Efficient methods for analysis of genome scale data

by

Liming Liang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2009

Doctoral Committee:

Professor Gonçalo R. Abecasis, Chair
Professor Michael Boehnke
Professor Peter X.K. Song
Associate Professor Noah A. Rosenberg
Assistant Professor Sebastian Zöllner

2009

To my parents, Biling & Sophie

Acknowledgements

I would like to thank everyone who ever helped me! In particular, I am extremely grateful to my advisor Dr. Gonçalo Abecasis for his knowledgeable guidance, insightful inspiration, tireless encouragement to achieve high standards and also for being a role model academically and professionally. I deeply appreciate my dissertation committee members, Dr. Michael Boehnke, Dr. Peter Song, Dr. Noah Rosenberg and Dr. Sebastian Zöllner for their invaluable guidance, constructive suggestions, enthusiasm and unlimited patience for insisting upon high quality research when doing projects together, writing the manuscript and anticipating future directions.

I am also very grateful to be working with such brilliant and dedicated mentors as Dr. William Cookson, Dr. Mark Lathrop as well as their group members, Dr. Miriam Moffatt, Dr. Anna Dixon, Dr. Nilesh Morar, Dr. Simon Heath and others. Without them, the exciting eQTL analyses presented here would have been impossible. It is my honor to be in this team and learn from

them.

I thank my colleagues, Dr. Pak Sham, Weihua Guan, Dr. Wei-Min Chen and Wei Chen. They are all excellent scientists and wonderful people. I am lucky to have the opportunity to cooperate with them and know them in person.

I thank Terry Gliedt for his remarkable work to maintain a stable and efficient cluster without which the vast amount of statistical analysis and simulations on huge datasets could not have been done in time.

I thank my friends in the department who make life itself a joy.

Finally I want to thank my parents and my wife Biling who have always supported my various endeavors and encouraged me to pursue my dream!

Table of Contents

List of Tables

List of Figures

Chapter I

INTRODUCTION

1.1 Genetics of complex diseases

In the past decades, geneticists have been remarkably successful in identifying genes for monogenic diseases that follow simple Mendelian inheritance (Botstein & Risch 2003). Most diseases or phenotypic traits, however, are affected by a combination of environmental factors, mutations in multiple genes, and even genetic variants outside genes (Lander & Schork 1994). Some examples of multi-factorial diseases (also called complex diseases) include asthma, autoimmune diseases such as inflammatory bowel disease and type I diabetes, cancers, type II diabetes, heart disease, hypertension, and others. Among the broadly four categories of genetics disorders (Human Genome Project Information), namely (1) monogenic (2) multi-factorial (3) chromosomal and (4) mitochondrial, complex diseases have the most impact on human populations and pose the most difficult challenges for scientists that aim to identify the genes involved (Chakravarti 1999, Risch 2000).

Although complex disorders usually cluster in families, the pattern of inheritance is unclear. Many factors affect the development of complex diseases, including effects from different genes, environmental effects, gene-gene and gene-environment interactions.

This makes it difficult to determine the risk of passing on the disease and identify the genes involving in disease pathways. The majority of genetic factors involved in complex disorders have not yet been identified and replicated (Hirschhorn et al. 2002). However, the rapid progress in identifying genetic variants and of genotyping technologies in the last few years has make it possible for geneticists to collect data on hundreds of microsatellite markers or 300,000 to 1,000,000 single nucleotide polymorphism (SNP) markers on thousands of individuals (Sachidanandam 2001, The International HapMap Consortium 2007, Eberle MA et al. 2007, McCarroll et al. 2008). Many genetic loci for different types of diseases have been identified by using two major strategies: linkage analysis and genome-wide association study (Jimenez-Sanchez et al. 2001, Carlson et al. 2004, Hirschhorn & Daly 2005). Linkage analysis looks for the co-segregation of a chromosomal region with a trait of interest in the family. It locates a rough position of disease gene related to know genetic markers with resolution down to 10-20Megabase. In addition to dichotomous disease status, it can handle quantitative traits as well (Ott 1999). Linkage analyses have been widely used to identify many important genes for different diseases, especially for monogenic diseases (Jimenez-Sanchez et al. 2001).

Genome-wide association studies test for the association of marker alleles with a trait of interest. Since the completion of the Human Genome Project in 2003 and the first phase of the International HapMap Project in 2005, association tests on a genomic scale have become possible. The revolution of commercial genotyping platforms and the success of the HapMap project have made genome-wide association studies a productive strategy for gene mapping of complex diseases in recent years (Carlson et al. 2004, Hirschhorn & Daly 2005, McCarthy et al. 2008, Hardy & Singleton 2009).

1.2 Challenges to complex disease mapping

Despite the dramatic increase in genomic discoveries involving complex diseases, the majority of genetic factors involved in common diseases have not been identified. The chasing of genetic variants responsible for risk of diseases is a systematic work. It requires breakthroughs in every perspective of the study: range from biologically meaningful and clinically accurate phenotypes to complete and reliable genotypes, from statistical inference and estimation based on real data to performance evaluation and probability sampling based on simulations, from establishing statistical evidence to seeking biological interpretations. Some major specific challenges tackled here include getting accurate measures of phenotype of interest, population heterogeneity, genome coverage of commercial genotyping platform and functional interpretation of identified disease loci. These specific challenges were chosen because they are among the most pressing problem areas faced by geneticists.

The first challenge we addressed here is inaccurate phenotyping that blurs the definition of disease status or results in large measurement error in a quantitative trait, which can greatly decrease the power to detect genetic variants for the trait of interest (Levy et al. 2000). Using medical records instead of questionnaires and identifying disease subtypes may help to obtain accurate disease status (Hallmayer et al. 2005). To improve quantitative measures, one could take multiple measurements or estimate the noise shared by multiple traits of interest, for example, all transcripts from a gene expression microarray (Leek & Storey 2007, Stegle et al. 2008).

Another challenge is population heterogeneity (also known as population stratification) that can inflate false positive or decrease power (Li 1972). As genotyping techniques become more and more affordable, larger and larger samples are collected in the hope of increasing power. It becomes more difficult to guarantee all individuals in the sample share the same ancestry (Freedman et al. 2004). Fortunately, genetic markers on genome scale are available from these studies and hence provide adequate genetic information to detect and correct for population stratification (Pritchard et al., 2000, Price et al., 2006, Luca et al., 2008).

The third challenge arises even when current commercial genotyping platforms can type as many as 1,000,000 SNPs on the genome and have fairly good coverage of the HapMap SNPs, they still miss most identified SNPs and so have limited coverage of the genome (Pe'er et al. 2006, Barrett & Cardon 2006, Hao et al. 2008, Bhangale et al. 2008). Genetic loci that harbor disease causal variants but are not in strong linkage disequilibrium of the typed markers will have low power to be detected by genome-wide association studies. However, existing data with denser genetic markers, such as those generated from the HapMap project, can help to perform statistical inference on genotypes of untyped markers (Scheet & Stephens 2006, Servin & Stephens 2007, Marchini et al 2007, Li et al. 2008). The approaches are commonly used to increase the power and coverage of individual genome-wide association studies and to facilitate meta-analysis of data across studies that relied on different commercial genotyping platforms (for early examples, see Willer et al. 2008, Sanna et al. 2008, Scott et al. 2007, The Wellcome Trust Case Control Consortium 2007). Simulation experiments and detailed genotyping within selected regions show that this strategy should result in

imputed genotypes that are highly accurate and that the analysis of imputed genotypes increases power for association studies (Li et al. 2008, Marchini et al 2007). Still, a large scale assessment of the accuracy of genotype imputation and, particularly, of its impact on power remains lacking.

Because most genetic loci identified from linkage analyses and genome-wide association analyses do not have immediate functional interpretations, the next challenge comes that biologically relevant genes are not easy to determine purely based on the proximity of the detected loci. Systematically generated unbiased functional data, such as the regulators of global gene expression, could aid in interpretation of results from the disease mapping (Dixon et al. 2007; Moffatt et al., 2007, Libioulle et al. 2007, Cookson et al. 2009).

Finally, the performance of any methods trying to tackle the challenges of complex disease mapping should be evaluated by large scale simulation studies. Existing software packages based on coalescent theory, such as ms (Hudson 2002) and cosi (Schaffner et al. 2005), are suitable for short genomic segments (<2-3Mb) but become very slow for larger regions (>100Mb). Efficient tools are needed to generate datasets on genome scale that follow desired parameters such as population histories and disease penetrance.

1.3 The scope of the dissertation

The continuous breakthroughs in biological and computational techniques have pushed the field of genetic research to move quickly. Now we have the material base to address each perspective of gene mapping studies and we need to address all these

upcoming open questions in order to move forward. In my dissertation, I propose efficient methods to tackle the above challenges from complex disease gene mapping studies and implement into software packages that are freely available to the community.

In chapter II, I extended the variance components approach (Jacquard 1972, Lange et al. 1976, Amos 1994, Almasy & Blangero 1998) to model repeated measures in a quantitative trait linkage study. I show that for balanced designs where each subject has the same number of measurements, a standard linkage test that takes the average of measures as the trait of interest is identical to the linkage test based on our extension of the variance components model. I derive general formulas of optimal sample size and number of repeated measures for a given power or cost. Finally, I carry out analytical calculations and perform simulations to compare power for different sample sizes and number of repeated measures across several scenarios. My results show that modeling repeated measurements can provide substantial power improvements across genetic models. I give recommendations on whether to take repeated measures or recruit additional subjects for different levels of measurement errors and ratios of genotyping, subject recruitment and phenotyping costs.

In chapter III, I developed a novel discrete-generation framework and an efficient software package, called GENOME, to simulate genomic scale sequences from a population based on the coalescent model (Kingman 1982, Hudson 1983 & 1990, Donnelly and Tavaré 1995). In contrast to existing packages that implement the coalescent approach which are designed to simulate short genomic segments (~1Mb), GENOME can simulate much larger regions (>100Mb). As genome-wide studies become a reality, the proposed program should help geneticists to investigate sampling properties

of statistics that is evaluated on a genome-wide study and to compare the performance of different methods that may be applied to genomic scale data. In addition to features of standard coalescent simulators, the program allows for recombination rates to vary along the genome and accommodate flexible population histories. I show that GENOME provides the same LD patterns and frequency spectra as other coalescent simulators and conforms to theoretical predictions. The discrete-generation framework can be extended to incorporate features generally not available in the standard coalescent approach, including constraints on mating patterns and detailed models of regional gene-flow. Importantly, the framework still retains the computational efficiency of coalescent based simulators.

In chapter IV, working with my colleague Weihua Guan, I proposed a method for efficient matched analysis of cases and controls to account for unknown population stratification (Li, 1972) after genotyping a large number of markers in a genome-wide association study or large-scale candidate gene association study. Our method has three steps: 1) calculating similarity scores for pairs of individuals using genotype data; 2) matching sets of cases and controls based on the similarity scores; 3) using conditional logistic regression to perform association tests. Through computer simulations we show that our strategy correctly controls false positive rates, improves power to detect true disease predisposing variants and outperforms standard methods, such as the genomic-control method. We illustrate our method with genome-wide association data from the Pritzker Consortium bipolar study (Scott et al. 2009).

In chapter V, by using genome-wide association analysis, I generated a large scale map of genetic variants that influence the level of specific mRNAs. This map integrates

information on expression levels for 54,675 transcripts evaluated using Affymetrix arrays and genotypes for 408,273 SNPs from 400 individuals. The database created provides a general tool to investigate whether SNPs associated with any disease/trait alter transcription of genes in *cis* or *trans*. It has already proven useful in the study of several diseases/traits, including asthma, Crohn's disease, type II diabetes and fetal hemoglobin expression (Dixon et al. 2007, Cookson et al. 2009). Using the data, I also evaluated new strategies and methods for the analysis of gene-mapping data. For example, I show that integrating our data with publicly available resources (such as the HapMap genotypes) allowed us to estimate the effect of ~2 million untyped polymorphisms and identify variants that regulate the expression of 15% more genes than could be mapped with observed genotypes alone. In an ongoing analysis, I am expanding the eQTLs database to incorporating data on 47,293 transcripts measured with Illumina BeadChips and on 306,207 SNPs for an independent sample of 550 individuals.

1.4 References

Amos CI. 1994. Robust variance-components approach for assessing genetic linkage in pedigree. Am J Hum Genet. 54: 535-43

Almasy L, Blangero J. 1998. Multipoint quantitative trait linkage analysis in general pedigree. Am J Hum Genet. 62: 1198-211

Bhangale TR, Rieder MJ, Nickerson DA. 2008. Estimating coverage and power for genetic association studies using near-complete variation data. 40, 841 - 843

Barrett JC, Cardon LR. 2006. Evaluating coverage of genome-wide association studies. Nat Genet. 38, 659 - 662

Botstein, D. & Risch, N. 2003. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat Genet. 33 (suppl.): 228–237

Carlson CS, Eberle MA, Kruglyak L, Nickerson DA. 2004. Mapping complex disease loci in whole-genome association studies. Nature 429: 446-452

Cookson WOC, Liang L, Abecasis GR, Moffatt MF, Lathrop M. 2009. Mapping complex disease traits with global gene expression. Nat Rev Genet. 10: 184-194

Dixon AL, Liang L, Moffatt MF et al. 2007. A genome-wide association study of global gene expression. Nat Genet. 39:1202-7

Donnelly P., Tavaré S. 1995. Coalescents and genealogical structure under neutrality. Annu. Rev. Genet. 29: 401-421

Eberle MA, Ng PC, Kuhn K et al. 2007. Power to detect risk alleles using genome-wide tag SNP panels. PLoS Genet. 3(10): 1827–1837

Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, Patterson N, Gabriel SB, Topol EJ, Smoller JW, Pato CN, Pato MT, Petryshen TL, Kolonel LN, Lander ES, Sklar P, Henderson B, Hirschhorn JN, Altshuler D. 2004. Assessing the impact of population stratification on genetic association studies. Nat Genet. 36: 388-393

Hao K, Schadt EE, Storey JD. 2008. Calibrating the Performance of SNP Arrays for Whole-Genome Association Studies. PLoS Genet. 4(6): e1000109. doi:10.1371/journal.pgen.1000109

Hardy J, Singleton A. 2009. Genomewide Association Studies and Human Disease. N Engl J Med. 360:1759-1768

Hallmayer JF et al. 2005. Genetic Evidence for a Distinct Subtype of Schizophrenia

Characterized by Pervasive Cognitive Deficit. Am. J. Hum. Genet. 77:468–476

Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. 2002. A comprehensive review of genetic association studies. Genetics in Medicine 4(2):45-61

Hirschhorn JN, Daly MJ. 2005. Genome-wide association studies for common diseases and complex traits. Nat Rev Genet. 6(2):95-108

Hudson R.R. 1983. Properties of a neutral allele model with intragenic recombination. Theoretical Population Biology. 23:183–201

Hudson R.R. 1990. Gene genealogies and the coalescent process, Oxford Surveys in Evolutionary Biology 7: 1-4

Human Genome Project Information:
http://www.ornl.gov/sci/techresources/Human_Genome/medicine/assist.shtml

Jacquard A. 1972. Genetic information given by a relative. Biometrics 28:1101-14

Jimenez-Sanchez G, Childs B & Valle D. 2001. Human disease genes. Nature 409: 853–855

Kingman J.F.C. 1982. The coalescent. *Stochastic Process. Appl*. 13:235-248

Lander ES, Schork NJ. 1994. Genetic dissection of complex traits. Science 265:2037–2048

Chakravarti A. 1999. Population genetics–making sense out of sequence. Nat Genet 21:56–60

Lange K, Westlake J, Spence MA. 1976. Extensions to pedigree analysis. III. Variance components by the scoring method. Ann Hum Genet. 39:485-91

Leek JT & Storey JD. 2007. Capturing Heterogeneity in Gene Expression Studies by Surrogate variable Analysis. PLoS Genetics V3(9): e161. 1724-1735

Levy, D. et al. 2000. Evidence for a gene influencing blood pressure on chromosome 17. Genome scan linkage results for longitudinal blood pressure phenotypes in subjects from the Framingham Heart Study. Hypertension 36: 477–483

Li CC. 1972. Population subdivision with respect to multiple alleles. Ann Hum Genet. 33:23-29

Li Y., Willer C.j., Ding J., Scheet P. & Abecasis G.R. 2008. Markov model for rapid haplotyping and genotype imputation in genome wide studies. Submitted for publication; manuscript available from G.R.A. (email: goncalo@umich.edu)

Libioulle, C. et al. 2007. Novel Crohn Disease Locus Identified by Genome-Wide Association Maps to a Gene Desert on 5p13.1 and Modulates Expression of PTGER4. PLoS Genet 3(4): e58, 538-543

Luca D, Ringquist S, Klei L, Lee AB, Gieger C, Wichmann HE, Schreiber S, Krawczak M, Lu Y, Styche A, Devlin B, Roeder K, Trucco M. 2008. On the use of general control samples for genome-wide association studies: genetic matching highlights causal variants. Am J Hum Genet. 82:453-463

Marchini J., Howie B., Myers S., McVean G. & Donnelly P. 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 39: 906-13

McCarroll SA, Kuruvilla FG, Korn JM et al. 2008. Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat Genet. 40(10): 1166–1174

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JPA & Hirschhorn JN. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 9: 356-369

Moffatt MF, Kabesch M, Liang L et al. 2007. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. Nature 448: 470-3

Ott J.1999. Analysis of Human Genetic Linkage. Johns Hopkins University Press

Pe'er I, de Bakker PIW, Maller J, Yelensky R, Altshuler D, Daly MJ. 2006. Evaluating and improving power in whole-genome association studies using fixed marker sets. Nat Genet. 38, 663 - 667

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 38:904-909

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. 2000. Association mapping in structured populations. Am J Hum Genet. 67:170-181

Risch NJ. 2000. Searching for genetic determinants in the new millennium. Nature 405:847–856

Sachidanandam R et al. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409: 928–933

Sanna S. et al. 2008. Common variants in the GDF5-UQCC region are associated with variation in human height. Nat Genet 40:198-203

Scott LJ, et al. 2009. Genome-wide association and meta-analysis of bipolar disorder in individuals of European ancestry. Proc Natl Acad Sci USA 106:7501-7506

Scott L.J. et al. 2007. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science 316: 1341-5

Servin B., Stephens M. 2007. Imputation-based analysis of association studies: candidate regions and quantitative traits. PLoS Genet 3(7): e114

Scheet P., Stephens M. 2006. A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet. 78:629-44

Stegle O, Kannan A, Durbin R, Winn J. 2008. Accounting for Non-genetic Factors Improves the Power of eQTL Studies. Springer Berlin/Heidelberg, Volume 4955: 411-422

The International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851-861

The Wellcome Trust Case Control Consortium. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661-78

Willer C.J. et al. 2008. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat Genet. 40:161-9

Chapter II

VARIANCE COMPONENTS LINKAGE ANALYSIS WITH REPEATED MEASUREMENTS

2.1 Abstract

Background: When subjects are measured multiple times, linkage analysis needs to appropriately model these repeated measures. A number of methods have been proposed to model repeated measures in linkage analysis. Here, we focus on assessing the impact of repeated measures on the power and cost of a linkage study. ***Methods:*** We describe three alternative extensions of the variance components approach to accommodate repeated measures in a quantitative trait linkage study. We explicitly relate power and cost through the number of measures for different designs. Based on these models, we derive general formulas for optimal number of repeated measures for a given power or cost and use analytical calculations and simulations to compare power for different numbers of repeated measures across several scenarios. We give rigorous proof for the results under the balanced design. ***Results:*** Repeated measures substantially improve power and the proportional increase in LOD score depends mostly on measurement error and total heritability (if not otherwise defined, we use the term "heritability" as underline heritability which means the proportion of genetic variance out of total trait variance excluding measurement error) but not much on marker map, the number of alleles per

marker or family structure. When measurement error takes up 20% of the trait variability and 4 measures/subject are taken, the proportional increase in LOD score ranges from 38% for traits with heritability of 20% to 63% for traits with heritability of 80%. An R package is provided to determine optimal number of repeated measures for given measurement error and cost. Variance component and regression based implementations of our methods are included in the MERLIN package to facilitate their use in practical studies.

2.2 Introduction

In quantitative trait studies, taking repeated phenotype measures for each subject may increase the power. The approach is especially useful when measurement error is large or the relative cost of recruiting and genotyping additional subjects is high. It is important for a linkage analysis to appropriately take into account these repeated measures. Boomsma and Dolan [1] use structural equation modeling approach to analyze multivariate traits. Levy et al. [2] and de Andrade et al. [3] analyze longitudinal data by extending the standard variance components approach [4,5]. Although in principle repeated measurements can be treated as multivariate traits or longitudinal data [6,7,8,22], here we restrict our attention to modeling of repeated measurements for traits whose variance components do not change appreciably across time (except due to random measurement error). This allows us to focus the relationship between the power and cost of different study designs for quantitative trait linkage analysis and the number of repeated measures of the phenotype of interest taken for each subject. We also provide general implementations of these approaches, for both variance component [4] and

regression-based [18] linkage analysis, in our MERLIN software package.

To analyze repeated measures, summary statistics such as the average of observed measurements are usually used to take advantage of the models and implementations designed for single measure. In this case, standard packages such as SOLAR [5] and MERLIN [9] can then be used to analyze the averaged measurements. Unfortunately, when different numbers of measures are available for each subject, this approach is invalid and likely to result in a loss of efficiency.

Here repeated measures are modeled explicitly and we use asymptotic theorems to explore the power of QTL linkage tests. Combining these theorems with a cost function that summarizes phenotyping, genotyping and general fixed costs, the optimal number of repeated measures and sample size can be determined for a proposed study.

We consider three analytical strategies: (a) a full model that explicitly incorporates all measurements for all subjects; (b) a simplified model that uses only the average phenotypic measurement and the number of measurements taken for each subject; and (c) a further simplified model that only considers the average phenotypic measurement for each subject. We find that repeated measures provide substantial power improvements across genetic models. The proportional increase in expected LOD score depends mostly on measurement error and total heritability (if not otherwise defined, we use the term "heritability" as underline heritability which means the proportion of genetic variance out of total trait variance excluding measurement error) but not much on marker map or number of alleles per marker. Given a fixed sample size, analysis of repeated measures can have a dramatic impact on power. For example, when measurement error takes up 20% of the trait variability and 4 measures per subject are taken, the proportional increase

in expected LOD score ranges from 38% for traits with low heritability (e.g. 20%) to 63% for traits with high heritability (e.g. 80%). When 2 measures per subject are taken, the increase ranges from 23% to 36%. We identify the optimal number of repeated measures for different settings and show that when the number of measures is appropriately taken into account the average measure is a good balance between statistical power and computation efficiency.

2.3 Methods

In this section, we briefly review the variance component method for quantitative trait linkage analysis and then extend the model to accommodate repeated measures for arbitrary pedigrees, without inbreeding.

Variance Component Model

Let $Y = (Y_1,...,Y_n)'$ be the vector of quantitative trait values for a pedigree with *n* subjects and no inbreeding. Y is assumed to follow a multivariate normal distribution with mean $\mu = (\mu_1,...,\mu_n)'$ and variance-covariance matrix $\mathbf{\Omega}$. The effect of covariates can be modeled by letting $\mu = \mathbf{X}\beta$, where $\mathbf{X}$ is the design matrix for covariates and β are the coefficients for each covariate.

In general, $\mathbf{\Omega}$ will have the form: $\mathbf{\Omega} = \sum_i \sigma_i^2 \mathbf{\Omega}_i$, where σ_i^2 is a scalar variance component and $\mathbf{\Omega}_i$ is the corresponding covariance structure matrix which depends on the effect that σ_i^2 is representing. When major gene effect and polygenic effect are of interest, the $\mathbf{\Omega}$ can be defined as:

$$\mathbf{\Omega} = \mathbf{\Pi}\sigma^2_{mg} + 2\mathbf{\Phi}\sigma^2_{pg} + \mathbf{I}_n\sigma^2_e$$

where σ^2_{mg} is the additive genetic variance due to the major gene; the element π_{ij} of $\mathbf{\Pi}$ is the proportion of alleles shared IBD at the test locus between subjects *i* and *j*; σ^2_{pg} denotes the polygenic variance which is the genetic variance due to all residual additive effects not explained by the QTL; $\mathbf{\Phi}$ is a matrix of genetic kinship coefficients; σ^2_e is the subject-specific environmental variance and $\mathbf{I}_n$ is the $n \times n$ identity matrix [4,5,10, 11]. The model can be readily extended to include other effects of interest, such as genetic dominance.

The effects in variance component model can be assessed through likelihood ratio tests. For example, the test comparing $H_0 : \sigma^2_{mg} = 0$ *vs.* $H_1 : \sigma^2_{mg} > 0$ is used to assess evidence for a major gene impacting the quantitative trait.

Full Model with Repeated Measures

Let Y_{ij} be the j^{th} measurement of the i^{th} subject for the quantitative phenotype of interest. Assume m_i repeated measures are taken for subject *i*. Then, let:

$$\begin{aligned}
&Var(Y_{ij}) = \sigma^2_{mg} + \sigma^2_{pg} + \sigma^2_e + \sigma^2_m && i = 1,...,n \qquad j = 1,...,m_i \\
&Cov(Y_{ij_1}, Y_{ij_2}) = \sigma^2_{mg} + \sigma^2_{pg} + \sigma^2_e && j_1 \neq j_2, \forall i \qquad (1) \\
&Cov(Y_{i_1 j_1}, Y_{i_2 j_2}) = \pi_{i_1 i_2}\sigma^2_{mg} + 2\phi_{i_1 i_2}\sigma^2_{pg} && i_1 \neq i_2, \forall j_1, j_2
\end{aligned}$$

Here, σ^2_m represents the error specific to each measurement. This model is rather general. The covariance between repeated measurements of the same subject follows the compound symmetry structure [12]. This model is valid when measurement errors within

a subject are (a) independent or (b) equally correlated. In the latter setting the correlation between measurements is absorbed by the σ_e^2 component.

Under the assumption of normality and because the variance-covariance structure of residuals does not involve the fixed effect parameters β, the distribution of the likelihood ratio statistics about a variance component does not depend on the fixed effects β [13]. Although our model assumes no time effect in the variance-covariance matrix, if the time effect were included as a fixed effect, the results of this paper remain unchanged. Longitudinal data can therefore be accommodated in this limited manner by specifying time dependent covariates as the fixed effects. For simplicity and without loss of generality we assume the mean of quantitative trait is zero, with no covariate effects. Hence all the phenotypic variation can be explained through the similarity between relatives and the variance components σ^2_{mg}, σ^2_{pg}, σ^2_e and σ^2_m.

Model for Average Measures

An alternative to using the model specified in (1) above is to use the average measurement for each subject (e.g. [2]) instead of individual measurements. This approach results in smaller variance-covariance matrices and thus requires less computation.

Let $\overline{Y}_i = \frac{1}{m_i}\sum_{j=1}^{m_i} Y_{ij}$ be the average phenotype of subject *i*, for $i = 1,...,n$. Using these averages, the model for the variances and covariances becomes:

$$Var(\overline{Y}_i) = \sigma^2_{mg} + \sigma^2_{pg} + \sigma^2_e + \sigma^2_m / m_i \qquad i = 1,...,n \qquad (2)$$

$$Cov(\overline{Y}_{i_1}, \overline{Y}_{i_2}) = \pi_{i_1 i_2}\sigma^2_{mg} + 2\phi_{i_1 i_2}\sigma^2_{pg} \qquad i_1 \neq i_2$$

For balanced designs, where each subject has the same number of repeated measures,

it can be shown that, although model (2) requires less computation, models (1) and (2) give identical estimates of genetic variance components (excluding the environmental and measurement error variance component, which are not identifiable) and lead to the same value for linkage test statistics. Details of the equivalence proof are given in the Appendix 2.1.

Furthermore, when the number of repeated measures $m_i = m$ for all i, the standard variance component model:

$$Var(\overline{Y}_i) = \sigma^2_{mg} + \sigma^2_{pg} + \sigma^{2*}_e \qquad i = 1,...,n \qquad (3)$$

$$Cov(\overline{Y}_{i_1}, \overline{Y}_{i_2}) = \pi_{i_1 i_2}\sigma^2_{mg} + 2\phi_{i_1 i_2}\sigma^2_{pg} \qquad i_1 \neq i_2$$

can be used to construct linkage test without loss of efficiency, where $\sigma^{2*}_e = \sigma^2_e + \sigma^2_m / m$ and $m_i = m$ for $i = 1,...,n$. Therefore, standard software packages for QTL linkage analysis can be used.

When m_i's are not all equal, as in unbalanced designs, the standard variance component model (3) is not valid because σ^{2*}_e will be different across subjects, potentially distorting estimates of the genetic variance components and test statistics. Model (2), which takes into account different numbers of measures for each subject, remains a valid model. Through simulation, we show that it is slightly less efficient than the full model (1).

Analytical NCP for Balanced Design

For simplicity we based our analytical calculation on the balanced design. Under general regularity conditions, classical properties of likelihood ratio tests can be used to calculate the power of the test for a given sample size or the sample size required to

achieve a desired power [15].

Under the Null hypothesis when there is no major gene effect, the likelihood ratio test statistics is asymptotically distributed as $\frac{1}{2}\chi_1^2 : \frac{1}{2}(0)$, a mixture of a chi-squared distribution with one degree of freedom and a unit point mass at zero [25]. Under the alternative hypothesis, the likelihood ratio test statistics approximately follow a non-central chi-squared distribution with non-centrality parameter $\sum_{f=1}^{F}\delta_f$, where δ_f is the non-centrality contributed by the f^{th} of F families and

$$\delta_f = \log\left|(\sigma_{mg}^2 + \sigma_{pg}^2)2\mathbf{\Phi}_f + \sigma_e^{2*}\mathbf{I}_{n_f}\right| - E_\pi \log\left|\sigma_{mg}^2\mathbf{\Pi}_f + \sigma_{pg}^2 2\mathbf{\Phi}_f + \sigma_e^{2*}\mathbf{I}_{n_f}\right|. \quad (4)$$

Here, n_f is the size of the f^{th} family and E_π denotes an expectation over all possible allele-sharing states that can be calculated by averaging over all possible inheritance vectors [14,21]. The power of the test is then given by:

$$Power = \Pr(\chi^2_{1,\sum_{f=1}^{F}\delta_f} > C_\alpha)$$

where $\chi^2_{1,\sum_{f=1}^{F}\delta_f}$ follows a one degree of freedom chi-squared distribution with non-centrality parameter $\sum_{f=1}^{F}\delta_f$ and C_α is the 100(1-α) percentile of $\frac{1}{2}\chi_1^2 : \frac{1}{2}(0)$. To simplify the presentation, we consider F families with the same pedigree structure and denote $\delta = \delta_f$ for all f, so that $\sum_{f=1}^{F}\delta_f = F\delta$. For any desired power the required number of families F or of repeated measures m can then be solved numerically.

Cost-effectiveness

Formula (4) allows us to compare analytically power for different studies, each

characterized by a specific family structure, the number of families examined, F, and the number of repeated measures, m, for each subject. To study cost-effectiveness of different designs, we first introduce a cost function for each design. Let:

C_0 = Fixed cost of the study

C_s = Cost per subject recruited and genotyped (total Fn subjects)

C_p = Cost per phenotype measurement (m measures per subject) Total cost

$$C = C_0 + F \cdot n \cdot C_s + F \cdot n \cdot m \cdot C_p$$

From the last section, we know that the power is determined by $F\delta$ the non-centrality parameter and that δ depends on m through σ_e^{2*}. We denote δ as $\delta(m)$.

For any two combinations of m and F: (m_1, F_1) and (m_2, F_2), maintaining the same power requires $\delta(m_1)F_1 = \delta(m_2)F_2$. Without loss of generality, we assume $m_1 > m_2$ so that $\delta(m_1) > \delta(m_2)$. The total costs for the first design and the second design are $C_0 + F_1 \cdot n \cdot C_s + F_1 \cdot n \cdot m_1 \cdot C_p$ and $C_0 + F_2 \cdot n \cdot C_s + F_2 \cdot n \cdot m_2 \cdot C_p$, respectively. By simple algebra, taking m_1 (more) measures will provide the same power but a lower cost than taking m_2 (less) measures per subject when the following inequality holds:

$$\frac{C_s}{C_p} > CR_{m_1,m_2} \doteq \frac{m_1 - m_2 \cdot \delta(m_1)/\delta(m_2)}{\delta(m_1)/\delta(m_2) - 1} \tag{5}$$

CR_{m_1,m_2} defined above is called the break-event for cost ratio C_s / C_p, where taking m_1 measures is as cost-effective as taking m_2 measures per subject. When this cost ratio is higher (e.g. when phenotyping costs are relatively small compared to subject recruitment and genotyping costs), designs that take more measures per subject are favored.

Note that, for a given total cost (or power), the combination of m and F that maximizes power (or minimizes the total cost) can be identified numerically.

For unbalanced designs, CR can be approximated through simulation by using the ratio of expected LOD (ELOD) scores to replace $\delta(m_1)/\delta(m_2)$ in formula (5).

Simulation

We performed simulations to compare power for different number of repeated measures across several scenarios (varying distance between markers from ~0 to ~10 cM, considering SNP and microsatellite markers, and varying major gene heritability, total heritability and measurement error from 2% to 20%, 8% to 80% and 0% to 60% of trait variability, respectively).

For unbalanced designs, we attempted to mimic designs we have encountered in actual studies. For example, we simulated a situation where subjects with an extreme initial measurement were measured a second time. Thus, we first simulated one measurement for every subject. Next, we ordered subjects based on their simulated measurement and generated an additional measurement for $\alpha/2$ subjects (α is the proportion of subjects to get a second measure) at the top and $\alpha/2$ subjects at the bottom of the list. This design reflects the "intuition" that it may be more fruitful to focus effort on measuring extreme subjects. In this design, the average number of measurements per subject is $1+\alpha$. We let α=20% and α=10%. In an alternative unbalanced design, referred to as the random design, the number of measures for each subject follows an exponential distribution. This mimics the situation where measurements are missing completely at random. For each subject, we draw independent

random number (rounding to the nearest greater integer) from an approximate exponential distribution with mean equal to 0.5, 1 and 2, respectively. The maximum number of measurements per subject was set to 4.

In each simulation, we simulated 1000 families and the results are based on 2000 simulations. The average of LOD scores at the QTL is used to estimate the ELOD. Power is measured by the proportion of likelihood ratio test p-values less than 0.001. The cost-effectiveness break-event for cost ratio, CR_{m_1,m_2}, is also presented to facilitate comparison between different designs.

2.4 Results

Analytical Results

Based on the average model (formulas 2 & 4), we can examine the ELOD (hence the power) for different settings under the balanced design and assuming markers are fully informative. Figure 2.1 shows how the ELOD changes as the heritability, defined as $(\sigma^2_{mg} + \sigma^2_{pg})/(\sigma^2_{mg} + \sigma^2_{pg} + \sigma^2_e)$, increases for different numbers of repeated measures. For example, when the heritability is 40%, increasing the number of measures from 1 to 3 results a 2-fold increase in ELOD. We also note that taking more repeated measures results in more rapid increases in ELOD for simulated traits with greater heritability.

According to (5) we can determine the optimal number of repeated measures for different ratios of genotyping and phenotyping cost and degrees of measurement error. Figure 2.2 shows the contour plot for the optimal number of repeated measures when the cost ratio C_s / C_p ranges from 0.01 to 50 and measurement error variance ranges from 0.11 to 1.5 (corresponding to 10% – 60% of the total trait variance). For example, when

measurement error variance is 0.4 (corresponding to 28.6% of the total trait variance), taking 2 measurements per subject is cost-effective if the cost ratio is between 1.11 and 4.17. When the ratio of genotyping and recruitment costs to phenotyping costs is <1.11, it is preferable to take a single measurement and collect more subjects. When this ratio is >4.17, it is preferable to take additional measurements and collect fewer subjects. When the cost ratio is between 9.09 and 15.62, taking 4 measurements per subject is the best. The ranges of figure 2.2 should include a variety of realistic scenarios. For example, chip based genotyping for genome-wide linkage studies typically costs a few hundred dollars per subject whereas phenotyping costs are widely variable, ranging from a few dollars per subject (for mail-in questionnaires [24]) to several hundred dollars (for expensive imaging measures or biological assays). The measurement error as well as the intra-individual environmental variance could range from very low (5%), for anthropometric measures such as height, to quite high (40%), for traits such as micro-array summaries of gene expression and questionnaire based assessments of personality.

Simulation Results

We simulated three scenarios: (1) One microsatellite marker with 20 alleles and 0 cM between the marker and the QTL to approximate a fully informative marker. (2) Ten microsatellite markers each with 4 alleles and with 10 cM separating consecutive markers; the QTL placed in the middle of the markers. (3) Fifty SNPs and 2 cM between consecutive markers; the QTL again placed in the middle of the SNPs. For each scenario, the trait variance excluding measurement error was fixed at 100, that is $\sigma^2_{mg} + \sigma^2_{pg} + \sigma^2_e = 100$. The major gene effect σ^2_{mg} was set at 20. Polygene effects σ^2_{pg}

ranged from 0 to 60. Measurement error variance σ_m^2 ranged from 11 to 150 (corresponding to 10% – 60% of the total trait variance). In each independent sample, we simulated 1000 nuclear families with 4 offspring each. Relative power for designs with different numbers of measurements varied only slightly for different family structures (table 2.4, which includes sibships with 2-6 siblings and cousin pedigrees) and so our presentation focuses on nuclear families with 4 offspring.

Simulation results again show repeated measures can provide substantial power improvements (table 2.1, figure 2.3). Table 2.1 shows the ELOD and power of balanced designs for a simulated microsatellite panel (scenario 2). Taking 2 measures per subject increases ELOD by 52% to 75% and power at $\alpha = 0.001$ by 63% to 78%. Figure 2.3 shows the average LOD score profile for the microsatellite panel (scenario 2, major gene effect 20 (or 12% total variance), polygene effect 40 (or 24% total variance), and measurement error 67 (or 40% total variance). In this case, taking 1 measure per subject results in an average peak LOD score of only 2.22. Taking repeated measures increases the average peak LOD to 3.69 (2 measures) and 5.04 (4 measures).

Since IBD estimation does not affect the accuracy of estimates of measurement error variance, the proportional increase in expected LOD score (ELOD Ratio) depends mostly on measurement error and total heritability but not much on marker map or number of alleles per marker (table 2.2), which mostly impact the precision of QTL effect size estimates. This suggests that the optimal design (in terms of optimal number of repeated measures) is relatively insensitive to the genotyping platform selected. Table 2.2 shows the average ELOD ratios for 4 repeated measures under three scenarios. Based on the ELOD ratio and the condition to maintain the same power, $\delta(m_1)F_1 = \delta(m_2)F_2$, we can

calculate the savings in sample size when using 4 repeated measures. For example, for the first setting when measurement error variance is 11 (10% total variance) and total heritability is 20%, the sample size (number of subjects) required when taking 4 measures per subject is 85% (1/1.17) of the sample size required when taking 1 measure per subject.

When the ELOD ratios are available, it is possible to calculate the break-event CR_{m_1,m_2} for cost ratio C_s / C_p using (5). For example, when measurement error variance is 25 (20% of the total variance) and heritability is 20%, if genotyping and recruitment costs per subject are more than 6.83 higher than phenotyping costs, taking 4 measures per subject is more cost-effective than taking 1 measure per subject. The cost ratio needs to exceed 14.44 so that taking 4 measures is better than taking 2 measures per subject (table 2.2).

For the unbalanced design where 20% (or 10%) of subjects with an extreme first measurement are measured one more time, the cost ratios can be calculated in a similar way because the total number of measures is fixed. We denote these two designs as "m=1.2" and "m=1.1" respectively. Now we can compare different designs using the cost ratio CR_{m_1,m_2}. The results are summarized in table 2.3. The cost ratio $CR_{1.1,1}$ is relative large, $CR_{1.1,1} = \infty$ in the first row means designs with m=1.1 are never more cost-effective than taking one measure per subject when the measurement error is small. Note that since $CR_{2,1.2} < CR_{1.2,1}$ and $CR_{2,1.1} < CR_{1.1,1}$, the unbalanced designs can always be outperformed by a balanced design that involves either 2 or 1 measures per subject depending on the cost ratio C_s / C_p. So in terms of cost-effectiveness, balanced designs are always better than these particular unbalanced designs no matter what the cost ratio

C_s / C_p. Using the data in table 2.3, we can draw a similar contour plot as figure 2.2. This plot is presented in figure 2.4. The parameter settings are equivalent to figure 2.2. The plot shows the theoretical result (figure 2.2) is consistent with the simulation result (figure 2.4).

Comparing efficiency between the full model and the average model

Using simulation, we next compared the efficiency between the full model (1) and the average model (2) for unbalanced designs. Both models take into account the different number of measurements across individuals, give valid likelihood functions and control type I error rate adequately.

Figure 2.5 shows the ELOD ratio of the full model vs. the average model. For both unbalanced designs, the full model did not provide substantially more efficiency than the average model (only in the extreme design, the full model increases ELOD by 1% on average across all scenarios). The largest increase in ELOD was 9% in settings where the measurement error was large and individuals with an initial extreme measurement were reassessed.

2.5 Discussions

When subjects are measured multiple times, it is important for a linkage analysis to appropriately take into account these repeated measures. In this study, we extend the variance components approach to model repeated measures in a quantitative trait linkage study. Our model can explicitly relate the power and cost of different sampling designs. We give the general formulas of optimal sample size and number of repeated measures for a given power or cost. We show for the case of a balanced design where the same

number of measurements is taken for each subject, a standard linkage test that takes the average of measures as the trait of interest is identical to a linkage test based on an appropriate extension of the variance components model.

In our model, the covariance between repeated measures of the same subject follows the compound symmetry structure. This model is valid when measurement errors within a subject are either independent or else equally correlated. It is one of the most commonly used covariance structures in the repeated measures literature. When necessary it should be possible to refine our model to include dominance effects, twin environment or other variance-covariance components or even to incorporate covariate effects into the variance-covariance matrix. In particular, time effects can be introduced into the variance-covariance structure to allow for longitudinal changes in variance components[3].

Through both analytical calculation and simulation, we find that repeated measures provide substantial power improvements across genetic models. The proportional increase in expected LOD score (ELOD Ratio) depends mostly on measurement error and total heritability but not much on marker map or number of alleles per marker. This suggests that the optimal design (in terms of optimal number of repeated measures) will be similar for a range of genotyping strategies (provided they are similar in cost). We give contour plots to help investigators decide on the optimal number of repeated measures for different levels of measurement errors and ratios of genotyping, subject recruitment and phenotyping costs. The R code to help determine the optimal number of repeated measures is available from our website.

Precise trade-offs can be obtained by examining Figure 2.2 and the R package. Still,

our results allow us to make some general recommendations. When measurement error is high, accounting for ~50% of the trait variance, it is typically cost effective to collect 2 or more measures per subject when the ratio of phenotyping to genotyping costs per subject is <16 fold. If genotyping is carried out using a commercially available SNP array that typically costs $100 - $200 per subject, it will almost always be worthwhile to phenotype each individual multiple times, given that most phenotyping assays cost <$1600 - $3200 per measurement. When measurement error is small, accounting for ~10% of the trait variance, it is only cost effective to collect 2 or more measures per subject when phenotyping is relatively inexpensive, costing no more than 0.154 times the cost of genotyping. With the same genotyping costs as above, this would correspond to $15 - $30 per measurement and would only be worthwhile for the most inexpensive phenotypes (such as those that rely on mail-in questionnaires or very simple trait measurements). In other situations, it will be more efficient to collect additional subjects.

For unbalanced designs, a standard linkage test that takes the average measurement as the trait of interest and ignores the number of measures is not valid. A model that uses the average measurement as the trait but takes into account the different number of measures for each subject, i.e. model (2), is a valid alternative to the full model. The advantage of model (2) is that it is less computationally intensive and, typically, only slightly less powerful than the full model. We implemented both the average model and the full model in the MERLIN package [9,23]. We also assessed the effect of ignoring the imbalance and taking the average as a single trait. Table 2.5 shows simulations where a random half of subjects were measured 2, 4, or 10 times while the other half were measured only once. The results suggest that ignoring imbalance could lead to

approximately correct type I error but could lose power (at pvalue<0.001) by 2-5% or decrease in ELOD by 3-15%.

While the average measure is widely used as a useful summary statistics of repeated measures, an alternative is the median of repeated measures from the same individual. Median has the advantage of being robust to outliers and has the asymptotic normal distribution when the number of repeated measures per individual is large, in this study we focus on the average measure because (1) it is often used by researchers in the community, (2) it has a known exact distribution and it is the sufficient statistics for the major gene effect, polygene effect and subject-specific environmental effect (appendix 2.3), (3) the number of repeated measures per individual is often small thus asymptotic theorem of the median does not hold. As a result, the asymptotic distribution of the likelihood ratio test of variance components would be hard to derive.

In our simulations, parental genotypes were used to help estimate IBD sharing between pairs of relatives. We also investigated the effect of parental phenotypes on power. Figure 2.6 shows the expected LOD scores with and without using parental phenotypes at a fully informative marker under the same scenario of figure 2.1. For a simulated trait with relatively low heritability, the additional measures from parents only slightly increase the expected LOD scores, suggesting that phenotyping parents is unlikely to be cost effective. For highly heritable traits, parental phenotypes do substantially increase the expected LOD scores especially for larger number of repeated measures. In this case, there will be a trade-off between phenotyping the parents and collecting more offspring genotypes and phenotypes.

While rigorous proof or comprehensive simulation is required to draw a solid

conclusion, we could get some hints about the gain in power in association tests by using repeated measures. Assuming a standard linear model, the maximum likelihood estimate of the regression coefficient follows a normal distribution with variance proportional to the total variance of the trait of interest. We also know that the maximum likelihood estimate of the variance of the model is the sample variance and it follows a scaled Chi-square distribution with the scale equal to the variance of the trait. Hence the variance of the sample variance is proportional to the square of the trait variance. So the gain in power by using repeated measures will be larger for test about the variance component (test for linkage) compared to the test about the regression coefficient (test for association). For example, if using repeated measure reduces the total trait variance by 20%, the variance of regression coefficient estimate will be reduced by 20% but the variance of the variance component estimate will be reduced by 36%. We should emphasize that the difference in gain of power will be depended on the underline model and need to be addressed in a rigorous way.

In cases of non-normality of the trait distribution and selected sampling, robust statistics such as score statistics [16,17] or regression-based statistics [18] can help adequately control the type I error and increase power. Intensive simulations [17,18] have shown that the regression-based model implemented in MERLIN-REGRESS [18] is robust to violations of normality, selected sampling and population parameter misspecification while achieving high power. Nash et al. 2004 discussed the treatment of average repeated measures in the regression-based model [20]. We take another approach which leads to simpler formulation and hence easier implementation of the software. We show that the regression-based model can be extended to incorporate individual repeated

measures as well as average measures [Appendix 2.2] and this alternative is implemented in MERLIN-REGRESS [18].

2.6 Appendices

Appendix 2.1: Equivalence between full model and average measurement model for balanced number of measurements

When each subject is measured the same number M of times, it can be shown that the full model (1) and average measurement model (3) are equivalent.

Let vector Y_j represent all measurements for subject j. In full model (1), the variance for vector Y_j is $Var(Y_j) = (\sigma^2_{mg} + \sigma^2_{pg} + \sigma^2_e)\mathbf{11'} + \sigma^2_m\mathbf{I}$ and the covariance between Y_j for subject j and Y_k for subject k is $Cov(Y_j, Y_k) = (\pi_{jk}\sigma^2_{mg} + 2\phi_{jk}\sigma^2_{pg})\mathbf{11'}$, where vector **1** consists of all 1's and **I** is the identity matrix.

We first apply a linear transformation **T** on multiple measurements Y_j

$$Y_j^* = \mathbf{T}(Y_{j1}, Y_{j2}, ..., Y_{jM})' \tag{A1}$$

where

$$\mathbf{T} = \begin{pmatrix} \frac{1}{M} & \frac{1}{M} & \cdots & \frac{1}{M} & \frac{1}{M} \\ -\frac{1}{2} & \frac{1}{2} & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \frac{1}{2} & 0 \\ 0 & 0 & \cdots & -\frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$

Thus the covariance matrix for the transformed vector Y_j^* is

$$\begin{aligned} Var(Y_j^*) &= \mathbf{T}Var(Y_j)\mathbf{T'} \\ &= (\sigma^2_{mg} + \sigma^2_{pg} + \sigma^2_e)\mathbf{T11'T'} + \sigma^2_m\mathbf{TT'} \end{aligned} \tag{A2}$$

$$\begin{aligned} Cov(Y_j^*, Y_k^*) &= \mathbf{T}Cov(Y_j, Y_k)\mathbf{T'} \\ &= (\pi_{jk}\sigma^2_{mg} + 2\phi_{jk}\sigma^2_{pg})\mathbf{T11'T'} \end{aligned} \tag{A3}$$

Simple algebra gives $\mathbf{T11'T'} = \begin{pmatrix} 1 & 0 \\ 0 & 0_{(M-1)\times(M-1)} \end{pmatrix}$, $\mathbf{TT'} = \begin{pmatrix} 1/M & 0 \\ 0 & \mathbf{A}_{(M-1)\times(M-1)} \end{pmatrix}$ where **A** is some *(M-1)* by *(M-1)* matrix. Let Z_j and Y^*_{jm} denote the first and the rest of the elements in the transformed vector Y^*_j respectively. Then, $Z_j = \sum_{m=1}^{M} Y_{jm}/M$, and $Y^*_{jm} = \frac{Y_{jm} - Y_{j(m-1)}}{2}$ for *m=2,...,M*. Covariances (A2) and (A3) imply

$$\begin{aligned} &Var(Z_j) = \sigma^2_{mg} + \sigma^2_{pg} + (\sigma^2_e + \sigma^2_m/M) \\ &Cov(Z_j, Z_k) = \pi_{jk}\sigma^2_{mg} + 2\phi_{jk}\sigma^2_{pg}, \text{ for } j \neq k \\ &Var(Y^*_{jm}) = \sigma^2_m/2, \text{for } m = 2,...,M \\ &Cov(Y^*_{jm}, Y^*_{j(m+1)}) = -\sigma^2_m/4, \text{for } m = 2,...,M-1 \end{aligned} \tag{A4}$$

and all other covariances are 0. Thus, the full model (1) implies model (A4). The reverse is also true since the transformation (A1) is not singular. Now we assume in the average model (3), $\sigma^{2*}_e = \sigma^2_e + \sigma^2_m/M$. By comparing model (A4) and model (3), we can see model (A4) implies model (3).

Let $\mathbf{\Omega}_Z$ denote the variance-covariance matrix of vector $z = (Z_1,...,Z_J)'$, and $\mathbf{\Omega}_*$ denotes the variance-covariance matrix of $y^* = (y^*_2,...,y^*_M)'$, where $y^*_m = (Y^*_{1m},...,Y^*_{Jm})'$ for *m=2,...,M*. Model (A4) shows vector z and y^*_m are orthogonal, indicating the variance-covariance matrix of $(z, y^*)'$ with form $\begin{pmatrix} \mathbf{\Omega}_Z & 0 \\ 0 & \mathbf{\Omega}_* \end{pmatrix}$. Thus, the likelihood of a family is

$$L = \frac{1}{(\sqrt{2\pi})^{M*J}} \left| \begin{matrix} \mathbf{\Omega}_Z & 0 \\ 0 & \mathbf{\Omega}_* \end{matrix} \right|^{-1} \exp\left\{ -\frac{1}{2} (z', y^{*'})' \begin{pmatrix} \mathbf{\Omega}_Z^{-1} & 0 \\ 0 & \mathbf{\Omega}_*^{-1} \end{pmatrix} \begin{pmatrix} z \\ y^* \end{pmatrix} \right\}$$

$$= \frac{1}{(\sqrt{2\pi})^{J}} \frac{1}{\left|\mathbf{\Omega}_Z\right|} \exp\left\{ -\frac{1}{2} z' \mathbf{\Omega}_Z^{-1} z \right\} \times \frac{1}{(\sqrt{2\pi})^{(M-1)*J}} \frac{1}{\left|\mathbf{\Omega}_*\right|} \exp\left\{ -\frac{1}{2} y^{*'} \mathbf{\Omega}_*^{-1} y^* \right\}.$$

The first part of the above likelihood is exactly the likelihood in model (3). Since the second part in the last expression of the likelihood only contains information about σ_m^2 and does not carry any information about σ_{mg}^2, σ_{pg}^2 and σ_e^{2*}, the maximum likelihood estimates for $(\sigma_{mg}^2, \sigma_{pg}^2, \sigma_e^2)$ in the average measurement model are identical to those for $(\sigma_{mg}^2, \sigma_{pg}^2, \sigma_e^{2*})$ in the full model. Therefore, for balanced data, the average of the repeated measurements can be treated as the actual trait, and the standard variance components analysis is the equivalent to the full model. When the number of measurements is not balanced, the equivalence between the above two models (1) and (3) does not hold anymore and the full model uses more information than the average measurement model (3).

Appendix 2.2: Extension of the regression model for linkage analysis in Sham et al. 2002 [18] to accommodate repeated measures

To incorporate repeated measures into the regression model, we only need to re-specify the form for the expectation and covariance that involves the squared sum S and squared difference D, other terms in the model will be identical to Sham et al. 2002. In fact, the regression model can be extended to model individual measures as well as the average measures and the relative performance of models using all available measurements, the average measurement and the count of measurements for each subject, or just the average measurement is analogous to the performance of formulas (1)-(3) in the variance component model.

Let c be the within-subject correlation $\frac{\sigma_{mg}^2+\sigma_{pg}^2+\sigma_e^2}{\sigma_{mg}^2+\sigma_{pg}^2+\sigma_e^2+\sigma_m^2}$ and H^2 be the total heritability $\frac{\sigma_{mg}^2+\sigma_{pg}^2}{\sigma_{mg}^2+\sigma_{pg}^2+\sigma_e^2+\sigma_m^2}$. Assuming the full model (1), all pairs of individual measures standardized by their population mean μ and variance $\sigma^2=\sigma_{mg}^2+\sigma_{pg}^2+\sigma_e^2+\sigma_m^2$ are considered. The vector of squared sums is $[S_{i_1j_1,i_2j_2}=(\frac{Y_{i_1j_1}-\mu}{\sigma}+\frac{Y_{i_2j_2}-\mu}{\sigma})^2]$ and similarly the vector of squared differences is $[D_{i_1j_1,i_2j_2}=(\frac{Y_{i_1j_1}-\mu}{\sigma}-\frac{Y_{i_2j_2}-\mu}{\sigma})^2]$. In the expectation and covariance of the squared sums S and squared differences D, only the form of correlation needs to be changed and it is equal to:

$$r_{i_1j_1,i_2j_2}=\mathrm{cov}(\frac{Y_{i_1j_1}-\mu}{\sigma},\frac{Y_{i_2j_2}-\mu}{\sigma})=\begin{cases}c & \text{if } i_1=i_2\\ 2\phi_{i_1i_2}\mathrm{H}^2 & \text{otherwise}\end{cases}$$

The parameter *c* as well as the population mean, μ, variance, σ^2, and total heritability H^2 will need to be specified by the user.

The remaining terms need to be considered are the covariance between S, D and $\hat{\pi}$: $\{\mathrm{Cov}_1(S_{i_1j_1,i_2j_2}, \hat{\pi}_{k_1l_1,k_2l_2})\}$ and $\{\mathrm{Cov}_1(D_{i_1j_1,i_2j_2}, \hat{\pi}_{k_1l_1,k_2l_2})\}$. For $i_1 \neq i_2$ and $k_1 \neq k_2$, these terms remain unchanged. For $k_1 = k_2$, $\hat{\pi}_{k_1l_1,k_2l_2} = 1$ so the covariance is 0. For $i_1 = i_2$ and $k_1 \neq k_2$, since the joint distribution of $(Y_{i_1j_1}, Y_{i_2j_2})$ does not involve π the covariance is again 0. This suggests that we only need to include the pair of measures that involve different subjects; greatly reducing the dimension of mean vectors and covariance matrixes. More importantly, all formulas in Sham et al. 2002 can be directly applied if we only include pairs of measures that are from different subjects.

Assuming model (2) for average measures under unbalanced designs, the variance for each average measure will be different. Unlike the treatment in [20], we propose to standardize the average measures $\{\overline{Y}_i\}$ by the population mean μ and their own variances $\{\sigma_i^2 = \sigma_{mg}^2 + \sigma_{pg}^2 + \sigma_e^2 + \sigma_m^2 / m_i = c\sigma^2 + (1-c)\sigma^2 / m_i\}$ so that they are multivariate normal with mean 0 and variance 1 and results in Appendix A of Sham et al. 2002 can apply. Hence the formulae for covariances of the squared sums S and squared differences D remain unchanged. Only the correlation between a pair of standardized average measures needs to be changed to:

$$r_{ij} = \mathrm{cov}\left(\frac{\overline{Y}_i - \mu}{\sigma_i}, \frac{\overline{Y}_j - \mu}{\sigma_j}\right) = 2\phi_{ij}\mathrm{H}^2 \frac{\sigma^2}{\sigma_i\sigma_j}$$

For covariance between S, D and $\hat{\pi}$, following a similar derivation to Drigalenko 1998 [19], we have:

$$\mathrm{Cov}_1(S_{ij},\hat{\pi}_{kl})=\frac{2Q\mathrm{Cov}_1(\hat{\pi}_{ij},\hat{\pi}_{kl})}{\sigma_i\sigma_j} \quad \text{and} \quad \mathrm{Cov}_1(D_{ij},\hat{\pi}_{kl})=\frac{-2Q\mathrm{Cov}_1(\hat{\pi}_{ij},\hat{\pi}_{kl})}{\sigma_i\sigma_j}$$

Other equations will be identical to Sham et al. 2002.

Analogous to model (3) for average measures with balanced designs, the average measures $\{\overline{Y}_i, i=1...n\}$ can be treated as an actual trait and standardized by the population mean μ and the variance $\sigma^2_{mg}+\sigma^2_{pg}+\sigma^2_e+\sigma^2_m/m=c\sigma^2+(1-c)\sigma^2/m$. So the model in Sham et al. 2002 can apply.

Appendix 2.3: Sufficient statistics for unbalanced design

When subjects have different number of repeated measures, it can be shown that $(\overline{Y}_1,..,\overline{Y}_n)$ is the sufficient statistics for σ_{mg}^2, σ_{pg}^2 and σ_e^2.

Use the same non-singular linear transformation in appendix 2.1 and slightly different notation to reflect the difference in the number of repeated measures, we define:

$$\begin{cases} Z_{i1} = \overline{Y}_i = \dfrac{1}{m_i}\sum_{i=1}^{m_i} Y_{ij} \\ Z_{ij} = \dfrac{Y_{i,j-1} - Y_{ij}}{2} \text{ for } j = 2..m_i \end{cases}$$

By simple algebra:

$$Var(Z_{i1}) = \sigma_{mg}^2 + \sigma_{pg}^2 + \sigma_e^2 + \sigma_m^2 / m_i$$

$$Cov(Z_{i1}, Z_{i'1}) = E(Z_{i1}Z_{i'1}) = \frac{1}{m_i m_{i'}}\sum_{j=1}^{m_i}\sum_{j'=1}^{m_{i'}} E(Y_{ij}Y_{i'j'}) = \pi_{ii'}\sigma_{mg}^2 + 2\phi_{ii'}\sigma_{pg}^2 \text{ for } i \neq i'$$

$$Cov(Z_{i1}, Z_{ij}) = \frac{1}{2m_i}(E(Y_{i,j-1}^2) + \sum_{j' \neq j-1} E(Y_{ij'}Y_{i,j-1}) - E(Y_{i,j}^2) - \sum_{j' \neq j} E(Y_{ij'}Y_{i,j})) = 0 \text{ for } j = 2,..,m_i$$

$$Cov(Z_{i1}, Z_{i'j}) = \frac{1}{2m_i}(\sum_{j'=1}^{m_i} E(Y_{ij'}Y_{i',j-1}) - \sum_{j'=1}^{m_i} E(Y_{ij'}Y_{i',j})) = 0 \text{ for } i \neq i' \text{ and } j = 2,..,m_{i'}$$

$$Var(Z_{ij}) = \sigma_m^2 / 2, \text{ for } j = 2,...,m_i$$

$$Cov(Z_{ij}, Z_{ij'}) = -\sigma_m^2 / 4, \text{ for } j,j' > 1 \text{ and } |j - j'| = 1$$

$$Cov(Z_{ij}, Z_{ij'}) = 0, \text{ for } j,j' > 1 \text{ and } |j - j'| > 1$$

$$Cov(Z_{ij}, Z_{i'j'}) = \frac{1}{4}(E(Y_{i,j-1}Y_{i',j'-1}) - E(Y_{i,j-1}Y_{i',j'}) - E(Y_{i,j}Y_{i',j'-1}) + E(Y_{i,j}Y_{i',j'})) = 0, \text{ for } i \neq i'$$

Because $(Z_{i1},...Z_{n1})$ and $((Z_{ij}))_{i=1,..,n,j>1}$ both follow multivariate normal distribution and they have zero correlation between each other, they are independent. The likelihood of the full model can be written as: Likelihood[$(Z_{i1},...Z_{n1})$]*Likelihood[$((Z_{ij}))_{i=1,..,n,j>1}$]. Since the distribution of $((Z_{ij}))_{i=1,..,n,j>1}$ does not involve σ_{mg}^2, σ_{pg}^2 and σ_e^2, $(Z_{i1},...Z_{n1})$ is the sufficient statistics for parameters σ_{mg}^2, σ_{pg}^2 and σ_e^2.

2.7 Tables and figures

Table 2.1 Power increment by taking repeated measures (scenario 2)

Polygene Effect (% total var.)	No Measurement Error or M=∞			M=4			M=2			M=1	
	ELOD	Power	Ratio	ELOD	Power	Ratio	ELOD	Power	Ratio	ELOD	Power
0.0 (0%)	4.88	0.94	2.71	3.58	0.80	1.99	2.74	0.64	1.52	1.80	0.36
0.2 (12%)	5.91	0.97	2.93	4.16	0.87	2.06	3.11	0.70	1.54	2.02	0.41
0.4 (24%)	7.48	1.00	3.35	5.01	0.94	2.25	3.63	0.80	1.63	2.23	0.48
0.6 (36%)	10.30	1.00	4.17	6.32	0.98	2.56	4.32	0.88	1.75	2.47	0.54

Measurement Error Variance = 67 (40% of the total trait variance). M = the number of repeated measures. The ratio is the ELOD ratio between M measures and 1 measures per subject. Scenario (2): Ten microsatellite markers each with 4 alleles and spaced 10 cM apart; the QTL placed in the middle of the markers.

Table 2.2 Cost-effectiveness analysis for 4 Repeated Measures vs. 1 Measure

Measurement Error Var. (% total var.)	Heritability (% total var.)	Ave ELOD Ratio	Sample Size Savings	$CR_{4,1}$	$CR_{4,2}$
11 (10%)	0.20 (18%)	1.17	0.15	16.31	31.20
	0.60 (54%)	1.23	0.19	12.04	26.75
25 (20%)	0.20 (16%)	1.38	0.28	6.83	14.44
	0.60 (48%)	1.51	0.34	4.84	10.41
67 (40%)	0.20 (12%)	2.01	0.50	1.97	4.55
	0.60 (36%)	2.28	0.56	1.34	3.33
150 (60%)	0.20 (8%)	3.07	0.67	0.45	1.33
	0.60 (24%)	3.39	0.71	0.25	0.78

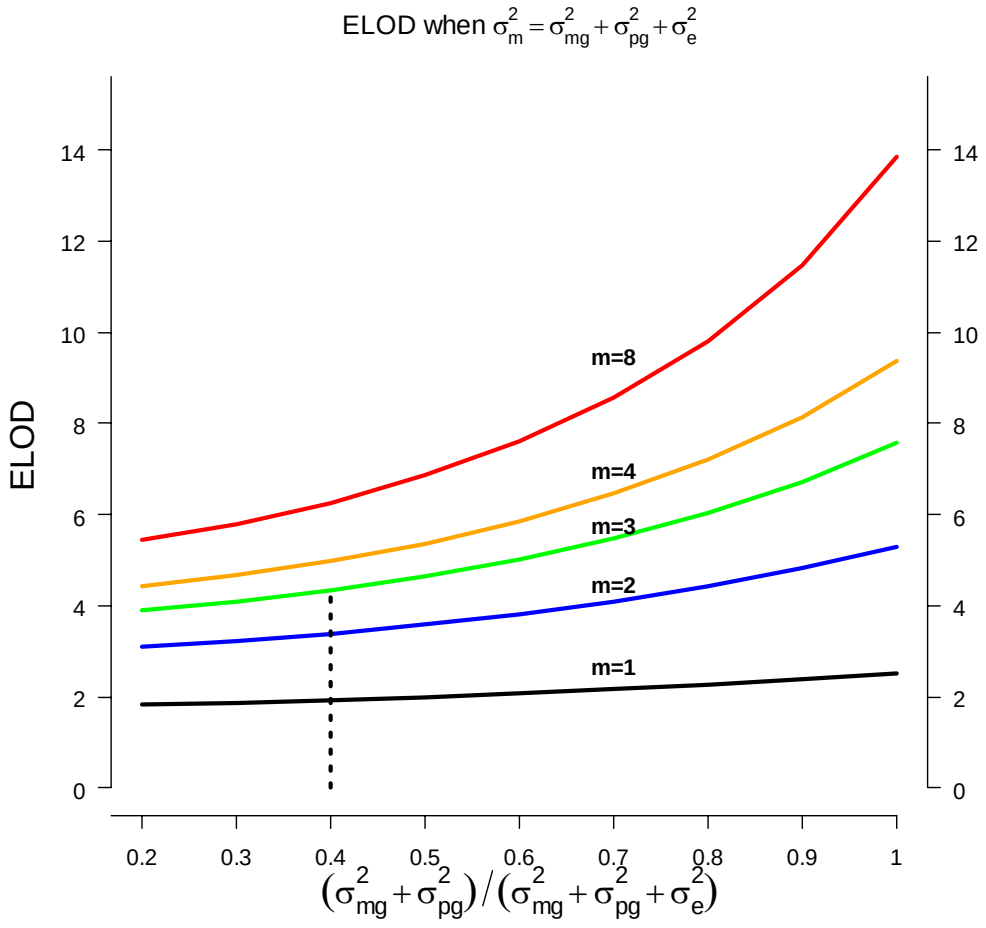
CR_{m_1,m_2} is defined in (5). $CR_{4,2}$ is also listed here for comparison purpose. When $C_s / C_p > CR_{m_1,m_2}$, taking m_1 measures is better than taking m_2 measures per subject. Heritability is defined as $(\sigma^2_{mg} + \sigma^2_{pg})/(\sigma^2_{mg} + \sigma^2_{pg} + \sigma^2_e)$ where $\sigma^2_{mg} + \sigma^2_{pg} + \sigma^2_e = 100$ and the major gene effect σ^2_{mg} is fixed to 20. Average of ELOD ratio is the average across three scenarios that give similar results: (1) a highly informative microsatellite marker with 20 alleles and 0 cM between the marker and the QTL. (2) Ten microsatellite markers each with 4 alleles and spaced 10 cM apart; the QTL placed in the middle of the markers. (3) Fifty SNPs and spaced 2 cM apart; the QTL again placed in the middle of the SNPs.

Table 2.3 Cost Ratios for the Comparison between Different Designs

Measurement Error Var.	Heritability (% total var.)	$CR_{4,2}$	$CR_{2,1}$	$CR_{2,1.2}$	$CR_{2,1.1}$	$CR_{1.2,1}$	$CR_{1.1,1}$
11 (10%)	0.20 (18%)	31.20	8.38	7.16	7.34	19.00	∞
	0.60 (54%)	26.75	5.67	5.26	5.22	7.57	14.00
25 (20%)	0.20 (16%)	14.44	3.29	2.77	2.97	6.50	9.00
	0.60 (48%)	10.41	2.30	1.91	2.00	4.45	9.00
67 (40%)	0.20 (12%)	4.55	0.85	0.53	0.65	2.75	5.00
	0.60 (36%)	3.33	0.52	0.36	0.38	1.07	2.00
150 (60%)	0.20 (8%)	1.33	0.09	0	0	0.76	1.31
	0.60 (24%)	0.78	0.03	0	0	0.40	0.43

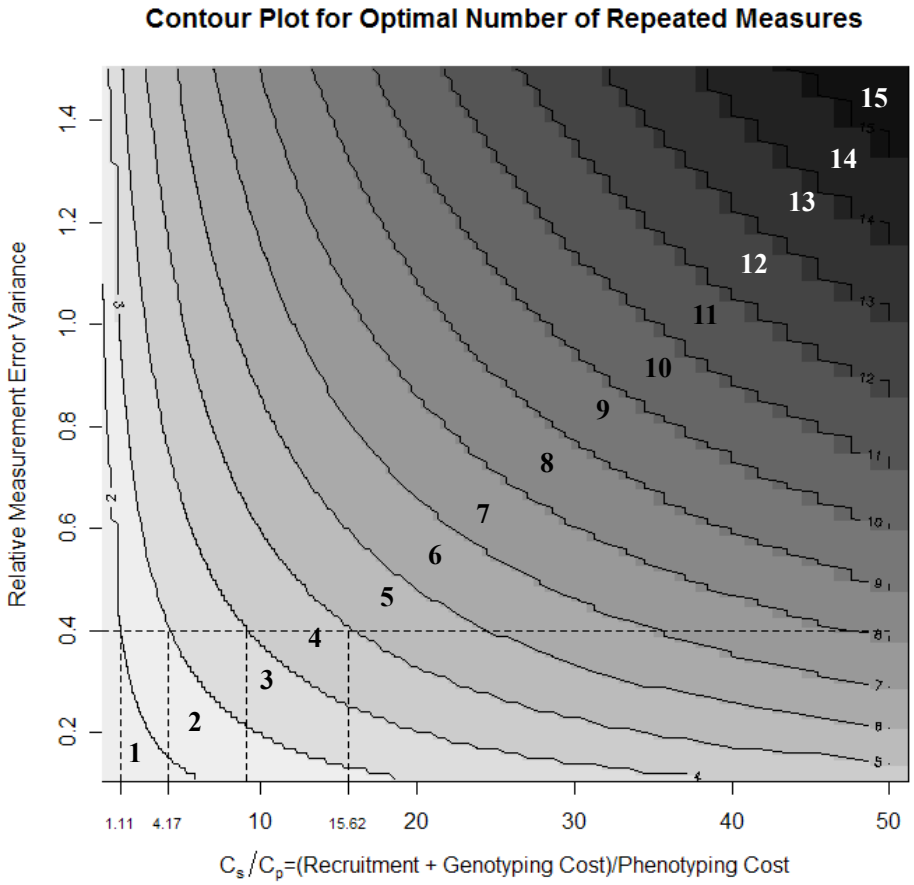
CR_{m_1,m_2} is defined in (5). When $C_s / C_p > CR_{m_1,m_2}$, taking m_1 measures is better than taking m_2 measures per subject. Heritability is defined as $(\sigma^2_{mg} + \sigma^2_{pg})/(\sigma^2_{mg} + \sigma^2_{pg} + \sigma^2_e)$

Figure 2.1 Expected LOD score for 1000 nuclear families with 4 offspring.



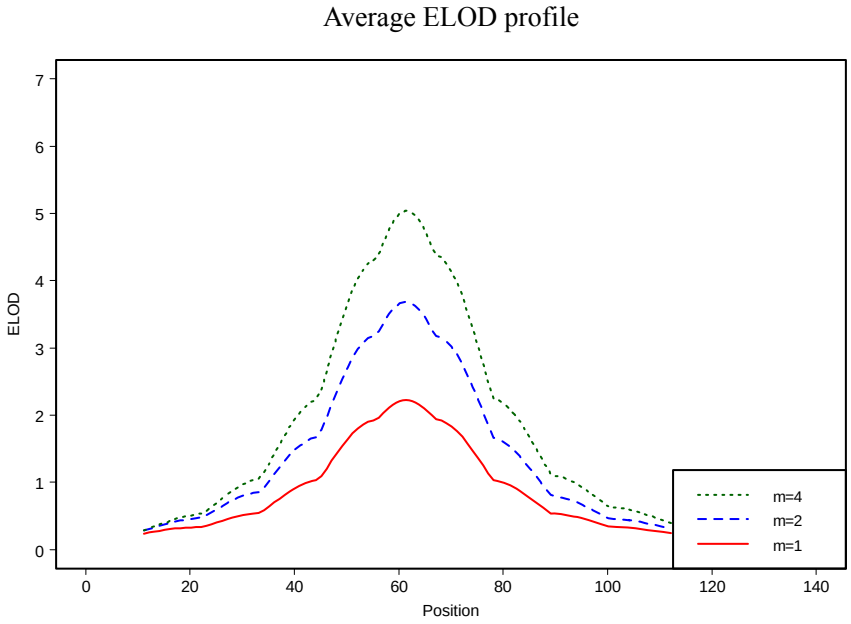
Where $\sigma^2_{mg} = 0.2$, $\sigma^2_{pg} = 0,...,0.8$, $\sigma^2_e = 0.8 - \sigma^2_{pg}$ and $\sigma^2_m = \sigma^2_{mg} + \sigma^2_{pg} + \sigma^2_e = 1$. m = the number of repeated measures.

Figure 2.2 Contour plot for optimal number of repeated measures



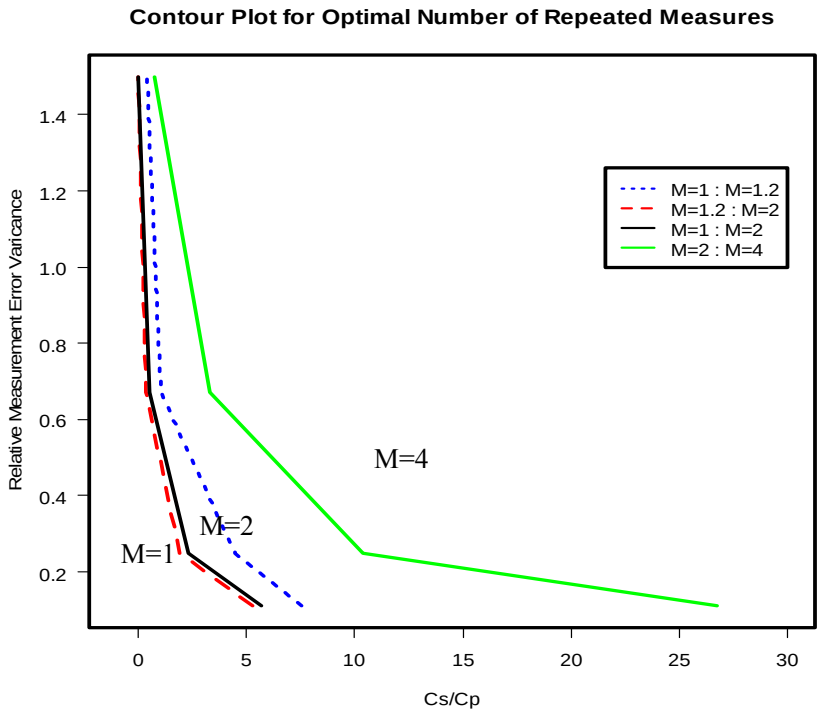
Cost ratio ranges from 0 to 50 and σ_m^2 ranges from 0.11 to 1.5 (10-60% of total trait variance). Trait variance excluding measurement error is fixed to 1 ($\sigma_{mg}^2 = 0.2$, $\sigma_{pg}^2 = 0.4$, $\sigma_e^2 = 0.4$). The numbers on the plot indicate the optimal number of repeated measures.

Figure 2.3 Average LOD score profile for balanced design simulation (scenario 2).



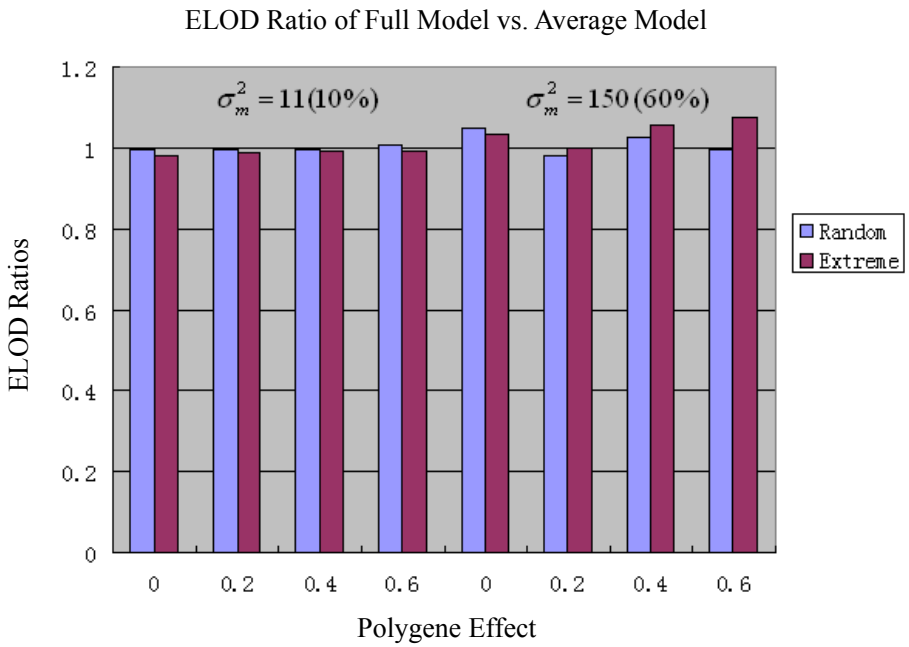
$\sigma_m^2 = 67$ (40% total variance), $\sigma_{pg}^2 = 40$ (24%). Results based on 500 simulation replications and plotted at every 1.0 Mb grid point.

Figure 2.4 Contour plot for optimal number of repeated measures



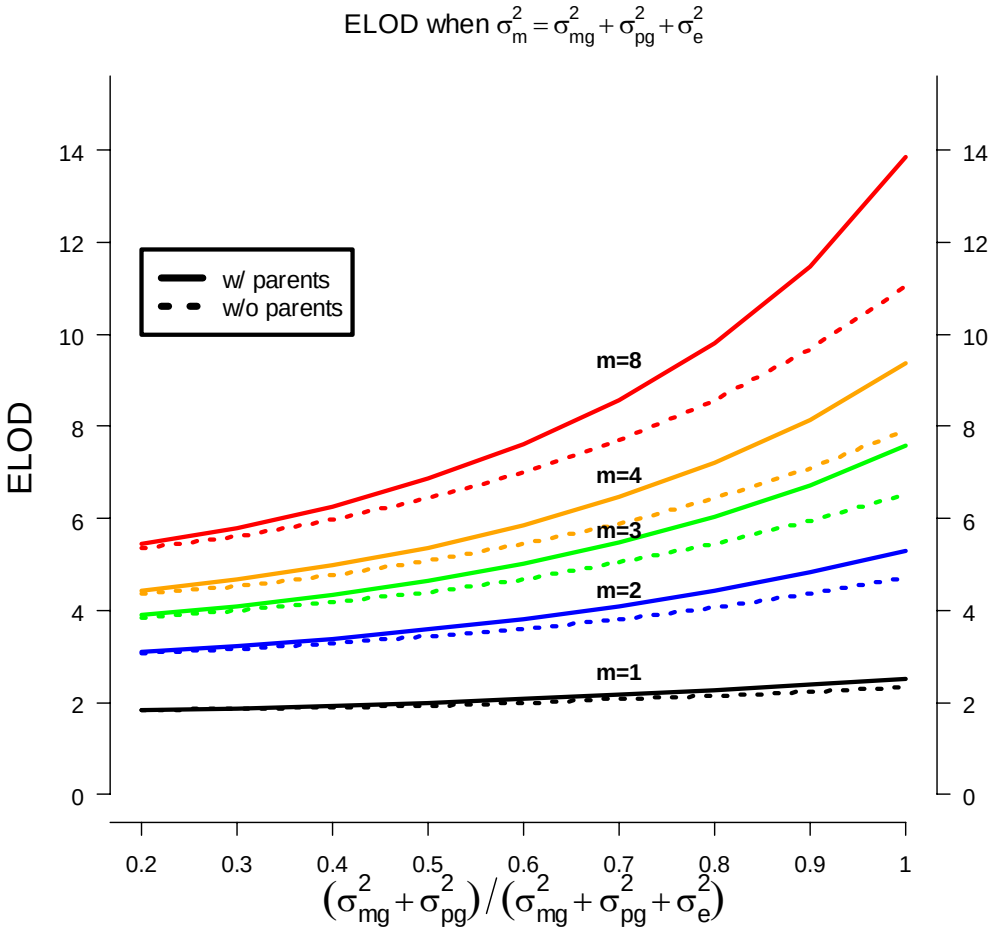
Cost ratio ranges from 0 to 30 and σ_m^2 ranges from 11 to 150 (10-60% total variance). Trait variance excluding measurement error is fixed to 100 ($\sigma_{mg}^2 = 20$, $\sigma_{pg}^2 = 40$, $\sigma_e^2 = 40$). This setting is equivalent to the setting in figure 2. Each line separates two regions in which one design is better than the other. For example, to the left of the (blue) dot line, balanced design m=1 is better than the unbalanced design m=1.2; on the right side of the line, the unbalanced design m=1.2 is better than balanced design m=1. Note that the (blue) dot line is to the right of the (red) dash line, thus balanced designs are superior to unbalanced designs in any situation. For region to the right of the grey (green) solid line, the optimal design is balanced design m=4; for region between the black solid line and the grey (green) solid lines, the optimal design is balanced design m=2; for region to the left of the black solid line, the optimal design is balanced design m=1.

Figure 2.5 ELOD ratio of full model vs. average model for unbalanced design.



Setting is scenario 2. Left 4 pairs of bars are for $\sigma_m^2 = 11$ (10% of total variance). Right 4 pairs of bars are for $\sigma_m^2 = 150$ (60% of total variance). Random design: the number of repeated measures follows an exponential distribution. Extreme design: 20% Subjects with extreme first measure have an additional measurement.

Figure 2.6 Expected LOD score for 1000 nuclear families with 4 offspring with and without using parental phenotypes



Where $\sigma^2_{mg} = 0.2$ (10% total variance), $\sigma^2_{pg} = 0,...,0.8$ (0-40% total variance), $\sigma^2_e = 0.8 - \sigma^2_{pg}$ and $\sigma^2_m = \sigma^2_{mg} + \sigma^2_{pg} + \sigma^2_e = 1$ (50% total variance). m = the number of repeated measures.

Table 2.4 ELOD ratios across different pedigree structures

	2sibs		4sibs		6sibs		cousin	
m	ELOD	ratio	ELOD	ratio	ELOD	ratio	ELOD	ratio
1	0.841		2.841		5.065		3.738	
2	1.367	1.626	4.824	1.698	8.530	1.684	6.029	1.613
4	1.841	2.190	6.566	2.311	11.847	2.339	8.119	2.172
8	2.206	2.624	7.977	2.808	14.274	2.818	9.731	2.603

The ratio is comparing the ELOD for m=2, 4, 8 with m=1. We simulated nuclear families with 2, 4, 6 offspring, and a family structure with 2 2nd-generation offspring and each has three 3rd generation offspring (the "cousin" scenario). The total number of individuals in all scenarios was set to the same so as to facilitate power comparison between different family structures. We fixed $\sigma^2_{mg} + \sigma^2_{pg} + \sigma^2_e = 100$, σ^2_{pg} = 40 (24% total variance), σ^2_m = 67 (40% total variance) and simulated a fully informative marker with σ^2_{mg} =20 (12% total variance).

Table 2.5 Power lost and Type I error when ignoring imbalance

	Power					Type I error	
	ELOD			% pvalue < 0.001		% pvalue < 0.05	
m	Correct average model	Ignoring imbalance	Ratio of ELOD	Correct average model	Ignoring imbalance	Correct average model	Ignoring imbalance
2	3.76	3.64	1.03	82.5%	80.6%	4.8%	4.8%
4	4.54	4.16	1.09	90.8%	87.5%	4.4%	4.7%
10	5.17	4.49	1.15	95.0%	90.5%	3.7%	4.6%

Half of the samples are randomly selected to take a specific number of repeated measures (m=2, 4 or 10), other samples will be measured only one time. Results are based on 2000 simulations. We fixed $\sigma^2_{mg} + \sigma^2_{pg} + \sigma^2_e = 100$, σ^2_{pg} = 40 (24% total variance) and σ^2_m = 67 (40% total variance). In the simulation for power, at a fully informative marker, σ^2_{mg} =20 (12% total variance); in the simulation for type I error, σ^2_{mg} =0.

2.8 References

1. Boomsma DI, Dolan CV: A comparison of power to detect a QTL in sib-pair data using multivariate phenotypes, mean phenotypes, and factor scores. Behav Genet 1998; 28:329-40.
2. Levy D, DeStefano AL, Larson MG, O'Donnell CJ, Lifton RP, Gavras H, Cupples LA, Myers RH: Evidence for a gene influencing blood pressure on chromosome 17. Genome scan linkage results for longitudinal blood pressure phenotypes in subjects from the Framingham heart study. Hypertension 2000; 36:477-83.
3. de Andrade M, Gueguen R, Visvikis S, Sass C, Siest G, Amos CI: Extension of variance components approach to incorporate temporal trends and longitudinal pedigree data analysis. Genet Epidemiol 2002; 22:221-32.
4. Amos CI: Robust variance-components approach for assessing genetic linkage in pedigree. Am J Hum Genet 1994; 54:535-43.
5. Almasy L, Blangero J: Multipoint quantitative trait linkage analysis in general pedigree. Am J Hum Genet 1998; 62: 1198-211.
6. Gauderman WJ, Macgregor S, Briollais L, Scurrah K, Tobin M, Park T, Wang D, Rao S, John S, Bull S: Longitudinal data analysis in pedigree analysis. Genet Epidemiol 2003; 25:S18-S28.
7. Schmitz S, Cherny SS, Fulker DW: Increase in power through multivariate analysis. Behav Genet 1998; 28:357-63.
8. Boomsma, DI, Molenaar, PCM: The genetic analysis of repeated measures. I. Simplex models. Behav Genet 1987; 17, 111-23.
9. Abecasis, GR, Cherny SS, Cookson WO, Cardon LR: Merlin: rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 2002; 30: 97-101.
10. Lange K, Westlake J, Spence MA: Extensions to pedigree analysis. III. Variance components by the scoring method. Ann Hum Genet 1976; 39:485-91.
11. Jacquard A. Genetic information given by a relative. Biometrics 1972; 28:1101-14.
12. Maxwell SE, Delaney HD: *Designing Experiments and Analyzing Data: A Model Comparison Perspective*, ed 2. Mahwah, NJ, Lawrence Erlbaum, 2003, pp 525-572.
13. Wald A: Tests of statistical hypotheses concerning several parameters when the number of observations is large. Trans Amer Math Soc 1943; 54:426-82.
14. Rijsdijk FV, Hewitt JK, Sham PC: Analytic power calculation for QTL linkage analysis of small pedigrees. Eur J Hum Genet. 2001; 9:335-40.
15. Chen WM, Abecasis GR: Estimating the power of variance component linkage analysis in large pedigrees. Genet Epidemiol 2006; 30:471-84.
16. Chen, WM, Broman, KW, and Liang, KY: Quantitative trait linkage analysis by generalized estimating equations: unification of variance components and Haseman-Elston regression. Genet Epidemiol 2004; 26:265-72.
17. Bhattacharjee S, Kuo CL, Mukhopadhyay N, Brock GN, Weeks DE, Feingold E: Robust score statistics for QTL linkage analysis. Am J Hum Genet 2008; 82:1-16.
18. Sham PC, Purcell S, Cherny SS and Abecasis GR: Powerful regression-based quantitative-trait linkage analysis of general pedigrees. Am J Hum Genet 2002; 71:238-53.
19. Drigalenko E: How sib pairs reveal linkage. Am J Hum Genet 1998; 63:1242-5.
20. Nash MW, Huezo-Diaz P, Williamson RJ, Sterne A, Purcell S, Hoda F, Cherny SS,

Abecasis GR, Prince M, Gray JA, Ball D, Asherson P, Mann A, Goldberg D, McGuffin P, Farmer A, Plomin R, Craig IW and Sham PC: Genome-wide linkage analysis of a composite index of neuroticism and mood-related scales in extreme selected sibships. Hum Mol Genet 2004; 13:2173-82.

21. Sham PC, Cherny SS, Purcell S, Hewitt JK: Power of Linkage versus Association Analysis of Quantitative Traits, by Use of Variance-Components Models, for Sibship Data. Am J Hum Genet 2000; 66:1616–1630.
22. Bauman LE, Almasy L, Blangero J, Duggirala R, Sinsheimer JS, Lange K. Fishing for pleiotropic QTLs in a polygenic sea. Ann Hum Genet 2005; 69:590-611.
23. Abecasis GR and Wigginton JE. Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers. Am J Hum Genet 2005; 77:754-67.
24. Fullerton J., et al. Linkage Analysis of Extremely Discordant and Concordant Sibling Pairs Identifies Quantitative-Trait Loci That Influence Variation in the Human Personality Trait Neuroticism. Am J of Hum Genet 2003; 72:879-890.
25. Self SG, Liang KY. Large sample properties of the maximum likelihood estimator and the likelihood ratio test on the boundary of the parameter space. Journal of the American Statistical Association 1987; 82:605-611.

Chapter III

DISCRETE GENERATION FRAMEWORK FOR COALESCENT SIMULATION OF GENOME-WIDE SCALE DATA

3.1 Abstract

Summary: We developed a rapid coalescent-based framework to simulate whole genome data. The proposed simulator, called GENOME, can simulate sequences that follow the Wright-Fisher model and from a region of more than 100Mb long, which is not practical for standard coalescent approach. In addition to features of standard coalescent simulators, the program allows for recombination rates to vary along the genome and for flexible population histories. Within small regions, we have evaluated samples simulated by GENOME to verify that GENOME provides the expected LD patterns and frequency spectra. The program can be used to study the sampling properties of any statistic for a whole genome study.

Availability: The program and C++ source code are available online at: http://www.sph.umich.edu/csg/liang/genome/

3.2 Introduction

The coalescent approach (Kingman 1982, Hudson 1983 & 1990, Donnelly and Tavaré 1995) is an efficient way to sample of sequences from a theoretical population that

follows the Wright-Fisher neutral model (Ewens 1979). Simulations based on coalescent models have also been used to study the sampling properties of interesting statistics or evaluate new methods. Applications include the inference of population history (Weiss & von Haeseler 1998), the study of positive selection (Przeworski 2002, Voight et al. 2006) and whole genome linkage disequilibrium mapping of common disease genes (Kruglyak 1999, Zöllner & von Haeseler 2000). Existing software packages, such as ms (Hudson 2002) and cosi (Schaffner et al. 2005), implement the standard coalescent approach which simulates genealogical events backward in time. Simulated events typically include the coalescence of two sequences into a single ancestral lineage, recombination within a sequence, or migration between populations. Since all these events are typically rare, coalescent simulators assume that they never occur simultaneously and assume many generations pass between consecutive events. Time between events is explicitly modeled and used to skip over generations with no genealogical events of interest. The algorithm proceeds until all sequences coalesce to their most recent common ancestor and the resulting genealogy is used to place mutation events along the various sequences.

The standard approach is extremely efficient when simulating short sequences. As sequences get longer, many more coalescent, recombination and migration events occur and the time intervals between them diminish. For longer sequences and large sample sizes, little computational efficiency is gained by skipping over uninteresting generations and substantial computational effort is expended tracking recombination events and their positions, and allocating memory to track the many ancestral fragments of each sequence as they repeatedly recombine and coalesce with each other. Overall, the standard

coalescent approach which is suitable for short genomic segments (<2-3Mb) becomes very slow for larger regions (>100Mb).

As genome-wide studies become a reality, efficient tools for simulating large sequences are essential to study the sampling properties of arbitrary statistics that might be evaluated on a genome-wide association study and to compare the performance of different methods that may be applied to genome-wide scale data. For example, in an ongoing study we are evaluating the distribution of stretches of haplotype shared among a majority of individuals with disease and need an efficient coalescent framework to evaluate the null distribution of the statistic. There is great interest in developing fast coalescent simulators to address this and similar problems. One potential speedup involves making further simplifying assumptions about the genealogy (Marjoram and Wall 2006). Here, we propose an alternative framework for the coalescent that allows efficient simulation of genealogies for long sequences and still fully captures the complexity of the genealogy. In our approach, the genealogy of sampled sequences is simulated backwards in time, one generation at a time, in a procedure that is computationally efficient and removes the bifurcate tree approximation (in the standard approach, each coalescence event involves exactly two sequences that coalesce to a common ancestor but, using our approach, multiple sequences can coalesce to a common ancestor simultaneously). When multiple sub-populations are simulated, the program allows for migration among subpopulations, and for user specified demographic events such as population bottlenecks and expansions or population merges and splits. We allow recombination rates to vary so as to mimic the pattern of hotspots along the genome. As in the standard coalescent approach, mutations are simulated assuming an infinite-sites

model.

3.3 Methods

As in the standard coalescent approach, we simulate the genealogy of a sample of sequences, conditional on parameters such as the population size, the recombination rate, and rates of migration between subpopulations. Instead of simulating the time to next event, we simulate the coalescent and recombination events at every generation proceeding backwards in time (Figure 3.1). For each generation, the sequences are stored in a sparse matrix where rows correspond to individuals and columns correspond to short stretches of sequence. The matrix is sparse because only portions of sequence with a descendant in the final generation are tracked. We allocate two sparse matrices in memory (the current and the previous generation, which are reused) together with a separate structure summarizing coalescent events for each portion of the sequence. To allow for population stratification or other constraints on mating, we define a set of rules that can be used to relate each individual sequence (a row in one of the sparse matrices) to its ancestors in the previous generation (one or more rows in the second sparse matrix). Since we simulate all intervening generations, these rules can be quite sophisticated – to enforce multiple populations, geographic proximity between subpopulations, diploid individuals (so that each sequence has exactly two ancestors), etc. These features are commonly only found in forward simulators, which are computationally much less efficient.

Because our approach can simulate multiple coalescent and recombination events in the same generation, it naturally accommodates situations where the number of sequences

sampled approximates the effective population size or where the sequences are very long. Conditional on the genealogical tree, mutations are placed on the branches. The number of mutations on each branch follows a Poisson distribution with mean equal to the product of the mutation rate and the branch length. The infinite-site mutation model is assumed. As with many other coalescent simulators, we also allow the number of mutations to be fixed so that the probability that a mutation occurs on a particular branch is proportional to its length. Varying recombination rate and population histories can also be specified by parameter files. The output distinguishes ancestral state and derived alleles and is similar to the output format of the ms program. The genealogy trees for each fragment in Newick tree format (see web link in reference) can be output, ready for plotting with PHYLIP (Felsenstein 2005) or for use with seq-gen, a sequence evolution program by Rambaut and Grassly (1997). Detailed instructions and examples are available on our website (http://www.sph.umich.edu/csg/liang/genome/).

3.4 Results when standard coalescent approach can apply

To evaluate our simulator, we first compared the generated allele frequency spectra with theoretical expectations. Using a goodness of fit test, we observed no significant differences between the expected spectra and those generated by GENOME (Figure 3.2). We have compared our simulated samples with those generated by Hudson's ms (simulating a 2Mb region). The two simulators provide similar LD patterns and frequency spectra (Figure 3.3-3.9). When simulating long regions, GENOME is substantially faster than ms (Table 3.1). For example, when simulating a sample of 1200 chromosomes, each 150 Mb long, from two populations of size 10,000, GENOME requires ~66 minutes,

compared to >12 hours for Hudson's ms (using a standard 2.8 GHz Pentium CPU). The scaled rates of mutation, recombination and migration were set to $4N\mu=60000$, $4Nr=60000$ and $4Nm=10$ in the simulation described. As expected, GENOME also outperforms cosi in runtime, a coalescent simulator similar to ms but allows for flexible recombination rates and is somewhat slower than ms (Table 3.2).

GENOME is written in C++ and is portable to a variety of operating systems, including Windows, Linux and MacOS. The Mersenne Twister Code (Matsumoto & Nishimura 1998) is used as the source of random deviates. In addition to a stand-alone version, our simulator is also provided as a C++ function "genome()", that can be incorporated as a module in other programs.

3.5 Difference between the proposed method and standard coalescent approach

In summary, our method differs from the standard coalescent model in the following aspects: (1) our method simulates every generation instead of skipping generations that do not have coalescent, recombination or migration events. The time to an event is integer instead of continuous in standard coalescent approach. (2) Our method does not assume population size to be much larger than sample size. In fact, sample size can be as large as the population size. (3) Our method allows multiple events (coalescent, recombination and migration) to occur in the same generation and on the same sequence so that the bifurcate tree approximation is not needed. The above features ensure that our proposed framework can be used to simulate the exact Wright-Fisher model and possibly incorporate useful features that otherwise are only available in forward simulators.

When sample size is close to the population size, many coalescent events will occur

at the first few generations. Allowing for general genealogy and discrete generation increases the possibility of singleton mutations. To assess this effect, we carried out simulations for two settings: (1) simulate 1,000 sequences from a population of 20,000. Each sequence consists of 2,000 independent loci each of 1kb long. Mutation rate is 10^{-8} per base pair per generation, (2) simulate 1,000 sequences from a population of 1,000. Each sequence consists of 40,000 independent loci each of 1kb long with the same mutation rate in (1). To assess the effect on mutations, we did not simulate recombination in both settings. The standard coalescent approach could not distinguish the two settings because they have the same total scaled mutation rate. We simulated settings 1 and 2 using GENOME and Hudson's ms program. Under the coalescent theory, we expected to see that 13.4% (sd=0.44% for 6006 SNPs) of polymorphisms are singleton (Hudson 1990). At setting 1, GENOME produced 13.3% singletons out of 5949 SNPs and the ms program produced 12.3% singletons out of 5988 SNPs. At setting 2, the ms program produced 13.1% singletons out of 5898 SNPs but GENOME produced 15.2% singletons out of 6006 SNPs, which is significantly higher than expected ($p=2.1*10^{-5}$ for one side test of excess singletons). The 13% increase in the number of singleton from GENOME is similar to the observation of exact coalescent developed in Fu 2006. For more common SNPs, the difference is smaller. For example, under the standard coalescent theory, we expected to see 37.8% (sd=0.63% for 6006 SNPs) SNPs with MAF<1%. GENOME at setting 2 produced 39.1% SNPs in this category but the difference is marginal significant ($p=0.02$ for one side test of excess rare SNPs). As expected, the fraction of SNPs with MAF<1% are 37.4% for GENOME at setting 1, 36.7% for ms at setting 1 and 36.5% for ms at setting 2.

GENOME separates the whole sequence into small segments and does not simulate recombination event within a segment. It would be interesting to see if this feature has an effect on the linkage disequilibrium (LD) pattern in small regions. We simulated 500 sequences of a 2Mb region from a population of 20,000 sequences. Each sequence consists of 20,000 segments of 100 bps. Under the coalescent theory, this setting will generate 5,432 SNPs on average. About 7% of the segments are expected to have more than two SNPs in the same segment. We simulated 100 datasets using GENOME and ms, respectively, and calculated the average R-square (Δ^2) by distance which is defined as the number of intervening SNPs. The absolute difference of R-square by distance from GENOME and ms were plotted in figure 3.10. When it is ~50 SNPs away, GENOME and ms produce similar R-square. Within the ~50 SNPs window, however, GENOME seems to produces smaller R-square in absolute difference but not in relative difference (figure 3.11). Note that the absolute difference is well within the 95% confidence limit (figure 3.10). A much larger scale of simulation is required to distinguish the subtle difference in LD pattern but overall GENOME agrees with standard coalescent approach very well when conditions for Kingman coalescent applied. When sample size is close to population size, GENOME also provides the opportunities to assess the effect on LD pattern and migrations between different populations.

3.6 Tables and figures

Table 3.1 Run time comparison of GENOME and Hudson's ms

Length of Region	Number of Populations	GENOME (seconds)	ms (seconds)
150Mb	1	1556 (25.9 mins)	13416 (3.7 hrs)
150Mb	2	3964 (66.1 mins)	45138 (12.5 hrs)

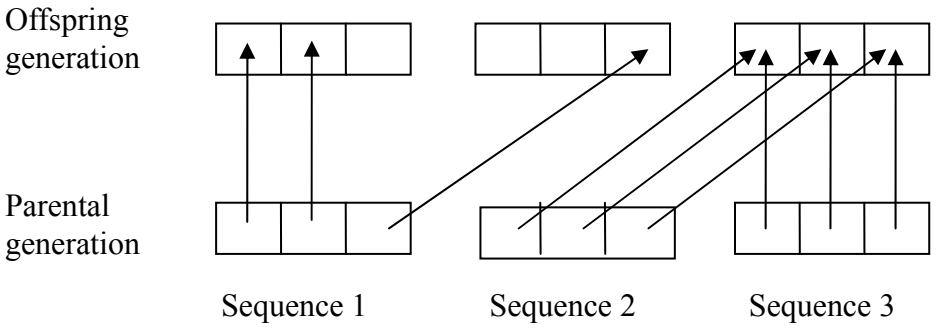
Settings:
Effective population size, N=10000 diploid individuals (for 2 pops, N=the size of each subpopulation)
1200 chromosomes (600/600 for 2 populations)
15000 fragments for 150Mb region.
Migration rate = $2.5*10^{-4}$ per generation, (4Nm=10)
Mutation rate = 10^{-8} per base pair (4Nu=60000 for 150 Mb)
Recombination rate =10^{-8} per base pair (4Nr=60000 for 150Mb)

Table 3.2 Run time comparison of GENOME and COSI

	COSI v1.1 (seconds)	**GENOME (seconds)**
40M	3089	294
30M	1658	240
20M	677	150

We simulate 300 sequences of 20Mb, 30Mb or 40Mb using COSI v1.1 and GENOME. COSI crashes when simulating 50Mb and 300 sequences or for longer region and more sequences. Observed that COSI took about 2.8G memory when simulating 30Mb region.

Figure 3.1 Discrete generation implementation



In the example, each sequence is divided into three fragments. There is a recombination event between the 2nd and the 3rd fragments in sequence 1. Sequences 2 and 3 coalesce.

Figure 3.2 Allele frequency spectra generated by GENOME compared with theoretical expectations

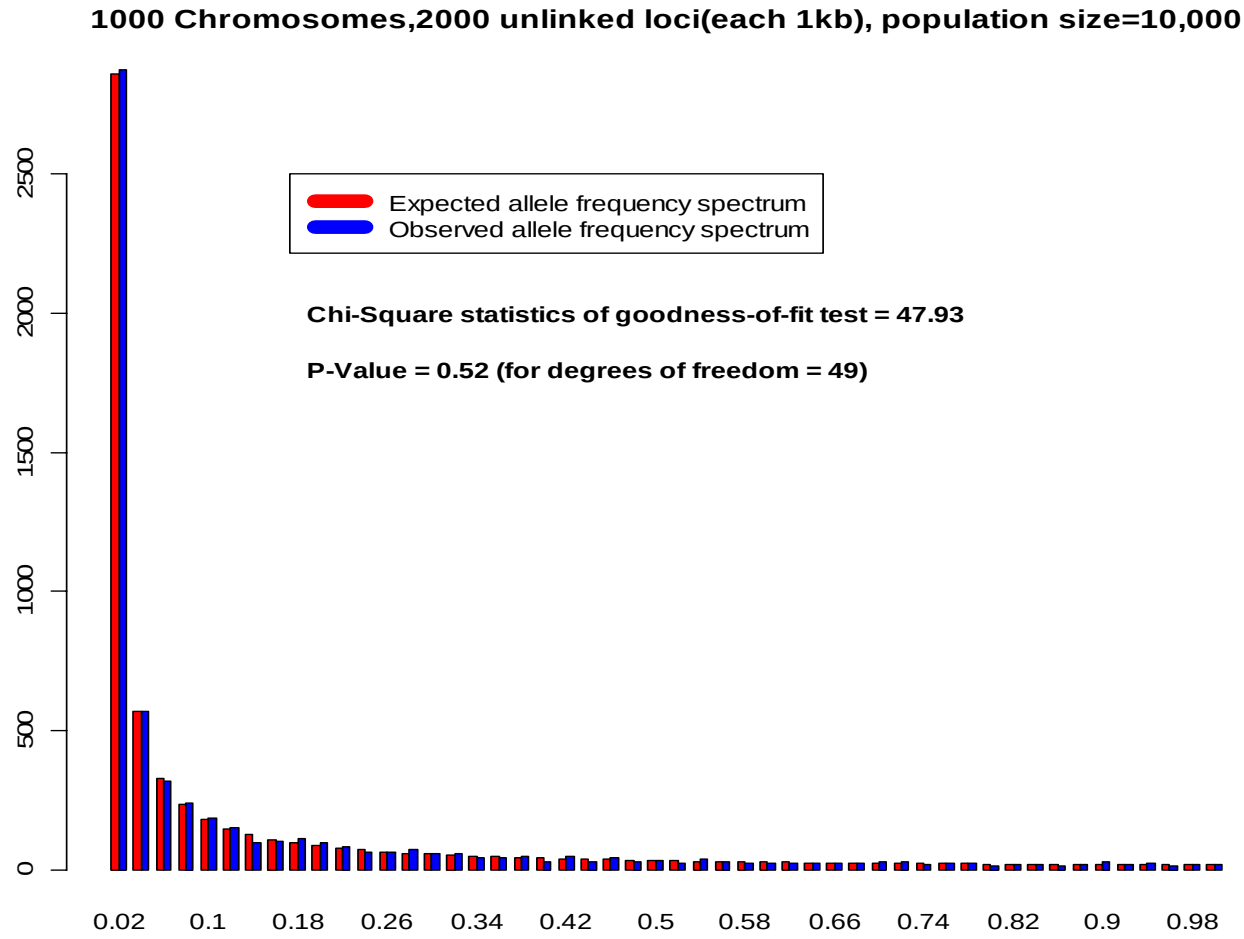
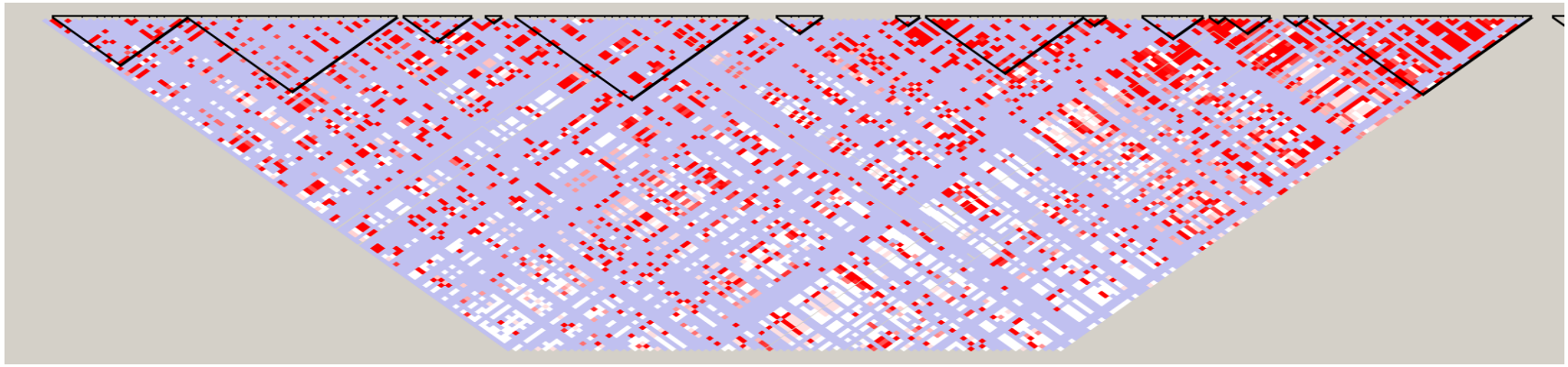
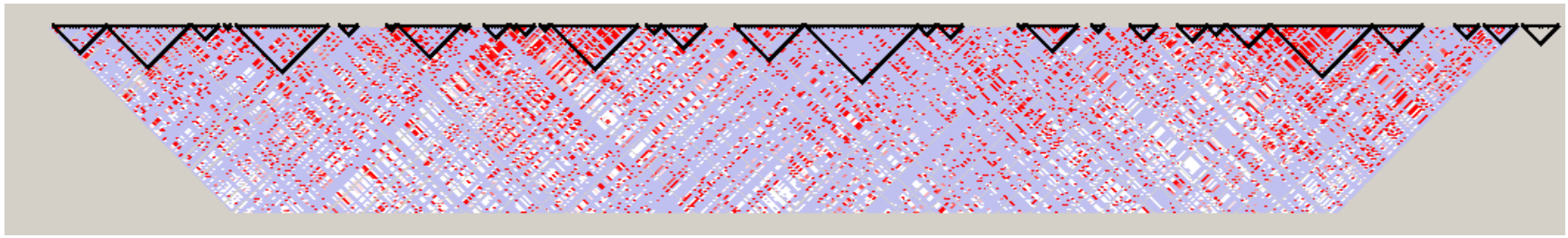


Figure 3.3 Haploview Plot for a 2Mb region simulated by GENOME.

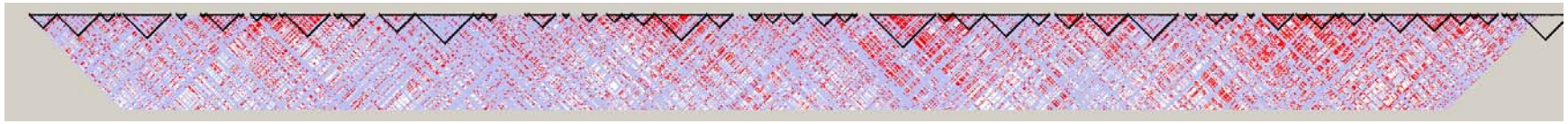
Marker1-200



Marker 1-500



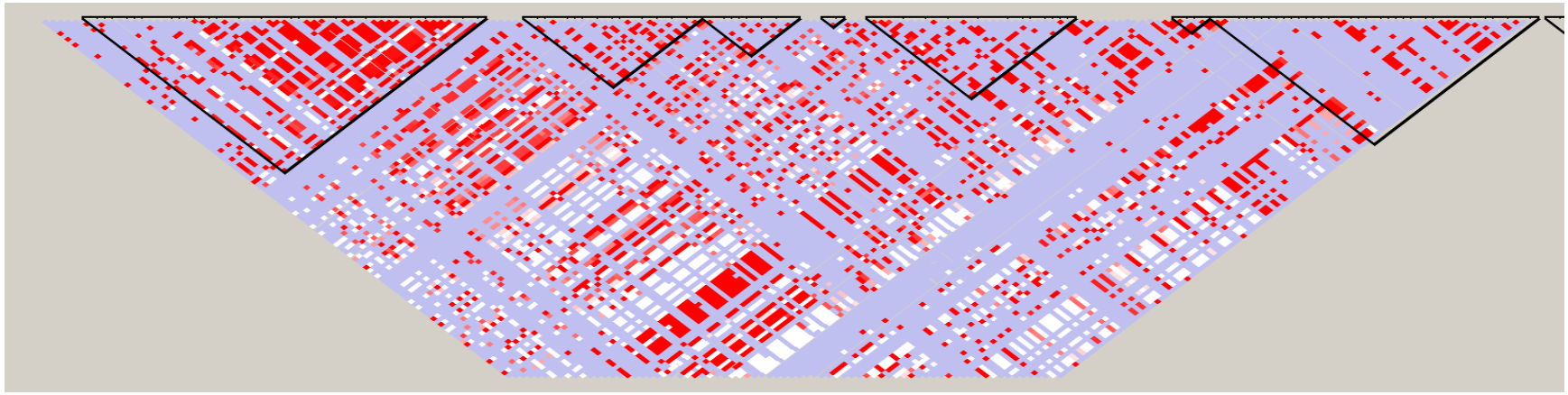
Marker 1-1000



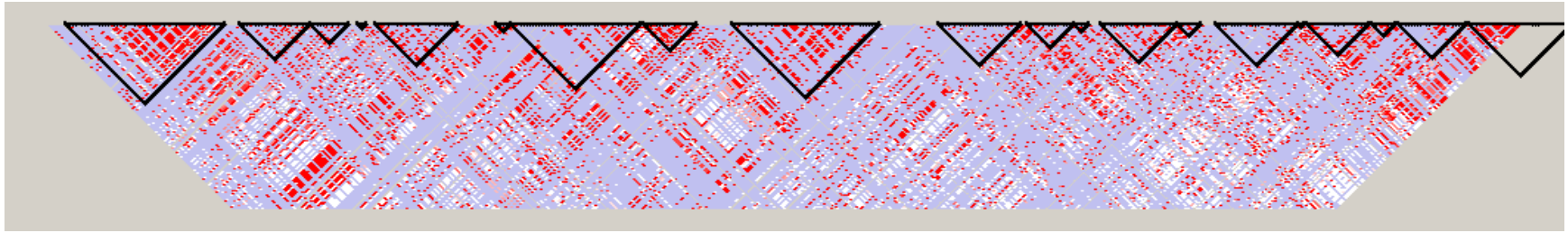
N=10000 diploid individuals, n=200, nPOP=1, fragment=20000, length=100, numChr=1, #SNP=Poisson (result=4943), rec=1e-6, mut=1e-8

Figure 3.4 Haploview Plot of a 2Mb region simulated by Hudson's ms

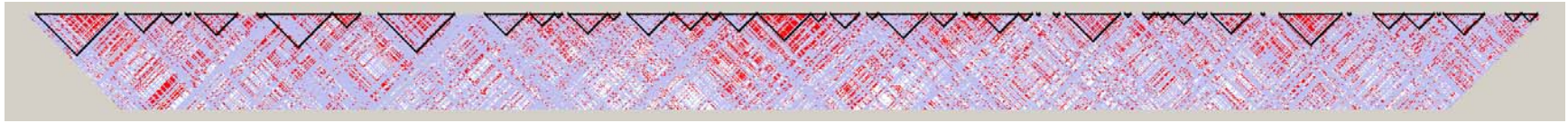
Marker1-200



Marker 1-500



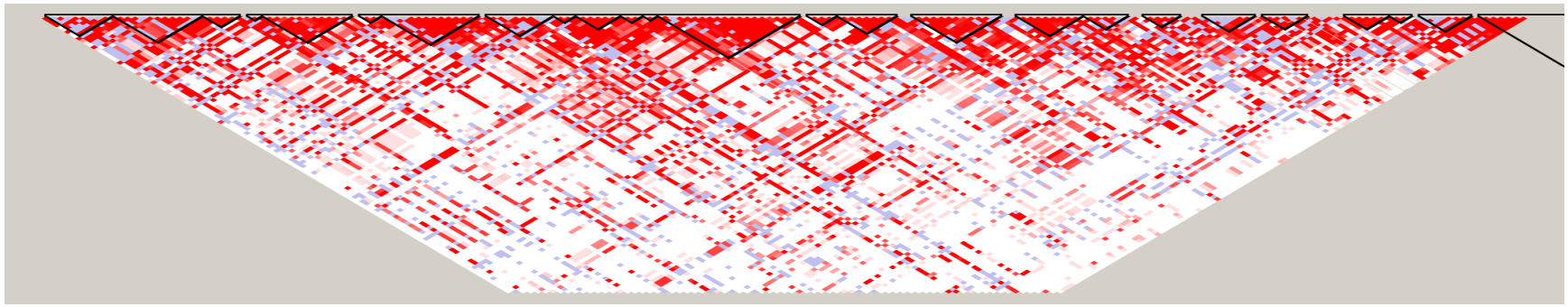
Marker 1-1000



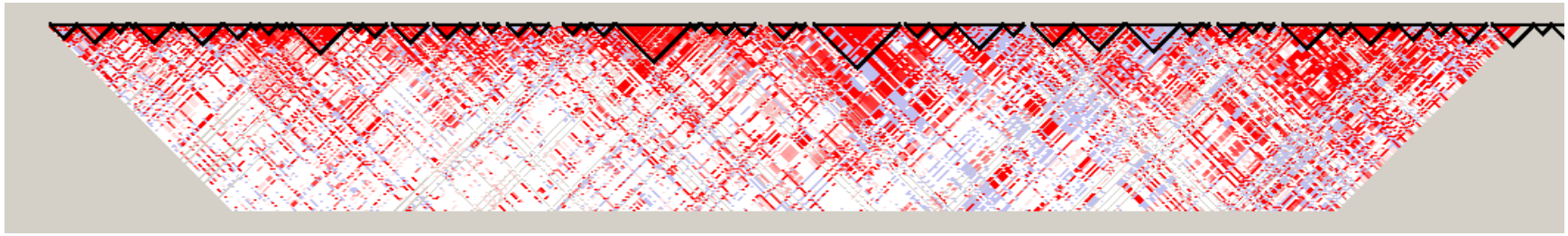
Command line: ms 200 1 -t 800 -r 800 20000, 4Nr=800, 4Nu=800, fragment=20000, result=4658 SNPs (setting equivalent to figure 3.3)

Figure 3.5 Haploview Plot of a 2Mb region (SNPs with MAF > 0.05, 2597 common SNPs) generated by GENOME

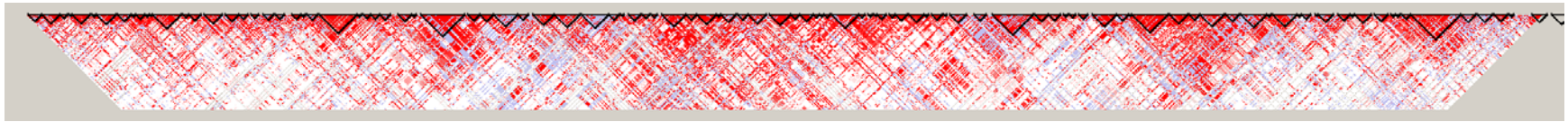
Marker1-200



Marker 1-500



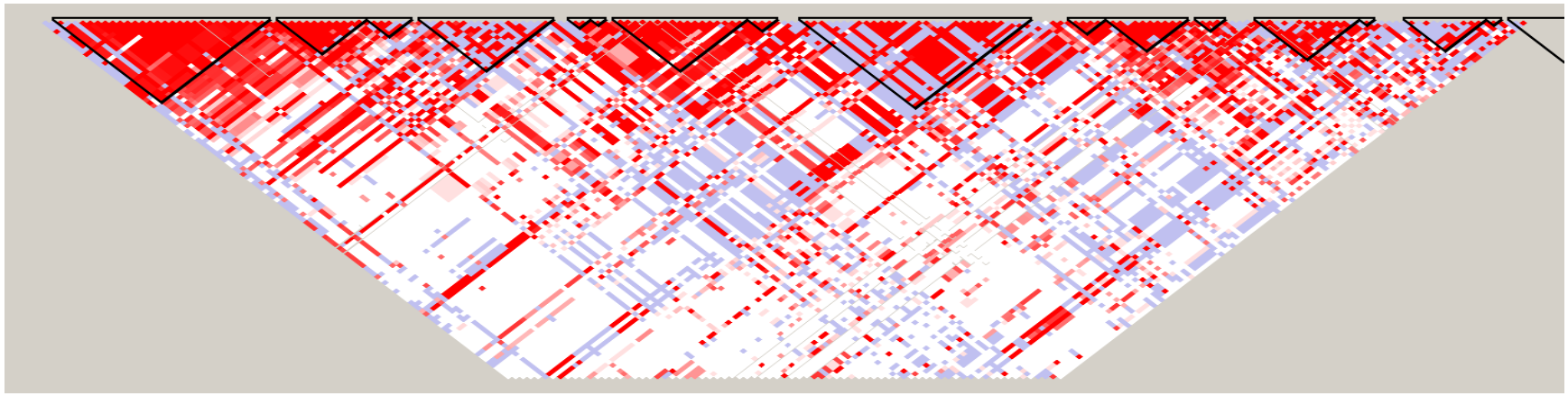
Marker 1-1000



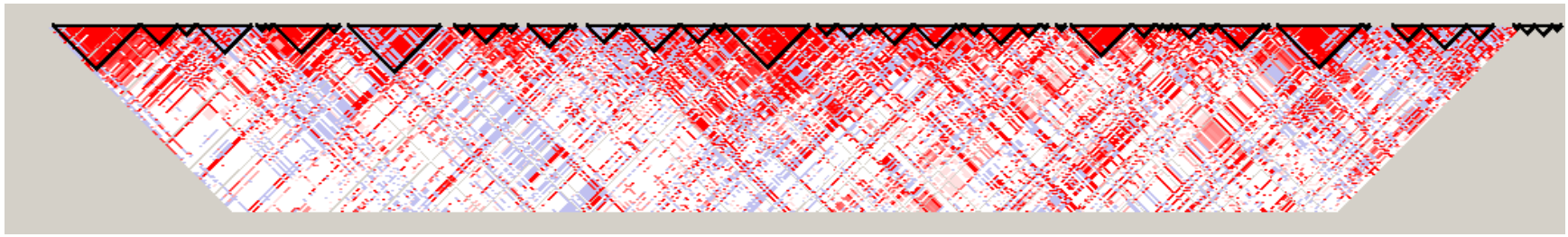
N=10000 diploid individuals, n=200, nPOP=1, fragment=20000, length=100, numChr=1, #SNP=Poisson (result=4943), rec=1e-6, mut=1e-8

Figure 3.6 Haploview Plot of a 2Mb region (SNPs with MAF > 0.05, 2426 common SNPs) simulated by Hudson's ms

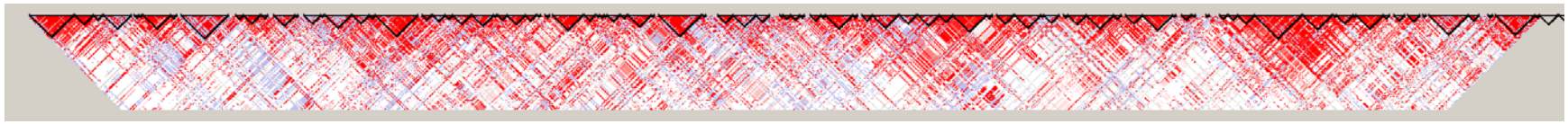
Marker1-200



Marker 1-500



Marker 1-1000



Command line: ms 200 1 -t 800 -r 800 20000, 4Nr=800, 4Nu=800, fragment=20000, result=4658 SNPs

Figure 3.7 Allele frequency spectra generated by GENOME and Hudson's ms with equivalent settings for a 2Mb region

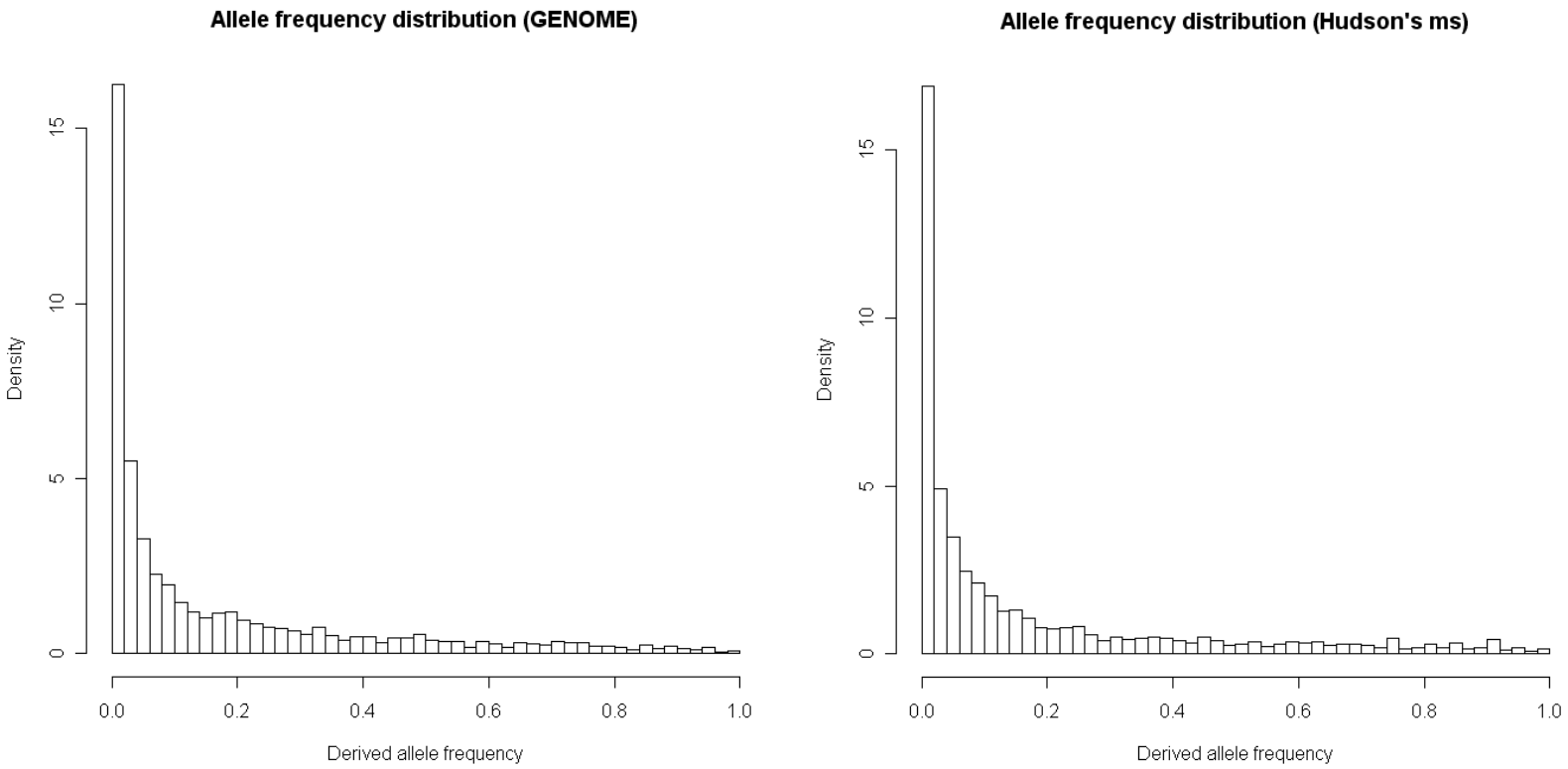
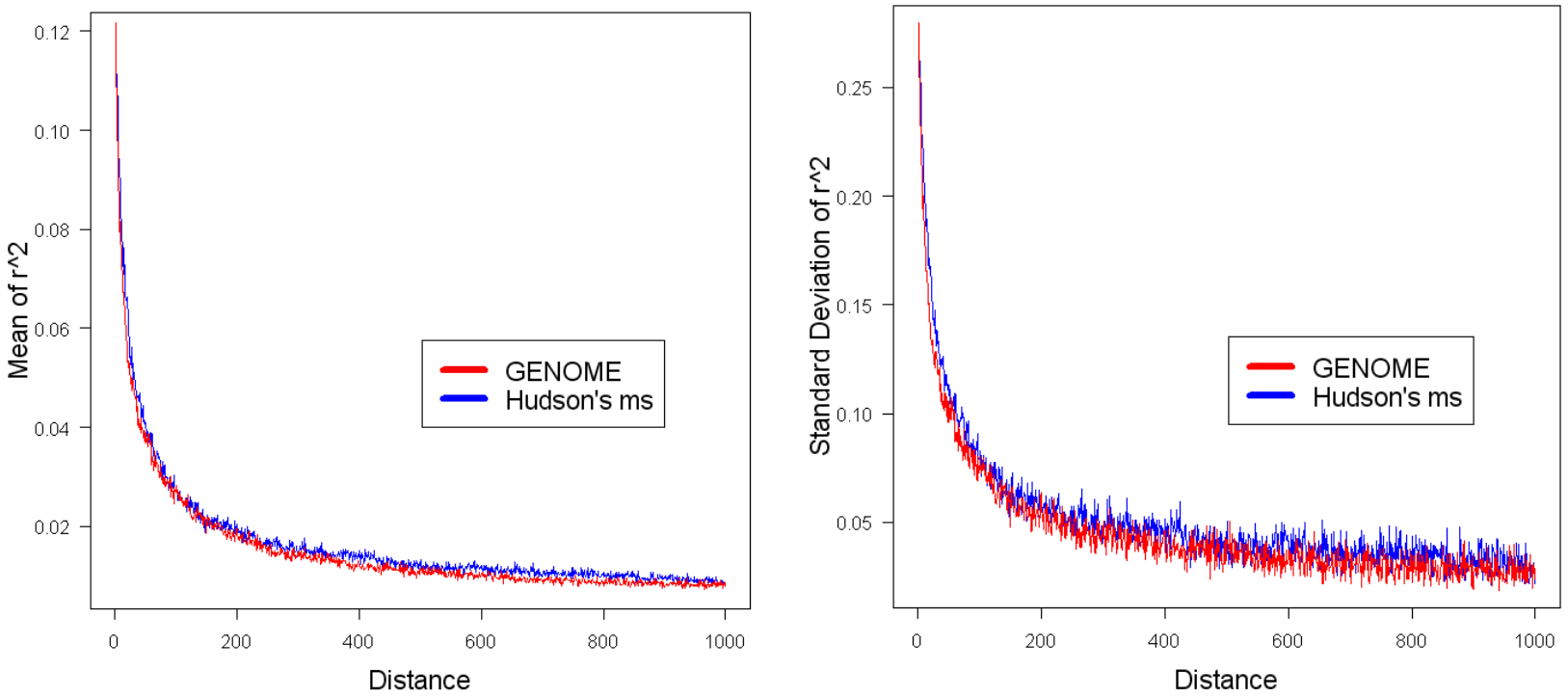
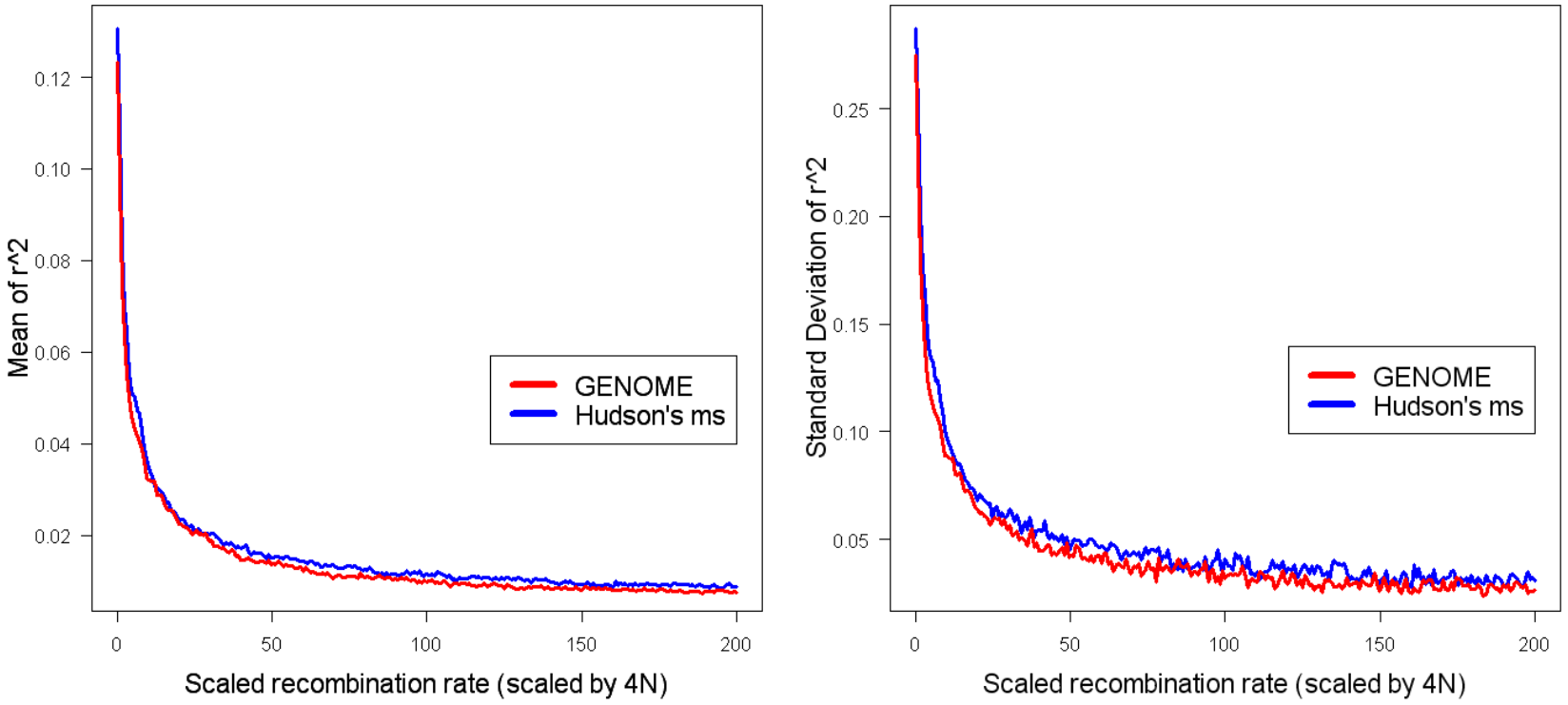


Figure 3.8 Distribution of LD by physical distance generated by GENOME and Hudson's ms



Equivalent settings for a 2Mb region. Physical distance is defined as the number of intervening SNPs.

Figure 3.9 Distribution of LD by genetic distance generated by GENOME and Hudson's ms



Equivalent settings for a 2 Mb region.

Figure 3.10 Difference in LD by distance simulated by GENOME and ms

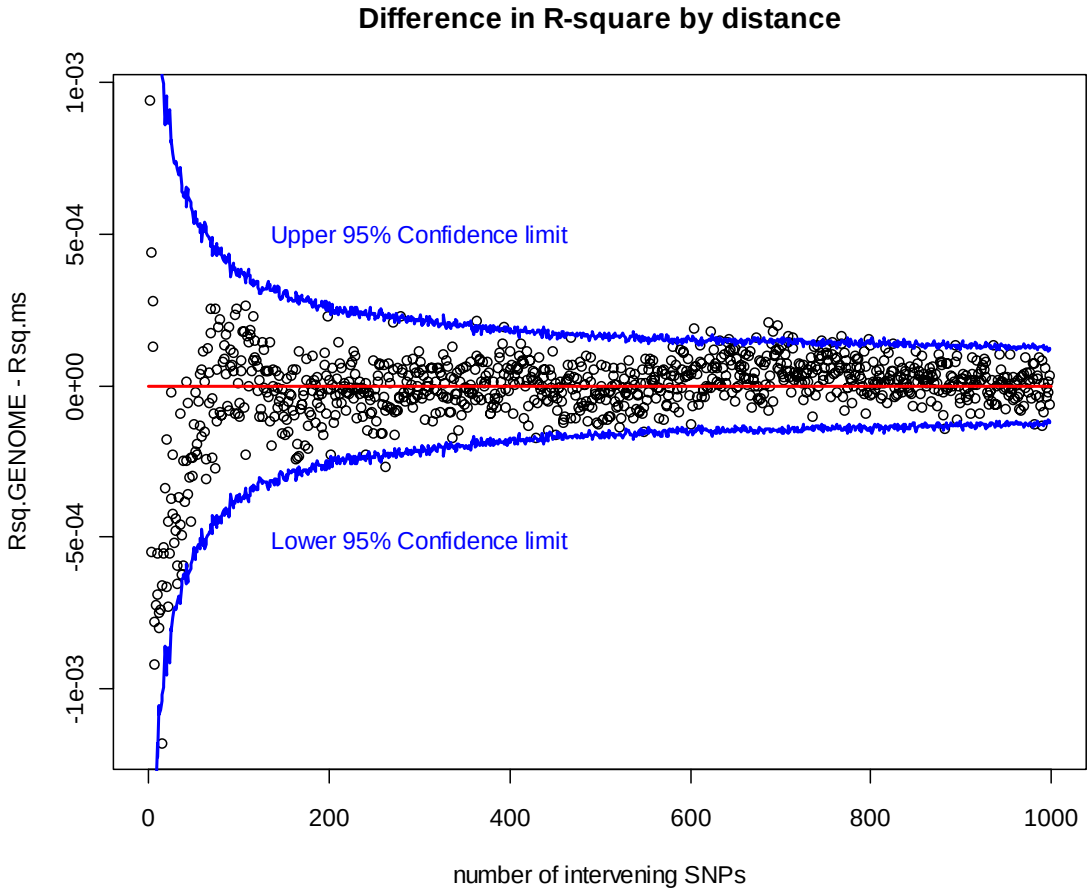
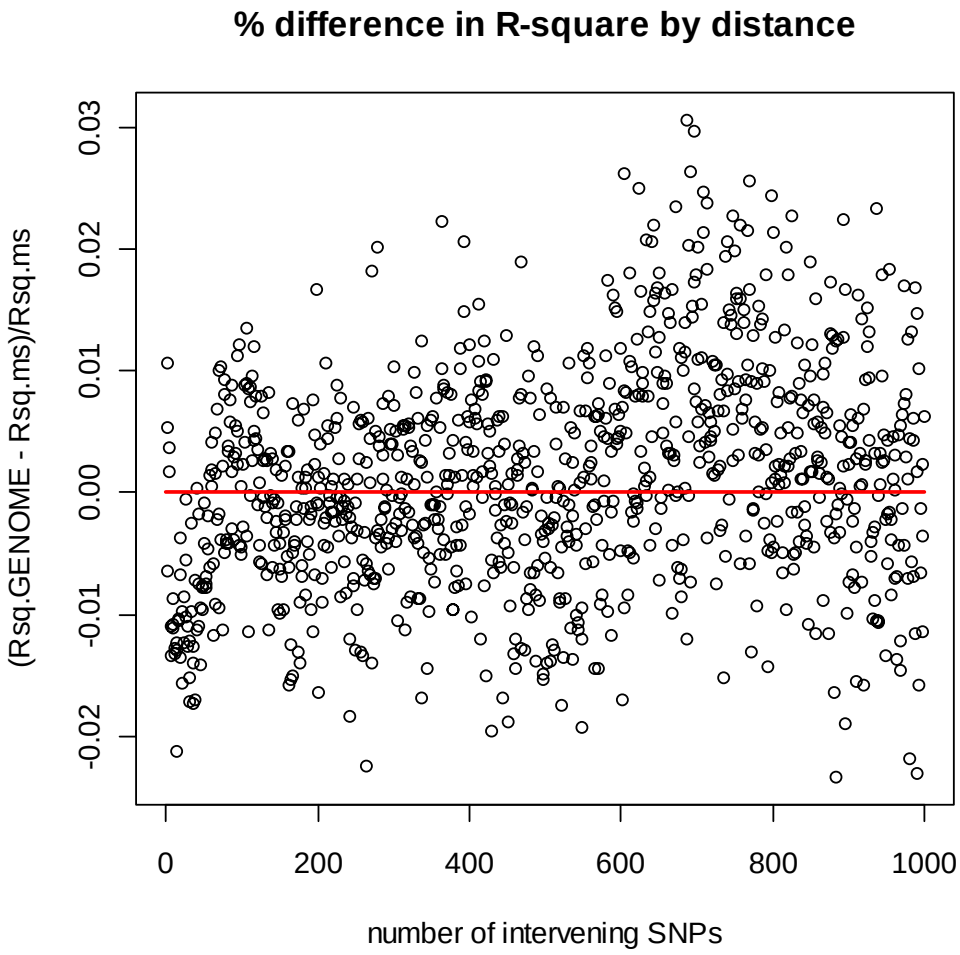


Figure 3.11 Relative differences in LD by distance simulated by GENOME and ms



3.7 References

Donnelly P., Tavaré S. (1995) Coalescents and genealogical structure under neutrality. *Annu. Rev. Genet.* 29: 401-421

Ewens WJ (1979), *Mathematical Population Genetics*. Springer, Berlin

Felsenstein J. (2005) PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle

Fu Y.X. 2006. Exact coalescent for the Wright-Fisher model. Theoretical Population Biology 69: 385-394.

Hudson R.R. (1983) Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, 23:183–201

Hudson R.R. (1990) Gene genealogies and the coalescent process, *Oxford Surveys in Evolutionary Biology* Vol 7: 1-4

Hudson R.R. (2002) Generating samples under a Wright-Fisher neutral model. *Bioinformatics* 18:337-378

Kingman J.F.C. (1982) The coalescent. *Stochastic Process. Appl*. 13:235-248

Kruglyak L. (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat Genet*. 22(2):139-44

Marjoram P, Wall JD (2006) Fast "coalescent" simulation. *BMC Genetics* 7:16

Matsumoto M., Nishimura T. (1998) Mersenne Twister: A 623-dimensionally equidistributed uniform pseudorandom number generator. *ACM Trans. on Modeling and Computer Simulation* 8(1):3-30

Newick tree format. http://evolution.genetics.washington.edu/phylip/newicktree.html

Przeworski M. (2002) The signature of positive selection at randomly chosen loci. *Genetics*. 160(3):1179-89

Rambaut A., Grassly N.C. (1997) Seq-gen: An application for the monte carlo simulation of dna sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.* 13:235–238

Schaffner S.F., Foo C., Gabriel S., Reich D., Daly M.J., Altshuler D. (2005) Calibrating a coalescent simulation of human genome sequence variation. *Genome Research* 15:1576–1583

Voight B.F., Kudaravalli S., Wen X., Pritchard J.K.(2006) A map of recent positive selection in the human genome. *PLoS Biol.* 4(3):e72

Weiss G., von Haeseler A. (1998) Inference of population history using a likelihood approach. *Genetics.* 149(3):1539-46

Zöllner S, von Haeseler A. (2000) A coalescent approach to study linkage disequilibrium between single-nucleotide polymorphisms. *Am J Hum Genet*. 66(2):615-28

Chapter IV

GENOTYPE-BASED CASE CONTROL MATCHING TO CORRECT FOR POPULATION STRATIFICATION

4.1 Abstract

Genome-wide association studies are helping to dissect the etiology of complex diseases. Although case-control association tests are generally more powerful than family-based association tests, population stratification can lead to spurious disease-marker association or mask a true association. Several methods have been proposed to match cases and controls prior to genotyping, using family information or epidemiological data, or using genotype data for a modest number of genetic markers. Here, we describe a genetic similarity score matching (GSM) method for efficient matched analysis of cases and controls in a genome-wide or large-scale candidate gene association study. GSM is comprised of three steps: 1) calculating similarity scores for pairs of individuals using the genotype data; 2) matching sets of cases and controls based on the similarity scores so that matched cases and controls have similar genetic background; and 3) using conditional logistic regression to perform association tests. Through computer simulation we show that GSM correctly controls false positive rates and improves power to detect true disease predisposing variants. We compare GSM to

genomic control using computer simulations, and find improved power using GSM. We suggest that initial matching of cases and controls prior to genotyping combined with careful re-matching after genotyping is a method of choice for genome-wide association studies.

4.2 Joint work with Weihua Guan

This chapter was a joint work with Weihua Guan, PhD candidate in the Department of Biostatistics at the University of Michigan. While all work were done interactively with discussion, exchanging ideas, motivations from findings of each other and sharing codes with each other, I have been focusing on the design, evaluation and implementation of matching scores, simulations of genome-scale case-control data with desired population structure and admixture parameters and the likelihood ratio test of conditional logistic regression.

4.3 Introduction

With the success of the International HapMap Project [The International HapMap Consortium, 2007], a dense set of single nucleotide polymorphisms (SNPs) throughout the human genome is now available for genetic studies of complex diseases, and many genome-wide association studies are being undertaken and published [Klein et al., 2005; Maraganore et al., 2005; Cheung et al., 2005; Sladek et al., 2007; Scott et al., 2007; Saxena et al., 2007; Zeggini et al., 2007].

Although case-control association tests are in principle more powerful for detecting disease variants than family-based association tests, population stratification can lead to spurious disease-marker association or mask true association [Li, 1972]. In genome-wide association studies, thousands of samples are typically used to ensure adequate power to identify disease predisposing variants, making it difficult to guarantee genetic homogeneity of the sample [Freedman et al., 2004]. Ancestry information on the sampled individuals may be unavailable to the researchers, and even when available, may not fully specify the underlying population genetic structure, due to vague definitions of ancestry groups and imperfect accuracy of self-report information.

Several methods have been proposed to adjust for the possible confounding effects of population substructure. Family-based association tests, such as the transmission/disequilibrium test (TDT) [Spielman et al., 1993], assess the transmission of alleles from parents to affected offspring. Comparisons are made within parent-offspring trios, and the resulting association test is immune to potential genetic heterogeneity between families. However, collecting trios can be difficult and expensive, and may simply be impractical for late-onset diseases. For unrelated case-control samples, approaches have been proposed to adjust the standard chi-square contingency test statistics according to a non-central χ^2 distribution [Devlin et al., 1999; Gorroochurn et al., 2006], to infer population structure [Pritchard et al., 2000], or to cluster the similarity estimates into several components [Zhang et al., 2002]. A few more recent approaches [Price et al., 2006; Epstein et al., 2007; Kimmel et al., 2007; Luca et al., 2008] focus specifically on genome-wide association studies.

In this paper, we propose a different approach, genetic similarity score matching

(GSM), to correct population stratification using individual-based matching rather than clustering. The huge amounts of data in genome-wide association studies have the potential to provide extremely accurate matching of individuals who share similar ancestries. We match cases with controls based on genetic (dis)similarity scores calculated from the genotype data available in a genome-wide association study or a large-scale candidate gene study and test the resulting matched sets for disease-marker association by conditional logistic regression. This matching-association framework builds on our previous work [Guan et al. 2005] and is similar to that of Luca et al. [2008]. Luca et al. [2008] derive the dissimilarity (distance) scores based on principal components of the variance matrix of genotypes, while our approach obtains the dissimilarity scores based on identity-by-state (IBS) measures. Simulations show that GSM results in false positive rates at the desired nominal level while retaining high power to detect disease associated markers. We find that with large-scale association data, the calculated genetic similarity scores differentiate subpopulations well, and that matching can be done with high accuracy even for samples that are mixtures of genetically similar populations. We further demonstrate that when population stratification is present, association tests based on GSM-matched case-control data can have a higher power than those that rely on either the standard trend test or the genomic-control method.

4.4 Methods

Outline

GSM includes three basic components:

1) *Genetic similarity score:* We calculate genetic similarity scores between pairs of cases and controls across all loci. Large scores should reflect pairs with similar genetic backgrounds.

2) *Matching:* Based on the matrix of similarity scores calculated in 1), we conduct optimal full matching [Rosenbaum, 2002] which groups one case with one or more controls, or one control with one or more cases to maximize the overall similarity of matched cases and controls.

3) *Association tests:* We use conditional logistic regression to assess the association between candidate markers and disease status. For ease of exposition, we consider here only single marker association tests, but other genetic or environmental factors can be easily incorporated into the regression.

Genetic Similarity Score

We define a genetic similarity score for a pair of individuals which measures the degree of similarity of their genotype data. Individuals with similar genetic backgrounds will generally have higher scores. For simplicity, we consider M biallelic genetic markers each with alleles "A" and "a"; the scores can easily be generalized to multiallelic markers. We consider three similarity scores.

The first score calculates the proportion of marker alleles shared identical by state (IBS). If IBS_k is the number of alleles shared at marker k (Table 4.1), then

$$S_{IBS} = \frac{1}{2M^*} \sum_{k=1}^{M^*} IBS_k \qquad (1)$$

where $1 \le M^* \le M$ is the number of markers that are successfully genotyped in both individuals.

While S_{IBS} has the virtue of simplicity, we may want to allow different markers to make different contributions to measure similarity. For example, we may wish to weight sharing a rare allele more strongly than sharing a common allele. We define our second score as:

$$S_{freq} = -\frac{1}{2M^*} \sum_{k=1}^{M^*} \sum_{i \in \{A,a\}} IBS_{k,i} \cdot \log(q_{k,i}) \qquad (2)$$

where $q_{k,i}$ is the frequency of allele i at marker k, and $IBS_{k,i}$ is the number of copies of allele i at marker k shared by the pair of individuals (Table 4.1). We can estimate $q_{k,i}$ using our sample or from the results of previous studies.

In a random mating population, markers are expected to follow Hardy-Weinberg Equilibrium (HWE). When population subdivision is present, tests of HWE tend to be significant owing to excess homozygosity. Our third score takes advantage of this by weighting markers based on their one-sided (excess homozygosity) HWE test p-value p_k [Wigginton et al., 2005]:

$$S_{HWE} = -\frac{1}{2M^*} \sum_{k=1}^{M^*} IBS_k \cdot \log(p_k) \qquad (3)$$

To avoid the impact of genotyping error that may lead to strong deviation from HWE, we exclude the markers that fail quality control; practically speaking, this might mean using markers with HWE p-value satisfying $p > 10^{-6}$.

As an example, suppose 3 cases and 3 controls are genotyped at 3 loci, as listed in

Table 4.2. Then the similarity scores S_{IBS} are as listed in Table 4.3.

For matching, we may use all genotyped markers, or a selected subset. For example, we might pick the markers with the smallest p-values in a HWE test for excess homozygosity, excluding those that fail quality control, in the hope that the selected markers provide maximal information about population stratification in the sample. Further, to avoid selecting markers which are highly correlated, we might choose at most one marker in every *n*-marker window or per linkage disequilibrium group.

In our analyses, matching relies on a transformed dissimilarity score, defined as:

$$D_{ij} = f(S_{ij}) = \left(\frac{\max - S_{ij}}{\max - \min} \right)^2 \qquad (4)$$

where max = $\max_{i,j} S_{ij}$ and min = $\min_{i,j} S_{ij}$, the maximum and minimum similarity scores among all case-control pairs.

Matching

We use the chosen (dis)similarity score to identify optimal matches between cases and controls. The simplest matching scheme is a 1:1 match in which each case is matched to a unique control. This approach is widely used but has obvious drawbacks. For example, when the numbers of cases and controls are not equal, some subjects must be discarded, resulting in a loss of information. Further, samples from various subpopulations often are not equally represented among the cases and controls, leading to forced mismatches if only 1:1 matching is allowed.

Instead, we consider an optimal matching approach that minimizes the total dissimilarity score:

$$T = \sum_{s=1}^{S} \sum_{i \in A_s, j \in B_s} D_{ij}$$

Here, A_s and B_s are the sets of cases and controls in a matched set s, and S is the total number of matched sets. It has been shown that an optimal solution to this minimization problem is a full matching, in which each matched set contains one case and one or more controls, or one control and one or more cases, that is, a 1:m or m:1 matching [Rosenbaum, 1991]. Given n cases and n controls, the summation can in principle contain as few as n terms for 1:1 matching to as many as 2(n-1) terms for 1:n-1 and n-1:1 matching. Since large sets result in larger numbers of terms, optimization tends to favor small matched sets. This helps mitigate any potential power loss due to unbalanced matching, i.e., 1:m or m:1 matching with m>>1 (see Discussion).

The problem of minimizing the total dissimilarity score T is analogous to the classic minimum cost flow (MCF) problem in computer science [Rosenbaum, 1991; Hansen, 2004; Hansen et al., 2006] (Appendix 4.1), and can be solved using the RELAX-IV algorithm [Bertsekas et al., 1994; Frangioni et al., 2006]. Given pre-calculated dissimilarity scores and an upper bound on m, determining the optimal matched set takes on the order of $n^3 \log n$ operations, where n is the total number of subjects. The choice of parameter m constrains the size of matched sets and is somewhat arbitrary; we typically require $m \leq 5$ when numbers of cases and controls are comparable (see Discussion). Prior to matching, we may exclude a few individuals with maximum similarity scores that are extremely small (this is the *caliper* parameter recommended by Hansen et al., 2006). In datasets including ~2,000 individuals, the matching typically takes <1 minute on a modern PC workstation.

To continue with the previous example, we calculate the dissimilarity scores in Table 4.3, and perform both 1:1 matching and optimal matching. In 1:1 matching, the best

match yields three pairs: (1, 4), (2, 5), and (3, 6). The total dissimilarity score is 1/36 + 16/36 + 0 = 17/36. In contrast, the optimal full match has two matched sets: (1, 2, 4) and (3, 5, 6). The matched sets include 4 case-control pairs: (1, 4), (2, 4), (3, 5), and (3, 6). The total dissimilarity score is 1/36 + 0 + 1/36 + 0 = 2/36. In this example, the individuals within group (1, 2, 4) and (3, 5, 6) are similar to each other, and less similar to the individuals in the other group. Full matching offers an obvious matching advantage over 1:1 matching here. In the general case, full matching is guaranteed to produce a total dissimilarity score that is no greater than that obtained using 1:1 matching.

Conditional Logistic Regression

Once matching is done, a natural choice for matched-set analysis is to use conditional logistic regression to test for disease-marker association. We employ an additive model for association by assigning values of 0, 1, and 2 to genotypes AA, Aa, and aa, respectively. Other genotyping coding schemes could be considered, corresponding for example to dominant, recessive, or general models. The regression can easily incorporate genotype, covariate, and interaction effects.

In a genome-wide association scan, we apply conditional logistic regression analysis to each marker separately. The multiple testing problem can be addressed using Bonferroni correction, permutation, or false-discovery rates.

Simulation

We simulated case-control data influenced by genotypes at a disease locus with alleles D and d, under six additive disease models (Table 4.4). We assumed sampling from a population that consisted of two subpopulations. We randomly sampled 500 cases and 500 controls from this mixed population. For each model, the relative risk (RR) of

the predisposing variant allele is set to be the same in different populations. For models 1 and 2, the disease prevalences $K_1=K_2$ and predisposing variant allele frequencies $q_1=q_2$; these models represent the scenario of no population stratification. For models 3 and 4, $K_1<K_2$, creating population stratification in the simulated data. For models 5 and 6, $K_1<K_2$ and $q_1\neq q_2$. For model 5, the first population has lower prevalence but higher predisposing variant allele frequency (K_1=.07, q_1=.55), than the second population (K_2=.13, q_2=.45). For model 6, the population with higher prevalence also has higher predisposing variant allele frequency (K_2=.13, q_2=.55) than the other population (K_1=.07, q_1=.45). For each model, we simulated 500 datasets.

We simulated autosomal SNPs using GENOME, a coalescent-based simulator [Hudson, 1983; Hudson, 1990; Donnelly et al., 1995; Liang et al., 2007]. Assuming discrete generations, GENOME simulates the genealogy of a sample of sequences. As the algorithm proceeds backwards in time, coalescence, recombination, and migration events are simulated. Multiple events can occur in the same generation. We set the effective population size as 10,000, the recombination rate as 10^{-8} per base pair, and the mutation rate as 10^{-9} per base pair, assuming the infinite-site mutation model [Kimura, 1969]. We set the rate of migration between subpopulations to .0025 per individual per generation, which resulted in a distribution of allele frequency differences similar to that observed when comparing HapMap Han Chinese (HCB) and Japanese (JPT) samples (www.hapmap.org). In particular, the mean allele frequency difference between the two simulated populations is .0470, compared to .0477 between the HCB and JPT samples. The simulated genome scans surveyed autosomal genomes of ~2866 Mb comprised of 22 chromosomes, whose lengths approximate the actual lengths of the human autosomes

(NCBI build 33, www.ncbi.nlm.nih.gov/genome/seq/). We randomly selected 300,000 SNPs with minor allele frequencies > .05, and choose a disease liability locus with the desired allele frequencies.

To calculate the similarity scores, we used 10,000 markers with the smallest one-sided HWE p-values, choosing no more than one marker from each 10-marker window. We set the maximum size of matched groups (m) to 6. We compared the type I error and power of GSM, the trend test, genomic control, and EIGENSTRAT for each simulated setting. Given that the simulated samples were drawn from two subpopulations, we used the first principal component to adjust for stratification in EIGENSTRAT; using additional principal components gave similar results. The estimated type I error rates are the proportion of simulated SNPs in which the association test p-value is less than the nominal value 10^{-6}, a significance threshold similar to that typically used in genome-wide scans. In this evaluation of type I error rates, we only considered SNPs that were effectively unlinked to the disease locus. We calculated power as the proportion of simulated replicates where the empirical p-value is $< 10^{-6}$ at the disease locus using a threshold obtained by inspection of test statistics at the null loci.

Bipolar data

We applied GSM to genome-wide association data from the Pritzker Consortium bipolar study (unpublished data). We selected 717 independent bipolar I European American cases and 779 independent European American controls from NIMH Human Genetics Initiative (www.nimhgenetics.org); controls were carefully matched to cases by self-reported ethnicity prior to genotyping. In addition, we downloaded genotype data on 3,182 independent European American controls from Illumina iControlIDB database

(www.illumina.com/pages.ilmn?ID=231). All individuals were genotyped using the Illumina HumanHap550 BeadChip. 505,796 autosomal SNPs passed quality-control criteria in the Prtizker bipolar study: 1) HWE p-values > 10^{-5}; 2) genotype call rate > 95%; and 3) no more than 1 non-mendelian inheritance or inconsistency among 15 father-mother-offspring trios and 30 duplicate samples. Of these, we excluded 1,632 SNPs due to allele frequency differences > .05 between the Illumina and Pritzker control samples. We applied GSM and trend tests for association on the Pritzker samples alone and then on the combined Pritzker and Illumina samples. In GSM, we used the 100,000 markers that passed quality control and have the smallest p-values from the one-sided HWE test to calculate the similarity scores. Given the relatively large control:case ratio of 3,961/717≈5.5, we set the upper limit of the group sizes (m) to 30.

4.5 Results

Similarity score performance in HapMap

We first examined the performance of our similarity scores in the HapMap dataset. We calculated our three similarity scores for all pairs of the 89 independent Han Chinese (CHB) and Japanese (JPT) individuals in the HapMap sample, using 100,000 HapMap phase I autosomal SNPs with MAF>.05, selected based on one-sided Hardy-Weinberg equilibrium test p-values of 4.3×10^{-6} to .11. In Figure 4.1, we showed plots from using multidimensional scaling (MDS) on the similarity score matrices. All three scores showed good separation between the two populations, except for one JPT individual residing in between the two clusters in the plots. The same individual is at a similar position in

principal component analysis when plotting the first two principal components. While S_{IBS} and S_{freq} provided similar separation, S_{HWE} provided less separation with that JPT individual much closer to the CHB cluster instead of the JPT cluster. The relatively poorer performance of S_{HWE} arises because of the heavy weighting of the small subset of markers with very small p-values from the one-sided HWE test, even after we have excluded markers with HWE p-value<10^{-6}.

Our experiences in simulations and real data (unpublished results) suggest that S_{freq} may perform slightly better than S_{IBS} in matching the samples. In the following simulations and analyses, we report results using S_{freq} as our measure of genetic similarity. Although the p-values from one-sided HWE test may not be the best weights for the similarity score as in S_{HWE}, they can still be employed to select a subset of markers for the score computation. In so doing we assume that markers with small HWE p-values but still passing quality control provide more information about population heterogeneity than randomly selected markers. In the following analyses, the matching is usually based on a subset of markers (10,000-100,000 markers) which had the smallest p-values from one-sided HWE test among those passing quality control filters.

False positive rate and power

For the six simulation models, mismatch rates are calculated as the proportion of individuals from population 1 matched to individuals from population 2. The minimal degree of mismatch in the simulations (Table 4.5) suggests accurate matching given the similarity measures and numbers of markers used.

In the absence of population stratification (models 1 and 2), all three methods give false positive rates close to the nominal value of 10^{-6}. The power of our GSM method is

typically ~2% lower than the trend test and genomic control, assumedly due to the unnecessary grouping of samples. When population stratification is present (models 3-6), the type I error rate of the trend test is ~30 times greater than the nominal value, while GSM and genomic control maintain the type I errors at or lower than the nominal value. Using empirical type I error rates, the power of the trend test is equal to that of genomic control, but significantly lower than that of GSM for models 3-4. For models 5 and 6, where population stratification is present, the variation of disease variant frequency may mask the association (model 5) or increase the power to detect association (model 6). For model 5, power of the trend test and genomic control drop ~30% compared to model 3, while GSM maintains the same level of power. For model 6, although the type I error is inflated, the trend test has adjusted power comparable to that of GSM. EIGENSTRAT has power similar to GSM in all simulation settings examined.

We also compared the frequency with which the disease variant is the most strongly associated marker, or among the most strongly associated 10, 100, and 1000 markers, in the trend test or GSM (Figure 4.2). The results are consistent with the observations above. In the absence of population stratification (models 1 and 2), the trend test identifies the disease variant slightly more frequently than GSM. When population stratification is present, GSM picks the correct disease variant more frequently for models 3-5. For model 6, GSM picks the correct disease variant almost as frequently as the trend test.

Bipolar data

We first applied standard trend tests to the Pritzker bipolar case and control samples. The estimated genomic control variance inflation factor λ of the test statistics was 1.03, close to the expected value of 1 when there is no population stratification [Devlin et al.,

1999], arguing that the matching based on self-reported ethnicity resulted in a sample with only limited population stratification. Applying GSM reduced the estimated λ slightly to 1.02. However, when we added the Illumina control samples to the analysis, the estimated λ from standard trend tests became 1.51, indicative of strong population stratification between the cases and controls. We then applied our GSM method on the combined samples, excluding one Illumina control sample that had a noticeably high similarity score with one Pritzker case sample (S_{IBS}=0.85), consistent with a first degree relationship. Using GSM, the estimated λ dropped to 1.072 when we used S_{freq} as our similarity measure and 1.088 using S_{IBS}, suggesting that GSM using either score provided good correction for the stratification problem. Using S_{freq}, each of the 712 cases was matched to one or more controls (i.e., 1:*m* matching only): 316 cases were matched to 1 control, 207 cases to 2-5 controls, 79 cases to 6-10 controls, and 115 cases to 10-30 controls. To check the appropriateness of setting the maximum number of controls (*m*) at 30, we repeated our analysis by changing *m* to 10 or 50, resulting in estimated λ values of 1.23 and 1.067, respectively. This suggests that some controls may be matched to dissimilar cases when we only allow up to 10 controls per case, while increasing *m* from 30 to 50 resulted in little improvement on the matching. Since the combined sample contains many more controls than cases, we considered removing some controls with relatively high dissimilarity by restricting the total number of controls to be matched from 3,960 to 3,500, and the estimated λ dropped slightly to 1.065. We also repeated the matching using 50,000 markers instead of 100,000, and in this setting the estimated λ increased slightly to 1.086, as expected.

As a comparison, we also applied EIGENSTRAT and another principal

component-based method (Luca et al. [2008], GEM) to the bipolar data, using 10 principal components. Without removing any potential outliers, EIGENSTRAT gave an estimated λ of 1.074, comparable to our results. GEM removed 132 samples as outliers and gave a slightly better estimated λ of 1.063. When we applied our method to the same set of samples used in GEM, we obtained an estimate λ of 1.065. Although the removal of these samples decreased the inflation of type I error rates, its impact on power requires further investigation.

4.6 Discussions

Population stratification, which can result in high false-positive rates and mask true associations, poses a potential problem for case-control association studies. In this paper, we propose GSM, a practical approach to correct for population stratification for large-scale association studies that uses information at thousands of genotyped genetic markers to group case and control subjects according to their similarity. Simulation studies show that GSM can control the false positive rates in the presence of population substructure, while maintaining power to detect disease loci.

GSM is computationally efficient. The computational time for similarity score calculation is linear in the number of markers used and in the number of all case-control pairs, and the time for matching is approximately cubic in the number of individuals.

We have compared the performance of GSM to the commonly used genomic control method [Devlin et al., 1999]. Genomic control assumes that a scaled test statistic (dividing the standard test statistic by a global correction factor λ) has an approximate

central χ^2 distribution. When stratification is modest, the genomic control procedure is able to control the false-positive rate at the nominal level through λ, but does not change the relative order of the test statistics along the genome. As shown in our simulations (model 3-5), when stratification masks the association, genomic control can be quite conservative. Another popular approach to correct for population stratification is structured association [Pritchard et al., 2000] which infers population structure using a set of independent makers. We did not evaluate this method in our simulations due to its computational intensity. Structured association also requires an assumption about the number of underlying subpopulations in the sample. EIGENSTRAT [Price et al., 2006] is an approach for genome-wide association studies based on principal components analysis (PCA). It has been shown that the K-1 principal components can be related to the solution to the K-way clustering solution [Ding et al., 2004]. EIGENSTRAT is less sensitive to the number of components than structured association (if the number is sufficiently large) because of orthogonality of the axes of variation, but the interpretation of the axes is less intuitive.

Our new GSM method tackles the stratification problem by matching at the individual level, without assuming an explicit population structure. Effectively, it treats every sample as a single population and compares it to the most similar counterparts. For samples from clearly distinguished subpopulations, such as the HapMap HCB and JPT populations or the two subpopulations in our simulations, GSM performs almost as well as cluster-based matching or EIGENSTRAT, with little loss of power. In real GWA studies, where sampled individuals may often derive from continuous mixtures of ancestral populations, the individual-based matching in GSM should be more flexible

than cluster-based matching. Luca et al. [2008] (GEM) also applied full matching to correct for population stratification, but used a different score calculated from the top eigenvectors from PCA. They showed that outliers may greatly inflate type I errors of association tests using EIGENSTRAT and need to be carefully removed beforehand. The similarity scores in GSM can be used like the GEM scores to identify outliers, but are more intuitive in measuring genetic similarity, compared to the abstract measures from eigenvectors used in GEM. In addition, PCA analysis is very sensitive to the independence of samples, while GSM can actually help to identify related samples through IBS scores. In our Pritzker study example, we found one pair of individuals with large similarity score of 0.85 (S_{IBS}), which strongly suggested a potential first-degree relative. Although the two samples showed strong correlation in their PC scores, they were not identified as outliers by EIGENSTRAT or GEM because their scores did not show strong deviation from the center of the score distributions in the top 10 PCs.

The success of our GSM procedure depends on the accuracy of matching. Incorrectly grouping individuals from different populations could inflate the type I error rate, decrease the power to detect the susceptibility genes, or both. To ensure correct matching, a well-defined similarity measure and a substantial number of markers in which to compute this measure are both important. Our simulations analysis and practical experience, show that similarity measures derived from the distribution of IBS between pairs of individuals, which are simple to calculate and do not require much computing power, provide an effective means of matching individuals. Furthermore, we found that weighting IBS estimates by a function of the marker allele frequencies (S_{freq}) improved the accuracy of matching. Other score metrics also exist and can be easily incorporated

into our approach to substitute the IBS-based scores presented. As an experiment, we considered similarity scores based on pairwise IBD estimates calculated using an E-M algorithm, the average mismatch rates using IBD-based scores were slightly higher than those for IBS-based scores. A weakness of IBD based scores is that they are truncated at zero: when many pairs of individuals are assigned IBD ~0, it becomes difficult to select optimal pairings. Figure 4.3 demonstrates the relationship between the IBD scores and IBS scores (S_{freq}) computed on the HapMap HCB and JPT samples.

The number of markers used in score calculation is another factor that affects the matching. We prefer to calculate the scores based on a large set of markers (typically including 10,000 – 100,000 SNPs). However, using too many markers increases the computational load while not necessarily improving the accuracy of matching. In our simulations, 10,000 markers with the smallest p-values from one-sided HWE test can correctly match the individuals from closely related populations such as Han Chinese and Japanese, with zero or almost zero mismatch (Table 4.5). In this example, using 30,000 markers worked as well as using 10,000 markers, while using only 1,000 markers led to incorrect grouping of individuals from different populations with up to ~10% mispaired individuals. For samples with subtle differences in genetic ancestry, such as the European American samples in the bipolar data, more markers (50,000 to 100,000, passing quality control) may help to obtain better matching. Inspecting the genomic control parameter λ on its closeness to the expected value of 1 from different analysis strategies can help to determine the appropriate number of markers for controlling stratification. To select the subset of markers, we usually prefer those with smaller p-values from one-sided HWE tests, because they tend to be more informative about population structure. However, we

need to be cautious regarding data quality, since markers with high error rates may show strong deviation from HWE and then give incorrect information about the genetic background of sampled individuals. A reasonable compromise is to exclude SNPs with extreme deviations from HWE (say, $p<10^{-6}$) but focus on those with mild deviations (say, $10^{-2}<p<10^{-6}$) to evaluate stratification. GSM does not require that *all* markers should be independent of disease status, since in a typical genome-wide setting the vast majority of markers will meet this criterion and the impact of disease-associated markers on the similarity scores is negligible and can be ignored. Furthermore, since our similarity scores are a function of the mean (weighted) IBS values across a large number of markers, it is also not critical that the assessed SNPs should be independent of each other.

We chose not to include X-linked markers in our matching scheme to avoid any possible biases due to differences by gender. Given genome-wide association data, the autosomal markers provide ample information for accurate matching.

When there is no population stratification, our simulations showed a small loss of power in GSM due to unnecessary matching. Studies have shown that when the population is indeed homogeneous, random matching by pairs (1:1) can do almost as well as the unmatched test [Chase, 1968]. Additional power may be lost when the matching is not balanced, so that multiple controls are compared to a single case subject or multiple cases are compared to a single control (i.e., 1:m or m:1 when $m > 1$). However, when stratification is present, larger values of m are preferred to decrease the chance of matching errors. It is then a trade-off of efficiency and bias that we need to consider in practice. In our GSM method, the objective function (T) we choose for optimal matching favors smaller groups, minimizing loss of efficiency. Although the original optimal

matching [Rosenbaum, 1991] is unconstrained ($m = \infty$) so that all controls are allowed to be matched to a single case or all cases to a single control, Hansen [2004] showed that the matching with restriction on m can reduce the variance of estimated parameters with little increase in bias, and suggested a linear search for good values of m that are as close to 1 as possible. In our simulations, a large proportion of the matched sets are 1:1 matches even when the proportions of the two populations in cases and controls are not equal, and the average size of matched sets does not vary much for different values of the upper bound of m. For example, for simulated setting 3, the average matched set size is 2.44 and 2.47 when the upper limits of m are set as 2 and 5, respectively.

Although the full matching scheme is flexible, cases (or controls) from a population without a corresponding partner among the controls (or cases) will decrease power and may lead to spurious association if matching is forced. Further, 1:1 matching is more efficient than m:1 for m>1. Therefore, we still strongly encourage careful sample selection during the study design. Skol et al. [2005] showed that the self-reported ethnicity can be a good predictor for population structure, consistent with our results based on the NIMH case and control samples alone.

In summary, we propose a new framework to match case and control samples by their genetic similarity and adjust for the underlying population substructure. Our GSM method is specifically designed to use the full information provided by the large number of genotypes in genome-wide association studies or large-scale candidate gene studies. Our method can correctly control the false positives, while maintaining considerable power to detect the disease-marker association. Our individual-based matching scheme can reflect the continuous mixing of ancestral populations. By comparing each case to

one or more controls sharing the most genetic backgrounds, we hope our method may increase the chance to identify the genetic variants that influence disease risk. Our GSM software is available freely with C++ source code at http://www.sph.umich.edu/csg/liang/gsm/. The package allows the users to automatically calculate matching score matrices, conduct full matching with a range of parameter choices, and carry out association analyses. We expect our method will aid analyses of large-scale genome-wide association studies.

4.7 Acknowledgements

We thank the Pritzker Neuropsychiatric Disorders Research Consortium, which is supported by the Pritzker Neuropsychiatric Disorders Research Fund L.L.C. We thank NIMH Human Genetics Initiative for providing the case and control samples, and the Illumina database and its contributors for providing the additional control samples. This research was supported by grants HG000376 (MB) and HG02651 and HL084729 (GRA) from the National Institutes of Health.

4.8 Appendix

Appendix 4.1 Minimum cost flow (MCF) problem

In a minimum cost flow (MCF) problem, we define a directed graph consisting of nodes, $i \in \mathrm{N}$, and arcs connecting the nodes, $(i, j) \in \mathrm{A}$. For each arc (i, j), an integer a_{ij} denotes the cost and a positive integer c_{ij} the capacity. For each node i, an integer s_i denotes the exogenous supply. A solution of the MCF problem is a set of arc flows x_{ij} that minimizes:

$$\sum_{(i,j)\in \mathrm{A}} a_{ij} x_{ij}$$

subject to the constraints on capacity:

$$\sum_{\{j|(i,j)\in \mathrm{A}\}} x_{ij} - \sum_{\{j|(j,i)\in \mathrm{A}\}} x_{ji} = s_i, \quad \text{for all } i \in \mathrm{N}$$
$$0 \le x_{ij} \le c_{ij}, \quad \text{for all } (i,j) \in \mathrm{A}$$

It is easy to see the equivalence between the MCF and the optimal matching (Figure 4.4). The nodes in a directed graph correspond to the cases and controls, a_{ij} is the dissimilarity measure between i and j, and the capacity of the flow, c_{ij}, is 1 between case and control nodes, and 0 between two cases or two controls. The optimal solution of the MCF problem is equivalent to an optimal matching. The nodes connected by arcs with non-zero flow are assigned to the same matched set.

In full matching, the numbers of case-control pairs vary across matched sets, so the supply of nodes (s_i) cannot be predetermined. To deal with this complication, we include an "overflow" node to the graph to balance the flows from or to the case or control nodes. Parameters U and $\mathrm{U_c}$ control the maximum flows going to "overflow" from each node, which correspond to the maximum number of cases or controls allowed in each matched set, i.e., the upper limit of m in 1:m or m:1 match. For each case node, there are m

connected control nodes and U-*m* arcs connecting it to “overflow”; for each control node, there are *m* connected case nodes and *m* arcs connecting to “overflow”. The cost for arcs entering “overflow” is set as 0, so these extra arcs do not affect the total cost. Similarly, another node, “sink”, may also be added to control the total number of controls to be matched, and the cost for arcs entering “sink” is also 0 (Hansen et al., 2006).

The translation is demonstrated in Figure 4.4. The MCF problem is then solved by iteratively updating a dual cost vector and the flow vector $\boldsymbol{x}$ (Bertsekas et al., 1994; Frangioni et al., 2006).

4.9 Tables and figures

Table 4.1 Values of IBS_k and $IBS_{k,i}$ for calculation of similarity scores

Genotype Pair	IBS_k	$IBS_{k,A}$	$IBS_{k,a}$
aa aa	2	0	2
aa Aa	1	0	1
aa AA	0	0	0
Aa Aa	2	1	1
Aa AA	1	1	0
AA AA	2	2	0

Table 4.2 Example genotypes

Cases		Controls	
Individual	Genotype	Individual	Genotype
1	aa, aa, AA	4	aa, aa, Aa
2	aa, aa, Aa	5	Aa, AA, aa
3	AA, AA, aa	6	AA, AA, aa

Table 4.3 Similarity (dissimilarity) scores for individuals in table 4.2

	Controls		
Cases	4	5	6
1	5/6 (1/36)	1/6 (25/36)	0 (1)
2	1 (0)	2/6 (16/36)	1/6 (25/36)
3	1/6 (25/36)	5/6 (1/36)	1 (0)

Table 4.4 Characteristics of simulated disease models

Model	Population 1			Population 2		
	K_1	p_1	RR_1	K_2	p_2	RR_2
1	.10	.5	1.6	.10	.5	1.6
2	.10	.2	1.6	.10	.2	1.6
3	.07	.5	1.6	.13	.5	1.6
4	.07	.2	1.6	.13	.2	1.6
5	.07	.55	1.6	.13	.45	1.6
6	.07	.45	1.6	.13	.55	1.6

Samples drawn from two subpopulations in 1:1 ratio
K_i: disease prevalence in population *i*.
p_i: predisposing variant allele frequency in population *i*.
RR_i: relative risk of the predisposing variant allele in population *i*.

Table 4.5 Average false positive rate and power of GSM, trend test and genomic control

Setting	Mismatch (%)	$\lambda^{\S}$	Average false positive rate ($\times 10^{-6}$)				Power*		
			GSM	Chisq	GC	EIGEN	GSM	GC	EIGEN
1	0	1.01	1.08	1.29	1.19	0.93	.80	.82	.82
2	0	1.01	1.10	1.16	1.10	0.97	.55	.56	.56
3	0.016	1.39	1.17	31.8	0.73	1.03	.75	.53	.76
4	0.015	1.38	1.15	30.7	0.47	1.07	.54	.28	.55
5	0.010	1.37	1.14	31.2	0.64	0.90	.72	.22	.72
6	0.010	1.38	1.09	33.0	0.66	0.87	.79	.78	.81

500 cases and 500 controls, 300,000 SNPs with MAF > .05, significance level = 10^{-6}
Chisq represents Trend test and GC represents genomic control

§. The global correction parameter in genomic control (GC), averaged over simulation replicates.

*. Power adjusted for the nominal false positive rates.

Figure 4.1 Multidimensional Scaling plots using dissimilarity scores as distance measure (calculated from 100,000 SNPs) for Han Chinese (HCB) and Japanese (JPT) HapMap samples

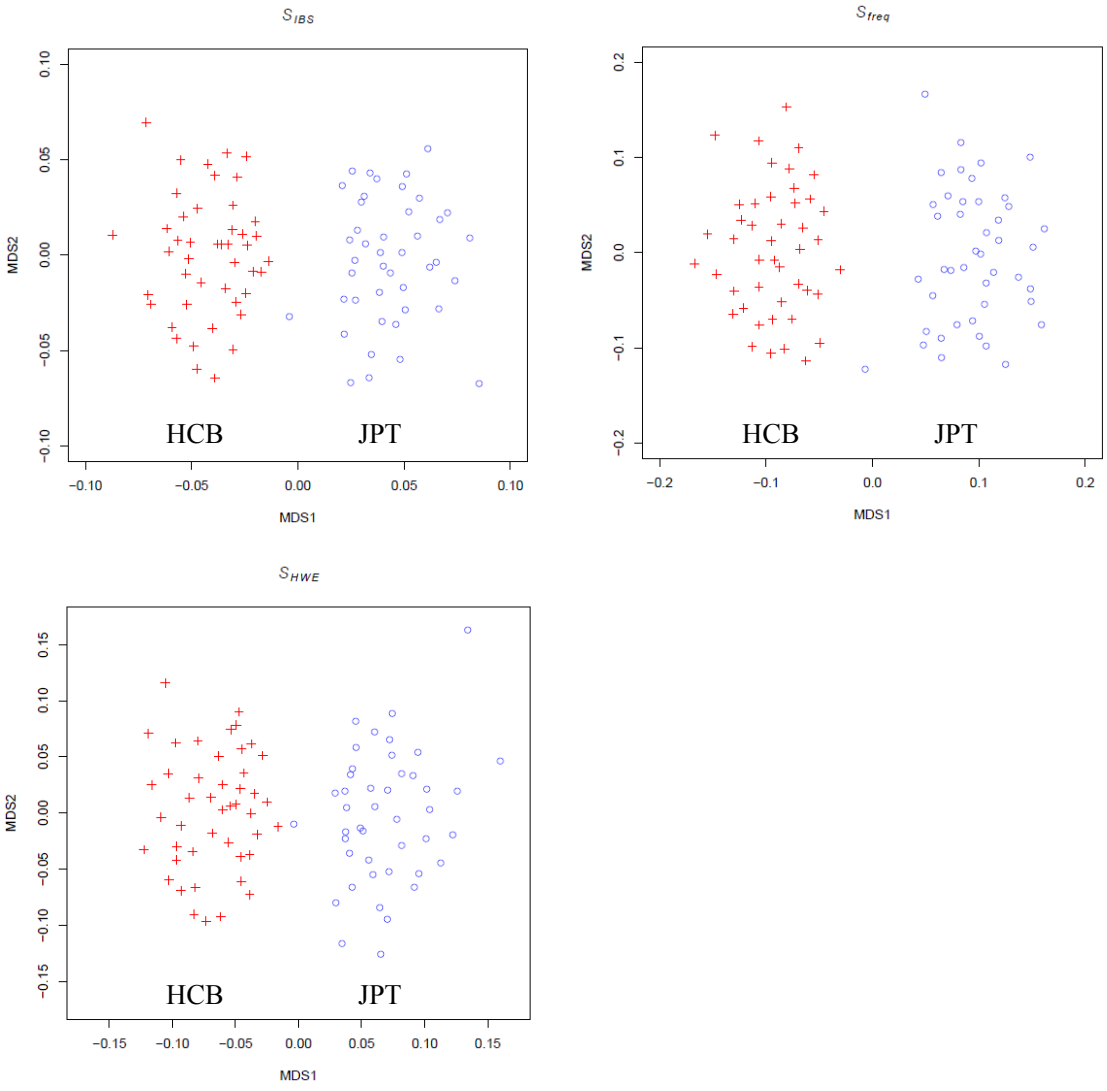


Figure 4.2 The frequencies of disease predisposing variant being identified among the best markers by similarity score matching method (GSM), EIGENSTRAT and trend test (Chisq)

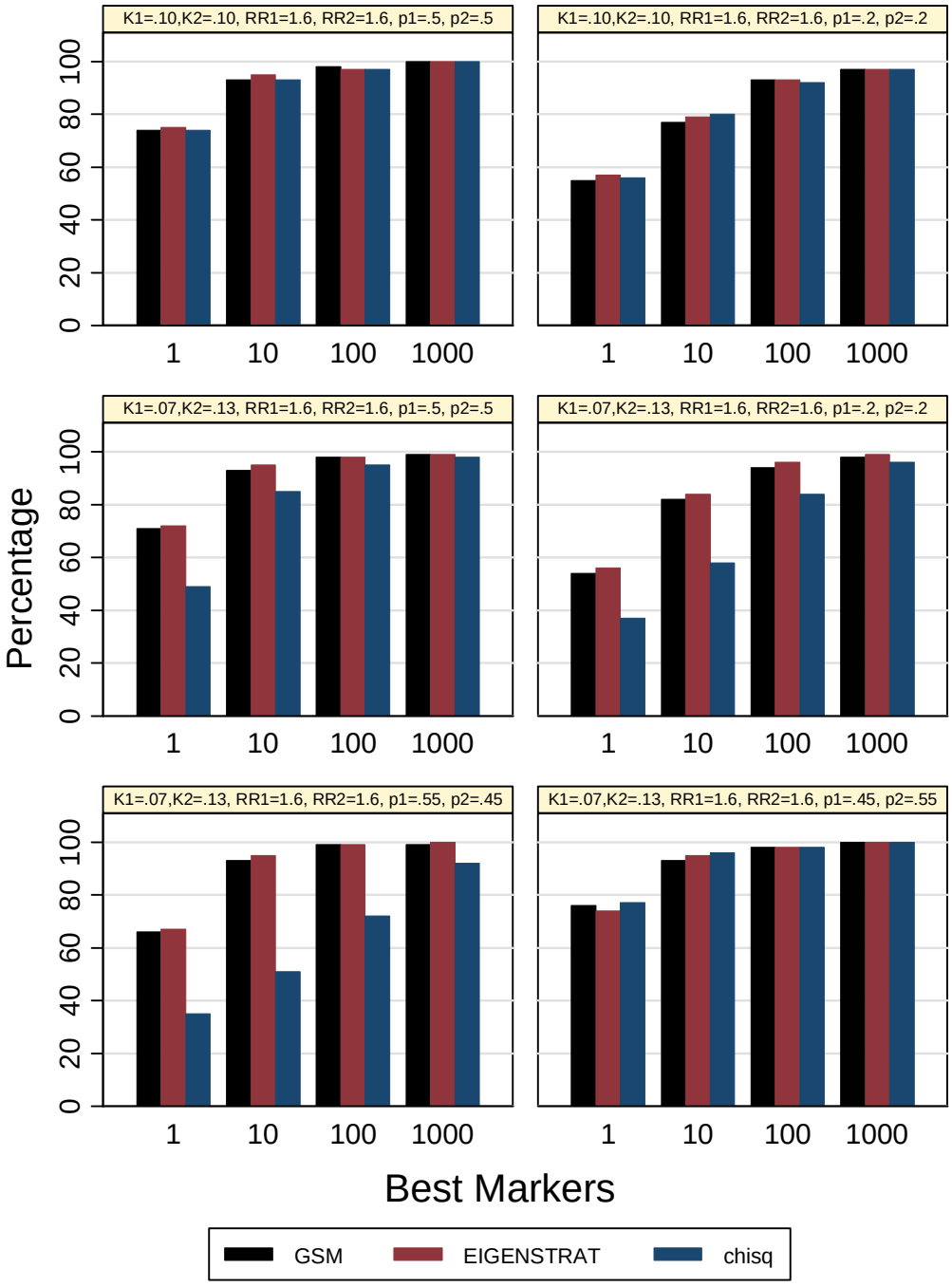
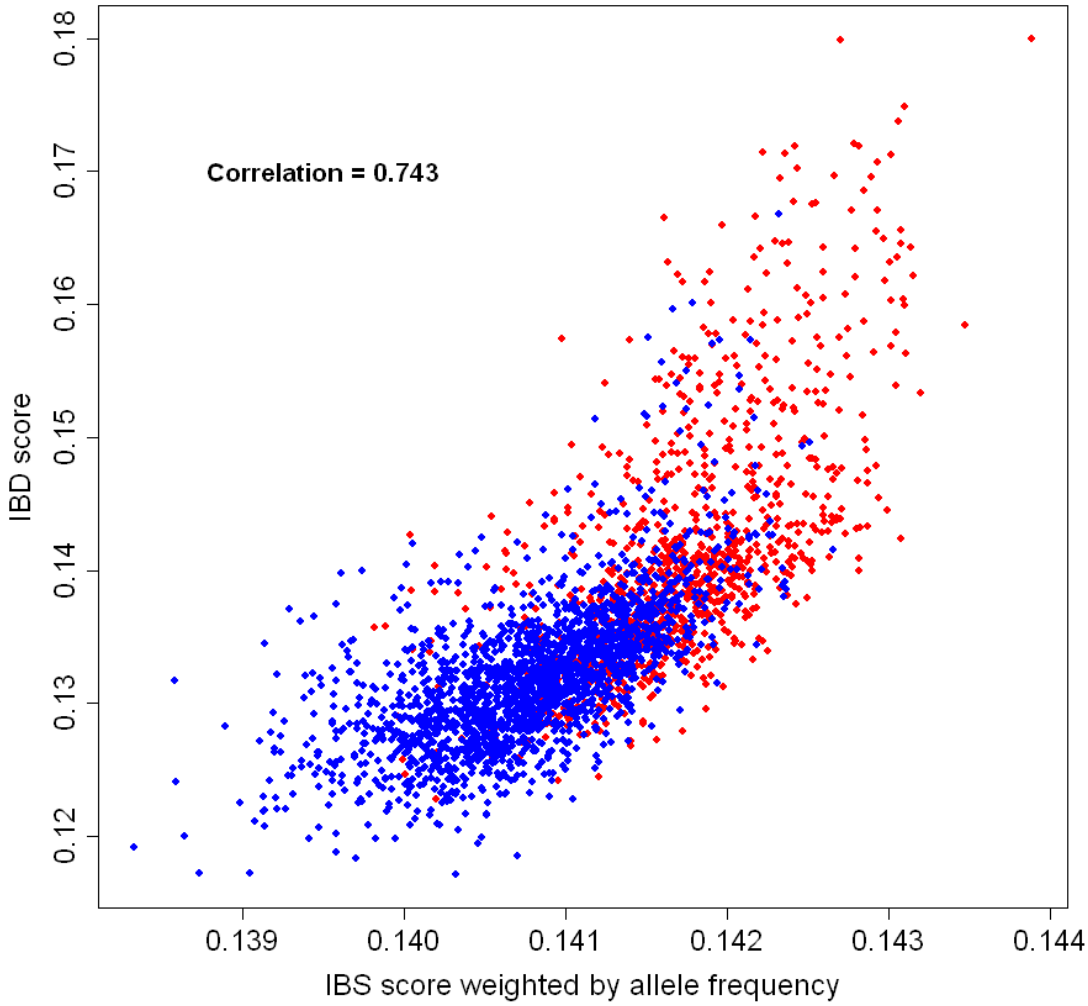
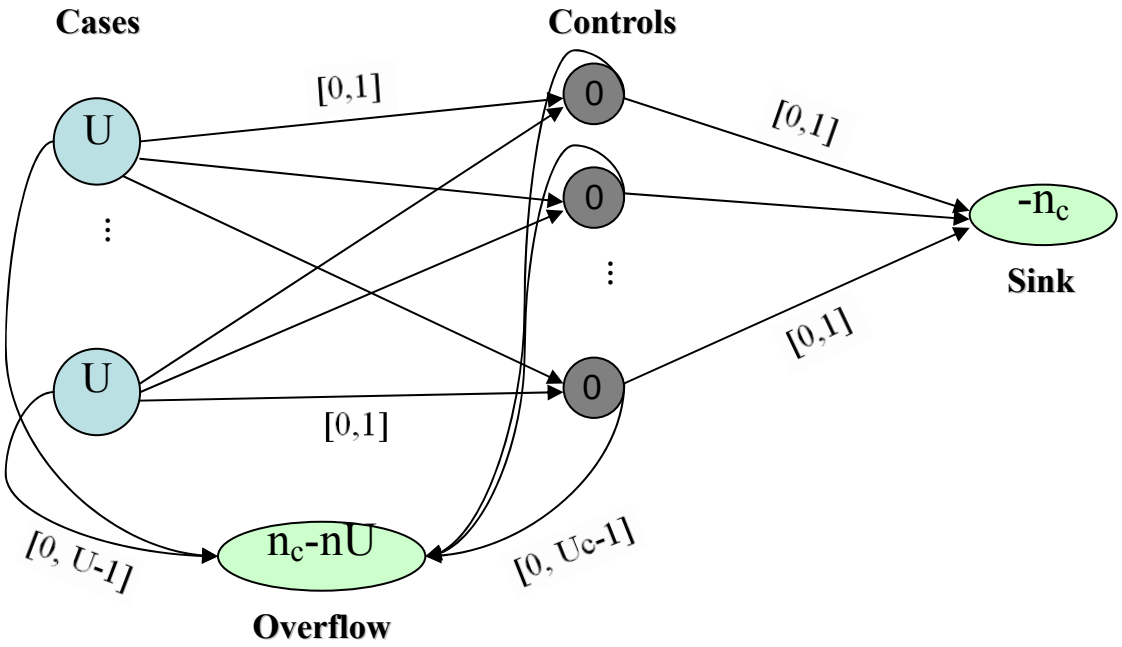


Figure 4.3 Similarity scores (calculated from 888,071 SNPs) between each pair of Han Chinese (HCB) and HCB-Japanese (JPT) in HapMap



Red: HCB-HCB pair; Blue: HCB-JPT pair.

Figure 4.4 Solve optimal full matching problem as a minimum cost flow (MCF) problem



U denotes the maximal number of controls each case can match, U_c the maximal number of cases each control can match, n_c the number of controls to match, and n the total number of cases and controls.

4.10 References

Bertsekas DP, Tseng P. 1994. RELAX-IV: A faster version of the RELAX code for solving minimum cost flow problems. Technical report LIDS-P-2276, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA.

Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT. 2005. Mapping determinants of human gene expression by regional and genome-wide association. Nature 437:1365-1369.

Ding C, He X. 2004. K-means clustering via principal component analysis. Proc Intl Conf Machine Learning (ICML 2004). p 225-232.

Donnelly P, Tavaré S. 1995. Coalescents and genealogical structure under neutrality. Annu Rev Genet 29:401-421.

Devlin B, Roeder K. 1999. Genomic control for association studies. Biometrics 55:997-1004.

Epstein MP, Allen AS, Satten GA. 2007. A simple and improved correction for population stratification in case-control studies. Am J Hum Genet 80:921-930.

Frangioni A, Manca A. 2006. A computational study of cost reoptimization for min cost flow problem. INFORMS J Comput 18:61-70.

Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, Patterson N, Gabriel SB, Topol EJ, Smoller JW, Pato CN, Pato MT, Petryshen TL, Kolonel LN, Lander ES, Sklar P, Henderson B, Hirschhorn JN, Altshuler D. 2004. Assessing the impact of population stratification on genetic association studies. Nat Genet 36:388-393.

Gorroochurn P, Heiman GA, Hodge SE, Greenberg DA. 2006. Centralizing the non-central chi-square: a new method to correct for population stratification in genetic case-control association studies. Genet Epidemiol 30:277-289.

Gu XS, Rosenbaum PR. 1993. Comparison of multivariate matching methods: structures, distances, and algorithms. J Comput Graph Stat 2:405-420.

Guan W, Liang L, Boehnke M, Abecasis GR. 2003. Matching cases and controls using genotype data from a whole genome association study. ASHG 2005 Annual Meeting #2395 (poster).

Hansen BB. 2004. Full matching in an observational study of coaching for the SAT. J Am Stat Assoc 99:609-618.

Hansen BB, Klopfer SO. 2006. Optimal full matching and related designs via network

flows. J Comput Graph Stat 15:609-627.

Hudson RR. 1983. Properties of a neutral allele model with intragenic recombination. Theor Popul Biol 23:183-201.

Hudson RR. 1990. Gene genealogies and the coalescent process. Oxford Surv Evol Biol 7:1-44.

Khlat M, Cazes MH, Genin E, Guiguet M. 2004. Robustness of case-control studies of genetic factors to population stratification: magnitude of bias and type I error. Cancer Epidemiol Biomarkers Prev 13:1660-1664.

Kimmel G, Jordan MI, Halperin E, Shamir R, Karp RM. 2007. A Randomization Test for Controlling Population Stratification in Whole-Genome Association Studies. Am J Hum Genet 81:895-905.

Kimura, M. 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics 61:893-903.

Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, Sangiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J. 2005. Complement factor H polymorphism in age-related macular degeneration. Science 308:385-389.

Li CC. 1972. Population subdivision with respect to multiple alleles. Ann Hum Genet 33:23-29.

Liang L, Zöllner S, Abecasis GR. 2007. GENOME: a rapid coalescent-based whole genome simulator. Bioinfomatics 23:1565-1567.

Luca D, Ringquist S, Klei L, Lee AB, Gieger C, Wichmann HE, Schreiber S, Krawczak M, Lu Y, Styche A, Devlin B, Roeder K, Trucco M. 2008. On the use of general control samples for genome-wide association studies: genetic matching highlights causal variants. Am J Hum Genet 82:453-463.

Maraganore DM, de Andrade M, Lesnick TG, Strain KJ, Farrer MJ, Rocca WA, Pant PVK, Frazer KA, Cox DR, Ballinger DG. 2005. High-resolution whole-genome association study of Parkinson disease. Am J Hum Genet 77:685-693.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904-909.

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. 2000. Association mapping in structured populations. Am J Hum Genet 67:170-181.

Rosenbaum PR. 1991. A characterization of optimal designs for observational studies. J Roy Statist Soc Ser B 53:597-610.

Rosenbaum PR. 2002. Observational Studies, 2nd ed. New York: Springer.

Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PI, Chen H, Roix JJ, Kathiresan S, Hirschhorn JN, Daly MJ, Hughes TE, Groop L, Altshuler D, Almgren P, Florez JC, Meyer J, Ardlie K, Bengtsson K, Isomaa B, Lettre G, Lindblad U, Lyon HN, Melander O, Newton-Cheh C, Nilsson P, Orho-Melander M, Rastam L, Speliotes EK, Taskinen MR, Tuomi T, Guiducci C, Berglund A, Carlson J, Gianniny L, Hackett R, Hall L, Holmkvist J, Laurila E, Sjogren M, Sterner M, Surti A, Svensson M, Svensson M, Tewhey R, Blumenstiel B, Parkin M, Defelice M, Barry R, Brodeur W, Camarata J, Chia N, Fava M, Gibbons J, Handsaker B, Healy C, Nguyen K, Gates C, Sougnez C, Gage D, Nizzari M, Gabriel SB, Chirn GW, Ma Q, Parikh H, Richardson D, Ricke D, Purcell S. 2007. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science 316:1331-1336.

Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, Erdos MR, Stringham HM, Chines PS, Jackson AU, Prokunina-Olsson L, Ding CJ, Swift AJ, Narisu N, Hu T, Pruim R, Xiao R, Li XY, Conneely KN, Riebow NL, Sprau AG, Tong M, White PP, Hetrick KN, Barnhart MW, Bark CW, Goldstein JL, Watkins L, Xiang F, Saramies J, Buchanan TA, Watanabe RM, Valle TT, Kinnunen L, Abecasis GR, Pugh EW, Doheny KF, Bergman RN, Tuomilehto J, Collins FS, Boehnke M. 2007. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science 316:1341-1345.

Skol AD, Xiao R, Boehnke M, Veterans Affairs Cooperative Study 366 Investigators. 2005. An algorithm to construct genetically similar subsets of families with the use of self-reported ethnicity information. Am J Hum Genet 77:346-354.

Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, Boutin P, Vincent D, Belisle A, Hadjadj S, Balkau B, Heude B, Charpentier G, Hudson TJ, Montpetit A, Pshezhetsky AV, Prentki M, Posner BI, Balding DJ, Meyre D, Polychronakos C, Froguel P. 2007. A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 445:881-885.

Spielman RS, McGinnis RE, Ewens WJ. 1993. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52:506-513.

The International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851-861.

Wigginton JE, Cutler DJ, Abecasis GR. 2005. A note on exact tests of Hardy-Weinberg equilibrium. Am J Hum Genet 76:887-893.

Zeggini E, Weedon MN, Lindgren CM, Frayling TM, Elliott KS, Lango H, Timpson NJ, Perry JR, Rayner NW, Freathy RM, Barrett JC, Shields B, Morris AP, Ellard S, Groves CJ, Harries LW, Marchini JL, Owen KR, Knight B, Cardon LR, Walker M, Hitman GA, Morris AD, Doney AS, McCarthy MI, Hattersley AT. 2007. Replication of genome-wide association signals in U.K. samples reveals risk loci for type 2 diabetes. Science 316:1336-1341.

Zhang S, Kidd KK, Zhao H. 2002. Detecting genetic association in case-control studies using similarity-based association tests. Statistica Sinica 12:337-359.

Chapter V

GLOBAL GENE EXPRESSION MAPPING AND GENOTYPE IMPUTATION TO ENHANCE GENOME-WIDE ASSOCIATION STUDIES

5.1 Abstract

Gene expression levels can be an important step between DNA variation and phenotypic manifestations. We constructed a genome-wide genetic map of genetic variants that influence global gene expression integrating data from two independent samples, gene expression data measured on 405 children using Affymetrix technique and data from 550 children using Illumina BeadChip. We identified genome-wide significant *cis* eQTLs for more than 1,000 genes from each dataset. The resulting comprehensive eQTL maps provide much information about biological regulation of gene expression and may serve as a general tool to aid in interpreting the results of disease association. Using this dataset, we perform systematic evaluation of accuracy and power of genotype imputation with respect to different aspects of the phenotypic traits of interest and genetic markers being tested. We carried out genome-wide association studies of global gene-expression using data for ~300,000 SNPs genotyped with Illumina arrays, before and after imputation. Analyses of imputed data increased the number of signals mapped in cis by 11.1% to 1,391 and maintained similar false discovery rates. The QTLs mapped after imputation (but missed before imputation) have a broad allele frequency spectrum

and, sometimes, represent quite large effects that were not well tagged by individual SNP in the original chip. To evaluate the accuracy of genotype imputation, we genotyped 58,819 SNPs in the same set of individuals using a different Illumina array. We observed high imputation accuracy and high correlation between the imputed and actual allele counts especially for common SNPs (average error rate is 0.033 with 80.85% SNPs having error rates < 0.05). We compared association results for imputed and genotyped SNPs at the 58,096 SNPs (MAF>2%) and found the correlation between LOD scores obtained from analysis of genotyped SNPs and their imputed counterparts was 0.952; we also found the estimated correlation between true and imputed genotypes (which can be calculated even when the true genotypes are unknown) to be a good predictor of the correlation between LOD scores for imputed and genotyped versions of the same SNP. In summary, we found that imputation based analysis can increase power of genome-wide association studies carried out using modern genotyping arrays. The results of our genome-wide association studies of global gene expression are available online to help investigators examine the functional consequences of interesting SNPs.

5.2 Introduction

Genome-wide association studies have detected many new loci for complex traits at stringent significance levels. However, it is not always clear how to connect association results to biological functions. Variation in transcription can mediate disease and might help to explain mechanism for many disease associated SNPs. Transcription abundance is directly regulated by genetic elements, and SNPs that modify these elements and are

associated with transcript levels can be mapped with high power (Schadt et al. 2003, Morley et al. 2004, Spielman et al. 2007, Stranger et al. 2007, Dixon et al. 2007). When the same genetic variant is associated with disease and transcript level, the gene could serve as a candidate for disease studies (Libioulle et al. 2005, Moffatt et al. 2007, Cookson et al. 2009). By measuring gene expression and genetic variants at genome-wide scale on a large number of individuals, statistical methods can be used to map genetic factors (*cis* or *trans*) for thousands of transcripts on the genome. The resulting comprehensive eQTL maps may serve as a general tool to aid in interpreting the results of disease association. The availability of the eQTL map database to the community might give immediate insight into the biological basis of disease associations from different genome-wide association studies of complex diseases. The systematic eQTL mapping on genome scale also may improve our understanding of the biological control of gene expression.

Our map provides a catalog of more than a thousand *cis* eQTLs. Since most eQTL loci are *cis*-regulators of gene expression, analysis of our dataset with different approaches provides a natural approach to evaluate the relative power of different gene mapping strategies (more powerful strategies tend to detect a larger number of *cis*-eQTL, at the same type I error rate). Here, our goal is to use our dataset to evaluate the power of genotype imputation. Genotype imputation approaches utilize a reference panel typed on millions of markers to estimate missing data in samples genotyped at a subset of these markers (Li et al. 2008, Marchini et al 2007, Servin & Stephens 2007, Scheet & Stephens 2006). The approaches are commonly used to increase the power and coverage of individual genome-wide association studies and to facilitate meta-analysis of data across

studies that are relied on different commercial genotyping platforms (for early examples, see Willer et al. 2008, Sanna et al. 2008, Scott et al. 2007, The Wellcome Trust Case Control Consortium 2007).

To date, most studies using genotype imputation have used one of the commercial genotyping arrays (from Illumina, Affymetrix, Perlegen Biosciences, among others) to genotype study samples and then used the HapMap samples as template reference panel (The International HapMap Consortium, 2007). Simulation experiments and detailed genotyping within select regions show that this strategy should result in imputed genotypes that are highly accurate and that the analysis of imputed genotypes increases power for association studies (Li et al. 2008). Still, a large scale assessment of the accuracy of genotype imputation and, particularly, of its impact on power remains unknown.

In this chapter, I assess the accuracy of genotype imputation by comparing imputed and experimentally derived genotypes on a genomic scale. Furthermore, I empirically evaluate the gain in power that results from genotype imputation by systematically contrasting the results of 15,084 genome-wide association scans for a series of mRNA transcript levels before and after imputation based analyses. This global assessment of mRNA transcript levels includes a variety of traits, each with its own (unknown) genetic architecture. Each trait is potentially influenced by a mix of common and rare variants, single nucleotide polymorphisms and copy number variants, genetic heterogeneity, etc. – a complex scenario that would be challenging to replicate in a simulation study. Despite the unknown genetic architecture underlying mRNA expression levels, we can contrast the power of different analytical strategies by tallying the number of *cis* signals that reach

genome-wide significance levels with these strategies.

5.3 Materials and Methods

Global gene expression data were measured by two techniques and in two independent samples. The first sample contains 405 children of British descent (Dixon et al. 2007). The 405 children are organized into 206 sibships, including 297 sib pairs and 11 half-sib pairs. The families were identified through a proband with childhood asthma and siblings were included regardless of disease status. Global gene expression in lymphoblastoid cell lines (LCLs) was measured using Affymetrix HG-U133 Plus 2.0 chips. LCL cultures were harvested at log phase in the first growth after Epstein-Barr virus (EBV) transformation. Robust multi-array averaging (RMA, Irizarry et al. 2003; Bolstad et al. 2003) was used for background correction, normalization and to compute expression values. All 405 children and their parents were genotyped using the Illumina Sentrix Human-1 Genotyping BeadChip (ILMN100K, including 105,713 autosomal SNPs) and 378 children were also genotyped using Illumina Sentrix HumanHap300 BeadChip (ILMN300K, including 307,981 autosomal SNPs) according to manufacturers' instructions (Dixon et al. 2007; Moffatt et al., 2007). Before analysis we excluded 4050 SNPs with call rate <95%, 96 SNPs with Hardy-Weinberg equilibrium p-value $<10^{-6}$ and 4310 SNPs with minor allele frequency (MAF) <2% from ILMN100K (a total of 8313 SNPs excluded), and 3921 SNPs with call rate <95%, 34 SNPs with Hardy-Weinberg equilibrium p-value $<10^{-6}$ and 483 SNPs with MAF <2% from ILMN300K (a total of 4420 SNPs excluded).

The second sample of 951 individuals from 320 families of British descent was genotyped using the Illumina Sentrix HumanHap300 Genotyping Beadchip (Gunderson et al. 2005, Steemers et al. 2006). The genotyped sample consisted of 347 subjects with asthma and 487 subjects with atopic dermatitis (259 subjects with both diseases). Of the 314,552 SNPs with annotation available in the UCSC genome browser (hg18, Mar 2006), 8,345 with less than 95% genotyping success rate or deviating from Hardy-Weinberg ($P<10^{-6}$) were excluded. We retained 306,207 SNPs and 296,533,535 genotypes (99.1% call rate) for further analyses. There were only 0.204 mendelian errors per SNP: these genotypes were excluded from subsequent analyses. Expression arrays using Illumina Human 6 BeadChips were available on 550 children (atopic dermatitis probands and their siblings). Expression values were estimated using BeadStudio (Illumina, San Diego) and bead summary data were used for downstream analysis. From the total of 47,293 probes, we excluded 30,806 probes called as "absent" (detection score less than 0.95) in more than 80% arrays to eliminate noise. We retained 16,487 probes representing 15,576 genes for analysis. The data were then normalized using quantile normalization (Bolstad et al. 2003). We performed parallel analysis on both samples and observed similar results. Results from the first sample (Dixon et al. 2007) will be presented in the remaining sections.

An inverse normal transformation was applied on each transcript to avoid the effect of outliers. Briefly, the procedure involves first transforming all observations to ranks and then converting these ranks to deviates from a standard normal distribution. Narrow-sense heritability for each transcript was estimated by using a variance component model and a variance component based score test was used to evaluate the

evidence for association at each SNP (Chen and Abecasis, 2007). This variance component based association analysis results in an estimate of the additive genetic effect at each SNP and accounts for the correlation in phenotypes between siblings. Both procedures are implemented in MERLIN (Abecasis et al, 2002; Abecasis and Wigginton, 2005).

The analyses identify hundreds of loci that are strongly associated with mRNA expression levels. As in other studies of global gene expression, most of the strongly associated loci map in *cis* (typically within a megabase or less) of the transcripts they regulate. We reasoned that more powerful analyses should increase the number of *cis* association signals identified while maintaining overall false positive rates.

We used the ILMN300K genotypes to mimic the data that might be used in a typical genome-wide association study and to impute the polymorphic SNPs in the Phase II HapMap. The ILMN100K SNPs were not used for imputation or for our initial analysis, instead we used genotypes for markers in the ILMN100K panel that were also not present in the ILMN300K panel to assess the accuracy of imputed genotypes and of association analysis results. In this way, we were able to assess not only the accuracy of imputed genotypes but also to directly assess the impact on power of using genotyped or imputed SNPs.

We imputed genotypes for all polymorphic HapMap SNPs by using a hidden Markov model programmed in MACH (Li et al. 2008). The method combines genotypes from the 378 study samples with the HapMap CEU sample (July 2006 phased haplotype release) and identifies the stretches of haplotype shared between the study samples and the HapMap sample. For each individual, the genotype at the untyped SNP can be

summarized by taking (1) the most likely genotype according to the posterior probability of the three possible genotypes and (2) allele dosage, the expected number of copies of the reference allele (a fractional value between 0 and 2).

5.4 Results

Global gene expression

We took expression level at each probeset as an individual trait. Since many genes are represented by multiple probesets, we also performed parallel analysis where we took the average expression level across all probesets in the same gene. We found similar conclusions by using probeset and gene level data and report summaries based on probeset level data except when interpretation is helped by using gene level results.

The narrow sense heritability H^2 for all the expression levels after RMA and quantile normalization ranged between 0.0 – 1.0, with a mean of 0.203 and a 3rd quartile (Q3) of 0.317. We applied an arbitrary H^2 threshold of 0.3 to filter transcripts for downstream analyses (figure 5.1a). We did not apply a threshold filter for transcript abundance because we felt that genetic regulation of transcripts with low abundance might still occur and could be biologically relevant. Nevertheless, we note that the correlation between mean expression levels and heritability was substantial ($r = 0.45$, $p = 2.2*10^{-16}$).

Human EBVL provide general information about gene expression, even for genes whose primary function is not in these cells **(**Schadt et al. 2003, Yan et al. 2002, Cheung et al. 2003, Gretarsdottir et al. 2003**)**. Although the EBVL used in our analysis were derived from children with and without asthma, only 10 of 54,675 transcripts (~0.018%)

differed significantly between asthmatics and non-asthmatics with P < .0001, and no differences were significant after adjustment for the number of comparisons. This result is not unexpected as we measured expression in cultured, unchallenged cell lines; many of the changes in transcript abundances previously observed in asthma cells and tissues are the result of challenge with environmental and pro-inflammatory stimuli. We consequently expect our experiment to inform the genetics of gene-expression not only for studies of asthma, but more generally.

Genome-wide eQTL maps

Our 408,273 genotyped SNPs included 372,821 common SNPs (MAF>0.05) from the HapMap database. These covered 1,794,828 HapMap SNPs (including the 372,821) at R2 > 0.8, so that the total coverage of the 2,236,212 HapMap common SNPs was 80.3%. We tested for association between the SNPs and expression levels including sex in the model. Based on 100 randomly selected transcripts, the genomic control parameter for the 378 samples is 1.012 for experimental genotype and 1.002 for imputed genotypes (Dixon et al, 2007). We found that the 14819 traits with annotation entries in the UCSC browser and H^2 > 0.3 had a minimum peak LOD score for association of 3.683, and a maximum of 59.128 (median 4.853, Q3 5.339) (Figure 5.1b). We estimated the threshold for genome wide significance to be a LOD score > 6.076 (equivalent to P = 0.05 for Bonferroni correction of 408,273 SNPs). Accounting for all possible transcript-SNP pairs, we found the false discovery rate (FDR) for a LOD score of 5.5 to be 0.152, for a LOD of 6 to be 0.056, for a LOD of 7.0 to be 0.0067, and for a LOD of 8 to be 0.0008.

The mean H^2 explicable by association to the SNP showing strongest association to each trait was 0.077 (SD 0.049, max 0.707) compared to 0.429 for the overall H^2 (SD 0.103, max 1.0), indicating that on average the peak SNP accounts for 18.2% of the H^2 in these traits. For the 1,989 transcripts where the peak LOD was >6, the mean H^2 explicable by association to the SNP showing strongest association was 0.157 and the average overall H^2 was 0.479, indicating 32.9% of the H^2 in these traits can be explained by the peak SNP. The proportion of peak SNPs exceeding the LOD > 6 significance threshold rose with the H^2 of the underlying trait, so that 81% of traits with $H^2 > 0.8$ were associated with at least one SNPs with LOD > 6 (Figure 5.2).

Sample size and power

Previous studies have shown the power of eQTL mapping, but have examined limited numbers of transcripts or markers in a small number of CEPH pedigrees (Schadt et al. 2003, Morley et al. 2004, and Cheung et al. 2003). In order to investigate the impact of sample size, we repeated our analyses using only the first 50 sibships in our sample. We identified only 503 associations (for 106 transcripts) in this subset that exceed our threshold of 6 for genome-wide significance. Using 100 sibships we found 4,923 such associations (for 736 transcripts) and in our full data set of 206 sibships we found a total of 16,098 such associations (for 1,989 transcripts). These results clearly suggest that further increases in sample size will enable even more regulators of gene expression to be mapped with statistical confidence.

Dominance and interaction

We explored the heritability that was not explained by association by testing for dominance and interaction effects on association amongst the 13,095 transcripts with H2 > 0.3 that could not be mapped (maxLOD < 6) under the additive model. We identified 699 transcripts under a dominant model with P < 6.12x10-8 (Bonferroni correction for 2*408273 tests). This was however less than the 1,097 transcripts that we observed in simulated null genotype data, suggesting that in these subjects dominance had a minimal effect on gene transcription.

We further tested for interactions amongst the top 100 SNPs for each of the 13,095 transcripts with high heritability but no genome-wide significant SNP associations. We found 600 had a P < 6x10-8 for the interaction term (Bonferroni correction for 2*408273+10000 tests), compared to 219 in a permuted genome-wide association scan dataset. Although many of the interactions were between SNPs in the same chromosome (and could simply point to a haplotype effect), we observed an excess of interacting SNPs even after removing these. Thus, our data suggest that genetic interactions may have an important influence on regulation of expression for individual genes.

Cis and *trans* effects

Trans effects were weaker than those in *cis* (defined as a SNP within 100Kb up-stream and down-stream of a gene) and most LOD scores > 9 were in *cis*. (Figure 5.3) This is consistent with previous studies in humans (Schadt et al. 2003, Morley et al. 2004) and mice (Hubner et al. 2005). Despite the relative weakness of *trans* effects, numerous distant associations were observed (for example, the peak of association for

698 transcripts was on the same chromosome but >100Kb from the transcribed gene and for 10,382 transcripts the peak of association was on a different chromosome), and it may be anticipated that larger samples will define more precisely the extent of trans regulation of human transcripts.

Gene Ontology

We used Gene Ontology analyses to identify genes that were significantly enriched amongst highly heritable traits (Table 5.1). The most highly heritable GO biological process was "response to unfolded proteins". This group contained numerous chaperonins and heat shock proteins (*CRNN*, 7 *DNAJ* family members, *HERPUD1*, 16 *HSPA*, *B*, *C* or *D* family members, *SERPINH1*, *TOR1A* and *1B*, *TRA1* and *TXNDC4*)**.** The individual variation in response to unfolded proteins may represent an evolutionary response to cellular stress, and these genes could be candidates in the study of neurodegenerative diseases and aging processes.

Genes regulating progression through cell cycle, RNA processing, and DNA repair were also exceptionally heritable (Figure 5.4a). We speculate that expression of these genes is under very tight genetic control, with little stochastic noise, so that nearby polymorphisms can more easily influence expression in a detectable manner. The evolutionary advantage of individual variation in these genes is unclear. These genes may be relevant candidates for the investigation of inherited susceptibility to cancer.

It has been shown previously that genes expressed in EBVL are enriched in GO categories of immune response (Monks et al. 2004), and the significant heritability that we observed to these traits (Figure 5.4b) emphasizes the value of our data for the study of

infectious and inflammatory diseases. Genetic variation of the level of *HLA-DQ* expression has been observed previously (Beaty et al. 1995), but effects that we found on *HLA-DR* and *HLA-DP* are novel, as are smaller effects on *HLA-A* and *HLA-C* (Figure 5.4b). The strength of these effects suggests that associations of MHC class I and class II polymorphism with diseases may depend on the level of gene transcription as much as restriction of response to antigen.

eQTL database can help interpretation of GWAS

Our dataset has wide application to the study of genetic markers associated with disease or other biological phenotypes. We used the genome-wide SNP data to map a novel susceptibility locus for childhood asthma to non-coding SNPs residing within a 206 kb segment on chromosome 17q23 (Moffatt et al. 2007). Our expression database showed that transcripts from *ORMDL3*, one of the nineteen genes within and around this segment, were strongly ($P < 10\text{-}22$) and consistently positively associated to exactly the same SNPs showing association with childhood asthma. The correlation between the P values from the test statistics for association with asthma and *ORMDL3* expression for markers across the 206 kb segment was 0.67 ($P = 0.004$). These results focus attention on *ORMDL3* as a strong candidate gene in asthma, and illustrate how the combination of gene expression with genetic data can be much more powerful than differential gene expression alone in identifying candidate disease genes.

Our database has also been of use in the identification of a novel susceptibility locus from Crohn's disease on chromosome 5 (Libioulle et al. 2007). A GWA study had identified markers with a strong disease association within a 1.25 Mb gene desert.

Examination of our database showed that these markers are also associated with expression of *PTGER4*, a gene that resides on chromosome 5 outside of the 1.25 Mb segment. This led to the identification of *PTGER4* as the primary candidate gene for this disease susceptibility locus (Libioulle et al. 2007). Searching public GWAS results accumulated by NHGRI (http://www.genome.gov/gwastudies/), we identified many other disease associated SNPs that alter gene expression (Table 5.2).

Empirical analysis of genotype imputation

After QC filtering the ILMN300K SNPs (HWE $p<10^{-6}$, Mendelian error >5, <95% genotype completeness, MAF<2%, annotation availability in the University of California Santa Cruz genome browser), there remain 303,561 autosomal SNPs on the ILMN300K panel. We used 298,285 autosomal SNPs that presented on both the ILMN300K chip and HapMap, together with 2,557,252 polymorphic SNPs in the phased HapMap CEU chromosomes as input of the program MACH 1.0 (Li et al. 2008). We estimated the most likely genotype and the expected number of copies of the reference allele (allele dosage) at each of the 2.5M HapMap SNPs. To assess the quality of imputed genotypes, we compared the most likely genotypes with the genotypes obtained from the ILMN100K panel. To evaluate association analysis power, we found similar conclusions for using most likely genotypes or using allele dosage. The results in the later sections are based on the most likely genotypes.

Overall imputation accuracy.

After QC filtering the ILMN100K SNPs (HWE p<1e-6, Mendelian error >5, <95%

genotype completeness, Singleton in HapMap, Different comment allele in r21 and r23 of HapMap and tri-allele in HapMap and the ILMN100K genotype) and removing SNPs used for imputation or not presented in the HapMap panel, we used 58,819 autosomal SNPs from the ILMN100K panel for quality assessment. Genotypes from the ILMN100K panel were compared to the imputed genotypes at the same SNP. We observed an average genotype mismatch error rate of 0.033 (range from 0 to 0.818) with 80.85% SNPs having error rates < 0.05 and 35.27% SNPs having error rates < 0.01 (Figure 5.19).

Estimated Quality and R-square.

The MACH 1.0 program provides two useful measures to estimate the imputation accuracy at each marker. The first one is called "quality" which is the estimated probability of a correct genotype call. The second one is called "R-square" which is the estimated correlation between the imputed allele dosage and the actual allele counts. The R-square measure was suggested as a better measure for quality prediction and a threshold of 0.3 was suggested to filter SNPs for downstream analysis (Scott et al. 2007, Willer et al. 2008, Sanna et al. 2008, Li et al. 2008). On a per SNP basis, the estimated quality and R-square are strongly correlated with their actual values with correlations equal to 0.822 and 0.864 for quality and R-square respectively (figures 5.5 & 5.6). The estimated R-square was also strongly correlated with the actual error rate (correlation -0.683, figure 5.20).

Performance by Allele Frequency.

Allele frequencies from the imputed genotypes were very close to the actual values

(correlation 0.997, figure 5.7). SNPs with large difference between actual allele frequency and imputed allele frequency were associated with large error rates but substantial estimated R-square (104 SNPs have estimated R-square>0.3, figure 5.21). But these only account 0.18% of the total 58,819 SNPs. Table 5.3 and 5.4 categorize SNPs by minor allele frequency (MAF) and compared the estimated and actual values for error rate and R-square. We found that for common SNPs the estimated and actual values were closed to each other whereas for rare SNPs (MAF<1%) there is trend towards overestimating the R-square and underestimating the error rate (table 5.3 & 5.4). The error rates and R-square increase with minor allele frequency (figure 5.8, table 5.3 & 5.4). For SNPs that would be included in downstream analysis (estimated R-square>0.3), the error rates increase slightly with MAF but remain at low level while the R-square increases substantially (table 5.3).

Accuracy by LD.

The performance of imputation relied on the LD between the untyped SNPs and the SNPs used for imputation. For each of the 58,819 ILMN100K SNPs, we found the best tag-SNP from the 298,285 SNPs used for imputation and categorized the actual error rates and R-squares in Figure 5.9. The average best tag R-square is 0.83 (77.23% R-square>0.8 and 88.47% R-square>0.5). Even for mild to modest best tag R-square, the correlation between imputed and true allele counts were substantial (0.77 and 0.87 for best tag R-square in 0.1-0.5 and 0.5-0.8 respectively). This indicated the potential increase of power in imputation analysis compared to single marker tests using typed genotypes only.

Error Rate and Local Hotspot.

Finally we looked at the association between error rate and local recombination rate. We calculated the local recombination rate around each ILMN100K SNPs by summing the recombination rates between the SNP and its two flanking partners in HapMap. The error rates and local recombination rates were plotted along the genome (figure 5.10) and chromosome 10 (figure 5.11) as an example. The ILMN100K SNPs were evenly distributed along the genome and without apparent bias in error rates for different chromosomes. The error rates were associated with local recombination rates with modest correlation 0.463. Note the increase in error rates at the beginning and the end of each chromosome, where recombination rates are higher.

Overall, we observed high imputation accuracy and high correlation between the imputed and actual allele counts especially for common SNPs. The estimated R-square is shown to be a useful measure to predict imputation quality.

Reproducibility of Association Analysis.

We first evaluated the reproducibility of association analysis results using imputed genotypes. We removed the rare SNPs (MAF<2%) from the 58,819 ILMN100K SNPs to ensure the association analysis would not be affected by sparse categories. The 58,096 remaining SNPs were tested for association with the 15,084 transcripts that have more than 30% total heritability (Dixon et al. 2007). We then repeated the same analysis using the imputed genotypes on these SNPs and compare the LOD score for each transcript-SNP pair.

The estimated R-squares for these SNPs ranged from 0.0329 to 1 with mean 0.917

and 99.5% SNPs with estimated R-squares>0.3. This showed a high imputation quality. The LOD scores using imputed vs. observed genotypes were very close (correlation 0.952, figure 5.12). The higher the LOD scores the closer the imputation results to the results based on observed genotypes (table 5.5). As we would expect, the higher the estimated R-square, the more reliable the imputation results (table 5.6). Table 5.7 compared the results in terms of findings. We use a relative arbitrary threshold of LOD>6 to define significance. This is a threshold that could be used in a genome-wide association mapping for same among of traits (Dixon et al. 2007). We categorized the SNPs by different estimated R-square and counted the number of findings (significant transcripts/signals) in each category. The results also showed that SNPs with higher estimated R-square gave more reliable association results. When the estimated R-square > 0.3, imputed genotypes give similar findings as observed genotypes (table 5.7). Table 5.6 and 5.7 suggested that the threshold of estimated R-square > 0.3 is a reasonable choice to produce reliable association results.

Gain of Power.

Supported by the high imputation accuracy and reliability of association analysis based on imputed genotypes, we were able to evaluate the potential gain of power by using all imputed SNPs with good estimated accuracy, i.e. estimated R-square > 0.3. A total of 2,492,059 autosomal SNPs passed this threshold and were tested for association with the 15,084 transcripts. We used the 5% false discovery rate (FDR) for these 15,084*2,492,059 tests to a determine genome-wide significance threshold of LOD $\geq$ 6.222. Then we used the 303,561 ILMN300K SNPs to mimic the genotype data that

would be used in a genome-wide association study and tested for association with the same transcripts. The 5% FDR for these 15,084*303,561 tests required LOD $\geq$6.250.

Cis findings are usually to be stronger than *trans* and are more likely to be true signals (Dixon et al. 2007, Schadt et al. 2003, Morley et al. 2004). We used *cis* signals within the 1Mb window of the transcript to compare the performance of analysis using only the ILMN300K SNPs and analysis using also the imputed SNPs. Figure 5.13 shows the number of transcripts that can be mapped (the signal was within cis 1Mb and passed the 5% FDR threshold) by the ILMN300K panel and the imputed HapMap SNPs. In total 1,397 transcripts can be mapped. The majority of findings (88.48%) can be mapped by both observed and imputed genotypes due to the overlap between the ILMN300K and the HapMap panels and the optimized tagging of the ILMN300K SNPs to the HapMap phased I SNPs. For 5 of the 6 transcripts that can only be mapped by ILMN300K SNPs, the LOD scores of the same SNP using imputed genotypes were range from 5.795 to 6.04, just below the genome-wide significance level for that analysis. The remaining transcript (202086_at) was mapped by a SNP in the ILMN300K set (rs459498, LOD=11.4) but not in the HapMap panel. While mostly agreeing with the findings based on observed genotypes, imputation resulted in 11.1% (155 transcripts) more findings. There are 636 unique markers were significantly associated with these 155 transcripts. Among these SNPs, 23 SNPs were also typed in the ILMN100K panel only. For these, the correlation between LOD score by imputation and LOD score by the ILMN100K genotypes is 0.912.

Total Heritability, MAF and MaxLOD.

To further investigate the new findings by imputation, we plotted the distribution of

total heritability of these 155 transcripts. It shows that even highly heritable traits can be missed by a genome-wide scale chip (figure 5.14). With the same sample size and zero extra genotyping cost, imputation was able to map some of these missing signals. But note that only 1,397 out of the 15,084 transcripts were mapped. Sample size is still the more effective way to increase power than marker density (Dixon et al. 2007). The minor allele frequency of the new findings had a broad range (0 to 0.5). The largest category is the rare SNPs but common SNPs also take a substantial fraction (75.6% SNPs have MAF>5%, figure 5.15 and 92.9% top SNPs have MAF>5%, figure not shown). The strength of signals was modest for most new findings (figure 5.18). The majority max LOD is between 6.222 and 8. For some of the newly mapped transcripts, the max LOD scores obtained from the ILMN300K observed genotypes were just below the 5% FDR threshold. It suggests more samples are needed to obtain genome-wide significance. Still, some transcripts were mapped with strong signals only after imputation. For example, the transcript 219865_at (annotated to gene HSPC157) at 22.09Mb on chromosome 1 was associated to the top SNP rs2268177 with LOD 16.223 (figure 5.17 & 5.18). Interestingly, this SNP was also typed in the ILMN100K panel and the LOD score using the observed genotypes was 14.863. The small difference in the two LOD scores is probably due to the different way of imputations of missing genotypes used in MACH 1.0 and MERLIN. MACH 1.0 relies on a population reference panel while MERLIN relies only on the siblings of the individual being imputed.

False Discovery Rate.

One might concern about the false positive rate due to the increasing number of tests

on the imputed data. We addressed this question by comparing the *trans* signals while adjusting for the same number of *cis* signals. *Cis* findings are usually stronger and more likely to be true than *trans* findings. For a given number of *cis* findings (number of transcripts that can be mapped), the number of *trans* findings should be similar for testing procedures with similar false discovery rate. Table 5.8 shows that imputation gave similar number of *trans* findings compared to the observed genotypes when fixing the number of *cis* findings. It suggests that genotype imputation does not increase the false discovery rate even though more than 2M tests are performed for each gene expression trait.

In summary, we observed that genotype imputation could increase the power by more than 11% and maintain similar false discovery rates. The newly mapped transcripts have substantial heritability. The newly mapped eQTLs have broad allele frequency spectrum and most are modest signals.

Association Analysis with Additional Phenotyped Samples.

A total of 405 siblings were measured gene expression in our sample. Although only 378 of these individuals were typed using the ILMN300K panel and thus have the imputed genotypes, the additional siblings with expression values can be included in the analysis and their missing genotypes can be inferred probabilistically using the pedigree information (Chen & Abecasis 2007). Including these additional phenotypes lead to 33 more transcripts mapped in *cis* (<1Mb) at genome-wide significant level (5% FDR). In total, we observed 785 additional *cis* genome-wide significant signals compared to the database of Dixon et al. 2007. All association analysis results using imputed genotypes with the additional phenotypes can be browsed at our website

(http://www.sph.umich.edu/csg/liang/imputation/).

5.5 Discussion

We have systematically mapped eQTLs on the genome for global gene expression measured on two independent samples. The resulting eQTL maps provide much information about biological control of gene expression and may be used as general tools to investigate if disease associated SNPs alter gene expression in *cis* or *trans*. We have developed a database browser (the MRBS browser) that can be downloaded from http://www.sph.umich.edu/csg/liang/asthma/ for the interrogation of our data.

We find that the two commercial platforms used: Affymetrix array and Illumina BeadChip, provide complementary information. Among the 9487 significant *cis* associations ($p<10^{-7}$, SNP within 1Mb of gene) identified using Illumina expression dataset, 1460 (15.4%) were also identified in the Affymetrix expression dataset with similar significance cut-off. For a lower threshold $p<0.001$ for replication, this overlap increases to 2543 (26.8%). The difference could be due to a variety of potential reasons, including different designs of the two commercial chips, different length of the probes, different locations of probes in the gene, the complexity of microarray hybridization, heterogeneity in the samples, or power to replication.

Recent human studies of eQTL have been primarily focused on LCL because LCL were usually obtained as a source of nucleic acids for genetic studies. It has been shown that LCL may also carry information about gene expression for genes even if their primary function is not in these cells (Schadt et al. 2003, Yan et al. 2002, Cheung et al.

2003 and Gretarsdottir et al. 2003). The convenience and utility of LCL will be continued, however, eQTL studies should include RNA obtained from a variety of tissues. Overlap eQTL identified from multiple tissues may represent regulation of housekeeping genes while tissue specific regulation might only be found in the corresponding tissues. Combining disease association mapping with eQTL maps from related tissues may increase the power to identify disease relevant pathways.

As genotype imputation becomes increasing popular (Willer et al. 2008, Sanna et al. 2008, Scott et al. 2007, The Wellcome Trust Case Control Consortium 2007), it is important to know what would be expected to gain from imputation. We systematically evaluated the genotype imputation accuracy and potential gain of power using genome-wide scale dataset. Our findings suggested that imputation achieve high quality and can well predict the accuracy of each marker being imputed. The estimated R-square measure is important to filter poorly imputed SNPs for downstream analysis and will facilitate meta-analysis across studies. Imputation increases the marker density and coverage to the genome (Li et al. 2008) thus increased power of detection by more than 11% for gene expression traits. We expected that similar gain of power could be observed in traits have similar heritability with similar sample size.

Although in this paper the most likely genotypes were used to evaluate power in association analysis, we recommend using imputed allele dosage for association analysis whenever appropriate and the analysis tools support this format of data. The imputed allele dosage takes into account the uncertainty of each possible genotype and avoids the cumbersome handling of multiple imputations for each genotype. However, in the current practice there are situations that most likely genotypes might be preferred (because of the

limitations of existing tools; a more appropriate alternative is to perform full likelihood inference). For example, if some completely untyped relatives of a family have their phenotypic traits available as in our sample 405 siblings were measured gene expressions but only 378 were available on the imputed HapMap panel, these additional individuals could also be included in analysis and their missing genotypes could be inferred probabilistically according to the pedigree structure (Chen & Abecasis 2007) and hence potentially increase the power. In this paper we restricted all analysis for imputation accuracy and gain of power to the 378 typed individuals to ensure fair comparisons.

Our study also provides valuable additional results to the database developed by Dixon et al. These results not only include newly mapped transcripts but also more and stronger eQTLs for the transcripts that were mapped by the ILMN300K and ILMN100K SNPs. In total, 10,384,399 additional association signals with pvalue<0.001 were added to the database. The results have been incorporated in a web-based browser and can be freely accessed at http://www.sph.umich.edu/csg/liang/imputation/.

5.6 Tables and figures

Table 5.1 Gene Ontology of exceptionally heritable and non-heritable traits

GO-Biological Process	**GO ID**	H^2	**Z for** H^2
response to unfolded protein	6986	0.38	9.03
regulation of progression through cell cycle	74	0.26	8.20
RNA processing	6396	0.30	7.85
DNA repair	6281	0.29	7.81
protein folding	6457	0.30	7.80
immune response	6955	0.26	7.62
regulation of I-kappaB kinase/NF-kappaB cascade	43123	0.28	6.84
mitosis	7067	0.30	5.82
intracellular signaling cascade	7242	0.26	5.72
adenylate cyclase activation	7190	0.11	-3.55
sodium ion transport	6814	0.13	-3.63
phospholipase C activation	7202	0.12	-3.74
potassium ion transport	6813	0.14	-4.43
glutamate signaling pathway	7215	0.08	-4.52
synaptic transmission	7268	0.16	-5.64

The analysis compared the mean total H^2 of transcripts in an individual GO category with the mean total H^2 of all 54675 transcripts. Positive z-score indicates exceptionally heritable traits and negative z-score indicate non-heritable traits

Table 5.2 Disease-linked associations with significant expression quantitative loci from the literature and public databases

Study	Trait	Region	Candidate Gene	Transcript affected by SNP	Transcript Region	LOD
Gudbjartsson *et al* [84]	Height	7p22	*GNA12*	*GNA12*	7p22	13
		11q13.2	*Intergenic*	*CCND1*	11q13	7.4
		7q21.3	*LMTK2*	*C17orf37*	17q21	6.0
				HSD17B8	6	6.4
				NDUFS8	11	6.1
		3p14.3	*PXK*	*RPP14*	3	9.2
Libioulle *et al* [37]	Crohn's Disease	5p13	*Intergenic*	*PTGER4*	5p13	3.0
Hom *et al* [85]	SLE	8p23.1	*C8orf13, BLK*	*BLK*	8p23.1	20
				C8orf13		28
Harkonason e*t al* [86]	T1D	12q13	*RAB5B, SUOX, IKZF4*	*RPS26*	12q13	33
		1p31.3	*ANGPTL3*	*DOCK7*	1p31.3	16
WTCCC [87]	T1D	12q13.2	*ERBB3*	*RPS26*	12q13.2	43.2
Todd *et al* [88]	T1D	12q13.2	*ERBB3*	*RPS26*	12q13.2	30.3
Plenge *et al* [89]	Rheumatoid arthritis	9q34	*TRAF1-C5*	*LOC253039*	9q34	6.3
Moffatt *et al* [30]	Childhood asthma	17q21	*Intergenic*	*ORMDL3*	17	14
WTCCC [87]	Bipolar disorder	16p12	*PALB2, NDUFAB1, DCTN5*	*DCTN5*	16p12	9.2
		6p21	*NR*	*HLA-DQB1* *HLA-DRB4*	6p21	8.9 11
Di Bernardo *et al* [91]	Chronic lymphatic leukaemia	2q37	*SP140*	*SP140*	2q37	8.8

Table 5.3 SNPs selected for downstream analysis (estimated Rsq>0.3)

		Error Rate mean, range, sd		**R-square** mean, range, sd	
MAF	**N**	**Estimate**	**Actual**	**Estimate**	**Actual**
<1%	179	0.007	0.012	0.659	0.44
		0-0.045	0-0.065	0.304-0.995	0-1
		0.007	0.011	0.186	0.326
1-3%	960	0.011	0.018	0.752	0.629
		0-0.069	0-0.093	0.303-1	0-1
		0.01	0.014	0.183	0.269
3-5%	1233	0.015	0.024	0.816	0.741
		0-0.088	0-0.101	0.304-1	0.003-1
		0.014	0.021	0.17	0.232
5-10%	3725	0.017	0.024	0.888	0.855
		0-0.167	0-0.26	0.301-1	0.004-1
		0.021	0.03	0.137	0.18
10-20%	10469	0.025	0.027	0.919	0.912
		0-0.309	0-0.369	0.3-1	0.004-1
		0.033	0.039	0.107	0.123
>20%	41808	0.039	0.035	0.928	0.933
		0-0.414	0-0.818	0.302-1	0.002-1
		0.053	0.054	0.097	0.099
Total	58374	0.034	0.032	0.918	0.913
		0-0.414	0-0.818	0.3-1	0-1
		0.048	0.05	0.111	0.132

***** In each MAF category, the mean, range and standard deviation of each measure are listed at the 1st, 2nd and 3rd row respectively.

Table 5.4 SNPs not selected for downstream analysis (estimated R-square<=0.3)

		Error Rate mean, range, sd		**R-square** mean, range, sd	
MAF	**N**	**Estimate**	**Actual**	**Estimate**	**Actual**
<1%	77	0.015	0.013	0.144	0.102
		0-0.099	0-0.042	0-0.291	0-0.815
		0.018	0.008	0.08	0.19
1-3%	99	0.023	0.041	0.172	0.108
		0-0.433	0.013-0.574	0.003-0.298	0-0.798
		0.044	0.055	0.075	0.165
3-5%	62	0.03	0.072	0.195	0.113
		0.003-0.11	0.045-0.103	0.049-0.293	0-0.419
		0.024	0.013	0.065	0.094
5-10%	52	0.082	0.14	0.225	0.147
		0.01-0.24	0.077-0.239	0.075-0.3	0.002-0.521
		0.057	0.037	0.053	0.126
10-20%	45	0.181	0.257	0.223	0.146
		0.003-0.3	0.109-0.422	0.033-0.298	0.002-0.795
		0.101	0.055	0.066	0.132
>20%	110	0.349	0.46	0.238	0.14
		0.012-0.467	0.26-0.747	0.094-0.3	0.003-0.509
		0.123	0.086	0.046	0.105
Total	445	0.126	0.177	0.198	0.124
		0-0.467	0-0.747	0-0.3	0-0.815
		0.156	0.184	0.073	0.142

***** In each MAF category, the mean, range and standard deviation of each measure are listed at the 1st, 2nd and 3rd row respectively.

Table 5.5 Difference between LOD_Imp and LOD_100K by LOD_100K

LOD_100K	N	Correlation	Mean |LODImp – LOD100K|/LOD100K
LOD<3	401082	0.434	0.061
LOD 3-6	142004	0.710	0.083
LOD 6-10	2146	0.810	0.067
LOD 10-20	1290	0.904	0.057
LOD>20	388	0.963	0.038

* LOD_Imp and LOD_100K represent the LOD scores using imputed and observed genotypes respectively.

Table 5.6 Difference between LOD_Imp and LOD_100K by Rsq

Estimated R-square	N	Mean |LODimp – LOD100K|/LOD100K
Rsq<0.3	27	0.156
Rsq 0.3-0.5	534	0.142
Rsq 0.5-0.8	23297	0.126
Rsq 0.8-0.9	58069	0.103
Rsq>0.9	464983	0.059

* LOD_Imp and LOD_100K represent the LOD scores using imputed and observed genotypes respectively.

Table 5.7 Number of significant traits and signals by imputation and observed genotypes

	Significant Transcripts		Significant Signals	
	Imputation	Observed	Imputation	Observed
Rsq<0.3	1	8	1	8
Rsq 0.3-0.5	10	13	12	14
Rsq 0.5-0.8	100	113	119	137
Rsq>0.8	927	936	3694	3696
All	953	980	3826	3855
Rsq>0.3	953	975	3825	3847
Rsq>0.5	951	971	3813	3833
Rsq>0.8	927	936	3694	3696

* Significant is defined by an arbitrary threshold LOD>6.

Table 5.8 Number of *trans* signals from observed genotypes and imputed data while adjusting for the same number of *cis* signals

	Counts of Associations (peak association for each probeset)				
ILMN300K LOD Limit	40.557	24.626	19.413	12.347	6.485
cis	10	100	200	500	1200
Chr	1	4	7	11	26
trans	0	1	1	4	87
NA	0	2	6	11	25
Imputation LOD Limit	43.011	26.563	20.726	13.118	7.157
cis	10	100	200	500	1200
Chr	1	4	5	11	24
trans	0	1	1	3	47
NA	0	2	7	11	25

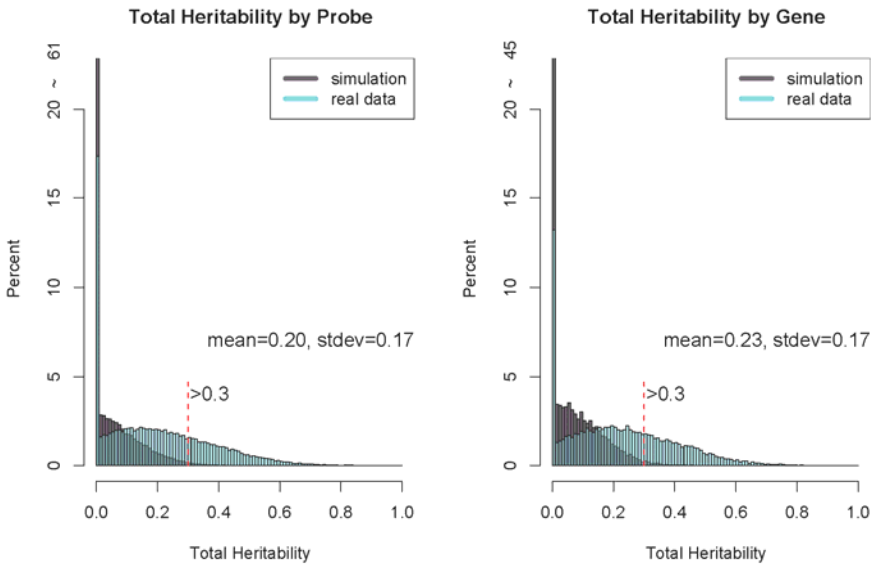
Suppose *cis* findings are more likely to be true than *trans* findings, this table suggests imputed data give a similar false positive rate as the observed genotypes.

LOD Limit: the LOD score cut-off to give the corresponding number of *cis* signals. This cut-off is then used to determine the number signals in the remaining categories.

cis: SNP within 1Mb of the probeset.
Chr: SNP on the same chromosome as the probeset but outside the 1Mb window.
trans: SNP on different chromosome as the probeset.
NA: position not available for SNP or probeset.

Figure 5.1 Total heritability and peak association of transcripts

a Total heritability of expression quantitative traits



b Distribution of lod scores for association between 14819 traits with annotation entries in the UCSC browser and $H^2 > 0.3$, and 408,273 SNP markers

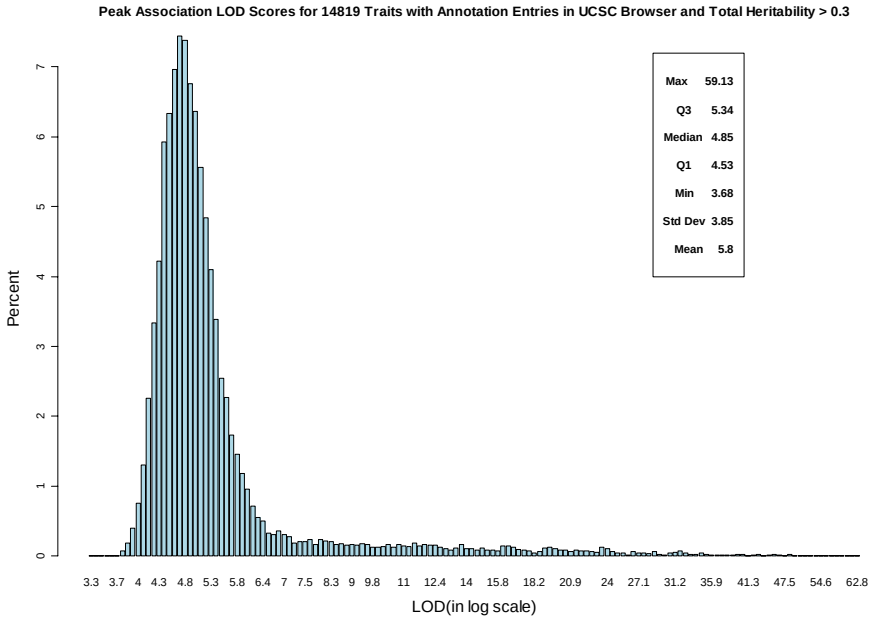


Figure 5.2 Proportion of significantly associated SNPs and expression trait heritability

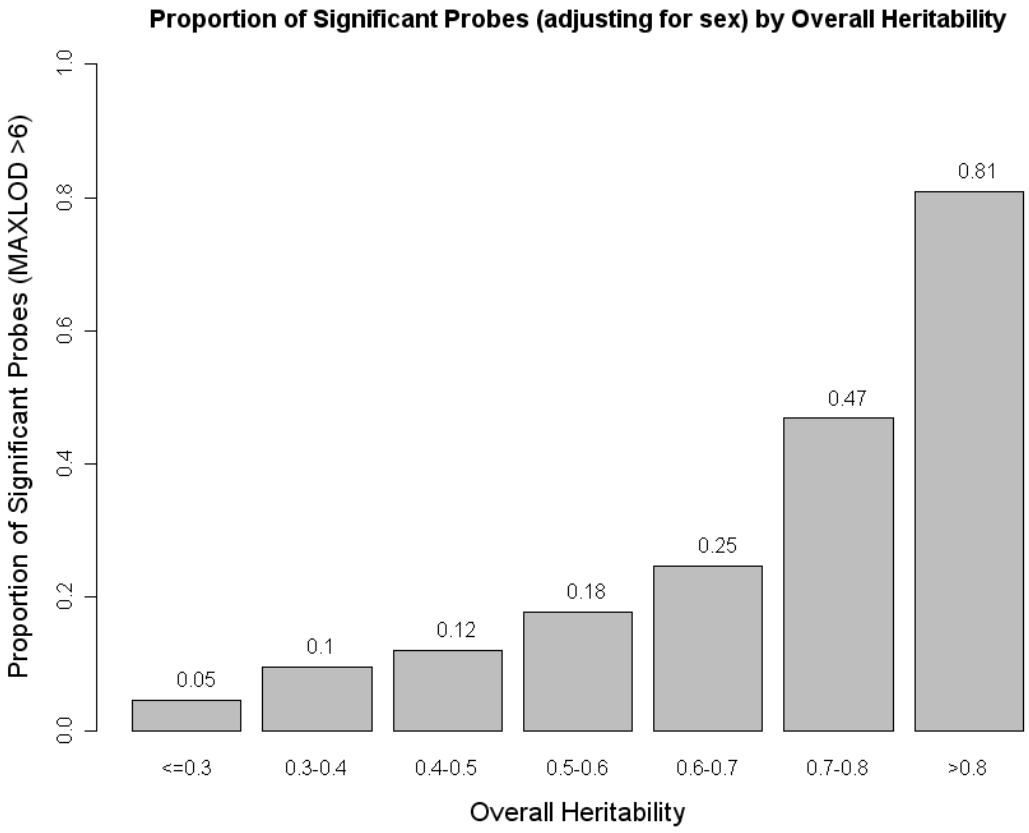
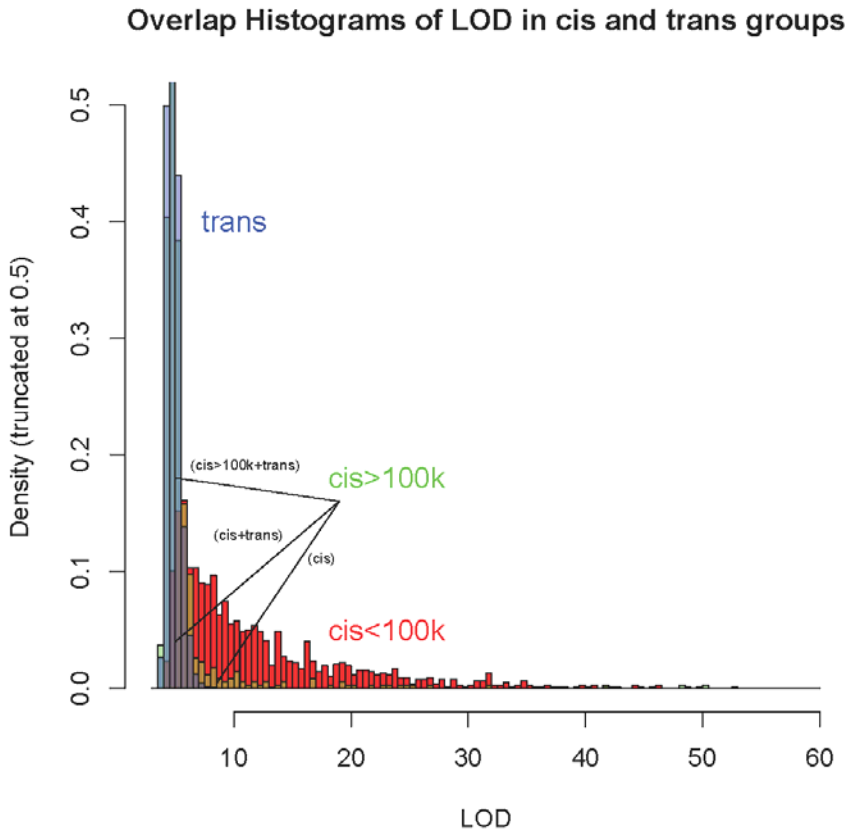


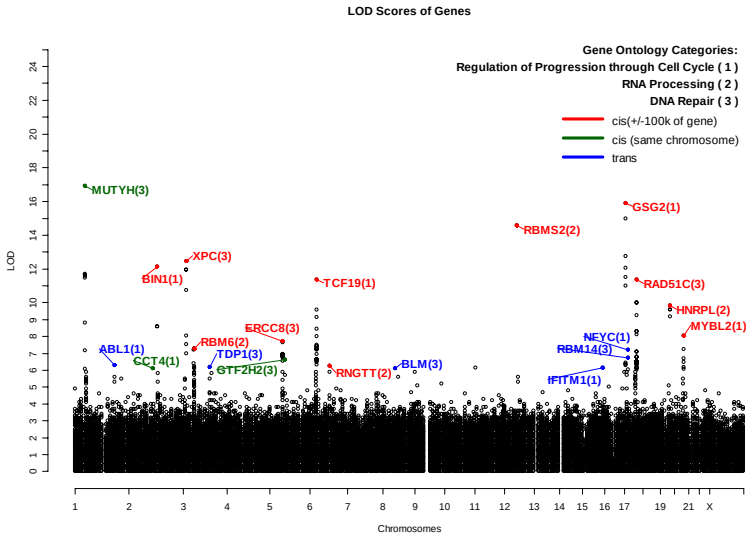
Figure 5.3 Associations in *cis* and *trans*



The density scale on the y axis is truncated at 0.5. Loci in *cis* <100Kb from the start of transcription are shown in red, loci in *cis* >100Kb from the start of transcription are shown in green, and loci in *trans* are shown in blue. (The overlap of the *cis* > 100Kb with the other distributions appears orange and grey).

Figure 5.4 Mapping of genes in highly heritable GO categories

a. Mapping of genes with GO-BP descriptors for cell cycle, DNA repair and RNA processing



b. Mapping of genes with GO-DP descriptors for immune responses

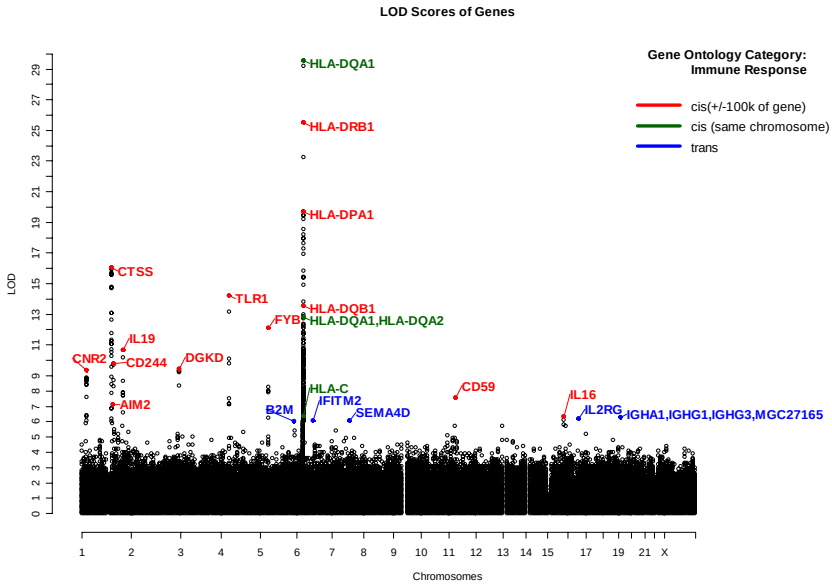
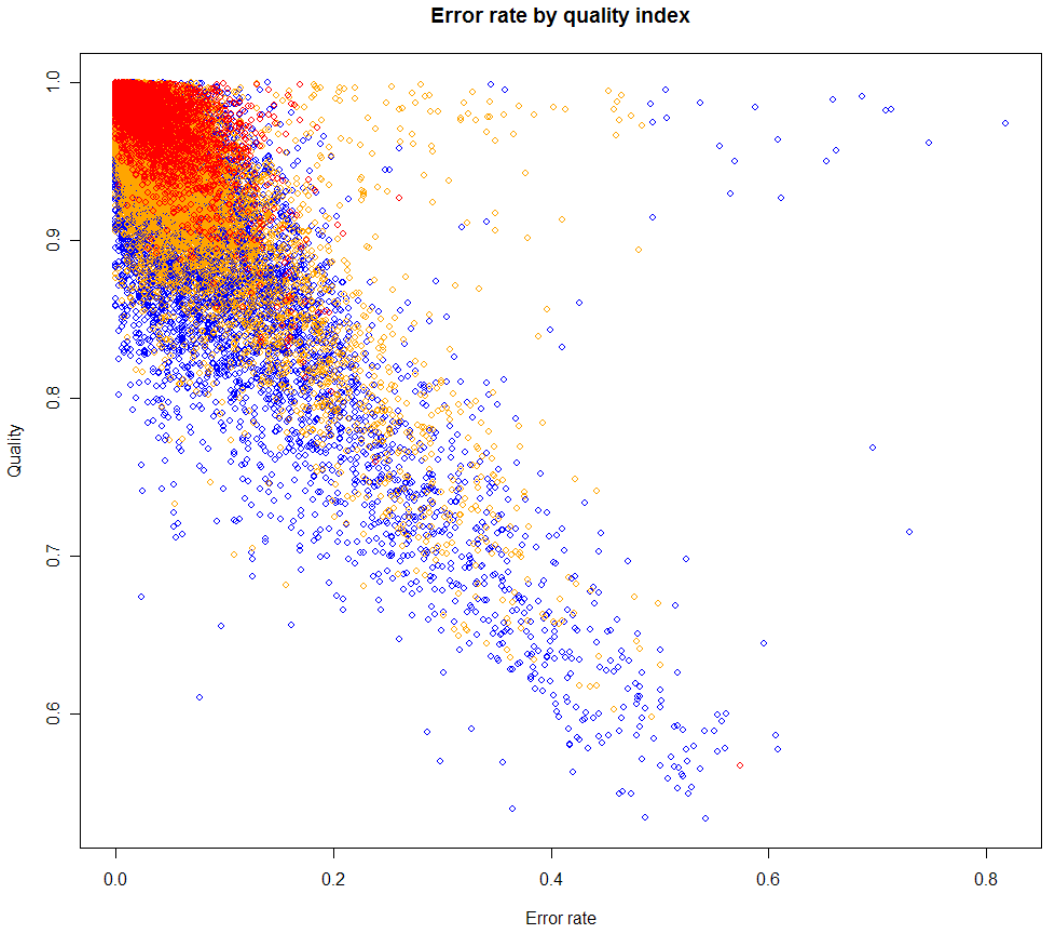


Figure 5.5 Estimated quality and the actual genotype mismatch error rate



Red: MAF<10%, Orange: MAF between 10% and 30%, Blue: MAF>30%

Figure 5.6 Estimated R-square and its actual value

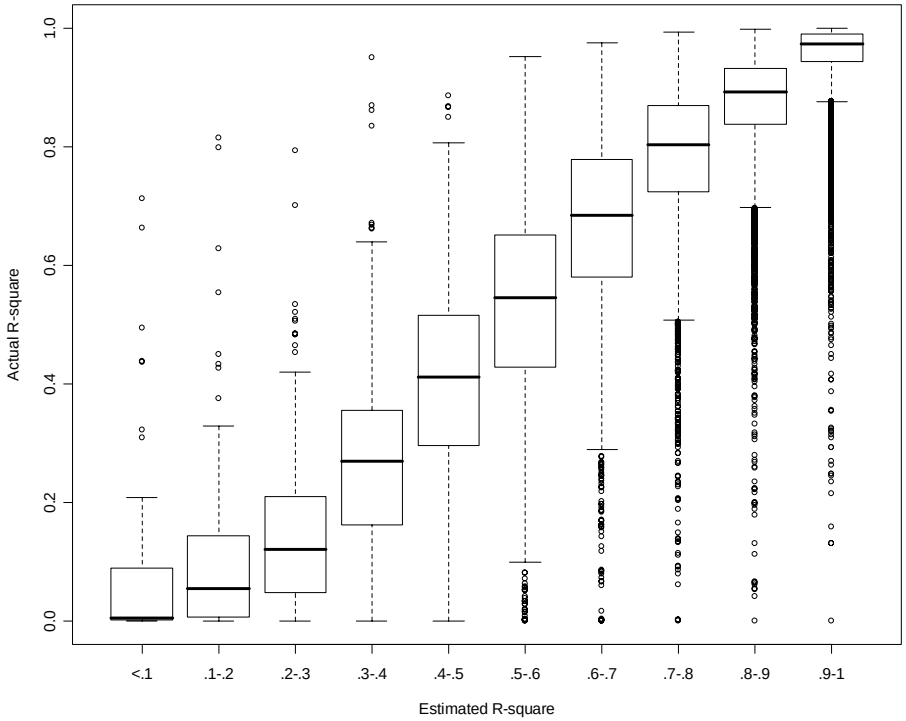


Figure 5.7 Allele frequency of imputed and actual genotypes

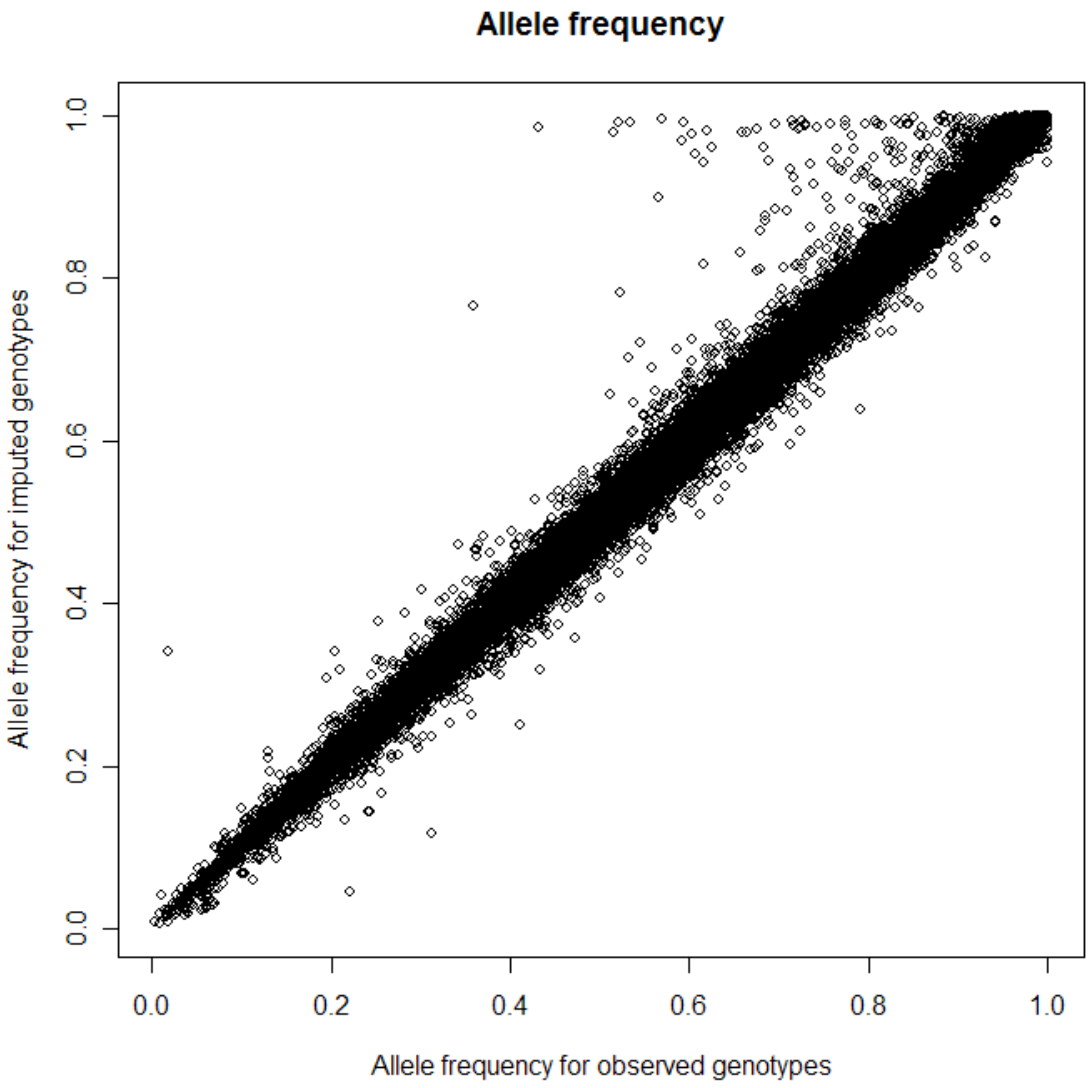


Figure 5.8 Minor allele frequency and mismatch error rate

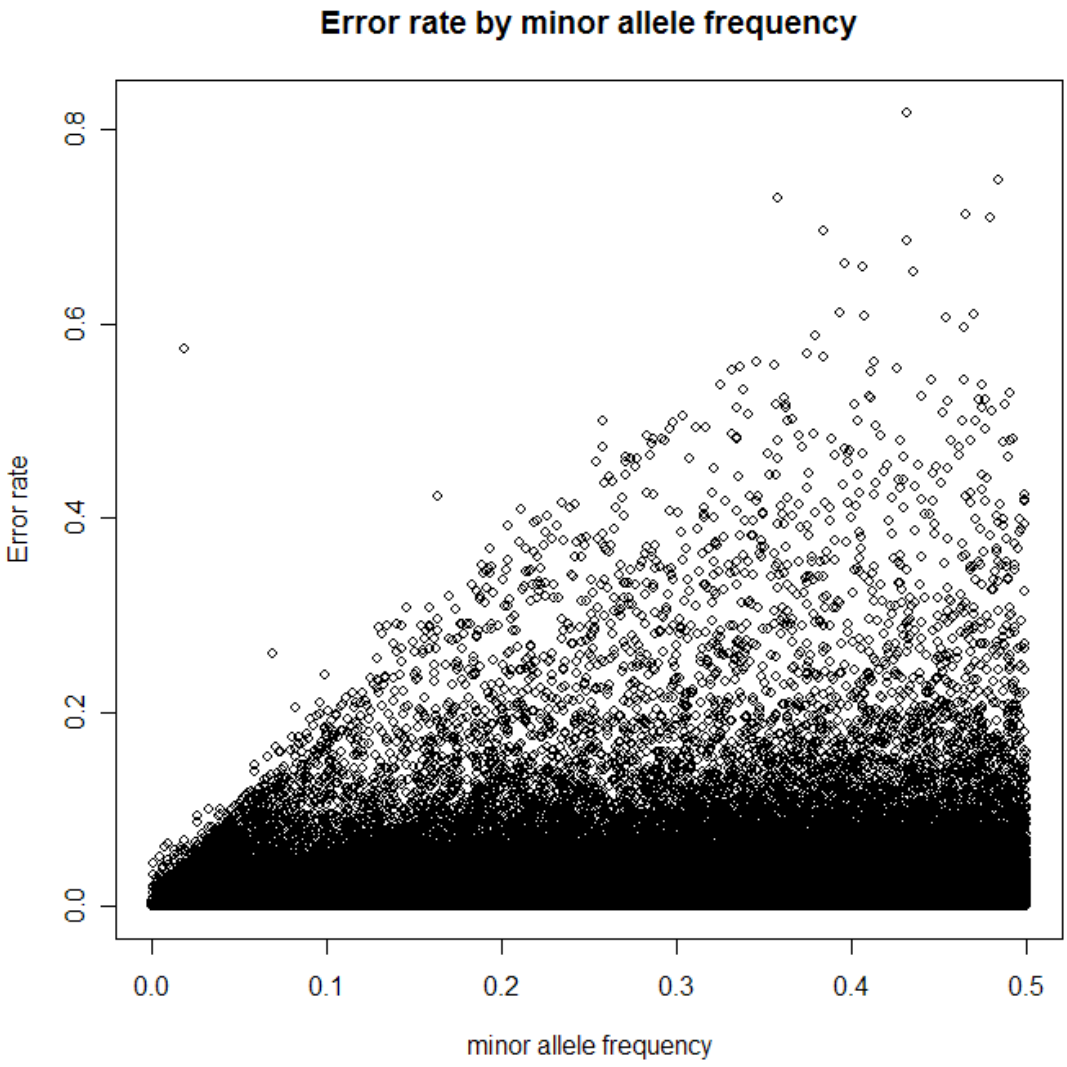
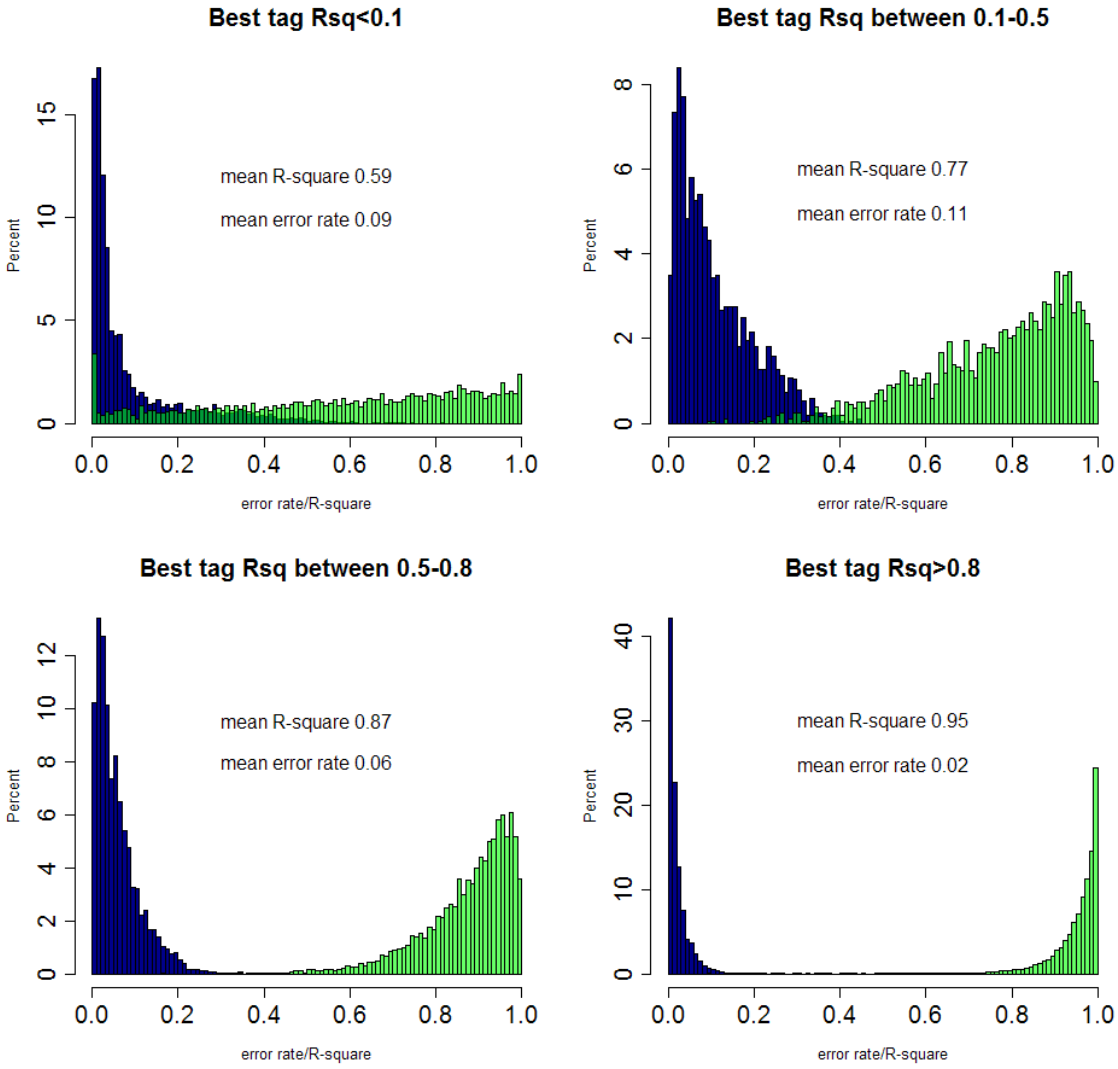


Figure 5.9 Actual error rate and R-square by best tagging R-square



* The green(right) histogram is for R-square and the blue(left) histogram is for error rate.

Figure 5.10 Error rate and local recombination rate along the genome

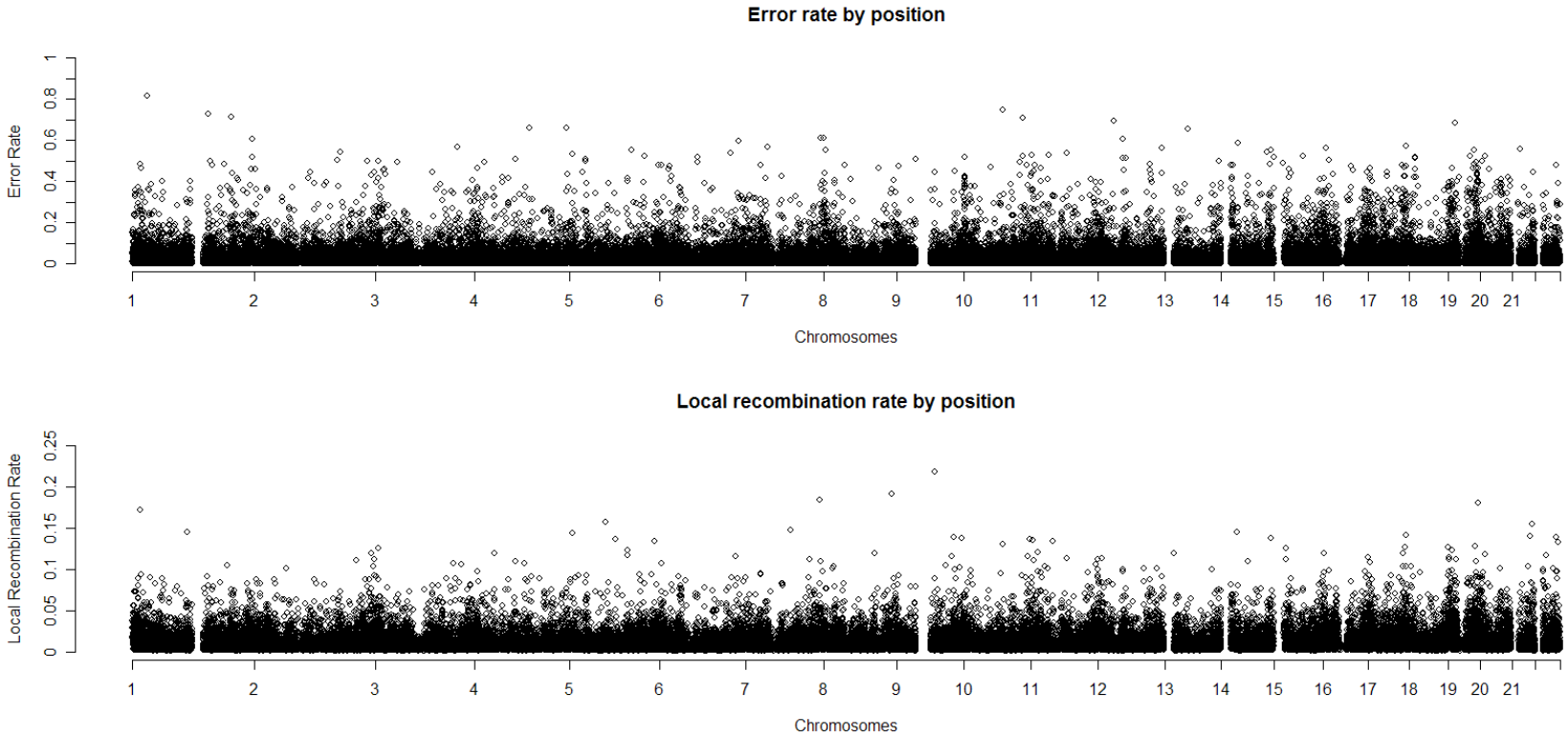


Figure 5.11 Error rate and local recombination rate on chromosome 10

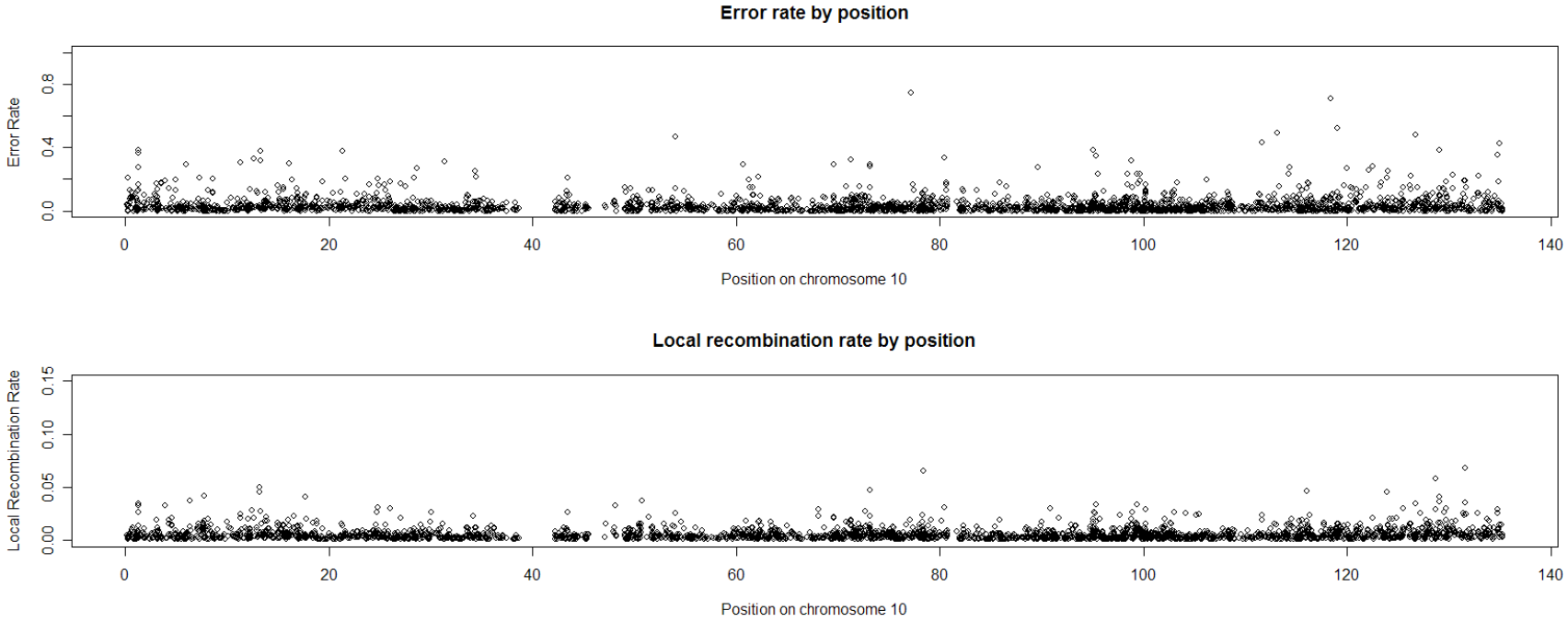


Figure 5.12 Association analysis using imputed vs. observed genotypes

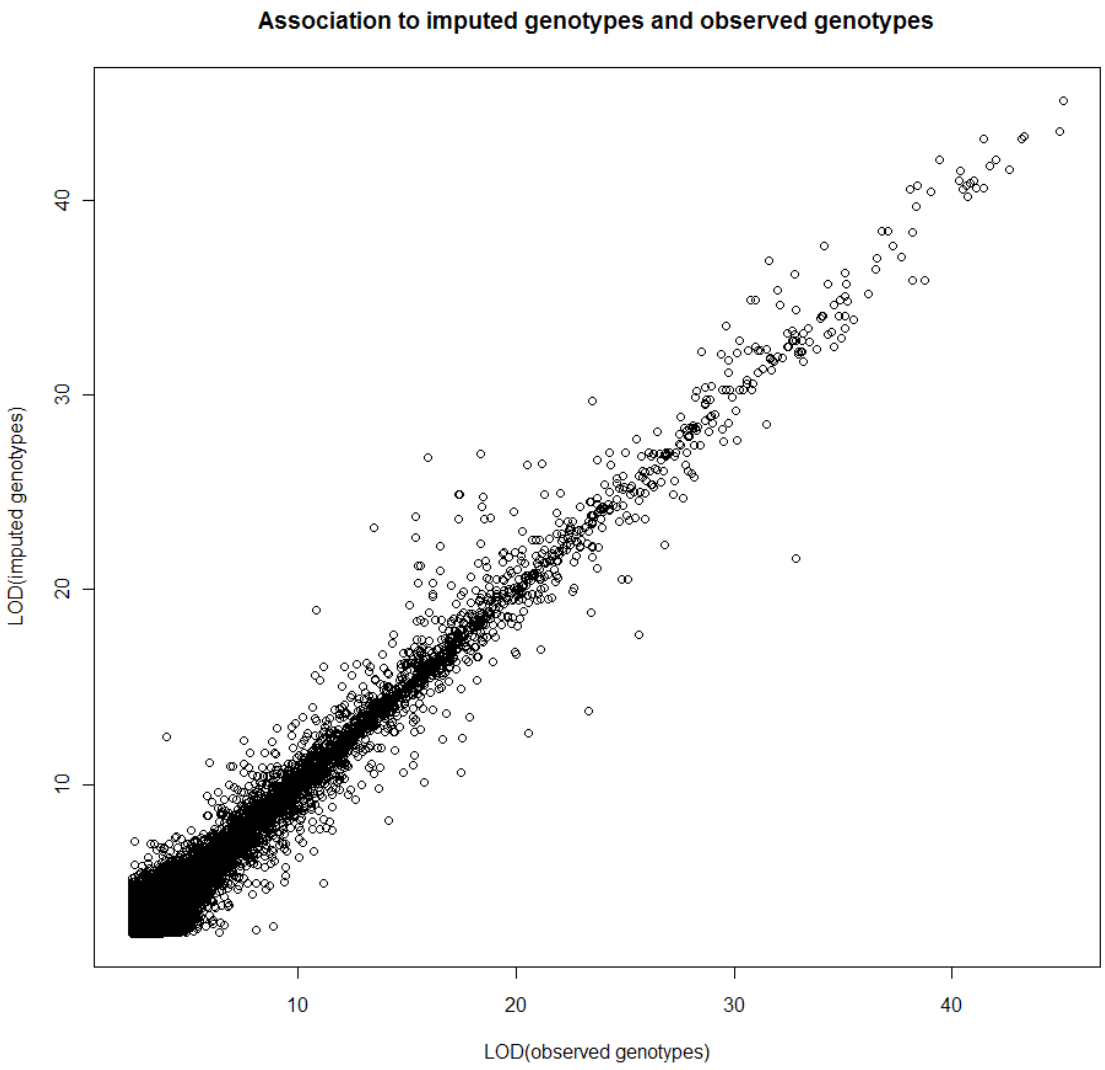
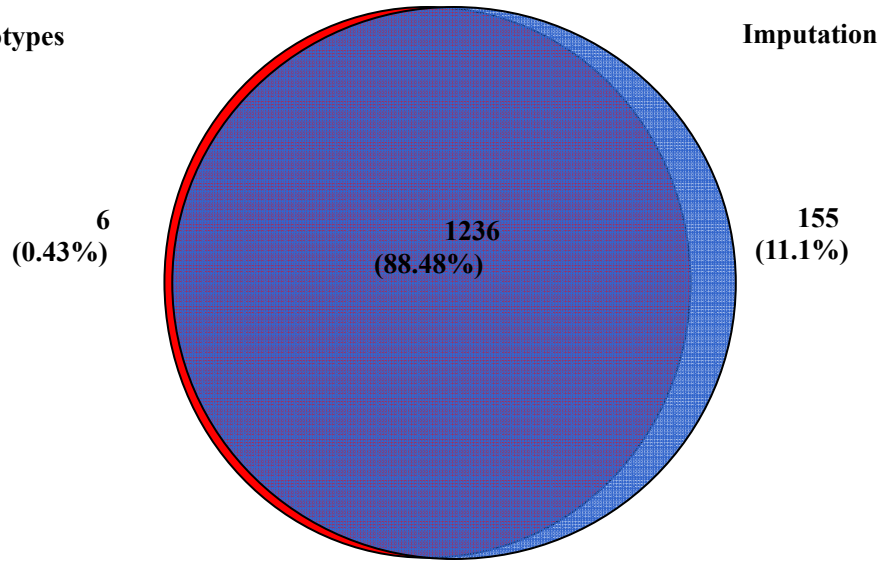


Figure 5.13 Venn diagram for the overlaps of findings between association analysis based on observed genotypes (300K), imputed HapMap SNPs



The number of transcripts that can be mapped in each category is shown in the Venn diagram.

Figure 5.14 Missing heritability mapped by imputation

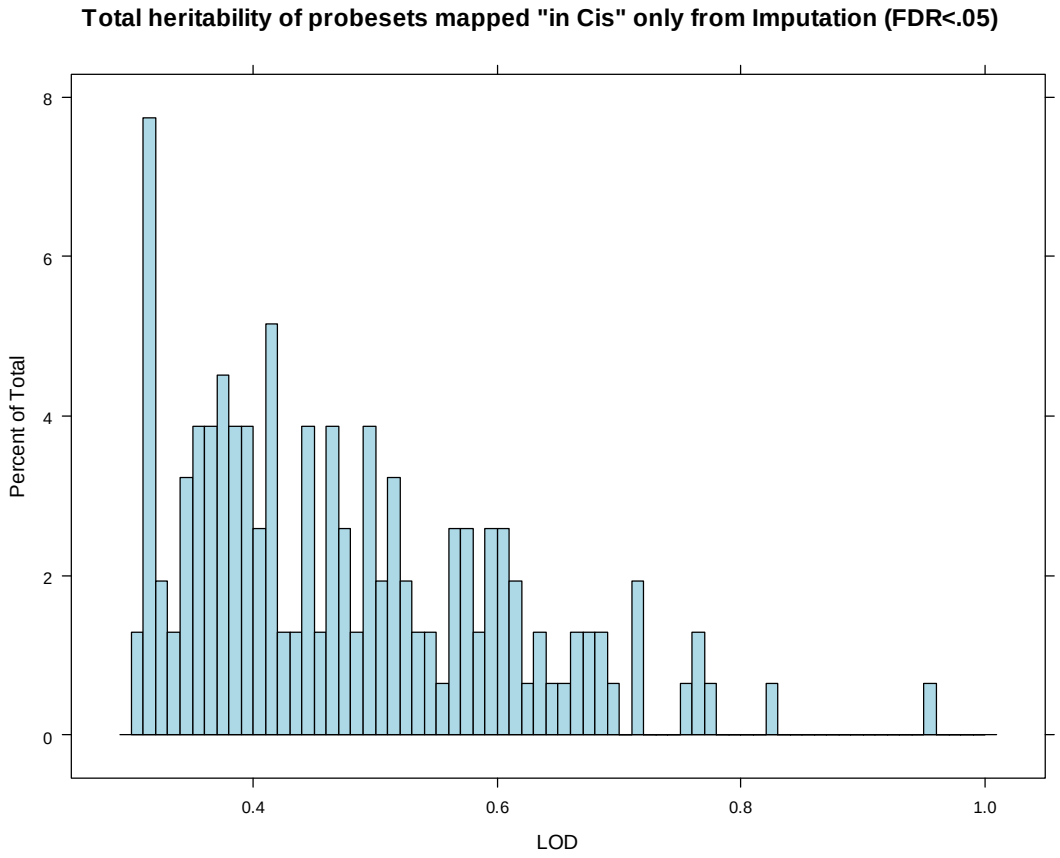


Figure 5.15 Allele frequency of eQTL mapped only from imputation

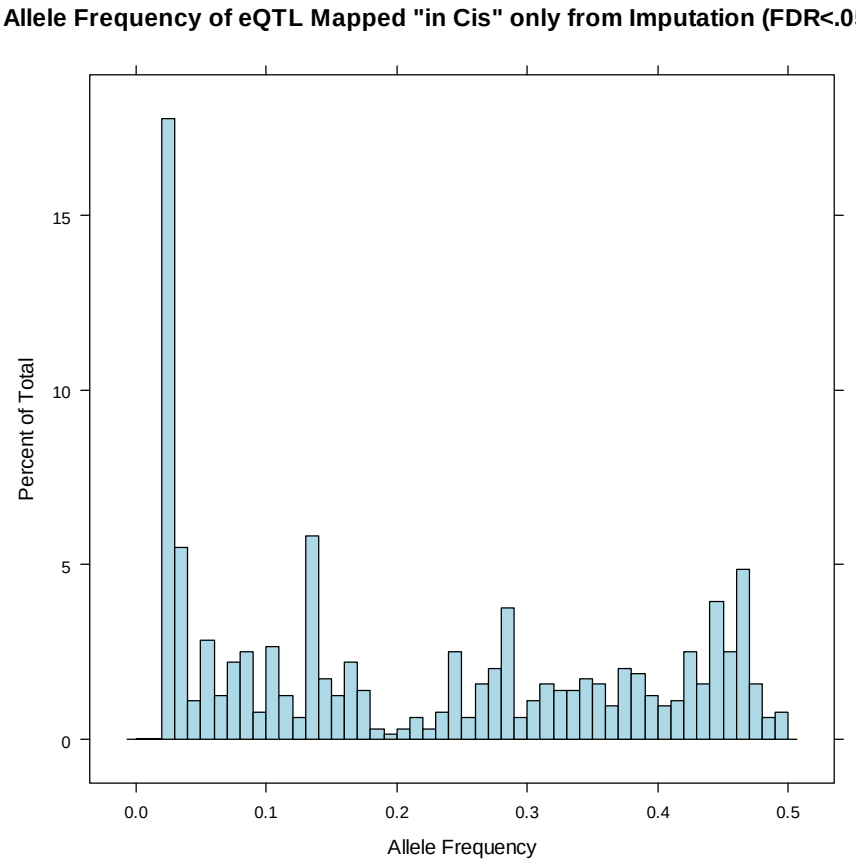


Figure 5.16 Max LOD of transcripts mapped only from imputation

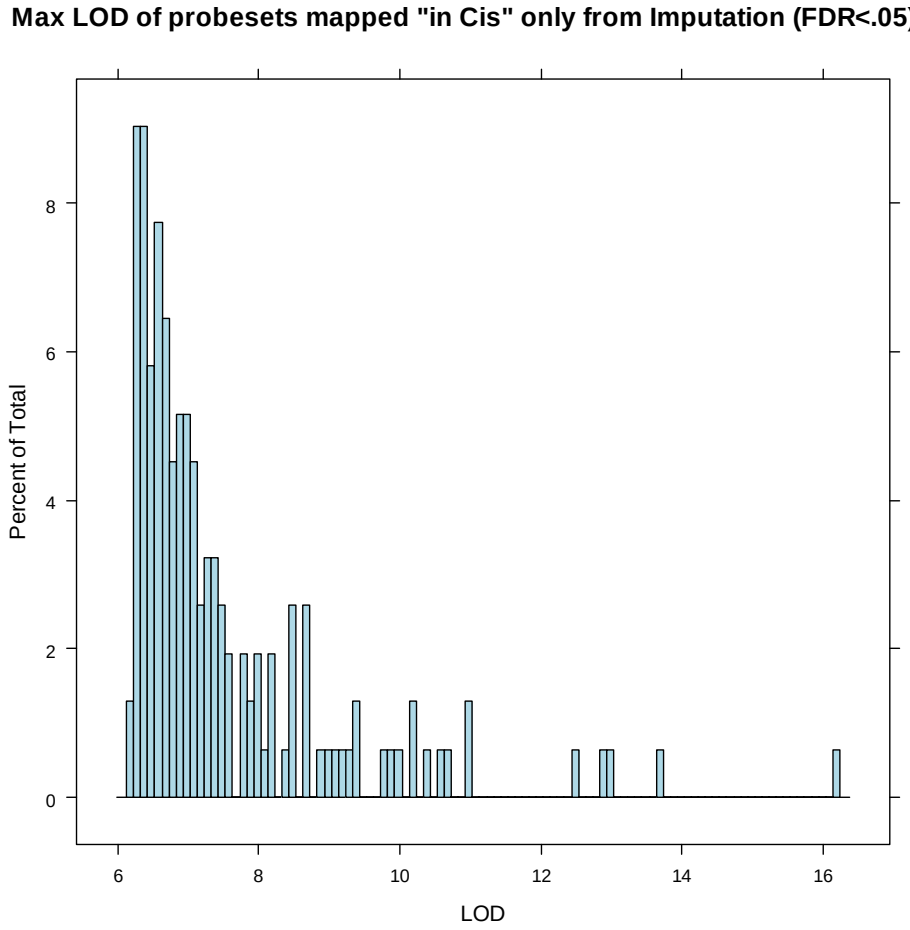


Figure 5.17 Association to the transcript 219865_at along the genome by different genotype panels

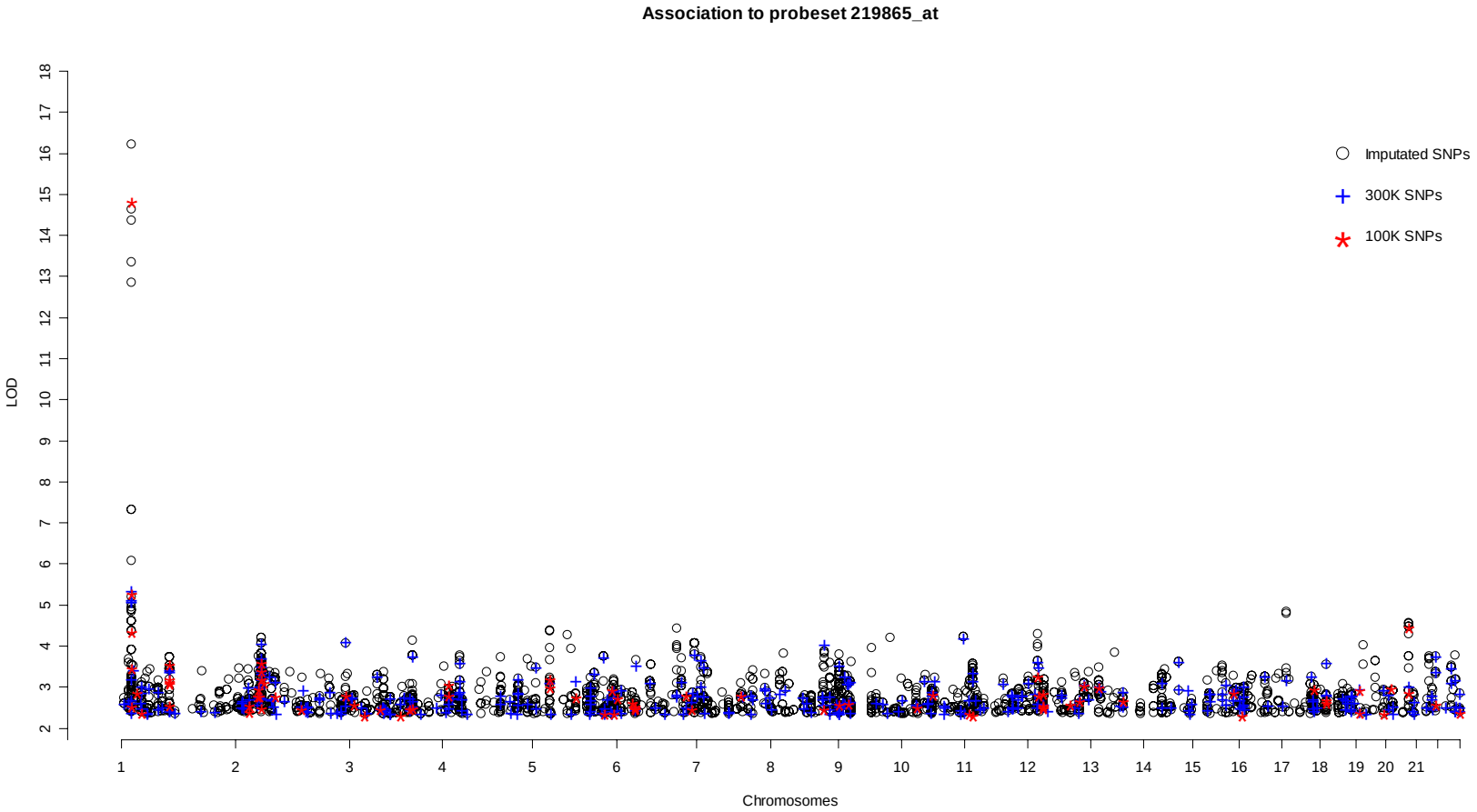


Figure 5.18 Association to transcript 219865_at on chromosome 1 region by different genotype panel

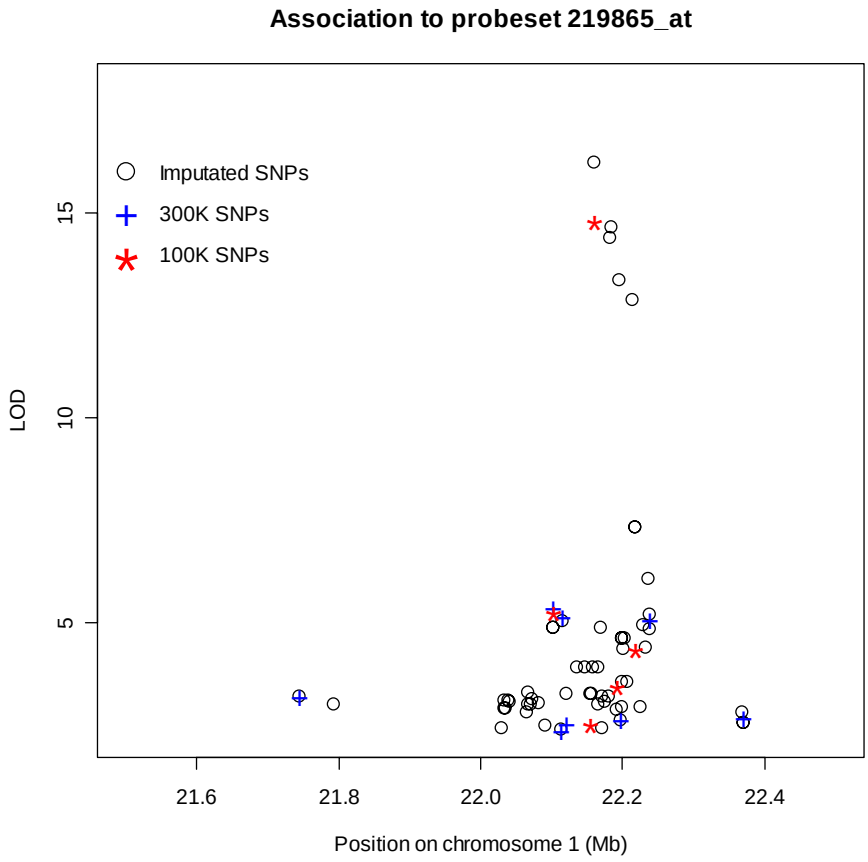


Figure 5.19 Distribution of genotype mismatch error rate

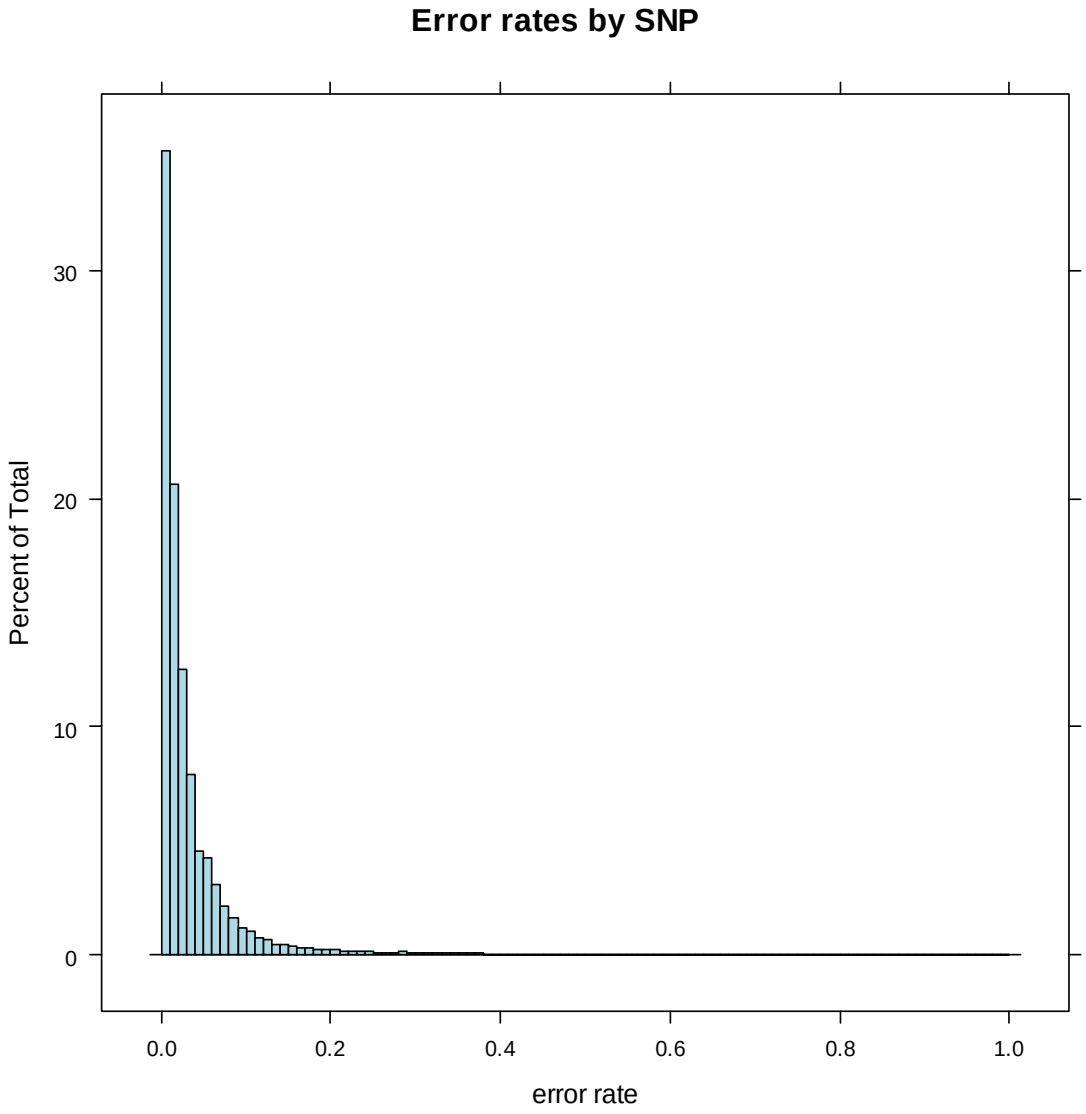


Figure 5.20 Correlation between estimate R-square and actual mismatch error rate

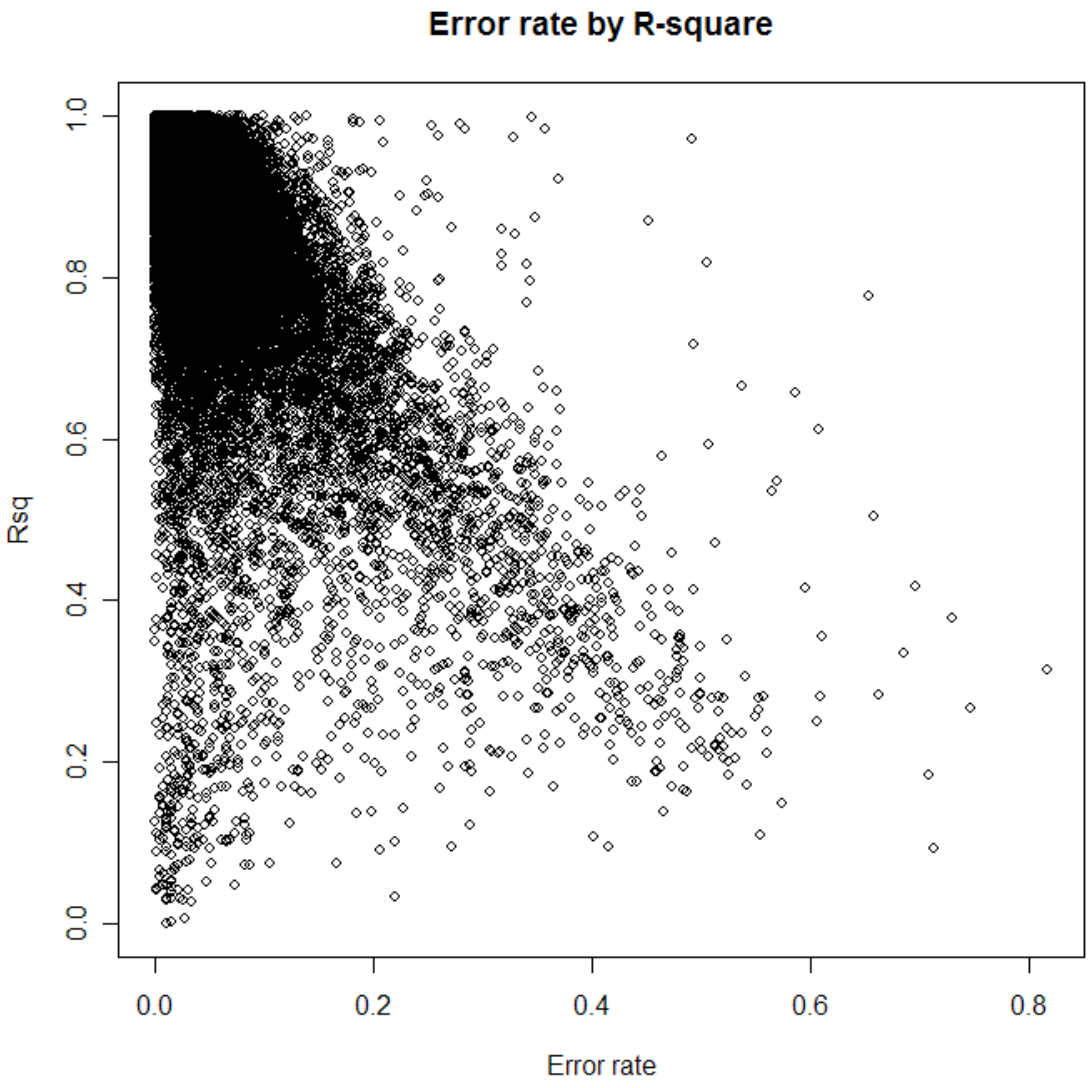
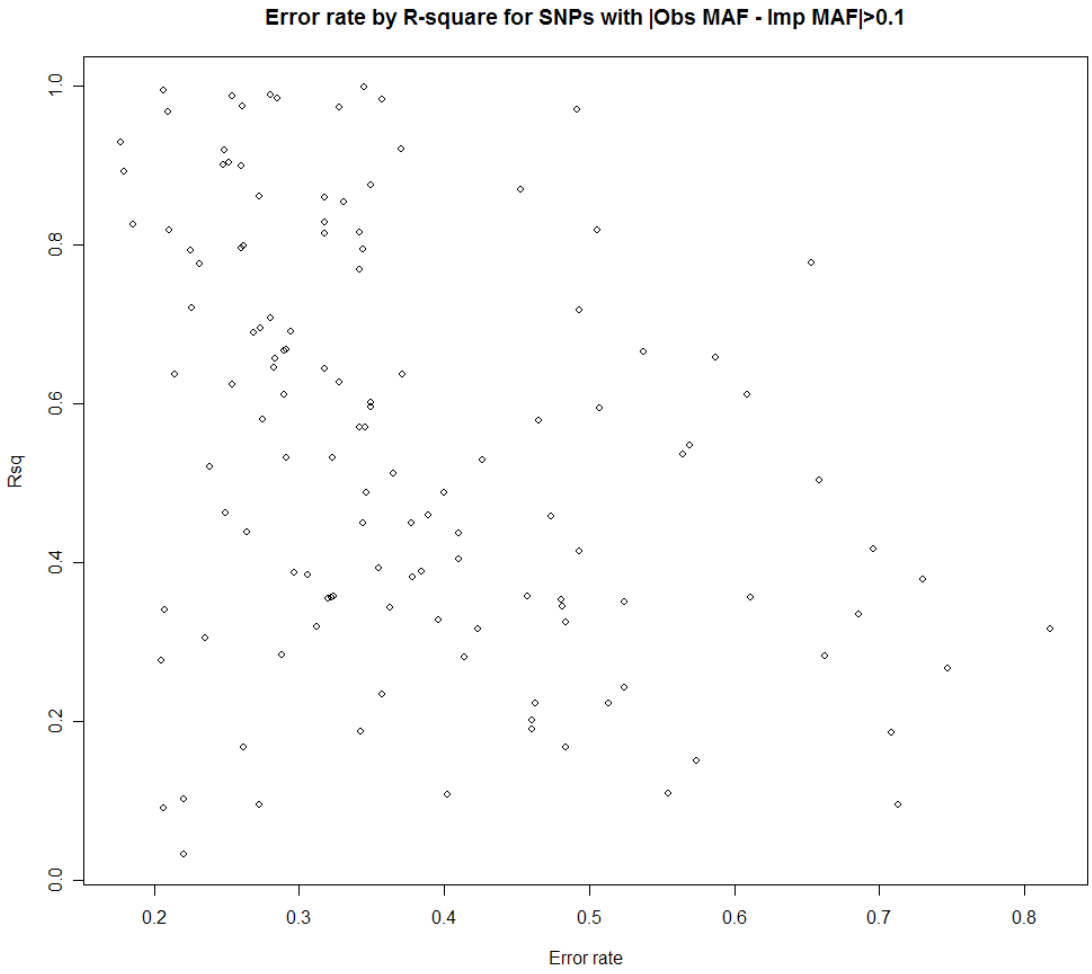


Figure 5.21 Error rate and estimated R-square for SNPs with large difference between actual minor allele frequency and imputed minor allele frequency



5.7 References

Abecasis G.R., Cherny S.S., Cookson W.O. & Cardon L.R. Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30, 97–101 (2002).
Chen W.M. and Abecasis G.R. Family-based association tests for genomewide association scans. Am J Hum Genet 81:913-26 (2007).

Beaty, J. S., West, K. A. & Nepom, G. T. Functional effects of a natural polymorphism in the transcriptional regulatory sequence of HLA-DQB1. Mol Cell Biol 15, 4771-82 (1995).

Bolstad, B.M., Irizarry, R.A., Astrand, M. & Speed, T.P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19, 185–193 (2003).

Cookson WOC, Liang L, Abecasis GR, Moffatt MF, Lathrop M. 2009. Mapping complex disease traits with global gene expression. Nat Rev Genet 10, 184-194.

Cheung, V. G. et al. Natural variation in human gene expression assessed in lymphoblastoid cells. Nat Genet 33, 422-5 (2003).

Dixon AL, Liang L, Moffatt MF et al. A genome-wide association study of global gene expression. Nat Genet 39:1202-7 (2007).

Gretarsdottir, S. et al. The gene encoding phosphodiesterase 4D confers risk of ischemic stroke. Nat Genet 35, 131-8 (2003).

Hubner, N. et al. Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. Nat Genet 37, 243-53 (2005).

Irizarry, R.A. et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4, 249–264 (2003).

Li Y., Willer C.j., Ding J., Scheet P. & Abecasis G.R. 2008. Markov model for rapid haplotyping and genotype imputation in genome wide studies. Submitted for publication; manuscript available from G.R.A. (email: goncalo@umich.edu).

Libioulle, C. et al. Novel Crohn Disease Locus Identified by Genome-Wide Association Maps to a Gene Desert on 5p13.1 and Modulates Expression of PTGER4. PLoS Genet 3, e58 (2007).

Marchini J., Howie B., Myers S., McVean G. & Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 39, 906-13 (2007).

Moffatt MF, Kabesch M, Liang L et al. Genetic variants regulating ORMDL3 expression

contribute to the risk of childhood asthma. Nature 448 :470-3 (2007).

Monks, S. A. et al. Genetic inheritance of gene expression in human cell lines. Am J Hum Genet 75, 1094-105 (2004).

Morley, M. et al. Genetic analysis of genome-wide variation in human gene expression. Nature 430, 743-7 (2004).

Sanna S. et al. Common variants in the GDF5-UQCC region are associated with variation in human height. Nat Genet 40:198-203 (2008).

Schadt, E. E. et al. Genetics of gene expression surveyed in maize, mouse and man. Nature 422, 297-302 (2003).

Scheet P., Stephens M. A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 78:629-44 (2006).

Scott L.J. et al. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science 316, 1341-5 (2007).

Servin B., Stephens M. Imputation-based analysis of association studies: candidate regions and quantitative traits. Plos Genet 3(7): e114 (2007).

Spielman, R.S. et al. Common genetic variants account for differences in gene expression among ethnic groups. Nat Genet 39, 226–231 (2007).

Stephens M., Scheet P. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. Am J Hum Genet 76:449-62 (2005).

Stephens M., Smith N., Donnelly P. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68:978-89 (2001).

Stranger, B.E. et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315, 848–853 (2007).

The International HapMap Consortium. The International HapMap Project. Nature 437, 1299-320 (2005).

The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661-78 (2007).

Willer C.J. et al. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat Genet 40:161-9 (2008).

Yan, H., Yuan, W., Velculescu, V. E., Vogelstein, B. & Kinzler, K. W. Allelic variation in

human gene expression. Science 297, 1143 (2002).

Chapter VI

CONCLUSION

This dissertation tackles a variety of challenges that arise in gene mapping studies of complex diseases, include the handling of repeated measures in quantitative trait linkage analysis, simulation of genome scale data, unknown population structure in case-control studies, and the genetics of gene expression and genotype imputation. We proposed appropriate statistical models and evaluated the methods using simulations and real data. The methods have also been implemented into efficient software packages that are available to the research community. While we show that our proposed methods achieve good performance; there are still opportunities for further extension.

6.1 Variance component linkage analysis for repeated measures

In the studies of repeated measures in quantitative linkage analysis, it is possible to refine our model to include dominance effects, twin environment or other variance-covariance components or even to incorporate covariate effects into the variance-covariance matrix (Lange et al. 1976, Lange and Boehnke 1983, Amos 1994, Almasy & Blangero 1998). In particular, time effects can be introduced into the variance-covariance structure to allow for longitudinal changes in variance components.

Besides simulation approaches, it is possible to compare balanced and unbalanced designs under theoretical framework and derive analytical results. One scenario when unbalanced designs might be preferred include those where there exists heterogeneity in measurement error variance among individuals, i.e. some individuals have higher measurement error variation and require additional measurements, while others with lower measurement error may require less measurements. A model incorporating heterogeneity in measurement error can be extended from our model and it is similar to weighted least square regression.

6.2 Discrete generation framework to simulate genome scale data

Our proposed discrete generation framework can be utilized to incorporate features that are not available in standard coalescent approaches. For example, it is easy to implement the simulation when multiple recombination events on the same sequence are a recombination of exactly two parental sequences. With this feature plus our discrete generation framework, our simulator can simulate sequences following exactly the three assumptions of the Wright-Fisher neutral model (Ewens 1979), namely, discrete generations, finite population size and random mating. Fu 2006 developed an exact probability model for coalescent events that follows the Wright-Fisher model and compared it to the Kingman approximation of coalescent (Kingman 1982) for the scenario where sample size is close to the population size. While the Kingman coalescent approximates the exact coalescent remarkably well, the author found that there is enough differences to justify the use of exact coalescent such as statistics that depends on the

topology of genealogy. Fu 2006 only considered the probability for coalescent event. Under more complex models, such as migration between subpopulations and recombination, the exact probabilities for different types of events remain to be developed. Our model accommodates the events of coalescent, recombination and migration naturally and simultaneously. It can be used to examine the accuracy of approximations of standard coalescent model to Wright-Fisher model, similar to Fu 2006 but can assess the effect of large sample size on recombination, migration as well as simple coalescence. We can then identify the situations when standard coalescent model is good enough and when exact simulation of the Wright-Fisher model is desired.

6.3 Matching-based analysis for genome-wide association studies

As commercial whole genome genotyping platforms gets more and more affordable, larger and larger samples are collected in genome-wide association studies. When samples are collected from a diverse population with a complex ancestry history, such as the United States, population substructure is likely to be present in the sample (Freedman et al., 2004). When samples are from an isolated population, there could be hidden relatedness among individuals in the sample (Lowe 2009). Unadjusted population structure and relatedness may increase the false positive rate or mask true associations in association studies. Genome-wide association studies collect genotypes on 100,000 to 100,000,000 of markers. This number of genotypes carries enough information to infer ancestry of individuals in the sample (Pritchard et al., 2000, Price et al., 2006, Luca et al., 2008). Matching cases and controls based on genotype simultaneously adjust the effect of

population structure and hidden relatedness in the sample because individuals from the same subpopulation or closed relatives tend to group together. Therefore, our method can be easily extended to family data by calculating the dissimilarity between families and then performing family-based matching instead of individual based matching.

Besides structured populations, matching can also be applied to admixed populations. Our simulations show that our method corrects the inflated false positive rate and still maintains power similar to EIGENSTRAT (Price et al., 2006). Another advantage of matching is the invariance to outliers in the data. Unlike principal component analysis, whose calculation depends on all individuals in the sample, pair-wise similarity scores only depend on a pair of individuals and will not be affected by outliers or unknown relatives in the sample. If any outliers need to be removed due to poor matching to the rest of the sample, no calculations need to be redone.

We applied our method on case-control data by using conditional logistic regression. Matching based analysis can also be extended to quantitative trait analysis. One idea is to group individuals with high similarity and model the group effect as a random intercept effect in a linear mixed model. By using only one degree of freedom in the model, one can adjust for heterogeneity in the mean of the quantitative trait among groups and reduce false positives that are due to heterogeneity in trait means and allele frequency difference among groups (or subpopulations). The approach should also reduce the variance in the residual errors and hence increase power to detect true genetic variants. Similarly, when there is heterogeneity in the effect size of the causal genetic variant, a random slope for the genetic variant can be used to further reduce residual errors and increase power. The idea of using linear mixed model for matched sets can then be extended to use

generalized linear mixed model for case-control data and adjust for population heterogeneity in disease prevalence (random intercept) and similarly for heterogeneity in genetic effect size (random slope). This gives an alternative method to the conditional logistical regression. It would be interesting to compare the performance of the two alternatives in the context of genome wide association studies.

6.4 Expression quantitative trait loci (eQTL) mapping and genotype imputation

It has been shown that the genetic map of gene expression can be used to help interpret findings from genome-wide association studies of complex diseases (Libioulle et al. 2007, Dixon et al. 2007, Cookson et al. 2009). The mapping of gene expression itself is an interesting genome-wide association study. Experiences gained from eQTL mapping may be borrowed to improve genetic mapping of other quantitative traits. We have examined the impact of genotype imputation on power. Genotype imputation (Scheet & Stephens 2006, Servin & Stephens 2007, Marchini et al 2007, Li et al. 2008) can reliably reproduce missing genotypes as well as association results. The quality of imputed genotype and the association statistics can be well predicted by the estimated correlation between imputed genotype and the true counterpart. We estimated that imputation using the HapMap SNP panel (The International HapMap Consortium 2007) increases the number of transcripts mapped in *cis* by ~10%. Encouraged by this finding, we have started to explore an even denser panel of SNPs. The 1000 Genomes project (www.1000genomes.org) has recently derived a panel of more than 8 million SNPs based on the shot-gun sequence data of ~120 individuals (HapMap CEU). Genotype imputation based on these 8 million SNPs leads to ~4-6% more *cis* eQTLs than imputation results

using the HapMap panel alone as a template. More than 1000 individuals will eventually be sequenced by the 1000 Genomes project. We expect to be able to impute more SNPs with higher quality and mapping even more eQTLs.

We show that sample size has a dramatic effect on study power (Dixon et al. 2007). However, even with hundreds of individuals in the sample plus imputation of a dense panel of SNPs, only ~13% transcripts with high total heritability (>30%) can be mapped in *cis* with genome-wide significance. Further increase in sample size will certainly increase the power but also increase in cost. The process of determining gene expression values is complex. For example, the stage of cell cycle, conditions when RNA is extracted and cDNA is synthesized, the variation among technicians, the design and production of chips, and the hybridization of the microarray could all affect the final gene expression value.

Repeated measures of expression for the same gene can help to reduce measurement error but still increase cost. Note that an individual array evaluates expression for tens of thousands of transcripts which undergo similar conditions such as similar cell cycle, conditions of RNA extraction and cDNA synthesis, the same technician, and the same microarray chip. Expression levels for these transcripts can be regarded as "repeated measures" of this experimental variability and can be used to summarize the effects of this experimental variability. One way to summarize information shared among all transcripts on the same chip is to use principal components given that genetic effect on the expression is much smaller than the above systematic effects (Leek & Storey 2007, Stegle et al. 2008).

Using the MRCA dataset, we found that the RNA extraction date, cDNA synthesis

date, IVT (*in vitro* transcription) date and the date that the sample was fragmented were significantly associated with the top 2nd to 4th principal components of gene expression. A few other principal components, such as the 11th, 18th and 22nd, were also associated with date of experiment. Adjusting the top 69 principal components as covariates in the model to account for unobserved systematic noises (such as batch effects), we observed a 3-fold increase in the *cis* eQTL. The number of transcripts mapped within 1Mb increased from 2219 to 6237, accounting for 37.3% highly heritable transcripts (H^2 >30%). We estimated that imputation of HapMap SNPs and SNPs from the 1000 Genomes project leads to a further increase of *cis* eQTL by 5.9% and 10.3%, respectively.

Sequence based techniques, such as RNA-seq (Wang et al. 2009), have been introduced to gene expression analysis. New methods are needed to utilize the power of this newly available data. For example, a new method could be used to give a much cleaner measures of gene expression abundance and to determine transcriptional events such as alternative splicing. In the meantime, array based techniques will still be used because of its relatively low cost and newly developed method that can better use the existing techniques.

In the chasing after causal genetic variants that are responsible for complex diseases, any breakthrough will make one step forward, from phenotypes with higher accuracy to genotypes with higher coverage, from establishing valid and powerful statistical evidence to obtaining sensible biological interpretation, from faster and larger scale simulations to better statistical inference based on real data. Besides these challenges addressed here, there are many others that need to be tackled.

6.5 References

Amos CI. 1994. Robust variance-components approach for assessing genetic linkage in pedigree. Am J Hum Genet. 54:535-43.

Almasy L, Blangero J. 1998. Multipoint quantitative trait linkage analysis in general pedigree. Am J Hum Genet. 62: 1198-211.

Cookson WOC, Liang L, Abecasis GR, Moffatt MF, Lathrop M. 2009. Mapping complex disease traits with global gene expression. Nat Rev Genet 10, 184-194.

Dixon AL, Liang L, Moffatt MF et al. 2007. A genome-wide association study of global gene expression. Nat Genet 39:1202-7.

Ewens WJ (1979), Mathematical Population Genetics. Springer, Berlin.

Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, Patterson N, Gabriel SB, Topol EJ, Smoller JW, Pato CN, Pato MT, Petryshen TL, Kolonel LN, Lander ES, Sklar P, Henderson B, Hirschhorn JN, Altshuler D. 2004. Assessing the impact of population stratification on genetic association studies. Nat Genet 36:388-393.

Fu Y.X. 2006. Exact coalescent for the Wright-Fisher model. Theoretical Population Biology 69: 385-394.

The International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449, 851-861.

Kingman J.F.C. 1982. The coalescent. Stochastic Process. Appl. 13:235-248.

Lange K, Westlake J, Spence MA. 1976. Extensions to pedigree analysis. III. Variance components by the scoring method. Ann Hum Genet. 39:485-91.

Lange K, Boehnke M. 1983. Extensions to pedigree analysis. IV. Covariance components models for multivariate traits. American Journal of Medical Genetics 14:513-524.

Leek JT & Storey JD. 2007. Capturing Heterogeneity in Gene Expression Studies by Surrogate variable Analysis. PLoS Genetics V3,9,e161: 1724-1735.

Li Y., Willer C.j., Ding J., Scheet P. & Abecasis G.R. 2008. Markov model for rapid haplotyping and genotype imputation in genome wide studies. Submitted for publication; manuscript available from G.R.A. (email: goncalo@umich.edu).

Libioulle, C. et al. 2007. Novel Crohn Disease Locus Identified by Genome-Wide Association Maps to a Gene Desert on 5p13.1 and Modulates Expression of PTGER4. PLoS Genet 3, e58.

Lowe JK, Maller JB, Pe'er I, Neale BM, Salit J, et al. 2009. Genome-Wide Association Studies in an Isolated Founder Population from the Pacific Island of Kosrae. PLoS Genet 5(2): e1000365. doi:10.1371/journal.pgen.1000365.

Luca D, Ringquist S, Klei L, Lee AB, Gieger C, Wichmann HE, Schreiber S, Krawczak M, Lu Y, Styche A, Devlin B, Roeder K, Trucco M. 2008. On the use of general control samples for genome-wide association studies: genetic matching highlights causal variants. Am J Hum Genet 82:453-463.

Marchini J., Howie B., Myers S., McVean G. & Donnelly P. 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 39, 906-13.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904-909.

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. 2000. Association mapping in structured populations. Am J Hum Genet 67:170-181.

Risch NJ. 2000. Searching for genetic determinants in the new millennium. Nature 405:847–856.

Servin B., Stephens M. 2007. Imputation-based analysis of association studies: candidate regions and quantitative traits. PLoS Genet 3(7): e114.

Scheet P., Stephens M. 2006. A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 78:629-44.

Stegle O, Kannan A, Durbin R, Winn J. 2008. Accounting for Non-genetic Factors Improves the Power of eQTL Studies. Springer Berlin / Heidelberg, Volume 4955: 411-422.

Wang Z, Gerstein M & Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10, 57-63.