}$ is $K = EE'$, and is proportional to its covariance. If the root of the phylogenetic tree changes, E is changed into $E + \mathbf{1}_t \mathbf{c}^T$, with $\mathbf{1}_t$ a vector of ones and another vector $\mathbf{c}$. However, the restricted likelihood does not change because $SZ\mathbf{1}_t = 0$ always holds.

In the experiments, we adjusted the genetic distance matrix using the F84 model [110, 59] from the genome-wide genotypes, and inferred the phylogenetic tree with the Fitch-Margoliash and least-squared distance method[62].

4.2.5 Statistical tests and multiple hypothesis testing

Once the ML or REML variance component $\hat{V} = \hat{\sigma_g}^2 K + \hat{\sigma_e}^2 I$ is estimated, a general F-statistic testing the null hypothesis $M\beta = 0$ for an arbitrary full-rank $p \times q$ matrix M can be constructed as suggested in [235, 106]

$$F = \frac{(M\hat{\beta})'(M(X'\hat{V}^{-1}X)^{-1}M')^{-1}(M\hat{\beta})}{p} \tag{4.13}$$

with p numerator degrees of freedom and $n - q$ denominator degrees of freedom. The Satterthwaite degree of freedom may also be computed avoiding computationally intensive matrix operations.

Likelihood ratio test can also be performed based on the estimated variance components under different fixed effects. The statistic asymptotically follows χ_p^2 distribution unless the estimated variation component meets boundary of parameter space. The genetic variance component explained by a SNP can be computed by comparing σ_g^2 under H_0 and H_1, otherwise a conventional way to compute the explained variance can be used.

When a large number of correlated SNPs are tested, Bonferroni correction may lead to too conservative Type I error control, and permutation tests can be used alternatively. In this case, the computational cost becomes even much larger but it can be reduced by reusing the spectral decomposition results for different set of permuted phenotypes. Since U_F, U_R is independent of y, it can be reused and only $U_R'y = [\eta_1, \eta_2, \cdots, \eta_{n-q}]$ has to be computed again in order to compute the full or restricted likelihood in linear time at each iteration. Thus, the computational cost for a cubic-time spectral decomposition at each permutation can be substituted by a square-time matrix-vector multiplication, reducing the overall time complexity from $O(t^3 + n^2t + rn)$ to $O(n^2 + rn)$.

In our results presented in the following sections, we applied the F-test with p-values computed from the asymptotic F distribution. The variance components are estimated via REML. The likelihood ratio test is also performed when comparing the different statistical methods. In this case, ML estimation is used with the full likelihood function described above.

4.2.6 Simulation studies

We performed two simulation studies for analyzing the statistical power of EMMA. First simulation is similar to those from other mixed model studies[235, 240]. A fixed effect based on a randomly chosen causal SNP across the genome with minor allele frequency greater than 10% is added to the existing phenotypes, and the statistical power is computed at the causal SNP. At each fixed effect, the simulation study was performed 1000 times to estimate the average power. The variance explained by a SNP is computed assuming that average minor allele frequency of the causal SNP is 0.3.

Next, we generated simulated phenotypes sampled from multivariate normal distribution with the kinship matrix as covariance using `mvrnorm` function in R `MASS` package. A random noise vector is added according to the contribution of genetic background to phenotypes, h_g^2. If h_g^2 is the fraction of variance due to genetic background excluding the SNP effect, then the covariance of the simulated data is simulated as $\mathrm{Var}(y) = (nh_g^2/\mathrm{tr}(S_0ZKZ'S_0))K + (1 - h_g^2)I)$ where $S_0 = I - \mathbf{11}'/n$. Similar to the first simulation study, a fixed effect based on a randomly chosen causal SNP is added to the simulated phenotypes and the average power is computed from 1000 times of independent simulations.

4.2.7 Derivation of restricted likelihood and derivatives

A derivation of Equation 4.6 and 4.7 from Equation 4.2 and 4.3 is presented in [157, 84]. However, its derivation is not straightforward, and it needs to be clarified how exactly it is related to spectral decomposition. Here we describe a more detailed description of obtaining Equation 4.6 and 4.7.

Plugging in the optimal parameters $\hat{\beta}$ and $\hat{\sigma}_F = R/n$ in Equation 4.2, it follows that

$$f_F(\delta) = l_F(\mathbf{y}; \hat{\beta}, \hat{\sigma}, \delta) = \frac{1}{2}\left[-n\log\frac{2\pi R}{n} - \log|H| - n\right] \tag{4.14}$$

From Equation 4.4, it is straightforward that $\log|H| = \sum_{i=1}^{n}\log(\xi + \delta)$. And R can be rewritten as follows

$$\begin{aligned} R &= (\mathbf{y} - X\hat{\beta})'H^{-1}(\mathbf{y} - X\hat{\beta}) && (4.15)\\ &= \mathbf{y}'(I - X(X'H^{-1}X)^{-1}X'H^{-1})'H^{-1}(I - X(X'H^{-1}X)^{-1}X'H^{-1})\mathbf{y} && (4.16)\\ &= \mathbf{y}'P'H^{-1}Py && (4.17) \end{aligned}$$

It is straightforward to show that

$$\begin{aligned} (SHS)(P'H^{-1}P)(SHS) &= SHS && (4.18)\\ (P'H^{-1}P)(SHS)(P'H^{-1}P) &= P'H^{-1}P && (4.19) \end{aligned}$$

using the fact $PS = S$ and $SP = S$. Consequently,

$$P'H^{-1}P = (SHS)^{+} = U_R \text{diag}\left[(\lambda_s + \delta)^{-1}\right] U_R' \tag{4.20}$$

where $(\cdot)^{+}$ denotes the pseudoinverse of a matrix. Therefore, it follows that

$$\begin{aligned} R &= \mathbf{y}'(P'H^{-1}P)\mathbf{y} & (4.21)\\ &= (U_R'\mathbf{y})'\text{diag}\left[(\lambda_s + \delta)^{-1}\right](U_R'\mathbf{y}) & (4.22)\\ &= \sum_{s=1}^{n-q} \frac{\eta_s^2}{\lambda_s + \delta} & (4.23) \end{aligned}$$

From Equation 4.14 and 4.23, it follows that

$$f_F(\delta) = \frac{1}{2}\left[n \log \frac{n}{2\pi e} - n \log\left(\sum_{s=1}^{n-q} \frac{\eta_s^2}{\lambda_s + \delta}\right) - \sum_{i=1}^{n} \log(\xi_i + \delta)\right] \tag{4.24}$$

The restricted likelihood of $\mathbf{y}$ is equivalent to computing the likelihood of $A\mathbf{y}$ where $S = AA'$ and $A'A = I$[84, 157].

$$(SHS)(SHS)^{+} = (SHS)(P'H^{-1}P) = SHP'H^{-1}P = SP = S \tag{4.25}$$

On the other hand,

$$(SHS)(SHS)^{+} = (U_R \text{diag}(\lambda_s + \delta)U_R')\left(U_R \text{diag}\left[(\lambda_s + \delta)^{-1}\right] U_R'\right) = U_R U_R' \tag{4.26}$$

Accordingly, $U_R U_R' = S$ and $U_R' U_R = I$ hold, and the restricted likelihood of $\mathbf{y}$ is equivalent to the likelihood of $U_R'\mathbf{y} \sim N(0, \sigma^2 \text{diag}(\lambda_s + \delta))$. By plugging in $\hat{\sigma}_R^2$ to σ^2, it immediately follows that

$$f_R(\delta) = \frac{1}{2}\left[(n-q) \log \frac{n-q}{2\pi e} - (n-q) \log\left(\sum_{s=1}^{n-q} \frac{\eta_s^2}{\lambda_s + \delta}\right) - \sum_{s=1}^{n-q} \log(\lambda_s + \delta)\right] \tag{4.27}$$

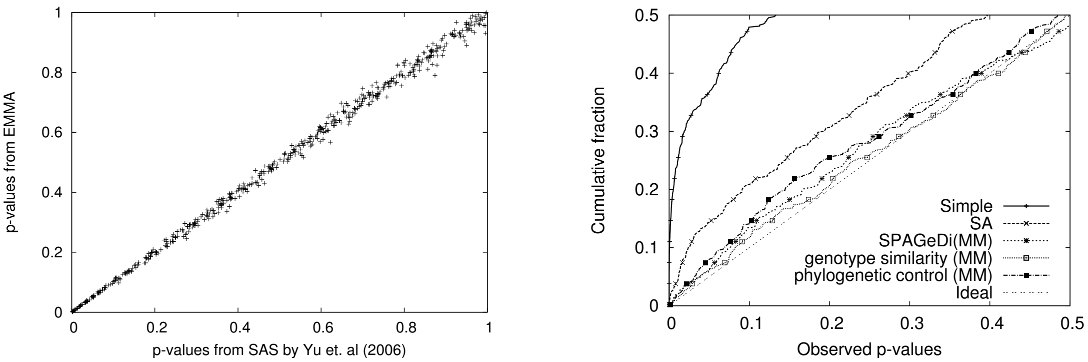
4.3 Results

4.3.1 Comparison with previous methods

We applied our EMMA method to the same maize panel data consisting of 553 SNPs and three phenotypes across 277 diverse inbred lines[65] analyzed with the unified mixed model [235]. The kinship matrix is inferred by SPAGeDi software[83], setting all negative coefficients to zero. The population structure coefficients are estimated from STRUCTURE[169] using 89 microsatellite loci for three subpopulations. Figure 4.1(a) shows the comparison of the p-values obtained from the previous unified mixed model (Structured Association (SA) + Mixed Model (MM) method) with those from EMMA for flowering time phenotypes. They are almost identical, differing only due to the differences in the numerical optimization procedure for the estimation of variance components.

While both the SAS and TASSEL implementations of mixed model [235] take nearly 2 hours for a single run over these datasets with Intel 2.8GHz Dual Core CPU, the execution time of our mixed model implementation is nearly hundred times faster, taking only 90 seconds. The results of our method are more reliable because we find the global REML estimate with guaranteed convergence properties. Previous implementations iteratively search for local optima with unknown convergence properties. Possibly due to the instability of the convergence properties, the TASSEL implementation could not compute p-values for several loci in the maize panel analysis.

Since the kinship matrix inferred from SPAGeDi software is not positive semidefinite, we explore other ways to estimate the variance components due to genetic background. We use a genotype similarity matrix and a phylogenetic control matrix which guarantee positive semidefiniteness. Haplotype similarity matrices are not applicable to this dataset due to sparse genotype density. We compared the goodness-of-fit of these kinship matrices in addition to the SPAGeDi-based kinship matrix over three maize phenotypes using Bayesian Information Criterion (BIC), which provides a measure of how well each model fits the data. Adjusting for the sample size and the number of free parameters, Table 4.1 shows that the goodness-of-fits of the three kinship matrices based on maximum likelihood estimates are com-



(a) Pairwise comparison of p-values between unified mixed model[235] and our method (EMMA with SA+MM) for flowering time association in maize panel

(b) Cumulative Distribution of p-values across different model

Figure 4.1: (a) Direct comparison of p-values between the SAS implementation of unified mixed model[235] and our method, computed from 553 SNPs of maize panel data and the flowering time phenotype. All p-values are extremely highly correlated, implying that two methods are almost identical in terms of accuracy. (b) Cumulative distribution of p-values across different models. Under the assumption that the SNPs are unlinked and there few true SNP association, the observed p-values are expected to be close to the cumulative p-values. A large deviation from the expectation implies that the statistical test may cause spurious associations. Simple is a simple t-test, SA is Structured Association, and MM is F test mixed model with specified kinship matrix

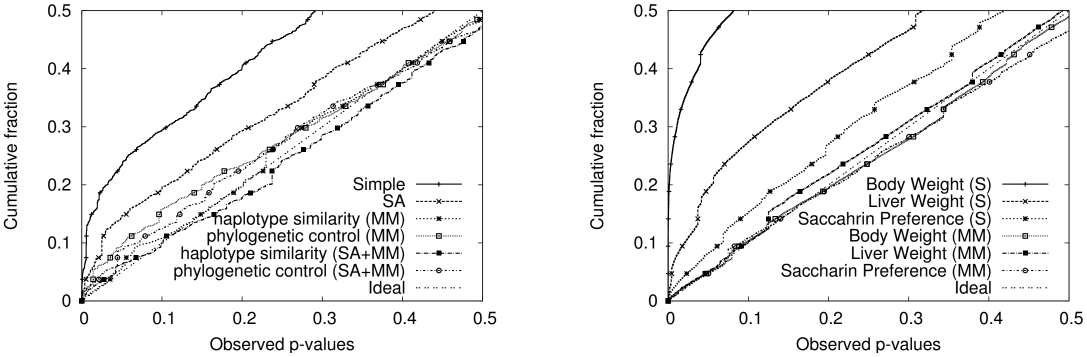
Table 4.1: Goodness of fit of different models and kinship matrices in explaining phenotypic variation of maize quantitative traits. Comparison of Maximum Likelihood (ML) and Bayesian Information Criterion (BIC) of each model with different kinship matrices for maize quantitative traits. The model with the smaller BIC is preferred. 'Simple' denotes simple linear model without adjustment for population effect. SA is the model using the output from STRUCTURE as covariates. MM is mixed model with different kinship matrices. The descriptions of kinship matrices are the same as in Figure 4.1

Method	Kinship Matrix	Flowering Time		Ear Height		Ear Diameter	
		-2*(ML)	BIC	-2*(ML)	BIC	-2*(ML)	BIC
Simple	N/A	1632.8	1643.9	2296.0	2307.1	1282.6	1293.5
MM	SPAGeDi	1524.3	1541.0	2237.7	2254.3	1254.2	1270.5
MM	Genotype Similarity	1527.5	1544.2	2243.1	2259.8	1266.6	1282.9
MM	Phylogenetic Control	1521.6	1538.6	2227.3	2243.9	1248.9	1265.2
SA	N/A	1525.7	1547.9	2248.9	2271.1	1276.9	1298.7
SA+MM	SPAGeDi	1494.9	1522.7	2220.3	2248.1	1253.6	1280.8
SA+MM	Genotype Similarity	1500.9	1528.7	2227.1	2254.9	1266.5	1293.7
SA+MM	Phylogenetic Control	1491.6	1519.4	2213.2	2241.0	1248.2	1275.4

parable, while all of them were significantly better than not using a mixed model.

The cumulative p-value distribution seen in Figure 4.1(b) show that the simple genotype similarity matrix corrects for genetic relatedness slightly better than the other two kinship matrices. There is a better reduction of false positive rates, especially within the region of small p-values. Since the simpler kinship matrices show comparable or better goodness-of-fit and false positive reduction results while guaranteeing positive semidefiniteness, we apply only these simple kinship matrices in the following sections.

We also applied our EMMA method to perform genome-wide association mapping of flowering time phenotype in which statistically significant associations are reported in previous studies. The cumulative distribution of p-values across 13,416 non singleton SNPs across 95 strains obtained from EMMA is shown in Figure 4.2(a). The cumulative distribution of p-values with haplotype similarity matrix nearly follows the expected distribution, implying that mixed models significantly outperform Structured Association in eliminating the inflation of false positives for this dataset. Phylogenetic control reduces a large portion of inflated false positives, but residual inflation is still observed. Structured Association and simple linear regression showed



(a) Arabidopsis

(b) Inbred Mouse Strains

Figure 4.2: Genome-wide cumulative distribution of observed p-values between (a) 13,416 Arabidopsis SNPs and flowering time phenotypes across 95 strains using various models. (b) 106,040 Mouse HapMap SNPs and three phenotypes, body weight(374 measurements over 38 strains), liver weight(304 measurements over 34 strains), and saccharin preference (280 measurements across 24 strains). S or Simple is a simple t-test, SA is Structure Association, and MM is F test with mixed model with specified kinship matrix. SA+MM is the unified mixed model using the output of STRUCTURE as additional fixed effects

much larger inflation of false positives, consistent to the previous studies. The previously known FRI gene was found to be significant at a nominal p-value $p = 10^{-5}$ across different kinship matrices. Our independent analysis are consistent to the more extensive results of Arabidopsis association mapping recently published[240].

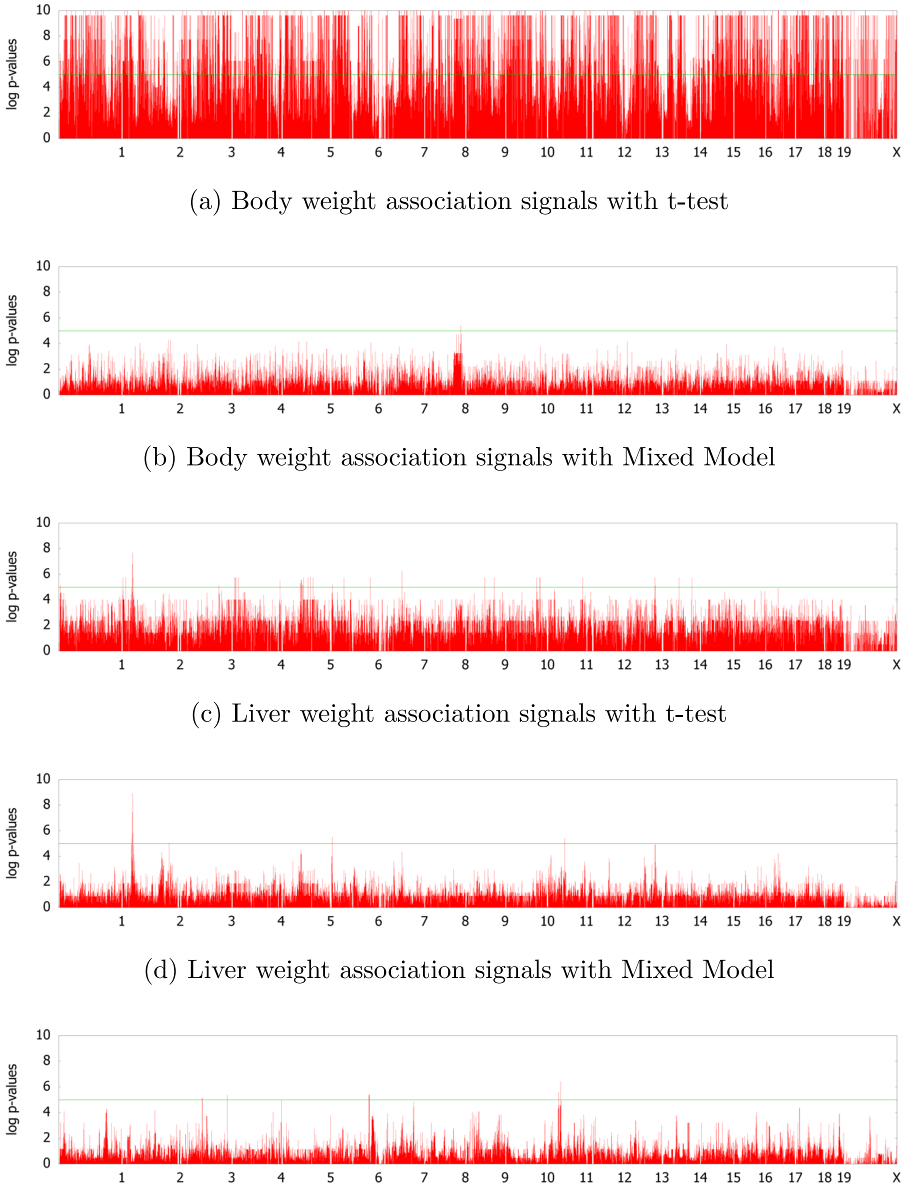
4.3.2 High resolution henome-wide association mapping in inbred mouse strains

We performed a high resolution genome-wide association mapping study using our mixed model method over inbred mouse strains. We used the Broad Mouse HapMap SNPs, containing nearly 140,000 SNPs expected to cover most of genetic variation among 48 inbred strains. For phenotypes, we used initial body weight and liver weight phenotypes downloaded from Jackson Laboratory Mouse Phenome Database[74]. In addition, we used a saccharin preference phenotype where statistically significant associations were identified in a previous study[176]. Among 48 genotyped strains, 38, 34, and 24 strains had phenotype values available for body weight, liver weight, and saccharin preference, respectively. Each phenotype has on average 10 multiple measurements across different individual mice per strain

The cumulative distributions of observed p-values in Figure 4.2 shows that, without correcting for population structure, the rate of false positives are very high. In particular, the body weight phenotype has a substantial inflation of false positives. When our mixed model is used, the inflation of the statistics are significantly reduced in all three phenotypes.

Figure 4.3 shows genome-wide association signals for the three phenotypes. Comparing Figure 4.3(a) and 4.3(b), it is obvious that, without correcting for population structure, many false positives are observed at a genome-wide level of significance due to inflated p-values. Without correcting for population structure, we were able to identify nearly 6000 SNPs at a nominal p-value of 10^{-6}, and 283 SNPs with p-values less than 10^{-10}. However, none of them are significant after applying the mixed model. This strongly supports that most of the significant associations without correcting for population structure are indeed false positives. Interestingly, although the strongest signals for the body weight with the mixed model are not genome-wide significant, they are concentrated in the region around 114Mb in chromosome 8. This region almost exactly falls into the LOD peak of a previously known body weight QTL *Bwq3*[7]. The p-value of most significant locus is 3.8×10^{-6} with F test. explaining 49% of the overall phenotypic variance and 39% of the phenotypic variation due to genetic variance component. Although it is slightly below the genome-wide significance threshold with conservative Bonferroni correction, if utilizing the results from previous QTL studies, the locus can be declared as significant over the region of known body weight QTLs.

For the liver weight phenotype, we identified a genome-wide significant association around the region of 34.5Mb in chromosome 2. This falls into a previously known liver weight QTL *Lvrq1*[182]. The region also contains many potentially relevant QTLs such as organ weight (*Orgwq2*[116]), spleen weight (*Sp1q1*[182]), heart weight (*Hrtq1*[182]), lean body mass (*Lbm1*[134]). The pointwise p-value of the most significant SNP was 1.2×10^{-9}, which explains 59% of the genetic variance component. When comparing the genome-wide p-values between simple t-test and mixed models in Figure 4.3(c) and 4.3(d), we observe that the inflation of p-values is reduced, but the signals are even more significant around the significant SNP at



(a) Body weight association signals with t-test

(b) Body weight association signals with Mixed Model

(c) Liver weight association signals with t-test

(d) Liver weight association signals with Mixed Model

(e) Saccharin Preference association signals with Mixed Model

Figure 4.3: Genome-wide scans for association with initial body weight, liver weight, and saccharin preference using simple t-test and F test with mixed models, based on a kinship inferred from haplotype similarities

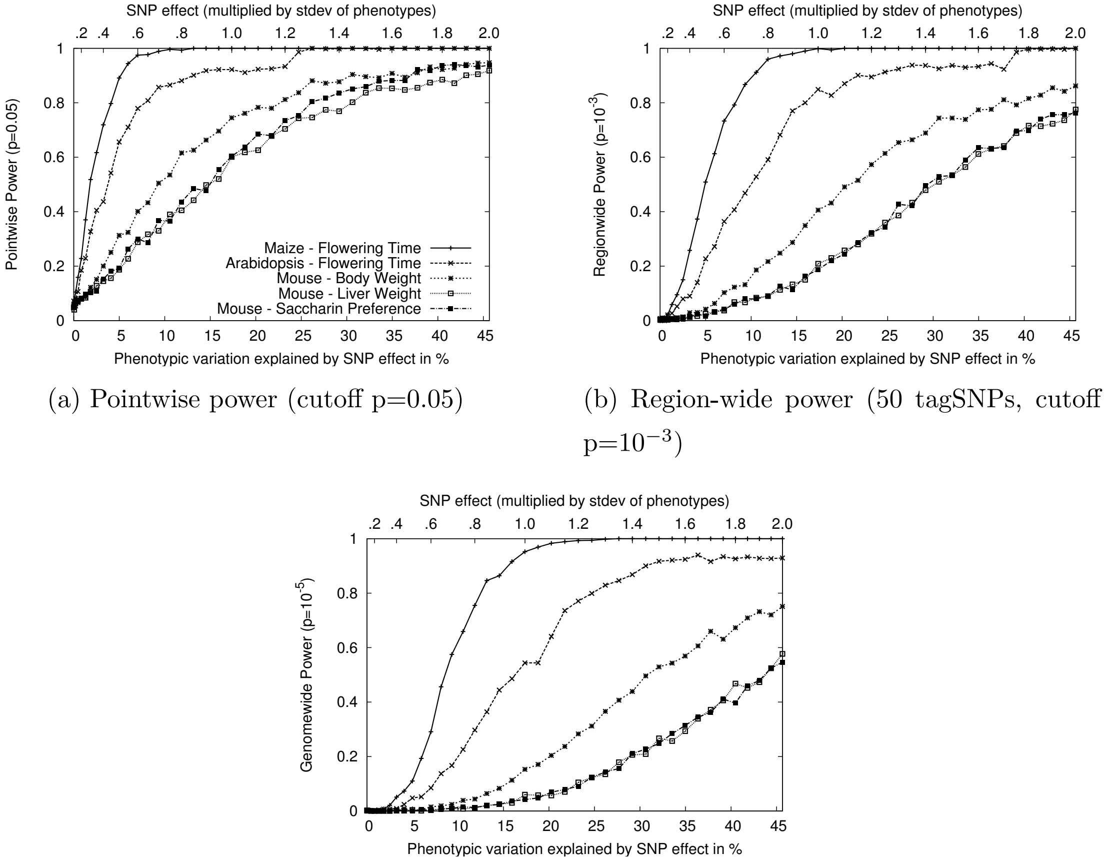
chromosome 2. This demonstrates that mixed model association mapping can, not only reduce the inflated false positives, but also reveal significant associations that have remained unidentified using conventional statistical methods in the case when the associated SNP is not highly correlated with population structure.

For the saccharin preference phenotype, we were able to identify a SNP 30kb away from the *Tas1r3* gene that is perfectly correlated with the SNP previously reported to have significant association with the phenotype [176]. It explains 51% of the genetic variance component, and the p-value is 1.0×10^{-5}. The SNP is neither genome-wide significant nor the most significant. We believe this is due to the limited power of the study with small number of strains.

4.3.3 Power of inbred association mapping

We evaluated the statistical power of association mapping of inbred model organisms in two different ways. First, we simulated an additive effect of causal SNP over the existing phenotypes for mouse, Arabidopsis, and maize strains, similar to previous studies. Such simulation studies evaluate the SNP effect on the power maintaining the existing correlation structure of phenotypes. However, they do not allow to change the effect of the genetic background or the number of multiple measurements, and no random variable other than the SNP is involved in the power simulation. As an alternative model driven method for simulation studies, we generated simulated phenotypes randomly sampled from multivariate normal distribution with various effects of population structure on the phenotypic variation. A SNP effect is simulated on the randomly generated samples, and the statistical power is evaluated. In this way, we can not only change the SNP effect in the simulation studies but also the effect of genetic background on the phenotypes as well as the number of replicated measurements. We believe that our simulation analysis provide a more extensive understanding of the statistical power of association mapping with model organisms based on mixed models.

Figure 4.4 shows the statistical power with respect to the additive SNP effect on the Arabidopsis and maize flowering time phenotypes and three inbred mouse phenotypes used in this chapter. The maize panel dataset consisting of 277 strains



(a) Pointwise power (cutoff p=0.05)

(b) Region-wide power (50 tagSNPs, cutoff p=10^{-3})

(c) Genome-wide power (cutoff p=10^{-5})

Figure 4.4: Comparisons of the genome-wide power of the EMMA method applied to inbred mouse association for real phenotypes with additional SNP effect.

have high statistical power, achieving 80% with SNP effect explaining 5% of phenotypic variation. Genome-wide significance can also be achieved with high power with 10% of SNP effects. For Arabidopsis dataset consisting of 95 strains, the statistical power is decreased, and roughly twice the SNP effect would be needed compared to the maize panales in order to achieve the same statistical power. For the inbred mouse phenotypes, genome-wide power is achievable only when the SNP explains a very large portion of phenotypic variance. In our results, the plausible significant associations explained more than 35% of the phenotypic variance. The power to achieve genome-wide power is largely dependent on the number of available strains. Table 4.2 summarize the most plausible associations in these three phenotypes.

Next, we performed simulation studies by sampling phenotypes from multivariate normal distribution based on the kinship matrix of 48 inbred mouse strains

Table 4.2: List of plausible associations in the mouse association mapping. Comparisons of the statistical power of the EMMA method across three different inbred mouse phenotypes and flowering time of Arabidopsis and maize, by randomly selecting causal SNPs across the genome-wide SNPs. (a) Pointwise power denotes the power to identify causal SNPs at a nominal p-value 0.05 (b) Region-wide power assumes 50 hypothetical tagSNPs in a genomic region. With 20kb between tagSNPs, the genomic region covers up to 1Mb. (c) Genome-wide power is the power to achieve genome-wide significance using the p-value threshold 10^{-5}, which is conservative compared to the permutation based genome-wide significance thresholds using the original phenotypes. The phenotypic variation explained by SNP effect is computed assuming a minor allele frequency (MAF) of 0.3.

Phenotype	Chr	Position	p-value		σ^2 Expl'd (%)		Alleles	MAF	Notes
			F test	LR test	Overall	Genetic			
Body Weight	chr8	113,588,970	3.9×10^{-6}	1.9×10^{-5}	49.0	38.7	A/C	0.27	300kb from the LOD peak of *Bwq3* QTL
Liver Weight	chr2	34,499,435	1.2×10^{-9}	1.4×10^{-7}	39.1	58.6	G/C	0.50	genome-wide significant within *Lvrq1* QTL
Saccharin Preference	chr4	154,883,600	1.0×10^{-5}	7.5×10^{-5}	35.9	50.6	G/A	0.31	30kb from *Tas1r3* gene

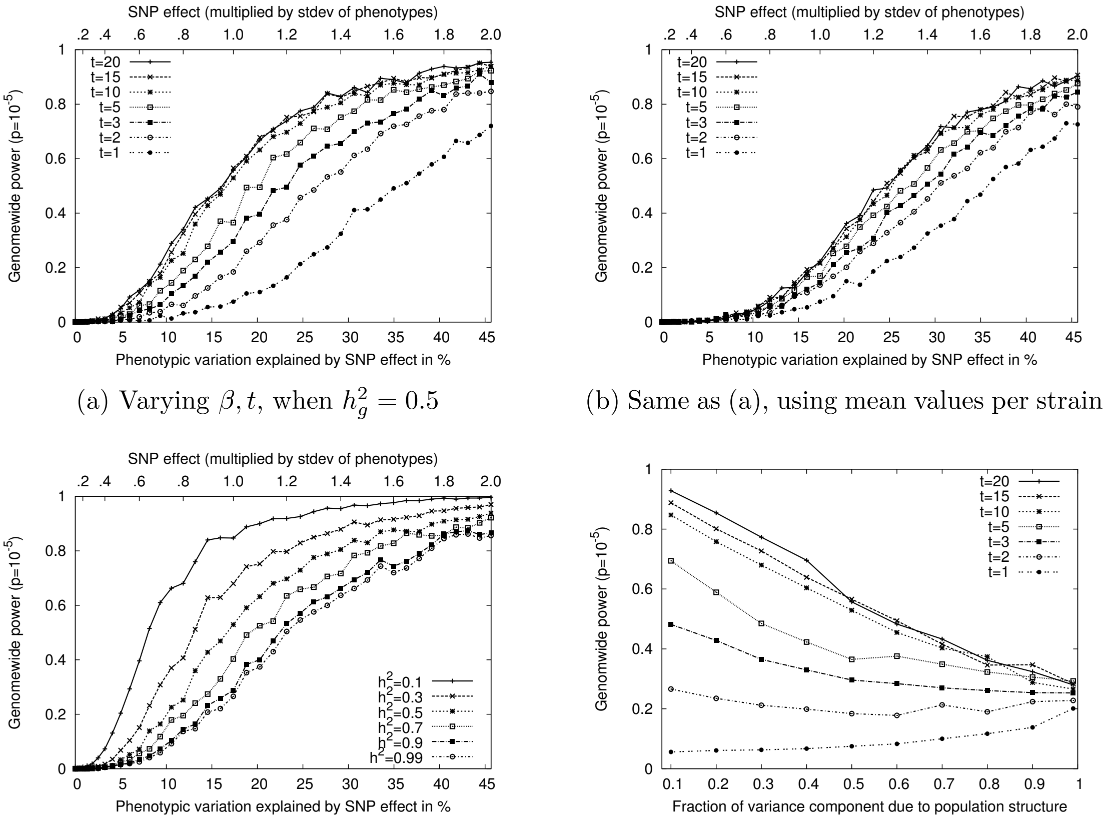
with different effect of genetic background due to population structure. We observed a significant increase of power when multiple measurements are used. Figure 4.5(a) shows the effect of multiple measurements on the statistical power when the variance from the genetic component and the residual component are the same. It suggests that using just a single measurement per strain may result in a significant decrease in power. Even though multiple measurements are used, if only the phenotypic mean is used in the analysis and the individual measurements are not taken into account, the statistical power would decrease significantly. Comparing Figure 4.5(b) with 4.5(a) clearly shows the advantage of using individual measurements over the phenotypic mean in the statistical analysis. It shows that the statistical power may be differ up to by a factor of two between the two methods. Other mixed model association mapping studies use only the mean values in their analysis, not fully utilizing the potential strength of statistical power with individual measurements.

Figure 4.5(c) shows that a large relative effect from genetic background reduces the statistical power. As the genetic background contributes larger portion of phenotypic variance, the within-strain variance becomes relatively smaller than the between-strain variance, and this limits contribution of multiple measurements to the statistical power[17]. For example, in an extreme case, when $h_g^2 = 1$, the residual variance is zero and the replicated measurement does not increase the power since there is no variability of phenotype allowed within the strains.

Figure 4.5(d) shows more clearly the effect of genetic background and multiple measurements at a glance. When a SNP explains a fairly large fraction (17%) of phenotypic variance, the genome-wide significance level can be achieved with high power only when the phenotype has very small effect of population structure and the number of replicates is large enough. As the effect from genetic background becomes larger, the advantage of using multiple measurements decrease significantly.

4.4 Discussion

In this chapter, we proposed an efficient statistical method to perform association mapping with structured samples based on linear mixed model. Our results with maize and Arabidopsis panel show that EMMA robustly reduces the inflated false



(a) Varying β, t, when $h_g^2 = 0.5$

(b) Same as (a), using mean values per strain

(c) Varying β, h_g^2, when $t = 10$

(d) Varying h_g^2, t, when $\beta = \sigma$

Figure 4.5: Comparisons of the genome-wide power of the EMMA method applied to inbred mouse association for simulated phenotypes with various SNP effect, genetic background effect, and the number of multiple measurements. The significance threshold is $p = 10^{-5}$. t is the number of multiple measurements per strain, and h_g^2 is the fraction of the variance explained by genetic background among overall phenotypic variance when SNP effect is not added. (a) With $h_g^2 = 0.5$, varying β and t. (b) Same as (a) using mean phenotype value per strain instead of individual measurements. (c) With 10 multiple measurements per strain, varying β and h_g^2 (d) With $\beta = \sigma$, varying t and h_g^2. The effect of population structure is varied by changing the ratio of two variance components, and number of multiple measurements are simulated with (a) ten measurements and (b) single measurement per strain.

positives under structured population similar to currently available mixed model implementations. The accuracy and stability of the numerical optimization in EMMA is greater than others due to global optimization of likelihood function and guaranteed convergence properties with smaller search space. Our presentation of the EMMA method is focused on a particular case of a mixed model where two variance components are involved because this is the typical model which previous studies assume, and it is straightforward to correct population structure via one kinship matrix inferred from genome-wide markers.

The computational efficiency of EMMA is orders of magnitude greater than other widely used implementations. When multiple measurement per strain is used across different individuals, the relative efficiency is further increased. This is of a great importance when the computational cost may be a bottleneck in the statistical analysis of high-throughput data such as genome-wide gene expressions. For example, the single run of genome-wide association mapping of mouse body weight phenotypes with multiple measurements would take up to a month of CPU time with other implementations, while EMMA takes only a single CPU hour. When hundreds and thousands of phenotypes be available such as in the analysis of eQTL data, the computational cost of previous implementations is prohibitive even with high performance computing. It should be noted that there are other techniques developed for improving computational efficiency of the numerical estimation in a more general context of linear mixed models such as average information REML[72], but these techniques would not provide us with the same improvements on the efficiency of each iterative procedure.

Our results of inbred mouse association mapping show the potential and limitations of genome-wide inbred mouse association studies. It is remarkable that we were able to identify significant associations at a genome-wide level without inflation of false positives, under the limited statistical power of the method. Although there is a possibility that residual confounding still remains with mixed model association, we believe that the most significant SNP associated with liver weight is likely to be a true positive because it explains a large portion of phenotypic variations between the strains beyond genetic background effect so that the conservative Bonferroni adjusted p-value still remains significant. The SNP associated with body weight looks also plausible, but it could possibly due to residual confounding that is not completely captured by a kinship matrix. Likewise, other significant associations can possibly be due to residual confounding not captured by kinship matrix, so the identified associations must be verified through independent analysis.

In a more general context of association mapping which require the use of multiple variance components, the computational advantage of EMMA are not applicable since EMMA can only effectively solve a model with one correlated variance component. For example, when allowing heterozygous alleles for outbred individuals, the full model typically takes both additive and dominant variance components in the linear mixed model[127, 9]. Likewise, if strain-specific environmental random effects or other additional random effects are to be considered, multiple variance components need to be used. In such cases where EMMA is not directly applicable, computational bottlenecks may be the biggest obstacles in analyzing large amounts of data. EMMA can still be applied in this case if a reasonable approximation is combined with other standard mixed model methods taking multiple variance components. Under null hypothesis, it is possible estimate the ratio between multiple variance components using full model, and EMMA can be applied under alternative hypothesis assuming that the ratio between variance components is preserved. Since variance component estimation under null hypothesis needs to be done once across a larger number of alternative hypotheses for each marker, such an approximation procedure provides a large amount of computational efficiency essentially equivalent to EMMA with one variance component. Although the approximated test may lose statistical power slightly, the false positive rates would not be inflated.

There have been several genome-wide association mapping studies with inbred mouse strains. To the best of our knowledge, our results are the first whole genome association mapping of inbred mice that takes the genetic relatedness into account via a statistical method supported by asymptotic theory. Previous studies either do not take the population structure into account[32], or apply heuristics to reduce the confounding effect from population structure. For example, the weighted version of F statistic[166] does not follow the asymptotic null distribution. Redefining the significance level based on empirical null distribution given heritibility parameter[124] or weighted permuation[137] only rescales the p-values similar to Genomic Control, and will suffer from a lack of power as the genetic background effect becomes larger.

Our power simulation studies provide assistance to the design of the association study under the effect of population structure. Multiple factors are involved in determining the the condition of identifying a loci, and it cannot be simply represented by a single value such as phenotypic variance explained by the SNP. Our results show the importance of multiple measurements of phenotypes from multiple animals for each strain, and of directly using the individual measurements in the statistics for association mapping. Taking individual measurements into account within the association mapping is much more computationally intensive EMMA provides a method for efficiently handling individual measurements. In addition, our results also demonstrate the effect of genetic background on the statistical power. As the population structure explains larger phenotypic variance, the power using multiple measurements becomes lower.

Our results show that phylogenetic control can control for population structure as effectively as the linear mixed model based on genetic similarity matrix in some datasets despite the limited ability of the model to represent complex genetic relatedness. Since genetic similarity matrices are better models accounting for recombination and hybridization, and also are easier to compute, phylogenetic control is not preferred in association mapping in model organisms. However, it is possible to compute the likelihood of phylogenetic control model in linear time[58], and this may be useful when a very large number of individuals are to be tested.

Chapter 4 was published in Genetics, Volume 178, pp 1709-23, 2008. 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”. The dissertation author was the primary investigator and author of this paper.

Chapter 5

Accounting for sample structure in large scale genome-wide association studies using a variance component model

5.1 Motivation

Genomewide association studies (GWAS) have played a prominent role in mapping efforts in recent years. They rely on high density genotyping to identify polymorphisms whose alleles are correlated with disease status or quantitative trait levels. While family based designs are possible and in some cases desirable[18], a large portion of current GWAS are conducted using a population sample. Ideally, these would be unrelated individuals that share the same population background; in practicality, however, it is not easy to thoroughly control these variables at the sampling stage. The term "population stratification" is used to indicate the situation where the subjects are unknowingly sampled from different populations. "Hidden relatedness" or "cryptic relatedness" indicates the possibility that some of the individuals in the sample may be related, unknowingly to the investigators and the subjects[211, 221].

The effect of ignoring either of these departures from the ideal design has been well documented[148, 85]. While sampling entirely unrelated individuals may be difficult or impossible, the genotype data provides valuable information on the sample structure that can be used to inform the analysis. For example, Pritchard et al.[170] suggest investigating the presence of sub-populations on the basis of genotype data, and to subsequently carry out association tests within the identified strata. To eliminate the effects of hidden relatedness, instead, one can estimate the proportion of genes identical by descent between any pair of individuals in the sample and exclude from the analysis those that appear strongly related[223]. Conducting association tests within strata can be effective when the sample under study is really composed by a set of separate populations, and the exclusion of a few "closely" related individuals is certainly efficient when these represent a very small percentage of the study subjects. Population stratification and hidden relatedness, however, constitute just two extreme manifestations of what we will call "sample structure." It is then important to develop methodologies that allow one to account for generic departures from the assumption of independent individuals. Devlin and Roeder[50] and Bacanu et al.[12] develop the *genomic control* approach: the distribution of test statistics from the single marker analysis is used to estimate the "inflation factor," λ, with which the test-statistics are subsequently rescaled, constraining the risk of false positives. This is a very practical approach, the authors illustrate its validity under a number of hypothesis, and the literature has uniformly adopted the use of λ to quantify the effects of possible structure on the association tests. Price et al. and Patterson et al.[168, 158] illustrate how principal components of the genotypes can be used to detect and describe sample structure in a continuous fashion. This authors suggest including the main principal components as covariates in association models. The intuition is that these principal components capture high level sample structure such as spatial structure and ancestral population structure. This approach, supported by Novembre and Stephens[152], can be quite effective, even though it is somewhat lacking in interpretability. Thus, today's association studies apply a set of heterogeneous strategies, first identifying close relatives to remove the effects of cryptic relatedness, then applying principal components to correct for large scale sample structure, and finally estimating the residual inflation factors to quantify the remaining effects of sample structure. Since the applied techniques are designed for these extreme manifestation of sample structure, it is not surprising that, even after applying techniques to correct for both hidden relatedness and population structure, a significant amount of residual inflation is likely to remain due to forms of sample structure not captured by the above methods.

We are motivated by the analysis of genomewide association data for quantitative phenotypes in a Finnish cohort[184]. During the past year, approximately 5000 individuals from the 1966 Northern Finnish birth cohort (NFBC66)[175] have been genotyped with the Illumina 370K array and this dataset has been used to investigate the genetic underpinnings of a number of quantitative phenotypes. Sabatti et al.[184] describe in detail the results of genome wide association studies involving 4763 individuals from NFBC66 for triglycerides, cholesterol, glucose, C-reactive protein, insulin plasma levels, body mass index, and blood pressure (systolic and diastolic); this same sample has been used in meta-analysis by Aulchenko et al.[11] for lipid levels, Prokopenko et al.[171] for glucose, Willer[226] for body weight; as well as in candidate gene investigations for temperament traits[153] and height [191]. Indeed, the availability of rich phenotypic information makes the use of this cohort data particularly attractive. Another reason this sample is appealing to geneticists is the fact that it is obtained from a homogeneous population isolate, which is expected to minimize genetic heterogeneity, increasing the chances of mapping genes underlying traits of interest[209]. However, a detailed study[96] of the genomewide high density genotype of a subset of NFBC66, together with other samples from Finland and Sweden, revealed the presence of substantial population structure that could influence the results of association studies. The genomewide association studies that have used NFBC66 all adopt some methodology to correct for population structure, but to date we are still lacking an extensive analysis of the actual impact of the detected population structure in NFBC66 on association results and a comparison of the effectiveness of different methodologies to correct for the noted structure.

Using a newly obtained larger sample (5337 individuals), and the additional phenotype of height, we here explore these questions in detail. To account for sample structure that can reflect both the presence of somewhat distinct sub-populations as well as hidden relatedness across individuals, we explore a methodology that relies on the use of the polygenic model[61] and its application to association mapping[155].

Unlike this classical setting, however, we do not assume the degree of relatedness among the individuals in the sample to be known a priori, but we roughly estimate it from the genotype data. A similar approach has been used successfully in animal models[235, 240, 103]. Capitalizing on the characteristics of quantitative traits in humans, we make a few simplifying assumptions that allow us to dramatically increase the speed of computations, making our approach readily applicable to genomewide association studies for quantitative traits with tens of thousands of samples. In addition, we extend our method to case-control studies and demonstrate its robustness using the Wellcome Trust Case Control Consortium (WTCCC) data over seven common complex diseases.

5.2 Materials and methods

5.2.1 Variance component model to account for sample structure

In his polygenic model, Fisher[61] describes a trait T as the result of an environmental component and a large number of different genetic factors, each with possibly an additive and dominant effect. He also considers the possibility of interactions between these factors (epistasis). We consider here the simplest form of this model where genetic factors act independently and additively. Let Z_{ik} be the contribution of factor k to person i, then we assume that the phenotype can be modeled as

$$y_i = \sum_{k=1}^{M} Z_{ik} + \epsilon_i, \qquad \mathrm{E}(\epsilon_i) = 0, \qquad \mathrm{Cov}(\epsilon_i, \epsilon_j) = 0 \text{ if } i \neq j \tag{5.1}$$

Note that in model (5.1), and throughout the chapter, there are no other relevant variables besides the genetic factors. This is purely a convenience assumption, to simplify notation.

With model (5.1) Fisher was able to explain the correlation between relatives and to estimate the proportion of the total trait variance due to genetic factors (heritability). Let the vector $Y = \{y_i, \ldots, y_n\}$ contain the phenotypes of the individuals in a pedigree; then assuming that the environmental components are uncorrelated, the variance covariance structure of Y depends on the amount of genes shared among subjects. In absence of dominance effects, we have

$$\mathrm{Var}(Y) = 2\sigma_a^2\Phi + \sigma_e^2 I, \tag{5.2}$$

where Φ is the matrix of kinship coefficients between each pair of individuals in the pedigree. An analysis of variance with random effects leads to the estimates of σ_a^2 and σ_e^2, and in turn to the evaluation of heritability $h^2 = \sigma_a^2/(\sigma_a^2 + \sigma_e^2)$.

In linkage studies, this decomposition of variance is carried one step further. By tracking the transmission of marker genes in the vicinity of locus k, one can calculate the conditional kinship coefficients (Φ_k, probabilities that two genes sampled from two individuals at locus k are IBD), and decompose the variance $\mathrm{Var}(Y)$ to emphasize the contribution of the k-th locus

$$\mathrm{Var}(Y) = 2\sigma_{ak}^2\Phi_k + 2\Phi\sigma_a^2 + \sigma_e^2 I. \tag{5.3}$$

To investigate the contribution of locus k to the phenotype, one tests the null hypothesis that $\sigma_{ak}^2 = 0$. The values of the variance parameters are estimated with maximum likelihood procedures[115].

In association studies, based on a much denser set of genotypes, we aim to associate the phenotypes directly to the alleles at marker loci; in other words we are aiming to estimate fixed effects. Assuming additive effects only, model (5.1) can be translated to the following regression framework:

$$y_i = \beta_0 + \sum_{k=1}^{M} \beta_k X_{ik} + \epsilon_i, \tag{5.4}$$

with $\mathrm{Var}(\epsilon) = \sigma_e^2 I$, and X_k the individuals' minor allele counts at locus k (for all markers are considered biallelic). Let the $n \times p$ matrix X contain allele counts on a $p << M$ set of markers, and let β a $p \times 1$ vector of corresponding coefficients. Our goal is to identify which elements in the vector β are different from 0.

While model (5.4) is fundamentally a multivariate one, association studies are typically carried out by testing the hypothesis $H_0 : \beta_k = 0$, for each of the M loci, one locus at the time, on the basis of model

$$y_i = \beta_0 + \beta_k X_{ik} + \eta_i. \tag{5.5}$$

With respect to model (5.4), model (5.5) is misspecified: relevant regressors are omitted, or, in other words, we ignore the polygenic background of the trait.

The appropriate statistical methods to make inference on β_k in (5.5) depends on the nature of the sample. If the n individuals are related, with known degree of relatedness, the variance covariance of η_i in model (5.5) can be described as in (5.2). That is, the effect of the genotype at locus k can be modeled as a main effect, while the relationships among all individuals are taken into account by means of variance components of random polygenic effects[155]. This model is sometimes referred to as a "mixed effect" model.

If the n individuals are unrelated and there is no dependence across the genotypes, so that the η_i are i.i.d. (independently and identically distributed), a simple linear regression would make appropriate inference. Unfortunately, these conditions are not easily met. Firstly, because of linkage disequilibrium, X_k corresponding to markers with close genomic position are correlated. More significantly, neither the homogeneity of population background nor the level of relatedness are easily controlled in the sampling stage. If the n individuals in the sample belong to distinct populations, or are (albeit distantly) related, one can expect a substantial correlation between the rows and columns of X. This translates to bias in the estimate of β_k from model (5.5) and in a distribution for $\hat{\beta}_k$ that is different from what is assumed in standard linear regression (i.e. the η_i in (5.5) are not i.i.d.). [Note that it is reasonable to assume that environmental exposure would vary across populations and family group, introducing further confounding]. In this context, the suggestion of Price et al.[168] of including principal components of the genotypes in the model can be understood as an attempt to include in (5.5) proxies of the missing variables. Similarly, approaches which remove closely related individuals can be understood as an attempt to bring the η_i closer towards i.i.d.

With the advent of dense, genomewide genotype data, it has become possible to estimate the degree of relationship between independently ascertained subjects[127, 55, 203] in the absence of genealogical information. With an estimated kinship matrix one can in principle use variance component techniques in linear mixed models (as in Ober et al.[155]) to analyze population samples. This approach has been indeed successfully adopted in Yu et al.[235] and Zhao et al.[240] in the analysis of small structured samples of model organisms. Its application to the analysis of human association samples has been hindered by the increased computational costs associated with large sample sizes and number of genotypes. Kang et al.[103] propose a more efficient variance component estimation procedure that allows the authors to analyze a mouse dataset including hundreds of thousands of SNPs. Unfortunately, its straightforward application to human studies that include thousands of individuals is still computationally too costly.

The experience accumulated in the first couple of years of genomewide association studies suggest that some shortcuts may be possible. Human quantitative trait association studies so far, appear to fully vindicate Fisher's polygenic model, with each of the loci involved responsible for a very small portion of the variance[128]; if the contribution of each SNP to the total trait variance is almost ignorable, the variance components for η_i in (5.5) may not need to be estimated separately for each SNP. Instead, one might estimate the values σ_a^2 and σ_e^2 from a variance decomposition model as in (5.2), keep them fixed, and then estimate the parameter β_k in model (5.5) using a generalized least squares (GLS) procedure. Additionally, it appears that using the simple identity by state (IBS) between individual, rather then the more laboriously constructed kinship coefficients, may be sufficient, and in some cases more appropriate, to model the dependency in the sample. Zhao et al.[240] observe this phenomena in *Arabidopsis* and suggest that while IBD is preferable to describe recent relatedness, IBS may be more apt to describe very distant relationships between individuals, that indeed blend into population level differences. Along these lines, Kang et. al[103] precisely reflect the polygenic background under the assumption that each SNP is equally likely to contribute to the quantitative trait at a very small level.

On the basis of these observations, we employed the following procedure to analyze human population samples in association studies for quantitative traits. Let n be the sample size, p the total number of genotyped SNPs, and Y the vector of observed phenotypes

1. Use the genotype data to calculate the $n \times n$ matrix $\hat{S}$ of identity by state between individuals, and normalize $\hat{S}$ to have sample variance 1 using a Gower's centered matrix[136].

$$\hat{S}_N = \frac{\hat{S}}{\mathrm{Tr}(P\hat{S}P)} \tag{5.6}$$

where $P = I - \mathbf{1}\mathbf{1}'/n$ and $\mathbf{1}$ is vector of ones.

2. Use a variance component model to estimate the restricted maximum likelihood (REML) parameters (or alternatively, maximum likelihood parameter) of σ_a and σ_e in:

$$\mathrm{Var}(Y) \;=\; \sigma_a^2\hat{S}_N + \sigma_e^2 I \tag{5.7}$$

Test the hypothesis $H_0 : \sigma_a^2 = 0$. If the null hypothesis is rejected, proceed to step 3; otherwise use ordinary least square (OLS).

3. Use GLS to estimate the effects β_k

$$y_i = \beta_0 + \beta_k X_{ik} + \eta_i \qquad \mathrm{Var}(\eta) = V \propto \hat{\sigma}_a^2\hat{S}_N + \hat{\sigma}_e^2 I. \tag{5.8}$$

The above model can be easily extended to have additional confounding variables by substituting β_0 for a multi-column matrix containing the confounding variables such as sex and age. Note that these additional confounding variables should be included in the procedure of REML estimation of the variance component parameters. For the variance component estimation procedure in step 2, we use EMMA (Efficient Mixed Model Association)[103]. We term our method as EMMAX (EMMA eXpedited), because our method dramatically reduces the computational cost compared to the original EMMA by avoiding the repetitive variance component estimation procedure per marker.

5.2.2 Estimating marker specific inflation factor

Devlin and Roeder[50] showed that, under some condition, one can expect the variance of the test statistics to be inflated by a constant across the genome. Bacanu et al.[12] extended these results to quantitative trait association mapping. Voight and Pritchard[211] developed a formal model of cryptic relatedness based on the coalescent theory. Typically, one expects a constant inflation across the genome when the sample structure is entirely due to cryptic relatedness. However, with a more complex genealogical relationship among individuals, it is not clear how the inflation of test statistics will behave. For example, each SNP may have different level of inflation due to differential effects from ancestry or age differences among the SNPs[129, 177, 14]. Indeed, alternative approaches to account for population structure such as Structured Association[170] or principal component analysis[168], lead to different correction for each marker. Model (5.8) gives us the opportunity to quantify the amount of inflation due to population structure on each specific marker, allowing us to shed some light on the discussion above.

Assuming that model (5.5) is true with $V = \mathrm{Var}(\eta)$ and marker k has no effect on the phenotype, we define the inflation factor for marker k as the ratio between the expectation of the F statistics calculated from OLS for a model that includes k, to the expectation of the F statistics for the same model calculated from GLS. In fact, we do not compute this ratio explicitly, but simply provide an approximation. If one considers that as $n \longrightarrow \infty$, the expectation of the GLS F statistics under arbitrary V, as long as V is non singular, converges to 1; hence we simply need an approximation for the numerator of the ratio.

Specifically, let us assume, to simplify notation, that Y and X_k are centered to have zero sample mean so that $\hat{\beta}_0 = 0$ holds. In such a case, $V = \mathrm{Var}(\eta)$ has to be centered to $V_C = PVP$ where $P = I - \mathbf{1}\mathbf{1}'/n$. In addition, for convenience purposes, we standardize X_k to satisfy $X_k^T X_k = n - 1$, where n is the number of individuals. Then the F-test statistic based on OLS becomes

$$F_{OLS} = \frac{((X_k'X_k)^{-1}X_k'Y)^2(X_k'X_k)(n-2)}{Y'(I - X_k(X_k'X_k)^{-1}X_k')Y} \tag{5.9}$$

$$= \frac{(X_k'Y)^2(n-2)}{nY'Y - (X_k'Y)^2}. \tag{5.10}$$

If $V = \sigma^2 I$, then F_{OLS} follows a F-distribution with (1, $n-2$) degree of freedom. Then if n is large, F_{OLS} asymptotically converges to chi-square distribution with 1 degree of freedom. While the distribution of F_{OLS} is difficult to calculate when V has off-diagonal elements, the expected values of numerator and denominator in F_{OLS} are relatively easy to compute. The expectation of denominator becomes $n\mathrm{Tr}(V_C) - X_k'V_CX_k$, and the expectation of numerator becomes $(n-2)X_k'V_CX_k$.

We can then take as operational definition of the marker specific inflation factor ζ_k at marker k,

$$\zeta_k = \frac{(n-2)X_k' V_C X_k}{(n-1)\mathrm{Tr}(V_C) - (X_k' V_C X_k)} \tag{5.11}$$

$$\approx \frac{X_k' V_C X_k}{\mathrm{Tr}(V_C)} \tag{5.12}$$

Note that when $V = \sigma^2 I$, then $\zeta_k = 1$ holds regardless of the values of X_k. Let $\hat{S}_C = P\hat{S}_N P$. When we take for V the specific form assumed in (5.8), we can further simplify the expression above:

$$\begin{aligned}
\zeta_k &= \frac{(n-2)X_k'(\sigma_a^2 \hat{S}_C + \sigma_e^2 P)X_k}{(n-1)\mathrm{Tr}(\sigma_a^2 \hat{S}_C + \sigma_e^2 P) - (X_k'(\sigma_a^2 \hat{S}_C + \sigma_e^2 P)X_k)} \\
&= \frac{\sigma_a^2 (n-1) X_k' \hat{S}_C X_k + \sigma_e^2 (n-1)(n-2)}{\sigma_a^2 \left[(n-1)^2 - X_k' \hat{S}_C X_k\right] + \sigma_e^2 (n-1)(n-2)} \\
&\approx \frac{\sigma_a^2 X_k' \hat{S}_C X_k/(n-1) + \sigma_e^2}{\sigma_a^2 + \sigma_e^2} \\
&= h_a^2 X_k' \hat{S}_C X_k/(n-1) + (1 - h_a^2) \qquad (5.13)
\end{aligned}$$

where $h_a^2 = \sigma_a^2/(\sigma_a^2 + \sigma_e^2)$.

We are now in the position to discuss the meaning and implication of the marker specific inflation factors we defined. The introduced marker-specific inflation factors essentially estimate the effect of misspecification of variance component by using OLS in the place of GLS. From expression (5.13) it is clear that the amount of inflation at any given marker depends on the level of correlation between the marker genotypes and the GLS variance-covariance matrix. This validates the common intuition that cryptic population structure may affect differently tests at different markers. It also rationalizes the the direction of this variability. Expression (5.13) also clarifies how the same level of sample structure will affect differently the association tests for different phenotypes. The inflation will be stronger the higher is the ratio of σ_a^2 to σ_e^2, while for a trait that does not follow the polygenic model $\sigma_a^2 = 0$, no amount of population structure will have any impact on the association tests. Finally, it is useful to recall that the inflation factors ζ_k, while marker specific, are calculated independently of the observed association between marker and phenotype, being based on expectations of test statistics under the null model.

More generally, if multiple confounding variables need to be accounted for in addition the to intercept under the null model, Equation (5.10) can be rewritten in a general form of F statistic to get the expectation of numerator and denominator. Such a procedure is asymptotically equivalent to centering an arbitrary variance component V to $V_C = (I - G(G'G)^{-1}G)V(I - G(G'G)^{-1}G)$, given a non-singular matrix of confounding variables G that includes the intercept. In this case, the SNP vector X_k also needs to be regressed out with respected to G, and $(n-2)$ in Equation (5.10) needs to be replaced with $(n-q-1)$, where q is the number of columns in G.

This method can also be extended for estimating the effect of misspecified variance component or errors in the variance component estimation. Before running GLS, let $U = \hat{\sigma_a}^2 \hat{S}_N + \hat{\sigma_e}^2 I$ be the estimated variance component when V is the true variance component. Assuming that Y and X_k are centered, the F test statistics for GLS is

$$\begin{aligned} F_{GLS} &= \frac{((X_k'U_C^{-1}X_k)^{-1}X_k'U_C^{-1}Y)^2(X_k'U_C^{-1}X_k)(n-2)}{Y'(U_C^{-1} - U_C^{-1}X_k(X_k'U_C^{-1}X_k)^{-1}X_k'U_C^{-1})Y} && (5.14) \\ &= \frac{(X_k'U_C^{-1}Y)^2(n-2)}{(X_k'U_C^{-1}X_k)Y'U_C^{-1}Y - (X_k'U_C^{-1}Y)^2} && (5.15) \end{aligned}$$

The ratio between expected numerator and denominator provides the inflation factor with misspecified variance component.

$$\begin{aligned} \zeta_k &= \frac{X_k'U_C^{-1}V_CU_C^{-1}X_k(n-2)}{(X_kU_C^{-1}X_k)\mathrm{Tr}(U_C^{-1}V_C) - X_k'U_C^{-1}V_CU_C^{-1}X_k} && (5.16) \\ &\approx \frac{(n-1)X_k'U_C^{-1}V_CU_C^{-1}X_k}{(X_kU_C^{-1}X_k)\mathrm{Tr}(U_C^{-1}V_C)} && (5.17) \end{aligned}$$

5.2.3 Accounting for large effect sizes at some SNPs

The accuracy of EMMAX depends on the assumption that the effect of each SNP on the phenotype is so small to be ignorable for the purpose of estimating σ_a^2 and σ_e^2 in model (5.8). When few SNPs have relatively large effects, this assumption is ungrounded, and the strategy described so far impractical. This is the case, for example, of several common diseases that are known to be associated with certain HLA alleles with relative risks of 4 or greater[44].

In fact, it is possible to use EMMAX even in this context, provided that one prior conditions on the effects of the known associated SNPs. Specifically, one can regress out the effects of the implicated SNPs from the original phenotype and use the residuals of this regression as a new phenotype for analysis. It is crucial, then, to decide on the effect of which SNPs one should condition upon.

If we know *a priori* the identity of associated loci with strong effect, such as the MHC region in the above example, the choice will be obvious. Alternatively, we may condition on the effects of SNPs with highly significant p-values. It is important to use a very stringent significance threshold at this level, to avoid loss of power. Recommended values are $p < 10^{-10}$ or $p < 10^{-15}$ for typical sample sizes ($2000 \sim 10000$). Note that this conditioning procedure is really recommended only if (1) there are a few genomic regions or sets of loci which largely explain the phenotypic variance, and (2) significant over-dispersion or under-dispersion of test statistics is observed after applying EMMAX.

5.2.4 Application to case control datasets

Although EMMAX was developed with quantitative traits in mind, it can also be adapted to the analysis of case-control datasets. Since the case-control phenotypes do not follow a normal distribution, applying a generalized linear mixed model using logit or probit link function is preferable to a linear mixed model. However, the computational cost of a generalized linear model with a correlated variance component is much higher, and currently available algorithms can not handle thousands of individuals simultaneously[138].

When the hypothesis of additive model appears reasonable (as it is the case when we do not have prior information on mode of inheritance), the Armitage trend test (Armitage, 2005) can be used to test for the presence of a genetic effect. (See, for example, Devlin and Roeder[50] and note the equivalence of an Armitage test to a score test in logistic regression for $H_0 : \beta = 0$[1]). The Armitage test can be described as testing the significance of the slope coefficient in a linear regression of a 0-1 variable representing case/control status on the additively coded genotypes. Armitage[10] suggested using a χ^2_1 test that is slightly different from the square of a standard t test in linear regression. The statistic proposed by Armitage is $\chi_0^2 = \beta^2/var(\beta)$, but instead of estimating the variance of the error terms using the residuals from the regression, we estimate it using the variance of the response variable. Therefore χ_0^2 is equal to the square of the correlation between the response and the genotype variables, multiplied by the number of samples.

Despite this suggestion, Armitage indicated that the standard t statistic may be preferable, especially to construct confidence intervals. Therefore, it seems that one can carry out tests in the spirit of Armitage simply using a standard linear regression framework with a 0-1 quantitative response variable representing the case control status. Adopting this approach, we can immediately translate the problem to the methodology suggested for quantitative traits.

5.2.5 Genotype and phenotype data

We analyzed two datasets: one that contains measurements on quantitative traits (NFBC66), and one on disease statuses (WTCCC).

Genotype data were available for 5,544 Finnish subjects from NFBC66, all with genotyping completeness >95%. Subjects were excluded from further analysis because they had withdrawn consent (15), had discrepancy between reported sex and sex determined from the X chromosome (13), were sample duplications (2), were too related to another subject (66), or had no phenotype data, leaving 5,337 subjects for analysis. For the relatedness criterion, all pairs of subjects with probability of IBD $> 20\%$ were identified, and one subject from each such pair was included in further analyses. In most cases, the subject with the most non-missing phenotype data was chosen for analysis. If the two subjects had an equal amount of missing phenotype data, the subject with the most non-missing genotype data was used.

Using these 5,337 subjects, the 339,017 SNP markers were examined for Hardy Weinberg Equilibrium (HWE, exact test), genotyping completeness, and minor allele frequency (MAF). Markers were excluded for HWE p-value< 0.0001 (3932), genotyping completeness $< 95\%$ (67) and MAF $< 1\%$ (7667), leaving 328401 markers for analysis (some SNP markers failed quality checks on more than one criterion).

We focused on the analysis of 10 phenotypes: triglycerides (TG), low density lipoprotein (LDL), high density lipoprotein (HDL), glucose (GLU), C-reactive protein (CRP), insulin plasma levels (INS), body mass index (BMI), systolic (SBP) and diastolic (DBP) blood pressure, and height. The first 9 phenotypes were adjusted for sex, pregnancy status and use of oral contraceptive, as described in Sabatti et al.[184]. Height was adjusted for sex only.

The NFBC66 database contains information on the birth locations of subjects and their parents. This can be used to derive ancestry information. Sabatti et al.[184] describe how 6 distinct linguistic/geographical groups can be identified in the Northern provinces of Finland. Given the patterns of internal migrations and their variation over time, we can assign individuals in NFBC66 to one one these groups when both parents were born in a municipality within the same group. Approximately 50% of the sample can be assigned this way and these individuals are used to compare the results of population stratification analysis based on genotypes.

We also obtained the genotypes of the Wellcome Trust Case Control Consortium (WTCCC) subjects collected for the study of seven common disease[223]. We applied the same quality control criteria as suggested in the original paper. We also excluded the SNPs that the original studies excluded in their analysis. A total of 404,862 SNPs were considered after the quality control across 2,938 shared controls and 13,241 case individuals across seven diseases.

5.3 Results

5.3.1 NFBC66

Revisiting principal component analysis

To investigate the nature of the possible structure in the sample, we used PCA analysis of the genotype matrix and analysis of the IBS matrix from NFBC66 samples[168]. In Sabatti et al.[184] it was shown how the first two coordinates identified by multidimensional scaling analysis of the IBS matrix correlate well with geographical location of the linguistic groups. The first two principal components in the current sample correlate well with latitude and longitude of parental birth places for the subset of individuals with known ancestry (Figure 5.1). Indeed, we

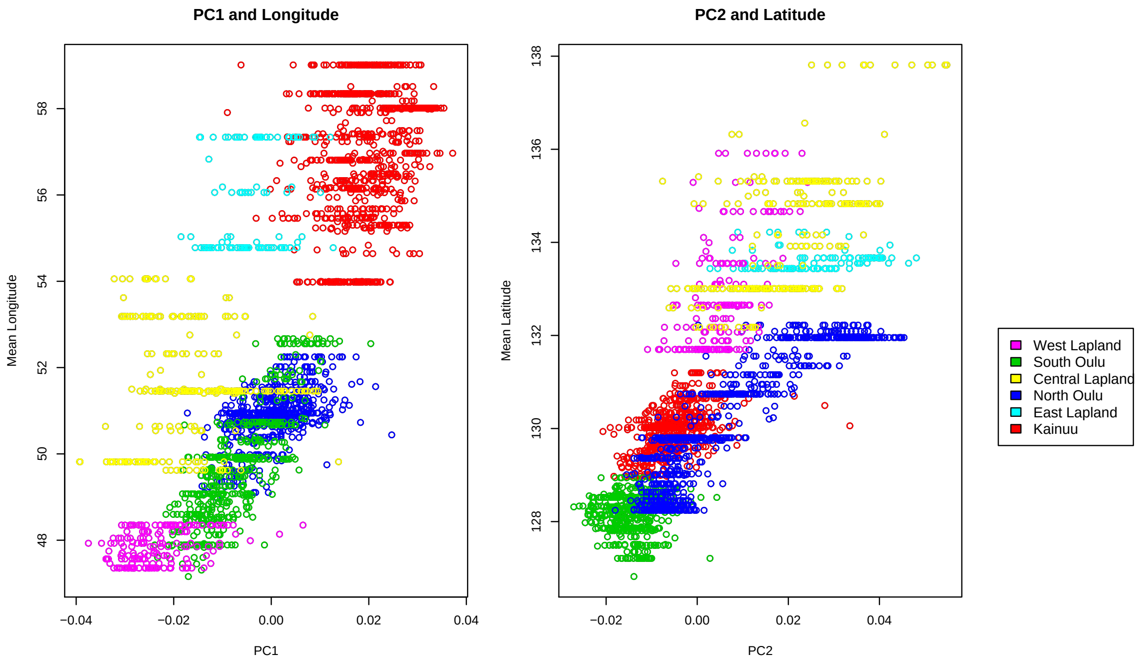


Figure 5.1: Scatter plots of loading for the first two principal components vs Latitude and Longitude. Only individuals of known ancestry are included in the plot. Latitude and longitude are defined as the average latitude and longitude of the parents birthplaces. Different colors indicate different linguistic/geographic subgroups.

noted that PCA analysis of genotypes and classical MDS of the IBS matrix lead to very similar results. The first 5 principal components separate to varying degree the linguistic/geographic groups (Figure 5.2).

It is also of interest to investigate how the phenotype varies across the linguistic/geographic groups. Four phenotypes show significant variability across the different groups (LDL, HDL, GLU, SBP). After including the first two principal components, there is no significant variability in LDL and SBP across the different groups, GLU does not show significant variability after accounting for multiple comparison, while HDL is still significantly variable (in both cases there is only one group that appears different, Eastern Lapland, comprising 90 individuals in total). Despite the clear correlation between geographical region of origin and the first two principal components (or principal coordinates of the IBS matrix), clustering analysis of the IBS matrix failed to identify separate subgroups. This suggests that an analysis of the data using the structured association approach may not be appropriate.

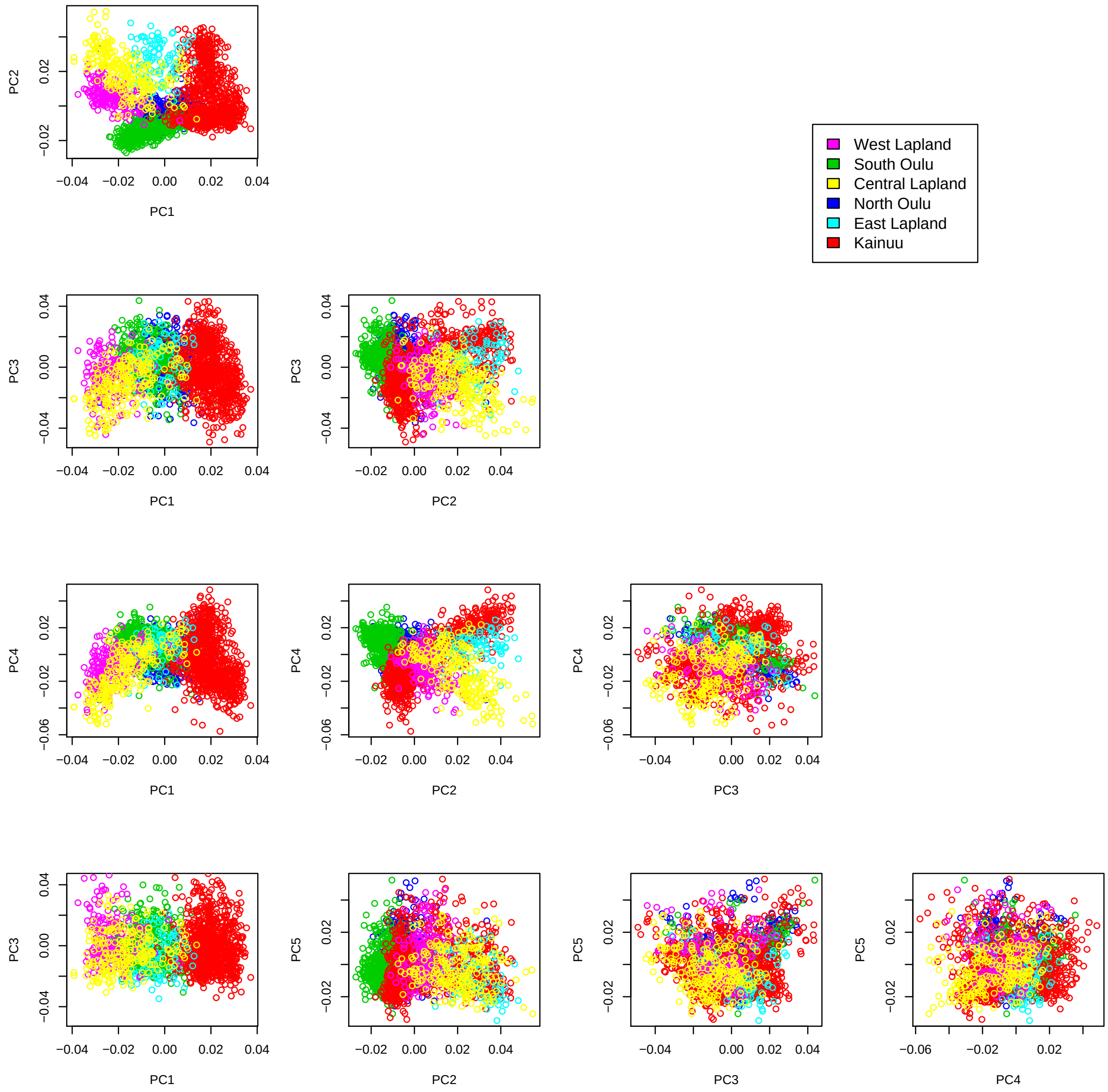


Figure 5.2: Scatterplot of the first 5 principal components for individuals of known ancestry. The different linguistic/geographic subgroups are color-coded.

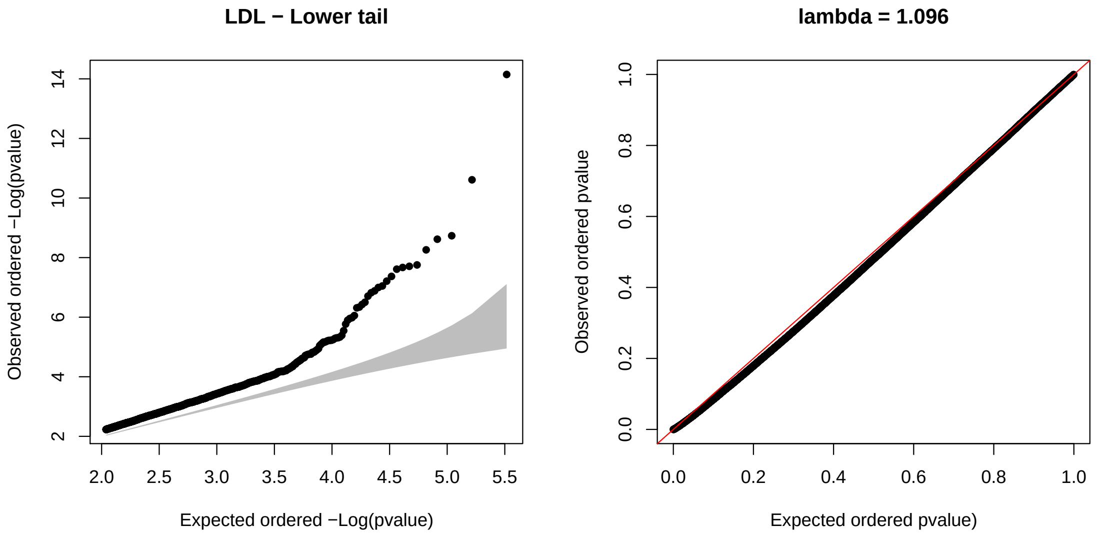


Figure 5.3: QQ plot of p-values from the association tests for LDL. On the left hand side, we focus on the 3000 most significant p-values and to increase readability, we plot -Log10(p-values). The shaded area represent pointwise 95% confidence intervals calculated assuming independence across tests and relying on a Beta approximation of the order statistics. On the right hand side, we present the p-value distribution in its original scale.

Association analysis and attempts to correct for population structure

When performing a simple uncorrected association test for each of the 9 phenotypes originally examined in [184], the following estimates of the genomic control parameters λ were observed: BMI 1.036, CRP 1.012, DBP 1.033, GLU 1.045, HDL 1.054, INS 1.026, LDL 1.093, SBP 1.063, TG 1.024. These values are all higher than the ones obtained with the smaller sample size in Sabatti et al.[184], and higher then what one would expect in a sample with no structure in our population. The additional phenotype of height, led to the highest λ value, 1.19. For reference note that conservative estimate of the 95% confidence interval of the inflation factor is between 0.992 and 1.008, assuming independence between the markers. Figure 5.3 presents the QQ plot of the p-values for association tests with LDL, the trait which exhibits the highest λ amongst our original traits.

As hidden relatedness is a possible cause of inflated genomic control parameters, we re-ran the analysis, after excluding a larger number of possibly related

Table 5.1: Comparison of genomic control inflation factor obtained with different models; ES stands for EIGENSTRAT.

Phenotypes	**uncorrected**	**IBD > 0.1 excluded**	**ES (100 PCs)**	**EMMAX**
BMI	1.036	1.028	1.024	1.001
CRP	1.012	1.020	1.020	0.994
DBP	1.033	1.025	1.029	1.010
GLU	1.045	1.025	1.030	1.009
HDL	1.054	1.041	1.037	1.003
INS	1.026	1.026	1.015	1.005
LDL	1.093	1.089	1.040	1.002
SBP	1.063	1.054	1.021	1.004
TG	1.024	1.021	1.018	0.999
HEIGHT	1.193	1.152	1.080	1.002

subjects (genomewide percentage of alleles identical by descent $> 10\%$ was taken as a cut-off, and 709 additional individuals were excluded). This resulted in a slight reduction of λ for some phenotypes (Table 5.1).

Following the suggestion of Price et al.[168], we explored the effect of including a variable number of principal components in the association tests. Including 2 or 5 PCA has a considerable effect on the λ values, however, augmenting the number of principal components beyond this point, does not result in a substantial decrease in the genomic control parameter (Figure 5.4). It is often suggested that only principal components having predictive power on the phenotype should be included in the regression. Table 5.2 reports the principal components, for each phenotype, that have a t-test p-value <0.005 as predictors for each of the phenotypes and the results of their inclusion in the association tests are reported in Figure 5.4.

Analysis with EMMAX

Unlike a traditional variance component model which uses IBD coefficients estimated from the pedigree[155], our suggested method uses an IBS matrix to capture the relatedness between the individuals. Yu et al.[235] suggested estimating IBD coefficients from multi-locus genotypes, but Zhao et al. and Kang et al.[240, 103] demonstrated that a simple IBS matrix more robustly corrects for the over-dispersion than the estimated IBD matrix from structured model organism samples. However,

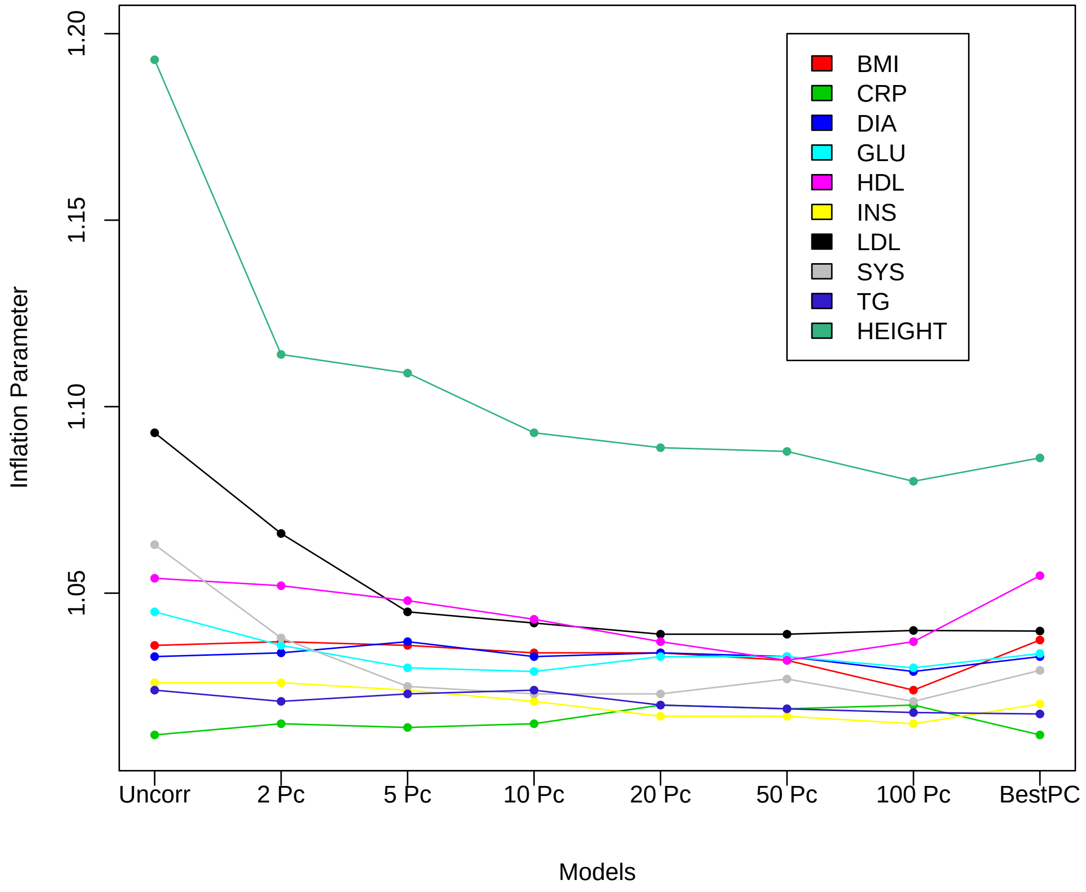


Figure 5.4: Illustration of how the genomic control parameters for 10 traits change as the number of principal components used for adjustment changes.

Table 5.2: Principal components that result associated with each of the considered phenotypes

Phenotype	**Significant PC**
BMI	PC23, PC50, PC83
CRP	none
DBP	none
GLU	PC1, PC23
HDL	PC10, PC39, PC41
INS	PC9, PC11
LDL	PC2, PC3, PC4, PC5
SPB	PC1, PC4
TG	PC17, PC87
HEIGHT	PC1, PC2, PC4,PC6, PC7, PC24, PC25, PC50, PC67, PC80, PC85

the effectiveness of IBS matrix has not been comprehensively examined in a large-scale human association mapping studies, where the genetic diversity among the samples is significantly smaller than those among the strains of model organisms. For this reason we present our results in detail.

We first compared the IBS coefficients with the estimated IBD coefficients obtained by PLINK[172, 142] across genome-wide markers of NFBC66 subjects. The pairwise plot of these two estimates shown in Figure 5.5A suggest that these two estimates are highly correlated when the IBD estimates are positive. However, when IBD estimates are zero, as is the case for 64% of the pairs of individuals, the IBS estimates still show a considerable amount of variation. In fact, the variance of IBS estimates amongst the individuals with zero IBD estimates is even larger than those with non-zero IBD estimates (Figure 5.5B). It is possible that the variability in the IBS estimates among pairs of individual with IBD=0 is simply due to random noise, and the IBD estimates better reflect the true underlying sample structure. To explore the validity of this hypothesis, we partitioned the genome-wide markers into two disjoint sets and we re-estimated IBS between the pairs of individuals with IBD=0, using each of these sets. These IBS estimates are highly correlated (Figure 5.5C,D), which suggests that the IBS values are not due to random noise, but reflect true underlying differences in genetic similarities.

While we relied on PLINK to estimate IBD, we would like to point out that

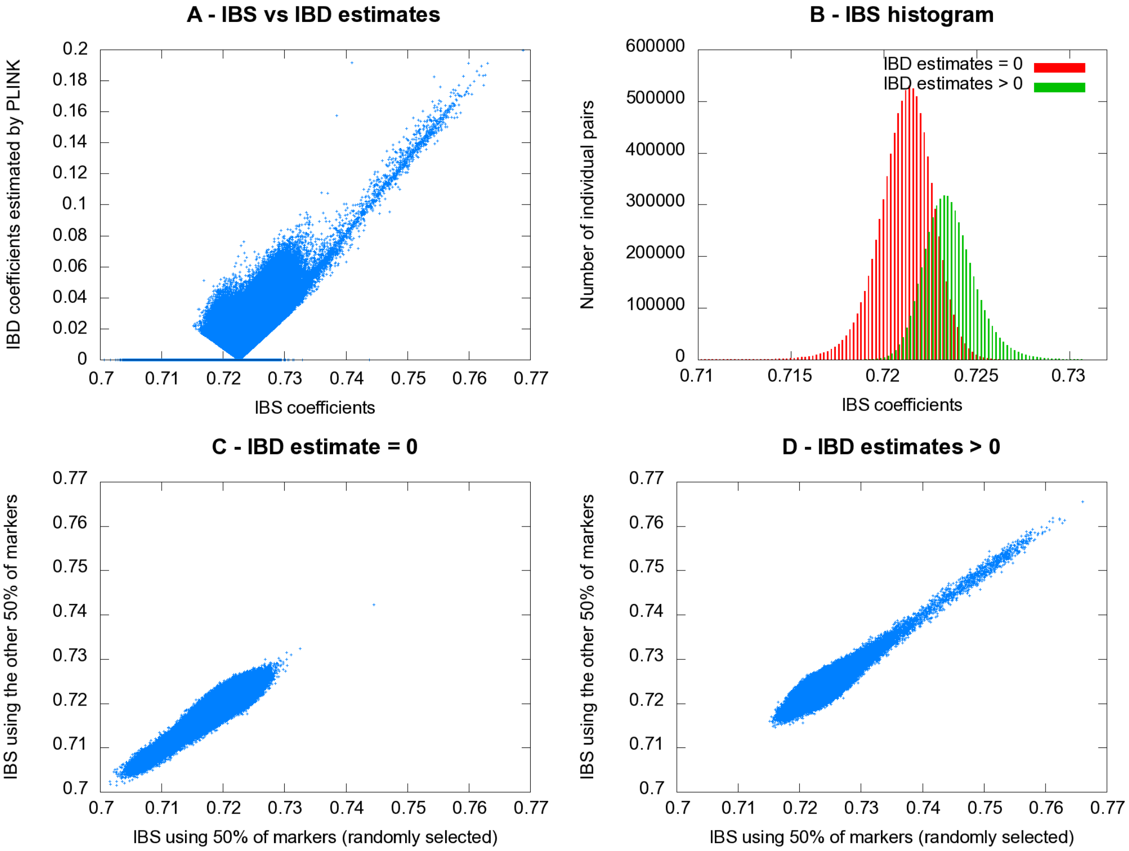


Figure 5.5: Comparisons between (A) the IBS coefficients and IBD estimates computed by PLINK (B) The distribution of IBS coefficients in (A) when the IBD estimates are zero, or positive (C) Two different IBS coefficient by randomly partitioning the SNPs into two different sets when IBS estimates are zero (D) and when IBS estimates are positive

Table 5.3: P-values for test of the null hypothesis $\sigma_a = 0$ for all traits; ratio of the estimates $\sigma_a^2/(\sigma_a^2 + \sigma_e^2)$, and heritability estimates from Kosrae population[125]. * Lowe et al. collected fasting plasma glucose, which is a different variant of glucose measurement from the one collected in NFBC66 subjects.

Phenotype	**P-value for** $\sigma_a^2 = 0$	$\sigma_a^2/(\sigma_a^2 + \sigma_e^2)$	h^2 **from Lowe et al.[125]**
BMI	2.3×10^{-6}	0.293	0.473
CRP	7.2×10^{-3}	0.156	0.245
DBP	2.7×10^{-4}	0.193	0.289
GLU	1.1×10^{-5}	0.252	0.188*
HDL	4.6×10^{-11}	0.388	0.391
INS	9.7×10^{-4}	0.204	N/A
LDL	8.8×10^{-18}	0.454	0.414
SBP	6.1×10^{-8}	0.281	0.243
TG	1.2×10^{-8}	0.191	0.274
HEIGHT	1.5×10^{-49}	0.782	0.790

there exist more sophisticated algorithms for relatedness estimation, with direct application to association mapping [139, 15, 37]. It remains true that the IBS matrix has the additional practical advantage to be positive definite, which makes it trivial to use it to estimate σ_a^2 and σ_e^2.

Relying then on the estimated IBS matrix, we then explored the strategy described in the method section. The values of $\hat{\sigma}_a^2$ and $\hat{\sigma}_e^2$ estimated for each phenotype are reported in Table 5.3, together with the p-value for an F test for the hypothesis $H_0 : \sigma_a^2 = 0$, and the heritability estimate for the traits obtained from the recent studies using an isolated population from the island of Kosrae[125]. One important observation is that the relative value $\hat{\sigma}_a^2/(\hat{\sigma}_a^2 + \hat{\sigma}_e^2)$, changes with phenotype, consistent with the observation that ignoring dependence between the individuals in the sample has different effects for each trait. The attentive reader would have noted that the ratio $\hat{\sigma}_a^2/(\hat{\sigma}_a^2 + \hat{\sigma}_e^2)$ closely resembles the one used in defining heritability, as it provides an estimates of phenotypic variance explained by the normalized IBS matrix $\hat{S}_N$. It is important to note how step 1 of our methodology, however, it is not directly interchangeable with the heritability of the trait; the IBS matrix does not correspond exactly to the kinship coefficients, and, most importantly, our sample

Table 5.4: For the traits that lead to the identification of at least one locus, we compare the most significant p-value obtained running EMMAX with the corresponding value obtained using EMMA.

phenotype	**SNP**	**EMMAX**	**EMMA**	**Diff**	**Ratio**
CRP	rs2794520	3.98×10^{-23}	3.95×10^{-23}	2.33×10^{-25}	1.01
GLU	rs560887	5.73×10^{-12}	5.33×10^{-12}	3.96×10^{-13}	1.07
HDL	rs3764261	6.11×10^{-33}	6.08×10^{-33}	3.14×10^{-35}	1.01
LDL	rs646776	1.21×10^{-14}	1.19×10^{-14}	1.57×10^{-16}	1.01
SBP	rs782602	3.54×10^{-7}	3.44×10^{-7}	1.02×10^{-8}	1.03
TG	rs1260326	1.32×10^{-10}	1.36×10^{-10}	-3.96×10^{-12}	0.971
HEIGHT	rs6719545	4.53×10^{-7}	4.46×10^{-7}	7.34×10^{-9}	1.02

may not contain sufficiently related individuals. Nonetheless, these values are quite concordant with the previous heritability estimates.

Using the estimated $\hat{\sigma}_a$ and $\hat{\sigma}_e$ we proceeded with a generalized least square estimates of the β_k and test statistics. The genomic control parameters we obtain are much lower than both the original ones and the ones from regression analysis including 100 PCs (Table 5.1) using EIGENSTRAT[168]. Figure 5.6 provides a graphical illustration, using QQ plots of the different p-values distributions from these three tests.

Because EMMAX does not re-estimate the σ_a^2 and σ_e^2 for each SNP, one may suspect, that the significance of SNPs contributing to the trait may be reduced. To assess the seriousness of this concern, we run the original EMMA which uses a full mixed effect model on the SNPs that achieved a p-value lower than 5×10^{-7}. These SNPs are the ones for which we are more likely to see a substantial difference between the results of a mixed effects model and GLS. Overall, as expected, the p-values from the full mixed effect model tend to be more significant then the p-value from the GLS model (Figure 5.7 and Table 5.4). However, the magnitude of this difference is very limited, as shown by examining the logarithm of their ratios.

While the p-value distribution, and the λ values, are quite different in the simple analysis and EMMAX, these two approaches do not diverge substantially in terms of loci declared to be genomewide significant. The distribution across the genome of p-values from the uncorrected screen with p-values obtained from EMMAX

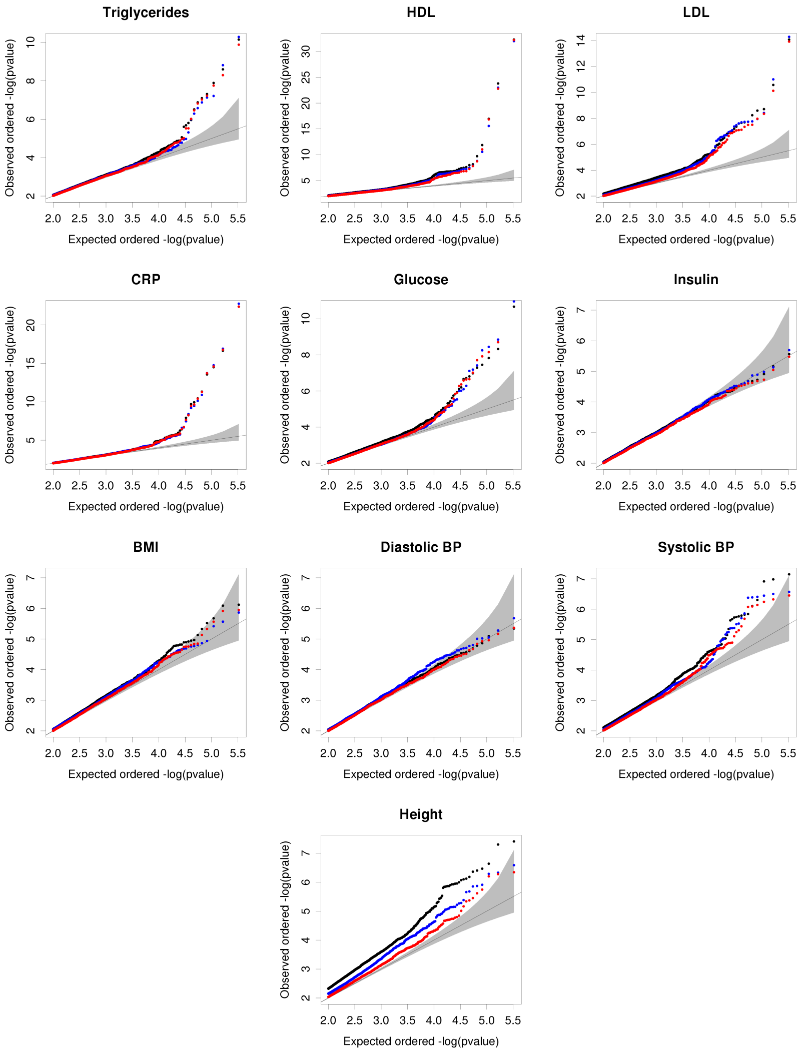


Figure 5.6: QQ-plots on the log10 scale of the association p-values obtained for nine traits according to there different models. In black, results from the unadjusted analysis; in blue results from the analysis conducted using 100 PC, and in red results from EMMAX.

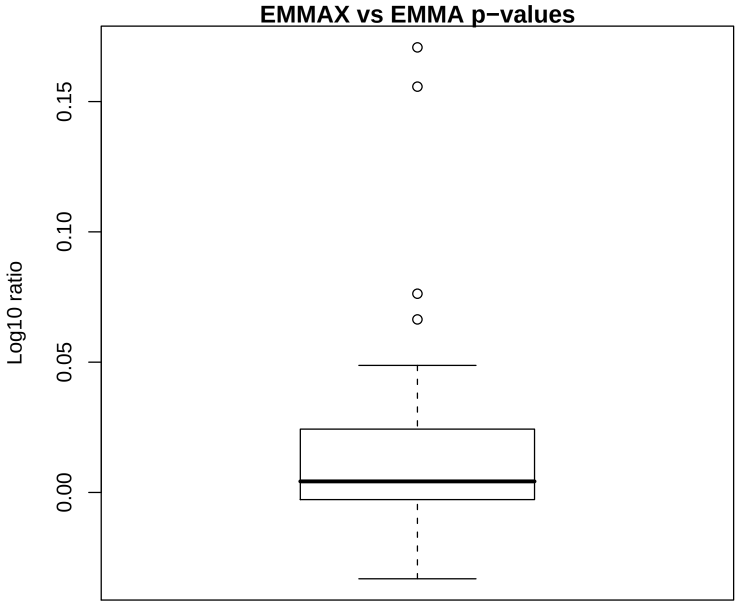


Figure 5.7: We compare the values of p-value obtained running EMMAX with the corresponding value obtained using EMMA for the SNPs whose p-value under EMMAX was smaller than 5×10^{-7}.

shows clear agreement between the identified locations (Figure 5.8). In fact, in general, adopting the a threshold of 5×10^{-7} on the p-value, EMMAX and the uncorrected analysis lead to the identification of the same loci; however, some of these loci do not meet this significance threshold when we apply the genomic control correction.

Unlike genomic control, the EMMAX model alters the rank of the statistics of SNPs. This is especially important in light of the fact that many follow up and multi-stage design studies take the approach of genotyping all SNPs exceeding some predefined threshold[53, 202, 2]. We examine the extent to which the adoption of the EMMAX model changes the ranking with respect to the use of simple linear regression and regression that includes PCAs. We took the top k markers, from the results of EMMAX, the uncorrected method, and PCA, for k between 10 and 10000. For each of these sets we calculated the number of SNPs shared between the lists and the fraction of these shared SNPs relative to the number of unique SNPs in both lists. We also identified the ranks in the uncorrected and PCA analyses in the top

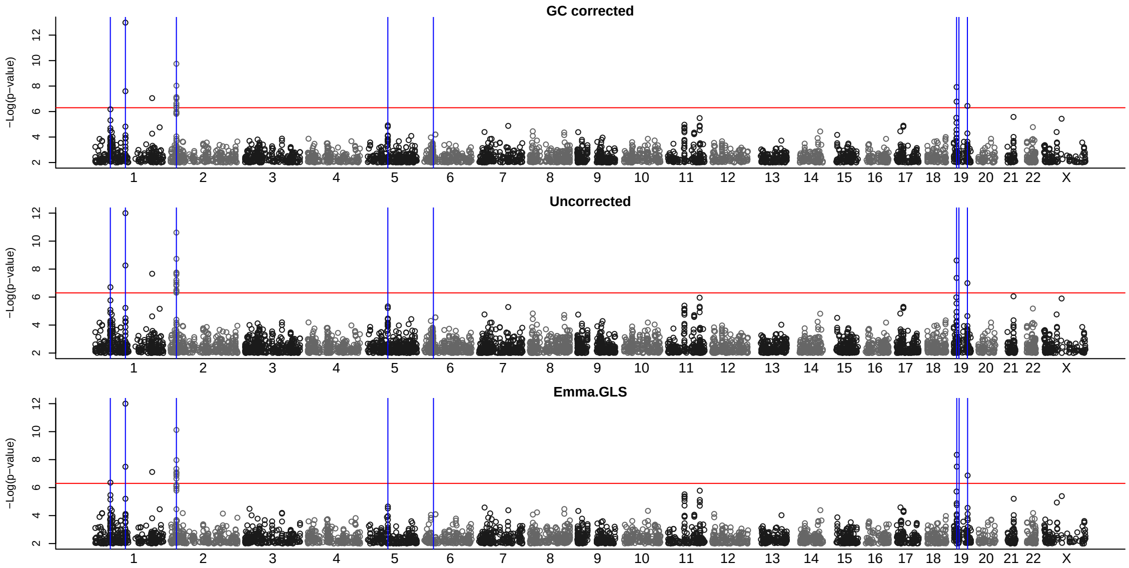


Figure 5.8: From top to bottom, the plots present the association p-values for LDL obtained with three methods: the uncorrected analysis, the genomic control correction, and EMMAX. Genomic position is on the x-axis (Chromosome number is indicated at the bottom of the plot). The negative of log10 of the association p-value is on the y-axis. Only p-values lower then 10^{-2} are displayed. The horizontal red line corresponds to a p-value of 5×10^{-7}. Blue vertical lines indicate position of loci recently identified in GWAS.

Table 5.5: Comparison of top 2000 hits obtained with EMMAX, PCA, and uncorrected analysis. The numbers in second to fourth column represents the proportion of shared SNPs between each pair of analysis, when selecting top 2,000 SNPs in each analysis.

Phenotype	**t-test vs EMMAX**	**t-test vs PCA**	**PCA vss EMMAX**	**t-test λ**
CRP	0.877	0.629	0.665	1.012
TG	0.854	0.599	0.642	1.026
INS	0.833	0.547	0.602	1.033
DBP	0.853	0.621	0.644	1.045
BMI	0.788	0.553	0.603	1.036
GLU	0.769	0.524	0.584	1.045
HDL	0.706	0.500	0.582	1.054
SBP	0.692	0.472	0.589	1.063
LDL	0.616	0.476	0.610	1.093
HEIGHT	0.448	0.384	0.498	1.193

2000 EMMAX hits. Results demonstrate that a great portion of top associated SNPs agrees between these methods, but a considerable amount of discordance is also observed (Table 5.5 and Figure 5.9). In general, EMMAX results become similar to uncorrected t-test when the effect from sample structure is small, but they becomes more similar to PCA results as the effect from sample structure increases. Interestingly, PCA method generally shows larger departure from uncorrected method than EMMAX does. For example, when the over-dispersion of test statistics is relatively negligible such as CRP, only 63% of top 2000 hits were concordant between PCA and uncorrected t-test, while 88% were concordant between EMMAX and t-test. The discordance becomes greater as larger over-dispersion of test statistics is observed, but the concordance between PCA and EMMAX drops relatively slowly compared to the other pairs, because of the increasing confounding effects from the sample structure on the uncorrected t-test results.

Finally, we compared the estimated effects β_k in the simple analysis and EMMAX. Again, we focused on the set of "significant loci", identified as those that obtain a p-value lower then 5×10^{-7} with the EMMAX analysis. Most estimates of β_k are very similar between the two methods (Figure 5.10) .

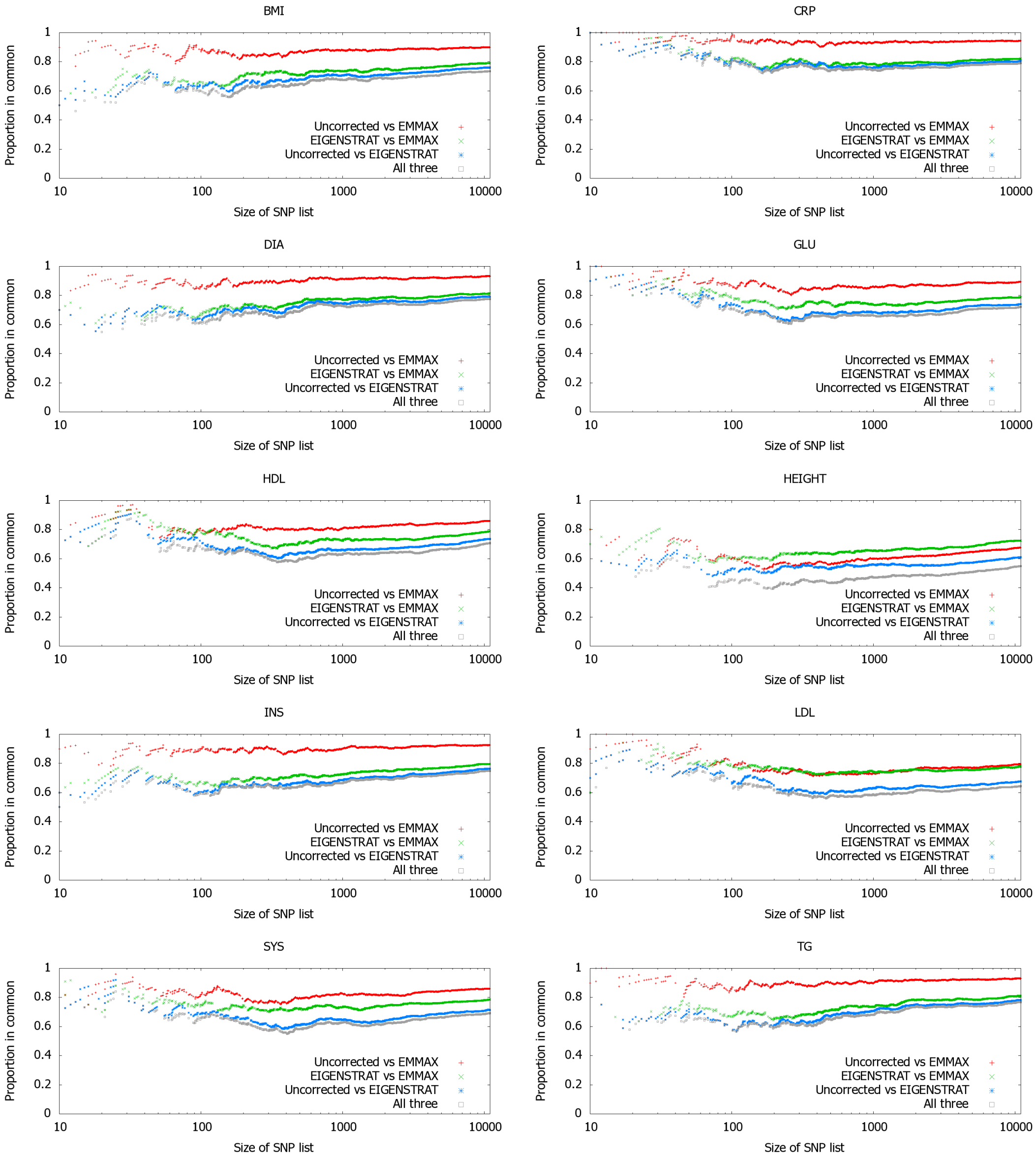


Figure 5.9: Concordance of strongly associated SNPs between different methods across 10 phenotypes with NFBC66 data

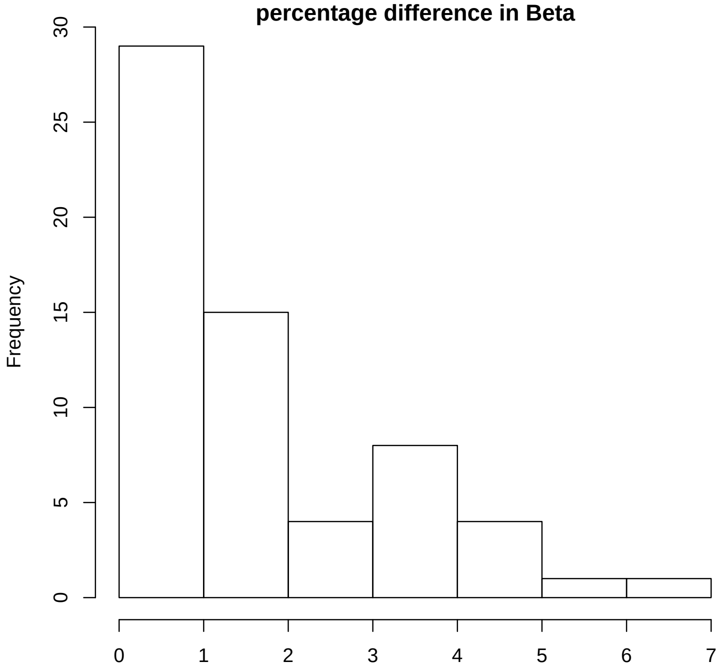


Figure 5.10: Histogram of the percentage differences in β estimates for EMMAX and uncorrected analysis. This is temporary, it exclude height for which we have not re-run the comparison yet.

5.3.2 Application to WTCCC case-control data

We applied our method to the Wellcome Trust Case Control Consortium (WTCCC) data. As described in the Materials and Methods section in more detail, the case-control phenotypes are encoded as 0 and 1 and a linear model is used in the spirit of Armitage's test. We performed association mapping over the seven disease phenotypes using various methods in the same setting to the original study. We observed a very similar level of inflation factors λ to the original study when the test statistics are uncorrected: bipolar disease (BD) 1.11, coronary artery disease (CAD) 1.06, Crohn's disease (CD) 1.10, hypertension (HT) 1.06, rheumatoid arthritis (RA) 1.03, type 1 diabetes (T1D) 1.04, and type 2 diabetes (T2D) 1.07. Consistent with our observations over the NFBC66 data, correcting for 100 principal components only partially reduced the inflation factors (Table 5.6). When EMMAX is applied, none of the phenotypes showed significant inflation of test statistics.

However, we noticed that two of the phenotypes, RA and T1D, show signif-

Table 5.6: Comparison of genomic control inflation factor obtained with different models in seven WTCCC phenotypes. * represents the inflation factor when the variance component parameters ($\hat{\sigma}_a^2$ and $\hat{\sigma}_e^2$) are estimated by conditioning on the large-sized SNP effects within the extended MHC region with $p < 10^{-10}$

Phenotypes	**uncorrected**	**ES with 100 PCs**	**EMMAX**
BD	1.105	1.071	0.998
CAD	1.063	1.048	1.006
CD	1.098	1.055	1.000
HT	1.055	1.051	0.997
RA	1.028	1.031	0.965 (0.991*)
T1D	1.043	1.028	0.946 (1.011*)
T2D	1.065	1.042	0.996

icant deflation of test statistics beyond the 95% confidence interval ($\lambda = 0.965$ for RA, $\lambda = 0.946$ for T1D). At the same time, we noticed that the QQ plot of the test statistics show much stronger associations than what we observed with metabolic phenotypes (Figure 5.11). RA and T1D contained a large number of strong associations, most of which are located at the MHC region of chromosome 6. Ninety nine and 280 SNPs within the extended human MHC region[44] are strongly associated at $p < 10^{-10}$, with RA and T1D respectively. The other five diseases showed no significant SNPs in the same region. These SNPs account for 47% and 63% of the phenotypic variance, respectively[223]. Such strong SNP effects may lead to an inaccurate estimation of variance component, resulting in over-dispersion or under-dispersion of test statistics. we re-estimated the variance component by conditioning on these strong SNPs within the extended MHC region as described in the method section. As a result, the genomic control λ became 0.991 for RA, and 1.011 for T1D.

5.3.3 Marker specific inflation factors

To illustrate the value and meaning of the marker specific inflation factors defined in the method section, we calculated them wit reference to the 10 phenotypes of the NFBC. Across all 10 NFBC phenotypes, the distribution of marker specific inflation factors greatly vary (Figure 5.12A). The genomic control inflation factors *lambda* from these phenotypes are quite concordant with the marker specific inflation

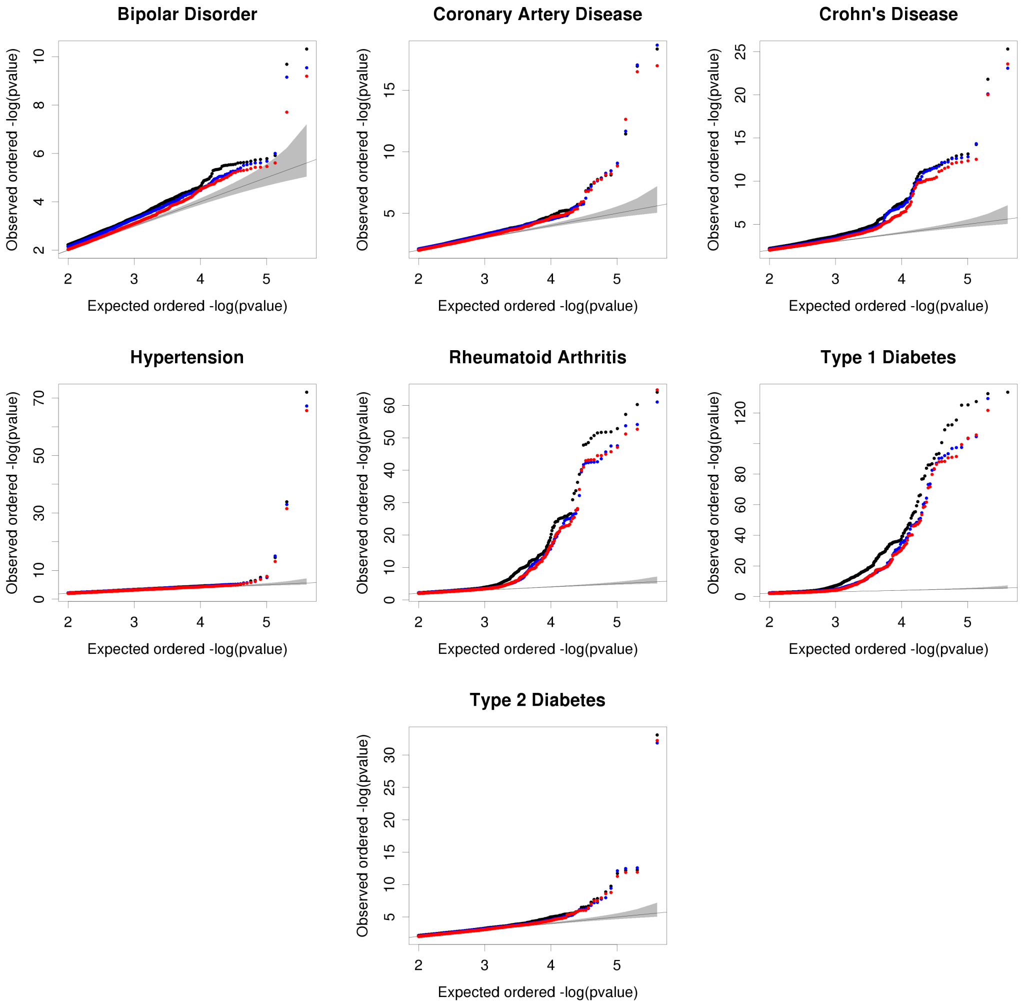


Figure 5.11: QQ-plots on the log10 scale of the association p-values obtained for seven disease phenotypes according to three different models. In black, results from the unadjusted analysis; in blue results from the analysis conducted using 100 PC, and in red results from EMMAX

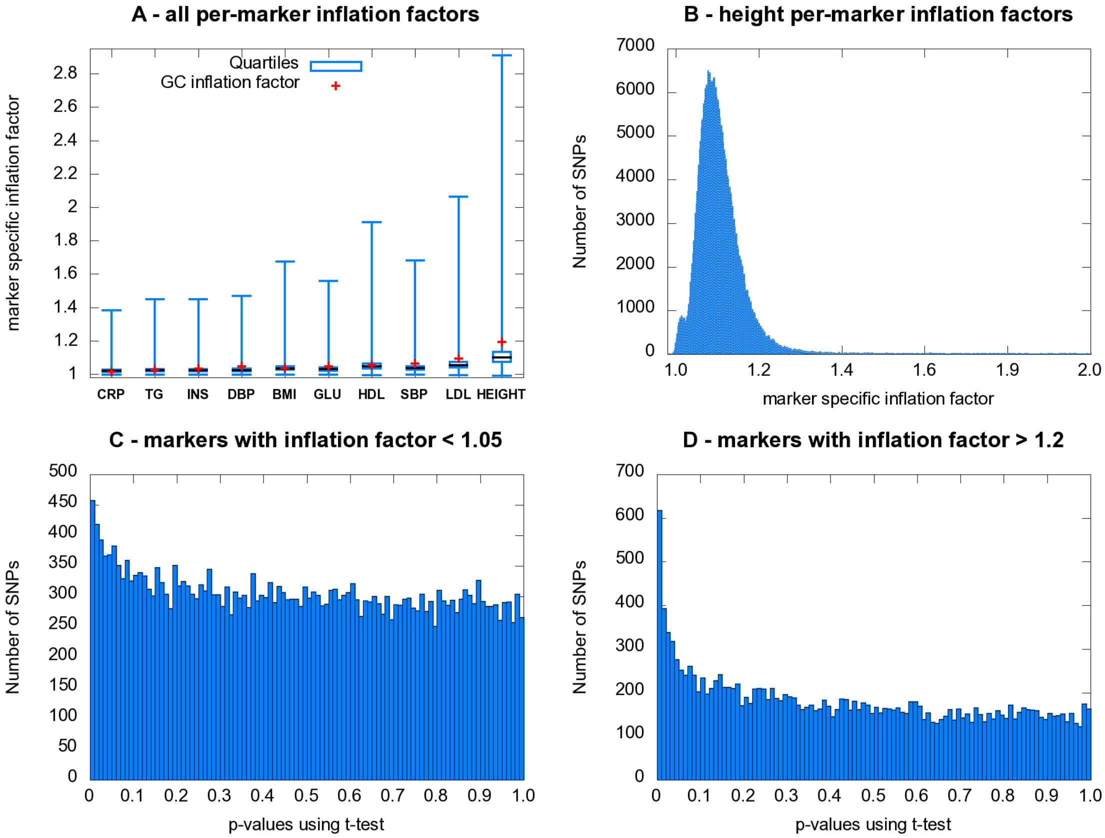


Figure 5.12: Distribution of the marker specific inflation factors from NFBC66 data sets. (A) Box plots of the marker specific inflation factors across 10 phenotypes, in addition to the genomic control inflation factor for each phenotypes. (B) The distribution of marker specific inflation factors for height phenotype. (C) The p-values of the height phenotype association when the estimated per-marker inflation factors are less than 1.05 (30,679 SNPs) versus when they are (D) greater than 1.2 (18,204 SNPs).

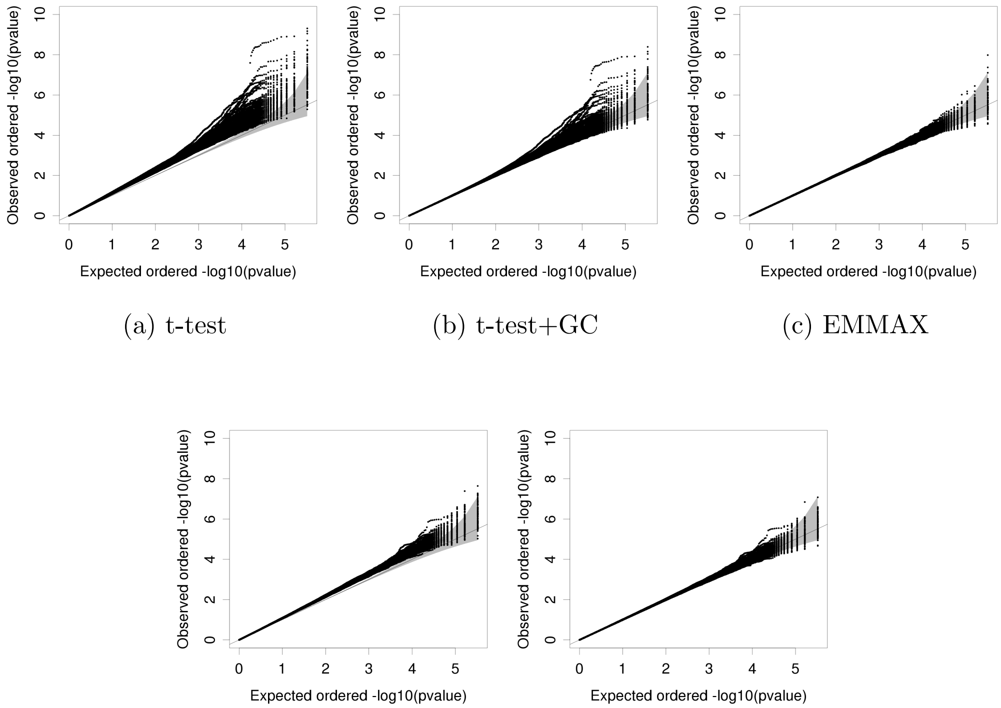
factors. The distribution of marker specific inflation factors estimated from the height phenotype showed mean 1.111, standard deviation of 0.087, and median value of 1.098 (Figure 5.12B).

Recall that the inflation factors are estimated on the basis of correlation of the genotypes with the IBS matrix, without reference to their association to the phenotype. It is interesting, then, to explore the predictive power of marker specific inflation factors for association p-values of uncorrected analysis. The distribution of height association p-values for SNPs with inflation factor $<$ 1.05 appears reasonably close to uniform (Figure 5.12C), while the distribution of association p-values for SNPs with inflation factor > 1.20 shows a clear excess of small p-values (Figure 5.12D). Given that SNPs with inflation factor > 1.20 are identified without consideration of their possible association with the phenotype, one may safely assume that this excess of small p-values is an over-dispersion effect.

These results underscore how applying the same deflation value to all test statistics may be inappropriate, possibly reducing power and not sufficiently controlling for false positives. To further illustrate this point, we run a simple simulation experiment assuming that the polygenic model at the basis of EMMAX (5.8), and that none of the tested SNPs has a significant effect. Clearly, generating data under this model puts our method at an advantage, but our goal here is not to illustrate its superiority, but simply that under some circumstances uniformly deflating p-values may be inappropriate.

We randomly simulated 100 sets of phenotypes solely from the polygenic background (i.e. $\beta_k = 0, \sigma_e^2 = 0$), and examined the QQ plots before and after genomic control. While the inflation is fairly well resolved in most of the SNPs after applying genomic control, we observed a significant amount of fluctuation of the test statistics at the tail of the distribution (Figure 5.13A,B). More than 20% of the phenotypes showed inflation beyond the 95% confidence interval at the tail of the distribution, and 5% showed deflation outside the confidence interval. On the other hand, when EMMAX is used over the same sets of random phenotypes without applying genomic control, the distribution of p-values closely follows the expected distribution (Figure 5.13C). This is because the SNPs with higher per-marker inflation factors tend to have residual inflation after applying genomic control, which divides every test statistic by a constant factor. When PCA is used, residual inflation was still observed, consistent to previous results (Figure 5.13D). When PCA and genomic control is combined, the inflation at the tail has been fairly well reduced, but it at least one phenotype showed a noticeable departure from the expected distribution. (Figure 5.13E).

The results showing that a per-marker inflation factor greatly varies per SNP have a considerable impact on the meta-analysis and multi-stage analysis. Many current meta-analysis and multi-stage analysis combine the test statistics from after



(a) t-test

(b) t-test+GC

(c) EMMAX

(d) PCA

(e) PCA+GC

Figure 5.13: QQ plots of 100 randomly generated phenotypes under the variance component model using a (a) t-test, (b) t-test after genomic control, (c) EMMAX, (d) PCA with 100 PCs and (e) PCA after genomic control

correcting for potential inflation using Genomic Control[238, 202, 2]. We evaluated the concordance of marker specific inflation factors between different data sets. Since WTCCC control samples consist of two sets of data, 1958 British Birth Cohort (58C) and UK Blood Service Control (NBS)[223], we first compared the marker specific inflation factors between these data sets. It will be interesting to understand how concordant these marker specific inflation factors are across different data sets that are essentially collected from the same population. We observed a very strong correlation ($r = 0.95$) of the expected per-marker inflation factors between the two data sets (Figure 5.14A). We further compared the concordance of marker specific inflation factors between the NFBC66 samples and WTCCC control samples. Since these two data sets are genetically farther apart than between the two WTCCC control sets, we expect higher disconcordance of the marker specific inflation factors. The discordance may further increase due to the fact that different genotyping array platforms are used to collect the genotypes in different data sets. Nonetheless, when we compared the marker specfic inflation factors of the 50,298 SNPs shared between the two data sets (Figure 5.14B), we observed a strong correlation ($r = 0.70$), which suggest that the marker specific inflation factors may be correlated across multiple data sets in meta-analysis or multi-stage analysis. Because of this, the standard approaches may suffer from the accumulated residual confounding effect across multiple studies, especially at the tail of the distribution, considering the simulation results presented in the previous paragraph.

5.4 Discussion

We proposed an expedited mixed model approach for large human genome wide association samples. Compared to principal component analysis, which identifies and corrects for population structure using the major axes of pairwise genetic relatedness matrix, our method accounts for the entire relatedness matrix by means of a linear mixed model, which has been demonstrated to be effective in model organism association mapping with complex sample structure. Compared to genomic control, which only rescales the test statistics by a constant inflation factor across the genome, our method alters the rank of marker associations through marker-specific

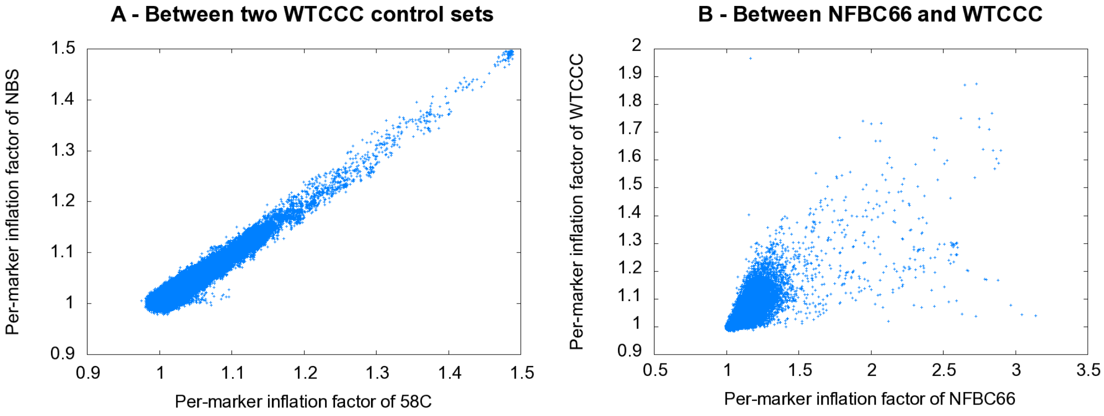


Figure 5.14: Concordance of per-marker inflation factor (A) between two different control sets (58C and NBS) in WTCCC data set, and (B) between NFBC66 samples and WTCCC control samples using the 50,298 overlapping markers

correction for sample structure. Compared to the approach combining the above two methods, our test statistics is mathematically more tractable. Our method achieves much higher computational efficiency than previous mixed model methods by avoiding separate variance component estimations across different markers. Our method robustly corrects for over-dispersion of test statistics consistently across various quantitative and dichotomous phenotypes in NFBC66 and WTCCC data.

Accounting for marker specific effects from sample structure can be advantageous over genomic control in reducing both false positives and false negatives. We have demonstrated that the over-dispersion of test statistics may still exist at the tail of distribution after applying genomic control through simulation studies. And the over-dispersion may be exacerbated in meta-analysis or multi-stage analysis. It was shown that the mixed model have higher statistical power than the genomic control and PCA methods when the effect from sample structure is large such as in model organisms[235, 103]. Although the gain in the power would be smaller for a human population, the statistical power of our method is expected to be higher than conventional methods, especially when the effect from sample structure is large.

Of course, our methodology is not the only one that leads to marker specific correction. In particular, we have analyzed the result of including principal components in the least squares association, and confirmed that it can be effective. This approach can successfully account for large-scale genetic difference due to a distinct number of sub-populations. However, it is not designed to correct for cryptic relatedness. To make this point clear, let us consider an extreme example. To handle the levels of familial relatedness typical of a sib-pair studies, we would need $n/2$ principal components. Traditional variance component models, based on the kinship matrix, can indeed deal with such familial relatedness. Our method can be viewed as an extension of the variance component modes, to account for undocumented genetic relationships amongst the study subjects, by leveraging the information from the high-density genotypes. Kang et al.[102] presented an simulation study on eQTL mapping using principal component analysis and a mixed model. They simulated different types of confounding effects resembling the population structure effect between two different clusters of individuals, and familial relatedness effects between pairs of individuals. The results show that principal components are only effective in correcting for the former, while mixed models are effective in correcting for both effects.

A number of recent studies have underscored the advantages of capitalizing on genetic similarity as measured by the IBS status of individuals across the genome, or in specific genomic locations. For example, multimarker haplotype association tests can be constructed by taking into account the similarities between the haplotypes[224, 187, 204, 205]. From another viewpoint, Guan et al.[77] proposed using a similarity matrix to achieve a better matching of cases and controls prior to testing. Adopting an approach very similar to the one proposed here, Zhao et. al[241] applied the R package `glmmPQL` which uses a generalized linear mixed model with a penalized quasi-likelihood to family-based study samples[28]. The results, however, showed a significant level of under-dispersion of the test statistics. It is possible that this unsatisfactory performance is due to the quasi-likelihood estimation procedure, as we do not experience it with our data and algorithm.

The effective application of our method depends on the an appropriate estimate of the variance components $V = \sigma_a^2 \hat{S}_N + \sigma_e^2 I$. The IBS matrix appears to better capture the long distance relationships that result on variations at the population levels than IBD estimates. However, when the structure of the sample at hand is better described in terms of fairly recent cryptic relatedness, methods based on the estimation of IBD may have an advantage. The example of RA in the WTCCC dataset exemplifies the difficulties encountered by EMMAX when there are SNPs strongly contributing to the phenotype. In such cases, one can carry out conditional analysis, alleviating the problem. In principle, our approach is also suitable for association mapping in a sample comprised of individuals from different populations and with admixed background. In such cases, it is important to consider SNP ascertainment bias in estimating the degree of relatedness between individuals. Since many SNP probes in genotyping arrays are selected from European populations, the IBS distance between two individuals may appear to be larger between unrelated European samples than between unrelated individuals from other population. In order to resolve the ascertainment bias, each SNP may be differently weighted when computing the IBS similarity matrix. Kang et. al[103] presented a general framework to compute the similarity matrix with different weight per marker. The effect of ascertainment bias may be reduced if each marker is weighted by the number of HapMap SNPs taggable by the marker. Different weighting schemes can be used to account for heterogeneous genetic effects per marker or per genomic region.

Finally, while the analysis presented in this chapter relies on decomposing the variance in two terms, $V = \sigma_a^2 \hat{S}_N + \sigma_e^2 I$, it is straightforward to account for multiple variance components to more precisely model heterogeneous relatedness matrix, such as additive and dominant effects. In expression quantitative trait loci (eQTL) mapping, for example, one may want to add additional variance components to account for technical bias using an additional variance component[102]. When multiple variance components are involved, one would need to resort to algorithms as PROC MIXED implemented in SAS, since EMMA is developed for two variance components only; this would increase the running time of the first step of our procedure. However, since the same variance components estimates will be used in GLS testing across the genome, the the overall computational time should still be acceptable.

Chapter 5 is currently in submission for publication for the material. Hyun Min Kang, Jae-hoon Sul, Susan Service, Noah Zaitlen, Sit-yee Kong, Nelson Freimer, Chiara Sabatti, and Eleazar Eskin. “Accounting for sample structure in large scale genome-wide association studies using a variance component model”. The dissertation author and Jae-hoon Sul are the primary investigators and authors of this paper.

Chapter 6

Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots

6.1 Motivation

Genome wide analysis of gene expression data in segregating populations has been widely conducted to understand the genetic basis of regulation in many organisms including yeast[25], Arabidopsis[107], mouse[34, 29] and human[35, 198]. In order to understand the complex regulatory network, numerous statistical analysis methods have been proposed including clustering of co-regulated genes[236], multipoint linkage analysis[195, 26], prediction of regulatory modules[117, 71], and pathway enrichment analysis [199, 234].

Among these “genetical genomics” approaches, the most widely used statistical analysis is expression quantitative trait loci (eQTL) mapping between genetic variation and gene expression levels[27]. The goal of these studies is to identify associations between an individual genetic variation and the differential expression of a gene that might help explain the transcriptional regulation of the gene. Many recent studies have identified a large number of *cis* associations between eQTLs and the expression of genes in close proximity. They have also identified many more *trans* associations between eQTLs and the expression of genes in other regions of the genome[236, 34, 89]. An interesting observation consistent across multiple datasets is that hundreds or even thousands of genes are *trans*-regulated by a small number of genomic regions called "regulatory hotspots"[107, 34] and these associations appear as "*trans*-regulatory bands" in eQTL plots regardless of the normalization method used[89, 34, 29, 160, 229, 34].

Recent genetical genomics studies of yeast have provided much evidence supporting the existence of global regulators that induce *trans*-regulatory bands[161, 66]. For mammalian expression datasets, although a large numbers of regulatory hotspots have consistently been observed, the locations of these regulatory hotspots are inconsistent between different datasets[89, 34, 160]. Simulation studies suggest that spurious regulatory hotspots may be frequently observed in outbred populations[174, 46, 216].

Building on previous studies we examine two first-generation expression datasets of recombinant inbred (RI) mice where their regulatory hotspots have been shown to poorly replicate in previous studies[160]. Due to the high degree of systematic confounding inherent in these datasets, it is particularly challenging to distinguish true genetic effects from the spurious associations. The availability of biological replicates in these datasets allows us to compare the level of true positives between different methods. Two observations suggest that many *trans*-regulatory bands previously identified in these datasets correspond to "spurious" regulatory hotspots not real genetic effects. First, the locations of regulatory hotspots are inconsistent across disjoint sets of biologically replicated samples. Second, stronger *trans*-regulatory bands frequently appear with randomly permuted SNPs. To understand the cause of this phenomenon, we carefully examined these datasets and identified a surprising pattern of inter-sample correlation where the pairwise correlations of expression arrays between different strains are often stronger than between replicates of the same strain.

Previous studies have shown that many factors contribute to the spurious correlation between microarray samples including systematic bias from sources such as technical variation in microarray manufacturing[38, 3], variations introduced during sample preparation such as the time postmortem a sample is collected, and variations introduced during expression measurements such as the batch of reagents used or laboratory ozone levels[57, 24]. Such spurious inter-sample correlation are usually not completely resolved by randomized design of the experiment[38] or through low-level normalization techniques[94, 233].

We suspect that when a SNP, by chance, segregates the strains in a manner consistent with the inter-sample correlation, the p-values of associations between that SNP and the transcripts are inflated leading to spurious associations and in extreme cases, a "*trans*-regulatory band." To verify this phenomenon, we constructed a set of simulated data to intuitively show how complex inter-sample correlation structure inherent in expression data leads to associations between genetic loci and a large number of gene transcripts inducing spurious regulatory hotspots. When we generate random expression data using the same inter-sample correlation structure found in the recombinant inbred expression data, we observed exactly the same regulatory hotspots.

Two types of computational approaches have previously been proposed to reduce the effects of confounding factors in gene expression experiments. The first type are methods that correct for known confounding factors, including ComBat[100], which directly estimates the location and scale model parameters that represent the batch effect using an empirical Bayes (EB) approach. The second type are methods that correct for unknown confounding factors, including Surrogate Variable Analysis (SVA)[118], which identifies, estimates and corrects for principal components of expression heterogeneity.

We propose a statistical method that corrects for confounding effects induced by complex inter-sample correlation of expression measurements in eQTL mapping using a linear mixed model. Our Inter-sample Correlation Emended (ICE) eQTL mapping directly incorporates the complex correlation structure into the statistical model as a variance component accounting for random effects. Compared to ComBat, our approach is not limited by prior knowledge of confounding factors and is capable of capturing the complex correlation structure introduced by multiple known and unknown effects. Compared to SVA which projects confounding effects onto several distinct single dimensional vectors each treated as a fixed effect in the statistical model, our random effects model is not limited by the number of confounding variables because it does not explicitly infer and correct for each confounding variable. Instead, our method only needs to estimate the total correlation between samples and corrects for the cumulative effects over all confounding factors using the inter-sample correlation structure. Furthermore, the statistical power of SVA decreases as the number of confounding variables increases due to the loss of degrees of freedom while our method always uses only one additional degree of freedom for the inter-sample variance component. As a result, our method has the advantage that it is able to correct for a mixture of strong and moderate confounding effects as shown in our simulation studies while SVA is only able to correct for a number of strong confounding factors.

To gain some intuition as to why our random effects model corrects for confounding factors, consider a pair of samples with a differing expression level for a given gene and a marker SNP which segregates the pair of samples. If the remaining gene expression values have similar expression values between the samples, intuitively this pair of samples provides more evidence that the SNP is associated with the gene's expression level than if the remaining gene expression values differ greatly between the samples. In the later case, the expression difference of the gene between the pair of samples is less informative given the large amount of global differences in expression values between the samples, which may be due to a confounding factor such as a batch effect.

We applied our statistical model to expression data from two mouse tissues (hematopoetic stem cell and whole brain). In both cases, ICE eQTL mapping outperformed ComBat and SVA in eliminating the spurious *trans*-regulatory bands while increasing the number of identified *cis* associations. The remaining *trans* associations are more likely to be real genetic effects because they are concordant between tissues and between replicates. In yeast, where global regulators have been previous identified, a separate permutation analysis showed that most of the regulatory hotspots are likely to correspond to real genetic effects. Even though yeast regulatory hotspots are likely to be genuine, they globally influence the expression levels and may seriously confound the identification of gene-specific *cis* or *trans* associations[195, 118]. After applying ICE eQTL mapping to correct for the confounding effects from regulatory hotspots, the number of *cis* associations almost doubled and the concordances of *cis* and *trans* associations between disjoint subsets significantly improved. Finally in human lymphoblastoid cell lines, where other known batch effects have been suggested[3], our analysis identified more real *cis* associations than methods that explicitly correct for the batch effects. Our method is publicly available as an R package at http://mouse.cs.ucla.edu/ice

6.2 Results

6.2.1 Spurious regulatory hotspots in recombinant inbred mice

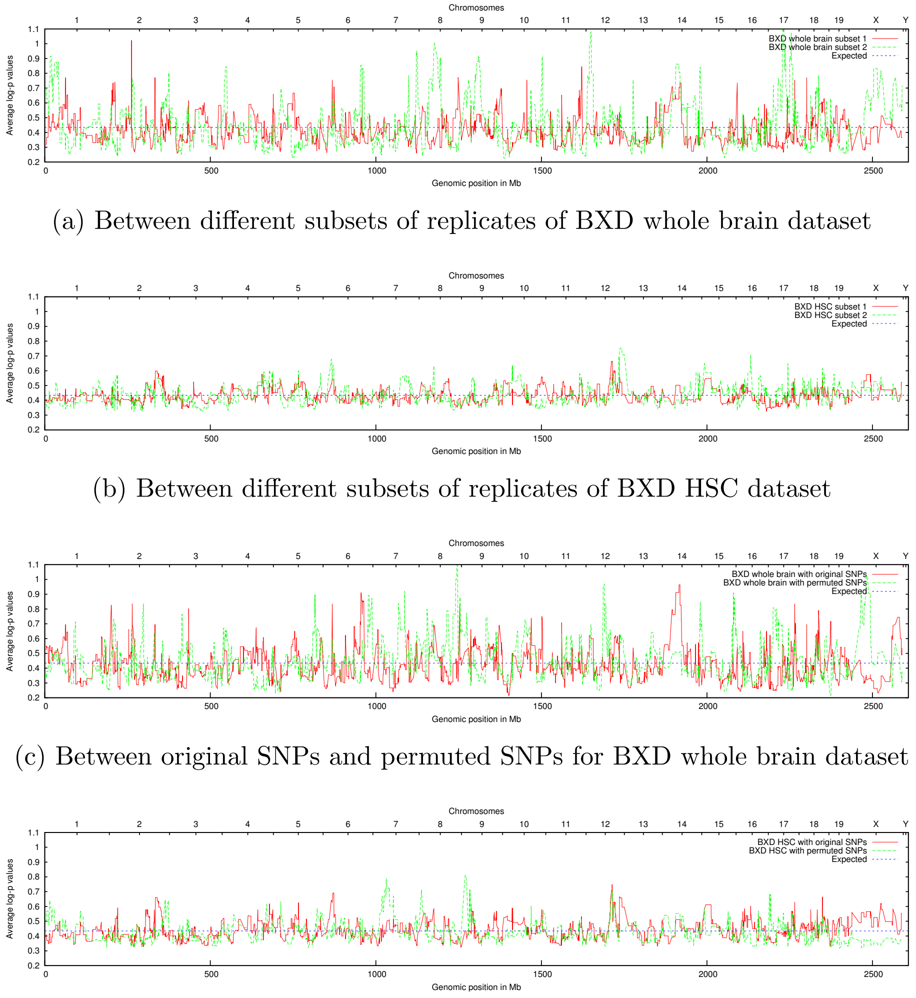
We analyzed the expression data from hematopoetic stem cells (HSC) and whole brain tissue collected from BXD mice where prominent *trans*-regulatory bands have previously been observed[34, 29]. However, most *trans*-regulatory bands found in these first-generation mouse expression datasets do not overlap with *trans*-regulatory bands from independent studies[130, 160]. We selected these datasets to evaluate and correct for the systematic confounding effects for two reasons. First, the presence of biological replicates allows us to quantify the level of systematic confounding effects that are heavily imprinted in the datasets. Second, we demonstrate that even in the presence of many complex systematic confounding effects, our method is able to recover true genetic signals better than competing approaches.

We first examined the reproducibility of *trans*-regulatory bands between different sets of biological replicates. We defined a metric to quantify the strength of a regulatory band allowing us to compare regulatory patterns between datasets. We performed standard eQTL mapping using the *t*-test and defined the average log p-values across all genes as the regulatory enrichment score. The resulting eQTL map shows that this score correlates well with the prominence of a *trans*-regulatory band (Figure 6.2a,6.2c). We created two disjoint subsets of expression experiments by picking one of the replicates per strain and compared the enrichment scores between them. Interestingly, the observed patterns of *trans*-regulatory bands are inconsistent between the subsets (Figure 6.1a, 6.1b). The enrichment scores between the replicates are uncorrelated (Spearman $r = -0.0067$) in the BXD brain dataset. On the other hand, the HSC dataset shows relatively high correlation of the enrichment scores (Spearman $r = 0.30$) due to the batch effect shared between two groups of strains.

Next, we examined whether regulatory hotspots are likely to be observed at random SNPs by chance. If we observe stronger *trans*-regulatory bands with randomly permuted SNP sets than the original *trans*-regulatory bands, then it suggests that the original hotspots may not correspond to real genetic effects, but are rather caused by co-regulation of a large number of transcripts[174, 216] or other systematic confounding factors. Out of 1,000 random permutations, we observed hotspots with higher enrichment scores than the strongest hotspot in the original dataset 890 and 643 times, corresponding to genome-wide adjusted p-values of 0.89 and 0.643, for whole brain and HSC dataset, respectively. (Figure 6.1c, 6.1d). The inconsistencies between biologically replicated samples and the occurrence of strong *trans*-regulatory bands with permuted SNPs suggest that the observed *trans*-regulatory bands correspond to spurious regulatory hotspots which do not correspond to real genetic effects.

6.2.2 Inter-sample correlation as signatures of systematic confounding effects

The question remains as to how systematic confounding effects cause spurious regulatory hotspots. To gain intuition of this phenomenon, we examined the pairwise correlations between expression arrays or the inter-sample correlation structure. After normalizing each gene's expression levels across strains, we computed the correlation between each strain pair and each replicate pair. The normalization ensures that the correlation between truly unrelated strain pairs is expected to approach zero, while the replicated pairs are likely to have higher correlation between them. We observed that most of the inter-sample correlations in recombinant inbred mouse strains do not correspond to real genetic effects. Correlation maps between intra and inter strain replicates show that the diagonals are not pronounced, providing striking evidence that replicated strain pairs are not correlated (Figure 6.3a,6.3b).



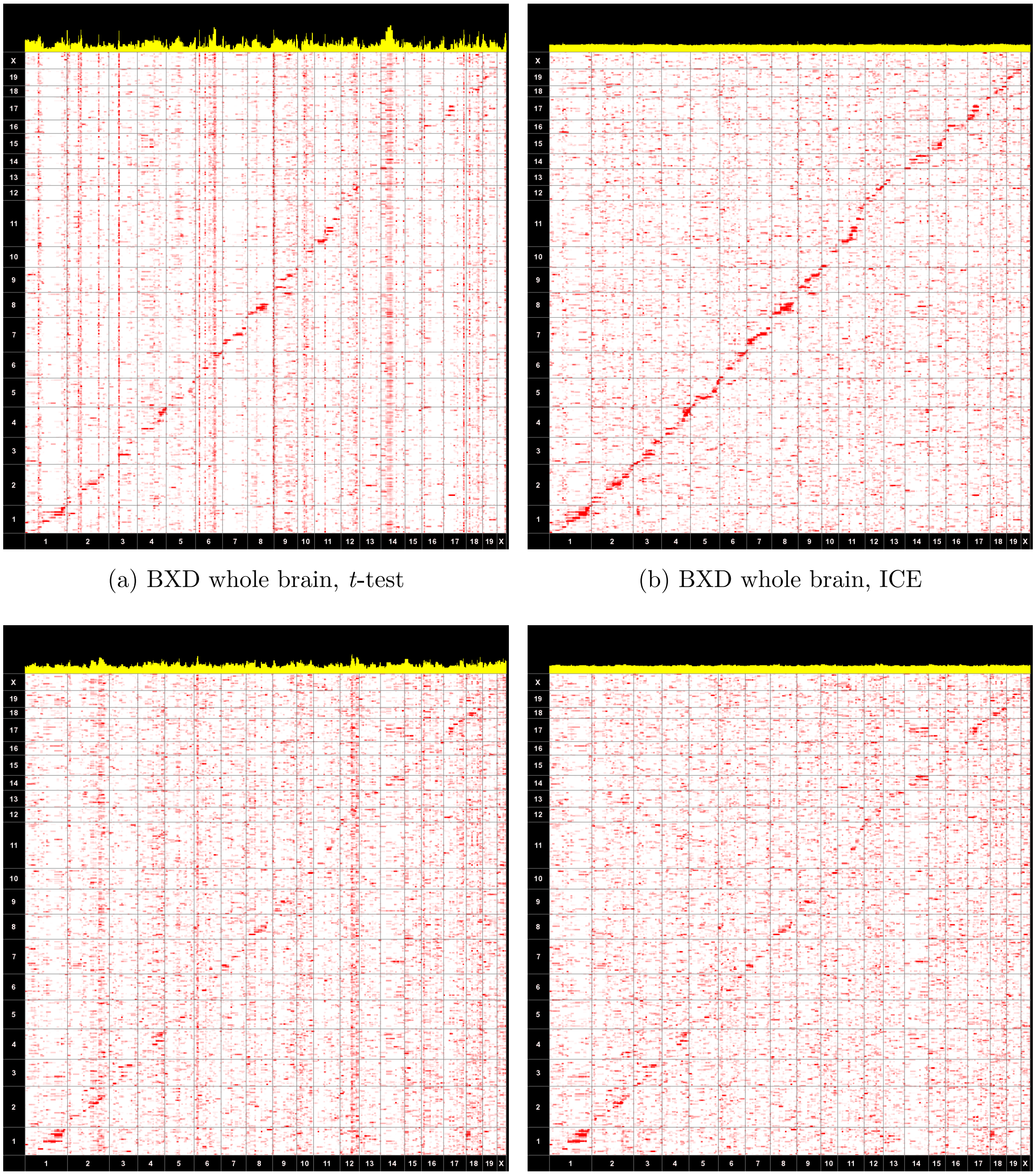
(a) Between different subsets of replicates of BXD whole brain dataset

(b) Between different subsets of replicates of BXD HSC dataset

(c) Between original SNPs and permuted SNPs for BXD whole brain dataset

(d) Between original SNPs and permuted SNPs for BXD HSC dataset

Figure 6.1: Comparison of regulatory hotspots in BXD datasets (A) and (B) compare the strength of *trans*-regulatory bands between replicated subsets, showing that the hotspots are inconsistent between replicates. (C) and (D) compare between using permuted SNPs and using original SNPs, illustrating that even stronger *trans*-regulatory bands are frequently observed using permuted SNPs. The horizontal axis is the genomic positions of the markers in megabases, and the vertical axis is the strength of regulatory hotspots quantified as the average log-p values at each marker across all genes.



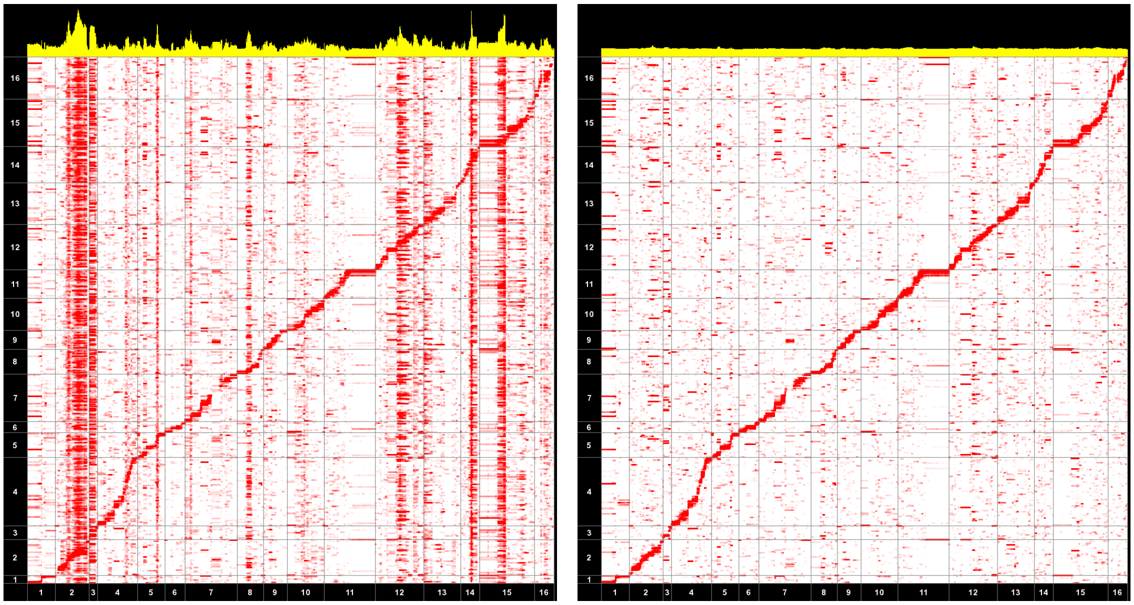
(a) BXD whole brain, *t*-test

(b) BXD whole brain, ICE

(c) BXD HSC, *t*-test

(d) BXD HSC, ICE

Figure 6.2: (Continued to the next page)



(e) Yeast, *t*-test

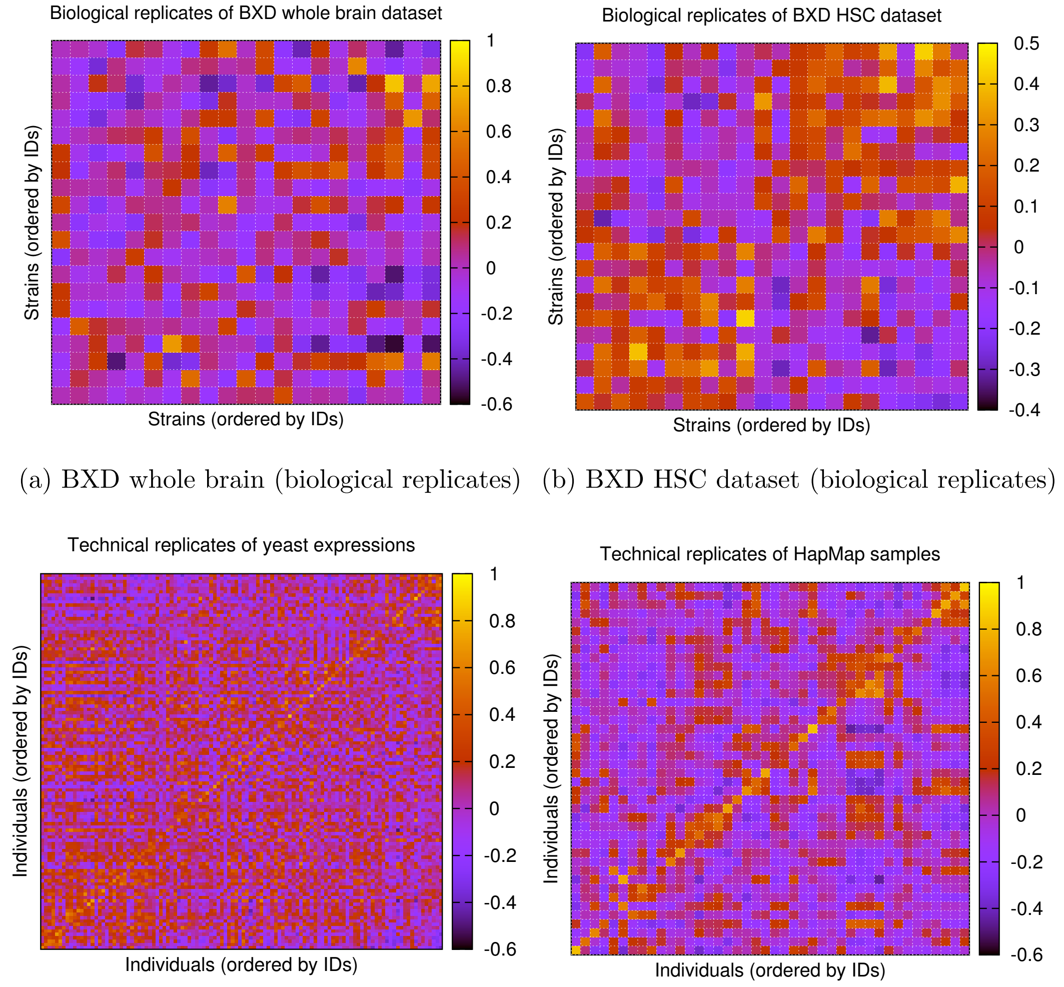
(f) Yeast, ICE

Figure 6.2: Genome wide eQTL maps from (A,B) BXD recombinant inbred whole brain dataset, (C,D) BXD recombinant inbred HSC dataset, and (E,F) yeast dataset, using (A,C,E) standard *t*-test and (B,D,F) ICE eQTL mapping. The SNPs (horizontal) and probes (vertical) are sorted according to genomic position, mapping each pixel to the linkage between a marker and a gene. The color of each pixel represents the strength of the linkage signal, with red being the strongest signal and white being the weakest. The yellow graph on top represents the strength of *trans*-regulatory bands, quantified as the average log-p values at each SNP across all genes. There are clear vertical *trans*-regulatory bands using the standard *t*-test to perform eQTL mapping. Those bands are eliminated using ICE eQTL mapping. A total of 8,596 probes and 7,413 SNPs are mapped in the two mouse datasets, and 5,534 probes and 2,956 SNPs in the yeast dataset.

Furthermore, the results show many unrelated strain pairs with much stronger correlation than expected by chance. In the HSC correlation map, there is also a clear division of two groups where members within a group are highly correlated. Upon further analysis, we discovered that the expression measurements for the two groups of individuals were collected in two batches three months apart.

In order to verify that the inter-sample correlation structure effectively captures the systematic confounding effects inducing spurious regulatory hotspots, we created a simulated expression dataset preserving the inter-sample correlation structure. In this dataset, each SNP corresponds to one simulated transcript with *cis*-regulatory effects accounting for 4% of the variance explained by the SNP. Standard eQTL mapping with simulated data shows almost identical *trans*-regulatory bands as the original data (Figure 6.10). The reason for this is that the SNPs which segregate the strains in a manner consistent with the inter-sample correlation structure are more likely to be associated with many expression transcripts. This result strongly supports that most of the *trans*-regulatory bands are explained by the complex inter-sample correlation structure inherent in expression data.

Furthermore, we evaluated how many transcripts are explained by the inter-sample correlation structure using a variance component model (see Methods). At a false discovery rate (FDR) of 0.05, we observed that 94.1% and 47.9% of the transcripts are significantly associated with the correlation structure in the whole brain and HSC datasets respectively. Since the HSC dataset has an obvious batch effect, we also tested how many transcripts are differentially expressed between the two batches using a *t*-test. At the FDR threshold of 0.05, only 20.0% of the transcripts are differentially expressed. These results suggest that a significant portion of confounding effects in the HSC dataset are not captured by the known batch effect. When applying SVA to test the significance of surrogates variables explaining the expression levels, 88.7% and 40.9% of the transcripts were significantly associated with the five and six identified surrogate variables of whole brain and HSC dataset, respectively, demonstrating that inter-sample correlation captures more of the systematic confounding than what is captured by surrogate variables with fewer degrees of freedom.



(a) BXD whole brain (biological replicates)

(b) BXD HSC dataset (biological replicates)

(c) Yeast dataset (technical replicates)

(d) HapMap dataset (technical replicates)

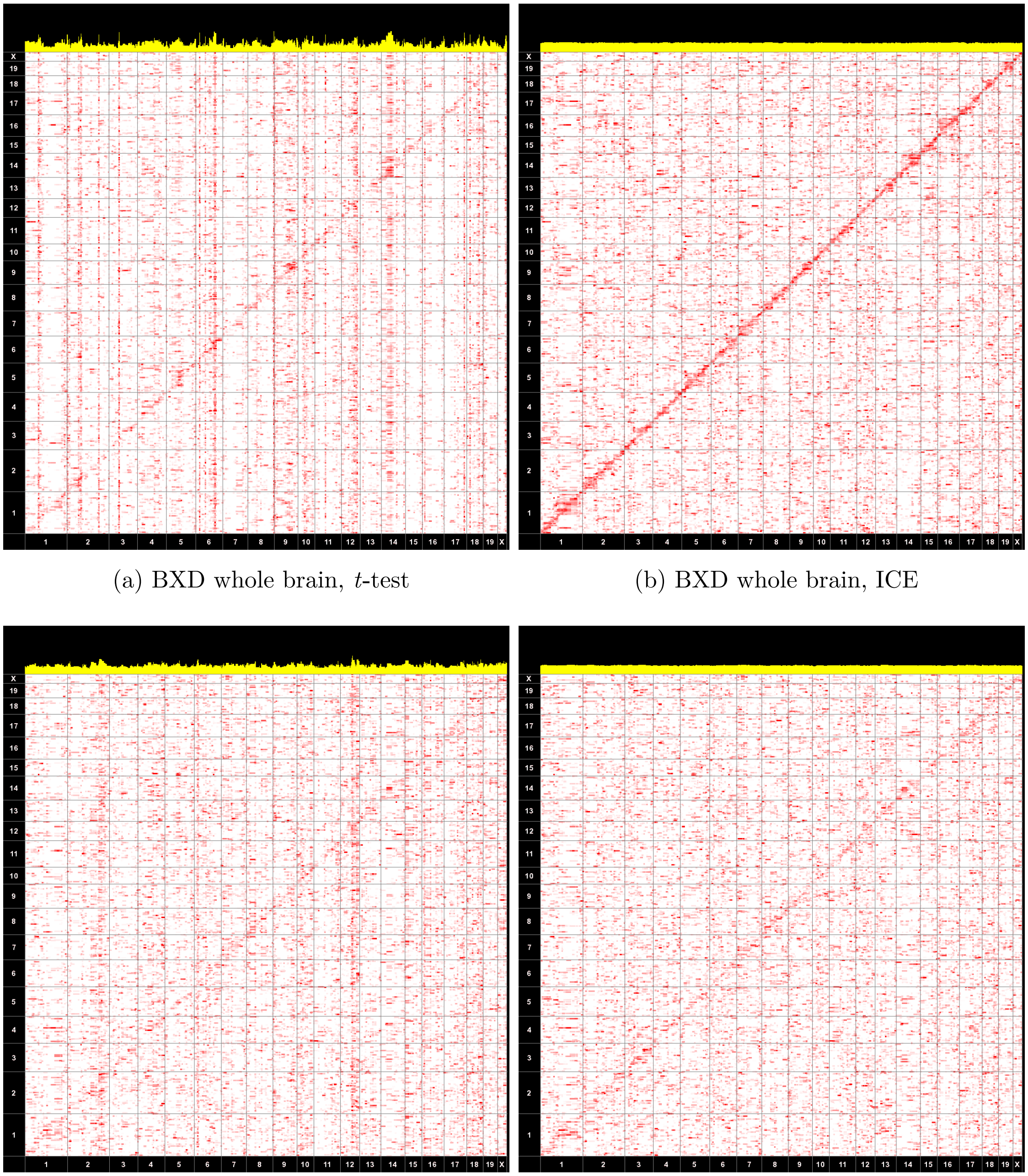
Figure 6.3: Genome wide correlation coefficients were computed for each pair of samples after standardizing each gene across the samples in (A) BXD RI whole brain and (B) HSC datasets, (C) yeast dataset and (D) HapMap dataset. The x-axis represents one subset of replicates and the y-axis represents the other. Each axis is ordered by strain (mouse), segregant (yeast) or individual name (human). Each diagonal element represents the strength of correlation between the replicated samples. Lower-triangular and upper-triangular regions show the correlation coefficients among two disjoint subsets of replicated samples.

6.2.3 Inter-sample Correlation Emended (ICE) eQTL mapping

Motivated by our observation of inter-sample correlation, we propose a new statistical method for identifying eQTLs based on a linear mixed model. Our method first estimates the pairwise correlation between samples which can be accurately estimated since there are thousands of probes in each sample. Instead of assuming independent random variations of expression levels between samples, our method assumes that a gene in pair of samples with globally correlated expression pattern is more likely to have similar expression values than a gene in a pair of globally uncorrelated samples. As a result, the variance component of expression levels at each gene is estimated as a mixture of inter-sample correlation and independent errors. A marker SNP is considered to be significantly associated with a transcript only if it predicts the expression beyond the level suggested by the inter-sample correlation (see Methods). Since spurious regulatory hotspots appear at marker SNPs consistent with the inter-sample correlation structure, accounting for this correlation in the null model significantly reduces these hotspots.

To demonstrate the effectiveness of our method, we first applied it to the simulated expression datasets presented in the previous section. Although the simulated datasets contain only *cis*-acting eQTLs, traditional eQTL mapping identified both the *cis*-acting band and spurious *trans*-regulatory bands (Figure 6.4a, 6.4c). The ICE eQTL map shows no *trans*-regulatory bands and a much stronger *cis*-regulatory band (Figure 6.4b, 6.4d). At a FDR level of 0.05, ICE eQTL mapping recovered 8.4% of the simulated whole brain *cis*-acting eQTLs, which was more than an one hundred fifty fold increase over the standard t-test and more than a three fold increase over SVA. These results illustrate that our method not only eliminates suspicious *trans*-regulatory bands but also has higher statistical power to recover real eQTLs that might be masked by the correlation structure.

To better understand the relative performance of random effects models versus fixed effects models on this problem, we analyzed our simulated data using the simple t-test, SVA and ICE eQTL mapping. At a SNP effect explaining 5% of phenotypic variations and a systematic confounding effect of 75%, we see that both the fixed



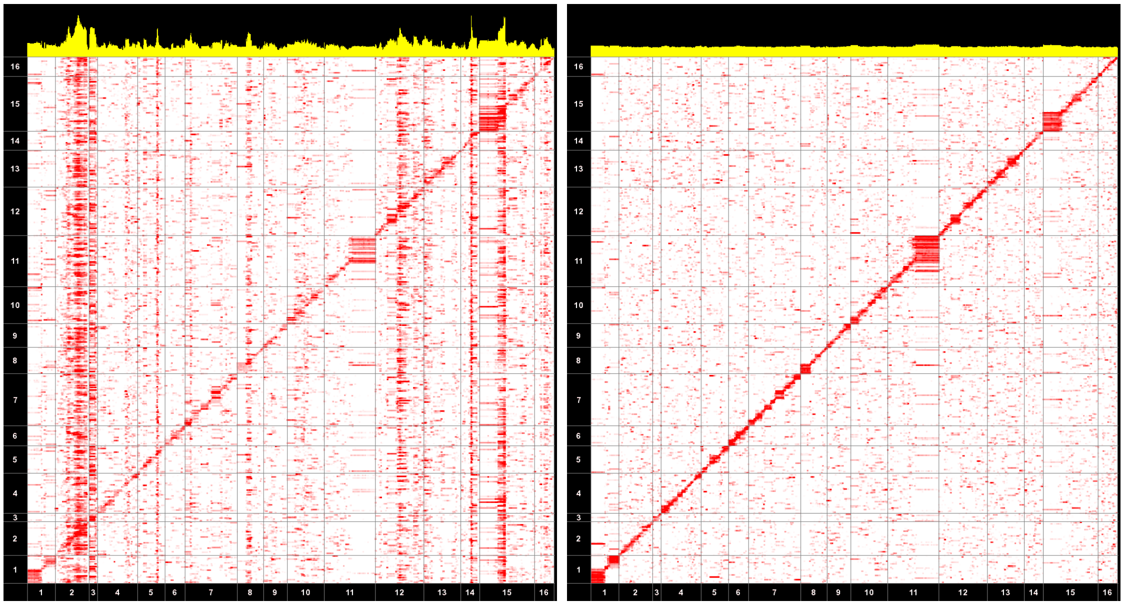
(a) BXD whole brain, *t*-test

(b) BXD whole brain, ICE

(c) BXD HSC, *t*-test

(d) BXD HSC, ICE

Figure 6.4: Refer to next page for details.



(e) Yeast, *t*-test

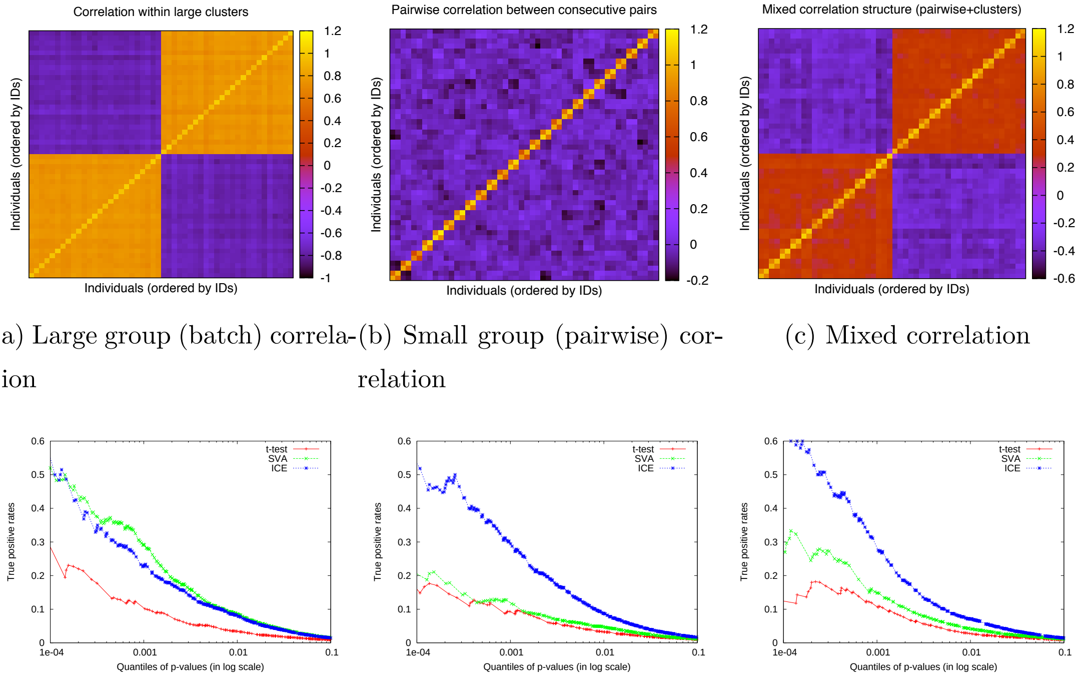
(f) Yeast, ICE

Figure 6.4: Simulated expression datasets are generated preserving inter-sample correlation structure of expressions, and traditional and ICE eQTL mapping is applied to the datasets. Six eQTL maps are plotted from (A,B) BXD recombinant inbred whole brain dataset, (C,D) BXD recombinant inbred HSC dataset, and (E,F) yeast dataset, using (A,C,E) standard *t*-test and (B,D,F) ICE eQTL mapping. There are almost identical pattern of regulatory hotspots observed in the original dataset with standard *t*-test, and they are eliminated after applying ICE eQTL mapping. Each SNP (horizontal) and probe (vertical) are sorted according to genomic positions, mapping each pixel to the linkage between a marker and a gene. The color of each pixel represents the strength of linkage signal. The yellow graph on the top represents the strength of *trans*-regulatory bands, quantified as the average log-p values at each SNP across all the genes. A total of 7,413 probes and 7,413 SNPs are simulated and mapped in two mouse datasets, and 2,956 probes and 2,956 SNPs in the yeast dataset.

effect model (SVA) and our random effects model (ICE eQTL) outperforms the simple t-test in discovering true positives when the samples are correlated in two batches (Figure 6.5d). When the samples are correlated within smaller groups of size two, we see that ICE eQTL outperforms SVA and the simple t-test (Figure 6.5e). In the mixture of large group and small group effects, which we expect to see in real datasets, we again see that ICE eQTL outperforms SVA and the simple t-test (Figure 6.5f). Under complex systematic confounding effects, because the fixed effects model requires a large number of confounding variables to completely correct for the confounding, it loses many degrees of freedom and the estimation of confounding variables becomes less accurate, resulting in the loss of statistical power.

It should be noted that ICE fundamentally differs from traditional mixed model methods such as MANOVA in that it estimates the variance component directly from the expression data. By leveraging the massive number of probes, ICE can accurately estimate the inter-sample correlation. Although we have used block-structure variance components as examples of systematic confounding in the above simulations, the estimated variance components typically have a much more complex structure. On the other hand, MANOVA uses predefined variance components which are usually block-structured to model random effects specific to groups of samples such as batches, cages, cohorts, or strains, depending on the context of the statistical analysis. Since these variance components are predefined, MANOVA can not correct for unknown confounding factors.

We next applied ICE eQTL mapping to real whole brain and HSC expression datasets from BXD RI mice. In both cases, ICE eQTL mapping eliminated the *trans*-regulatory bands while enhancing the *cis*-regulatory bands (Figure 6.2b,6.2d). The number of significant *cis*-acting eQTLs discovered increased dramatically. The enrichment in *cis*-acting eQTLs serves as a good indicator of the statistical power to identify differential expressions due to true genetic effects, even though some of the *cis*-associations might be due to polymorphic SNPs residing in the probe sequences[214]. For the whole brain dataset, ICE eQTL mapping identified nearly three times as many genes with *cis*-acting eQTLs (120) as the t-test (43) and 52% more than SVA (79) at a significance level of ten false positives per genome (Table 6.1, Figure 6.7). ICE eQTL mapping of the HSC dataset showed fewer significant



(a) Large group (batch) correlation

(b) Small group (pairwise) correlation

(c) Mixed correlation

(d) Large group correlation power

(e) Small group correlation power

(f) Mixed correlation power

Figure 6.5: Statistical power under systematic confounding from (A) large group (batch) correlation, (B) small group (pairwise) correlation and (C) a combination of batch and pairwise correlation structures in gene expression, using t-test, SVA, and ICE eQTL mapping, presented in (D), (E), and (F), respectively. All the p-values in the eQTL map are ranked and the fraction of true positives is plotted across different quantile of the p-values. For example, in among top 0.1% of p-values in (D), 27% of the signals are true positives with ICE-eQTL mapping, and 12% and 10% are true positives with SVA and *t*-test

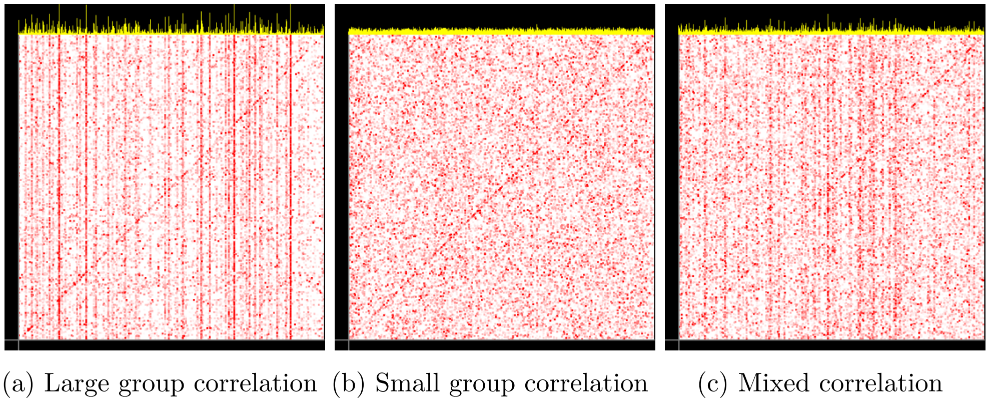

(a) Large group correlation (b) Small group correlation (c) Mixed correlation

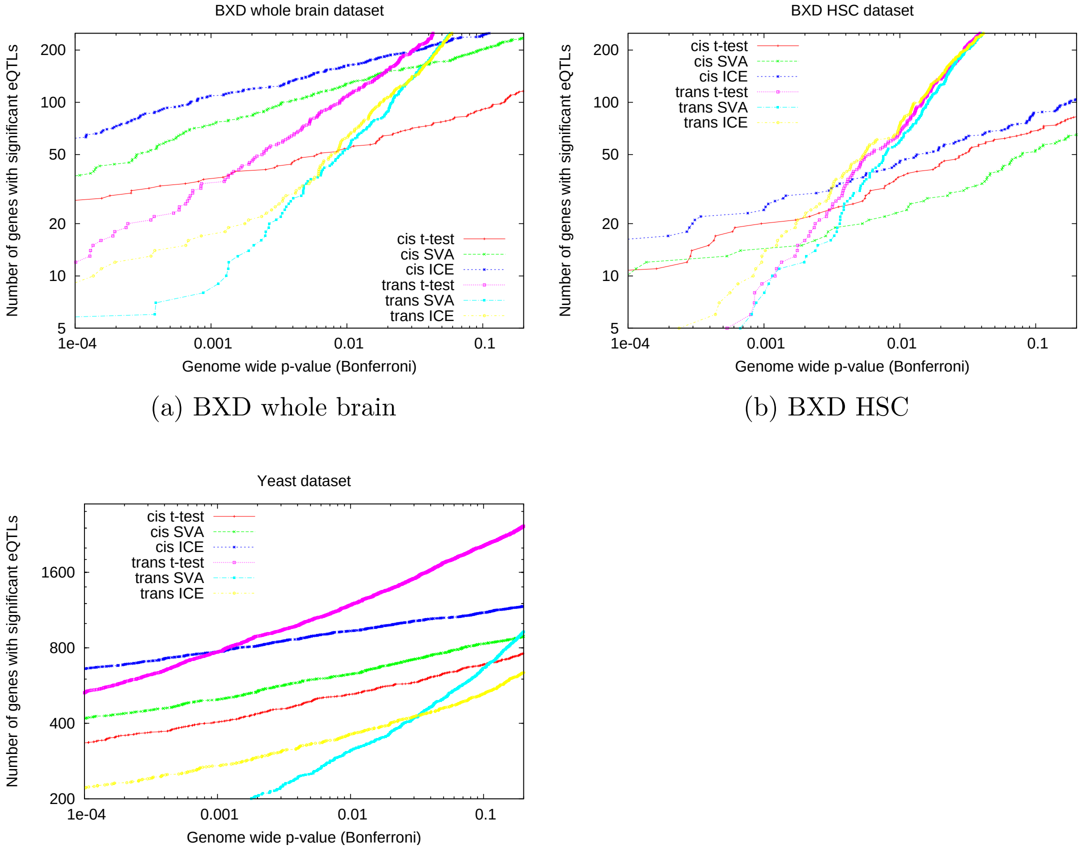
Figure 6.6: Different systematic confounding effects lead to different pattern of *trans*-regulatory bands. (A) large group (batch) correlation, (B) small group (pairwise) correlation and (C) a combination of pairwise and batch correlation structures. The combination of pairwise and batch correlation effects result in an eQTL map similar to those observed in real data sets.

eQTLs due to the reduced power of having a limited number of strains in the dataset. Nevertheless, similar to the whole brain results, our method consistently identified more genes with *cis*-acting eQTLs (23) than the *t*-test (19) and SVA (14). In this case SVA was outperformed by the *t*-test because having a large number of surrogate variables significantly reduced the degrees of freedom.

Similar to how the number of *cis* associations detected is a good measure of increased power to identify true genetic effects, another measure is the concordance of association between biologically replicated samples. We leveraged the replicated samples of the BXD datasets to measure the concordances of *cis* and *trans* eQTLs between replicates. After ordering the transcripts according to the strength of association for each replicated set, we plotted the concordances of *cis* and *trans* associations between the sets using CAT concordance plots[95]. In HSC and brain, both *cis* and *trans* eQTLs between replicates are significantly more concordant with ICE eQTL mapping (Figure 6.8a, 6.8b) than the *t*-test and SVA. Finally, we compared the results between whole brain and HSC datasets to see if the *trans*-acting eQTLs are replicable across different tissues. Previous studies have suggested that most *trans*-regulatory elements are tissue-specific because they have not been replicated in different tissues[34]. We postulate that most *trans*-regulatory elements were not

Table 6.1: Number of genes with significant *cis* and *trans* eQTLs in three datasets at different number of expected false positives. Total of 5,534 genes are tested in the yeast dataset, and 8,596 genes are tested in two BXD mouse datasets.

Dataset	type	method	E(false positives)		
			100	10	1
Yeast	*cis*	*t*-test	646	506	398
		SVA	817	617	490
		ICE	1118	935	757
	trans	*t*-test	1845	1067	738
		SVA	612	293	167
		ICE	539	363	263
BXD whole brain	*cis*	*t*-test	71	43	32
		SVA	148	79	42
		ICE	193	120	69
	trans	*t*-test	201	61	22
		SVA	82	11	5
		ICE	131	21	11
BXD HSC	*cis*	*t*-test	40	19	10
		SVA	28	14	12
		ICE	50	23	13
	trans	*t*-test	94	8	1
		SVA	102	11	1
		ICE	103	9	1



(a) BXD whole brain

(b) BXD HSC

(c) Yeast

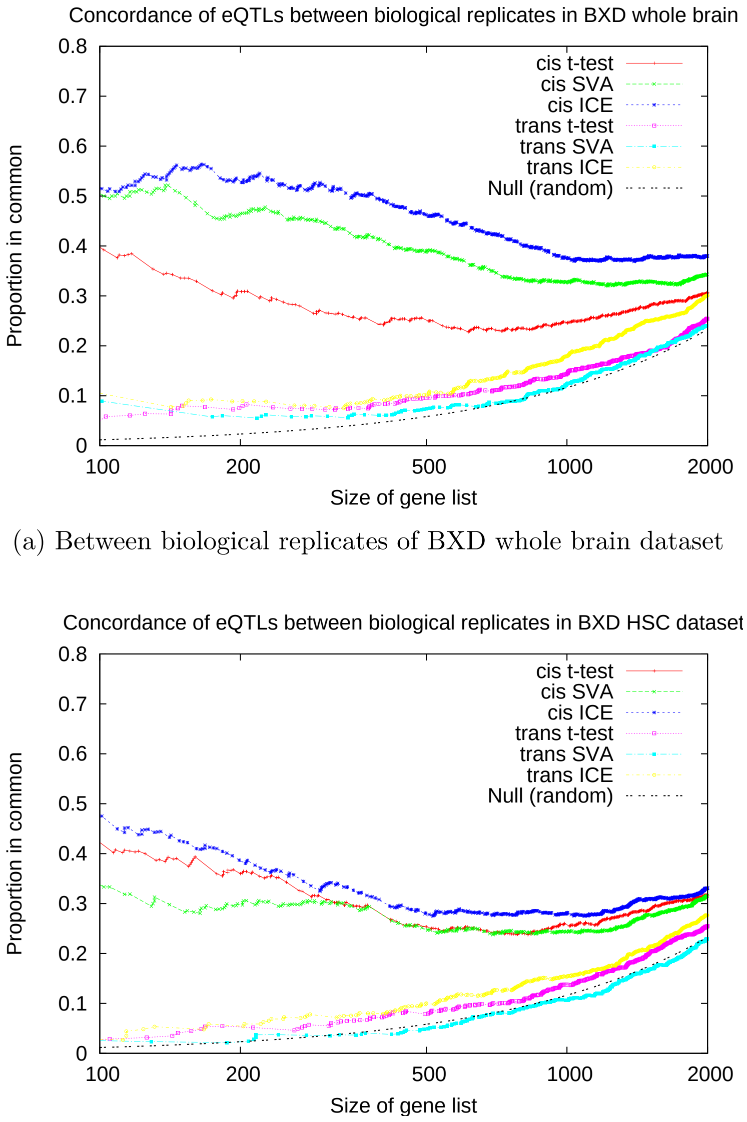
Figure 6.7: Number of genes with significant *cis*-acting and *trans*-acting eQTLs at various p-value thresholds in the (A) BXD whole brain dataset, (B) BXD HSC dataset, (C) yeast dataset. The horizontal axis represents the genome wide p-values of the most significant *cis* association or *trans* association for each gene, adjusted by Bonferroni correction using the number of non-redundant SNPs. The vertical axis represents the number of genes with significant associations at a given p-value threshold. In all three datasets, ICE eQTL mapping outperformed traditional eQTL mapping by consistently finding more *cis* associations at all p-value cutoffs. It also consistently found more real *trans* associations at low p-value cutoffs while fewer spurious *trans* associations at moderate p-value cutoffs in mouse.

replicated across tissues in previous studies because they are spurious associations caused by confounding factors. We ordered the transcripts based on the strength of *trans*-acting eQTLs for each dataset, and computed the Spearman's rank correlation between the two datasets. The p-values of correlation obtained from the standard t-test show a slightly negative correlation ($r = -0.012$), with a one-sided p-value of 0.857. However, the ICE eQTLs show much higher rank-concordance between the tissues with a p-value of 1.1×10^{-7} ($r = 0.056$). The CAT concordance plot[95] also shows that ICE eQTL mapping results are significantly more concordant between tissues (Figure 6.8c). This suggests that a significant fraction of real *trans*-acting eQTLs are not tissue-specific, and many of the previously identified *trans*-acting eQTLs did not replicate since they are largely confounded by spurious associations.

6.2.4 Some *trans*-regulatory bands in high quality datasets are likely to correspond to real genetic effects

In the previous sections, we demonstrated the ability of ICE eQTL mapping to obtain reliable and consistent associations in first generation mouse datasets that have been previously shown to have little reproducibility between independent data sets[160]. Second generation datasets collected using better protocols and newer expression chips such as Affymetrix M430v2 are of higher quality, resulting in much higher correlation between replicated samples than between unrelated pairs. Nevertheless, not only do these studies still suffer from moderate levels of inter-sample correlation between unrelated pairs (Figure 6.9), potentially genuine regulatory hotspots globally affect the expression levels and may confound the identification of gene-specific *cis* or *trans* associations[195, 118]. In this section, we analyze one of the classic genetical genomics dataset in yeast where global regulators have been previously reported by several studies[161, 66, 27, 242]. Under the confounding effects from such genuine regulatory hotspots, we demonstrate that ICE eQTL mapping identifies more *cis* and *trans* associations that are consistent between disjoint datasets.

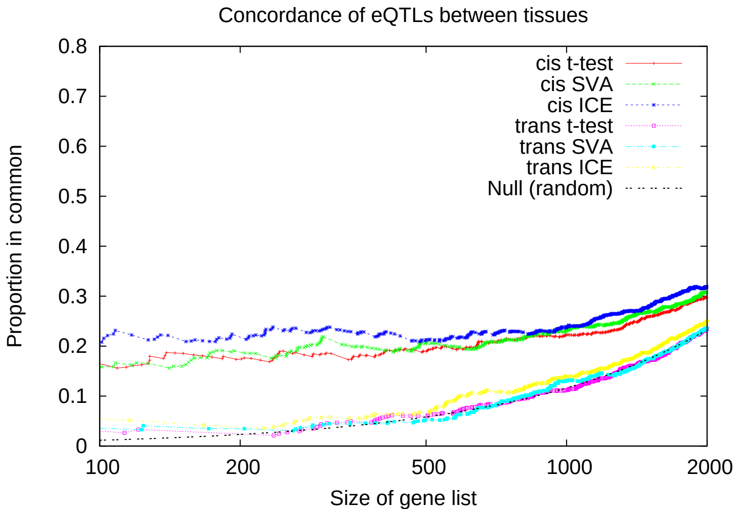
Yeast expression profiles and genotypes were collected from 112 segregants derived by crossing the lab isogeneic BY4716 strain with the wild RM11-1A isolate[27,



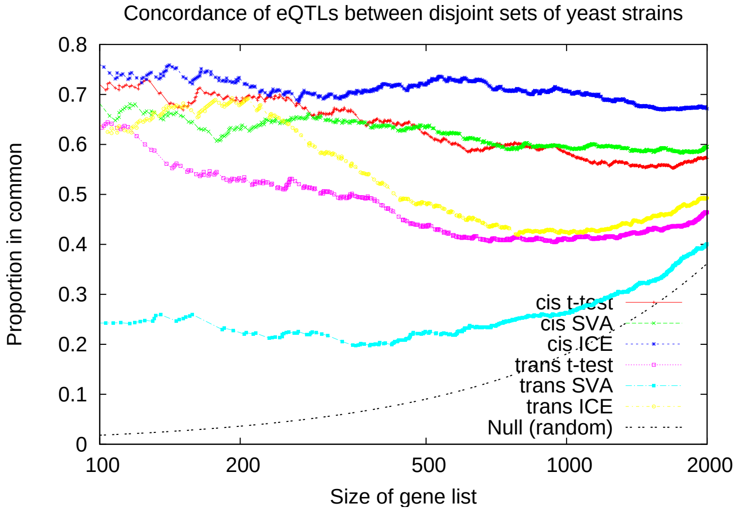
(a) Between biological replicates of BXD whole brain dataset

(b) Between biological replicates of BXD HSC dataset

Figure 6.8: Concordance of eQTLs between replicates, tissues, and populations. (Continuted on next page)

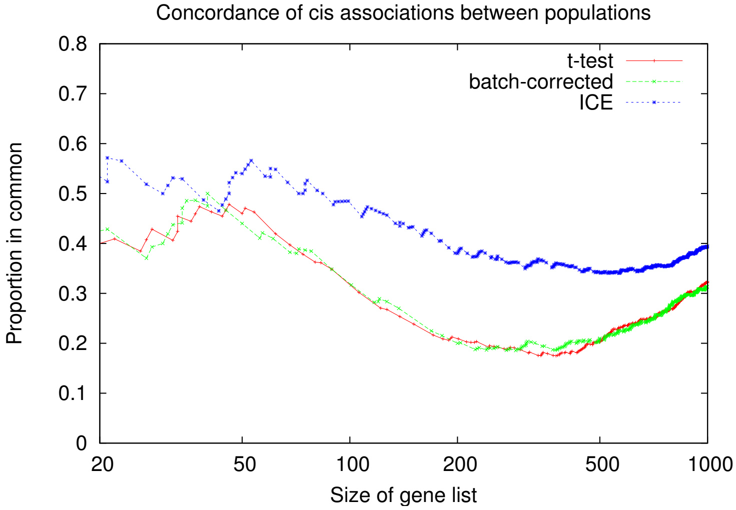


(c) Between different tissues (whole brain and HSC)



(d) Between different subsets of yeast strains

Figure 6.8: Concordance of eQTLs between replicates, tissues, and populations. (Continuted on next page)



(e) Between different HapMap populations (*cis*-only)

Figure 6.8: Each gene is ranked according to the strength of *cis* and *trans* associations, and the fraction of genes concordantly ranked within a certain rank is plotted in CAT format[95]. The horizontal axis represents the size of top ranking genes according to the strongest *cis* or *trans* eQTLs, and the vertical axis represents the number of genes concordantly ranked within the size divided by the size of the list. If the ranks are completely independent, the curve will follow the null distribution line. The ICE eQTL mapping shows higher concordance of both *cis* and *trans* eQTLs than the standard t-test across all four comparisons, including (A) between biological replicates of BXD whole brain dataset, (B) between biological replicates of BXD HSC dataset, (C) between whole brain and HSC of BXD datasets, (D) between disjoint subsets of yeast datasets, and (E) between European and Asian HapMap populations (*cis* associations only).

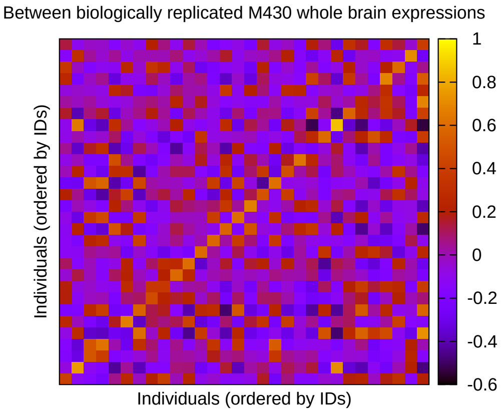
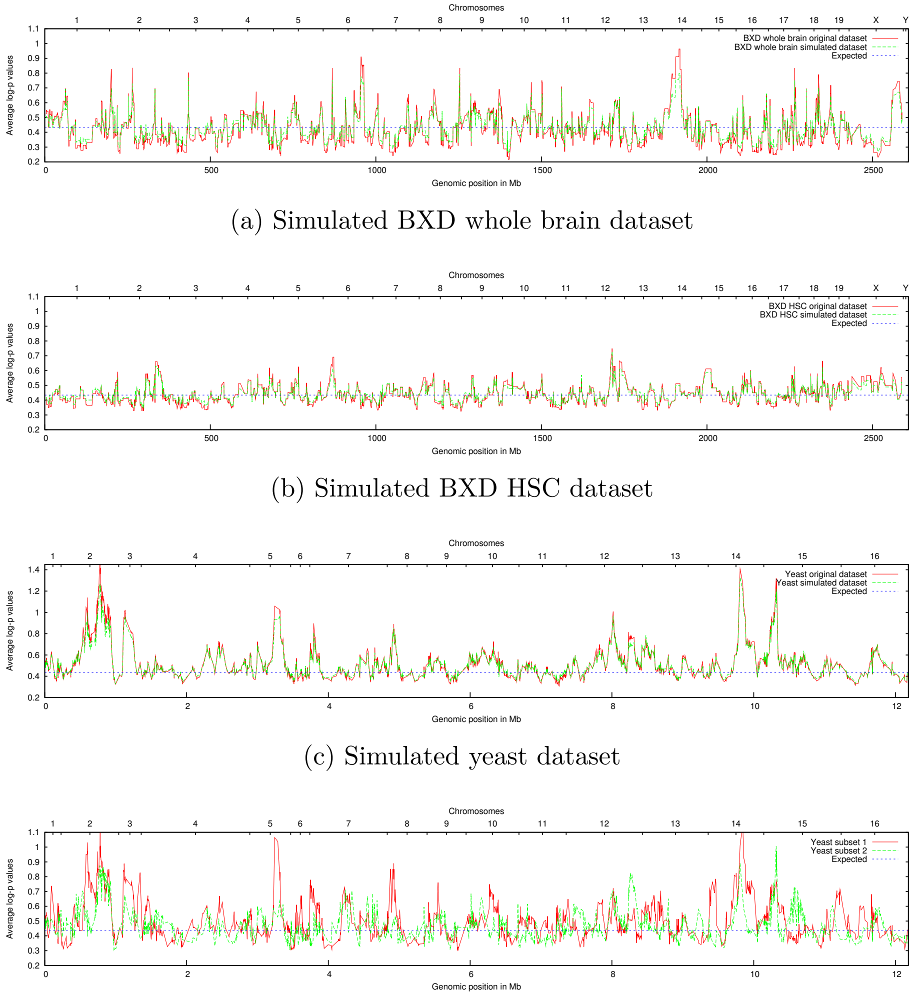


Figure 6.9: Genome wide correlation coefficients were computed for each pair of samples after standardizing each gene across the samples in BXD RI whole brain datasets using M430v2 arrays (Williams RW, unpublished).The x-axis represents one subset of replicates and the y-axis represents the other. Each axis is ordered by strain (mouse), and diagonal elements represent the strength of correlation between the replicated samples. Lower-triangular and upper-triangular regions show the correlation coefficients among two disjoint subsets of replicated samples, covering 30 biologically replicated strains. The diagonal is moderately pronounced suggesting that these dataset is of higher quality than the previous dataset.

25]. There are several differences between the yeast and the BXD RI datasets. First, the technological differences between the cDNA arrays used for yeast and the Affymetrix GeneChips used for mice may lead to very different patterns of systematic bias. Second, having a larger number of strains increases the number of eQTLs expected at the same significance level due to increased power. Third, since biological replicates are not available in the yeast dataset, it is difficult to determine whether the appearance of regulatory hotspots is caused by a systematic bias or a real genetic effect. Although the dye-swap results provide us with technical replicates, they are not suitable for verifying real hotspots because a dye-swap pair tend to have much stronger correlation than a pair of biological replicates due to smaller environmental or sampling biases between replicates than between unrelated samples (Figure 6.3c). This correlation may lead to the biased conclusion that most regulatory hotspots are highly reproducible.

After applying traditional eQTL mapping, we observed strong *trans*-regulatory bands, many of which are consistent with the inter-sample correlation structure (Figure 6.2e). However, unlike in the BXD recombinant inbred strains, several of the bands remained significant after performing permutation analysis. Three genomic regions in chromosome 2 (521 584kb), 14 (418 502kb), and 15 (171 193kb) had genome wide significant p-values of less than 0.05 with the most significant $p = 2 \times 10^{-4}$. This suggests that these *trans*-regulatory bands may be the result of real genetic effects rather than confounding effects. Recently, linkage studies of small-molecule drug response traits with the same set of yeast strains have shown that most of the QTL hotspots of these traits fall into the same genomic region where the bands occur[161]. Since the yeast dataset does not have biological replicates, we instead randomly divided the 112 segregants into two disjoint sets to perform eQTL mappings separately. If the regulatory hotspots are not real genetic effects, it would be unlikely that the same regulatory hotspot consistently appear between the disjoint sets. However, most of hotspots between the sets coincide, suggesting that they correspond to real genetic effects (Figure 6.10d).

We further tried to understand the biological importance of those significant *trans*-regulatory bands in the yeast data. We listed all 61 genes within 10kb of the significant regulatory hotspots and queried the set of genes in the Comprehensive



(a) Simulated BXD whole brain dataset

(b) Simulated BXD HSC dataset

(c) Simulated yeast dataset

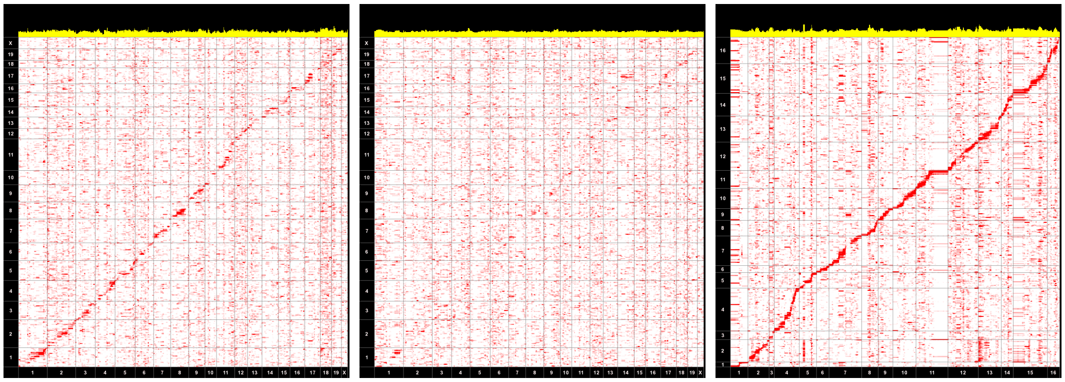
(d) Between disjoint subsets of yeast strains

Figure 6.10: Expression datasets were simulated preserving inter-sample correlation structure of expressions, and the regulatory hotspots computed from the simulated dataset were compared to those from the original dataset, in (A) the BXD whole brain dataset (B) the BXD HSC dataset, and (C) the yeast dataset. (D) is the comparison between disjoint subsets of yeast strains, showing that regulatory hotspots are reproduced with independent sets of samples. The simulated datasets almost perfectly reproduced the original regulatory hotspots, suggesting that the inter-sample correlation is the primary source of spurious regulatory hotspots. The horizontal axis is genomic positions of the markers in megabases, and the vertical axis is the strength of regulatory hotspots quantified as the average log-p values at each marker across all the genes. (See Figure 6.1).

Yeast Genome Database (CYGD)[81, 180]. Interestingly, the three regions with significant hotspots on chromosomes 2, 14 and 15, contain IRA1, RAS2, and IRA2, respectively. It has been known that IRA1 and IRA2 genes negatively regulate the RAS2 protein activation state from multiple studies[201, 42]. The probability of three genes appearing in a random set of 61 genes is 1.3×10^{-6}. There are also other genes that encode small GTP-binding proteins of the RAS superfamily such as ARL1, RHO2 and YPT53 near the significant regulatory hotspots. It is possible that the variations in those regions may change the mRNA levels of a large number of genes by perturbing the RAS GTP-binding signal transduction pathway. However, under this interpretation, it is not certain why a portion of mRNA levels are up-regulated while others are down-regulated by the same variant in those regulatory hotspots. In addition, a recent study suggest that MKT1 is the causal regulator that may be responsible for the regulatory hotspot in chromosome 14[242].

Even though many *trans*-regulatory bands in yeast are likely to be real genetic effect, they globally influence the expression levels and may seriously confound the identification of gene-specific *cis* or *trans* associations[195, 118], resulting in the loss of power to identify real *cis* and *trans* associations. Correcting for the inter-sample correlation induced by genuine regulatory hotspots may eliminate true *trans*-regulatory bands, but also can reveal many true regulatory signals obscured by the hotspots. We compared the power of different eQTL mapping methods at identifying true genetic effects by randomly partitioning the dataset as described above. In each partition, the transcripts are ordered by the strength of *cis* or *trans* associations, and the concordance between the disjointly partitioned datasets are illustrated using the CAT plot (Figure 6.8d). The results show that ICE eQTL mapping have higher concordance than the *t*-test and SVA both for *cis*-acting and *trans*-acting eQTLs, despite the loss of true regulatory hotspots.

We applied ICE eQTL mapping to the entire yeast dataset and observed that the *trans*-regulatory bands are eliminated while the genes with significant *cis* associations is nearly doubled (Figure 6.2f, Table 6.1, Figure 6.7). The number of genes with *trans*-acting eQTLs are significantly reduced using ICE eQTL mapping due to eliminated regulatory bands, but many new *trans*-acting genes that have not been identified by the *t*-test are discovered. For example, among the 363 significant



(a) BXD whole brain, SVA (b) BXD HSC, SVA (c) Yeast, SVA

Figure 6.11: Genome wide eQTL maps from (A) BXD recombinant inbred whole brain dataset, (B) BXD recombinant inbred HSC dataset, and (C) yeast dataset, using Surrogate Variable Analysis[118]. The SNPs (horizontal) and probes (vertical) are sorted according to genomic position, mapping each pixel to the linkage between a marker and a gene. The color of each pixel represents the strength of the linkage signal, with red being the strongest signal and white being the weakest. The yellow graph on top represents the strength of *trans*-regulatory bands, quantified as the average log-p values at each SNP across all genes. Unlike ICE eQTL mapping, some vertical *trans*-regulatory bands remain using SVA. But even more problematic is the elimination of *cis*-associations on chromosome 12 of the yeast data due to over correction. A total of 8,596 probes and 7,413 SNPs are mapped in the two mouse datasets, and 5,534 probes and 2,956 SNPs in the yeast dataset.

trans-acting genes identified by ICE eQTL mapping at the significance of ten false positives per genome, 25% (89) of them are not identified by the *t*-test at the same threshold. On the contrary, only 7% (35) of the 506 significant *cis*-associated genes identified by the *t*-test are not identified by ICE eQTL mapping at the same significance level. ICE eQTL mapping outperforms SVA in discovering *cis*-acting eQTLs across different significance thresholds. For *trans*-acting eQTLs, ICE identifies larger number of eQTLs for conservative significance threshold of FDR less than 0.1, while SVA identifies more eQTLs for higher thresholds. This may be due to the effects from the moderate regulatory hotspots that have not been captured by surrogate variables as appeared in Figure 6.11c. In this dataset, SVA appears to over correct for the *trans*-regulatory bands and eliminated even the the *cis*-acting eQTLs in the middle of chromosome 12. (Figure 6.11c).

6.2.5 Correcting for confounding effects in human lymphoblastoid cell line expression

Finally, we applied our method to a human genetical genomics study of the HapMap individuals where the goal was to determine whether differentially expressed genes between CEU and JPT+CHB populations are caused by allelic or population differences. It is known that the HapMap expression experiments were conducted on different dates for the CEU and JPT+CHB populations and the problems introduced by this batch effect have recently been addressed [3]. While the original paper claimed that 26% of genes are differentially expressed between European and Asian samples at a genome wide Sidak-corrected $p < 0.05$, none of them were identified to be significant after controlling for the year in which the sample was processed. In fact, with respect to this batch effect, 28% of the genes were differentially expressed.

We applied ICE eQTL mapping to identify differentially expressed genes. Our method is able to control for the inflated false positives of differentially expressed genes without the prior knowledge of batch information. The p-value distribution appears to be almost uniform (Figure 6.12). Spielman et. al. provided POMZP3 as an example of a differentially expressed gene between the two populations to demonstrate that not all of their findings were false positives[3]. The gene was associated with a *cis*-regulatory SNP, whose allele frequency was significantly different between the two populations. We examined how strongly the POMZP3 gene is differentially expressed using three different methods. Without correcting for confounding effects, the gene is significant at a p-value of 1.91×10^{-6}. However, since numerous other genes are identified to be significant, the strength of the signal is ranked only 943th (23.4%) out of 4,030 genes. After explicitly correcting for the year of the experiment using ComBat[100], the gene is no longer significant at a p-value of 0.309. However, the signal is ranked relatively high, 434th (10.8%) out of 4,030 genes. After correcting using SVA, it is ranked only 1992th (49.4%) with a p-value of 0.352. After correcting for the inter-sample correlation pattern using our method, the gene is ranked 6th (0.15%) at a p-value of 3.1×10^{-4}. Using the same approach, we examined the top 5 genes among the 11 genes reported as differentially expressed genes with concordant *cis*-eQTLs between populations. Correcting for inter-sample correlation consistently outperformed the other methods at identifying those genes as differentially expressed with higher ranks (Table 6.2).

We next performed ICE eQTL mapping and compared the *cis* associations with those obtained from *t*-test based mapping and batch-corrected mapping. We analyzed a total of 3942 genes within 500kb of at least one of the 2 million HapMap SNPs. In both CEU and JPT+CHB populations, the number of genes with *cis* associations increased significantly with our method (Figure 6.14a, 6.14b). eQTL mapping performed after correcting for the known batch effect using ComBat did not significantly outperform the *t*-test. Furthermore, the concordance of *cis*-acting genes between populations significantly increased as well, suggesting that ICE association mapping has higher power to identify real genetic effects (Figure 6.8e).

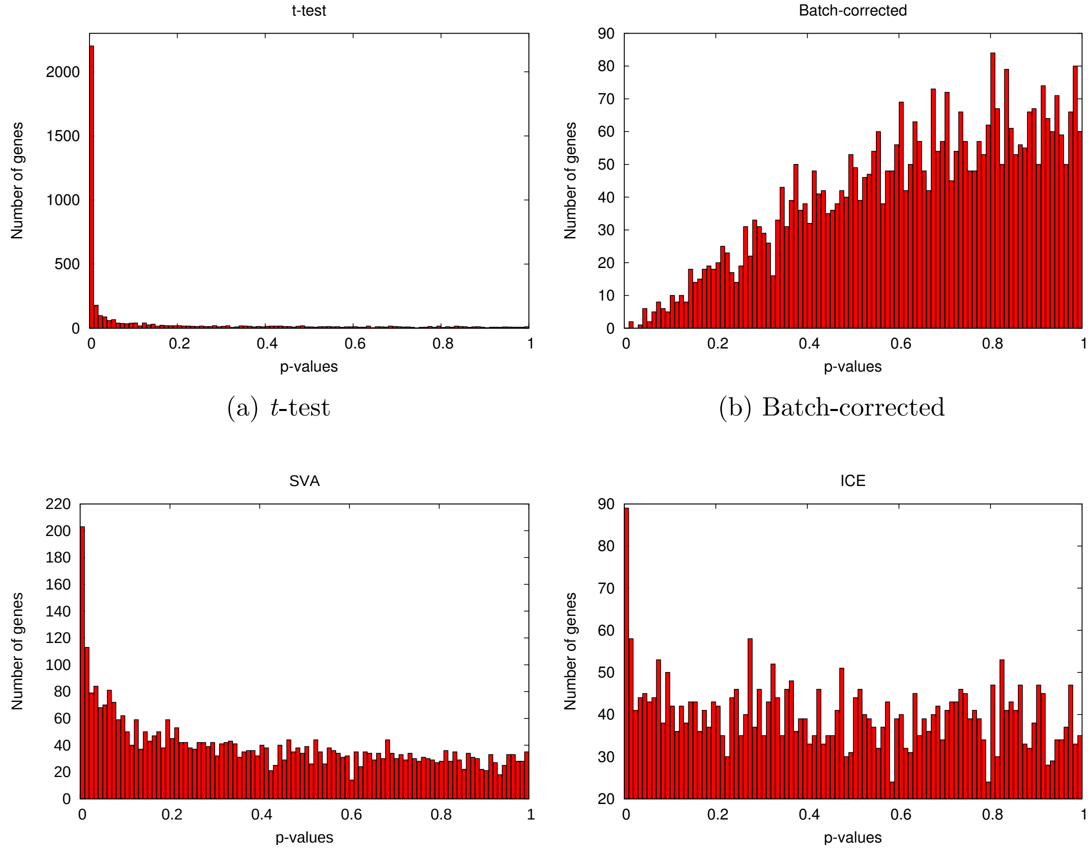
Finally, we applied our method to identify differentially expressed genes with evidence of concordant *cis*-acting SNP between populations. We applied a more stringent threshold than previous studies[3] by requiring the *cis*-acting SNP to have a genome wide p-value of less than 2.5×10^{-8} in at least one population and a strong p-value of less than 10^{-5} in the other after Bonferroni correction. In addition, we required the minor allele frequency of the SNP to differ by at least 0.1, and the strength of differential expression to be ranked in the top 10% of all genes. Using these stringent criteria, only two genes are identified using the *t*-test, and three genes are identified after explicitly correcting for the batch effect. On the other hand, ICE association mapping successfully identified 10 differentially expressed genes including four previously unreported (Table 6.3).

6.2.6 Comparison with previous methods

A key difference of ICE association mapping from the previous methods using singular value decomposition[5, 118] is that previous methods project the systematic confounding onto several distinct single dimensional vectors as fixed effects while ICE association mapping directly incorporates the pairwise correlation as random effects into the statistical model. For previously known confounding variables such as batch effects, both methods can incorporate them as fixed effects in the statistical model. While the singular value decomposition methods infers a number of confounding

Table 6.2: The relative strengths of differential expression in two different populations of the top 5 genes reported by Spielman et. al.[3], with significant *cis*-eQTLs and significant MAF differences between the populations

Gene	Quantiles (Rank)								MAF of *cis*-acting SNP	
Name	*t*-test		batch-corrected		SVA		ICE		CEU	CHB+JPT
UGT2B17	1.4%	(58)	7.2%	(292)	0.024%	(1)	0.024%	(1)	0.68	0.15
POMZP3	23%	(943)	11%	(434)	49%	(1992)	0.15%	(6)	0.28	0.06
PEX6	21%	(866)	19%	(773)	3.6%	(147)	0.84%	(34)	0.59	0.36
PSPHL	28%	(1136)	6.2%	(249)	31%	(1268)	1.2%	(47)	0.17	0.45
CSTB	14%	(559)	6.9%	(277)	10.6%	(429)	1.2%	(48)	0.63	0.19



(a) *t*-test

(b) Batch-corrected

(c) SVA-corrected

(d) ICE-corrected

Figure 6.12: Differential expressions between two HapMap populations are tested using three different methods: (A) t-test using the original uncorrected expressions, (B) Batch correction by the year of experiment, (C) Surrogate Variable Analysis, and (D) ICE association mapping. The horizontal axis represents the p-values of differential expression for each gene using three different methods, and the vertical axis is the frequency of each interval out of 4,030 genes.

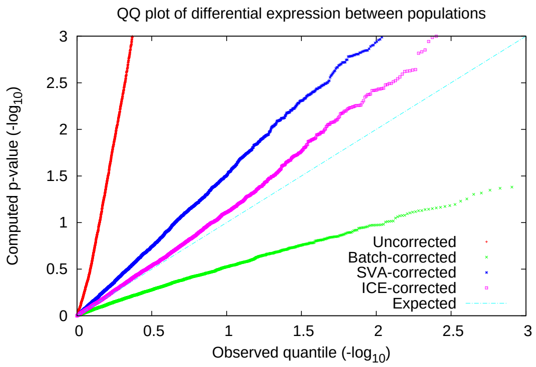
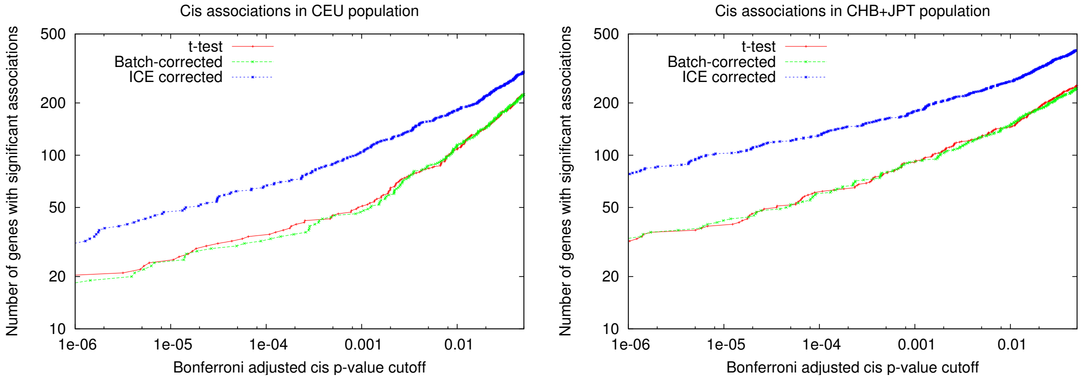


Figure 6.13: Q-Q plot of differential expression between two HapMap populations tested using the standard *t*-test, batch-corrected *t*-test, and ICE *t*-test. The horizontal axis is the negative logarithm of the quantiles of the observed p-values, and the vertical axis represents the negative log p-values of differential expression.



(a) CEU

(b) CHB+JPT

Figure 6.14: Number of genes with significant *cis* associations at various p-value threshold for (A) European-derived and (B) Asian-derived populations among HapMap samples. The standard *t*-test without correction, *t*-test after correcting for the previously known batch effect (year of experiment), and ICE association mapping are compared. The horizontal axis represents the Bonferroni adjusted p-values of the most significant *cis* association for each gene, and the vertical axis represents the number of genes with significant associations at a given p-value threshold. In both populations, ICE eQTL mapping outperformed traditional eQTL mapping by consistently finding more *cis* associations at all p-value cutoffs. Batch corrected eQTL mapping does not outperform traditional eQTL mapping.

Table 6.3: List of differentially expressed genes with concordant *cis* associations across populations. The bold genes have (1) *cis* associations with $p < 2.5 \times 10^{-8}$ in at least one population, (2) *cis* associations with $p < 10^{-5}$ at the same SNP in the other population, (3) the difference of minor allele frequency (MAF) between populations is greater than 0.1, and (4) the strength of differential expressions between populations ranked among the top 10% out of 4,030 genes, using three different methods. Only two and three genes pass the criteria using the standard t-test and t-test after batch correction, respectively. 10 genes are identified to be significant using ICE eQTL approach, including four previously unreported differentially expressed genes. The genes are ordered by the difference of MAF between populations.

* : Previously identified by Spielman et. al.[3]

† : If multiple methods identify the same gene with different strongest SNP, the SNP is selected according to the method providing the strongest rank in differential expression.

** : Although two methods have ranked the gene as among the top 10% in terms of differential expressions, no *cis* associations pass the p-value threshold described above.

Gene	Quantiles (Rank)						rsID† of	MAF†		*cis* eQTL p-values	
Name	t-test		batch-corrected		ICE		*cis* SNP	CEU	ASI	CEU	ASI
TECA1	21%	(840)	30%	(1227)	**0.55%**	**(22)**	rs11996666	0.12	0.67	3×10^{-7}	2×10^{-10}
UGT2B17*	**1.4%**	**(58)**	**7.2%**	**(292)**	**0.024%**	**(1)**	rs3100645	0.68	0.15	1×10^{-11}	9×10^{-37}
CSTB*	14%	(559)	**6.9%**	**(277)**	**1.2%**	**(48)**	rs4818868	0.38	0.82	2×10^{-8}	9×10^{-10}
DFNA5	7.0%**	(284)	2.4%**	(96)	**2.6%**	**(106)**	rs754555	0.13	0.49	9×10^{-8}	1×10^{-13}
PEX6*	21%	(866)	19%	(773)	**0.84%**	**(34)**	rs2296804	0.59	0.36	2×10^{-15}	8×10^{-11}
POMZP3*	23%	(943)	11%	(434)	**0.15%**	**(6)**	rs17718068	0.29	0.06	3×10^{-22}	1×10^{-19}
RABGGTA	50%	(2013)	24%	(985)	**3.8%**	**(152)**	rs3940231	0.35	0.58	2×10^{-9}	4×10^{-17}
RRM1	16%	(633)	12%	(487)	**5.8%**	**(232)**	rs10500594	0.91	0.75	1×10^{-8}	1×10^{-18}
AP3S2*	**6.5%**	**(262)**	**5.8%**	**(235)**	10%	(414)	rs9920421	0.31	0.16	8×10^{-9}	2×10^{-6}
NUBP2*	10%	(416)	11%	(452)	**6.4%**	**(259)**	rs344359	0.18	0.32	3×10^{-9}	3×10^{-11}
DNAJC15*	19%	(771)	10%	(405)	**3.4%**	**(136)**	rs9594865	0.76	0.90	2×10^{-9}	1×10^{-11}

factors strongly affecting the variations in expression, those with relatively moderate effects may remain uncorrected. Our mixed model approach ICE does not suffer from this shortcoming since it does not explicitly infer a limited number of confounding variables. Instead, the confounding effects from various unknown sources are assumed to be intrinsically imprinted in the expression profiles, specifically as inter-sample correlation.

The simulation results under various types of confounding effects we presented above (Figure 6.5) are largely consistent with those seen with applying random models versus fixed models on the related problem of correcting for population structure in association studies. Previous studies showed that the random effects model corrects for the heterogeneous population structure better than a fixed effects model based on principal component analysis such as EIGENSTRAT for model organisms association mapping[235, 240, 103]. Although EIGENSTRAT can robustly correct for population structure in human association mapping where an admixture model assuming a small number of distinct ancestral populations accurately describes the structure of the data[170, 168], in the model organism association mapping involving multi-level population structure, such methods only partially capture the population structure resulting in an inflated number of false positives[8]. Similarly, we see that fixed effects models can effectively correct for inter-sample correlation where there's relatively simple confounding structure such as batch effects while the random effects model performs much better when we have more complex and multi-leveled confounding structures we see in simulated and real data sets.

A second intuition why mixed models outperform SVD methods in this case is that a large number of surrogate variables or eigengenes are required to capture the complex expression heterogeneity, resulting in a significant increase in the degrees of freedom which affects the statistical power. These effects can be substantial especially for those datasets with a limited number of samples. For example, in the HSC dataset containing only 22 strains, SVA was shown to be even less powerful than the *t*-test in identifying *cis*-acting eQTLs.

Both approaches have potential risk of over correcting true genetic effects, especially for those *trans*-acting eQTLs corresponding to true regulatory hotspots. The concordance plots between replicated and disjointly partitioned datasets consistently show that our ICE association mapping provides higher concordance than standard t-test at identifying both *cis*-acting and *trans*-acting eQTLs, while SVA method consistently shows lower concordance of *trans*-acting eQTLs than standard t-test (Figure 6.8). Although some genuine regulatory hotspots may have been eliminated using ICE eQTL mapping particularly in the yeast dataset, we were able to identify some of the regulatory hotspots as significant through the analysis of replicates and our SNP permutation approach.

In terms of the computational cost, the running time of ICE association mapping is twice as fast as SVA using Efficient Mixed Model Association (EMMA)[103].

6.3 Discussion

We have proposed a novel statistical method, Inter-sample Correlation Emended (ICE) eQTL mapping which corrects for the systematic confounding effects inherent in expression datasets. Using the first-generation RI mouse expression dataset where the problem of systematic confounding effects has already been documented, we have demonstrated that most *trans*-regulatory bands in the dataset correspond to spurious regulatory hotspots through the analysis of biological replicates and the permutation analysis. Using simulated data that preserves the inter-sample expression correlation structure, we have shown that the inter-sample correlation effectively characterize the systematic biases that are responsible for the spurious associations. Using the same methods, we demonstrated that a number of *trans*-regulatory bands in yeast correspond to genetic variation in global regulators.

From both differential expression analysis in human and association analysis in recombinant inbred mice and yeast, we conclude that our method is more robust at correcting for systematic confounding factors than previous methods including an explicit batch correction method, ComBat[100] and an automated method that corrects for unknown confounding factors, Surrogate Variable Analysis[118]. Not only did ICE eQTL mapping identify more *cis*-acting eQTLs than both methods, those identified *cis*-acting and *trans*-acting eQTLs also showed higher concordance between replicated datasets (BXD RI strains), different tissues (BXD RI strains), and disjoint subsets (yeast). These results suggest that our method has higher power to identify associations corresponding to real genetic effects.

Our results also highlight the importance of obtaining independent replicates of expression measurements and the utility of these replicates for analyzing and validating eQTLs. We have shown that different strategies for obtaining replicates have profound effects on the correlation structure between replicates. Technical replicates obtained by either performing a dye-swap (Figure 6.3c) or running multiple expression arrays in the same sample (Figure 6.3d) exhibit much higher correlation between replicate pairs than full biological replicates (Figures 6.3a and 6.3b). We suspect that confounding factors in the sample preparation are largely responsible for the higher pairwise correlation observed among technical replicates reducing their utility in analysis and validation. Preparing multiple samples from the same individual can help reduce the effect of these confounding factors. In many eQTL studies, it is possible to independently measure expression from genetically identical individuals which can further reduce the effects of these confounding factors.

6.4 Materials and methods

6.4.1 Gene expression data and genetic maps

We obtained the yeast expression dataset over 112 segregants across 6,216 probes from the GEO database with accession number GSE1990[25]. Each of them has two replicates, and the values are represented as the log ratio between the expression and the average expression of the reference(BY) strains. 5,534 genes with validated genomic annotations were mapped onto the genome to draw the genome wide eQTL maps. For BXD RI datasets, we obtained the hematopoetic stem cell (HSC) data from the GEO database with accession number GSE2031, and the whole brain dataset by request from the authors. Both datasets use the Affymetrix U74Av2 GeneChip platform and contain 12,422 probes. 8,596 probes were mapped onto NCBI build 34 version of the mouse genome using refSeq to draw the eQTL maps. The HSC dataset contains the expression data over 22 strains with duplicates for each strain, and the whole brain dataset contains 64 samples over 28 strains, varying one to four measurements per strain. The second generation whole brain dataset using M430v2 arrays downloaded from GeneNetwork (`http://www.genenetwork.org`) contains expression profiles over 45102 probes across 30 BXD RI strains with up to 6 replicates per strain. Their expression values were normalized using RMA[94].

We obtained the human lymphoblastoid cell line expression data over the HapMap individuals from the GEO database with accession number GSE5859 [192]. There are a total of 141 samples, 60 from CEU and 81 from JPT+CHB. Although the Affymetrix Genome Focus Array contains 8,500 annotated genes, we focused on the 4,030 that are expressed in lymphoblastoid cell lines defined the same way as Spielman, et. al [192].

The genetic map of 2,956 SNPs of yeast segregants were obtained by request from the authors. The BXD RI SNPs were obtained from the Wellcome Trust Center, containing 13,270 SNPs over the genome, of which 7,413 SNPs are polymorphic between the two parental strains. Sixty-one and 25 SNPs with minor allele frequency less than 5% were discarded in the HSC and the whole brain datasets respectively. A very small portion of heterogeneous alleles in the RI strains were assumed to have additive effects, and the missing SNPs were not resolved. The genetic map for the HapMap samples were obtained using release 22 of the human HapMap[93]. We examined a total of 3942 genes that are within 500kb of at least one of the 2 million HapMap SNPs.

6.4.2 Traditional eQTL mapping and genome wide eQTL maps

Traditional eQTL mapping was performed by taking the average of expression values of each strain and performing *t*-test between each marker SNP and each transcript. The eQTL mapping using either of the replicates was performed in the same way except that the samples were divided into two disjoint sets of expression experiments by randomly picking one of the replicates. For the seven strains that have only one measurement in the BXD whole brain dataset, they were included in both sets. Missing SNPs or missing expression values were excluded in the test only for the corresponding marker-transcript pair, and the p-value was obtained from the asymptotic *t*-distribution.

Genome-wide eQTL maps were plotted based on the relative genomic positions of each transcripts and marker SNPs. Since the previously suggested transcriptome map[34] may create artificial horizontal bands due to non-uniform genomic densities of the probes, we used an eQTL map based on the relative positions of markers and probes, simply corresponding each marker-transcript association to one pixel. The degree of redness of each pixel is proportional to the log p-values.

6.4.3 Explicit batch effect correction and Surrogate Variable Analysis

We explicitly corrected for known batch effects using the ComBat R package [100]. We used the default settings and the batch corrected expression levels were used to perform traditional eQTL mapping using the t-test.

For surrogate variable analysis, we used SVA R package downloaded from the author's website, identifying surrogate variables ignoring the genotype data as suggested [118]. The p-values are obtained using a linear model after correcting for the surrogate variables.

6.4.4 Genome wide inter-sample correlation

An inter-sample correlation matrix from a expression dataset is generated as follows. Let Y be a $m \times n$ expression matrix with n arrays for m genes, then the inter-sample correlation matrix is generated as follows. Let μ_i, σ_i be the mean and standard deviation of expression values of i-th genes, $(Y_{i1}, Y_{i2}, \cdots, Y_{in}$. Let Z be a $m \times n$ matrix with each element $Z_{ij} = (Y_{ij} - \mu_i)/\sigma_i$, then the inter-sample correlation matrix defined as the covariance matrix of Z. It should be noted that we used the covariance matrix $K = \mathrm{Cov}(Z)$ instead of the correlation matrix because the variances are not homogeneous across the strains. Such heterogeneous distribution of variances can be an additional source of systematic confounding but is not emphasized in the main text for the sake of simplicity.

In order to visually compare the consistency between replicated pairs and unrelated pairs, we used the correlation matrix of Z for each replicated dataset because the correlation matrix can be more intuitively to understood than the covariance matrix. Each diagonal element represents the pairwise correlation of a replicated pair. In the upper-triangular region, the correlations between unrelated pairs in one subset of replicates was computed and visualized. In the lower-triangular region, the other subset was computed and visualized. The seven strains without replicates in the BXD whole brain dataset were omitted in the heatmap visualization.

6.4.5 Simulation studies

The eQTL mapping with permuted SNPs was performed by permuting the SNPs across the individuals, thereby preserving the correlation between each pair of SNPs. For generating the simulated expression data preserving the genome wide correlation pattern, we assumed the following generalized linear model.

$$\mathbf{y} = \alpha\mathbf{g} + \mathbf{u} \tag{6.1}$$

where $\mathbf{g}$ represents SNPs encoded by 0 and 1, and $\mathbf{u}$ is a multivariate normal random variable sampled from $N(0, K)$. $K = \mathrm{Cov}(Z)$ is the inter-sample correlation matrix defined in the previous section. α is set so that *cis*-regulatory effects account for 4% of the phenotypic variation explained by each causal SNP. The number of significant *cis*-eQTLs are counted using a conservative FDR estimate with $\pi_0 = 1$, considering only the SNP and simulated gene pair where the *cis*-regulatory effects are simulated.

For the comparison of various systematic confounding effects, we simulated expression datasets of 500 genes over 50 samples from three different inter-sample correlation structure described in Figure 6.5, with 75% of phenotypic variations explained by the confounding effects using a multivariate normal distribution[103]. We generated a random SNP of minor allele frequency of 0.3 for each gene, and added a SNP effect explaining 5% of phenotypic variation. We performed eQTL mapping using t-test, SVA, and ICE eQTL mapping for all 500 × 500 SNP-gene pairs, and computed the true positive rates at each p-value cutoff.

6.4.6 Variance component test

We applied the following variance component model to assess the statistical significance of each association in the presence of genome wide correlation.

$$\mathbf{y} = X\beta + \mathbf{u} + \mathbf{e} \tag{6.2}$$

where X, β is the fixed effects of known confounding variables and their coefficients, and $\mathbf{u} \sim N(0, \sigma_g^2 K)$ and $\mathbf{e} \sim N(0, \sigma_e^2 I)$ are random variables accounting for unknown confounding and random errors. $K = \mathrm{Cov}(Z)$ is the inter-sample correlation matrix and I is an identity matrix. σ_g^2 and σ_e^2 are coefficients of the two variance components. Under the null hypothesis, $\sigma_g^2 = 0$ is assumed. Under the alternative hypothesis, $\sigma_g^2 > 0$ is tested. Only the mean is used as the fixed effect in the analysis above. The asymptotic null distribution of the likelihood ratio test statistic follows a 1:1 mixture of χ_0^2 and χ_1^2 distributions[197]. Efficient Mixed Model Association (EMMA) R package was applied for rapid estimation of variance components and maximum likelihood to perform likelihood tests[103].

We used standard t-test to test for the known batch effect for the BXD HSC dataset. When testing the significance of surrogate variables, a standard F test is performed to assess the significance of all surrogate variables using a linear model. In all tests above, FDR is conservatively estimated with $\pi_0 = 1$.

6.4.7 ICE eQTL mapping

ICE QTL mapping models the expression levels as the following linear mixed model:

$$\mathbf{y} = G\alpha + X\beta + \mathbf{u} + \mathbf{e} \tag{6.3}$$

where X, β are the fixed effects of known confounding variables and their coefficients, and $\mathbf{u}$ and $\mathbf{e}$ are random variables accounting for unknown confounding and random errors as described above. G represents the SNPs or other predictor variables to be tested with the coefficients of α. EMMA is applied to test for the significance of α using F test as previously suggested based on REML estimates of variance component[235, 240, 103].

We classified the eQTLs as *cis*-acting when the SNP and the probe are no farther than 50kb for yeast, 5Mb for BXD mouse RI strains and 500kb for human. *Trans*-acting eQTLs are stringently classified with the distance larger than 250kb for yeast, and 50Mb for mouse. The number of expected false positives were computed from 10 null randomized runs of each eQTL mapping setting $\pi_0 = 1$, as suggested in previous studies [196, 25].

6.4.8 Assessing the statistical significance of *trans*-regulatory bands

The statistical significance of a *trans*-regulatory band was quantified as the average of log p-values across all probes. We performed 10,000 random permutations of the SNP set with family-wise error correction to evaluate the genome wide corrected p-value of the strength of *trans*-regulatory bands.

Chapter 6 is published in Genetics, Volume 180, pp 1909-25, 2008. Hyun Min Kang, Chun Ye, and Eleazar Eskin, “Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots”. The dissertation author and Chun Ye were the primary investigators and authors of this paper.

Chapter 7

A High Resolution Association Mapping Panel for the Dissection of Complex Traits in Mice

7.1 Motivation

Over the past few years, human complex trait genetic studies have been revolutionized by the ability to carry out association studies on a genome-wide basis. Such genome-wide association (GWA) studies have now been applied to numerous complex traits and have resulted in the identification of hundreds of novel genes for traits such as diabetes, cancer, and various inflammatory diseases[6, 131] previously undetected in linkage studies. This success can be attributed to many factors including technological developments in collection of high-throughput genotype data, development of catalogues of common human variation such as the HapMap, and development of methodologies for association studies[45]. These developments allowed the human genetics community to leverage the increased power and resolution of association studies as compared to linkage analyses. Despite these successes, the fraction of the genetic component that is explained has been relatively modest for most traits. Thus, for example, traits such as type 2 diabetes and lipoprotein levels have relatively high heritabilites, in the range of 50%, and yet the many genes discovered for these traits explain in aggregate less than 10% of the risk. This can likely be attributed to many factors which include that the effects of single loci on a disease trait are very weak for common variation and GWA studies have low power to detect rare variation involved in disease [40, 69]. In addition, environmental factors likely are involved in disease and current GWA the lack power to examine gene-by-environment interactions.

Mouse models have a several advantages over human studies along these lines. In particular, complex traits in mouse strains have higher heritability and genetic loci often have stronger effects on the trait compared to humans partially due to their genetic history[120, 213, 230, 231]. In mouse studies, the environment can more carefully controlled and measurements can be replicated in genetically identical animals reducing the environmental effect on traits, thus increasing the portion of the variability explained by genetic variation.

Mouse models have been used very effectively for the identification of genes contributing to simple Mendelian traits, but unfortunately there have been few successes for genes contributing to complex, multigenic traits. Traditional genetic analysis in mice involves crossing different inbred strains and mapping the traits of interest using linkage analysis. An important problem has been the lack of resolution in identifying the causal loci from the results of a linkage study. Linkage analysis of a Mendelian trait is relatively straightforward since recombinant mice in the region of interest can be identified and fine mapping achieved given enough backcross or intercross progeny. In contrast, mapping of genes underlying complex traits results in only statistical evidence for the effect of an allele at a particular locus. Fine mapping in such cases generally requires the construction of congenic strains in which the region of interest from one strain is transferred onto the background of the second strain by a series of crosses. This essentially "Mendelizes" the trait, allowing the scoring of recombinations. But this as well frequently proves difficult because the alleles contributing to complex traits generally exhibit the very subtle effects that approach the levels of noise[4, 63], and several closely link genes may influence the trait at a given locus.

Naturally, buoyed by the prospects and success of human association studies having advantages over linkage studies, several groups have proposed mouse genome wide association studies[166]. These initial pioneering studies demonstrated the potential of mouse genome wide association studies with their early successes, but have also raised some challenges which include complex population stratification among the mouse strains and concerns about the lack of power to detect loci with modest effects. In fact, these two issues are intimately related. Population structure causes an inflation of the association statistics both creating spurious associations as well as artificially increasing the apparent strength of true association signals. The initial mouse genome wide association studies reported a tremendous number of genome wide significant signals some of which overlapped with known loci. This combined with the knowledge that mouse strains have high heritability for traits suggested that mouse association studies had sufficient statistical power. However, these initial studies did not adequately correct for population structure which when taken into account, eliminates the vast majority of predicted associations. Thus the inability to correct for population structure of the initial studies led them to severely overestimate their statistical power.

Recent studies on the confounding effects from population structure and genetic relatedness suggested that the complex pattern of sample structure can effectively be captured and corrected for by means of linear mixed model[235, 240, 103], and it has been demonstrated that both the false positive and false negative rates considerably decrease when applying linear mixed model compared to principal component analysis[168] and genomic control[50] that have been popularly used in human association studies. The availability of Efficient Mixed Model Association (EMMA) method enabled us to perform a large scale genome-wide association mapping between hundreds of thousands of SNPs and tens of thousands of phenotypes which includes whole genome expression profiles. Using the same method, we were able to perform simulations to measure power of mouse strain association studies under the effect from sample structure, which led us to the conclusion that previous studies which typically incorporated on the order of 30 strains are underpowered to detect moderate loci involved in complex traits[32, 75, 137]. Our ability to perform simulations which can accurately estimate statistical power allowed us to explore a wide range of possible designs for mouse association studies. Through these simulations, we observed that by incorporating permanent recombinant inbred strains which were derived by inbreeding progeny from crosses of two different inbred strains, we are able to achieve power to detect loci which segregate the strains and have a moderate affect on the trait. Building on this observation, we developed a combined set of inbred strains, which we term the "Hybrid Mouse Diversity Panel" (HMDP) that includes 100 commercially available inbred strains consisting of 30 classical inbred strains and 3 sets of recombinant inbred strains.

We show in this report that this panel has sufficient power to map traits that contribute to less than 5% of the overall variance. Importantly, the resolution of the panel is, in some cases, two orders of magnitude better than that achievable using linkage analysis of complex traits. Additional practical advantages of the approach include the need for costly genotyping is eliminated since approximately 100 strains have now been genotyped at over 135,000 SNPs. In addition, each strain is renewable and, therefore, diverse molecular and phenotype data can be collected *ad infinitum*. Both of these points highlight the significant advantages of an inbred panel over other non-renewable populations which can be used for high resolution mapping, such as heterogeneous or outbred stocks[64]. In addition to greatly increased ability to narrowly map and identify genes for complex traits, this panel should be useful for the analysis of gene-by-environment interactions where multiple individuals of the same genotype need to be studied. Moreover, the fact that the data involving clinical traits, expression traits, proteomic traits, and metabolomic traits is cumulative, makes this resource ideal for systems biology.

7.2 Results

7.2.1 Design principles of mouse association studies

Our goal is to develop a panel of inbred mouse strains for performing association studies which have adequate statistical power and resolution for mapping of complex traits. However, compared to human studies, estimating statistical power in mouse association studies is more complex. Since humans are an outbred population, a reasonably collected association study cohort can be assumed to be unrelated. Under such assumptions, statistical power of association studies only depends on a few factors including the effect size, the minor allele frequency, and the significance threshold. Due to the population structure of inbred mouse strains, statistical power of mouse association studies is much more complicated and depends on many more factors including phenotypic specific factors. These include not only the effect size and significance threshold, but also the inter-strain relatedness, the distribution of the causal SNP among the strains, and the background genetic effect. These latter factors also affect inflation of association statistics due to population structure and as expected the statistical power is intricately related to population structure.

We can leverage our ability to correct for population structure using EMMA to estimate the statistical power of any strain set through simulation studies. The basic idea behind a simulation study is that we generate simulated phenotypic data that capture the genetic background effects. In this data, strains which are genetically closer to each other are more likely to have more similar phenotype values than strains which are further apart. We then pick a random SNP and modify the phenotype values by assuming that the SNP affects the phenotype. We then apply EMMA to see if we would detect the phenotype at a given significance threshold. Using this approach, we can estimate the statistical power of a given set of strains.

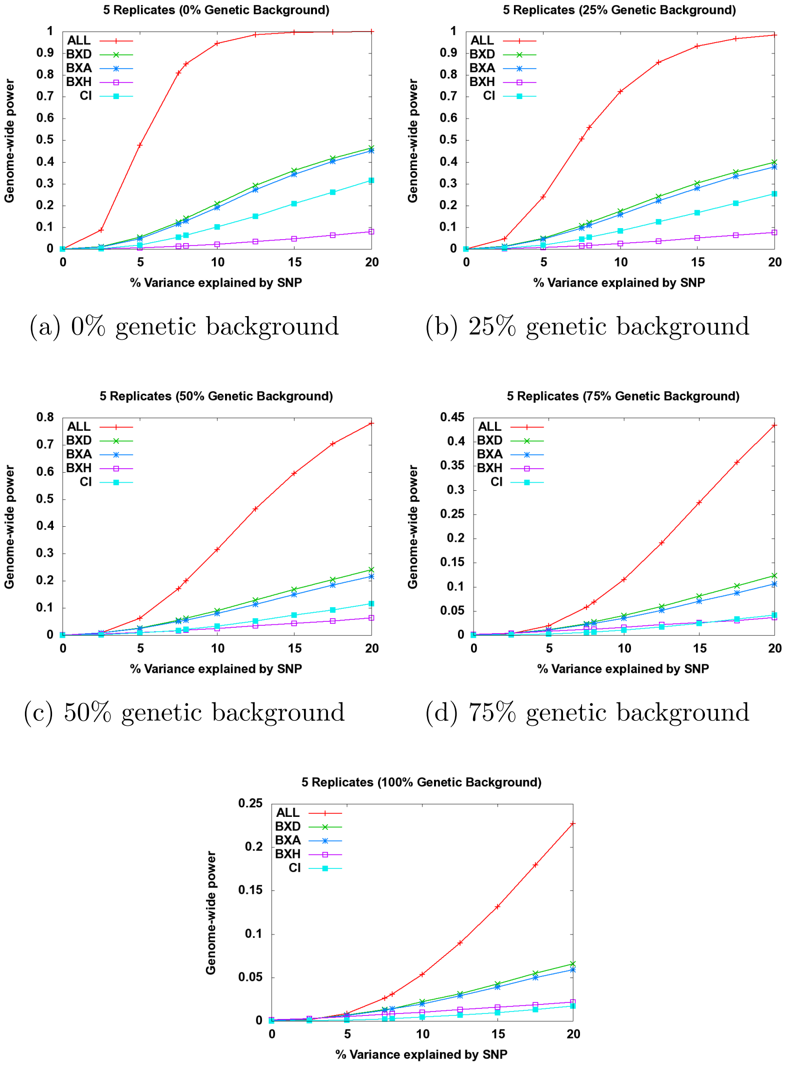
7.2.2 Strain selection for the Hybrid Mouse Diversity Panel

While hundreds of inbred strains have been derived, a relatively small fraction of these are useful for an association panel and we can use several intuitions to guide our choices for the inbred strains. Certain strains, such as congenics and closely related members of a family of strains (for example, many members of the C57BL family) are minimally informative because of their largely identical genetic ancestry. These strains are only informative for the small number of loci that differ and we include only one representative of each of these strains in our panel. On the opposite spectrum, “wild-derived” strains such as CAST, SPRET, and MOLF, are so widely diverged from the classical inbred strains that they differ at many loci which are not polymorphic among the classical inbred strains. These “wild-derived” strains have dramatically different phenotypes, and much of the genetic contribution to this difference stems from these loci. Since these loci are not polymorphic in the majority of the strains, we have very little power to identify associations at these loci due to small minor allele frequencies. Moreover, including such a diverged set of strains increases the size of polygenic background effect, possibly resulting in loss of power in identifying the effect of a single locus. For these reasons, we do not include "wild-derived" strains in our panel. Altogether, we selected for the panel 29 classical inbred strains (Table 7.1). This panel is representative of previous mouse association studies that were performed[166]. We carried out power calculations to estimate the level of SNP effect that could be detected by the inbred panel under various conditions of heritability and p-value cutoff (Figure 7.1 and Table 7.2). These analyses indicated that 29 strains was not sufficient to detect typical loci for various complex traits in mice, which generally explain less than 10 percent of the total trait variance and is consistent with previous estimates[64].

Where our approach differs from previous approaches is that we additionally include in the our panel 71 recombinant inbred (RI) strains (Table 7.1). These are derived by crossing a pair of inbred parental strains and then deriving a set of inbred progeny through brother-sister mating for 20 or more generations. These strains consist of roughly 50% genetic contribution from each of the parental strains. Each allele which is polymorphic among the parents is present in about 50% of the strains in the RI set. Since the RI strains do not suffer from population structure due to the way they were constructed, this genetic structure is maximally informative for detecting associations at these polymorphic loci, yet provide additional power to detect loci only polymorphic between the parental strains. The power can further increased by combining multiple RI sets, considering that the complex genetic relatedness among the strains are accounted by the availability of high-density markers. We select several RI sets to cover a significant fraction of the SNPs in our panel. The RI strains we selected were derived from crosses between C57BL/6J (B) and either DBA/2J (D), A/J (A), or C3H/HeJ (H). RI substantially add to the overall power to detect loci with small effects (Figure 7.1 and Table 7.2). Thus, for example, in this set of inbred strains [the Hybrid Mouse Diversity Panel (HMDP)], we have approximately 80% power to detect SNPs that contribute to about 10% of the overall variance of a complex trait, depending on the heritability of the trait and the number of mice examined.

Table 7.1: Inbred and RI strains to be used in mouse whole genome association

Inbred strains				Recombinant inbred					
Strain	(n)	Strain	(n)	Strain	(n)	Strain	(n)	Strain	(n)
129X1/SvJ	10	LG/J	7	AXB1/PgnJ	6	BXA4/PgnJ	6	BXD28/TyJ	8
A/J	10	LP/J	10	AXB10/PgnJ	5	BXA7/PgnJ	6	BXD29/TyJ	6
AKR/J	10	MA/MyJ	6	AXB12/PgnJ	6	BXA8/PgnJ	6	BXD31/TyJ	6
C57BL6/J	11	NOD/ShiLtJ	10	AXB13/PgnJ	4	BXD1/TyJ	4	BXD32/TyJ	4
BALB/cJ	10	NON/ShiLtJ	13	AXB15/PgnJ	6	BXD5/TyJ	6	BXD33/TyJ	4
BTBR T+ tf/J	10	NZB/BINJ	7	AXB19/PgnJ	6	BXD6/TyJ	6	BXD34/TyJ	6
BUB/BnJ	9	NZW/LacJ	10	AXB19a/PgnJ	5	BXD8/TyJ	6	BXD36/TyJ	6
C3H/HeJ	10	PL/J	10	AXB19b/PgnJ	8	BXD9/TyJ	6	BXD38/TyJ	6
C57L/J	11	RIIIS/J	10	AXB2/PgnJ	6	BXD11/TyJ	6	BXD39/TyJ	6
C58/J	10	SEA/GnJ	11	AXB23/PgnJ	4	BXD12/TyJ	4	BXD40/TyJ	6
CBA/J	10	SJL/J	2	AXB24/PgnJ	5	BXD13/TyJ	4	BXD42/TyJ	6
CE/J	9	SM/J	10	AXB4/PgnJ	6	BXD14/TyJ	2	BXH10/TyJ	5
DBA/2J	10	SWR/J	10	AXB5/PgnJ	2	BXD15/TyJ	6	BXH14/TyJ	9
FVB/NJ	5			AXB6/PgnJ	6	BXD16/TyJ	4	BXH19/TyJ	1
8 I/LnJ	9			AXB8/PgnJ	6	BXD18/TyJ	6	BXH2/TyJ	9
KK/HiJ	8			BXA1/PgnJ	2	BXD19/TyJ	5	BXH20/KccJ	8
				BXA11/PgnJ	6	BXD2/TyJ	2	BXH22/KccJ	1
1				BXA12/PgnJ	5	BXD20/TyJ	6	BXH4/TyJ	1
1				BXA13/PgnJ	6	BXD21/TyJ	6	BXH6/TyJ	1
1				BXA14/PgnJ	6	BXD22/TyJ	5	BXH7/TyJ	1
0				BXA16/PgnJ	7	BXD24a/TyJ	5	BXH8/TyJ	1
1				BXA2/PgnJ	6	BXD24b/TyJ	6	BXH9/TyJ	1
0				BXA24/PgnJ	6	BXD27/TyJ	4	B6Cc3-1/KccJ	6
				BXA25/PgnJ	6	BXD26/PgnJ	6		



(a) 0% genetic background

(b) 25% genetic background

(c) 50% genetic background

(d) 75% genetic background

(e) 100% genetic background

Figure 7.1: Power calculations. Estimated power when the SNP effect is 10%, and genetic background effect is 30% with 5 replicates per strain. We estimated the power for the 29 inbred strains, the individual RI panels (BXD, AXB/BXA, and BXH) and the combined HMDP.

Table 7.2: Statistical power of HMDP, RI and classical inbreds

SNP effect : 10% variance, 5 replicates per strain averaged power across 107k SNPs					
Strain sets	HMDP	BXD	BXA	BXH	CI
$h_g^2 = 0.00$	0.945	0.209	0.191	0.022	0.101
$h_g^2 = 0.25$	0.725	0.174	0.158	0.025	0.083
$h_g^2 = 0.50$	0.315	0.090	0.080	0.024	0.033
$h_g^2 = 0.75$	0.116	0.040	0.035	0.017	0.010
$h_g^2 = 1.00$	0.053	0.022	0.020	0.010	0.004
SNP effect : 20% variance, 5 replicates per strain averaged power across 107k SNPs					
Strain sets	HMDP	BXD	BXA	BXH	CI
$h_g^2 = 0.00$	0.999	0.465	0.452	0.079	0.315
$h_g^2 = 0.25$	0.984	0.399	0.378	0.077	0.254
$h_g^2 = 0.50$	0.779	0.242	0.217	0.063	0.115
$h_g^2 = 0.75$	0.435	0.123	0.106	0.037	0.041
$h_g^2 = 1.00$	0.227	0.066	0.059	0.022	0.017

7.2.3 Validating the statistical power of the HMDP through mapping metabolic clinical traits

We phenotyped the HMDP strains, using 6 to 12 males per strain, for a variety of metabolic traits, including total cholesterol, HDL cholesterol, free fatty acids, body fat, and weight. For the association, we analyzed nearly 107,000 of the informative mouse HapMap SNPs for associations with these quantitative phenotypes. Genetic association studies in inbred model organisms, including inbred strains of mice, are confronted by the problem of inflated false positive rates due to population structure and genetic relatedness. Inbred strains of mice have a complex genealogical history that results in a wide range of differences in the extent of genetic relatedness[67]. Conventional statistical tests of independence between a genetic marker and a phenotype are prone to spurious associations because the marker and the phenotype are likely to be correlated due to population structure, which violates the independence assumption under the null hypothesis. This population structure can result in false positives unless corrected. We previously developed a new method, termed Efficient Mixed Model Association (EMMA) which corrects for such population structure and genetic relatedness in model organism mapping[103]. We have used EMMA to perform whole genome association mapping with the HMDP (see Methods). We conservatively estimated using permutation analysis (see Methods) that a p-value of 4.1×10^{-6} was significant at a genome wide level.

Due to our long interest in complex traits, such as those contributing to metabolic syndrome and atherosclerosis, we focus on plasma lipids to demonstrate the overall approach. Several of these coincide with loci previously identified using quantitative trait locus (QTL) linkage analysis (Table 7.2). Previous linkage approaches in mice have identified over 50 QTL for plasma HDL levels[218]. We compared our associations to those available from the Mouse Genome Informatics database and we observed considerable overlap between our results in the HMDP and this comprehensive list of HDL QTL. In each of these studies loci are identified based on the individual variations between 2 strains of mice. In the current approach, using 100 strains, we are able to 'recover' a significant portion of all previously reported HDL QTL. The signals not replicated in the in the HMDP where either identified using a wild-derived strain, fed an atherogenic diet or where found either in females or a cross combining females and males.

To validate the association approach, we asked whether we could detect a previously identified common variation among inbred strains affecting HDL-cholesterol levels. We have previously shown that variations of the apolipoprotein A2 (ApoA2) gene locus affecting APOAII protein levels in the plasma occur commonly among inbred strains and that these significantly influence HDL-cholesterol levels[31, 51, 220]. We observed a total of 38 SNPs on distal chr. 1 associated with HDL-cholesterol at $p < 4.1 \times 10^{-6}$ (Figure 7.2D). One of the peak SNPs in the region is located 50 kilobases upstream of the ApoA2 gene at 173.12 Mb. Surprisingly, the peak snp at 172.4 Mb, within an intron of the gene Nos1ap, is the peak HDL associated snp . In contrast, linkage analysis of HDL cholesterol in one large cross (shown in Figure 7.2E), identified a very large region, containing approximately 300 genes associated with HDL cholesterol. Thus, the resolution achieved in the association analysis is more than two orders of magnitude better than that achieved by linkage analysis.

To better assess the contribution of each individual set of mice on the overall performance of GWAS we performed EMMA on the inbred, combined RI panel and

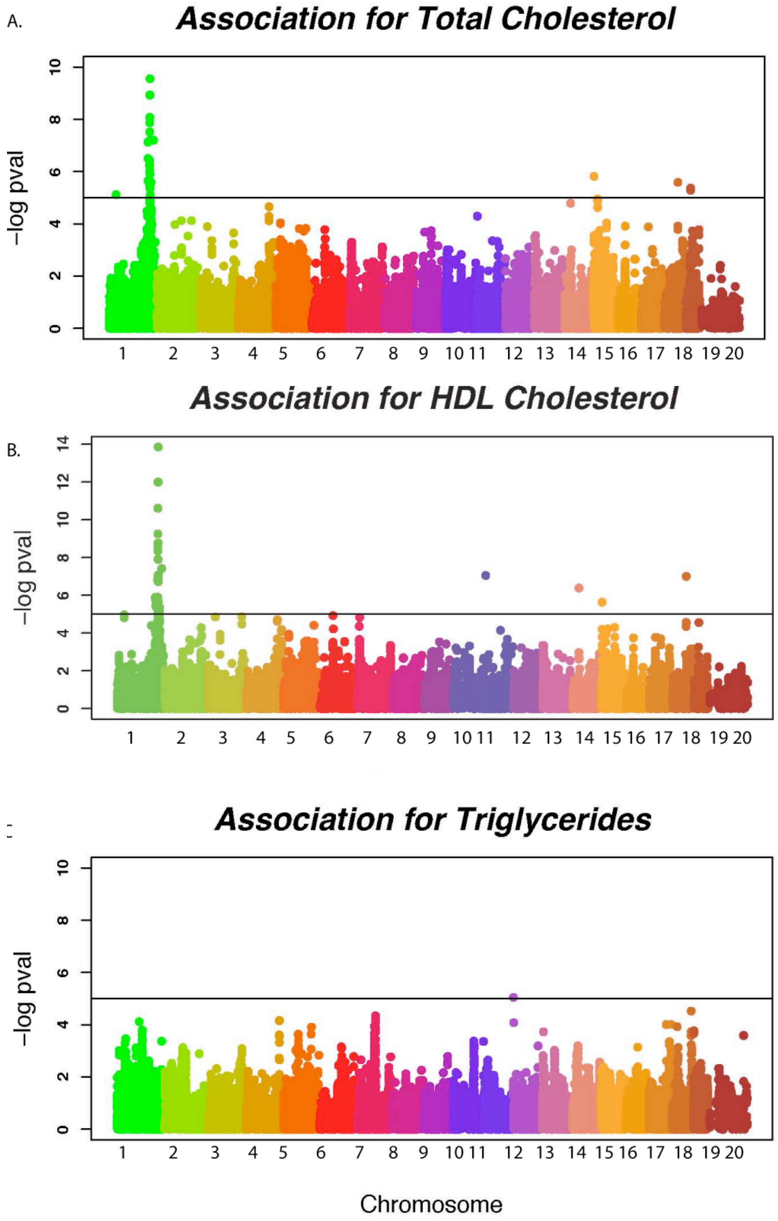


Figure 7.2: Detection of associations for plasma lipids in HMDP strains coincide with a corresponding QTL in C57BL/6 x C3H/HeJ F2 crosses. (Continued on the next page)

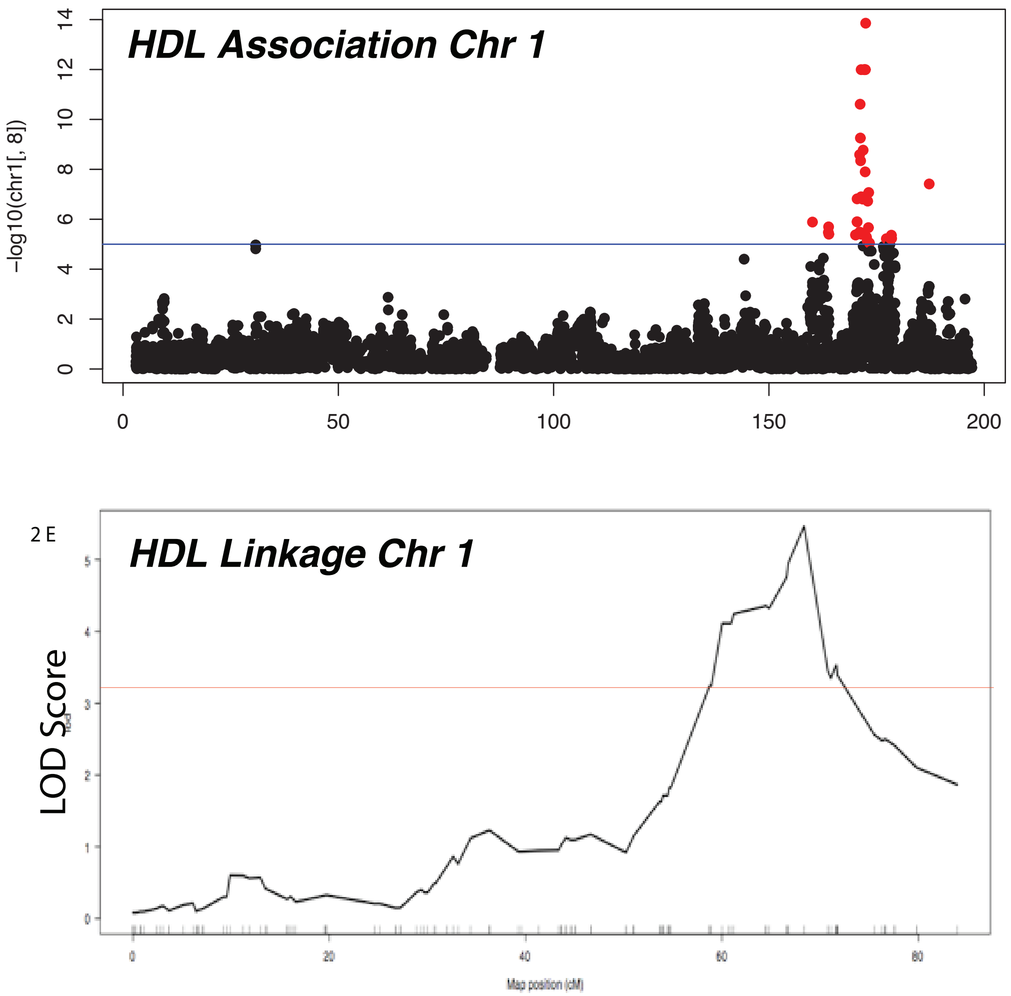


Figure 7.2: Detection of associations for plasma lipids in HMDP strains coincide with a corresponding QTL in C57BL/6 x C3H/HeJ F2 crosses. Panels A-C represent the GWAS for total plasma cholesterol (B), and triglycerides (C), respectively. Lod score curve for HDL cholesterol for chromosome 1, with a significant QTL present at the distal region. The middle panels (D and E) compare association results with linkage results for chromosome 1. Panel D is a plot of p-values for plasma HDL from association testing of SNPs using 100 strains of the HMDP and panel E shows LOD scores for plasma HDL from an F2 cross of C3H/HeJ and C57BL6/J. These results demonstrate the power of the HMDP to detect associations for QTL observed in the F2 cross, and also highlight the vastly improved resolution of association testing with the MDP.

the individual RI sets. The increase in positive signal in the RI set reflects the reduced resolution of the RI panels. Due to the complex nature of the hMDP we wanted to confirm the additional peaks on chr 1. Using the genotype for the peak marker at the ApoA2 locus as a covariate in our EMMA analysis completely ablated the remaining HDL signal on chr 1 This indicates that the additional peaks near ApoA2 where likely due to the large LD block introduced by the inclusion of the BXH RI panel in the HMDP. Below, using expression traits, we further consider this issue of LD.

In addition to the apolipoprotein A2 locus, several other significant associations with plasma Lipoproteins were observed. Importantly, several of these loci were previously identified by QTL analysis of various crosses of inbred strains of mice as noted above[218]. In addition to the Apoa2 locus, there are at least two other significant loci on mouse chromosome 1 as well as loci on chromosomes 11, 18, 14, and 15 (summarized in Table 2). Of particular interest is the association on chromosome 15 at 58.6 MB for unesterified cholesterol. This corresponds to the same region, within 1 MB, as the novel plasma lipid genes, Trib1 and Nsmce2, recently identified in human genome-wide association studies[225]. A considerable advantage of murine studies is the availability of peripheral tissues for transcriptional, proteomic and matabolomic profiling. For example, the expression of Trib 1 is under cis-regulation (p-value) and is significantly correlated with Total cholesterol, Free fatty Acids, and VLDL/LDL levels (need to include R values). Another novel human GWAS candidate recovered with the current approach is for Amac1 the homologue of AMACL1[105]

7.2.4 Resolution of mouse association studies

An important criterion for the effectiveness of a mouse association panel is the mapping resolution or the size of the region which we can detect as associated with a trait. Due to many population bottlenecks in the history of the inbred mouse strains, long regions of linkage disequilibrium are common throughout the mouse genome. RI strains contain even longer regions of linkage disequilibrium since there are a limited numbers of recombinations that occur when they are being derived. Intuitively, by adding the classical inbred strains to the RI strains, we can improve the mapping

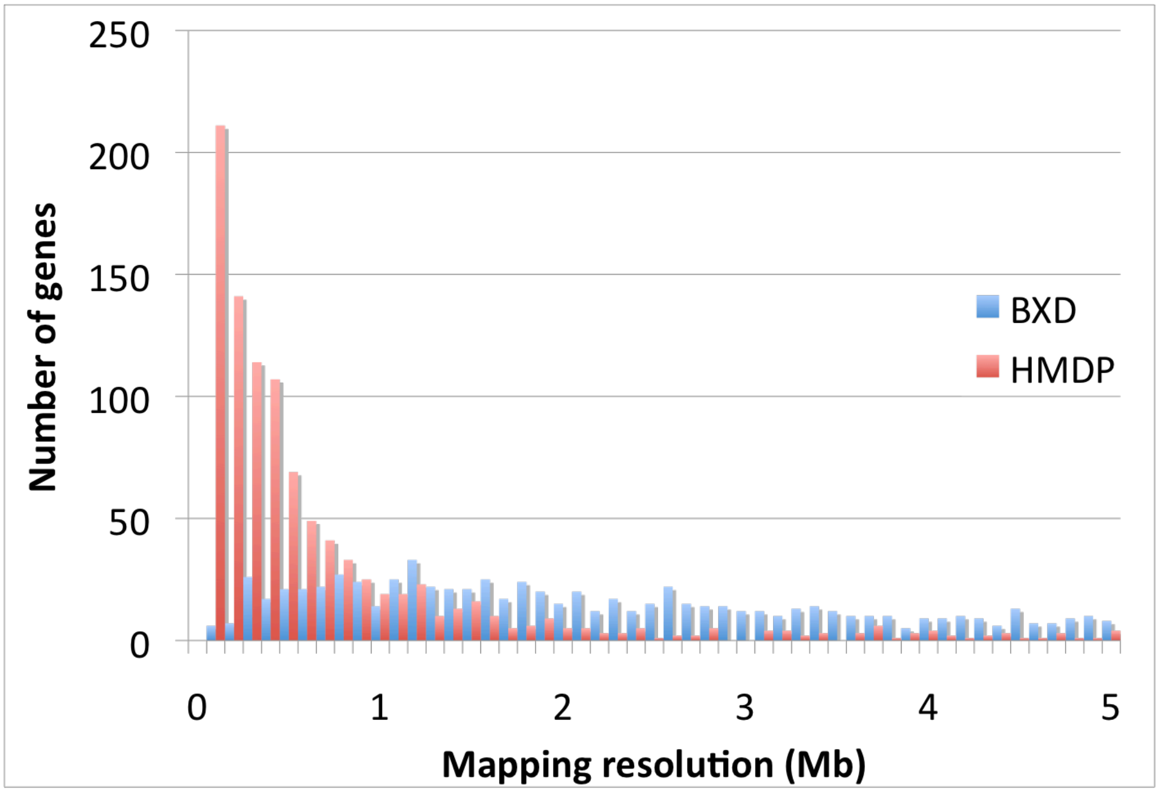


Figure 7.3: Expression Traits Demonstrate high resolution of HMDP. Distance between peak cis-eSNPs and the transcription start site of the corresponding gene in the HMDP. The majority of cis-eSNPs map within 500 kb of the transcription start site of the corresponding gene

resolution over only using RI strains. The cis-eSNPs provide a convenient measure for the overall resolution of the HMDP. Thus, it is reasonable to assume that the majority of causal DNA variations contributing to cis-eSNPs would reside within 500 kb of the gene itself. Thus, the distance between the peak eSNP and the 5' or 3' end of the gene provides a measure of the accuracy of our association analysis. The results, presented in Figure 7.3, indicate that the peak SNPs usually occur within 500 kb of either end of the gene. We also repeated this analysis using only data from recombinant inbred mice or the combined HMDP using the top 1000 local eSNP and as expected found a dramatic increase in the number of local eSNP mapping within 1Mb of the genes transcription start site with the HMDP, 80% in the HMDP vs. 18% in the BXD strains (Figure 7.3).

7.2.5 Application of the HMDP by mapping metabolic clinical traits

A major goal of the HMDP is to achieve high-confidence, high-resolution genetic data contributing to complex phenotypes. Of particular interest are phenotypes related to human disease, such as those contributing to metabolic syndrome and atherosclerosis, and we focus on plasma lipids to demonstrate the overall approach. We phenotyped the HMDP strains, using 6 to 12 males per strain, for a variety of metabolic traits, including total cholesterol, HDL cholesterol, free fatty acids, and unesterified cholesterol. For the association, we analyzed nearly 107,000 of the informative mouse HapMap SNPs and we have used EMMA to perform whole genome association mapping with these quantitative phenotypes with the HMDP. We conservatively estimated using simulation analysis (see Methods) that a p-value of 4.1×10^{-6} was significant at a genome wide level.

The GWAS plots for plasma phenotypes are presented in Figure 7.2A-C and a number of loci approaching or exceeding significance were observed. Figure 7.4 shows the dramatic reduction of p-value inflation following application of EMMA, many of which are false positive signals due to a series of confounders discussed below [159].

To validate the association approach, we asked whether we could detect a previously identified common variation among inbred strains affecting HDL-cholesterol levels. We and others have previously shown that variations of the apolipoprotein A2 (ApoA2) gene locus affecting APOAII protein levels in the plasma occur commonly among inbred strains and that these significantly influence HDL-cholesterol levels [31, 51, 63, 218, 220] We observed a total of 38 SNPs on distal chr. 1 associated with HDL-cholesterol at $p < 4.1 \times 10^{-6}$ (Figure 7.2D). One of the peak SNPs in the region is located 50 kilobases upstream of the ApoA2 gene at 173.12 Mb. Surprisingly, the peak SNP at 172.4 Mb, within an intron of the gene Nos1ap, is the peak HDL associated SNP . In contrast, linkage analysis of HDL cholesterol in one large cross, identified a very large region, containing approximately 300 genes associated with HDL cholesterol. Thus, the resolution achieved in the association analysis is more than two orders of magnitude better than that achieved by linkage analysis.

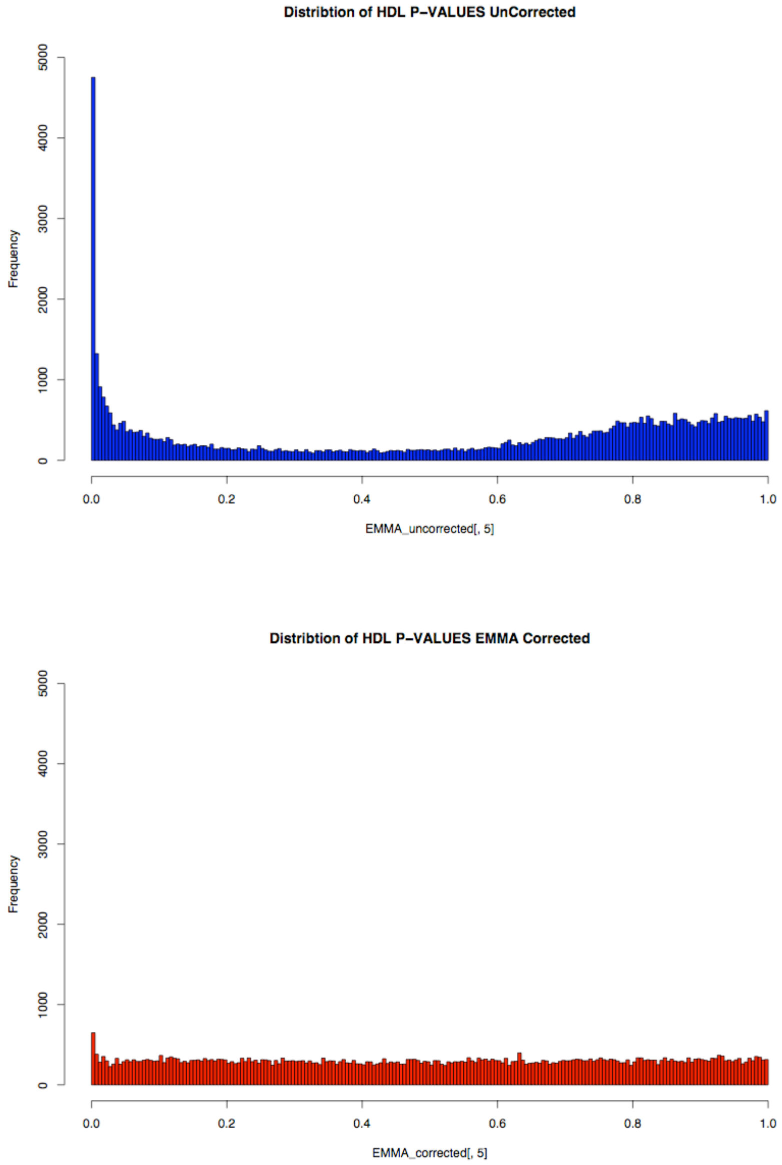


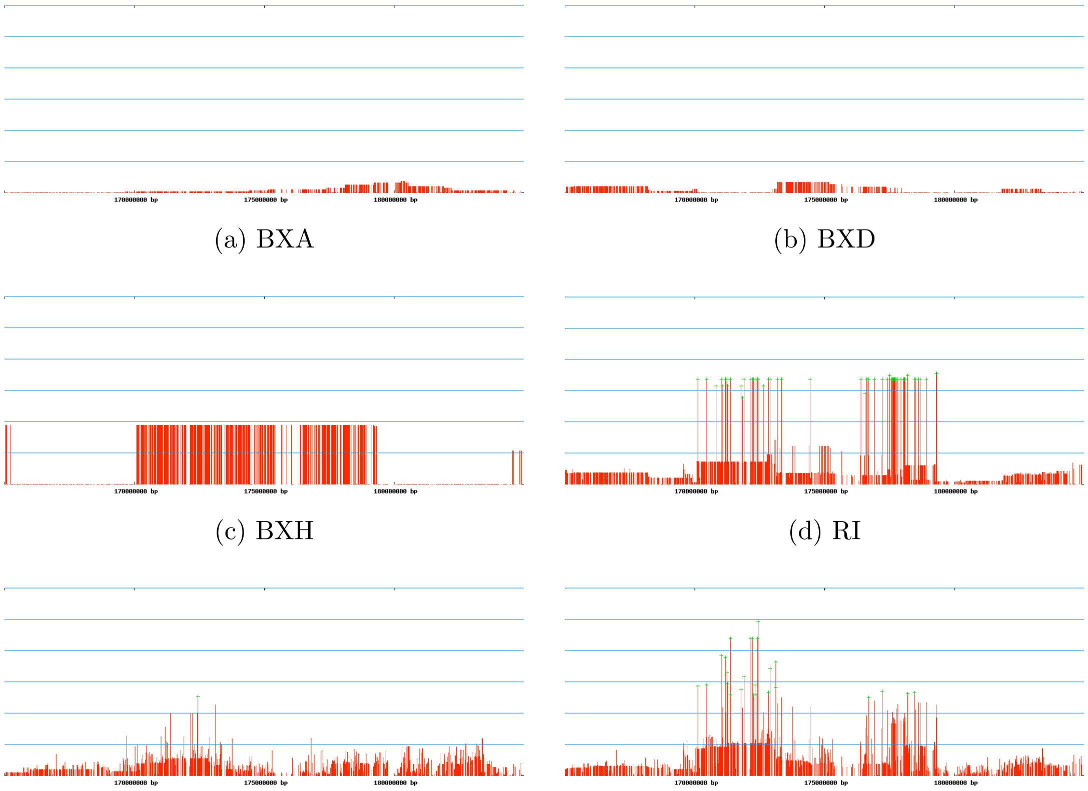
Figure 7.4: Correcting for population structure dramatically reduces false positives in murine association studies Panel A is a histogram of uncorrected p-values for plasma HDL and panel B is a histogram EMMA corrected p-values

In addition to the apolipoprotein A2 locus, several other significant associations with plasma lipoproteins were observed. Previous linkage approaches in mice have identified multiple QTL for plasma HDL levels[219]. We compared our associations to those available from the Mouse Genome Informatics database and we observed considerable overlap between our results in the HMDP and this comprehensive list of HDL QTL. In each of these studies loci were identified based on the individual variations between 2 strains of mice. In the current approach, using 100 strains, we are able to 'recover' a significant portion of all previously reported HDL QTL. The signals not replicated in the HMDP where either identified using a wild-derived strain, fed an atherogenic diet or were found either in females or a cross combining females and males.

In addition to plasma HDL levels, we found significant associations for total cholesterol, triglycerides, free fatty acids and unesterified cholesterol. Several of these are of particular interest because they demonstrate how murine studies complement human associations. For example the signal on chromosome 15 at 58.6 MB for unesterified cholesterol is within 1 MB of the novel human GWA plasma lipid genes, Trib1 and Nsmce2[225]. A considerable advantage of murine studies is the availability of peripheral tissues for transcriptional, proteomic and metabolomic profiling. For example, the expression of Trib 1 in liver is under cis-regulation (1×10^{-5}) and is negatively correlated with total cholesterol ($R = -0.27$), HDL($R = -0.23$), free fatty acids($R = -0.36$) and unesterified cholesterol ($R = -0.30$) levels. Conversely Nsmce2 is under distant regulation (1.6×10^{-6}) and is also significantly correlated with total cholesterol, HDL and unesterified cholesterol. Another novel human GWAS candidate recovered with the current approach is for Amac1 the homologue of AMACL1, which is candidate near the HDL and total cholesterol peak on chr 11. Unlike the Trib 1 locus, Amac1 does not have a cis-eQTL nor is its expression correlated with plasma lipid traits.

7.2.6 Comparison to previous mouse association studies

The HMDP contains several large divergent cohorts of mice. In order to comprehensively identify the genetic component of the association signal identified



(a) BXA

(b) BXD

(c) BXH

(d) RI

(e) CI

(f) HMDP

Figure 7.5: Comparison of individual strain sets on phenotype association mapping in the HMDP

for clinical traits we performed a series of mapping studies. We analyzed the inbred and each individual RI sets independently. The classical inbred set simulates the design of many of the original studies on inbred strains. Figure 7.5 shows the GWAS for the HDL trait using only the 29 inbred strains. As we can see, no loci are reported as significant after correcting for population structure. Figures 7.5 also shows the association for each RI set. This simulates the design of an mouse association study proposed by Williams et al[227]. In this case, several loci are associated, however they show very poor resolution. These analysis highlight the lack of power or resolution to map complex traits in either the classical inbred panels or the individual RI panels of mice.

7.3 Discussion

We have utilized a “hybrid” strategy for association mapping in mice that combines classical inbred strains as well as RI strains. We specifically designed this study population to address several key limitations of complex genetic mapping in mice: low resolution of linkage approaches, the high degree of false positive signals found in murine association mapping and the critical need for permanent resources for systems based approaches. In this strategy, the inbred strains provide mapping resolution while the RI strains provide power. Our results clearly indicate that the approach is capable of mapping complex traits with high resolution. We have validated the approach by mapping both complex metabolic traits as well as expression traits. The HMDP approach should be useful for gene discovery, analysis of gene-by-environment interactions, and systems biology. An important advantage of the approach is that the strains are commercially available to all investigators and that the data is cumulative. We have now established a database (`http://whap.cs.ucla.edu/emmaCorrectionServer/`) that allows investigators to utilize our data in a straightforward manner.

A number of previous studies have utilized association analysis in mice in an attempt to improve mapping resolution. Clearly, outbred, heterogeneous stocks of mice can be used if corrected for population structure[63, 70, 206]. A major disadvantage of the approach is the cost of high-density genotyping and the fact that each mouse is unique and thus can be studied for a limited number of phenotypes. Nevertheless, Valdar and colleagues[206] identified hundreds of significant associations for 97 typed traits with an average of 95% confidence interval of 2.8 Mb.

There have also been a number of studies that have exploited the mosaic structure of common inbred mouse strains to perform association mapping [32, 75, 79, 80, 112, 119, 123, 124, 137, 144]. The methods have proved effective for localizing genes with large effects but not for genes with effect sizes less than 10%, as is usually observed with complex traits[32, 63]. In addition, due to the lack of power, population structure is a major problem that can produce false positives[32, 103]. Recently, previously identified susceptibility genes using murine association mapping have failed to replicate in linkage studies designed to confirm these novel loci under-scoring the importance of developing methods designed to improve power and reliability of murine association mapping. We believe that the methods used to develop the HMDP address the limitations of these previous studies. Another limitation has been the finding of sporadic correlations across the genome and is discussed below.

As a first test of the approach for mapping genes underlying complex traits, we have typed eight mice for each strain in the HMDP for a variety of complex metabolic traits. Relatively few genes contributing to common complex variations in mice have been identified for reasons discussed in the Introduction. However, one gene that has been shown in a number of studies to contribute to complex variation is the apolipoprotein A2 gene on distal mouse chromosome 1[140, 173, 220]. We showed that the resolution for mapping HDL cholesterol to the Apoa2 locus is more than 2 orders of magnitude better than is achieved by linkage analysis in a cross with over 300 progeny (Figure 7.6). In addition to Apoa2, we have identified a number of additional loci contributing to metabolic traits. Importantly, many of these correspond to previously mapped QTL loci for the corresponding traits. Some also correspond to peak associations identified for these traits in human genome wide association studies. For example, association was observed with Trib1 and Nsmce2 locus on mouse chromosome 15, and Amac1 on mouse chromosome 11, and the corresponding human loci were observed in a meta-analysis of genome-wide association studies for lipid traits[225]. Interestingly, both Trib1 and Nsmce2 exhibit cis-eSNPs in our study, consistent with their role in HDL and triglyceride metabolism while Amac1 was under trans-regulation.

The strongest validation of the approach came from analysis of gene expression traits. A variety of âĂIJgenetical genomicâĂİ studies in humans, rats, mice, and plants have shown that genetic variations influencing gene expression are very common in natural populations[54, 165, 167]. Trans-acting loci contributing to transcript levels have proven difficult to validate due to the problem of multiple comparisons, but cis-acting loci provide a relatively straightforward means of examining the power and resolution of our MDP association approach. Although such loci have been commonly termed cis-acting since the peak linkage occurs very near the expressed gene, a better term for these is âĂIJlocalâĂİ eQTL (or eSNPs) since there could be local trans-regulation as well as cis-regulation[183]. We and others have previously

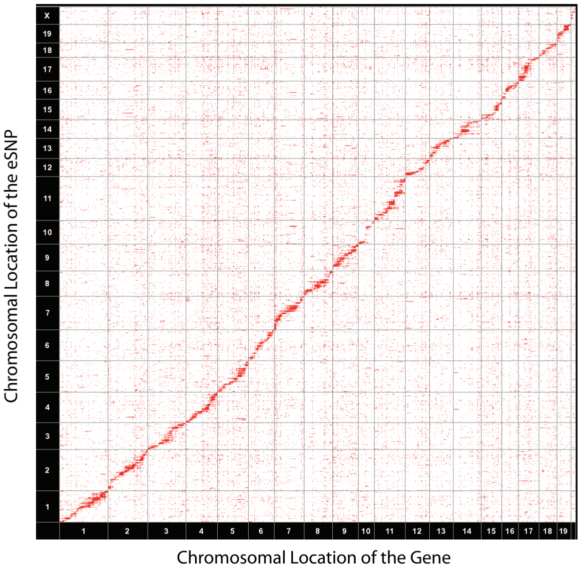


Figure 7.6: Expression SNPs from HMDP. Panel A: Transcript levels in liver of HMDP mice (3 RNA per strain) were profiled and significant associations are plotted according to chromosomal position (X-axis) versus the location of the structural gene (Y-axis). The strong diagonal line represents cis-eSNP, whereas the remainder are trans-eSNP signals

validated the cis-acting of the loci by quantitating transcript levels derived from each allele in heterozygous mice using coding polymorphisms[52]. Our data provide much better resolution than previous studies of mammalian cis-eQTL. Using whole genome expression array analysis in livers of the MDP strains, we identified over 2,200 cis-eQTL, comparable to the numbers identified in large crosses with hundreds of rodents.

One potential problem with the use of our association approach is a level of long-range LD. In particular, Paigen and colleagues have provided strong evidence of functional LD both within blocks and also between regions of separate chromosomes [164]. Thus, some association signals could represent such regions of distant LD. We have addressed this concern by testing for the presence of LD between loci identified for a given trait.

The identification of additional strains useful for our approach is important especially to identify SNPs with low effect size contributing to complex phenotypes. A very large set of RI strains, termed the Collaborative Cross, is being developed for the mapping and analysis of complex traits[39]. Unfortunately, that will not be completed until about 2012, and it is not anticipated to be capable of gene-level resolution[63]. Once even a subset of these strains are fully backcrossed to homozygosity and genotyped, these mice would complement the methods developed here and extend our approach.

Why is a mouse association resource important for the dissection of complex diseases? While human studies have identified many genes involved in complex diseases, these studies have important limitations. For example, there is limited ability to carry out experimental validation, and limited access to tissues. In some cases, due to linkage disequilibrium and distal regulatory elements, it is difficult to definitively identify the susceptibility genes at loci associated with the disease. Also, the genes are generally identified out of context, with limited information about how they lead to disease. Finally, although many genes have been identified in human association studies, the total amount of variance explained by the set of known genes is a small fraction of the total variance. High resolution mapping studies in mice should complement human association studies and also make possible the development of co-expression networks allowing functional annotation of the identified genes[126, 185].

This resource should also be valuable for examining gene-by-environment interactions. Direct examination of such gene-by-environment interactions in human populations is extremely difficult for several reasons. In particular, each human will have experienced a different set of environmental exposures over his lifetime and these would be almost impossible to quantitatively assess. The difficulty in understanding gene-by-environment interactions may partially explain why human association studies have only revealed a small fraction of the variance of complex traits. The analysis of such interactions using genetic crosses between inbred strains of mice has significant problems. While the power to detect linkage is very good, the loci identified tend to be very large and the identification of the underlying gene is extremely difficult. In addition, analysis of inbred strains rather than progeny from genetic crosses offers the following important advantages: 1) Multiple measurements can be carried out on each strain, allowing the measurements to more accurately reflect the true effect of the genetic factors. 2) Publicly available genotypes eliminate the need for individual researchers to genotype the animals independently. 3) The inbred strains can be directly purchased from suppliers such as The Jackson Laboratory, eliminating the need to breed the two generations of strains necessary to generate a cross. 4) Inbred strains allow more flexibility in collecting phenotypes, since the phenotypes can be studied in multiple phases of the study. 5) More tissues are accessible for expression and metabolomic analyses. For example, for some tissues in mice, not enough of the tissue exists in a single animal to obtain enough of a sample to reliably obtain expression proteomic, or metabolomic data. For inbred strains, pooling is possible. 6) The mouse HMDP allows improved sharing of data in the community. Since animals within an individual strain are genetically identical, phenotypes collected in different studies for different research groups may be combined taking care to correct for any research group effects on the phenotype measurement. 7) There is the ability to measure effects for multiple interventions.

Finally, the MDP resource is ideal for systems-based approaches. Systems biology uses technologies such as gene expression, microarrays, and mass spectrometry in combination with computational and statistical tools to address complex systems. This necessitates analysis of many different components to allow an understanding of their interactions. In this study, we have provided both clinical metabolic data and liver expression data for the HMDP strains. In the future, other kinds of clinical traits and expression data can be integrated with our datasets. Moreover, proteomic and metabolomic analyses that require additional tissues can be performed on the same set of inbred strains. The resulting extensive datasets can then be used to model causal interactions and construct biologic networks.

7.4 Materials and methods

7.4.1 Animals

Male mice from the hybrid MDP panel were purchased from the Jackson Labs (Bar Harbor, ME). Mice were between 6 and 10 weeks of age and to ensure adequate acclimatization to a common environment the mice were aged until 16 weeks of age. All mice were maintained on a chow diet (Ralston-Purina Co, St. Louis, Mo) until sacrifice at 16 weeks of age the mice. A complete list of strain included in the study is listed in Table 7.1.

7.4.2 Phenotypes/ phenotyping protocols

Body composition

At 16 weeks of age whole body fat, fluids and lean tissue mass of isoflurane-anaesthetised mice were determined using a Bruker Optics Minispec nuclear magnetic resonance (NMR) analyzer (The Woodlands, TX, USA) according to the manufacturer's recommendations.

Plasma measurements

Following a 16 hour fast. Mice were bled retro-orbitally under isoflurane anaesthesia. Plasma lipids were determined as previously described[140]. Glucose levels were determined using commercially available kits from Sigma (St Louis, MO, USA). Insulin levels were measured using commercial ELISA kits from ALPCO Diagnostics. All measurements were performed in duplicate or triplicate according to the manufacturer's instructions.

Sacrifice

Mice were anethesized with isoflourane, cervically dislocated and the mass of individual tissues depots (heart, kidney, retroperitoneal fat pad, epididymal fat pad, subcutaneous fat pad, and omental fat pad) were determined by dissecting and weighing each fat pad separately after the mice were euthanized.

7.4.3 Genotyping

Inbred strains were previously genotyped by the Broad Institute (Kirby et al, submitted), and they are combined with the genotypes from Wellcome Trust Center for Human Genetics (WTCHG). Genotypes of recombinant inbred strains at the Broad SNPs were inferred from WTCHG genotypes by interpolating alleles at polymorphic SNPs among parental strains, calling ambiguous genotypes missing. Of the 140,000 SNPs available at the Broad Institute, 107,145 were informative with an allele frequency greater then 5% and where used for GWAS. To calculate the linkage disequilibrium between markers (LD), we calculated the Pearson's pair-wise correlation coefficient between all pairs of markers for each chromosome. LD blocks were defined as groups of SNPs with a R2 greater then 0.7. To determine the average correlation between markers for each chromosome, we generated a distribution of mean r2 values for all pairs of informative markers with allele frequency greater than 5% at various distances from each other, using increasing window sizes of 100kb bins. We then took the mean for each window size using all 20 mouse chromosomes to determine the average r2 for each window size across the genome. The average correlation in chromosome 1 for markers 100kb apart is $r^2 = 0.7$, for markers 1Mb apart is $r^2 = 0.5$.

7.4.4 RNA isolation and expression profiling

Initial profiling studies were performed on liver tissue. Flash frozen samples were weighed and homogenized in Qiazol according to the manufacturerâĂŹs protocol. Following homogenization livers were isolated in RNeasy 96 columns (Qiagen) using the manufacturers protocol. RNA integrity was confirmed using the Agilent 2100 Bioanalyzer (Agilent, Palo Alto, CA).

7.4.5 Gene expression analysis

Gene expression was measured on 3 mice from each stain in the HMDP using affymetrix 430A microarray. The Microarray Suite 5.0 software (Affymetrix) was used to analyze image data and make the absolute call for each measurement. The array data were normalized with the Robust MultiarrayRMA method (RMA). Each probeset was treated as an individual trait association analysis was performed and corrected for using EMMA. Since EMMA is orders of magnitude faster than other implementations commonly used, we were able to perform statistical analyses for all pairs of transcripts and genome wide markers in a few hours using a cluster of 50 processors. Of the 107k SNPs in the MDP, those with less then 5% allele frequency were removed from the analysis. We further characterized the eQTL into distal or local eQTL. We defined an eQTL as local if the peak association signal was within a 10Mb sliding window of the gene(s) physical location. We then calculated the average distance between these cis eSNPs and transcription start site of the corresponding gene(s) transcription start site.

7.4.6 Genome-wide association mapping accounting for population structure

We applied the following linear mixed model to account for the population structure and genetic relatedness among strains in the genome-wide association mapping [103].

$$y = \mu + x\beta + u + e \tag{7.1}$$

where μ represents mean, x represents SNP effect, u represents random effects due to genetic relatedness with $\mathrm{Var}(u) = \sigma_g^2 K$ and $\mathrm{Var}(e) = \sigma_e^2 I$, where K represents IBS (identity-by-descent) matrix across all genotypes. A restricted maximum likelihood (REML) estimate of σ_g^2 and σ_e^2 are computed using EMMA (Efficient Mixed Model Association), and the association mapping is performed based on the estimated variance component with a standard F test to test $\beta \neq 0$.

7.4.7 Estimation of power and mapping resolution

We evaluated the statistical power of the HMDP through simulation studies, with various parameters including the variance explained by SNP, variance explained by genetic background, and variance explained by random errors, and the number of repeated measurement per strain. For the comparison of power with single RI set or classical inbred only studies, we selected subset of the simulated phenotypes for each RI or CI set and evaluated the power in the same way. Since there are 8 possibilities of SNPs being polymorphic among three sets of RI strains, the putative causal SNPs are categorized into 8 classes and power is evaluated for each class.

The mapping resolution is evaluated using the cis-acting eQTLs. In addition to the eQTL mapping described above, we selected a subset of expression dataset involving BXD strains only, and performed association mapping. We evaluated the mapping resolution using the top 1,000 probes with strongest cis-acting eSNPs as the distance between the gene and the strongest cis-eSNPs. When multiple SNPs are perfectly linked, the maximum distance is considered as the mapping resolution.

7.4.8 Genome-wide significance threshold

Genome-wide significance threshold in genome-wide association mapping is determined by the family-wise error rate (FWER) as the probability of observing one or more false positives across all SNPs per phenotype. Since the nearby SNPs are highly correlated with each other, applying Bonferroni correction imposing independence assumption among SNPs will lead to overly conservative estimate of significance threshold. Permutation test is a standard procedure to accurately account for multiple testing, but under the effect from population structure, permutation will break the relationship between the phenotype and the population structure and may lead an anti-conservative estimate of significance threshold. We used parametric bootstrapping to estimate the genome-wide threshold under various levels of population structure effect. It has been previously shown that parametric bootstrapping provide almost the same estimates of significance threshold[45]. We confirmed it by comparing the genome-wide significance levels between permutation and parametric bootstrapping where the phenotypes are simulated by multivariate normal distribution. We ran 100 different sets of permutation test and parametric bootstrapping of size 1,000, and observed that the mean and standard error of the genome-wide significance threshold at FWER of 0.05 were $3.9 \times 10^{-6} \pm 0.3 \times 10^{-6}$, and $4.0 \times 10^{-6} \pm 0.3 \times 10^{-6}$, respectively. This is approximately an order of magnitude larger than the significance threshold obtained by Bonferroni correction (4.6×10^{-7}). We also performed parametric bootstrapping under simulated the genetic background effect from population structure using EMMA. With 50% and 100% of variance explained by genetic background, the thresholds were determined to be $1.6 \times 10^{-6} \pm 0.2 \times 10^{-6}$ and $1.7 \times 10^{-6} \pm 0.2 \times 10^{-6}$. The reduction in the significance threshold compared to no genetic background effect is due to the fact that inter-SNP correlation due to long-range LDs reduces when conditioning on the population structure. Because LD spans longer for RI strains than classical inbreds only or the

HMDP panels, the significance threshold for a subset of the strains can dramatically differ. We used the parametric bootstrapping to estimate the significance threshold for each set of RI strains and classical inbreds. The estimated genome-wide significance threshold was $7.5 \times 10^{-5} \pm 0.5 \times 10^{-5}$ for BXD, $7.5 \times 10^{-5} \pm 0.4 \times 10^{-5}$ for AXB/BXA, and $1.1 \times 10^{-4} \pm 0.1 \times 10^{-4}$ for BXH, and $1.7 \times 10^{-6} \pm 0.2 \times 10^{-6}$ for classical inbreds. We used these thresholds to estimate the genome-wide power of each subset of strains in Figure 7.1 and Table 7.2

7.4.9 Validation of clinical and expression associations

We downloaded all available QTL studies for high density lipoproteins and compared these QTL to our association results. We also compare eQTL and clinical HDL QTL from 2 previously reported, independent crosses of C3H/HeJ and C57BL6/J to our HMDP results to demonstrate the improved resolution of the approach[217]

Chapter 7, is currently being prepared for submission for publication of the material. Brian J. Bennett, Charles R. Farber, Luz Orozco, Hyun Min Kang, Anatole Chanzalpour, Todd Kirchgessner, Peter Gargalovic, 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”. The dissertation author, Brian J. Bennett, Charles R. Farber, and Luz Orozco are the primary investigators and authors of this paper.

Chapter 8

Conclusion and future work

8.1 Summary and conclusion

In this thesis, I first focused on understanding the structure of genetic variation among different organisms such as human and mouse. Understanding the patterns of genetic variation within a population is one of the key steps for identifying the genetic variants associated with complex traits. I presented an extensive haplotype analysis of the mouse HapMap resource across 94 commonly used inbred mouse strains, in conjunction with the resequencing-based resource containing 8.27 million SNPs among 15 common inbred strains[67]. I also developed an adaptive and efficient algorithm that allows very accurate imputation of unobserved genotypes from resequenced strains. The algorithm shows ten times smaller error than the previous mouse imputation method, and it is also shown to be robust in human genotype imputation.

I also presented statistical methods to account for confounding effects due unmodeled factors to in association mapping and expression studies. In genome-wide association mapping, it is widely known that the population structure and genetic relatedness confounds the association mapping, significantly inflating false positive rates. Such confounding effects are especially substantial in association mapping among inbred mouse strains complexly related with each other, and it is known that mixed models accounting for the entire genetic relatedness matrix as an additional variance component robustly correct for the inflated false positives. However, mixed models were not efficient enough to be practically used for genome-wide association studies. I presented an efficient mixed model association (EMMA) method that is several thousand times more efficient than previous methods by leveraging a single spectral decomposition to dramatically reduce the time complexity in the numerical optimization procedure[103]. It is shown to robustly and efficiently correct for various types of confounding effects due to heterogeneous population samples, from cryptic relatedness among carefully stratified samples, to complex familial relatedness among genetically isolated population. In the analysis of expression data, I demonstrated that the technical confounding effects often result in inflation of spurious associations or trans-regulatory bands, and the signature of the confounding effects are shown to be effectively characterized by the inter-sample correlation matrix across all genes. Across various types of expression data sets, I showed that mixed models robustly account for the confounding effects when the inter-sample correlation is used as a variance component[102]. As a result, the true positive rates in the expression quantitative trait loci (eQTL) mapping significantly increased, and the concordance between independent studies greatly improved.

Finally, I presented an effective design of systems genetics studies in model organism, mostly focusing on the mouse genetics. The association and linkage mapping strategies in model organisms have a long history with strong theoretical background. While linkage studies tend to provide more replicable results than association studies, the mapping resolution usually spans more than several megabases which is not fine enough to identify candidate genes. Relatively recent in-silico mapping approach provides a higher mapping resolution using classical inbred strains, but the lack of power and increased false positive rates have recently been a great concern. The availability of near-complete genetic variation information across the common inbred lines enables us to perform a high-powered and high-resolution mapping of complex traits by combining multiple sets of recombinant inbred and classical inbred strains and by precisely accounting for their genetic relatedness using mixed model. This hybrid mouse diversity panel (HMDP) approach is shown to be effective in replicating previously known quantitative loci with higher resolution, and the expression studies based on the hybrid design also showed a great improvement both in power and in mapping resolution.

8.2 Future work

8.2.1 GWAS with unstratified populations

Large-scale genome-wide association studies (GWAS) have successfully identified many significant associations validated by previous studies or further replication studies. At the same time, there are still ongoing statistical challenges to identify associations with small effect size. Meta-analysis is shown to be a powerful approach to identify genetic variants with smaller effects. One of the challenges in the meta-analysis is the accumulative effect from cryptic relatedness across multiple data sets. The cryptic relatedness is referred to as the cause of over-dispersion of the test statistics in GWAS due to subtle level of distant familial relationship or population structure. Principal component analysis and Genomic Control has been widely used for correcting for the over-dispersion, but our recent simulation studies show that the fluctuations of test statistics at the end of distribution may be higher than expected, and that some SNPs are more vulnerable to the fluctuations than others. This phenomenon is more problematic in meta-analysis where the effects of the marker-specific fluctuations are accumulated across multiple studies, resulting in increased false positive rates. Mixed models are shown to be more robust against such fluctuations, and they can be used as an alternative method that is more suitable for meta-analysis.

While current large-scale association studies spend a lager amount of effort for avoiding population stratification issues, some of association study subjects are sampled from a heterogeneous population such as admixed population or partially inbred population with complex history among the founder individuals. Our recent application of mixed models on an isolated founder population from the Pacific Island of Kosrae suggests that the association mapping with heterogeneous population may be generally possible with the availability of high-density genotype information, even when the history of the population samples are unknown. This approach is readily applicable to many ongoing association studies involving heterogeneous or admixed populations. Moreover, multiple sets of association study samples with different population background may be combined together in a genotype level rather than a meta-analysis level so that the genetic relatedness between different groups may be accounted for more precisely.

8.2.2 Exploring multiple rare variants hypothesis

Although GWAS made a significant progress in identifying many novel associations, the overall phenotypic variance explained by the identified loci only explains a small fraction of the heritability estimated from family-based linkage studies. It has been suggested that such a big difference may be due to the assumption of common-disease-common-variant (CDCV) hypothesis which early stage of GWAS mainly relied on due to the technical limitation. Under CDCV hypothesis, it is assumed that relatively small number of causal variants explain the disease risk. However, recent studies on complex disease consistently suggest that the effect size per each genetic variant is likely to be marginal, and polygenic effects involving a large number of genes may more effective explain the genetic effect. Our results of variance component test on Finnish birth cohort samples and Kosrae population confirm that the entire genetic relatedness matrix largely explains the phenotypic variation at a similar level suggested by linkage studies.

With the advance of high-throughput sequencing technologies, it is becoming feasible to identify rare variants accounting for complex traits through various methods such as homozygosity mapping. In addition, many studies are searching for epistatic interactions between genes exhaustively or using multi-step approaches. My interest lies in developing statistical methods for testing the multipoint effects of a large set of genes or a large genomic region, where the variance component model can serve as an effective tool. More specifically, current large-scale association study samples can be used for testing the effect of a particular genomic region in a similar way to the traditional variance component tests by substituting the pedigree-based IBD (Identity-by-descent) matrix to the relatedness estimates from high-density genotypes. Similarly, a set of candidate genes can also be simultaneously tested using the variance component test, which may shed a light on understanding the structure of polygenic effects more systematically.

8.2.3 Capturing unmodeled confounding effects inherent in various high-throughput data

Current technologies for generating high-throughput biological data mainly rely on array hybridization procedure across various data types such as genotypes, CNVs (Copy Number Variations), gene expression, DNA methylation, and protein-binding arrays. It is widely known that the array hybridization technologies may be vulnerable to various types of systematic confounding effects such as bias in sample preparation, ozone levels, batch effects, and plate effects. Although randomization procedure greatly reduces the chances of confounding effects from such unmodeled factors, it has been shown that the confounding effects can significantly increase spurious signals with statistical tests with a large hypothesis space such as eQTL mapping.

My recent study demonstrated that the signature of technical confounding effects can be very well characterized as global correlation structure across the probes in expression studies, and mixed models are shown to be effective in correcting for the spurious associations induced by the global correlation. This approach can be extended to different types of array-based high-throughput data such as CNVs to correct for plate effects or batch effects. In this case, an important statistical challenge would be to precisely extract only the signature of the technical bias from the mixture of true biological effects and technical bias. CNV data sets can be a very effective application because there are many probes that can be used as negative controls, so the patterns of non-biological confounding effects can be captured by the negative controls.

8.2.4 Challenges in sequence-based association mapping

High-throughput sequencing technologies are expected to drive the next generation genetics and genomics research by providing more accurate and comprehensive profiling of genetic variation and other intermediate phenotypes such as gene expression. Many sequence-based genotypes resources are currently being collected in a large scale, and there are many known and unknown statistical and computational challenges to comprehensively understand the structure of genetic variation, to perform association mapping, and to identify genetic variants causally affecting complex disease traits.

Haplotype assembly is one of the important and challenging problems in the analysis of sequence-based genotype data. The development of longer read and paired-end ditags enables us to accurately infer the haplotype phase. The haplotype assembly has been shown to be effective even in the individual genome level when the read is very long, but the complexity of the problem significantly increases with shorter reads at the size of current technologies. In such cases, combining information from multiple individual genomes can considerably improve the accuracy of the haplotype assembly, but more computationally efficient algorithms are required in practice with minimum loss of accuracy. In addition, a systematic error model accounting for heterogeneous nature of sequencing errors can also greatly improve the accuracy of haplotype assembly.

Another important application with the availability of sequence-based genotype data is accurate genotype imputation with a large number of reference samples. I developed an adaptive and efficient genotype imputation method that can estimate the mutation and recombination parameters from the observed data using EM (Expectation-Maximization) algorithm, which can account for the individual level of difference. This approach can be combined with the imputation methods such as MACH (MArkov Chain Haplotyping), which accounts for the per-marker difference in the mutational and transitional parameters by leveraging a large number of genotyped samples. The combination of these two approaches is expected to provide us with accurate genotype imputation accounting for both individual and marker level differences of the parameters, which can be especially useful for imputation with heterogeneous or admixed population.

Bibliography

[1] A. Agresti and J. Wiley. *Categorical data analysis*. Wiley New York, 1990.

[2] Shahana Ahmed, Gilles Thomas, Maya Ghoussaini, Catherine S. Healey, Manjeet K. Humphreys, Radka Platte, Jonathan Morrison, Melanie Maranian, Karen A. Pooley, Robert Luben, Diana Eccles, D. Gareth Evans, Olivia Fletcher, Nichola Johnson, Isabel Dos Santos Silva, Julian Peto, Michael R. Stratton, Nazneen Rahman, Kevin Jacobs, Ross Prentice, Garnet L. Anderson, Aleksandar Rajkovic, J. David Curb, Regina G. Ziegler, Christine D. Berg, Saundra S. Buys, Catherine A. McCarty, Heather Spencer Feigelson, Eugenia E. Calle, Michael J. Thun, W. Ryan Diver, Stig Bojesen, BÃÿrge G. Nordestgaard, Henrik Flyger, Thilo DÃűrk, Peter SchÃijrmann, Peter Hillemanns, Johann H. Karstens, Natalia V. Bogdanova, Natalia N. Antonenkova, Iosif V. Zalutsky, Marina Bermisheva, Sardana Fedorova, Elza Khusnutdinova, SEARCH, Daehee Kang, Keun-Young Y. Yoo, Dong Young Noh, Sei-Hyun H. Ahn, Peter Devilee, Christi J. van Asperen, R. A. E. M. Tollenaar, Caroline Seynaeve, Montserrat Garcia-Closas, Jolanta Lissowska, Louise Brinton, Beata Peplonska, Heli Nevanlinna, Tuomas Heikkinen, Kristiina AittomÃďki, Carl Blomqvist, John L. Hopper, Melissa C. Southey, Letitia Smith, Amanda B. Spurdle, Marjanka K. Schmidt, Annegien Broeks, Richard R. van Hien, Sten Cornelissen, Roger L. Milne, Gloria Ribas, Anna GonzÃąlez-Neira, Javier Benitez, Rita K. Schmutzler, Barbara Burwinkel, Claus R. Bartram, Alfons Meindl, Hiltrud Brauch, Christina Justenhoven, Ute Hamann, The GENICA Consortium, Jenny Chang-Claude, Rebecca Hein, Shan Wang-Gohrke, Annika Lindblom, Sara Margolin, Arto Mannermaa, Veli-Matti M. Kosma, Vesa Kataja, Janet E. Olson, Xianshu Wang, Zachary Fredericksen, Graham G. Giles, Gianluca Severi, Laura Baglietto, Dallas R. English, Susan E. Hankinson, David G. Cox, Peter Kraft, Lars J. Vatten, Kristian Hveem, Merethe Kumle, Alice Sigurdson, Michele Doody, Parveen Bhatti, Bruce H. Alexander, Maartje J. Hooning, Ans M. W. van den Ouweland, Rogier A. Oldenburg, Mieke Schutte, Per Hall, Kamila Czene, Jianjun Liu, Yuqing Li, Angela Cox, Graeme Elliott, Ian Brock, Malcolm W. R. Reed, Chen-Yang Y. Shen, Jyh-Cherng C. Yu, Giu-Cheng C. Hsu, Shou-Tung T. Chen, Hoda Anton-Culver, Argyrios Ziogas, Irene L. Andrulis, Julia A. Knight, kConFab, Australian Ovarian Cancer Study Group, Jonathan Beesley, Ellen L. Goode, Fergus Couch, Georgia Chenevix-

Trench, Robert N. Hoover, Bruce A. J. Ponder, David J. Hunter, Paul D. P. Pharoah, Alison M. Dunning, Stephen J. Chanock, and Douglas F. Easton. Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2. *Nat Genet*, 3 2009.

[3] Joshua M. Akey, Shameek Biswas, Jeffrey T. Leek, and John D. Storey. On the design and analysis of gene expression studies in human populations. *Nat Genet*, 39(7):807–8; author reply 808–9, 7 2007.

[4] Hooman Allayee, Anatole Ghazalpour, and Aldons J. Lusis. Using mice to dissect genetic factors in atherosclerosis. *Arterioscler Thromb Vasc Biol*, 23(9):1501–9, 9 2003.

[5] Orly Alter, Patrick O. Brown, and David Botstein. Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. *Proc Natl Acad Sci U S A*, 100(6):3351–6, 3 2003.

[6] David Altshuler, Mark J. Daly, and Eric S. Lander. Genetic mapping in human disease. *Science*, 322(5903):881–8, 11 2008.

[7] R. V. Anunciado, M. Nishimura, M. Mori, A. Ishikawa, S. Tanaka, F. Horio, T. Ohno, and T. Namikawa. Quantitative trait loci for body weight in the intercross between sm/j and a/j mice. *Exp Anim*, 50(4):319–24, 7 2001.

[8] MarÃŋa JosÃľ Aranzana, Sung Kim, Keyan Zhao, Erica Bakker, Matthew Horton, Katrin Jakob, Clare Lister, John Molitor, Chikako Shindo, Chunlao Tang, Christopher Toomajian, Brian Traw, Honggang Zheng, Joy Bergelson, Caroline Dean, Paul Marjoram, and Magnus Nordborg. Genome-wide association mapping in arabidopsis identifies previously known flowering time and pathogen resistance genes. *PLoS Genet*, 1(5):e60, 11 2005.

[9] M. Arbelbide, J. Yu, and R. Bernardo. Power of mixed-model qtl mapping from phenotypic, pedigree and marker data in self-pollinated crops. *Theor Appl Genet*, 112(5):876–84, 3 2006.

[10] P. Armitage. Tests for linear trends in proportions and frequencies. *Biometrics*, pages 375–386, 1955.

[11] Yurii S. Aulchenko, Samuli Ripatti, Ida Lindqvist, Dorret Boomsma, Iris M. Heid, Peter P. Pramstaller, Brenda W. J. H. Penninx, A. Cecile J. W. Janssens, James F. Wilson, Tim Spector, Nicholas G. Martin, Nancy L. Pedersen, Kirsten Ohm Kyvik, Jaakko Kaprio, Albert Hofman, Nelson B. Freimer, Marjo-Riitta R. Jarvelin, Ulf Gyllensten, Harry Campbell, Igor Rudan, Asa Johansson, Fabio Marroni, Caroline Hayward, Veronique Vitart, Inger Jonasson, Cristian Pattaro, Alan Wright, Nick Hastie, Irene Pichler, Andrew A. Hicks, Mario Falchi, Gonneke Willemsen, Jouke-Jan J. Hottenga, Eco J. C. de Geus, Grant W. Montgomery, John Whitfield, Patrik Magnusson, Juha Saharinen,

Markus Perola, Kaisa Silander, Aaron Isaacs, Eric J. G. Sijbrands, Andre G. Uitterlinden, Jacqueline C. M. Witteman, Ben A. Oostra, Paul Elliott, Aimo Ruokonen, Chiara Sabatti, Christian Gieger, Thomas Meitinger, Florian Kronenberg, Angela DÃűring, H-Erich E. Wichmann, Johannes H. Smit, Mark I. McCarthy, Cornelia M. van Duijn, Leena Peltonen, and ENGAGE Consortium. Loci influencing lipid levels and coronary heart disease risk in 16 european population cohorts. *Nat Genet*, 41(1):47–55, 1 2009.

[12] Silviu-Alin A. Bacanu, Bernie Devlin, and Kathryn Roeder. Association studies for quantitative traits in structured populations. *Genet Epidemiol*, 22(1):78–93, 1 2002.

[13] Jeffrey C. Barrett, Sarah Hansoul, Dan L. Nicolae, Judy H. Cho, Richard H. Duerr, John D. Rioux, Steven R. Brant, Mark S. Silverberg, Kent D. Taylor, M. Michael Barmada, Alain Bitton, Themistocles Dassopoulos, Lisa Wu Datta, Todd Green, Anne M. Griffiths, Emily O. Kistner, Michael T. Murtha, Miguel D. Regueiro, Jerome I. Rotter, L. Philip Schumm, A. Hillary Steinhart, Stephan R. Targan, Ramnik J. Xavier, NIDDK IBD Genetics Consortium, CÃľcile Libioulle, Cynthia Sandor, Mark Lathrop, Jacques Belaiche, Olivier Dewit, Ivo Gut, Simon Heath, Debby Laukens, Myriam Mni, Paul Rutgeerts, AndrÃľ Van Gossum, Diana Zelenika, Denis Franchimont, Jean-Pierre P. Hugot, Martine de Vos, Severine Vermeire, Edouard Louis, Belgian-French IBD Consortium, Wellcome Trust Case Control Consortium, Lon R. Cardon, Carl A. Anderson, Hazel Drummond, Elaine Nimmo, Tariq Ahmad, Natalie J. Prescott, Clive M. Onnie, Sheila A. Fisher, Jonathan Marchini, Jilur Ghori, Suzannah Bumpstead, Rhian Gwilliam, Mark Tremelling, Panos Deloukas, John Mansfield, Derek Jewell, Jack Satsangi, Christopher G. Mathew, Miles Parkes, Michel Georges, and Mark J. Daly. Genome-wide association defines more than 30 distinct susceptibility loci for crohn's disease. *Nat Genet*, 40(8):955–62, 8 2008.

[14] Marc Bauchet, Brian McEvoy, Laurel N. Pearson, Ellen E. Quillen, Tamara Sarkisian, Kristine Hovhannesyan, Ranjan Deka, Daniel G. Bradley, and Mark D. Shriver. Measuring european population stratification with microarray genotype data. *Am J Hum Genet*, 80(5):948–56, 5 2007.

[15] Lara E. Bauman, Janet S. Sinsheimer, Eric M. Sobel, and Kenneth Lange. Mixed effects models for quantitative trait loci mapping with inbred strains. *Genetics*, 180(3):1743–61, 11 2008.

[16] J. A. Beck, S. Lloyd, M. Hafezparast, M. Lennon-Pierce, J. T. Eppig, M. F. Festing, and E. M. Fisher. Genealogies of mouse inbred strains. *Nat Genet*, 24(1):23–5, 1 2000.

[17] J. K. Belknap. Effect of within-strain sample size on qtl detection and mapping using recombinant inbred mouse strains. *Behav Genet*, 28(1):29–38, 1 1998.

[18] Lars Bertram, Christoph Lange, Kristina Mullin, Michele Parkinson, Monica Hsiao, Meghan F. Hogan, Brit M. M. Schjeide, Basavaraj Hooli, Jason Divito, Iuliana Ionita, Hongyu Jiang, Nan Laird, Thomas Moscarillo, Kari L. Ohlsen, Kathryn Elliott, Xin Wang, Diane Hu-Lince, Marie Ryder, Amy Murphy, Steven L. Wagner, Deborah Blacker, K. David Becker, and Rudolph E. Tanzi. Genome-wide association analysis reveals putative alzheimer's disease susceptibility loci in addition to apoe. *Am J Hum Genet*, 83(5):623–32, 11 2008.

[19] Tanmoy Bhattacharya, Marcus Daniels, David Heckerman, Brian Foley, Nicole Frahm, Carl Kadie, Jonathan Carlson, Karina Yusim, Ben McMahon, Brian Gaschen, Simon Mallal, James I. Mullins, David C. Nickle, Joshua Herbeck, Christine Rousseau, Gerald H. Learn, Toshiyuki Miura, Christian Brander, Bruce Walker, and Bette Korber. Founder effects in the assessment of hiv polymorphisms and hla allele associations. *Science*, 315(5818):1583–6, 3 2007.

[20] C. E. Bishop, P. Boursot, B. Baron, F. Bonhomme, and D. Hatat. Most classical mus musculus domesticus laboratory mouse strains carry a mus musculus musculus y chromosome. *Nature*, 315(6014):70–2, 1985.

[21] F. Bonhomme, J. L. Guenet, B. Dod, K. Moriwaki, and G. Bulfield. The polyphyletic origin of laboratory inbred mice and their rate of evolution. *Biological Journal of the Linnean Society*, 30(1):51–58, 1987.

[22] Ingrid B. Borecki and Michael A. Province. Linkage and association: basic concepts. *Adv Genet*, 60:51–74, 2008.

[23] Justin O. Borevitz, Samuel P. Hazen, Todd P. Michael, Geoffrey P. Morris, Ivan R. Baxter, Tina T. Hu, Huaming Chen, Jonathan D. Werner, Magnus Nordborg, David E. Salt, Steve A. Kay, Joanne Chory, Detlef Weigel, Jonathan D. G. Jones, and Joseph R. Ecker. Genome-wide patterns of single-feature polymorphism in arabidopsis thaliana. *Proc Natl Acad Sci U S A*, 104(29):12057–62, 7 2007.

[24] William S. Branham, Cathy D. Melvin, Tao Han, Varsha G. Desai, Carrie L. Moland, Adam T. Scully, and James C. Fuscoe. Elimination of laboratory ozone leads to a dramatic improvement in the reproducibility of microarray gene expression measurements. *BMC Biotechnol*, 7:8, 2007.

[25] Rachel B. Brem and Leonid Kruglyak. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc Natl Acad Sci U S A*, 102(5):1572–7, 2 2005.

[26] Rachel B. Brem, John D. Storey, Jacqueline Whittle, and Leonid Kruglyak. Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature*, 436(7051):701–3, 8 2005.

[27] Rachel B. Brem, GaÃńl Yvert, Rebecca Clinton, and Leonid Kruglyak. Genetic dissection of transcriptional regulation in budding yeast. *Science*, 296(5568):752–5, 4 2002.

[28] N. E. Breslow and D. G. Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, pages 9–25, 1993.

[29] Leonid Bystrykh, Ellen Weersing, Bert Dontje, Sue Sutton, Mathew T. Pletcher, Tim Wiltshire, Andrew I. Su, Edo Vellenga, Jintao Wang, Kenneth F. Manly, Lu Lu, Elissa J. Chesler, Rudi Alberts, Ritsert C. Jansen, Robert W. Williams, Michael P. Cooke, and Gerald de Haan. Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. *Nat Genet*, 37(3):225–32, 3 2005.

[30] Catarina D. Campbell, Elizabeth L. Ogburn, Kathryn L. Lunetta, Helen N. Lyon, Matthew L. Freedman, Leif C. Groop, David Altshuler, Kristin G. Ardlie, and Joel N. Hirschhorn. Demonstrating stratification in a european american population. *Nat Genet*, 37(8):868–72, 8 2005.

[31] Lawrence W. Castellani, Cara N. Nguyen, Sarada Charugundla, Michael M. Weinstein, Chau X. Doan, William S. Blaner, Nuttaporn Wongsiriroj, and Aldons J. Lusis. Apolipoprotein aii is a regulator of very low density lipoprotein metabolism and insulin resistance. *J Biol Chem*, 283(17):11633–44, 4 2008.

[32] Alessandra C. L. Cervino, Ariel Darvasi, Mohammad Fallahi, Christopher C. Mader, and Nicholas F. Tsinoremas. An integrated in silico gene mapping strategy in inbred mice. *Genetics*, 175(1):321–33, 1 2007.

[33] M. Chee, R. Yang, E. Hubbell, A. Berno, X. C. Huang, D. Stern, J. Winkler, D. J. Lockhart, M. S. Morris, and S. P. Fodor. Accessing genetic information with high-density dna arrays. *Science*, 274(5287):610–4, 10 1996.

[34] Elissa J. Chesler, Lu Lu, Siming Shou, Yanhua Qu, Jing Gu, Jintao Wang, Hui Chen Hsu, John D. Mountz, Nicole E. Baldwin, Michael A. Langston, David W. Threadgill, Kenneth F. Manly, and Robert W. Williams. Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nat Genet*, 37(3):233–42, 3 2005.

[35] Vivian G. Cheung, Richard S. Spielman, Kathryn G. Ewens, Teresa M. Weber, Michael Morley, and Joshua T. Burdick. Mapping determinants of human gene expression by regional and genome-wide association. *Nature*, 437(7063):1365–9, 10 2005.

[36] M. Chicurel. Faster, better, cheaper genotyping. *Nature*, 412(6847):580–2, 8 2001.

[37] Yoonha Choi, Ellen M. Wijsman, and Bruce S. Weir. Case-control association testing in the presence of unknown relationships. *Genet Epidemiol*, 3 2009.

[38] Gary A. Churchill. Fundamentals of experimental design for cdna microarrays. *Nat Genet*, 32 Suppl:490–5, 12 2002.

[39] Gary A. Churchill, David C. Airey, Hooman Allayee, Joe M. Angel, Alan D. Attie, Jackson Beatty, William D. Beavis, John K. Belknap, Beth Bennett, Wade Berrettini, Andre Bleich, Molly Bogue, Karl W. Broman, Kari J. Buck, Ed Buckler, Margit Burmeister, Elissa J. Chesler, James M. Cheverud, Steven Clapcote, Melloni N. Cook, Roger D. Cox, John C. Crabbe, Wim E. Crusio, Ariel Darvasi, Christian F. Deschepper, R. W. Doerge, Charles R. Farber, Jiri Forejt, Daniel Gaile, Steven J. Garlow, Hartmut Geiger, Howard Gershenfeld, Terry Gordon, Jing Gu, Weikuan Gu, Gerald de Haan, Nancy L. Hayes, Craig Heller, Heinz Himmelbauer, Robert Hitzemann, Kent Hunter, Hui-Chen C. Hsu, Fuad A. Iraqi, Boris Ivandic, Howard J. Jacob, Ritsert C. Jansen, Karl J. Jepsen, Dabney K. Johnson, Thomas E. Johnson, Gerd Kempermann, Christina Kendziorski, Malak Kotb, R. Frank Kooy, Bastien Llamas, Frank Lammert, Jean-Michel M. Lassalle, Pedro R. Lowenstein, Lu Lu, Aldons Lusis, Kenneth F. Manly, Ralph Marcucio, Doug Matthews, Juan F. Medrano, Darla R. Miller, Guy Mittleman, Beverly A. Mock, Jeffrey S. Mogil, Xavier Montagutelli, Grant Morahan, David G. Morris, Richard Mott, Joseph H. Nadeau, Hiroki Nagase, Richard S. Nowakowski, Bruce F. O'Hara, Alexander V. Osadchuk, Grier P. Page, Beverly Paigen, Kenneth Paigen, Abraham A. Palmer, Huei-Ju J. Pan, Leena Peltonen-Palotie, Jeremy Peirce, Daniel Pomp, Michal Pravenec, Daniel R. Prows, Zhonghua Qi, Roger H. Reeves, John Roder, Glenn D. Rosen, Eric E. Schadt, Leonard C. Schalkwyk, Ze'ev Seltzer, Kazuhiro Shimomura, Siming Shou, Mikko J. SillanpÃďÃď, Linda D. Siracusa, Hans-Willem W. Snoeck, Jimmy L. Spearow, Karen Svenson, Lisa M. Tarantino, David Threadgill, Linda A. Toth, William Valdar, Fernando Pardo-Manuel de Villena, Craig Warden, Steve Whatley, Robert W. Williams, Tim Wiltshire, Nengjun Yi, Dabao Zhang, Min Zhang, Fei Zou, and Complex Trait Consortium. The collaborative cross, a community resource for the genetic analysis of complex traits. *Nat Genet*, 36(11):1133–7, 11 2004.

[40] Jonathan C. Cohen, Robert S. Kiss, Alexander Pertsemlidis, Yves L. Marcel, Ruth McPherson, and Helen H. Hobbs. Multiple rare alleles contribute to low plasma levels of hdl cholesterol. *Science*, 305(5685):869–72, 8 2004.

[41] F. S. Collins, L. D. Brooks, and A. Chakravarti. A dna polymorphism discovery resource for research on human genetic variation. *Genome Res*, 8(12):1229–31, 12 1998.

[42] Sonia Colombo, Daniela Ronchetti, Johan M. Thevelein, Joris Winderickx, and Enzo Martegani. Activation state of the ras2 protein and glucose-induced signaling in saccharomyces cerevisiae. *J Biol Chem*, 279(45):46715–22, 11 2004.

[43] C. M. Crainiceanu and D. Ruppert. Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, pages 165–185, 2004.

[44] Paul I. W. de Bakker, Gil McVean, Pardis C. Sabeti, Marcos M. Miretti, Todd Green, Jonathan Marchini, Xiayi Ke, Alienke J. Monsuur, Pamela Whittaker, Marcos Delgado, Jonathan Morrison, Angela Richardson, Emily C. Walsh, Xiaojiang Gao, Luana Galver, John Hart, David A. Hafler, Margaret Pericak-Vance, John A. Todd, Mark J. Daly, John Trowsdale, Cisca Wijmenga, Tim J. Vyse, Stephan Beck, Sarah Shaw Murray, Mary Carrington, Simon Gregory, Panos Deloukas, and John D. Rioux. A high-resolution hla and snp haplotype map for disease association studies in the extended human mhc. *Nat Genet*, 38(10):1166–72, 10 2006.

[45] Paul I. W. de Bakker, Roman Yelensky, Itsik Pe'er, Stacey B. Gabriel, Mark J. Daly, and David Altshuler. Efficiency and power in genetic association studies. *Nat Genet*, 37(11):1217–23, 11 2005.

[46] Dirk-Jan J. de Koning and Chris S. Haley. Genetical genomics in humans and model organisms. *Trends Genet*, 21(7):377–81, 7 2005.

[47] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

[48] A. P. Dempster, D. B. Rubin, and R. K. Tsutakawa. Estimation in covariance components models. *Journal of the American Statistical Association*, pages 341–353, 1981.

[49] B. Devlin and N. Risch. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics*, 29(2):311–22, 9 1995.

[50] B. Devlin and K. Roeder. Genomic control for association studies. *Biometrics*, 55(4):997–1004, 12 1999.

[51] M. H. Doolittle, R. C. LeBoeuf, C. H. Warden, L. M. Bee, and A. J. Lusis. A polymorphism affecting apolipoprotein a-ii translational efficiency determines high density lipoprotein size and composition. *J Biol Chem*, 265(27):16380–8, 9 1990.

[52] Sudheer Doss, Eric E. Schadt, Thomas A. Drake, and Aldons J. Lusis. Cis-acting expression quantitative trait loci in mice. *Genome Res*, 15(5):681–91, 5 2005.

[53] Douglas F. Easton, Karen A. Pooley, Alison M. Dunning, Paul D. P. Pharoah, Deborah Thompson, Dennis G. Ballinger, Jeffery P. Struewing, Jonathan Morrison, Helen Field, Robert Luben, Nicholas Wareham, Shahana Ahmed,

Catherine S. Healey, Richard Bowman, SEARCH collaborators, Kerstin B. Meyer, Christopher A. Haiman, Laurence K. Kolonel, Brian E. Henderson, Loic Le Marchand, Paul Brennan, Suleeporn Sangrajrang, Valerie Gaborieau, Fabrice Odefrey, Chen-Yang Y. Shen, Pei-Ei E. Wu, Hui-Chun C. Wang, Diana Eccles, D. Gareth Evans, Julian Peto, Olivia Fletcher, Nichola Johnson, Sheila Seal, Michael R. Stratton, Nazneen Rahman, Georgia Chenevix-Trench, Stig E. Bojesen, BÃÿrge G. Nordestgaard, Christen K. Axelsson, Montserrat Garcia-Closas, Louise Brinton, Stephen Chanock, Jolanta Lissowska, Beata Peplonska, Heli Nevanlinna, Rainer Fagerholm, Hannaleena Eerola, Daehee Kang, Keun-Young Y. Yoo, Dong-Young Y. Noh, Sei-Hyun H. Ahn, David J. Hunter, Susan E. Hankinson, David G. Cox, Per Hall, Sara Wedren, Jianjun Liu, Yen-Ling L. Low, Natalia Bogdanova, Peter SchÃijrmann, Thilo DÃűrk, Rob A. E. M. Tollenaar, Catharina E. Jacobi, Peter Devilee, Jan G. M. Klijn, Alice J. Sigurdson, Michele M. Doody, Bruce H. Alexander, Jinghui Zhang, Angela Cox, Ian W. Brock, Gordon MacPherson, Malcolm W. R. Reed, Fergus J. Couch, Ellen L. Goode, Janet E. Olson, Hanne Meijers-Heijboer, Ans van den Ouweland, AndrÃľ Uitterlinden, Fernando Rivadeneira, Roger L. Milne, Gloria Ribas, Anna Gonzalez-Neira, Javier Benitez, John L. Hopper, Margaret McCredie, Melissa Southey, Graham G. Giles, Chris Schroen, Christina Justenhoven, Hiltrud Brauch, Ute Hamann, Yon-Dschun D. Ko, Amanda B. Spurdle, Jonathan Beesley, Xiaoqing Chen, kConFab, AOCS Management Group, Arto Mannermaa, Veli-Matti M. Kosma, Vesa Kataja, Jaana Hartikainen, Nicholas E. Day, David R. Cox, and Bruce A. J. Ponder. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*, 447(7148):1087–93, 6 2007.

[54] Valur Emilsson, Gudmar Thorleifsson, Bin Zhang, Amy S. Leonardson, Florian Zink, Jun Zhu, Sonia Carlson, Agnar Helgason, G. Bragi Walters, Steinunn Gunnarsdottir, Magali Mouy, Valgerdur Steinthorsdottir, Gudrun H. Eiriksdottir, Gyda Bjornsdottir, Inga Reynisdottir, Daniel Gudbjartsson, Anna Helgadottir, Aslaug Jonasdottir, Adalbjorg Jonasdottir, Unnur Styrkarsdottir, Solveig Gretarsdottir, Kristinn P. Magnusson, Hreinn Stefansson, Ragnheidur Fossdal, Kristleifur Kristjansson, Hjortur G. Gislason, Tryggvi Stefansson, Bjorn G. Leifsson, Unnur Thorsteinsdottir, John R. Lamb, Jeffrey R. Gulcher, Marc L. Reitman, Augustine Kong, Eric E. Schadt, and Kari Stefansson. Genetics of gene expression and its effect on disease. *Nature*, 452(7186):423–8, 3 2008.

[55] M. P. Epstein, W. L. Duren, and M. Boehnke. Improved inference of relationship for pairs of individuals. *Am J Hum Genet*, 67(5):1219–31, 11 2000.

[56] Charles R. Farber and Aldons J. Lusis. Integrating global gene expression analysis and genetics. *Adv Genet*, 60:571–601, 2008.

[57] Thomas L. Fare, Ernest M. Coffey, Hongyue Dai, Yudong D. He, Deborah A.

Kessler, Kristopher A. Kilian, John E. Koch, Eric LeProust, Matthew J. Marton, Michael R. Meyer, Roland B. Stoughton, George Y. Tokiwa, and Yanqun Wang. Effects of atmospheric ozone on microarray data quality. *Anal Chem*, 75(17):4672–5, 9 2003.

[58] J. Felsenstein. Phylogenies and the comparative method. *American Naturalist*, 125(1):1, 1985.

[59] J. Felsenstein and G. A. Churchill. A hidden markov model approach to variation among sites in rate of evolution. *Mol. Biol. Evol*, 13(1):93–104, 1 1996.

[60] S. D. Ferris, R. D. Sage, and A. C. Wilson. Evidence from mtdna sequences that common laboratory strains of inbred mice are descended from a single female. *Nature*, 295(5845):163–5, 1 1982.

[61] S. R. A. Fisher. The correlation between relatives on the supposition of mendelian inheritance. *Trans R Soc Edinb*, 52:399–433, 1918.

[62] W. M. Fitch and E. Margoliash. Construction of phylogenetic trees. *Science*, 155(760):279–84, 1 1967.

[63] Jonathan Flint and Richard Mott. Applying mouse complex-trait resources to behavioural genetics. *Nature*, 456(7223):724–7, 12 2008.

[64] Jonathan Flint, William Valdar, Sagiv Shifman, and Richard Mott. Strategies for mapping and cloning quantitative trait genes in rodents. *Nat Rev Genet*, 6(4):271–86, 4 2005.

[65] Sherry A. Flint-Garcia, Anne-CÃľline C. Thuillet, Jianming Yu, Gael Pressoir, Susan M. Romero, Sharon E. Mitchell, John Doebley, Stephen Kresovich, Major M. Goodman, and Edward S. Buckler. Maize association population: a high-resolution platform for quantitative trait locus dissection. *Plant J*, 44(6):1054–64, 12 2005.

[66] Eric J. Foss, Dragan Radulovic, Scott A. Shaffer, Douglas M. Ruderfer, Antonio Bedalov, David R. Goodlett, and Leonid Kruglyak. Genetic basis of proteome variation in yeast. *Nat Genet*, 39(11):1369–75, 11 2007.

[67] Kelly A. Frazer, Eleazar Eskin, Hyun Min Kang, Molly A. Bogue, David A. Hinds, Erica J. Beilharz, Robert V. Gupta, Julie Montgomery, Matt M. Morenzoni, Geoffrey B. Nilsen, Charit L. Pethiyagoda, Laura L. Stuve, Frank M. Johnson, Mark J. Daly, Claire M. Wade, and David R. Cox. A sequence-based variation map of 8.27 million snps in inbred mouse strains. *Nature*, 448(7157):1050–3, 8 2007.

[68] Kelly A. Frazer, Claire M. Wade, David A. Hinds, Nila Patil, David R. Cox, and Mark J. Daly. Segmental phylogenetic relationships of inbred mouse strains

revealed by fine-scale analysis of sequence variation across 4.6 mb of mouse genome. *Genome Res*, 14(8):1493–500, 8 2004.

[69] Ruth Frikke-Schmidt, BÃÿrge G. Nordestgaard, Gorm B. Jensen, and Anne Tybjaerg-Hansen. Genetic variation in abc transporter a1 contributes to hdl cholesterol in the general population. *J Clin Invest*, 114(9):1343–53, 11 2004.

[70] Anatole Ghazalpour, Sudheer Doss, Hyun Kang, Charles Farber, Ping-Zi Z. Wen, Alec Brozell, Ruth Castellanos, Eleazar Eskin, Desmond J. Smith, Thomas A. Drake, and Aldons J. Lusis. High-resolution mapping of gene expression using association in an outbred mouse stock. *PLoS Genet*, 4(8):e1000149, 2008.

[71] Anatole Ghazalpour, Sudheer Doss, Bin Zhang, Susanna Wang, Christopher Plaisier, Ruth Castellanos, Alec Brozell, Eric E. Schadt, Thomas A. Drake, Aldons J. Lusis, and Steve Horvath. Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS Genet*, 2(8):e130, 8 2006.

[72] A. R. Gilmour, R. Thompson, and B. R. Cullis. Average information reml: An efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics*, 51:1440–1450, 1995.

[73] H. U. Graser, S. P. Smith, and B. Tier. A derivative-free approach for estimating variance components in animal models by restricted maximum likelihood. *Journal of animal science*, 64(5):1362, 1987.

[74] Stephen C. Grubb, Terry P. Maddatu, Carol J. Bult, and Molly A. Bogue. Mouse phenome database. *Nucleic Acids Res*, 37(Database issue):D720–30, 1 2009.

[75] A. Grupe, S. Germer, J. Usuka, D. Aud, J. K. Belknap, R. F. Klein, M. K. Ahluwalia, R. Higuchi, and G. Peltz. In silico mapping of complex disease-related traits in mice. *Science*, 292(5523):1915–8, 6 2001.

[76] Xun Gu. Statistical framework for phylogenomic analysis of gene family expression profiles. *Genetics*, 167(1):531–42, 5 2004.

[77] Weihua Guan, Liming Liang, Michael Boehnke, and GonÃğalo R. Abecasis. Genotype-based matching to correct for population stratification in large-scale case-control genetic association studies. *Genet Epidemiol*, 1 2009.

[78] Kevin L. Gunderson, Frank J. Steemers, Grace Lee, Leo G. Mendoza, and Mark S. Chee. A genome-wide scalable snp genotyping assay using microarray technology. *Nat Genet*, 37(5):549–54, 5 2005.

[79] Yingying Guo, Peng Lu, Erin Farrell, Xun Zhang, Paul Weller, Mario Monshouwer, Jianmei Wang, Guochun Liao, Zhaomei Zhang, Steven Hu, John

Allard, Steve Shafer, Jonathan Usuka, and Gary Peltz. In silico and in vitro pharmacogenetic analysis in mice. *Proc Natl Acad Sci U S A*, 104(45):17735–40, 11 2007.

[80] Yingying Guo, Paul Weller, Erin Farrell, Paul Cheung, Bill Fitch, Douglas Clark, Shao-yong Y. Wu, Jianmei Wang, Guochun Liao, Zhaomei Zhang, John Allard, Janet Cheng, Anh Nguyen, Sharon Jiang, Steve Shafer, Jonathan Usuka, Mohammad Masjedizadeh, and Gary Peltz. In silico pharmacogenetics of warfarin metabolism. *Nat Biotechnol*, 24(5):531–6, 5 2006.

[81] U. GÃijldener, M. MÃijnsterkÃűtter, G. KastenmÃijller, N. Strack, J. van Helden, C. Lemer, J. Richelles, S. J. Wodak, J. GarcÃŋa-MartÃŋnez, J. E. PÃľrez-OrtÃŋn, H. Michael, A. Kaps, E. Talla, B. Dujon, B. AndrÃľ, J. L. Souciet, J. De Montigny, E. Bon, C. Gaillardin, and H. W. Mewes. Cygd: the comprehensive yeast genome database. *Nucleic Acids Res*, 33(Database issue):D364–8, 1 2005.

[82] Eran Halperin and Eleazar Eskin. Haplotype reconstruction from genotype data using imperfect phylogeny. *Bioinformatics*, 20(12):1842–9, 8 2004.

[83] O. J. Hardy and X. Vekemans. Spagedi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Molecular Ecology Notes*, 2(4):618–620, 2002.

[84] D. A. Harville. Biometrika, 1974.

[85] Agnar Helgason, BryndÃŋs YngvadÃşttir, Birgir Hrafnkelsson, Jeffrey Gulcher, and KÃąri StefÃąnsson. An icelandic example of the impact of population structure on association studies. *Nat Genet*, 37(1):90–5, 1 2005.

[86] C. R. Henderson. *Applications of linear models in animal breeding.* University of Guelph Guelph, CN, 1984.

[87] David A. Hinds, Laura L. Stuve, Geoffrey B. Nilsen, Eran Halperin, Eleazar Eskin, Dennis G. Ballinger, Kelly A. Frazer, and David R. Cox. Whole-genome patterns of common dna variation in three human populations. *Science*, 307(5712):1072–9, 2 2005.

[88] Lucy Huang, Yun Li, Andrew B. Singleton, John A. Hardy, GonÃğalo Abecasis, Noah A. Rosenberg, and Paul Scheet. Genotype-imputation accuracy across worldwide human populations. *Am J Hum Genet*, 84(2):235–50, 2 2009.

[89] Norbert Hubner, Caroline A. Wallace, Heike Zimdahl, Enrico Petretto, Herbert Schulz, Fiona Maciver, Michael Mueller, Oliver Hummel, Jan Monti, Vaclav Zidek, Alena Musilova, Vladimir Kren, Helen Causton, Laurence Game, Gabriele Born, Sabine Schmidt, Anita MÃijller, Stuart A. Cook, Theodore W. Kurtz, John Whittaker, Michal Pravenec, and Timothy J. Aitman. Integrated

transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nat Genet*, 37(3):243–53, 3 2005.

[90] T. Ideker, T. Galitski, and L. Hood. A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet*, 2:343–72, 2001.

[91] Folami Y. Ideraabdullah, Elena de la Casa-EsperÃşn, Timothy A. Bell, David A. Detwiler, Terry Magnuson, Carmen Sapienza, and Fernando Pardo-Manuel de Villena. Genetic and haplotype diversity among wild-derived mouse inbred strains. *Genome Res*, 14(10A):1880–7, 10 2004.

[92] SAS Institute. *SAS/STAT 9.1 User's Guide*. SAS Institute Inc. Cary, NC, 2004.

[93] InternationalHapMapConsortium. A second generation human haplotype map of over 3.1 million snps. *Nature*, 449(7164):851–61, 10 2007.

[94] Rafael A. Irizarry, Bridget Hobbs, Francois Collin, Yasmin D. Beazer-Barclay, Kristen J. Antonellis, Uwe Scherf, and Terence P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–64, 4 2003.

[95] Rafael A. Irizarry, Daniel Warren, Forrest Spencer, Irene F. Kim, Shyam Biswal, Bryan C. Frank, Edward Gabrielson, Joe G. N. Garcia, Joel Geoghegan, Gregory Germino, Constance Griffin, Sara C. Hilmer, Eric Hoffman, Anne E. Jedlicka, Ernest Kawasaki, Francisco MartÃŋnez-Murillo, Laura Morsberger, Hannah Lee, David Petersen, John Quackenbush, Alan Scott, Michael Wilson, Yanqin Yang, Shui Qing Ye, and Wayne Yu. Multiple-laboratory comparison of microarray platforms. *Nat Methods*, 2(5):345–50, 5 2005.

[96] Eveliina Jakkula, Karola RehnstrÃűm, Teppo Varilo, Olli P. H. PietilÃďinen, Tiina Paunio, Nancy L. Pedersen, Ulf deFaire, Marjo-Riitta R. JÃďrvelin, Juha Saharinen, Nelson Freimer, Samuli Ripatti, Shaun Purcell, Andrew Collins, Mark J. Daly, Aarno Palotie, and Leena Peltonen. The genome-wide patterns of variation expose significant substructure in a founder population. *Am J Hum Genet*, 83(6):787–94, 12 2008.

[97] Georg Jander, Susan R. Norris, Steven D. Rounsley, David F. Bush, Irena M. Levin, and Robert L. Last. Arabidopsis map-based cloning in the post-genome era. *Plant Physiol*, 129(2):440–50, 6 2002.

[98] R. C. Jansen and J. P. Nap. Genetical genomics: the added value from segregation. *Trends Genet*, 17(7):388–91, 7 2001.

[99] D. L. Johnson and R. Thompson. Restricted maximum likelihood estimation of variance components for univariate animal models using sparse matrix

techniques and average information. *Journal of Dairy Science*, 78(2):449–456, 1995.

[100] W. Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–27, 1 2007.

[101] Haja N. Kadarmideen, Peter von Rohr, and Luc L. G. Janss. From genetical genomics to systems genetics: potential applications in quantitative genomics and animal breeding. *Mamm Genome*, 17(6):548–64, 6 2006.

[102] Hyun Min Kang, Chun Ye, and Eleazar Eskin. Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics*, 180(4):1909–25, 12 2008.

[103] 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(3):1709–23, 3 2008.

[104] Elinor K. Karlsson, Izabella Baranowska, Claire M. Wade, Nicolette H. C. Salmon Hillbertz, Michael C. Zody, Nathan Anderson, Tara M. Biagi, Nick Patterson, Gerli Rosengren Pielberg, Edward J. Kulbokas, Kenine E. Comstock, Evan T. Keller, Jill P. Mesirov, Henrik von Euler, Olle KÃďmpe, Ake Hedhammar, Eric S. Lander, GÃűran Andersson, Leif Andersson, and Kerstin Lindblad-Toh. Efficient mapping of mendelian traits in dogs through genome-wide association. *Nat Genet*, 39(11):1321–8, 11 2007.

[105] K. Kathiresan, S. Manivannan, M. A. Nabeel, and B. Dhivya. Studies on silver nanoparticles synthesized by a marine fungus, penicillium fellutanum isolated from coastal mangrove sediment. *Colloids Surf B Biointerfaces*, 71(1):133–7, 6 2009.

[106] B. W. Kennedy, M. Quinton, and J. A. van Arendonk. Estimation of effects of single genes on quantitative traits. *J Anim Sci*, 70(7):2000–12, 7 1992.

[107] Joost J. B. Keurentjes, Jingyuan Fu, Inez R. Terpstra, Juan M. Garcia, Guido van den Ackerveken, L. Basten Snoek, Anton J. M. Peeters, Dick Vreugdenhil, Maarten Koornneef, and Ritsert C. Jansen. Regulatory network construction in arabidopsis by using genome-wide gene expression quantitative trait loci. *Proc Natl Acad Sci U S A*, 104(5):1708–13, 1 2007.

[108] J. F. C. Kingman. On the genealogy of large populations. *Journal of Applied Probability*, pages 27–43, 1982.

[109] Andrew Kirby, Hyun Min Kang, Claire M. Wade, Chris J. Cotsapas, Emrah Kostem, Buhm Han, Manuel Rivas, Molly A. Bogue, Kelly A Frazer, Frank M.

Johnson, Erica J. Beilharz, David R. Cox, Eleazar Eskin, and Mark J. Daly. A high density haplotype resouce of 94 inbred mouse strains.

[110] H. Kishino and M. Hasegawa. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from dna sequence data, and the branching order in hominoidea. *J Mol Evol*, 29(2):170–9, 8 1989.

[111] Hiroaki Kitano. Systems biology: a brief overview. *Science*, 295(5560):1662–4, 3 2002.

[112] Robert F. Klein, John Allard, Zafrira Avnur, Tania Nikolcheva, David Rotstein, Amy S. Carlos, Marie Shea, Ruth V. Waters, John K. Belknap, Gary Peltz, and Eric S. Orwoll. Regulation of bone mass in mice by the lipoxygenase gene alox15. *Science*, 303(5655):229–32, 1 2004.

[113] Nan M. Laird and Christoph Lange. Family-based methods for linkage and association analysis. *Adv Genet*, 60:219–52, 2008.

[114] U. Landegren, M. Nilsson, and P. Y. Kwok. Reading bits of genetic information: methods for single-nucleotide polymorphism analysis. *Genome Res*, 8(8):769–76, 8 1998.

[115] K. Lange. *Mathematical and statistical methods for genetic analysis*. Springer, 2002.

[116] Larry J. Leamy, Daniel Pomp, E. J. Eisen, and James M. Cheverud. Pleiotropy of quantitative trait loci for organ weights and limb bone lengths in mice. *Physiol Genomics*, 10(1):21–9, 7 2002.

[117] Su-In I. Lee, Dana Pe'er, AimÃl'e M. Dudley, George M. Church, and Daphne Koller. Identifying regulatory mechanisms using individual variation reveals key role for chromatin modification. *Proc Natl Acad Sci U S A*, 103(38):14062–7, 9 2006.

[118] Jeffrey T. Leek and John D. Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet*, 3(9):1724–35, 9 2007.

[119] Guochun Liao, Jianmei Wang, Jingshu Guo, John Allard, Janet Cheng, Anh Ng, Steve Shafer, Anne Puech, John D. McPherson, Dorothee Foernzler, Gary Peltz, and Jonathan Usuka. In silico genetics: identification of a functional element regulating h2-ealpha gene expression. *Science*, 306(5696):690–5, 10 2004.

[120] K. Lindblad-Toh, E. Winchester, M. J. Daly, D. G. Wang, J. N. Hirschhorn, J. P. Laviolette, K. Ardlie, D. E. Reich, E. Robinson, P. Sklar, N. Shah, D. Thomas, J. B. Fan, T. Gingeras, J. Warrington, N. Patil, T. J. Hudson, and E. S. Lander. Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. *Nat Genet*, 24(4):381–6, 4 2000.

[121] Kerstin Lindblad-Toh, Claire M. Wade, Tarjei S. Mikkelsen, Elinor K. Karlsson, David B. Jaffe, Michael Kamal, Michele Clamp, Jean L. Chang, Edward J. Kulbokas, Michael C. Zody, Evan Mauceli, Xiaohui Xie, Matthew Breen, Robert K. Wayne, Elaine A. Ostrander, Chris P. Ponting, Francis Galibert, Douglas R. Smith, Pieter J. DeJong, Ewen Kirkness, Pablo Alvarez, Tara Biagi, William Brockman, Jonathan Butler, Chee-Wye W. Chin, April Cook, James Cuff, Mark J. Daly, David DeCaprio, Sante Gnerre, Manfred Grabherr, Manolis Kellis, Michael Kleber, Carolyne Bardeleben, Leo Goodstadt, Andreas Heger, Christophe Hitte, Lisa Kim, Klaus-Peter P. Koepfli, Heidi G. Parker, John P. Pollinger, Stephen M. J. Searle, Nathan B. Sutter, Rachael Thomas, Caleb Webber, Jennifer Baldwin, and Eric S. Lander. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, 438(7069):803–19, 12 2005.

[122] M. J. Lindstrom and D. M. Bates. Newton-raphson and em algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83(404):1014–1022, 1988.

[123] Pengyuan Liu, Haris Vikis, Yan Lu, Daolong Wang, and Ming You. Large-scale in silico mapping of complex quantitative traits in inbred mice. *PLoS ONE*, 2(7):e651, 2007.

[124] Pengyuan Liu, Yian Wang, Haris Vikis, Anna Maciag, Daolong Wang, Yan Lu, Yan Liu, and Ming You. Candidate lung tumor susceptibility genes identified through whole-genome association analyses in inbred mice. *Nat Genet*, 38(8):888–95, 8 2006.

[125] Jennifer K. Lowe, Julian B. Maller, Itsik Pe'er, Benjamin M. Neale, Jacqueline Salit, Eimear E. Kenny, Jessica L. Shea, Ralph Burkhardt, J. Gustav Smith, Weizhen Ji, Martha Noel, Jia Nee Foo, Maude L. Blundell, Vita Skilling, Laura Garcia, Marcia L. Sullivan, Heather E. Lee, Anna Labek, Hope Ferdowsian, Steven B. Auerbach, Richard P. Lifton, Christopher Newton-Cheh, Jan L. Breslow, Markus Stoffel, Mark J. Daly, David M. Altshuler, and Jeffrey M. Friedman. Genome-wide association studies in an isolated founder population from the pacific island of kosrae. *PLoS Genet*, 5(2):e1000365, 2 2009.

[126] Aldons J. Lusis, Alan D. Attie, and Karen Reue. Metabolic syndrome: from epidemiology to systems biology. *Nat Rev Genet*, 9(11):819–30, 11 2008.

[127] M. Lynch and K. Ritland. Estimation of pairwise relatedness with molecular markers. *Genetics*, 152(4):1753–66, 8 1999.

[128] Brendan Maher. Personal genomes: The case of the missing heritability. *Nature*, 456(7218):18–21, 11 2008.

[129] Andrea Manica, William Amos, FranÃğois Balloux, and Tsunehiko Hanihara. The effect of ancient population bottlenecks on human phenotypic variation. *Nature*, 448(7151):346–8, 7 2007.

[130] Kenneth F. Manly, Jintao Wang, and Robert W. Williams. Weighting by heritability for detection of quantitative trait loci with microarray estimates of gene expression. *Genome Biol*, 6(3):R27, 2005.

[131] Teri A. Manolio. Cohort studies and the genetics of complex disease. *Nat Genet*, 41(1):5–6, 1 2009.

[132] Jonathan Marchini, Bryan Howie, Simon Myers, Gil McVean, and Peter Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet*, 39(7):906–13, 7 2007.

[133] E. P. Martins and T. F. Hansen. Phylogenies and the comparative method: a general approach to incorporating phylogenetic information into the analysis of interspecific data. *American Naturalist*, 149(4):646, 1997.

[134] Godfred L. Masinde, Xinmin Li, Weikuan Gu, Heather Davidson, Melanie Hamilton-Ulland, Jon Wergedal, Subburaman Mohan, and David J. Baylink. Quantitative trait loci (qtl) for lean body mass and body length in mrl/mpj and sjl/j f(2) mice. *Funct Integr Genomics*, 2(3):98–104, 8 2002.

[135] Hajime Matsuzaki, Shoulian Dong, Halina Loi, Xiaojun Di, Guoying Liu, Earl Hubbell, Jane Law, Tam Berntsen, Monica Chadha, Henry Hui, Geoffrey Yang, Giulia C. Kennedy, Teresa A. Webster, Simon Cawley, P. Sean Walsh, Keith W. Jones, Stephen P. A. Fodor, and Rui Mei. Genotyping over 100,000 snps on a pair of oligonucleotide arrays. *Nat Methods*, 1(2):109–11, 11 2004.

[136] B. H. McArdle and M. J. Anderson. Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology*, 82(1):290–297, 2001.

[137] Phillip McClurg, Jeff Janes, Chunlei Wu, David L. Delano, John R. Walker, Serge Batalov, Joseph S. Takahashi, Kazuhiro Shimomura, Akira Kohsaka, Joseph Bass, Tim Wiltshire, and Andrew I. Su. Genomewide association analysis in diverse inbred mice: power and population structure. *Genetics*, 176(1):675–83, 5 2007.

[138] Charles E. McCulloch. *Generalized linear mixed models*. Institute of Mathematical Statistics ; Alexandria, Va. : American Statistical Association,, Beachwood, Ohio, 2003.

[139] M. S. McPeek and L. Sun. Statistical tests for detection of misspecified relationships by use of genome-screen data. *Am J Hum Genet*, 66(3):1076–94, 3 2000.

[140] M. Mehrabian, J. H. Qiao, R. Hyman, D. Ruddle, C. Laughton, and A. J. Lusis. Influence of the apoa-ii gene locus on hdl levels and fatty streak development in mice. *Arterioscler Thromb*, 13(1):1–10, 1 1993.

[141] K. Meyer. Restricted maximum likelihood to estimate variance components for animal models with several random effects using a derivative-free algorithm. *Genetics selection evolution*, 21(3):317–340, 1989.

[142] Brook G. Milligan. Maximum-likelihood estimation of relatedness. *Genetics*, 163(3):1153–67, 3 2003.

[143] Karen L. Mohlke, Michael Boehnke, and GonÃğalo R. Abecasis. Metabolic and cardiovascular traits: an abundance of recently identified common genetic variants. *Hum Mol Genet*, 17(R2):R102–8, 10 2008.

[144] Jennifer L. Moran, Andrew D. Bolton, Pamela V. Tran, Alison Brown, Noelle D. Dwyer, Danielle K. Manning, Bryan C. Bjork, Cheng Li, Kate Montgomery, Sandra M. Siepka, Martha Hotz Vitaterna, Joseph S. Takahashi, Tim Wiltshire, David J. Kwiatkowski, Raju Kucherlapati, and David R. Beier. Utilization of a whole genome snp panel for efficient genetic mapping in the mouse. *Genome Res*, 16(3):436–40, 3 2006.

[145] MouseGenomeSequencingConsortium. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–62, 12 2002.

[146] Richard J. Mural, Mark D. Adams, Eugene W. Myers, Hamilton O. Smith, George L. Gabor Miklos, Ron Wides, Aaron Halpern, Peter W. Li, Granger G. Sutton, Joe Nadeau, Steven L. Salzberg, Robert A. Holt, Chinnappa D. Kodira, Fu Lu, Lin Chen, Zuoming Deng, Carlos C. Evangelista, Weiniu Gan, Thomas J. Heiman, Jiayin Li, Zhenya Li, Gennady V. Merkulov, Natalia V. Milshina, Ashwinikumar K. Naik, Rong Qi, Bixiong Chris Shue, Aihui Wang, Jian Wang, Xin Wang, Xianghe Yan, Jane Ye, Shibu Yooseph, Qi Zhao, Liansheng Zheng, Shiaoping C. Zhu, Kendra Biddick, Randall Bolanos, Arthur L. Delcher, Ian M. Dew, Daniel Fasulo, Michael J. Flanigan, Daniel H. Huson, Saul A. Kravitz, Jason R. Miller, Clark M. Mobarry, Knut Reinert, Karin A. Remington, Qing Zhang, Xiangqun H. Zheng, Deborah R. Nusskern, Zhongwu Lai, Yiding Lei, Wenyan Zhong, Alison Yao, Ping Guan, Rui-Ru R. Ji, Zhiping Gu, Zhen-Yuan Y. Wang, Fei Zhong, Chunlin Xiao, Chia-Chien C. Chiang, Mark Yandell, Jennifer R. Wortman, Peter G. Amanatides, Suzanne L. Hladun, Eric C. Pratts, Jeffery E. Johnson, Kristina L. Dodson, Kerry J. Woodford, Cheryl A. Evans, Barry Gropman, Douglas B. Rusch, Eli Venter, Mei Wang, Thomas J. Smith, Jarrett T. Houck, Donald E. Tompkins, Charles Haynes, Debbie Jacob, Soo H. Chin, David R. Allen, Carl E. Dahlke, Robert Sanders, Kelvin Li, Xiangjun Liu, Alexander A. Levitsky, William H. Majoros, Quan Chen, Ashley C. Xia, John R. Lopez, Michael T. Donnelly, Matthew H. Newman, Anna Glodek, Cheryl L. Kraft, Marc Nodell, Feroze

Ali, Hui-Jin J. An, Danita Baldwin-Pitts, Karen Y. Beeson, Shuang Cai, Mark Carnes, Amy Carver, Parris M. Caulk, Angela Center, Yen-Hui H. Chen, Ming-Lai L. Cheng, My D. Coyne, Michelle Crowder, Steven Danaher, Lionel B. Davenport, Raymond Desilets, Susanne M. Dietz, Lisa Doup, Patrick Dullaghan, Steven Ferriera, Carl R. Fosler, Harold C. Gire, Andres Gluecksmann, Jeannine D. Gocayne, Jonathan Gray, Brit Hart, Jason Haynes, Jeffery Hoover, Tim Howland, Chinyere Ibegwam, Mena Jalali, David Johns, Leslie Kline, Daniel S. Ma, Steven MacCawley, Anand Magoon, Felecia Mann, David May, Tina C. McIntosh, Somil Mehta, Linda Moy, Mee C. Moy, Brian J. Murphy, Sean D. Murphy, Keith A. Nelson, Zubeda Nuri, Kimberly A. Parker, Alexandre C. Prudhomme, Vinita N. Puri, Hina Qureshi, John C. Raley, Matthew S. Reardon, Megan A. Regier, Yu-Hui C. H. Rogers, Deanna L. Romblad, Jakob Schutz, John L. Scott, Richard Scott, Cynthia D. Sitter, Michella Smallwood, Arlan C. Sprague, Erin Stewart, Renee V. Strong, Ellen Suh, Karena Sylvester, Reginald Thomas, Ni Ni Tint, Christopher Tsonis, Gary Wang, George Wang, Monica S. Williams, Sherita M. Williams, Sandra M. Windsor, Keriellen Wolfe, Mitchell M. Wu, Jayshree Zaveri, Kabir Chaturvedi, Andrei E. Gabrielian, Zhaoxi Ke, Jingtao Sun, Gangadharan Subramanian, J. Craig Venter, Cynthia M. Pfannkoch, Mary Barnstead, and Lisa D. Stephenson. A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. *Science*, 296(5573):1661–71, 5 2002.

[147] J. A. Nelder and R. Mead. A simplex method for function minimization. *The computer journal*, 7(4):308, 1965.

[148] D. L. Newman, M. Abney, M. S. McPeek, C. Ober, and N. J. Cox. The importance of genealogy in determining genetic associations with complex traits. *Am J Hum Genet*, 69(5):1146–8, 11 2001.

[149] Caroline M. Nievergelt, Ondrej Libiger, and Nicholas J. Schork. Generalized analysis of molecular variance. *PLoS Genet*, 3(4):e51, 4 2007.

[150] Magnus Nordborg, Tina T. Hu, Yoko Ishino, Jinal Jhaveri, Christopher Toomajian, Honggang Zheng, Erica Bakker, Peter Calabrese, Jean Gladstone, Rana Goyal, Mattias Jakobsson, Sung Kim, Yuri Morozov, Badri Padhukasahasram, Vincent Plagnol, Noah A. Rosenberg, Chitiksha Shah, Jeffrey D. Wall, Jue Wang, Keyan Zhao, Theodore Kalbfleisch, Vincent Schulz, Martin Kreitman, and Joy Bergelson. The pattern of polymorphism in arabidopsis thaliana. *PLoS Biol*, 3(7):e196, 7 2005.

[151] Michael Nothnagel, David Ellinghaus, Stefan Schreiber, Michael Krawczak, and Andre Franke. A comprehensive evaluation of snp genotype imputation. *Hum Genet*, 125(2):163–71, 3 2009.

[152] John Novembre and Matthew Stephens. Interpreting principal component analyses of spatial population genetic variation. *Nat Genet*, 40(5):646–9, 5 2008.

[153] Emma S. Nyman, Anu Loukola, Teppo Varilo, Jesper Ekelund, Juha Veijola, Matti Joukamaa, Anja Taanila, Anneli Pouta, Jouko Miettunen, Nelson Freimer, Marjo-Riitta R. JÃďrvelin, and Leena Peltonen. Impact of the dopamine receptor gene family on temperament traits in a population-based birth cohort. *Am J Med Genet B Neuropsychiatr Genet*, 12 2008.

[154] Todd H. Oakley, Zhenglong Gu, Ehab Abouheif, Nipam H. Patel, and Wen-Hsiung H. Li. Comparative methods for the analysis of gene-expression evolution: an example using yeast functional genomic data. *Mol Biol Evol*, 22(1):40–50, 1 2005.

[155] C. Ober, M. Abney, and M. S. McPeek. The genetic dissection of complex traits in a founder population. *Am J Hum Genet*, 69(5):1068–79, 11 2001.

[156] N. Patil, A. J. Berno, D. A. Hinds, W. A. Barrett, J. M. Doshi, C. R. Hacker, C. R. Kautzer, D. H. Lee, C. Marjoribanks, D. P. McDonough, B. T. Nguyen, M. C. Norris, J. B. Sheehan, N. Shen, D. Stern, R. P. Stokowski, D. J. Thomas, M. O. Trulson, K. R. Vyas, K. A. Frazer, S. P. Fodor, and D. R. Cox. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, 294(5547):1719–23, 11 2001.

[157] H. D. Patterson and R. Thompson. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3):545–554, 1971.

[158] Nick Patterson, Alkes L. Price, and David Reich. Population structure and eigenanalysis. *PLoS Genet*, 2(12):e190, 12 2006.

[159] Bret A. Payseur and Michael Place. Prospects for association mapping in classical inbred mouse strains. *Genetics*, 175(4):1999–2008, 4 2007.

[160] Jeremy L. Peirce, Hongqiang Li, Jintao Wang, Kenneth F. Manly, Robert J. Hitzemann, John K. Belknap, Glenn D. Rosen, Shirlean Goodwin, Thomas R. Sutter, Robert W. Williams, and Lu Lu. How replicable are mrna expression qtl? *Mamm Genome*, 17(6):643–56, 6 2006.

[161] Ethan O. Perlstein, Douglas M. Ruderfer, David C. Roberts, Stuart L. Schreiber, and Leonid Kruglyak. Genetic basis of individual differences in the response to small-molecule drugs in yeast. *Nat Genet*, 39(4):496–502, 4 2007.

[162] Luanne L. Peters, Raymond F. Robledo, Carol J. Bult, Gary A. Churchill, Beverly J. Paigen, and Karen L. Svenson. The mouse as a model for human biology: a resource guide for complex trait analysis. *Nat Rev Genet*, 8(1):58–69, 1 2007.

[163] Petko M. Petkov, Yueming Ding, Megan A. Cassell, Weidong Zhang, Gunjan Wagner, Evelyn E. Sargent, Steven Asquith, Victor Crew, Kevin A. Johnson,

Phil Robinson, Valerie E. Scott, and Michael V. Wiles. An efficient snp system for mouse genome scanning and elucidating strain relationships. *Genome Res*, 14(9):1806–11, 9 2004.

[164] Petko M. Petkov, Joel H. Graber, Gary A. Churchill, Keith DiPetrillo, Benjamin L. King, and Kenneth Paigen. Evidence of a large-scale functional organization of mammalian chromosomes. *PLoS Genet*, 1(3):e33, 9 2005.

[165] Enrico Petretto, Jonathan Mangion, Michal Pravanec, Norbert Hubner, and Timothy J. Aitman. Integrated gene expression profiling and linkage analysis in the rat. *Mamm Genome*, 17(6):480–9, 6 2006.

[166] 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.

[167] Alkes L. Price, Nick Patterson, Dustin C. Hancks, Simon Myers, David Reich, Vivian G. Cheung, and Richard S. Spielman. Effects of cis and trans genetic ancestry on gene expression in african americans. *PLoS Genet*, 4(12):e1000294, 12 2008.

[168] Alkes L. Price, Nick J. Patterson, Robert M. Plenge, Michael E. Weinblatt, Nancy A. Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*, 38(8):904–9, 8 2006.

[169] J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–59, 6 2000.

[170] J. K. Pritchard, M. Stephens, N. A. Rosenberg, and P. Donnelly. Association mapping in structured populations. *Am J Hum Genet*, 67(1):170–81, 7 2000.

[171] Inga Prokopenko, Claudia Langenberg, Jose C. Florez, Richa Saxena, Nicole Soranzo, Gudmar Thorleifsson, Ruth J. F. Loos, Alisa K. Manning, Anne U. Jackson, Yurii Aulchenko, Simon C. Potter, Michael R. Erdos, Serena Sanna, Jouke-Jan J. Hottenga, Eleanor Wheeler, Marika Kaakinen, Valeriya Lyssenko, Wei-Min M. Chen, Kourosh Ahmadi, Jacques S. Beckmann, Richard N. Bergman, Murielle Bochud, Lori L. Bonnycastle, Thomas A. Buchanan, Antonio Cao, Alessandra Cervino, Lachlan Coin, Francis S. Collins, Laura Crisponi, Eco J. C. de Geus, Abbas Dehghan, Panos Deloukas, Alex S. F. Doney, Paul Elliott, Nelson Freimer, Vesela Gateva, Christian Herder, Albert Hofman, Thomas E. Hughes, Sarah Hunt, Thomas Illig, Michael Inouye, Bo Isomaa, Toby Johnson, Augustine Kong, Maria Krestyaninova, Johanna Kuusisto, Markku Laakso, Noha Lim, Ulf Lindblad, Cecilia M. Lindgren, Owen T. McCann, Karen L. Mohlke, Andrew D. Morris, Silvia Naitza, Marco OrrÃź, Colin

N. A. Palmer, Anneli Pouta, Joshua Randall, Wolfgang Rathmann, Jouko Saramies, Paul Scheet, Laura J. Scott, Angelo Scuteri, Stephen Sharp, Eric Sijbrands, Jan H. Smit, Kijoung Song, Valgerdur Steinthorsdottir, Heather M. Stringham, Tiinamaija Tuomi, Jaakko Tuomilehto, AndrÃľ G. Uitterlinden, Benjamin F. Voight, Dawn Waterworth, H-Erich E. Wichmann, Gonneke Willemsen, Jacqueline C. M. Witteman, Xin Yuan, Jing Hua Zhao, Eleftheria Zeggini, David Schlessinger, Manjinder Sandhu, Dorret I. Boomsma, Manuela Uda, Tim D. Spector, Brenda Wjh Penninx, David Altshuler, Peter Vollenweider, Marjo Riitta Jarvelin, Edward Lakatta, Gerard Waeber, Caroline S. Fox, Leena Peltonen, Leif C. Groop, Vincent Mooser, L. Adrienne Cupples, Unnur Thorsteinsdottir, Michael Boehnke, InÃłs Barroso, Cornelia Van Duijn, JosÃľe Dupuis, Richard M. Watanabe, Kari Stefansson, Mark I. McCarthy, Nicholas J. Wareham, James B. Meigs, and GonÃğalo R. Abecasis. Variants in mtnr1b influence fasting glucose levels. *Nat Genet*, 41(1):77–81, 1 2009.

[172] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I. W. de Bakker, Mark J. Daly, and Pak C. Sham. Plink: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, 81(3):559–75, 9 2007.

[173] D. A. Purcell-Huynh, A. Weinreb, L. W. Castellani, M. Mehrabian, M. H. Doolittle, and A. J. Lusis. Genetic factors in lipoprotein metabolism. analysis of a genetic cross between inbred mouse strains nzb/binj and sm/j using a complete linkage map approach. *J Clin Invest*, 96(4):1845–58, 10 1995.

[174] Miguel PÃľrez-Enciso. In silico study of transcriptome genetic variation in outbred populations. *Genetics*, 166(1):547–54, 1 2004.

[175] P. Rantakallio. Groups at risk in low birth weight infants and perinatal mortality. *Acta Paediatr Scand*, 193:Suppl 193:1+, 1969.

[176] D. R. Reed, S. Li, X. Li, L. Huang, M. G. Tordoff, R. Starling-Roney, K. Taniguchi, D. B. West, J. D. Ohmen, G. K. Beauchamp, and A. A. Bachmanov. Polymorphisms in the taste receptor gene (tas1r3) region are associated with saccharin preference in 30 mouse strains. *J Neurosci*, 24(4):938–46, 1 2004.

[177] David E. Reich, Stephen F. Schaffner, Mark J. Daly, Gil McVean, James C. Mullikin, John M. Higgins, Daniel J. Richter, Eric S. Lander, and David Altshuler. Human genome sequence variation and the influence of gene history, mutation and recombination. *Nat Genet*, 32(1):135–42, 9 2002.

[178] John P. Rice, Nancy L. Saccone, and Jonathan Corbett. Model-based methods for linkage analysis. *Adv Genet*, 60:155–73, 2008.

[179] N. Risch and K. Merikangas. The future of genetic studies of complex human diseases. *Science*, 273(5281):1516–7, 9 1996.

[180] Mark D. Robinson, JÃűrg Grigull, Naveed Mohammad, and Timothy R. Hughes. Funspec: a web-based cluster interpreter for yeast. *BMC Bioinformatics*, 3:35, 11 2002.

[181] Mark D. Robinson and Terence P. Speed. A comparison of affymetrix gene expression arrays. *BMC Bioinformatics*, 8:449, 2007.

[182] Joao L. Rocha, Eugene J. Eisen, L. Dale Van Vleck, and Daniel Pomp. A large-sample qtl study in mice: Ii. body composition. *Mamm Genome*, 15(2):100–13, 2 2004.

[183] Matthew V. Rockman and Leonid Kruglyak. Genetics of global gene expression. *Nat Rev Genet*, 7(11):862–72, 11 2006.

[184] Chiara Sabatti, Susan K. Service, Anna-Liisa L. Hartikainen, Anneli Pouta, Samuli Ripatti, Jae Brodsky, Chris G. Jones, Noah A. Zaitlen, Teppo Varilo, Marika Kaakinen, Ulla Sovio, Aimo Ruokonen, Jaana Laitinen, Eveliina Jakkula, Lachlan Coin, Clive Hoggart, Andrew Collins, Hannu Turunen, Stacey Gabriel, Paul Elliot, Mark I. McCarthy, Mark J. Daly, Marjo-Riitta R. JÃďrvelin, Nelson B. Freimer, and Leena Peltonen. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat Genet*, 41(1):35–46, 1 2009.

[185] Eric E. Schadt, Cliona Molony, Eugene Chudin, Ke Hao, Xia Yang, Pek Y. Lum, Andrew Kasarskis, Bin Zhang, Susanna Wang, Christine Suver, Jun Zhu, Joshua Millstein, Solveig Sieberts, John Lamb, Debraj GuhaThakurta, Jonathan Derry, John D. Storey, Iliana Avila-Campillo, Mark J. Kruger, Jason M. Johnson, Carol A. Rohl, Atila van Nas, Margarete Mehrabian, Thomas A. Drake, Aldons J. Lusis, Ryan C. Smith, F. Peter Guengerich, Stephen C. Strom, Erin Schuetz, Thomas H. Rushmore, and Roger Ulrich. Mapping the genetic architecture of gene expression in human liver. *PLoS Biol*, 6(5):e107, 5 2008.

[186] Paul Scheet and Matthew Stephens. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet*, 78(4):629–44, 4 2006.

[187] Nicholas J. Schork, Jennifer Wessel, and Nathalie Malo. Dna sequence-based phenotypic association analysis. *Adv Genet*, 60:195–217, 2008.

[188] Solveig K. Sieberts and Eric E. Schadt. Moving toward a system genetics view of disease. *Mamm Genome*, 18(6-7):389–401, 7 2007.

[189] L. M. Silver. *Mouse genetics: concepts and applications*. Oxford University Press, USA, 1995.

[190] Erin N. Smith and Leonid Kruglyak. Gene-environment interaction in yeast gene expression. *PLoS Biol*, 6(4):e83, 4 2008.

[191] Ulla Sovio, Amanda J. Bennett, Iona Y. Millwood, John Molitor, Paul F. O'Reilly, Nicholas J. Timpson, Marika Kaakinen, Jaana Laitinen, Jari Haukka, Demetris Pillas, Ioanna Tzoulaki, Jassy Molitor, Clive Hoggart, Lachlan J. M. Coin, John Whittaker, Anneli Pouta, Anna-Liisa L. Hartikainen, Nelson B. Freimer, Elisabeth Widen, Leena Peltonen, Paul Elliott, Mark I. McCarthy, and Marjo-Riitta R. Jarvelin. Genetic determinants of height growth assessed longitudinally from infancy to adulthood in the northern finland birth cohort 1966. *PLoS Genet*, 5(3):e1000409, 3 2009.

[192] Richard S. Spielman, Laurel A. Bastone, Joshua T. Burdick, Michael Morley, Warren J. Ewens, and Vivian G. Cheung. Common genetic variants account for differences in gene expression among ethnic groups. *Nat Genet*, 39(2):226–31, 2 2007.

[193] STARConsortium. Snp and haplotype mapping for genetic analysis in the rat. *Nat Genet*, 40(5):560–6, 5 2008.

[194] Lincoln D. Stein, Christopher Mungall, ShengQiang Shu, Michael Caudy, Marco Mangone, Allen Day, Elizabeth Nickerson, Jason E. Stajich, Todd W. Harris, Adrian Arva, and Suzanna Lewis. The generic genome browser: a building block for a model organism system database. *Genome Res*, 12(10):1599–610, 10 2002.

[195] John D. Storey, Joshua M. Akey, and Leonid Kruglyak. Multiple locus linkage analysis of genomewide expression in yeast. *PLoS Biol*, 3(8):e267, 8 2005.

[196] John D. Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A*, 100(16):9440–5, 8 2003.

[197] D. O. Stram and J. W. Lee. Variance components testing in the longitudinal mixed effects model. *Biometrics*, 50(4):1171–7, 12 1994.

[198] Barbara E. Stranger, Matthew S. Forrest, Mark Dunning, Catherine E. Ingle, Claude Beazley, Natalie Thorne, Richard Redon, Christine P. Bird, Anna de Grassi, Charles Lee, Chris Tyler-Smith, Nigel Carter, Stephen W. Scherer, Simon TavarÃĺ, Panagiotis Deloukas, Matthew E. Hurles, and Emmanouil T. Dermitzakis. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, 315(5813):848–53, 2 2007.

[199] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102(43):15545–50, 10 2005.

[200] Jin P. Szatkiewicz, Glen L. Beane, Yueming Ding, Lucie Hutchins, Fernando Pardo-Manuel de Villena, and Gary A. Churchill. An imputed genotype resource for the laboratory mouse. *Mamm Genome*, 19(3):199–208, 3 2008.

[201] K. Tanaka, M. Nakafuku, T. Satoh, M. S. Marshall, J. B. Gibbs, K. Matsumoto, Y. Kaziro, and A. Toh-e. S. cerevisiae genes ira1 and ira2 encode proteins that may be functionally equivalent to mammalian ras gtpase activating protein. *Cell*, 60(5):803–7, 3 1990.

[202] Gilles Thomas, Kevin B. Jacobs, Peter Kraft, Meredith Yeager, Sholom Wacholder, David G. Cox, Susan E. Hankinson, Amy Hutchinson, Zhaoming Wang, Kai Yu, Nilanjan Chatterjee, Montserrat Garcia-Closas, Jesus Gonzalez-Bosquet, Ludmila Prokunina-Olsson, Nick Orr, Walter C. Willett, Graham A. Colditz, Regina G. Ziegler, Christine D. Berg, Saundra S. Buys, Catherine A. McCarty, Heather Spencer Feigelson, Eugenia E. Calle, Michael J. Thun, Ryan Diver, Ross Prentice, Rebecca Jackson, Charles Kooperberg, Rowan Chlebowski, Jolanta Lissowska, Beata Peplonska, Louise A. Brinton, Alice Sigurdson, Michele Doody, Parveen Bhatti, Bruce H. Alexander, Julie Buring, I-Min M. Lee, Lars J. Vatten, Kristian Hveem, Merethe Kumle, Richard B. Hayes, Margaret Tucker, Daniela S. Gerhard, Joseph F. Fraumeni, Robert N. Hoover, Stephen J. Chanock, and David J. Hunter. A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (rad51l1). *Nat Genet*, 3 2009.

[203] S. C. Thomas and W. G. Hill. Estimating quantitative genetic parameters using sibships reconstructed from marker data. *Genetics*, 155(4):1961–72, 8 2000.

[204] Jung-Ying Y. Tzeng and Daowen Zhang. Haplotype-based association analysis via variance-components score test. *Am J Hum Genet*, 81(5):927–38, 11 2007.

[205] Jung-Ying Y. Tzeng, Daowen Zhang, Sheng-Mao M. Chang, Duncan C. Thomas, and Marie Davidian. Gene-trait similarity regression for multimarker-based association analysis. *Biometrics*, 2 2009.

[206] William Valdar, Leah C. Solberg, Dominique Gauguier, Stephanie Burnett, Paul Klenerman, William O. Cookson, Martin S. Taylor, J. Nicholas P. Rawlins, Richard Mott, and Jonathan Flint. Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nat Genet*, 38(8):879–87, 8 2006.

[207] T. Van de Casteele, P. Galbusera, and E. Matthysen. A comparison of microsatellite-based pairwise relatedness estimators. *Mol Ecol*, 10(6):1539–49, 6 2001.

[208] Laura J. van ’t Veer, Hongyue Dai, Marc J. van de Vijver, Yudong D. He, Augustinus A. M. Hart, Mao Mao, Hans L. Peterse, Karin van der

Kooy, Matthew J. Marton, Anke T. Witteveen, George J. Schreiber, Ron M. Kerkhoven, Chris Roberts, Peter S. Linsley, RenÃľ Bernards, and Stephen H. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–6, 1 2002.

[209] Teppo Varilo and Leena Peltonen. Isolates and their potential use in complex gene mapping efforts. *Curr Opin Genet Dev*, 14(3):316–23, 6 2004.

[210] A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory*, 13(2):260–269, 1967.

[211] Benjamin F. Voight and Jonathan K. Pritchard. Confounding from cryptic relatedness in case-control association studies. *PLoS Genet*, 1(3):e32, 9 2005.

[212] Claire M. Wade and Mark J. Daly. Genetic variation in laboratory mice. *Nat Genet*, 37(11):1175–80, 11 2005.

[213] Claire M. Wade, Edward J. Kulbokas, Andrew W. Kirby, Michael C. Zody, James C. Mullikin, Eric S. Lander, Kerstin Lindblad-Toh, and Mark J. Daly. The mosaic structure of variation in the laboratory mouse genome. *Nature*, 420(6915):574–8, 12 2002.

[214] Nicole A. R. Walter, Shannon K. McWeeney, Sandra T. Peters, John K. Belknap, Robert Hitzemann, and Kari J. Buck. Snps matter: impact on detection of differential expression. *Nat Methods*, 4(9):679–80, 9 2007.

[215] Jinliang Wang. An estimator for pairwise relatedness using molecular markers. *Genetics*, 160(3):1203–15, 3 2002.

[216] Shuang Wang, Tian Zheng, and Yuanjia Wang. Transcription activity hot spot, is it real or an artifact? *BMC Proc*, 1 Suppl 1:S94, 2007.

[217] Susanna S. Wang, Eric E. Schadt, Hui Wang, Xuping Wang, Leslie Ingram-Drake, Weibin Shi, Thomas A. Drake, and Aldons J. Lusis. Identification of pathways for atherosclerosis in mice: integration of quantitative trait locus analysis and global gene expression data. *Circ Res*, 101(3):e11–30, 8 2007.

[218] Xiaosong Wang, Naoki Ishimori, Ron Korstanje, Jarod Rollins, and Beverly Paigen. Identifying novel genes for atherosclerosis through mouse-human comparative genetics. *Am J Hum Genet*, 77(1):1–15, 7 2005.

[219] Xiaosong Wang and Beverly Paigen. Genetics of variation in hdl cholesterol in humans and mice. *Circ Res*, 96(1):27–42, 1 2005.

[220] C. H. Warden, A. Daluiski, X. Bu, D. A. Purcell-Huynh, C. De Meester, B. H. Shieh, D. L. Puppione, R. M. Gray, G. M. Reaven, and Y. D. Chen. Evidence for linkage of the apolipoprotein a-ii locus to plasma apolipoprotein a-ii and free

fatty acid levels in mice and humans. *Proc Natl Acad Sci U S A*, 90(22):10886–90, 11 1993.

[221] Bruce S. Weir, Amy D. Anderson, and Amanda B. Hepler. Genetic relatedness analysis: modern data and new challenges. *Nat Rev Genet*, 7(10):771–80, 10 2006.

[222] 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)*, pages 701–714, 1997.

[223] WellcomeTrustCaseControlConsortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–78, 6 2007.

[224] Jennifer Wessel and Nicholas J. Schork. Generalized genomic distance-based regression methodology for multilocus association analysis. *Am J Hum Genet*, 79(5):792–806, 11 2006.

[225] Cristen J. Willer, Serena Sanna, Anne U. Jackson, Angelo Scuteri, Lori L. Bonnycastle, Robert Clarke, Simon C. Heath, Nicholas J. Timpson, Samer S. Najjar, Heather M. Stringham, James Strait, William L. Duren, Andrea Maschio, Fabio Busonero, Antonella Mulas, Giuseppe Albai, Amy J. Swift, Mario A. Morken, Narisu Narisu, Derrick Bennett, Sarah Parish, Haiqing Shen, Pilar Galan, Pierre Meneton, Serge Hercberg, Diana Zelenika, Wei-Min M. Chen, Yun Li, Laura J. Scott, Paul A. Scheet, Jouko Sundvall, Richard M. Watanabe, Ramaiah Nagaraja, Shah Ebrahim, Debbie A. Lawlor, Yoav Ben-Shlomo, George Davey-Smith, Alan R. Shuldiner, Rory Collins, Richard N. Bergman, Manuela Uda, Jaakko Tuomilehto, Antonio Cao, Francis S. Collins, Edward Lakatta, G. Mark Lathrop, Michael Boehnke, David Schlessinger, Karen L. Mohlke, and GonÃğalo R. Abecasis. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat Genet*, 40(2):161–9, 2 2008.

[226] Cristen J. Willer, Elizabeth K. Speliotes, Ruth J. F. Loos, Shengxu Li, Cecilia M. Lindgren, Iris M. Heid, Sonja I. Berndt, Amanda L. Elliott, Anne U. Jackson, Claudia Lamina, Guillaume Lettre, Noha Lim, Helen N. Lyon, Steven A. McCarroll, Konstantinos Papadakis, Lu Qi, Joshua C. Randall, Rosa Maria Roccasecca, Serena Sanna, Paul Scheet, Michael N. Weedon, Eleanor Wheeler, Jing Hua Zhao, Leonie C. Jacobs, Inga Prokopenko, Nicole Soranzo, Toshiko Tanaka, Nicholas J. Timpson, Peter Almgren, Amanda Bennett, Richard N. Bergman, Sheila A. Bingham, Lori L. Bonnycastle, Morris Brown, NoÃńl P. Burtt, Peter Chines, Lachlan Coin, Francis S. Collins, John M. Connell, Cyrus Cooper, George Davey Smith, Elaine M. Dennison, Parimal Deodhar, Paul Elliott, Michael R. Erdos, Karol Estrada, David M. Evans, Lauren Gianniny, Christian Gieger, Christopher J. Gillson, Candace

Guiducci, Rachel Hackett, David Hadley, Alistair S. Hall, Aki S. Havulinna, Johannes Hebebrand, Albert Hofman, Bo Isomaa, Kevin B. Jacobs, Toby Johnson, Pekka Jousilahti, Zorica Jovanovic, Kay-Tee T. Khaw, Peter Kraft, Mikko Kuokkanen, Johanna Kuusisto, Jaana Laitinen, Edward G. Lakatta, Jian'an Luan, Robert N. Luben, Massimo Mangino, Wendy L. McArdle, Thomas Meitinger, Antonella Mulas, Patricia B. Munroe, Narisu Narisu, Andrew R. Ness, Kate Northstone, Stephen O'Rahilly, Carolin Purmann, Matthew G. Rees, Martin RidderstrÃěle, Susan M. Ring, Fernando Rivadeneira, Aimo Ruokonen, Manjinder S. Sandhu, Jouko Saramies, Laura J. Scott, Angelo Scuteri, Kaisa Silander, Matthew A. Sims, Kijoung Song, Jonathan Stephens, Suzanne Stevens, Heather M. Stringham, Y. C. Loraine Tung, Timo T. Valle, Cornelia M. Van Duijn, Karani S. Vimaleswaran, Peter Vollenweider, Gerard Waeber, Chris Wallace, Richard M. Watanabe, Dawn M. Waterworth, Nicholas Watkins, Wellcome Trust Case Control Consortium, Jacqueline C. M. Witteman, Eleftheria Zeggini, Guangju Zhai, M. Carola Zillikens, David Altshuler, Mark J. Caulfield, Stephen J. Chanock, I. Sadaf Farooqi, Luigi Ferrucci, Jack M. Guralnik, Andrew T. Hattersley, Frank B. Hu, Marjo-Riitta R. Jarvelin, Markku Laakso, Vincent Mooser, Ken K. Ong, Willem H. Ouwehand, Veikko Salomaa, Nilesh J. Samani, Timothy D. Spector, Tiinamaija Tuomi, Jaakko Tuomilehto, Manuela Uda, AndrÃľ G. Uitterlinden, Nicholas J. Wareham, Panagiotis Deloukas, Timothy M. Frayling, Leif C. Groop, Richard B. Hayes, David J. Hunter, Karen L. Mohlke, Leena Peltonen, David Schlessinger, David P. Strachan, H-Erich E. Wichmann, Mark I. McCarthy, Michael Boehnke, InÃłs Barroso, GonÃğalo R. Abecasis, Joel N. Hirschhorn, and Genetic Investigation of ANthropometric Traits Consortium. Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat Genet*, 41(1):25–34, 1 2009.

[227] R. W. Williams, J. Gu, S. Qi, and L. Lu. The genetic structure of recombinant inbred mice: high-resolution consensus maps for complex trait analysis. *Genome Biol*, 2(11):RESEARCH0046, 2001.

[228] Robert W. Williams. Expression genetics and the phenotype revolution. *Mamm Genome*, 17(6):496–502, 6 2006.

[229] Rohan B. H. Williams, Chris J. Cotsapas, Mark J. Cowley, Eva Chan, David J. Nott, and Peter F. R. Little. Normalization procedures and detection of linkage signal in genetical-genomics experiments. *Nat Genet*, 38(8):855–6; author reply 856–9, 8 2006.

[230] Tim Wiltshire, Mathew T. Pletcher, Serge Batalov, S. Whitney Barnes, Lisa M. Tarantino, Michael P. Cooke, Hua Wu, Kevin Smylie, Andrey Santrosyan, Neal G. Copeland, Nancy A. Jenkins, Francis Kalush, Richard J. Mural, Richard J. Glynne, Steve A. Kay, Mark D. Adams, and Colin F. Fletcher.

Genome-wide single-nucleotide polymorphism analysis defines haplotype patterns in mouse. *Proc Natl Acad Sci U S A*, 100(6):3380–5, 3 2003.

[231] B. Yalcin, J. Fullerton, S. Miller, D. A. Keays, S. Brady, A. Bhomra, A. Jefferson, E. Volpi, R. R. Copley, J. Flint, and R. Mott. Unexpected complexity in the haplotypes of commonly used inbred strains of laboratory mice. *Proc Natl Acad Sci U S A*, 101(26):9734–9, 6 2004.

[232] Hyuna Yang, Timothy A. Bell, Gary A. Churchill, and Fernando Pardo-Manuel de Villena. On the subspecific origin of the laboratory mouse. *Nat Genet*, 39(9):1100–7, 9 2007.

[233] Yee Hwa Yang, Sandrine Dudoit, Percy Luu, David M. Lin, Vivian Peng, John Ngai, and Terence P. Speed. Normalization for cdna microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*, 30(4):e15, 2 2002.

[234] Chun Ye and Eleazar Eskin. Discovering tightly regulated and differentially expressed gene sets in whole genome expression data. *Bioinformatics*, 23(2):e84–90, 1 2007.

[235] Jianming Yu, Gael Pressoir, William H. Briggs, Irie Vroh Bi, Masanori Yamasaki, John F. Doebley, Michael D. McMullen, Brandon S. Gaut, Dahlia M. Nielsen, James B. Holland, Stephen Kresovich, and Edward S. Buckler. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet*, 38(2):203–8, 2 2006.

[236] GaÃńl Yvert, Rachel B. Brem, Jacqueline Whittle, Joshua M. Akey, Eric Foss, Erin N. Smith, Rachel Mackelprang, and Leonid Kruglyak. Trans-acting regulatory variation in saccharomyces cerevisiae and the role of transcription factors. *Nat Genet*, 35(1):57–64, 9 2003.

[237] Noah Zaitlen, Hyun Min Kang, Eleazar Eskin, and Eran Halperin. Leveraging the hapmap correlation structure in association studies. *Am J Hum Genet*, 80(4):683–91, 4 2007.

[238] Eleftheria Zeggini, Laura J. Scott, Richa Saxena, Benjamin F. Voight, Jonathan L. Marchini, Tianle Hu, Paul I. W. de Bakker, GonÃğalo R. Abecasis, Peter Almgren, Gitte Andersen, Kristin Ardlie, Kristina Bengtsson BostrÃűm, Richard N. Bergman, Lori L. Bonnycastle, Knut Borch-Johnsen, NoÃńl P. Burtt, Hong Chen, Peter S. Chines, Mark J. Daly, Parimal Deodhar, Chia-Jen J. Ding, Alex S. F. Doney, William L. Duren, Katherine S. Elliott, Michael R. Erdos, Timothy M. Frayling, Rachel M. Freathy, Lauren Gianniny, Harald Grallert, Niels Grarup, Christopher J. Groves, Candace Guiducci, Torben Hansen, Christian Herder, Graham A. Hitman, Thomas E. Hughes, Bo Isomaa, Anne U. Jackson, Torben JÃÿrgensen, Augustine Kong, Kari Kubalanza,

Finny G. Kuruvilla, Johanna Kuusisto, Claudia Langenberg, Hana Lango, Torsten Lauritzen, Yun Li, Cecilia M. Lindgren, Valeriya Lyssenko, Amanda F. Marvelle, Christa Meisinger, Kristian Midthjell, Karen L. Mohlke, Mario A. Morken, Andrew D. Morris, Narisu Narisu, Peter Nilsson, Katharine R. Owen, Colin N. A. Palmer, Felicity Payne, John R. B. Perry, Elin Pettersen, Carl Platou, Inga Prokopenko, Lu Qi, Li Qin, Nigel W. Rayner, Matthew Rees, Jeffrey J. Roix, Anelli Sandbaek, Beverley Shields, Marketa SjÃűgren, Valgerdur Steinthorsdottir, Heather M. Stringham, Amy J. Swift, Gudmar Thorleifsson, Unnur Thorsteinsdottir, Nicholas J. Timpson, Tiinamaija Tuomi, Jaakko Tuomilehto, Mark Walker, Richard M. Watanabe, Michael N. Weedon, Cristen J. Willer, Wellcome Trust Case Control Consortium, Thomas Illig, Kristian Hveem, Frank B. Hu, Markku Laakso, Kari Stefansson, Oluf Pedersen, Nicholas J. Wareham, InÃłs Barroso, Andrew T. Hattersley, Francis S. Collins, Leif Groop, Mark I. McCarthy, Michael Boehnke, and David Altshuler. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet*, 40(5):638–45, 5 2008.

[239] Jinghui Zhang, Kent W. Hunter, Michael Gandolph, William L. Rowe, Richard P. Finney, Jenny M. Kelley, Michael Edmonson, and Kenneth H. Buetow. A high-resolution multistrain haplotype analysis of laboratory mouse genome reveals three distinctive genetic variation patterns. *Genome Res*, 15(2):241–9, 2 2005.

[240] Keyan Zhao, MarÃŋa JosÃľ Aranzana, Sung Kim, Clare Lister, Chikako Shindo, Chunlao Tang, Christopher Toomajian, Honggang Zheng, Caroline Dean, Paul Marjoram, and Magnus Nordborg. An arabidopsis example of association mapping in structured samples. *PLoS Genet*, 3(1):e4, 1 2007.

[241] Keyan Zhao, Magnus Nordborg, and Paul Marjoram. Genome-wide association mapping using mixed-models: application to gaw15 problem 3. *BMC Proc*, 1 Suppl 1:S164, 2007.

[242] Jun Zhu, Bin Zhang, Erin N. Smith, Becky Drees, Rachel B. Brem, Leonid Kruglyak, Roger E. Bumgarner, and Eric E. Schadt. Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nat Genet*, 40(7):854–61, 7 2008.

[243] Fei Zou, Jonathan A. L. Gelfond, David C. Airey, Lu Lu, Kenneth F. Manly, Robert W. Williams, and David W. Threadgill. Quantitative trait locus analysis using recombinant inbred intercrosses: theoretical and empirical considerations. *Genetics*, 170(3):1299–311, 7 2005.