STATISTICAL METHODS AND ANALYSIS IN GENOME WIDE ASSOCIATION STUDIES AND NEXT-GENERATION SEQUENCING

by

Wei Chen

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2011

Doctoral Committee:

Professor Gonçalo R. Abecasis, Chair
Professor Michael L. Boehnke
Professor Bin Nan
Research Assistant Professor Hyun Min Kang
Professor Anand Swaroop, National Eye Institute

To Ying and Claire

Acknowledgements

I would like to thank many people including my committee members, friends, colleagues and my family. Without the guidance and help from them, this dissertation will not be possible.

First of all, I would like to express my deepest gratitude to my thesis advisor, Dr. Gonçalo Abecasis, who brought me to this exciting area, led me to the challenging and interesting genetic problems and offered excellent guidance. During my Ph.D. study, I also benefit a lot from the wonderful group atmosphere and the worldwide collaborations, provided by my advisor. I would like to thank Dr. Michael Boehnke, who is always willing to help and gave me great advice on many things ranging from problem solving, presentation skills, job hunting and communication experience. I would like to thank Dr. Bin Nan and Dr. Hyun Min Kang for guiding my research for the past few years. My special thanks are delivered to Dr. Anand Swaroop, who provided unique opportunity for me to work on wonderful eye-related genetic problems and broaden my knowledge in biology and medical science.

I am very thankful to knowledgeable faculty members, CSG members and fellow students in the Department of Biostatistics at the University of Michigan. I enjoy being a graduate student and doing research in this nice community.

My deep gratitude is given to my beloved wife Ying. Ying keeps encouraging and supporting me at my good and bad times. My daughter Claire always gives me a lot of happiness and working energy. I would like to thank my parents for their caring and supporting and Ying’s parents for their encouragements.

TALBE OF CONTENTS

DEDICATION ii

ACKNOWLEDGEMENTS iii

LIST OF FIGURES vi

LIST OF TABLES vii

ABSTRACT viii

CHAPTER 1 INTRODUCTION AND THE SCOPE OF THIS DISSERTATION 1

1.1 GENETIC STUDY OF COMPLEX DISEASES 1
1.2 STATISTICAL AND COMPUTATIONAL CHALLENGES IN GWAS AND NGS 2
1.3 THE SCOPE OF THIS DISSERTATION 3

CHAPTER 2 GENOME-WIDE ASSOCIATION STUDY AND PREDICTION FOR AGE-RELATED MACULAR DEGENERATION 7

2.1 INTRODUCTION 7
2.2 GENOTYPING DATA 8
2.3 STATISTICAL METHODS 8
2.4 RESULTS AND DISCUSSION 11

CHAPTER 3 GRAPHICAL BROWSER FOR GWAS WITH HIGH-DIMENSIONAL PHENOTYPES 41

3.1 INTRODUCTION 41
3.2 FEATURES OF GWAS GUI BROWSER 42
3.3 EXAMPLES OF APPLICATION 44
3.4 IMPLEMENTATION 45

CHAPTER 4 GENOTYPE CALLING AND HAPLOTYPING IN PARENT-OFFSPRING TRIOS . 47

4.1 INTRODUCTION 47
4.2 MATERIAL AND METHODS 49
4.3 DATA SETS 55
4.4 RESULTS 57
4.5 DISCUSSION 63

CHAPTER 5 STATE SPACE REDUCTION MODEL FOR HAPLOTYPING AND GENOTYPE CALLING 80

5.1 INTRODUCTION 80
5.2 METHOD 81
5.3 RESULTS AND DISCUSSION 90

CHAPTER 6 SUMMARY AND DISCUSSION 101

6.1 SUMMARY AND FUTURE WORK OF CHAPTER 2-5 101
6.2 ONGOING WORK AND FUTURE PLAN 104
6.3 CONCLUSION 107

REFERENCES 110

LIST OF FIGURES

Figure 2.1 Summary of genomewide association scan results 36
Figure 2.2 Regional plots for association signals in five previously reported loci.......... 37
Figure 2.3 Regional plot for association signals in the three new loci 38
Figure 2.4 Regional plot of association signals in HDL-c and AMD............................... 39
Figure 2.5 Multi-locus genotypes and disease risk... 40
Figure 3.1 An illustration of the GWAS GUI browser interface...................................... 46
Figure 4.1 Workflow of SNP discovery and genotype calling 75
Figure 4.2 Cartoon view of LD-aware method for unrelated samples and parent-offspring trios 76
Figure 4.3 Frequency stratified mismatch rate at heterozygote sites at different depths for 30 trios, 60 unrelated and 90 unrelated samples at base error rate 0.01 77
Figure 4.4 Genotype distributions and disconcordance for heterozygotes, reference homozygotes and alternative homozygotes 78
Figure 5.1 Cartoon view of the standard hidden Markov model for imputation and haplotype inference... 96
Figure 5.2 The pattern of the number of unique haplotypes (for a fixed window of 10kb) 97
Figure 5.3 Decomposition of template haplotypes 98
Figure 5.4 Cartoon view of the state space reduction method in an exampled window.. 99
Figure 6.1 A comparison of some conventional methods, homogeneous HMM approach and non-homogeneous HMM approach... 108
Figure 6.2 Raw R^2 distribution from chromosome 22 109

LIST OF TABLES

Table 2.1 Summary description of discovery samples used in the genome-wide association and replication studies 22
Table 2.2 Confirmation of previously reported association signals in the discovery samples 23
Table 2.3 New Locus with Confirmed Association to AMD ($p < 5x10^{-8}$) 24
Table 2.4 Association of HDL-C loci with AMD 25
Table 2.5 Association results of some published candidate SNPs in our scan 26
Table 2.6 Complete Results for All SNPs Where Replication Attempted 27
Table 2.7 Sample by Sample Results for Newly Reported Loci 28
Table 2.8 Best genotyped proxy SNPs for reported loci 33
Table 2.9 Association Results in Discovery Sample for Different Analysis Models 34
Table 2.10 Evaluation of Association of Loci with $p < 5x10^{-8}$ Overall In Specific AMD Subtypes (OR, 95 C.I., p-value) 35
Table 4.1 Error rates for genotype calling 67
Table 4.2 Quality of estimated haplotypes in simulated 1Mb regions 68
Table 4.3 Family structures of the SardiNIA data sets 69
Table 4.4 Overall genotype concordance between metabochip and low-pass sequence data from Sardinia project 70
Table 4.5 Stratified genotype concordance between metabochip and low-pass sequence data from Sardinia project 71
Table 4.6 Improvement of genotype accuracy with phased input from Beagle 72
Table 4.7 Error rates stratified by frequency at heterozygotes for genotype calling 73
Table 4.8 Comparisons of mismatch rate between the child and the parents for the simulation of 30 trios 74
Table 5.1 Estimate of running time and memory for haplotype inference using MaCH. 93
Table 5.2 Comparisons of the computing time and memory in simulations 94
Table 5.3 Comparisons of the computing time and memory in real data 95

ABSTRACT

STATISTICAL METHODS AND ANALYSIS IN GENOME WIDE ASSOCIATION STUDIES AND NEXT-GENERATION SEQUENCING

by

Wei Chen

Chair: Gonçalo R. Abecasis

Genome-wide association studies (GWAS), which examine common genetic variants in thousands of individuals, have identified many genetic loci associated with a variety of complex diseases and phenotypes. New Next-Generation Sequencing (NGS) technologies allow us to extend these studies to rarer variants not typically evaluated by GWAS. In this dissertation, I present novel statistical methods and software to dissect the genetic basis of complex traits in the context of both GWAS and NGS.

First, I present a large-scale GWAS for Age-related Macular Degeneration (AMD). Our studies extend the catalog of AMD associated loci and provide clues about underlying cellular pathways. A novelty in our study is that I propose a prediction method using all susceptibility loci to help identify individuals at high risk of disease. The prediction can be extended to the general population with a weighted scheme combining both disease prevalence and case-control ratio in GWAS sample.

Second, I describe an interactive package that provides graphical overviews of the results of whole-genome association studies in datasets with rich multi-dimensional phenotypic information, such as global surveys of gene expression.

Third, I propose and implement an efficient Hidden Markov Model (HMM) based method for genotype calling and haplotype inference in parent-offspring trios. Our method considers both linkage disequilibrium (LD) patterns and the constraints imposed by the family structure in assigning individual genotypes and haplotypes. Using simulations and sequencing data from ongoing projects, I show that trios provide higher genotype calling accuracy across the frequency spectrum, both overall and at hard-to-call heterozygous sites. In addition, sequencing trios can provide greatly improved haplotype phasing accuracy.

Finally, I describe an efficient state space reduction method for haplotype inference and genotype calling. This method is motivated by the increasing computational challenge of HMM-based approaches used to describe haplotype sharing in GWAS and NGS data. Our method takes advantage of local similarity between haplotypes and reduces the HMM state space dynamically, while preserving the same accuracy of full state space method. Through simulation and real data analysis, I show that this method can have substantial savings in both memory and CPU time.

Chapter 1

Introduction and the Scope of this Dissertation

1.1 Genetic Study of Complex Diseases

Modern genetics originated from simple Mendelian disorders back to a century ago. In the past decades, more attention has been paid to complex traits, which are much more complicated and caused by both genetic and environmental factors. To understand the genetic basis of complex diseases and many common phenotypes, linkage analysis and association analysis, have made substantial progress in the past few decades [1, 2]. Due to the high experimental cost and limitations of biology technology, only a small fraction or a few short regions of the genome were studied and thus the localization of the disease-associated variants was very limited. Shortly after the completion of Human Genome Project, genome-wide association study (GWAS) became a feasible and powerful approach to uncover genetic variants with much better resolution by examining hundreds of thousands of Single Nucleotide Polymorphisms (SNPs) across the whole genome. Over one thousand loci susceptible to common diseases and phenotypes have been revealed [3-5] and have led to subsequent functional analysis in biomedical area. Despite the number of loci identified, GWAS are based on the hypothesis of common disease / common variant (CDCV) and a large proportion of heritability has not been explained [6]. Rare variants are believed to play an important role in the missing heritability. To study rare variants and pinpoint the causal alleles more accurately, sequencing technology has been advanced rapidly in the past few years, boosted by several worldwide giant projects such as 1000 Genomes Project (ww.1000genomes.org). Next-generation sequencing (NGS) allows us to detect more variants, including SNPs and structural variations, and explore rare variants systematically beyond common variants assayed by GWAS.

1.2 Statistical and Computational Challenges in GWAS and NGS

Compared to the genetic studies back to thirty years ago, GWAS and NGS generate a huge amount of data. For GWAS, thousands of samples are collected and genotyped using commercial or customized microarray chips examining hundreds of thousands to millions of variants across the genome. Statistical methods are crucial to analyze such data sets [7]. A few successful examples include but are not limit to a) power calculation for study design; b) genotype imputation for meta-analysis and examining more variants; c) methods for adjusting population stratification; d) methods for adjusting multiple testing. For NGS, the data scale grows to another level. The statistical and computational challenges mainly lie on two levels: a) lower level variants calling and haplotype inference from sequencing data; b) upper level association tests involving rare variants and structural variations. Unlike very accurate genotypes from GWAS, high throughput sequencing machine generates millions of short fragments of the genome. This process requires accurate and efficient statistical algorithms for mapping and genotype calling. The accuracy of genotypes is critical to follow-up association studies. New association tests are required in the context of NGS. For example, traditional association methods might be underpowered if disease is caused by multiple rare variants. The successful imputation and haplotype inference methods in GWAS become less feasible given greatly increased number of samples and DNA variants. Efficient computational methods and updated computer hardware are required to overcome these practical issues.

1.3 The Scope of this Dissertation

In this dissertation, I will present both methodologies and software developments about imputation, SNP calling, haplotype inference and multiple testing motivated by the challenges previously described; I will also present novel scientific findings on age-related macular degeneration and visualization tools developed for efficiently storing and displaying GWAS results with high-dimensional phenotypes.

In chapter 2, I present a genome-wide association study for age-related macular degeneration (AMD) in 2,157 cases and 1,150 controls [4]. Our results validate AMD susceptibility loci near *CFH* ($P < 10^{-75}$), *ARMS2* ($P < 10^{-59}$), *C2/CFB* ($P < 10^{-20}$), *C3* ($P < 10^{-9}$), and *CFI* ($P < 10^{-6}$). I compared our top findings with the Tufts/Massachusetts General Hospital genome-wide association study of advanced AMD (821 cases, 1,709 controls) and genotyped 30 promising markers in additional individuals (up to 7,749 cases and 4,625 controls). With these data, I identified a susceptibility locus near *TIMP3* (overall $P = 1.1 \times 10^{-11}$), a metalloproteinase involved in degradation of the extracellular matrix and previously implicated in early-onset maculopathy. In addition, our data revealed strong association signals with alleles at two loci (*LIPC*, $P = 1.3 \times 10^{-7}$; *CETP*, $P = 7.4 \times 10^{-7}$) that were previously associated with high-density lipoprotein cholesterol (HDL-c) levels in blood. Consistent with the hypothesis that HDL metabolism is associated with AMD pathogenesis, I also observed association with AMD of HDL-c—

associated alleles near *LPL* (P = 3.0 × 10^{-3}) and *ABCA1* (P = 5.6 × 10^{-4}). Multilocus analysis including all susceptibility loci showed that 329 of 331 individuals (99%) with the highest-risk genotypes were cases, and 85% of these had advanced AMD. In addition, I propose a novel method to facilitate the AMD prediction with GWAS data in general population. Our studies extend the catalog of AMD associated loci, help identify individuals with a high risk of developing the disease, and provide clues about underlying cellular pathways that could eventually lead to new therapies.

Following DNA microarray, NGS technologies allow us to explore more variants beyond GWAS. Much progress has been made for efficient and accurate genotype call from NGS data and simulations have been performed to study the efficiency and accuracy at different sample sizes and sequencing depths. However, most of the current variant calling algorithms can only handle unrelated samples; systematic evaluations of sequencing data of families are not available up to date. In Chapter 4, I propose an efficient and accurate method for genotype calling and haplotype inference in sequencing parent-offspring trios. This method combines both linkage disequilibrium (LD) patterns and family constraints within the trio together into a widely used hidden Markov model in imputation, which takes advantage of similar stretches of chromosomes shared between individuals. This method provides a tool of variant calling and haplotype phasing for many ongoing sequencing projects that have parent-offspring trios. In addition, I am able to explore the potential advantages of sequencing additional family members. I simulated shotgun sequencing data in genotype likelihood format (GLF) for trios and unrelated samples at various depths with two sequencing error rates. For same number of sequenced samples, our simulations show that sequencing trios are preferable to unrelated samples at low depth 1X, 2X and 4X in terms of detecting polymorphic sites. Generally, trios have higher calling accuracy across different frequency spectra. Furthermore, sequencing trios can greatly increase the haplotyping accuracy, which is crucial for follow-up imputation with existing GWAS data. However, at depth 8X and above, the gain of trios are limited and design of unrelated samples are more preferable in terms of variant calling. The method can be extended to the designs of nuclear family and general pedigree structure.

There have been great successes of linkage disequilibrium (LD) based imputation and haplotype haplotype inference methods in detecting additional analysis and performing meta analysis across different platforms [4, 8, 9]. A commonly used approach is based on a hidden Markov model treating sample haplotype as a mosaic of a pool of reference haplotypes [10-14]. The size of the reference panel is usually limited to less than one hundred. As the reference panel expands quickly (e.g. 1000 Genomes Project) to a few hundred individuals or even more, the computing cost, including time and memory, increases as well. The type of method using full state space thus becomes less feasible in practice. On the other hand, the method using approximated reduced state space can be applied but will result in losing efficiency and accuracy, especially for genotype calling of rare variants. In chapter 5, I propose a state space reduction method to overcome the above two limitations. The method defines a set of quantities in a reduced state space to keep track of all information in full state space by taking advantage of local similarity between different haplotypes. The results from simulation and real data sets show that the new method can substantially save computing memory and time while preserving the accuracy of the full sate space method.

In Chapter 6, I will summarize the work of Chapter 2-5 and discuss limitations and future plans. In addition, I will describe ongoing effort to address a multiple testing problem. I extend a hidden Markov model based FDR control procedure to account for non-homogeneous dependency structures. I developed a general EM algorithm for parameter estimation. I aim to explore its application in GWAS and sequencing, where hundreds of thousands of SNPs are examined with specific dependency structures.

Overall, my research will facilitate current biomedical research in different aspects including understanding the biology of retina, accurate genotype calling and haplotype inference program for next generation sequencing and improvement of existing computational software for imputation and sequencing.

Chapter 2

Genome-wide Association Study and Prediction for Age-related Macular Degeneration

The content of this chapter has been published in Chen et al. 2010 [4].

2.1 Introduction

AMD is a progressive neurodegenerative disease and a common cause of blindness in the elderly population, particularly in developed countries [15, 16]. The disease affects primarily the macular region of the retina, which is necessary for sharp central vision. An early hallmark of AMD is the appearance of drusen, which are extracellular deposits of proteins and lipids under the retinal pigment epithelium (RPE). As the disease progresses, drusen grow in size and number. In advanced stages of AMD, atrophy of the RPE (geographic atrophy) and/or development of new blood vessels (neovascularization) result in death of photoreceptors and central vision loss [15, 17, 18].

Multiple genetic linkage studies provided strong evidence of susceptibility loci, notably on chromosomes 1q31 and 10q36 [19-23]. Disease-associated variants near *CFH* (1q31) and in a cluster of genes near *ARMS2* (10q26) were first identified both through genomewide association studies (GWAS) [24, 25] and fine mapping of linkage signals [26-29]. Discovery of association between AMD and the *CFH* locus lead researchers to discovery of association signals near other complement genes, including *C2/CFB*, *C3*, and *CFI* [30-33].

2.2 Genotyping Data

The participants in genome-wide association study were mainly collected at the University of Michigan in Ann Arbor (collection coordinated by AS), at the University of Pennsylvania in Philadelphia (coordinated by DS), and at the Mayo Clinic in Rochester, Minnesota (coordinated by AE). Detailed information about the number of cases and controls and the distribution of age, sex and disease severity in each collection is summarized in Table 2.1.

Genotyping was performed at the Center for Inherited Disease Research (CIDR) at Johns Hopkins University using Illumina Human370 Bead Chips (Illumina, San Diego, CA, USA) and the Illumina Infinium II assay protocol [34]. Allele cluster definitions for each SNP were determined using Illumina BeadStudio Genotyping Module version 3.2.32 and the combined intensity data from 99% of the samples according to CIDR protocol; the resulting cluster definitions were then used on all samples. Genotypes were not called if the quality threshold (gencall score) was below 0.25. Genotypes were not released from CIDR for SNPs which failed technical filters for call rates less than 85%, more than 1 HapMap replicate error, more than a 4% (autosomal) or 5% (X chromosome) difference in call rate between sexes, more than 0.5% male heterozygote frequency for X chromosome. Y and XY SNPs were manually reviewed and clusters adjusted or genotypes dropped as appropriate. Genotypes were released from CIDR for 344,942 (99.46%) of the attempted SNPs. Blind duplicate reproducibility was 99.992%.

2.3 Statistical Methods

Population Stratification: The samples are all European descent. I used the software EIGENSTAT to adjust for the modest population stratification. After adjustment for the first two principal components of ancestry, the genomic control parameter was 1.007.

Genotype Imputation: To expand the genome coverage, I performed a genome-wide imputation using haplotypes from the HapMap CEU samples as templates (release 22). Imputation was done using MACH (Yun Li, www.sph.umich.edu/csg/abecasis/Mach/). For downstream analyses, I filtered out poorly imputed SNPs and focused on markers with estimated r^2 between imputed and true genotypes > 0.3.

Statistical Analyses: To investigate the association between each genotyped or imputed SNP and AMD, I first carried out a logistic regression for each SNP assuming an additive genetic model and adjusting for the top two eigenvectors from EIGENSTRAT. At $p < 10^{-6}$, I identified a total of seven independently associated SNPs in previously reported loci (*CFH*, *ARMS2*, *C3*, *C2/CFB* and *CFI*). These SNPs were included as covariates in logistic regression analyses designed to identify additional loci associated with AMD.

Analysis for Follow-up Study: To combine the statistics across different groups for replication, I first selected an arbitrary reference allele for each marker and then calculated a z-statistic summarizing the evidence for association in each study (summarizing both the p-value, in its magnitude, and the direction of effect, in its sign). I then calculated an overall z-statistic as a weighted average of the individual statistics and calculated the corresponding p-value. Weights were proportional to the square root of the number of individuals examined in each study and were selected such that the squared weights sum up to 1.0.

Association Testing: For samples including unrelated individuals only (all discovery samples, the Tufts/MGH samples and the Johns Hopkins, Oregon and Penn-NJ sample sets) the data were analyzed using simple logistic regression models with age and sex as covariates. For the discovery samples, the first two principal components of ancestry were used as covariates in all reported analyses and genotypes for the markers listed in Table 2.2 were used as covariates in a subset of the analyses (described in the text). For follow-up samples, genotypes at *CFH* and *ARMS2* were included as covariates where available. For samples including related individuals, the data were analyzed with the test of Thornton and McPeek [35].

Risk Prediction Approach. To evaluate the cumulative contribution of the alleles identified here to disease risk, I fitted a simple logistic regression model to the data. The effect of each genotype was modeled on a log-additive scale, with no interaction terms between genotypes: $\log it(y_i) = \alpha + \sum_j \beta_j x_{ij}$

Then fitted probability is calculated for each sample $\hat{y}_i = 1/(1+e^{-(\hat{\alpha}+\sum_j \hat{\beta}_j x_{ij})})$

I sorted samples according to their fitted probability of disease and organized individuals into deciles of fitted risk y_i. Then I counted the proportion of affected individuals in each risk decile. In a subsequent analysis, I assigned different weights to cases and controls, designed to reflect the fact that cases are enriched in our sample. The weight is defined as

$$w_i = \begin{cases} f_{case} / p_{case} & y_i = 1 \\ f_{control} / p_{control} & y_i = 0 \end{cases}$$

and weighted fitted probability is defined as $z_i = w_i \hat{y}_i$

where p_{case} = 0.65 and $p_{control}$ = 0.35 are the fractions of cases and controls in our sample and f_{case} = 0.20 and $f_{control}$ = 0.80 are the expected fractions of cases and controls in an elderly population at age ~75. Cases were assigned weight $f_{\mathrm{case}}/p_{case}$ and controls were assigned weight $f_{control}/p_{control}$. I sort z_i in an ascending order and denote as $z_{(1)},z_{(2)},\ldots,z_{(m)}$ with corresponding weight $w_{(1)},w_{(2)},\ldots,w_{(m)}$. Taking these weights into account, I then divided the sample into deciles ensuring that summed weights in each decile were identical. z_i is used to estimate the case fraction in general population.

2.4 Results and Discussion

The execution of progressively larger GWAS typically results in the gradual discovery of new susceptibility loci (see the examples of Crohn's disease [36], type 2 diabetes [37], obesity [38], and lipids [5, 39]). To identify additional susceptibility loci and biochemical pathways contributing to AMD, I performed GWAS in a large collection of cases and controls (Table 2.1) using a genotyping platform that captures >90% of common variants in European ancestry samples.

I genotyped study samples, including 75 blind duplicates, together with HapMap controls at the Center for Inherited Disease Research (CIDR, Johns Hopkins) with Illumina Human370 chips. After genotyping, I excluded 18 individuals with an unexpected 1st or 2nd degree relative in the dataset and 13 individuals with evidence for a non-European ancestry component[40] resulting in a total of 2,157 unrelated cases and 1,150 unrelated controls for analyses. I excluded markers with <95% call rate, minor allele frequency <1%, or evidence for deviation from Hardy-Weinberg equilibrium at $p<10^{-6}$, resulting in a total of 324,067 autosomal SNPs for analysis. The average call rate for analyzed markers and samples was 99.94%. I identified short stretches of haplotype shared between individuals in our study and those in the HapMap CEU [41] and used these to impute missing genotypes, expanding the number of analyzed SNPs to about 2.5 million imputed or genotyped SNPs. Complete GWAS data and results are available from dbGaP accession phs000182.v1.p1.

An initial comparison of allele frequencies between cases and controls resulted in a genomic control parameter [42] of 1.056; adjustment for the first two principal components of ancestry [PCA, 40] reduced this to 1.007. PCA can account for subtle differences among European ancestry samples (such as North-South or East-West gradients in allele frequency, see [43]) and provide a useful safeguard against population stratification. All results reported here refer to this PCA adjusted analysis.

Reassuringly, I observed strong evidence of association at established susceptibility loci (see Table 2.2, Figure 2.1 and 2.2); near *CFH* (strongest association at rs10737680, odds ratio 3.11(2.76, 3.51), with $p<1.6x10^{-75}$), near *ARMS2* (at rs3793917, OR=3.40 (2.94,3.94), $p=4.1x10^{-60}$), near complement component 2 (*C2*) and complement factor B (*CFB*) (at rs429608, OR=2.16 (1.84,2.53), $p=2.5x10^{-21}$), and near complement component 3 (*C3*) (at rs2230199, OR=1.74 (1.47,2.06), $p=1.0x10^{-9}$). Our study provides confirmation of a recently reported association between AMD and complement factor I (*CFI*) (at rs2285714, OR=1.31 (1.18,1.45), p=3.4x10^{-7}) [30]. Conditioning on the strongest associated variant at each of these loci identified additional, strong association signals near *CFH* (at rs1329424, p=6.4x10^{-16}) and in the *C2*/*CFB* locus (at rs9380272, p=2.3x10^{-8}), consistent with previous reports of multiple disease-associated alleles at the two loci [29, 31, 44, 45]. Where possible, I evaluated evidence for association at other previously suggested susceptibility loci using genotypes or imputed data. The results are summarized in Table 2.5; although none of these loci show p<.05 in our data, note that 8 of 9 signals trend in the same direction as the original report.

To identify new AMD susceptibility loci, I conditioned on the seven strongly associated SNPs (see Table 2.2) and repeated the genomewide analysis. No single SNP was significant at p<5x10^{-8} in this conditional analysis. Next, I exchanged initial results with the Tufts/MGH GWAS for 1358 SNPs that could be assayed directly with Affymetrix 6.0 genotyping arrays and that were significant at p<.001 in either study. Tufts/MGH GWAS results were adjusted for possible population stratification using genomic control[42], consistent with the analysis presented in the companion paper. After excluding 158 AREDS study participants that were genotyped in both studies, this allowed us to examine promising SNPs in an additional 821 cases with geographic atrophy or neovascularization and 1,709 controls. Twenty-five SNPs showing consistent evidence of association in both groups of participants and five other SNPs with strong evidence for association in our data alone were genotyped in additional samples (see Table 2.1). Summary results from follow-up experiments are presented in Table 2.6. Detailed results for the three most strongly associated loci (near *TIMP3, CETP*, and *LIPC)* and two other loci discussed below (*LPL, ABCA1*) are provided in Table 2.7.

To validate our results, I examined Hardy-Weinberg equilibrium statistics and evidence for heterogeneity at these new loci. I also genotyped a subset of the imputed SNPs in our discovery sample. At each of these loci, Hardy-Weinberg equilibrium p-values in cases, in controls and in the combined dataset were all >.20, suggesting no data quality problems. Furthermore, I found no evidence for heterogeneity at any of these loci (all Cochran's Q heterogeneity p-values >.20). Finally, when I genotyped a subset of the 1,161 samples for 6 of the imputed SNPs near *TIMP3* (our strongest new locus), I observed >99.4% concordance between imputed and genotyped alleles. Association results for this set of individuals were essentially the same whether imputed or actual genotypes were used for analysis. A comparison of results with genotyped and imputed SNPs at each locus is given in Table 2.8.

Our strongest new locus maps near *TIMP3* and *SYN3* on chromosome 22 (see Fig. 3, top panel, and Table 2.3). There, I found that very common alleles (frequency of ~.94 in controls) at rs9621532 and nearby markers were associated with increased risk of AMD (OR=1.41 (1.27,1.57), overall $p=1.1x10^{-11}$, one sided p-value in newly genotyped follow-up samples $p_{follow-up}=3.3x10^{-7}$). Consistent with the expectation that GWAS tend to estimate effect sizes (the “winner’s” curse effect), I found that odds-ratios estimates in the discovery samples were larger than in the follow-up samples [46]. Results at the *TIMP3* locus were robust to a variety of analysis models (including different combinations of PCA, adjustment for previously known loci, and inclusion of age and sex as covariates, see Table 2.9), are supported by nearby directly genotyped SNPs (see Table 2.8), and remain significant when data from the companion paper are excluded from analysis (overall $p=7x10^{-11}$ excluding all Tufts/MGH data).

Two other loci also exhibited strong evidence for association. Near *LIPC* on chromosome 15, the common allele at rs493258 (frequency of ~.53 in controls) was associated with increased risk of AMD (OR=1.14 (1.09,1.20), overall $p=1.3x10^{-7}$, $p_{follow-up}=.0012$). Near *CETP* on chromosome 16, the rare allele at rs3764261 (frequency ~.36 in controls) was associated with increased risk of AMD (OR=1.19 (1.12,1.27), overall $p=7.4x10^{-7}$, $p_{follow-up}=.009$). The signals near *CETP* and *LIPC* do not reach $p <5x10^{-8}$, corresponding to genomewide significance after adjustment for one million independent tests. However, note that: (a) both *LIPC* and *CETP* show nominally significant association in follow-up samples alone; (b) less than 0.3 loci per scan are expected to reach $p < 3x10^{-7}$ by chance, suggesting that one or both of these signals are real; (c) *LIPC* association with AMD reaches genomewide significance in a companion paper; (d) in a sample of Japanese individuals, top SNPs at *CETP* (p=.001), *LIPC* (p=.10) and *TIMP3* (p=.09) trend in the right direction (see Table 2.7).

Additional experiments will be required to identify the functional alleles at each locus and the genes/pathways they impact. The challenges in identifying functional alleles are illustrated by the controversy over causal alleles near *ARMS2* (where the *PLEKHA1*, *ARMS2*, and *HTRA1* genes have been implicated [25-27, 47, 48]) and *CFH* (where non-coding variants may contribute to disease [44, 45] independently of the Y402H coding variant that was the initial focus of attention). Despite these caveats, the new loci reported here suggest biological pathways influencing disease susceptibility and possibly new therapies.

Our top novel signal maps to a large intron of the synapsin III (*SYN3*) gene involved in neurotransmission and synapse formation [49]. The SNP is located about 100 kb upstream of *TIMP3*, a metalloproteinase encoded within the same intron of *SYN3*. *TIMP3* is involved in degradation of the extracellular matrix and mutated in Sorby's Fundus Dystrophy [50], an early onset macular degenerative disease that shares clinical features with AMD but typically presents before age 40. Sorby's is extremely rare, presents with a highly penetrant autosomal dominant family history, and unlikely to be misclassified as AMD. When I excluded all patients with age of onset <60 from our sample, evidence for association at *TIMP3* was essentially unchanged. Linkage of AMD to the *TIMP3/SYN3* region has been reported previously [22]. The effects of the common alleles reported here are too small to account for the observed linkage signal, but it is possible that missed rare high penetrance alleles could reside in the same locus and explain the linkage.

Outside known loci and *TIMP3*, our two strongest signals are located near the hepatic lipase (*LIPC*) gene on chromosome 15q22 (initial evidence of association at rs493258 came from Tufts/MGH GWAS) and the cholesterylester transfer protein (*CETP*) gene on 16q21. The AMD associated alleles at these loci have been associated with HDL-c levels in blood [5, 39]. This prompted us to examine whether other common HDL-c associated polymorphisms might contribute to AMD risk. The three common alleles showing strongest association to blood HDL-c levels in an analysis of 19,840 individuals [5] also reveal evidence of association with AMD in our discovery cohort (rs173539 near *CETP* with $p=2.4x10^{-6}$; rs12678919 near *LPL* with p=.0016; rs10468017 near *LIPC* with p=.0018). Table 2.4 and Suppl. Fig. 1 show that the same clusters of SNPs (colored) associated with HDL-C (each cluster has lead SNP with $p < 5x10^{-8}$) are associated with macular degeneration; association signals are sharper for HDL-C given the much larger sample sizes (and greater power) of that analysis. Multiple common alleles near *CETP* and *LIPC* are associated independently with HDL-c levels [5]. In our sample, I find modest association of the secondary HDL-associated alleles in each of these loci with AMD (rs289714 near *CETP* with p=.062; rs2070895 near *LIPC* with p=.051). Finally, HDL-associated alleles near *ABCA1* also show evidence of association with AMD (rs1883025, p=.0026). The probability that four or more of the 14 reported HDL-associated alleles [5] would show association with AMD with p<.0026 is extremely low ($4x10^{-8}$), and the probability that the top three HDL- associated alleles would reveal association with p<.0018 is $6x10^{-9}$ (the probability of p<.14 or better, as in the replication samples alone, is .003). Importantly, since I found association specifically for alleles with the largest impact on HDL levels, it seems likely that additional signals may have been missed due to lack of power. Just as for *CETP* and *LIPC*, association signals at *LPL* and *ABCA1* were consistent in follow-up samples and discovery samples; combining all available data I observed association with $p=3.0x10^{-3}$ near *LPL* and $5.6x10^{-4}$ near *ABCA1* (Table 2.4).

Cholesterol and lipids accumulate underneath the RPE with age [51] and are present in the drusen that characterize early AMD [52, 53]. Genetic variants that impact cholesterol levels in the macula and RPE might impact drusen formation and thus modulate the risk of AMD. Since alleles near *CETP* and *LPL* associated with decreased HDL-c levels in blood appear to *increase* the risk of AMD, but alleles near *LIPC* and *ABCA1* associated with decreased HDL-c levels in blood appear to *decrease* the risk of AMD, I speculate that some alleles impact cholesterol levels in blood and in the macula in opposite directions. For example, a variant that impacts cholesterol transport between tissues could facilitate transport of HDL-c from the macula to the blood (or vice-versa). *CETP* and *LPL* play major roles in the synthesis and degradation of HDL-c, whereas *LIPC* and *ABCA1* are involved in mediating the uptake of HDL-c at the cell surface. Previously, epidemiological studies have indicated a link between cardiovascular risk factors (including HDL-c) and incidence of AMD [54, 55], but these findings have not been definitive. Our data therefore suggest an important role for HDL-c metabolism in AMD pathogenesis but also that (a) blood HDL-cholesterol levels may be a poor surrogate for the impact of HDL-c on disease risk and that (b) further work is needed to characterize the relationship between AMD and HDL-c associated alleles. It would be particularly interesting to examine samples with information on AMD classification and direct measurements of HDL-c levels in the retina.

To investigate whether identified risk alleles contributed preferentially to one disease subtype, I carried out a series of subgroup comparisons (Table 2.10). When I compared different case subgroups, I found *ARMS2* risk alleles were more common in cases with neovascular disease than in cases with large drusen (OR=1.79 (1.50-2.13), p=4.3x10^{-11}) or with geographic atrophy (OR=1.36 (1.13-1.63), p=.0009). In contrast, *CFH* risk alleles were more common in cases with geographic atrophy than in those with large drusen (OR=1.38 (1.11-1.73), p=.0012) or neovascular disease (OR=1.32 (1.08-1.64), p=.009). Risk alleles near other complement genes appeared to be somewhat more common in cases with geographic atrophy than in those with neovascularization, whereas the reverse was true for risk alleles near *TIMP3* (differences not significant). In our discovery sample, I tested for, but did not find, evidence of interactions between associated alleles at the seven loci listed in Tables 2.2 and 2.3. I also tested for, but did not find, significant interactions of risk alleles with sex and smoking.

Although the other GWAS paper did not identify *TIMP3*, targeted follow-up of the markers identified in our scan confirms our findings. The difference in the initial results of the two studies derives from different choices of markers to follow-up after the initial GWAS: a costly experiment with maximum power would involve genotyping all discovery and follow-up samples for all markers. Practical considerations meant that each study could only examine a subset of interesting markers in available follow-up samples. Ultimately, I expect that further genotyping of follow-up samples and meta-analysis of our results with those of future GWAS will identify more disease susceptibility loci. The variants identified here have only a modest impact on the risk of age-related macular degeneration. However, they do point to potentially important biological targets (such as the *TIMP3* gene and HDL-cholesterol), whose effectiveness for therapeutic intervention remains to be evaluated. I note that genes like *IL23* and *HMGCR2* are extremely effective drug targets (for the treatment of psoriasis and for LDL-cholesterol lowering medications, respectively) despite the fact that naturally occurring common variants in the corresponding loci account for only small changes in the risk of psoriasis [56] and in blood lipid levels [5], respectively.

Genetic susceptibility variants may be used to predict individual risk of AMD [57, 58]. To evaluate the effectiveness of the approach, I fitted a simple logistic regression model to the data. The model included the SNPs listed in Tables 2.2, 2.3 and 2.4 as predictors. For each SNP, a single variable encoding the number of risk alleles was modeled; no interaction terms or dominance effects were considered. In effect, the model calculates a weighted sum of risk alleles for each individual (with weights proportional to the log odds-ratio for each allele) and assigns individuals with large weighted sums the largest risk. Among the 331 individuals (10% of our sample) with the highest risk genotypes only 2 are controls and 329 are cases (see Fig. 5, top panel, for information on other genotype risk bands). Assuming a disease prevalence of 20% at age ~75, I predict that ~80% of individuals with genotypes in the top decile of risk will develop AMD, but <5% of individuals in the bottom 3 deciles will develop disease (see Fig. 5, bottom panel). Furthermore, I find that, among cases, individuals with high risk genotypes will present with severe disease more often (in the top risk decile for our sample, 15% of our cases have large drusen, 22% geographic atrophy and 63% have neovascularization) than individuals with lower risk genotypes (in the bottom risk decile, 51% of cases have large drusen, 19% geographic atrophy only and 30% have neovascularization). A productive strategy to identify rare alleles that impact disease susceptibility might involve a detailed examination of DNA sequences in individuals with severe disease but whose common variant genotypes predict low disease risk.

Despite these encouraging contrasts between individuals with low and high risk genotypes, AMD susceptibility alleles must be evaluated in population-based cohorts before genotypes can become routine diagnostic tools [58]. While trends pointing to increased frequency of severe disease in individuals with high-risk genotypes should hold, the absolute risk of developing severe disease is difficult to estimate accurately in samples collected in tertiary clinics. In the meantime, our results point to new molecular pathways and encourage new directions in the search for treatment and prevention of this common blinding disease.

Table 2.1 Summary description of discovery samples used in the genome-wide association and replication studies

	Cases						Controls			
	N	Male (%)	Age (Average)	Large Drusen (%)	Geographic Atrophy (%)	Neovascular (%)	N	Male (%)	Age (Average)	**Total**
Discovery Samples										
Michigan	786	36.9	79.8	14.2	21.6	64.2	516	41.5	76.6	1,302
Mayo Clinic	535	36.1	77.3	46.5	13.6	39.8	433	46.7	70.2	968
AREDS	440	41.0	80.8	none genotyped	26.8	73.2	0	0	0	440
Pennsylvania	396	40.4	75.7	42.7	26.3	31.0	201	45.3	76	597
Total	**2,157**	**38.2**	**78.6**	**24.5**	**21.6**	**53.9**	**1,150**	**44.1**	**74.1**	**3,307**
Parallel Discovery Samples										
Tufts/MGH†	821	46.0	80.3	none genotyped	27.5	72.5	1,709	46.0	76.0	2,530
Replication Samples										
Pittsburgh*	1,308	36.7	69.9	9.7*	18.9	70.0	229	49.8	76.7	1,537
Miami/Duke/Vanderbilt	1,157	35.1	75.7	28.3	13.6	58.2	514	40.5	68.4	1,671
Tufts/MGH II	868	40.0	79.7	none genotyped	28.3	71.7	789	40.0	73.0	1,657
Johns Hopkins*	665	32.8	75.5	21.8*	12.4	57.2	131	31.3	74.7	796
Penn-NJ	556	39.8	79.8	19.1	6.8	65.5	347	47.0	75.6	903
Oregon	515	34.0	79.8	none genotyped	27.2	72.8	263	45.0	74.0	778
Massachusetts E. E. I.	391	40.4	76.0	10.5	1.3	73.6	194	44.6	75.4	585
Spain (IDIS-Sgo)	353	46.2	76.7	none genotyped	16.1	83.9	282	44.7	75.1	635
Case Western Reserve	1,258	43.5	78.5	32.6	9.2	40.5	1,540	50.7	72.5	2,798
Total	**7,071**	**41.1**	**76.2**	**14.9**	**14.0**	**65.8**	**4,289**	**45.5**	**73.0**	**11,360**
Non-European Samples										
Japan	678	69.0	74.8	none genotyped	0.0	100.0	336	42.0	74.2	1,014
Grand Total*	**10,727**	**40.9**	**77.0**	**15.7**	**16.6**	**64.0**	**7,484**	**45.3**	**73.9**	**18,211**

*Proportions of cases with large drusen, geographic atrophy, and neovascular disease do not add up to 100.0% because 8.6% of cases from Johns Hopkins and 0.4% of cases from Pittsburgh had intermediate drusen. †The Tufts/MGH samples used here exclude 158 AREDS samples that overlap with our discovery sample.

Table 2.2 Confirmation of previously reported association signals in the discovery samples

SNP	Chrom.	Position (basepair)	Notable Nearby Genes	Alleles (risk/non-risk)	Frequency (risk allele) Cases	Frequency (risk allele) Controls	OR	p-value	λ_{sib}
Primary Association Signals									
rs10737680*	1	194,946,078	*CFH*	A/C	0.801	0.566	3.11 (2.76, 3.51)	1.6×10^{-76}	1.24
rs3793917*	10	124,209,265	*ARMS2/HTRA1*	G/C	0.371	0.164	3.40 (2.94, 3.94)	4.1×10^{-60}	1.45
rs429608	6	32,038,441	*C2/CFB*	G/A	0.920	0.842	2.16 (1.84, 2.53)	2.5×10^{-21}	1.05
rs2230199*	19	6,669,387	*C3*	C/G	0.224	0.163	1.74 (1.47, 2.06)	1.0×10^{-10}	1.06
rs2285714	4	110,858,259	*CFI*	T/C	0.464	0.395	1.31 (1.18, 1.45)	3.4×10^{-7}	1.02
Secondary Association Signals									
rs1329424*	1	194,912,799	*CFH*	T/G	0.603	0.351	1.88 (1.68, 2.10)	6.4×10^{-16}	1.11
rs9380272*	6	32,013,989	*C2/CFB*	A/G	0.016	0.012	4.31 (2.76, 6.87)	2.3×10^{-8}	1.12

Association peaks at previously reported loci. For two of these loci (near CFH and C2/CFB), I found significant secondary signals after adjusting for the strongest initial signal. At C2/CFB locus, rs9380272 shows no significant association before adjusting for the primary signal because its risk allele is in linkage disequilibrium with the protective allele at rs429608. Conditioning on either of these two SNPs enhances the signal at the other SNP. The recurrence risk ratio λsib quantifies the increase in risk to siblings of affected individuals attributable to a specific allele. For example, a λsib of 1.24 implies that alleles at the first locus are responsible for 24% increase in risk to siblings of AMD patients compared to the general population. *(imputation r2 > 0.95).

Table 2.3 New Locus with Confirmed Association to AMD ($p < 5x10^{-8}$)

		Position	Notable Nearby	Alleles	Frequency (risk allele)				
SNP	**Chrom.**	**(basepair)**	**Genes**	**(risk/non-risk)**	**Cases**	**Controls**	**OR**	**p-value§**	**λ_{sib}**
rs9621532	22	31,414,511	*SYN3/TIMP3*	A/C					
		Discovery sample (2,157 cases, 1,150 controls) …			0.964	0.943	1.81 (1.42, 2.29)	3.9×10^{-5}	1.011
		Tufts/MGH sample (821 cases, 1,709 controls) …			0.959	0.947	1.31 (0.98, 1.74)	0.008 *	1.004
		De novo replication sample (7,071 cases, 4,289 controls) …			0.959	0.947	1.33 (1.17, 1.52)	3.3×10^{-7}	1.008
		Combined sample (10,049 cases, 7,148 controls) …			**0.960**	**0.946**	**1.41 (1.27, 1.57)**	$\mathbf{1.1 \times 10^{-11}}$	**1.008**

Cochran's Q Heterogeneity Test P-value = 0.245

This table summarizes results for a new confirmed association signal near TIMP3 (overall p < 5x10-8; corresponding to an adjustment for ~1 million independent tests). *Excluding overlapping AREDS samples in the Tufts/MGH study. §P-values for the discovery and combined samples are two sided. P-values for the Tufts/MGH and de novo replication samples are one sided.

Table 2.4 Association of HDL-C loci with AMD

SNP	Chrom.	Position (basepair)	Notable Nearby Genes	Alleles (risk/non-risk)	Frequency (risk allele) Cases	Controls	OR	p-value§
rs493258	15	56,475,172	*LIPC*	C/T				
Discovery sample (2,157 cases, 1,150 controls) …					0.564	0.528	1.21 (1.10, 1.34)	0.002
Tufts/MGH sample (821 cases, 1,709 controls) …					0.579	0.524	1.25 (1.11, 1.41)	2.8×10^{-4} *
De novo replication sample (5,914 cases, 3,775 controls) …					0.562	0.575	1.10 (1.03, 1.16)	0.001
Combined sample (8,892 cases, 6,634 controls) …					**0.563**	**0.564**	**1.14 (1.09, 1.20)**	**1.3×10^{-7}**
								Cochran's Q Heterogeneity Test P-value = 0.64
rs3764261	16	55,550,825	*CETP*	A/C				
Discovery sample (2,157 cases, 1,150 controls) …					0.364	0.314	1.36 (1.26, 1.46)	1.7×10^{-6}
Tufts/MGH sample (821 cases, 1,709 controls) …					0.356	0.329	1.13 (1.00, 1.28)	0.070
De novo replication sample (4,945 cases, 1,960 controls) …					0.347	0.317	1.10 (1.00, 1.22)	0.009
Combined sample (7,923 cases, 4,819 controls) …					**0.354**	**0.316**	**1.19 (1.12, 1.27)**	**7.4×10^{-7}**
								Cochran's Q Heterogeneity Test P-value = 0.65
rs12678919	8	19,888,502	*LPL*	G/A				
Discovery sample (2,157 cases, 1,150 controls) …					0.115	0.096	1.38 (1.17, 1.63)	0.002
De novo replication sample (3,333 cases, 1,288 controls) …					0.113	0.108	1.11 (0.92, 1.35)	0.140
Combined sample (5,490 cases, 2,438 controls) …					**0.114**	**0.102**	**1.26 (1.11, 1.43)**	**0.003**
								Cochran's Q Heterogeneity Test P-value = 0.893
rs1883025	9	106,704,122	*ABCA1*	C/T				
Discovery sample (2,157 cases, 1,150 controls) …					0.739	0.705	1.25 (1.12, 1.40)	0.003
De novo replication sample (4,982 cases, 3,022 controls) …					0.752	0.741	1.10 (1.00, 1.19)	0.019
Combined sample (7,139 cases, 4,172 controls)…					**0.747**	**0.731**	**1.15 (1.07, 1.23)**	**5.6×10^{-4}**
								Cochran's Q Heterogeneity Test P-value = 0.51

* Excluding overlapping AREDS samples in the Tufts/MGH study. Before excluding these samples, Tufts/MGH results differ slightly (for example, p-value at rs493258 was 2.2x10-5). §P-values for the discovery and combined samples are two sided. P-values for the Tufts/MGH and de novo replication samples are one sided.

Table 2.5 Association results of some published candidate SNPs in our scan

Gene	SNP	RiskAllele/Other	P-value In Original Report	Original Report	P-value in Discovery Sample	P-value in Discovery Sample, After Adjusting For Known Loci	Direction Of Effect, Vs. Original Report
TLR3	rs3775291	C/T	1.2×10^{-7}	Yang Z et al. NEJM 2008	0.526	0.885	opposite
TLR4	rs4986790	G/A	0.001	Zareparsi S et al. HMG 2005	0.552	0.091	same
SERPING1	rs2511989	G/A	7.5×10^{-8}	Ennis S et al. Lancet 2008	0.944	0.923	same
ERCC6	rs3793784	G/C	0.020	Tuo J et al. PNAS 2005	0.961	0.480	same
LRP6	rs7294695	C/G	0.020	Haines JL et al. IOVS 2006	0.543	0.867	same
CX3CR1	rs3732378	A/G	0.002	Tuo J et al. FASEB J. 2004	0.150	0.100	same
IL8	rs4073	T/A	0.037	Goverdhan SV et al. Br. J. Ophthalmol 2008	0.578	0.301	same
VEGF	rs2010963	C/G	0.020	Haines JL at al. Invest Ophthalmol Vis Sci. 2006	0.302	0.320	same
VLDLR	rs2290465	C/G	0.010	Haines JL at al. Invest Ophthalmol Vis Sci. 2006	0.782	0.402	same

Previously associated SNPs near APOE and ABCA4 are not listed because they were not genotyped in our sample and could not be imputed confidently using either 1000 Genomes or HapMap reference haplotypes.

Table 2.6 Complete Results for All SNPs Where Replication Attempted

SNP	Risk/ Nonrisk	GWAS	Tufts/MGH GWAS	Tufts/MGH Replication	JHU	Penn-NJ	Oregon	Spain IDIS	MEEI	Case Western	Pitt	Miami Duke Vanderbilt	Japan	Combined
rs9621532	**A/C**	**3.9×10^{-5}**	**0.008**	**0.175**	**0.005**	**0.001**	**0.018**	**0.249**	**0.060**	**0.150**	**0.006**	**0.037**	**0.093**	**1.1×10^{-11}**
rs493258	**C/T**	**2.1×10^{-3}**	**0.0003**	**0.062**	**0.045**	**0.229**	**0.118**	**0.456**	**0.441**	**0.095**	**0.052**	**--**	**0.101**	**1.3×10^{-7}**
rs3764261	**A/C**	**1.7×10^{-6}**	**0.070**	**--**	**0.866**	**0.153**	**0.114**	**0.126**	**0.166**	**--**	**0.530**	**0.007**	**0.004**	**7.4×10^{-7}**
rs2958154	C/T	3.8×10^{-5}	0.475	--	--	--	--	--	0.453	--	0.950	0.039	--	2.0×10^{-6}
rs11878133	T/C	3.5×10^{-4}	0.002	--	--	--	--	--	0.136	--	0.531	0.091	--	4.4×10^{-6}
rs2142541	T/G	6.5×10^{-5}	0.035	--	--	--	--	--	0.265	--	0.459	0.044	--	1.1×10^{-5}
rs17628762	A/C	8.6×10^{-3}	0.0002	--	--	--	--	--	0.080	--	0.001	0.932	--	2.4×10^{-5}
rs6022766	A/C	1.5×10^{-2}	0.0005	--	--	--	--	--	--	--	0.439	0.073	--	3.3×10^{-5}
rs9973159	C/T	2.0×10^{-3}	0.010	0.485	--	--	0.071	--	--	0.095	0.453	0.071	--	4.4×10^{-5}
rs2127740	A/G	1.6×10^{-3}	0.493	--	--	--	--	--	0.100	--	--	0.361	--	5.2×10^{-5}
rs6484926	A/G	6.5×10^{-5}	0.012	0.053	0.495	--	0.875	--	--	0.060	0.385	0.520	--	6.3×10^{-5}
rs6982567	T/C	9.5×10^{-7}	--	0.060	0.162	--	0.242	--	0.375	0.100	0.047	0.845	--	8.9×10^{-5}
rs10103849	A/G	5.2×10^{-6}	0.003	--	0.265	--	0.649	--	--	0.425	0.621	0.560	--	1.7×10^{-4}
rs 8052081	G/C	3.8×10^{-5}	0.024	--	--	--	--	--	--	--	0.417	0.636	--	2.0×10^{-4}
rs655464	G/A	1.9×10^{-3}	0.004	--	--	--	--	--	--	--	0.631	0.322	--	2.7×10^{-4}
rs13142235	A/G	6.9×10^{-5}	0.045	--	--	--	--	--	--	--	--	0.675	--	4.1×10^{-4}
rs1884807	G/A	8.3×10^{-4}	0.002	--	--	--	0.711	--	--	--	0.113	0.923	--	5.2×10^{-4}
rs1883025	C/T	2.6×10^{-3}	--	--	--	--	0.020	0.119	0.798	0.135	0.109	0.429	--	5.7×10^{-4}
rs7737931	C/G	8.6×10^{-5}	0.055	--	--	--	--	--	--	--	0.382	0.638	--	6.5×10^{-4}
rs12914520	T/C	1.3×10^{-3}	0.002	--	0.050	--	0.84	--	--	--	0.466	0.714	--	7.6×10^{-4}
rs7704053	A/G	5.7×10^{-2}	0.0001	--	--	--	--	--	--	--	0.631	0.353	--	7.7×10^{-4}
rs17121872	A/G	1.3×10^{-4}	0.003	--	--	--	--	--	--	--	0.908	0.825	--	1.0×10^{-3}
rs16848791	G/T	1.4×10^{-4}	0.006	0.343	0.139	--	0.319	--	0.841	0.305	0.834	0.549	--	1.3×10^{-3}
rs10468017	C/T	1.8×10^{-3}	--	--	--	0.170	0.105	0.671	0.365	--	0.050	0.484	--	1.5×10^{-3}
rs12678919	G/A	1.8×10^{-3}	--	--	--	--	0.392	0.416	--	--	0.243	0.193	0.334	3.2×10^{-3}
rs12001032	T/C	8.5×10^{-4}	0.023	0.751	--	--	0.302	--	--	--	0.112	0.779	--	5.4×10^{-3}
rs2892715	G/A	8.5×10^{-6}	0.711	--	--	--	--	--	0.872	--	0.642	0.401	--	2.0×10^{-2}
rs6445063	C/T	1.4×10^{-5}	0.814	--	--	--	--	--	--	--	0.446	0.667	--	2.8×10^{-2}

Table 2.7 Sample by Sample Results for Newly Reported Loci

PART 1/3

rs9621532 (A/C) near TIMP3

	Cases				Controls					
	A/A	A/C	C/C	P(A)	A/A	A/C	C/C	P(A)	OR	P
Discovery	2005	149	3	0.964	1022	125	3	0.943	1.81 (1.42, 2.29)	3.9×10^{-5}
Tufts/MGH	732	62	4	0.957	1466	163	3	0.947	1.31 (0.98, 1.74)	0.016
Tufts/MGH II	777	69	4	0.955	703	75	1	0.951	1.09 (0.85, 1.51)	0.350
Johns Hopkins	626	37	1	0.971	113	16	0	0.938	2.21 (1.22, 4.03)	0.008
Penn-NJ	510	46	0	0.959	295	52	0	0.925	1.90 (1.26, 2.86)	0.002
Oregon	452	24	0	0.975	229	23	0	0.954	1.88 (1.05, 3.37)	0.036
Spain(IDIS-Sgo)	330	17	0	0.976	259	17	0	0.969	1.27 (0.64, 2.50)	0.498
*Massachusetts E.E. I.***	*345*	*39*	*0*	*0.949*	*163*	*26*	*1*	*0.926*	*1.49 (0.90, 2.46)*	*0.119*
Case Western Reserve	1124	95	8	0.955	1370	147	3	0.950	1.12 (0.87, 1.44)	0.300
*Pittsburgh***	*169*	*10*	*0*	*0.972*	*130*	*10*	*1*	*0.957*	*1.55 (0.66, 3.63)*	*0.011*
*Miami/Duke/Vanderbilt***	*629*	*69*	*4*	*0.945*	*218*	*30*	*1*	*0.936*	*1.18 (0.77, 1.81)*	*0.074*
Japan	617	37	1	0.970	303	27	0	0.959	1.38 (0.84, 2.28)	0.195

Test of heterogeneity:	Q	d.f.	p.value
	11.47	9	0.2448

PART 2/3

rs493258 (C/T) near LIPC

	Cases				Controls					
	C/C	C/T	T/T	P(C)	C/C	C/T	T/T	P(C)	OR	P
Discovery	691	1053	413	0.564	323	569	258	0.528	1.21 (1.10, 1.34)	0.002
Tufts/MGH	260	391	147	0.579	470	782	380	0.524	1.25 (1.11, 1.41)	0.001
Tufts/MGH II	254	428	172	0.548	213	387	182	0.520	1.12 (0.98, 1.29)	0.124
Johns Hopkins	203	315	119	0.566	35	58	33	0.508	1.26 (0.96, 1.66)	0.090
Penn-NJ	193	273	90	0.593	110	179	58	0.575	1.08 (0.89, 1.31)	0.458
Oregon	167	228	104	0.563	78	111	63	0.530	1.14 (0.92, 1.42)	0.235
Spain(IDIS-Sgo)	104	164	79	0.536	82	128	64	0.533	1.01 (0.81, 1.27)	0.911
*Massachusetts E.E. I.***	*128*	*159*	*88*	*0.553*	*52*	*88*	*35*	*0.549*	*1.02 (0.79, 1.31)*	*0.822*
Case Western Reserve	366	595	217	0.563	404	726	300	0.536	1.12 (1.00, 1.24)	0.190
*Pittsburgh***	*66*	*70*	*39*	*0.577*	*52*	*64*	*35*	*0.556*	*1.09 (0.80, 1.49)*	*0.104*
*Miami/Duke/Vanderbilt***	*222*	*337*	*131*	*0.566*	*65*	*149*	*31*	*0.569*	*0.99 (0.80, 1.21)*	--
Japan	35	200	408	0.210	10	102	217	0.185	1.17 (0.94, 1.46)	0.202

Test of heterogeneity:	Q	d.f.	p.value
	6.96	9	0.6412

rs3764261 (A/C) near CETP

	Cases				Controls					
	A/A	A/C	C/C	P(A)	A/A	A/C	C/C	P(A)	OR	P
Discovery	296	979	882	0.364	118	486	546	0.314	1.36 (1.26, 1.46)	1.7 × 10^{-6}
Tufts/MGH	104	377	340	0.356	216	784	709	0.329	1.13 (1.00, 1.28)	0.140
Tufts/MGH II	--	--	--	--	--	--	--	--	--	--
Johns Hopkins	87	293	261	0.364	24	50	48	0.402	0.85 (0.70, 1.04)	0.268
Penn-NJ	58	251	247	0.330	31	151	165	0.307	1.11 (0.96, 1.29)	0.306
Oregon	60	252	197	0.365	26	117	110	0.334	1.15 (0.98, 1.34)	0.227
Spain(IDIS-Sgo)	33	145	170	0.303	22	107	147	0.274	1.15 (0.97, 1.37)	0.252
*Massachusetts E.E. I. ***	*45*	*178*	*163*	*0.347*	*17*	*87*	*86*	*0.318*	*1.14 (0.95, 1.37)*	*0.332*
Case Western Reserve	--	--	--	--	--	--	--	--	--	--
*Pittsburgh ***	*24*	*77*	*69*	*0.368*	*18*	*55*	*70*	*0.318*	*1.25 (0.99, 1.58)*	*0.940*
*Miami/Duke/Vanderbilt ***	--	--	--	--	--	--	--	--	--	--
Japan	31	228	395	0.222	17	80	236	0.171	1.39 (1.17, 1.65)	0.008

Test of heterogeneity:	Q	d.f.	p.value
	4.18	6	0.6524

PART 3/3

rs12678919 (G/A) near LPL

	Cases				Controls					
	G/G	G/A	A/A	P(G)	G/G	G/A	A/A	P(G)	OR	P^*
Discovery	23	448	1686	0.115	9	206	939	0.097	1.38 (1.17, 1.63)	0.002
Tufts/MGH	--	--	--	--	--	--	--	--	--	--
Tufts/MGH II	--	--	--	--	--	--	--	--	--	--
Johns Hopkins	--	--	--	--	--	--	--	--	--	--
Penn-NJ	--	--	--	--	--	--	--	--	--	--
Oregon	6	85	416	0.096	2	42	208	0.091	1.06 (0.73, 1.53)	0.783
Spain(IDIS-Sgo)	2	81	162	0.173	5	63	149	0.168	1.04 (0.74, 1.46)	0.832
*Massachusetts. E.E. I. ***	--	--	--	--	--	--	--	--	--	--
Case Western Reserve	--	--	--	--	--	--	--	--	--	--
*Pittsburgh ***	*1*	*32*	*141*	*0.098*	*1*	*21*	*127*	*0.077*	*1.30 (0.75, 2.27)*	*0.486*
*Miami/Duke/Vanderbilt ***	*5*	*139*	*555*	*0.107*	*3*	*40*	*203*	*0.093*	*1.17 (0.83, 1.66)*	*0.385*
Japan	10	141	496	0.124	6	64	253	0.118	1.06 (0.80, 1.42)	0.668

Test of heterogeneity:	Q	d.f.	p.value
	0.62	3	0.8926

rs1883025 (G/A) near ABCA1

	Cases				Controls					
	G/G	G/A	A/A	P(G)	G/G	G/A	A/A	P(G)	OR	P[*]
Discovery	1171	845	141	0.739	571	480	99	0.705	1.25 (1.12, 1.40)	0.003
Tufts/MGH	--	--	--	--	--	--	--	--	--	--
Tufts/MGH II	--	--	--	--	--	--	--	--	--	--
Johns Hopkins	--	--	--	--	--	--	--	--	--	--
Penn-NJ	--	--	--	--	--	--	--	--	--	--
Oregon	299	180	27	0.769	126	111	15	0.720	1.29 (1.01, 1.65)	0.039
Spain(IDIS-Sgo)	174	155	17	0.727	143	97	35	0.696	1.16 (0.91, 1.49)	0.238
*Massachusetts. E.E. I. ***	*205*	*138*	*42*	*0.712*	*98*	*79*	*10*	*0.735*	*0.89 (0.67, 1.17)*	*0.405*
Case Western Reserve	713	418	67	0.770	821	563	77	0.755	1.09 (0.96, 1.23)	0.270
*Pittsburgh ***	*104*	*66*	*7*	*0.774*	*89*	*45*	*12*	*0.764*	*1.06 (0.73, 1.53)*	*0.318*
*Miami/Duke/Vanderbilt ***	*378*	*275*	*47*	*0.736*	*130*	*98*	*20*	*0.722*	*1.08 (0.86, 1.36)*	*0.858*
Japan	--	--	--	--	--	--	--	--	--	--
Test of heterogeneity:	Q	d.f.	p.value							
	4.25	5	0.5137							

** Note that for datasets that include related individuals (Pittsburgh, Miami/Due/Vanderbilt and Massachusetts. E.E. I.), this samples counts include only unrelated individuals. Thus, the results differ from those in Table 3 in the main paper where all available samples were analyzed using the method of Thornton and McPeek. The tabulated p-values are calculated from the complete family data set. P values are two sided.

Table 2.8 Best genotyped proxy SNPs for reported loci

SNP	Chrom	Position	Gene	P-value at Imputed SNP	Best Genotyped Proxy	Allele1/ Allele2	Cases 1/1 1/2 2/2	Controls 1/1 1/2 2/2	Rsq	P-value at Genotyped SNP*
rs10737680	1	194,946,078	*CFH*	1.6×10^{-76}	rs1329428	A/G	86/685/1384	214/571/365	1.00	5.2×10^{-76}
rs3793917	10	124,209,265	*ARMS2/HTRA1*	4.1×10^{-60}	rs6585827	G/A	377/993/782	335/557/256	0.32	7.5×10^{-22}
rs429608	6	32,038,441	*C2/CFB*	2.5×10^{-21}	rs429608	A/G	18/311/1827	27/311/812	1.00	2.5×10^{-21}
rs2230199	19	6,669,387	*C3*	1.0×10^{-10}	rs2250656	G/A	139/775/1243	107/491/552	0.08	1.3×10^{-7}
rs2285714	4	110,858,259	*CFI*	3.4×10^{-7}	rs2285714	T/C	462/1076/617	187/534/429	1.00	3.4×10^{-7}
rs1329424	1	194,912,799	*CFH*	6.4×10^{-16}	rs2019724	G/A	271/998/886	432/546/172	0.79	1.3×10^{-14}
rs9380272	6	32,013,989	*C2/CFB*	2.3×10^{-8}	rs9332702	G/C	0/67/2089	0/27/1123	0.50	1.1×10^{-7}
rs9621532	22	31,414,511	*SYN3/TIMP3*	3.9×10^{-5}	rs135150	C/T	45/519/1592	32/330/787	0.14	0.001
rs493258	15	56,475,172	*LIPC*	2.1×10^{-3}	rs1532085	A/G	255/949/951	179/509/462	0.64	0.002
rs3764261	16	55,550,825	*CETP*	1.4×10^{-6}	rs3764261	T/G	296/979/882	118/485/546	1.00	1.4×10^{-6}

*The second cluster is conditional on the five SNPs in the first cluster. The third cluster is conditional on the 7 SNPs above. Marginally, the SNPs in second cluster are not significant.

Table 2.9 Association Results in Discovery Sample for Different Analysis Models

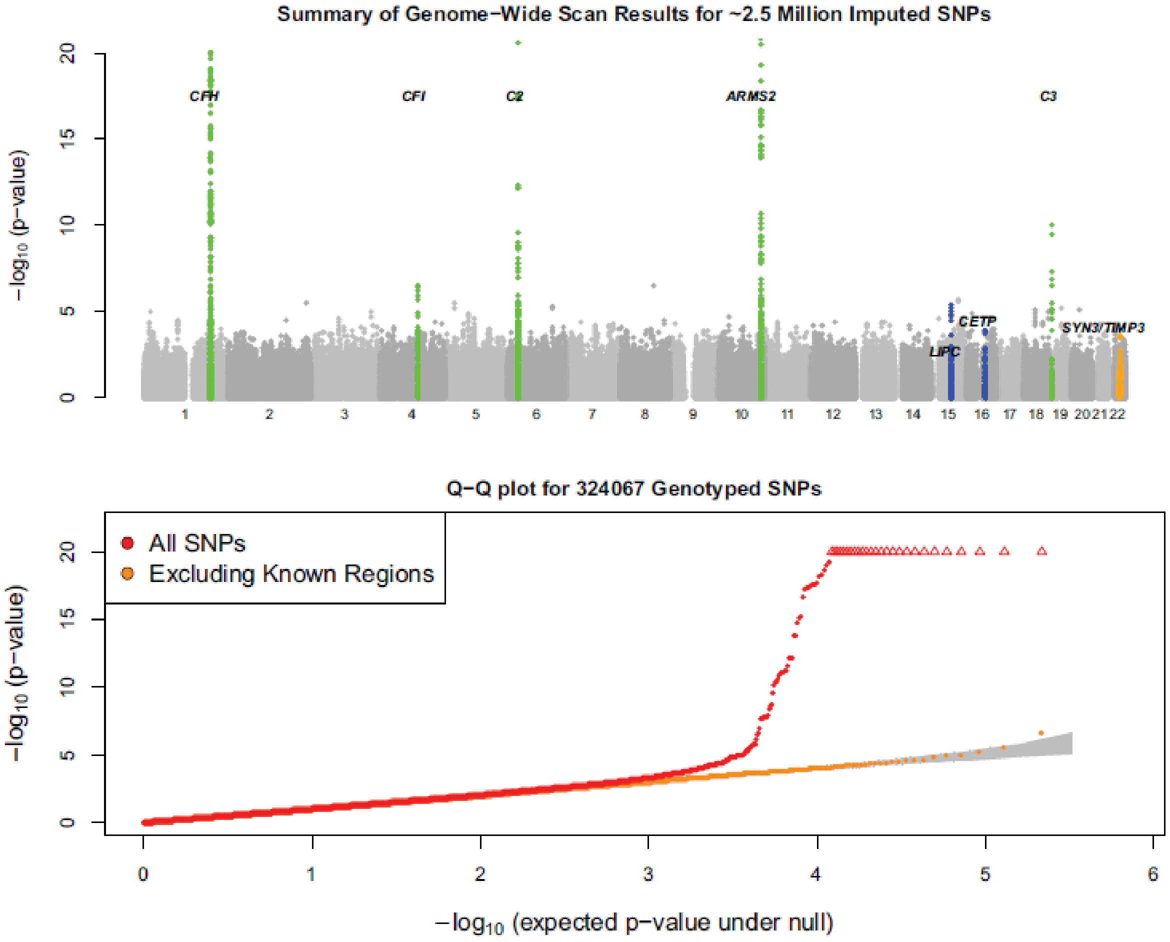
		Analysis Covariates			
SNP	Notable Nearby Genes	None	Principal Components of Ancestry (PCA)	PCA and Index SNPs at Previous Loci	PCA, Previous Loci Age and Sex
rs10737680	*CFH*	2.5×10^{-78}	1.6×10^{-76}	--	--
rs3793917	*ARMS2*	1.7×10^{-60}	4.1×10^{-60}	--	--
rs429608	*C2/CFB*	4.7×10^{-21}	2.5×10^{-21}	--	--
rs2230199	*C3*	3.6×10^{-11}	1.0×10^{-10}	--	--
rs2285714	*CFI*	8.0×10^{-8}	3.4×10^{-7}	--	--
rs9621532	*TIMP3*	5.9×10^{-5}	2.6×10^{-4}	4.5×10^{-5}	7.1×10^{-4}
rs493258	*LIPC*	5.1×10^{-3}	6.9×10^{-3}	3.6×10^{-3}	1.1×10^{-2}
rs3764261	*CETP*	5.8×10^{-5}	1.2×10^{-4}	4.6×10^{-6}	9.5×10^{-6}
rs12678919	*LPL*	1.7×10^{-2}	2.0×10^{-2}	4.0×10^{-3}	2.9×10^{-3}
rs1883025	*ABCA1*	3.4×10^{-3}	6.4×10^{-3}	5.2×10^{-3}	4.9×10^{-3}

Table 2.10 Evaluation of Association of Loci with $p < 5x10^{-8}$ Overall In Specific AMD Subtypes (OR, 95 C.I., p-value)

	rs10737680 (*CFH*) Alleles (A/C)	rs3793917 (*ARMS2*) Alleles (G/C)	rs429608 (*C2/CFB*) Alleles (G/A)	rs2230199 (*C3*) Alleles (C/G)	rs2285714 (*CFI*) Alleles (T/C)	rs9621532 (*TIMP3*) Alleles (T/C)
Large Drusen (529) vs Control (1150)	2.69 (2.27,3.20) 2.2×10^{-29}	2.36 (1.94,2.87) 4.4×10^{-26}	2.03 (1.59,2.59) 1.8×10^{-8}	1.66 (1.32,2.08) 1.2×10^{-5}	1.26 (1.08,1.45) 2.3×10^{-3}	1.47 (1.03,2.12) 0.03
GA (465) vs Control (1150)	**3.85** **(3.15,4.71)** **1.0×10^{-39}**	3.68 (3.07,4.42) 1.7×10^{-44}	**2.46** **(1.95,3.10)** **2.0×10^{-14}**	**2.00** **(1.62,2.46)** **6.3×10^{-11}**	**1.38** **(1.21,1.57)** **1.4×10^{-6}**	1.31 (0.91,1.88) 0.14
Neovascular (1163) vs Control (1150)	3.15 (2.73,3.63) 1.4×10^{-57}	**4.28** **(3.63,5.04)** **1.1×10^{-66}**	2.16 (1.79,2.61) 1.3×10^{-15}	1.67 (1.38,2.00) 7.9×10^{-8}	1.34 (1.19,1.50) 1.3×10^{-6}	**1.91** **(1.42,1.91)** **1.9×10^{-5}**
GA (465) vs Large Drusen (529)	1.38 (1.11,1.73) 4.3×10^{-3}	1.26 (1.02,1.55) 0.032	1.12 (0.81,1.55) 0.48	1.22 (0.93,1.60) 0.15	1.09 (0.91,1.30) 0.36	1.12 (0.72,1.73) 0.62
Neovascular (1163) vs Large Drusen (529)	1.13 (0.95,1.35) 0.16	1.79 (1.50,2.13) 4.3×10^{-11}	1.07 (0.83,1.39) 0.59	0.99 (0.80,1.24) 0.95	1.06 (0.92,1.23) 0.43	1.30 (0.88,1.92) 0.19
Neovascular (888) vs GA (465)	0.76 (0.61,0.93) 0.009	1.36 (1.13,1.63) 0.0009	0.90 (0.67,1.20) 0.47	0.78 (0.62,1.00) 0.046	0.95 (0.81,1.12) 0.54	1.39 (0.93,1.39) 0.11

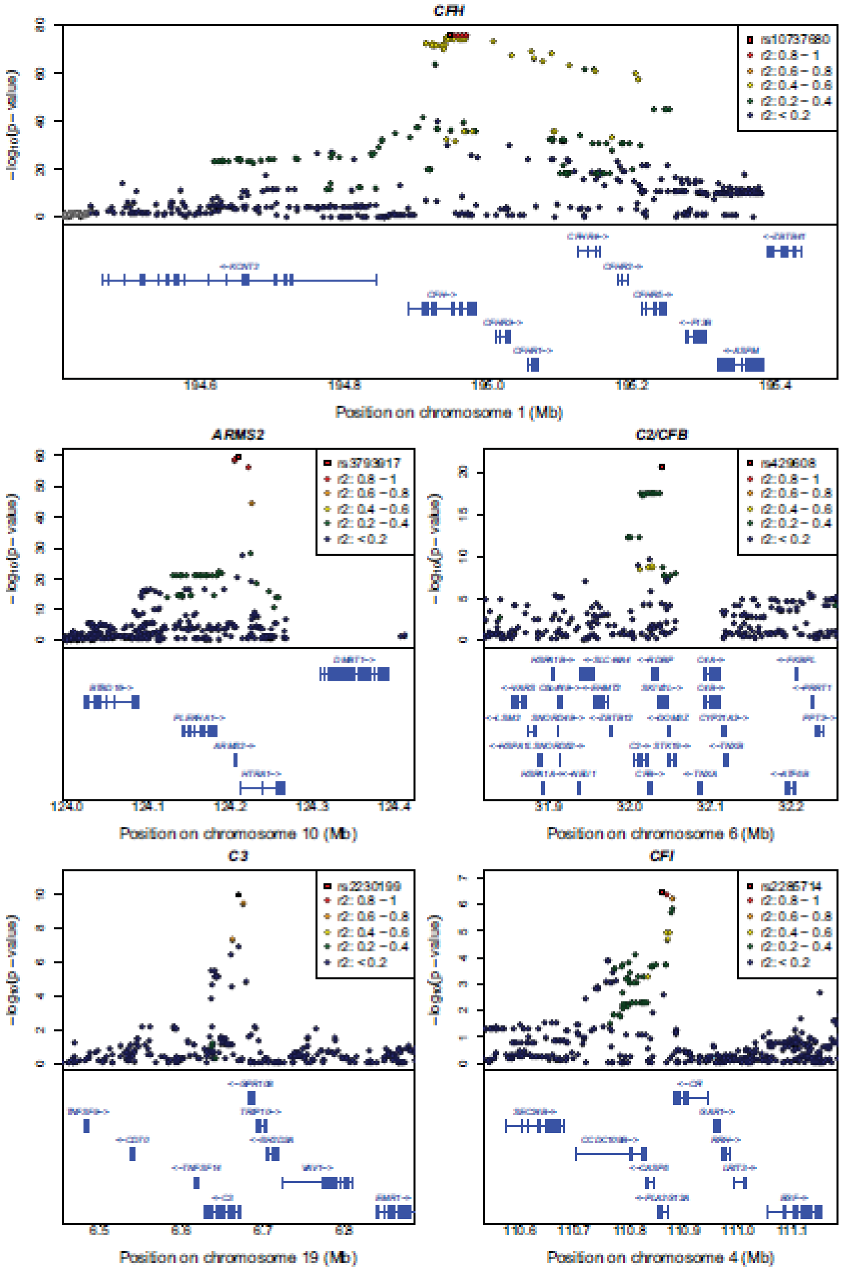
The entry corresponding to the largest odds ratio in each column is bolded.

Figure 2.1 Summary of genomewide association scan results



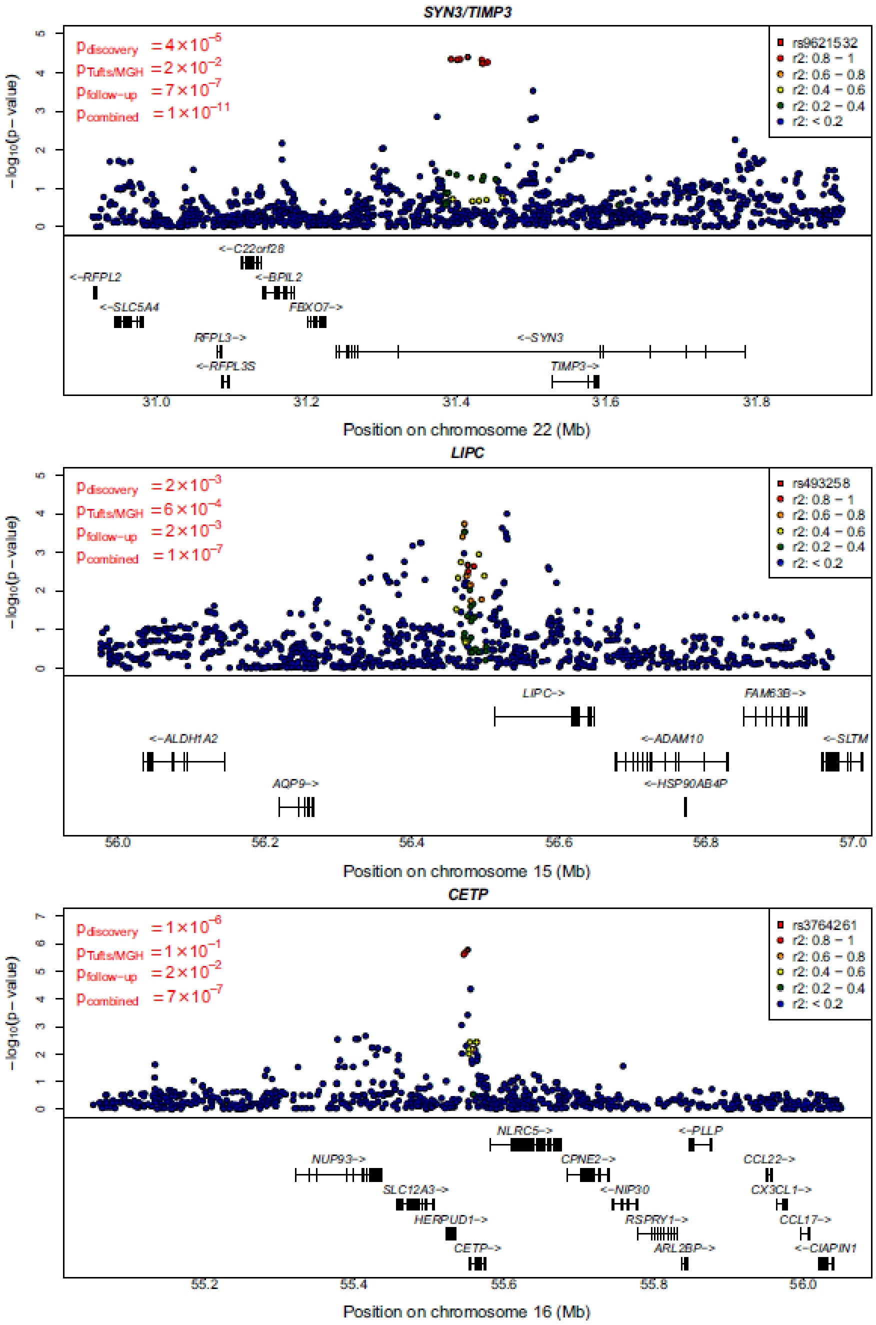
The top panel summarizes the significance of the association signal at each examined SNP in the discovery samples. The five known loci are highlighted in green. The three strongest loci after follow (*TIMP3*, *LIPC*, *CETP*) are highlighted in blue. The bottom panel displays a quantile-quantile plot for test statistics. The shaded region in the bottom panel corresponds to a 90% confidence interval for the test statistics.

Figure 2.2 Regional plots for association signals in five previously reported loci



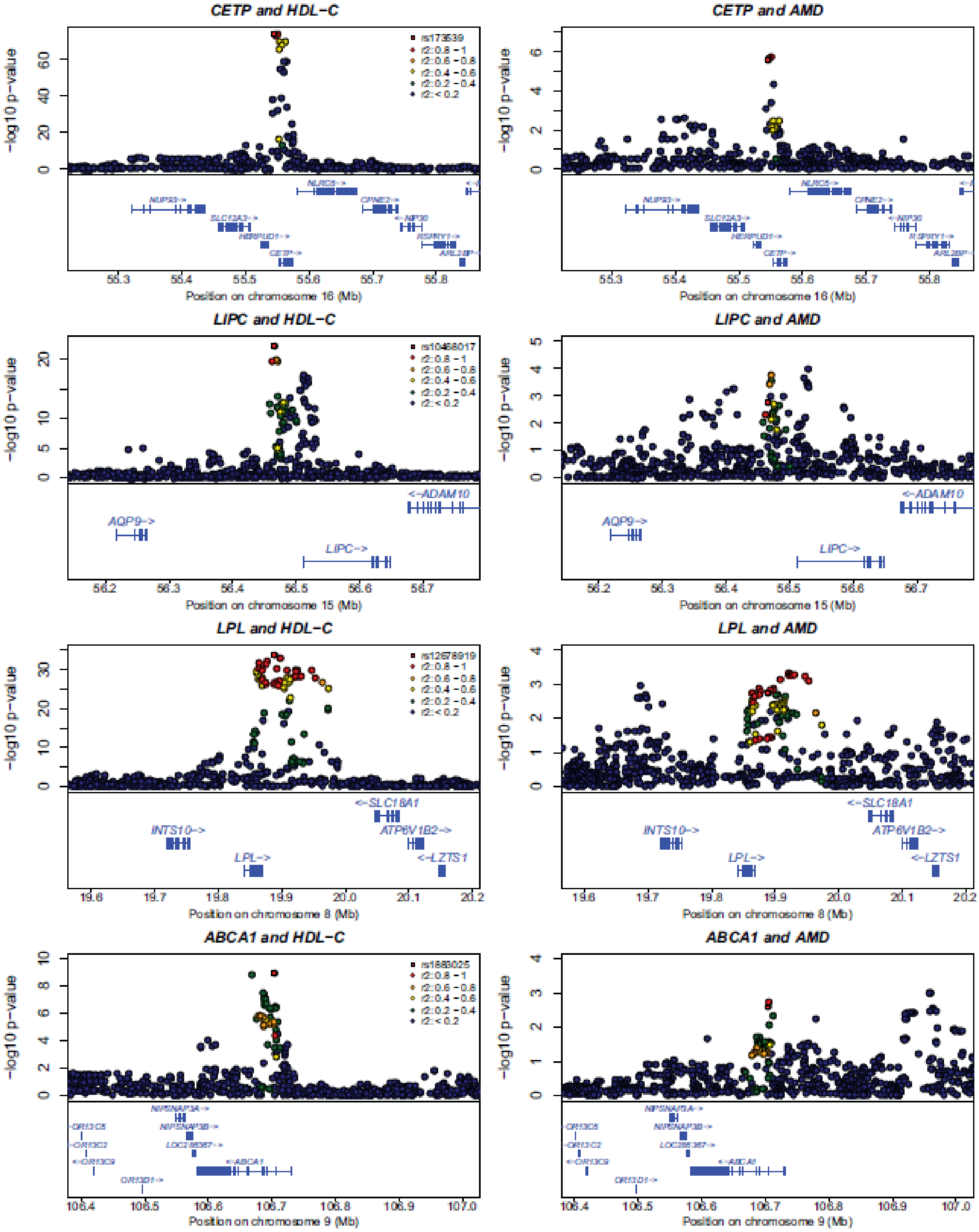
Detailed plots of association in the discovery samples in five confirmed regions (*CFH*, *ARMS*, *C2/CFB*, *C3* and *CFI*) are shown. The most significant SNP in each region is highlighted in a red square and other SNPs are drawn as colored circles reflecting linkage disequilibrium (LD) with the top selected SNP. Exons and transcript direction for genes in each region are indicated in bottom panel.

Figure 2.3 Regional plot for association signals in the three new loci



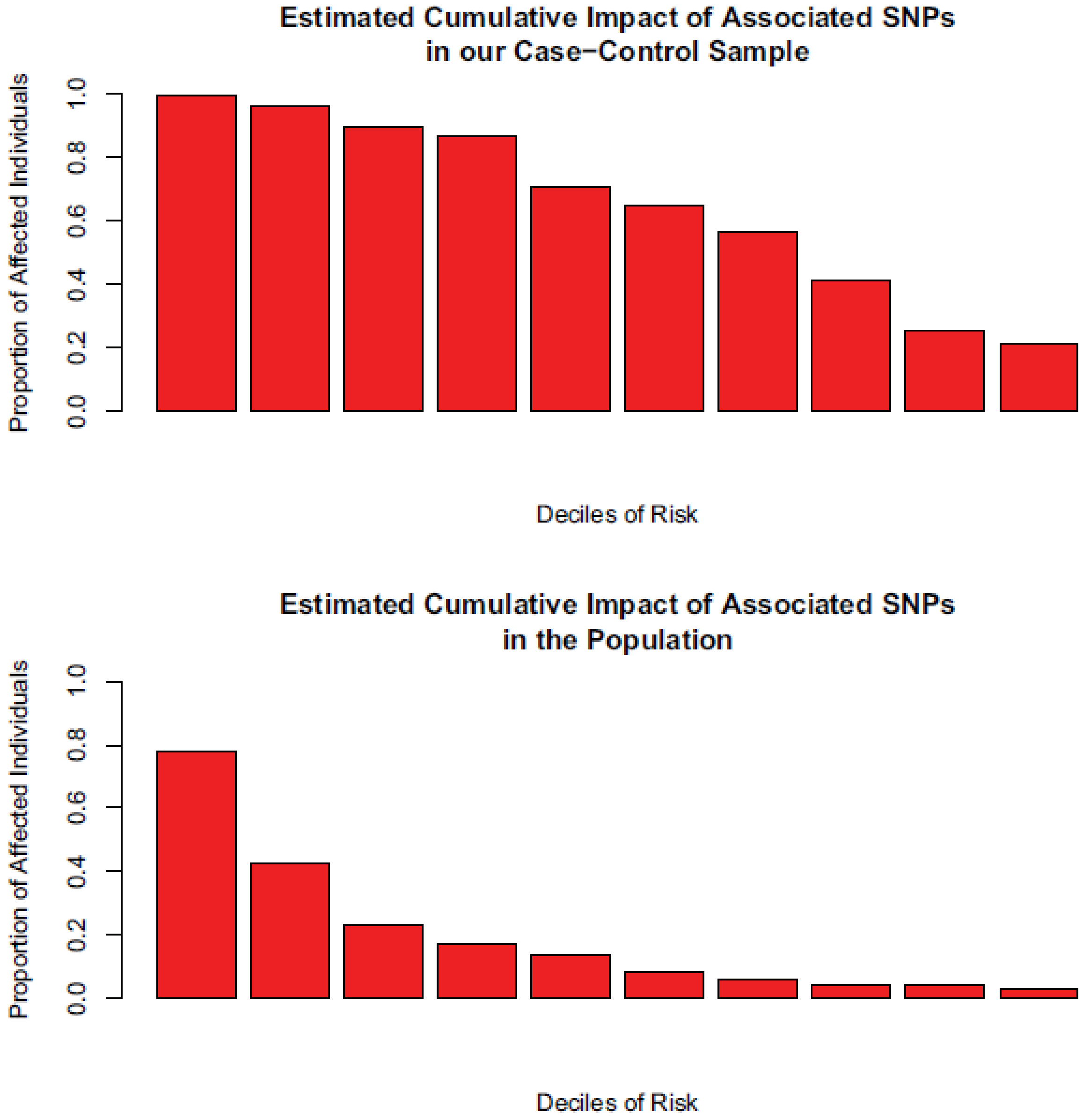
Detail plots for the regions surrounding the *SYN3/TIMP3*, *LIPC* and *CETP* regions. Original, follow-up, and combined p-values for the SNP selected for replication are indicated on the left. Discovery sample p-values for the index SNP and other nearby SNPs are plotted.

Figure 2.4 Regional plot of association signals in HDL-c and AMD



Detailed plots comparing HDL-cholesterol association signals (from the discovery sample of Kathiresan et al27; left column) and AMD association signals (from the discovery sample in the scan reported here; right column). The same marker and linkage disequilibrium proxies are highlighted in each row.

Figure 2.5 Multi-locus genotypes and disease risk



The top panel summarizes the proportion of affected individuals in each risk decile, with the highest risk decile on the left, when our sample is segregated according to the risk of disease predicted by a simple logistic regression model. The bottom panel makes equivalent predictions at the population level, after weighting cases and controls to take into account that our sample is enriched for cases (see methods for details).

Chapter 3
Graphical Browser for GWAS with High-dimensional Phenotypes

The content of this chapter has been published in Chen et al. 2009 [59].

3.1 Introduction

Recently, genome wide association scans (GWAS) have been used to successfully dissect a variety of complex traits, ranging from discrete clinical outcomes such as asthma and diabetes [60-62] to continuous traits as diverse as height, weight, global gene expression and blood lipid levels [63, 64]. The amount of information generated in these studies is staggering and interpreting their results requires efficient computational tools for data analysis and visualization. This challenge is most noticeable when high-dimensional data (such as microarray gene expression data or proteomics data) is analyzed. In this case, the results of whole genome association studies can include billions of data points [63, 65]. Realizing the full benefits of these studies requires an efficient way to share data among collaborators and with other researchers, both before and after the data is published. Here, I present a tool that facilitates interactive browsing of the results from whole genome association studies. To illustrate the capabilities of our browser, I used it to create an interactive interface for the results of a recent genome-wide association study of global gene expression. The objective of the Dixon et al (2007) study was to build a database that would allow researchers to systematically examine potential effects of disease-associated variants on transcript expression and our interactive browser makes it easy for many researchers to explore the data.

A diverse set of statistical methods can be used to examine the association between phenotypes of interest and SNP data. For example, chi-squared test statistics, p-values, effect size estimates and their standard errors, as well as SNP specific heritability estimates are all commonly reported in GWAS studies. When there are tens of thousands of phenotypic outcomes and hundreds of thousands SNPs, the result set is usually very large, containing several million statistics and easily totaling several gigabytes. These datasets can be integrated into specialized local databases for further investigation, but it can be challenging for researchers without extensive database or programming skills to access results. Our GWAS GUI (Graphic User Interface) is intended to provide a convenient tool for interacting with arbitrary GWAS result sets and to facilitate searches and displays of GWAS results in graph or tabular form. I hope our tool will facilitate data sharing within collaborative groups and with the public at large.

3.2 Features of GWAS GUI Browser

Our GWAS GUI browser is an interactive package that facilitates rapid interactive browsing of whole-genome association study results. It is designed to handle thousands of phenotypes, and thus can handle very rich datasets, such as those where global surveys of gene expression are combined with genome-wide SNP data. The browser also allows users to interact with the results of simpler scans, such as scans that focus on a single discrete outcome or a small number of related traits. To evaluate the program, I have applied it to several large datasets, including a study evaluating the association between 408,273 SNPs and the levels of 54,675 transcripts representing 20,599 known genes and assessed in lymphoblastoid cell lines from ~400 children [63]. After this initial evaluation, I released an early version of the program, named the mRNA by SNP browser (MRBS), when the Dixon et al (2007) paper was published. In addition to the visualization tool, the full GWAS GUI browser includes a data preparation tool that can be used to organize tabulated results into an indexed database for rapid browsing. There are two main browsing interfaces within our browser: (a) an interface that retrieves all results for a specific trait and (b) an interface that retrieves all results in a specific genomic region. In either view, results can typically be retrieved almost instantaneously. In the "trait-centric" view, the browser can tabulate and sort a summary of user provided association test results (e.g. effect size, standard error, heritability estimates, test statistics, and p-value) and quickly generate plots that summarize the distribution of a user specified test statistics along the genome. Alternatively, in the "position-centric" view, the browser can tabulate all significant association test statistics (using a user defined threshold) in a target region and plot the results for multiple traits. Optionally, information such as the location of nearby genes can also be displayed (see Figure 3.1). For convenience, both interfaces allow the browser to link the results to external databases chosen by the user, such as the University of California Santa Cruz (UCSC) genome browser, where users can examine the genomic context of each result in detail. When the user requests a SNP that is not included in the current dataset, LD and tag information from the International HapMap Consortium can be used to suggest a backup tag-SNP. Figure 3.1 is an illustration of the browser interface after searching for a specific SNP position using the “position-centric” view. Four SNPs of interest have been highlighted by the user in the tabular view (bottom-left) and are circled in the graphical view.

3.3 Examples of Application

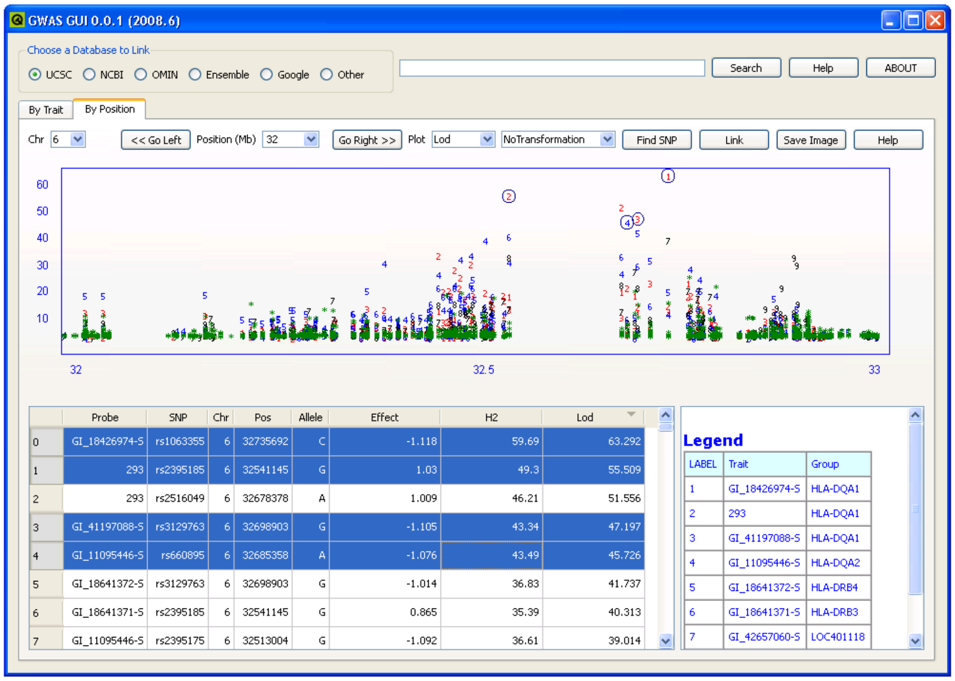
Allowing large groups of scientists to browse and interact with the results of large multi-dimensional GWAS can be extremely helpful. For example, prior to the publication of the Dixon et al. (2007) gene expression paper I used an early version of our browser to share preliminary results with several colleagues. This led to the observation that SNPs in an intergenic region on chromosome 5p13 that were associated with Crohn's Disease were also associated with transcript levels of PTGER4 suggesting that PTGER4 may be the primary candidate gene for Crohn’s disease on chromosome 5. The Crohn’s associated SNPs are more than 200 Kb away from the nearest annotated gene. The result is published and described in detail elsewhere [66]. Since then, many others have browsed our results resulting in several potential links between SNPs, human disease, and mRNA transcript levels.

The current version of the GWAS GUI browser program is not restricted to gene-expression data, but is intended as a general tool that provides graphical overviews of whole-genome association study results for arbitrary phenotypes. The extended program allows users to load their own data files, tests statistics and genomic annotation files into the browser in a standardized text format. Generally, the traits can be any outcomes of interest, such as case-control indicators, expression values, and many other continuous or categorical measurements. Arbitrary meta-data about each trait can be tracked and displayed. I expect the browser will be particularly helpful when multiple related traits are studied. In this setting, the browser simplifies the initial comparison of signals for different related traits in regions of interest.

3.4 Implementation

The GWAS GUI browser program was implemented in C++ using the Qt4 toolkit (open-source version 4.3 Trolltech Inc.). It has been tested on Windows, Linux and Mac workstations. The system requirements depend on the size of input data sets range which can range from a dataset examining a single trait dataset and hundreds of thousands of genetic markers to large scale genome wide gene-expression datasets with tens of thousands of traits and markers. On a a modern Windows Workstation, the initial indexing of a set of results generated by PLINK [67], MERLIN [68] or another whole genome analysis tools and including ~300,000 SNPs requires ~200 MB of RAM and five to ten minutes of computing time. After indexing, opening the same dataset and browsing the data should be nearly instantaneous and require only 60 MB RAM.

Figure 3.1 An illustration of the GWAS GUI browser interface



● UCSC ○ NCBI ○ OMIN ○ Ensemble ○ Google ○ Other

Chr 6 Position (Mb) 32 Plot Lod NoTransformation

	Probe	SNP	Chr	Pos	Allele	Effect	H2	Lod
0	GI_18426974-S	rs1063355	6	32735692	C	-1.118	59.69	63.292
1	293	rs2395185	6	32541145	G	1.03	49.3	55.509
2	293	rs2516049	6	32678378	A	1.009	46.21	51.556
3	GI_41197088-S	rs3129763	6	32698903	G	-1.105	43.34	47.197
4	GI_11095446-S	rs660895	6	32685358	A	-1.076	43.49	45.726
5	GI_18641372-S	rs3129763	6	32698903	G	-1.014	36.83	41.737
6	GI_18641371-S	rs2395185	6	32541145	G	0.865	35.39	40.313
7	GI_11095446-S	rs2395175	6	32513004	G	-1.092	36.61	39.014

Legend

LABEL	Trait	Group
1	GI_18426974-S	HLA-DQA1
2	293	HLA-DQA1
3	GI_41197088-S	HLA-DQA1
4	GI_11095446-S	HLA-DQA2
5	GI_18641372-S	HLA-DRB4
6	GI_18641371-S	HLA-DRB3
7	GI_42657060-S	LOC401118

This example demonstrates how to display the results for a specific region. The several largest statistics are highlighted in blue circles by selecting the corresponding rows. The top transcripts ordered by maximum statistics within that region are tabulated in the right panel with corresponding genes.

Chapter 4
Genotype Calling and Haplotyping in Parent-Offspring Trios

4.1 Introduction

In the past decade, genome-wide association studies (GWAS) have identified associations between >1,000 common variants and a variety of complex traits and diseases [1, 69]. Next generation sequencing technologies enable researchers to look beyond the common variants typically evaluated in genomewide association studies and systematically consider the contributions of rarer variation[70, 71]. The ability to systematically examine these rare variants may improve our understanding of complex traits, by identifying the underlying biological mechanisms more completely and by improving our ability to predict individual outcomes[6, 72].

Next generation sequencing technology can be used to study rare variation either by directly sequencing study samples or using genotype imputation approaches to impute variants observed in a small number of reference samples into larger sets of phenotyped individuals. In the first case, it is of primary importance to obtain accurate genotypes for each of the studied individuals. In the second case, it is also important to obtain accurate haplotypes for each sequenced individual, since these are a key reagent in the imputation based analyses that follow. Since short reads from massively parallel technologies typically contain errors, some degree of redundancy is required to ensure adequate estimates of genotypes and haplotypes and sequencing depth is a key variable in determining the accuracy of estimated genotypes and haplotypes [14, 73].

Most ongoing sequencing studies have focused on the analysis of unrelated samples. An example of the utility of sequencing related individuals is the work of Roach et al[74]. By sequencing a nuclear family, including two children with Miller syndrome and their parents, Roach et al [74] were able to identify the majority of sequencing errors and narrow their search for functional alleles. I reasoned that, by imposing constraints of Mendelian inheritance and by ensuring that many rare variants would be observed in multiple individuals, sequencing parent-offspring trios would improve genotype and haplotype calls, particularly in cases where each individual is sequenced at low depth[75, 76].

Here, I describe a new statistical method for estimating individual genotypes and haplotypes when next generation sequence data is available on parent-offspring trios. I organize this chapter as follows. First, I will describe how a hidden Markov model designed for analyses of sequence data in unrelated individuals can be extended to trios and parent offspring pairs in a computationally efficient manner. Second, I evaluate the model in a variety of simulated datasets – varying sequencing depth, sequencing error rates and sample sizes. Third, I evaluate my method in data from the ongoing Sardinia sequencing project. Our results show that my method substantially outperforms existing approaches that ignore familial relatedness.

4.2 Material and Methods

The Pipeline for SNP Discovery and Genotype Calling

SNP analyses with next generation sequencing data typically start with three key steps: read alignment, site discovery and genotype calling. In the first step, sequenced reads are mapped to human reference genome[77, 78] and the alignment is refined to calibrate base quality scores and account for known insertion-deletion polymorphisms (indels)[79]. Next, variant sites are identified by examining each overlapping base position in the genome and taking into account a population genetic model (that might describe a prior probability of polymorphism for each site, an allele frequency spectrum and a mutation spectrum, for example)[77]. Finally, genotypes at each site can be refined using linkage disequilibrium information [75, 76]. The complete process is illustrated Figure 4.1. Each step involves many challenges, but here I focus on the last step of genotype calling and haplotype phasing. A companion paper discusses the process of SNP discovery using family information [80].

Describing Chromosomes as Imperfect Mosaics

Hidden Markov models can be used to describe the haplotypes of each individual as imperfect mosaics of other haplotypes in the sample[81]. The approach is commonly used for genotype imputation[10, 12, 14] and can be extended to the analysis of short read sequence data[75]. In this section, I briefly review how these models can be used to model sequence data in unrelated individuals. First, haplotypes for each individual are initialized randomly – sampling a genotype consistent with observed read data at each position. Then, I iterate over individuals, updating the haplotypes of each individual using a Hidden Markov Model that describes the pair of haplotypes as an imperfect mosaic of other haplotypes in the sample.

To describe the model, it is sufficient to specify how haplotypes for one individual can be updated conditional on current haplotype estimates for all other individuals. The first step is to calculate *P(R$_i$|G$_i$)*, the likelihood of observed read data *R$_i$* given an hypothetical true genotype *G$_i$* at each site *i*. These likelihoods can be pre-calculated conveniently with existing tools[82] and can optionally incorporate sophisticated error models, for example, to account for correlated errors[77]. Assuming independent errors, a simple definition for these likelihoods might be:

For homozygous genotype 1/1

$$P(R_i = (\mathbf{B},\mathbf{E}) \mid G_i = \{1,1\}) = \prod_j (1-e_j)^{I(b_j=1)} (\tfrac{1}{3}e_j)^{I(b_j \neq 1)}$$

For heterozygous genotype 1/2

$$P(R_i = (\mathbf{B},\mathbf{E}) \mid G_i = \{1,2\}) = \prod_j \left\{ \tfrac{1}{2}(1-e_j)^{I(b_j=1)} (\tfrac{1}{3}e_j)^{I(b_j \neq 1)} + \tfrac{1}{2}(1-e_j)^{I(b_j=2)} (\tfrac{1}{3}e_j)^{I(b_j \neq 2)} \right\}$$

Here, **B** and **E** are vectors of base calls and associated error probabilities for bases overlapping position *i* in the current sample (*b$_j$* and *e$_j$* are corresponding elements) and I(*expression*) is an indicator function that returns 1 or 0 depending on whether the *expression* is true or false, respectively.

The next step, is to define *P(G$_i$|S$_i$)*, which is the probability of an underlying true genotype *G$_i$* given mosaic state *S$_i$*. To calculate this, I use the function *T(S$_i$)* which returns the number of variant alleles in the template haplotypes indexed by *S$_i$*. Consistent with Yun et al[14], I define:

$$P(G_i \mid S_i\ ) = \begin{cases} (1-\varepsilon_i)^2 & \{T(S_i)=0 \text{ or } T(S_i)=2\} \text{and } T(S_i)=G_i \\ \varepsilon_i(1\text{-}\varepsilon_i) & \{T(S_i)=0 \text{ or } T(S_i)=2\} \text{and } |T(S_i)-G_i|=1 \\ \varepsilon_i{}^2 & \{T(S_i)=0 \text{ or } T(S_i)=2\} \text{and } |T(S_i)-G_i|=2 \\ (1-\varepsilon_i)^2+\varepsilon_i{}^2 & T(S_i)=1 \text{and } T(S_i)=G_i \\ 2\varepsilon_i(1\text{-}\varepsilon_i) & T(S_i)=1 \text{and } T(S_i)\neq G_i \end{cases}$$

ε_i denotes the mosaic error rate at *i*th marker.

Together, *P(R$_i$|G$_i$)* and *P(G$_i$|S$_i$)* allow us to calculate *P(R$_i$|S$_i$)* as:

$$P(R_i \mid S_i) = \sum_{G_i} P(R_i \mid G_i) \times P(G_i \mid S_i) \qquad \text{(Equation 1)}$$

Finally, the last ingredient in the definition of our Hidden Markov Model is to define the transition probabilities *P(S$_{i+1}$|S$_i$)*.

$$\text{P}(S_{\text{i+1}} = (w,v) \mid S_i\ = (x,y)) = \begin{cases} \theta_i{}^2 / N^2 & x \neq \text{w and y} \neq \text{v} \\ (1\text{-}\theta_i)\theta_i / N + \theta_i{}^2 / N^2 & \text{Either } (x \neq \text{w and y = v) or (x = w and y} \neq \text{v)} \\ (1\text{-}\theta_i)^2 + 2(1\text{-}\theta_i)\theta_i / N + \theta_i{}^2 / N^2 & \text{x = w and y = v} \end{cases}$$

Here, *(x,y)* and *(w,v*) denote indexes for the template haplotypes at position *i* and *i+1* and θ_i denotes the mosaic transition rate.

These are all the ingredients needed to calculate *P(S$_i$|***R**), the probability of a specific mosaic state at any position along the chromosome. Calculating this probability for all possible values of *S$_i$* allows us to select a pair of ordered alleles for every position (either by selecting the most likely pair or by selecting a pair at random, for example). *P(G$_i$|***R**) can be obtained by the formula $P(G_i \mid R) = \sum_{S_i \in H(G_i)} P(G_i \mid S_i) \times P(S_i \mid R)$, *H(G$_i$)* is the space compatible with *G$_i$*. Because our model is Markovian, *P(S$_i$|***R**) can be conveniently calculated using Baum's forward/backward algorithm[83]. Thus, I first define recursive left and right probability functions.

The left probability function function L_{i+1} is defined as:

$$\begin{aligned}L_{i+1}(w,v) &= P(R_1,...,R_{i+1},S_{i+1}=(w,v)) = \sum_{x,y} P(R_1,...,R_i,S_i=(x,y)) \times P(S_{i+1}=(w,v) \mid S_i=(x,y)) \times P(R_{i+1} \mid S_{i+1}=(w,v)) \\ &= \sum_{x,y} L_i(x,y) \times P(S_{i+1}=(w,v) \mid S_i=(x,y)) \times P(R_{i+1} \mid S_{i+1}=(w,v)) \\ &= [L_i(w,v) \times (1-\theta_i)^2 + \sum_y L_i(w,y) \times (1-\theta)\theta / N + \sum_x L_i(x,v) \times (1-\theta)\theta / N \\ &\quad + \sum_{x,y} L_i(x,y) \times \theta^2 / N^2] \times P(R_{i+1} \mid S_{i+1}=(w,v)\end{aligned}$$

At the first variant site, the function is defined as

$L_1(w,v)=P(R_1,S_1=(w,v))=P(R_1|S_1=(w,v))\times P(S_1=(w,v))$, where $P(S_1 = (w,v))$ is typically assumed to be a constant.

The right probability $Q_{i+1}(w,v)$ function is defined as:

$$\begin{aligned}Q_{i+1}(w,v) &= P(R_{i+2},...,R_M \mid S_{i+1}=(w,v)) \\ &= \sum_{x,y} P(R_{i+3},...,R_M \mid S_{i+2}=(x,y)) \times P(S_{i+2}=(x,y) \mid S_{i+1}=(w,v)) \times P(R_{i+2} \mid S_{i+2}=(x,y)) \\ &= \sum_{x,y} Q_i(x,y) \times P(S_{i+2}=(x,y) \mid S_{i+1}=(w,v)) \times P(R_{i+1} \mid S_{i+1}=(x,y))\end{aligned}$$

At the last variant site M, the function is defined as $Q_M(w,v)=1$ for convenience.

Finally, I have $P(S_i=(w,v) \mid R) \propto P(S_i=(w,v),R) = L_i(w,v) \times Q_i(w,v)$.

Joint Modeling for Trios

The approach described in the previous section assumes all individuals are unrelated. If related individuals are sequenced, the method ignores important constraints on individual genotypes and haplotypes imposed by Mendel's laws. In this section, I propose a strategy for computationally efficient modeling of linkage disequilibrium and Mendelian inheritance constraints.

I denote R_f, R_m and R_c as the read data for the father, mother and child in a parent-offspring trio and the corresponding genotype likelihoods are $P(R_f|G_f)$, $P(R_m|G_m)$ and $P(R_c|G_c)$. In principle, I could extend the previous algorithm, which is designed to sample pairs of haplotypes in unrelated individuals, to sample four haplotypes at a time in trio parents. The main weakness of this extended model would be that it requires jointly iterating over 4 possible haplotypes, resulting in a substantial increase in computational burden (compute costs would be proportional to H^4 instead of H^2, where H is the number of haplotypes used for each update). Instead, I use an approximate but computationally tractable solution. First, I sample an ordered pair of template haplotypes and an ordered genotype for one of the trio parents conditional on the observed read data for the trio. Next, I sample an ordered pair of template haplotypes and an ordered genotype for the second parent conditional on observed read data *and* the sampled haplotypes for the first parent. For convenience, I assume the first allele in each ordered haplotype is transmitted to the child. In each iteration, the order in which parents are updated alternates randomly.

Let $\vec{R}_i = (R_{f(i)}, R_{m(i)}, R_{c(i)})$ denotes available read information for the father, mother and child at position i. When sampling the first parent, I now replace equation 1 with :

$$P(\vec{R}_i \mid S_i) = \sum_g P(\vec{R}_i \mid G_f = g) \times P(G_f = g \mid S_i)$$

Updates for the second parent, condition on the sampled genotype for the first parent in addition to read data and replace equation 1 with:

$$P(\vec{R}_i \mid S_i, G_m) = \sum_g P(\vec{R}_i \mid G_f = g, G_m) \times P(G_f = g \mid S_i)$$

To calculate the first quantity, I use:

$$\begin{aligned} P(\vec{R}_i \mid G_f = g) &= P(\vec{R}_i, G_f = g) / P(G_f = g) \\ &= \sum_{g_m} P(\vec{R}_i, G_f = g, G_m = g_m, G_c = transmit(g_f, g_m)) / P(G_f = g) \\ &= \sum_{g_m} P(\vec{R}_i \mid G_f = g, G_m = g_m, G_c = transmit(g_f, g_m)) P(G_f = g) P(G_m = g_m) / P(G_f = g) \\ &= \sum_{g_m} P(R_f \mid G_f = g) \times P(R_m \mid G_m = g_m) \times P(R_c \mid G_c = transmit(g_f, g_m)) \times P(G_m = g_m) \end{aligned}$$

Where the transmit function *transmit(G_f,G_m)* returns the child haplotypes implied by the ordered parental genotypes G_f and G_m.

To calculate $P(\vec{R}_i \mid S_i, G_m)$, I note that

$$P(\vec{R}_i \mid G_f = g, G_m = k) = \sum_{g_m} P(R_f \mid G_f = g) \times P(R_m \mid G_m = g_m) \times P(R_c \mid G_c = transmit(g_f, g_m))$$

When dealing with samples that include trios, our algorithm proceeds as follows:

a) Find an initial set of haplotypes that is consistent with available read data (see Appendix A).

b) Sample a new pair of template haplotypes and corresponding genotypes for each unrelated individual.

c) For each nuclear family, randomly pick one parent and sample a new pair of haplotypes for that parent. Then, sample a new pair of haplotypes for the other parent conditioning on both observed read data and the previous pair of sampled haplotypes.

d) Record sampled haplotypes for every individual

e) Update estimated recombination and error rates.

f) Repeat steps b through e).

Generating Consensus Haplotypes

Each round of updates generates a new pair of haplotypes for each sequenced individual. After a pre-defined number of rounds, a pair of consensus haplotypes for each unrelated individual is generated by finding the haplotype pair that minimizes switch error in relation to sampled haplotypes[14, 75]. For parent-offspring trios, where sampled haplotypes are ordered I don't attempt to minimize switch error and simply assign each consensus haplotype the most frequently sampled allele at that position.

4.3 Data Sets

Simulated Data

To evaluate the performance of our method, I start with simulated data sets, which mimic the real shotgun sequencing output. The advantage of simulated data is that I can have comprehensive evaluations by setting up different scenarios and assessing a wide range of possibilities in real studies. In addition, I can compare the results to truth, which is usually unknown for real studies. To be realistic, I simulated 10,000 haplotypes for multiple 1 Mb regions using a coalescent model mimicking realistic LD patterns, modeling population demographic history and local recombination rates similar to European ancestry[84]. Next, I randomly selected haplotypes for founders and generated haplotypes of offspring by family inheritance information. At each site, read depths followed a Poisson distribution and each base was simulated according to a specified per-base error rate. Finally, genotype likelihoods $P(R|G_i)$ were calculated based on the simulated reads R. More details and implemented software are described in the companion paper[80].

I simulated 30 trios, 60 and 90 unrelated samples at depth 1X, 2X, 4X and 8X with per-base error rate 0.01 (Q20) or 0.001 (Q30). Then I doubled the sample size to 60 trios, 120 and 180 unrelated samples. Recall that depth is defined as the average number of read covering each site. I repeated the simulation 100 times.

Real Data

I applied this method to ongoing Sardinia sequencing projects. Up to 2,000 Sardinia samples are being sequenced at an average depth of 3.7X at the University of Michigan. The pilot study consists of complete trios, parent-offspring pairs and unrelated samples (Table 4.3). Most of the samples are also genotyped in Metabo-Chip with high accuracy. It’s a perfect data set to evaluate the performance of our method in that I can compare the called genotypes to those sites also genotyped in the Metabo Chip. I focused on two recently generated data sets with sample size and structure listed in Table 4.3. I compared our genotype calling results to the Metabo Chip at overlapping samples and sites. In addition, I also applied two other methods - a) LD-based method ignoring relatedness and b) single marker caller developed in the companion paper - for comparisons.

Metrics of Performance

To have comprehensive evaluation of the algorithm, I defined a number of metrics to quantify the performance of the genotype calling and phasing results.

Genotype calling:

1. Overall genotype mismatch rate – the percentage of incorrectly called genotypes
2. Mismatch rate at true heterozygous sites – the percentage of incorrectly called heterozygous genotypes

Since the frequency affects above quantities substantially, I stratified the results according to population frequency spectrums.

Haplotype phasing:

1. Mismatched genotypes – the number of mismatched genotypes between inferred and true haplotypes in the simulated region
2. Flips - the number of switch errors between inferred and true haplotypes excluding mismatched genotypes
3. Perfectly predicted haplotypes – the number of inferred haplotypes without any flip excluding mismatched sites. The quantity is expected to be smaller when the region is enlarged. It will also depend on the number of mismatched sites. However, if the numbers of flips and mismatched sites in method A are both less than method B, method A is clearly superior to B in terms of phasing accuracy.

4.4 Results

Overall Performance

I evaluated the performance of the methods on the simulated and real sequencing data sets. I will show the relative performance pattern of different study designs. The absolute numbers are only specific to the current sample size, parameters used in simulation and methods performed. Given the various parameters, I will only discuss some examples in the table for general patterns.

As shown in Table 4.1, comparing 30 or 60 trios with 60 or 120 unrelated samples, sequencing an additional child can always detect more variants, as expected. For the same sequencing cost, comparing 30 trios or 60 trios with 90 or 180 unrelated samples, I could detect more variants for trios at low depth (1X and 2X) but unrelated samples can outperform trios at 8X in terms of genotyping accuracy. Both are comparable at 4X. The sequencing base error rate also affects the SNP discovery. Generally, I can call 10% more SNPs when base error rate is reduced from 0.01 to 0.001.

Henceforth, I will focus on evaluating the mismatch rates. For each sample category, the mismatch rate decreases as the depth increases and heterozygote sites grow more difficult to infer. Increasing the sample size could result much improvement. For example, at base error rate 0.01 and 2X coverage, the mismatch rate of 120 unrelated samples is 2.7% compared to 4.4% of 60 unrelated samples. The elevated sequencing error caused by multiple reasons also challenges the genotype calling. The observation in Table 4.1 indicates that the mismatch rate increases as the base error rate increases, but the impact is not big relative to the sample size and the depth. For instance, with the base error rate 1% and 1X, the mismatch rate of 30 trios is only 4.5% compared to 3.8% with the error rate 0.1%. Considering that the base error rate in current sequencing technologies is expected to lie between 0.001 and 0.01, increasing sample size and sequencing depth will more efficiently increase the genotype calling accuracy. Next, I compare the performance of trios to unrelated samples. For each set of comparisons, I compared trios to two sample sizes of unrelated samples, which correspond to the same number of independent samples and same sequencing cost as the trios. Generally, sequencing trios has lower mismatch rates at all sites and heterozygote sites for all categories. For instance, the mismatch rate of 30 trios at depth of 2X with error rate 0.001 is 1.1% compared to 2.4% and 3.2% for 60 and 90 unrelated samples. The gain at high depth of coverage (8X) is still remarkable, even though the genotypes can be inferred confidently from its own genotype likelihood based on the single marker.

Although I called genotypes for both parents and child jointly, mismatch rates are still slightly differences between them especially at high depth (Table 4.7). For example, the mismatch rate of child is 0.3%, comparing to 0.45% at the depth of 4X and error rate 0.01. It is expected in the reason that the two chromosomes of the child are actually double sequenced.

Performance by Stratified Frequency

The number of discovered SNP is always limited by the sample size. Given the sample size in the simulation, the evaluation of the method might be biased towards common variants since rare variants are difficult to discover before the genotype calling steps. Therefore, to have more comprehensive evaluations, I carried out the analysis on different allele frequency categories. More specifically, I categorized all SNPs into 10 even frequency bins. The frequency refers to the base allele in the reference genome. Examining only heterozygous sites can provide a more accurate evaluation of the method especially at rare variants. I will only focus the heterozygous sites. An evident pattern is shown in Figure 4.3 with 30 trios, 60 and 90 unrelated samples at 2X. The figures for other scenarios are very similar. Table 4.7 gives detailed information. The top panel is the overall mismatch rate at different depths. Sequencing more samples can yield a lower mismatch rate. 30 trios outperform 60 and 120 unrelated samples in all frequency categories at all depths. However, the absolute gain is marginal at 8X since all mismatch rates are already very slow. For low frequency categories, the major allele homozygotes dominate the performance, but the heterozygous sites are more difficult to infer. As I can see, the relative order of performance is unchanged. Trios are even more beneficial at heterozygous sites for low frequency categories. This behavior is expected because although a rare variant is not easy to sample from reference panel, it is actually double sequenced if it passes to offspring, which increases the likelihood of correctly inferring heterozygotes.

Accuracy of Haplotype Inference

Another important result of our method is the haplotype reconstruction, which is essential for follow-up imputation and population history inference. I evaluate the accuracy of our method by comparing three quantities jointly as defined in our method part: errors, flips and perfectly predicted haplotypes. I show the simulation results in Table 4.2. The region is 1 Mb long and the number of compared sites depends on each sample size. As expected from GWAS, a larger sample size can always increase phasing accuracy. For instance, at 4X, 90 unrelated samples yield 40 flips, while 60 unrelated samples have 60 flips. Trios have great advantage in haplotype inference. 30 trios only have 2 flips at 4X.

It is important to recall that mismatched sites were excluded from the haplotypes to calculate the flips. A better phasing method needs to have both lower mismatch rate and number of flips. The number of perfectly predicted haplotypes is not comparable for different depths, but the relative pattern reveals that trios have a great advantage in haplotype inference.

Performance on Sardinia Sequencing Data

The performance of our method in simulation data sets is encouraging. However, it has always depended on simulation models. Clear performance pattern could be demonstrated with a realistic model. It is more interesting to see what happens in real data sets. As described in the method part, three approaches were used to call genotypes. I summarized the comparison results at the available genotypes on chromosome 20 in Metabo Chip in Table 4.4. I present genotype mismatch rate on both overall and heterozygote sites and further stratify the results into different categories by minor allele frequency. I will focus on heterozygous sites. The general message is that heterozygous sites are more difficult to infer for all methods. For LD-based algorithm, the large sample size yields better genotype calling accuracy; the single marker caller variant sites. For instance, a single marker approach has high overall mismatch rate of 11.4% compared to 2.4% for the LD-based approach ignoring relatedness and 1.5% for our approach. Our approach outperforms the LD-based algorithm ignoring relatedness in all categories, reducing the mismatch rate from 4.2% to 2.4 % for the first data set and 2.4% to 1.6% for the second data set. Table 4.5 presents the stratified results. LD-aware method performs much better at common sites. At rare variants sites, more samples are needed. For 186-

sample data set, our method has best performance in all categories. Although our methods are promising, the limited counts and accuracy of Metabo chip at rare sites require more investigation when more data are generated.

Summary

Our simulations show that sequencing trios can have similar or even higher variant calling rates at both all and heterozygote sites compared to the same number of unrelated individuals at low depth 1X and 2X. For depths of 4X and 8X, although sequencing unrelated samples has more power to detect variants, sequencing trios have the advantage of higher sensitivities at heterozygous sites, which is crucial for individual genotype call of rare variants. As the sample size increases, I expect sequencing trios will have increased advantages. Generally, sequencing trios has the higher calling accuracy across different frequency spectrums. In addition, trios can greatly increase the haplotyping accuracy. Applying our method to Sardinia sequencing project leads to better genotype calling accuracy than an approach that ignores the relatedness.

Computational Complexity

Since our state space is ungrouped and in the square of the number of reference haplotypes, the complexity of the LD-based approach for SNP calling is $O(N^3)$, which increases rapidly as N increases. Sampling from the joint space of two parents will make the computation infeasible in practice. Calculating the marginal emission probabilities conditional on alternatively sampled parent keeps the same scale of computation as unrelated cases but incorporates the trio information simultaneously. Our approach also orders the genotypes so that the genotypes of the offspring are determined by the genotypes of the parents. This could reduce the size of reference panel and save computing time if there is big proportion of trios in the study.

4.5 Discussion

The new method this paper presents can accurately call genotypes and infer haplotypes for shotgun sequencing data. It can handle unrelated samples and parent-offspring pairs in a manner similar to common LD-based approaches for genotype calling and imputation of unrelated individuals. The family information is simultaneously taken into account in the hidden Markov model. This method can be used in many pipelines for shotgun sequencing data. In the sequencing project, once a set of polymorphic sites are called (e.g. family-based approach presented in the companion paper), our method can refine individual genotypes. The output of our program has both inferred genotypes and posterior probabilities, which could be used selectively in follow-up association tests.

SNP genotype imputation method has been widely used in GWAS and benefits many large scale meta-analyses [4, 38]. Some comprehensive evaluations of different imputation methods have been published [8, 85]. A common conclusion is that a LD-based algorithm performs well in common variants but has limitations in uncommon or rare variants. The main reason is that rare variants occur less frequently in study samples; LD information may often “correct” true rare heterozygous sites into homozygotes. A similar issue has appeared, although not comprehensively, in our simulation studies for shotgun sequencing in unrelated samples. Inheritance information from offspring can increase the likelihood of correctly inferring rare heterozygotes and avoid miscorrection by LD information from the population. I expect adding more family members will continuously improve genotype calling accuracy for rare variants. Furthermore, if the trio based design finds interesting rare variants in a region to be associated with some disease through the trio based design, I can sequence more family members, if available, and use the family-based caller[80] to confirm those rare genotypes. This two-step approach will eliminate the false positives more efficiently and increase power to detect true causal variants.

One big advantage of LD-aware caller is haplotype inference or phasing, which is not available from single marker based caller. Haplotype inference is crucial in follow-up studies such as haplotype association analysis or estimation of r^2 to assess the LD blocks. If all samples are unrelated, then the sampled haplotypes are basically copies of different mosaics of reference haplotypes in the population. Trio data sets impose the family constraints and guide the sampling steps in Markov model. More beneficially, samples in many ongoing sequencing projects are often a small portion of existing GWAS. Those sequenced samples serve as the reference panel to impute the polymorphic sites detected in the unsequenced samples in GWAS. Accurate phase information will benefit the imputation step and greatly reduce computational cost. .

I proposed two approaches for initial guess of haplotypes. I showed that our method achieved better accuracy than the single marker caller and the method that ignores the relatedness with a random initial guess of the haplotypes. In practice, I can even do better using the haplotypes inferred from other software. For example, I used the haplotypes inferred from BEAGLE as starting point of our method on a latest released data set with 508 samples. From Table 4.6, even after 30 rounds, our method can reach very low mismatch rate (0.49% at all sites, 1.04% at heterozygous sites).

The main focus of this paper is to evaluate the performance of our method in terms of accuracy of genotype calling and haplotype inference. Although sequencing trios can increase the accuracy of both genotype calling and haplotype inference, it does not imply that the design of sequencing trios is more preferable to unrelated samples with the fixed total sequencing cost. Sequencing more unrelated samples may have more power in SNP discovery[80] and follow-up association tests. This question is beyond the scope of this paper and deserves further investigation.

Although I have investigated various scenarios, many other interesting experimental setups do exist. For example, our simulation fixed the error rate for all bases and depths across all samples for each case. It might also be interesting to see what would happen if the base error rate varies following some realistic distribution or if offspring has more and less depths than parents. Some optimal design may lead to larger power given the total sequencing cost or other constraints.

The current approach could be potentially extended to the nuclear family: two parents with multiple offspring and general pedigree. A simple starting point might be splitting a nuclear family into multiple trios with duplicated parents, updating parents alternatively and summarizing the results across all split trios. A more complicated approach based on pedigree likelihood calculation is also possible but I need to recall that the computing complexity should be feasible given current computing power. Some approximation or genotype space reduction might be useful to speed up the calculation. Parallel computing and multiple threading techniques provide a potential solution in practice.

Web Resources

The URLs for data presented herein are as follows:

Triocaller: The C++ program based on the method described in this paper, www.sph.umich.edu/csg/weich/software/TrioCaller

Simulator: www.sph.umich.edu/csg/binghsan

Variant Calling: www.sph.umich.edu/csg/bingshan

Table 4.1 Error rates for genotype calling

						Mismatch Rate							
Base		Called Variants				All Sites				Heterozygous Sites			
Error	Sample	1X	2X	4X	8X	1X	2X	4X	8X	1X	2X	4X	8X
0.01	60 unrelated	2112	2548	3079	3757	.1040	.0438	.0120	.0021	.1563	.0563	.0147	.0033
	90 unrelated	2323	2778	3350	4178	.0809	.0324	.0092	.0016	.1262	.0424	.0112	.0024
	30 trios	2351	2827	3435	3993	.0380	.0151	.0040	.0008	.0523	.0175	.0048	.0011
0.001	60 unrelated	2448	2853	3447	4084	.0878	.0319	.0084	.0015	.1774	.0538	.0126	.0030
	90 unrelated	2616	3128	3796	4576	.0667	.0238	.0065	.0011	.1363	.0405	.0098	.0022
	30 trios	2641	3172	3773	4223	.0320	.0106	.0031	.0006	.0607	.0169	.0046	.0011
0.01	120 unrelated	2472	2923	3565	4529	.0687	.0265	.0076	.0013	.1087	.0344	.0093	.0021
	180 unrelated	2686	3156	3898	5041	.0537	.0203	.0060	.0011	.0863	.0266	.0075	.0016
	60 trios	2722	3253	4049	4866	.0264	.0104	.0027	.0005	.0371	.0120	.0033	.0008
0.001	120 unrelated	2780	3323	4063	4962	.0559	.0193	.0054	.0009	.1167	.0332	.0082	.0018
	180 unrelated	3034	3610	4516	5626	.0426	.0146	.0044	.0007	.0917	.0255	.0068	.0014
	60 trios	3081	3708	4530	5155	.0205	.0070	.0020	.0004	.0404	.0114	.0032	.0007

Error rates for genotype calling in samples of parent-offspring trios or unrelated individuals, as function of sequencing depth (1X, 2X, 4X or 8X) and per base error rate of the original sequence traces (0.01 or 0.001)

Table 4.2 Quality of estimated haplotypes in simulated 1Mb regions

Depth	1X			2X			4X			8X		
	Error[a]	Flips[b]	Perfect[c]	Error	Flips	Perfect	Error	Flips	Perfect	Error	Flips	Perfect
60 unrelated	220.2	46.9	0.2	111.7	58.5	0.3	37.1	59.9	0.4	7.8	60.8	0.2
90 unrelated	188.7	33.4	0.3	90.0	39.5	1.2	31.0	39.5	2.4	6.6	42.0	0.6
30 trios	89.5	6.0	6.9	42.8	2.8	26.6	13.8	1.5	47.0	3.2	0.7	68.3
120 unrelated	170.2	26.1	0.6	77.5	28.6	3.1	27.1	30.4	5.4	6.0	33.6	1.8
180 unrelated	144.4	17.5	2.0	64.1	18.6	12.5	23.4	20.5	14.9	5.4	23.7	6.2
60 trios	71.9	3.4	36.8	33.6	1.5	88.2	10.8	0.9	118.5	2.6	0.4	150.0

[a] Error: the number of mismatched genotypes per person between inferred and true haplotypes in the simulated region

[b] Flips: number of switch errors per person between inferred and true haplotypes excluding mismatched genotypes

[c] Perfect: the number of predicted haplotypes with no flips excluding mismatched sites

Table 4.3 Family structures of the SardiNIA data sets

	Data Set 1	Data Set 2
Unrelated samples	7	66
Complete Trio	13	25
1 Parent with 1 offspring	4	0
1 Parent with 2 offspring	4	15
Total	66	186
Samples genotyped	55	105

Table 4.4 Overall genotype concordance between metabochip and low-pass sequence data from Sardinia project

	66 Samples				186 Samples			
	Count	Single[a](%)	Thunder[b](%)	TrioCaller(%)	Count	Single(%)	Thunder(%)	TrioCaller(%)
Overall	107165	12.70	4.23	2.32	222049	12.18	2.37	1.51
Heterozygote	31339	28.79	8.69	5.19	60878	28.72	5.53	3.66
Alternative Homozygote	19412	12.09	3.18	1.59	37307	13.07	1.94	1.23
Reference Homozygote	56414	3.95	2.12	0.98	123864	3.9	0.96	0.55

[a] Single is a family-based genotype calling algorithm on single marker.

[b] Thunder is a LD-aware genotype calling algorithm ignoring the relatedness.

Table 4.5 Stratified genotype concordance between metabochip and low-pass sequence data from Sardinia project

MAF[a]	Nsample[b]	NSNP	Overall (%)			Heterozygots (%)		
			Single	Thunder	Triocaller	Single	Thunder	Triocaller
66 Samples								
All freq	55	1950	12.70	4.23	2.40	28.79	8.69	5.19
0 - 2%	55	75	1.92	2.64	2.32	30.82	16.78	16.08
2% - 5%	55	180	2.42	2.37	0.91	25.19	11.95	6.30
> 5%	55	1695	14.26	4.50	2.48	28.87	8.57	5.11
186 Samples								
All freq	105	2116	12.18	2.41	1.55	28.72	5.46	3.66
0 - 2%	105	120	1.34	1.42	1.09	34.43	14.47	13.84
2% - 5%	105	273	2.76	1.34	0.72	34.98	9.53	5.47
> 5%	105	1723	14.52	2.65	1.71	28.49	5.28	3.51

[a] MAF denotes the minor allele frequency, stratified in three categories.

[b] Nsample is the number of samples with genotypes available in Metabo chip.

Table 4.6 Improvement of genotype accuracy with phased input from Beagle

MAF	NSNP	Overall Mismatch Rate (%)			Heterzygotes Mismatch Rate (%)		
		Count	Beagle only	Beagle+Triocaller	Count	Beagle only	Beagle+Triocaller
all	2491	393346	0.68	0.49	102297	1.67	1.04
0 - 2%	233	36806	0.22	0.16	696	7.76	4.83
2% - 5%	328	51797	0.29	0.33	3233	2.88	1.69
> 5%	1930	304743	0.80	0.55	98368	1.59	0.99

Table 4.7 Error rates stratified by frequency at heterozygotes for genotype calling

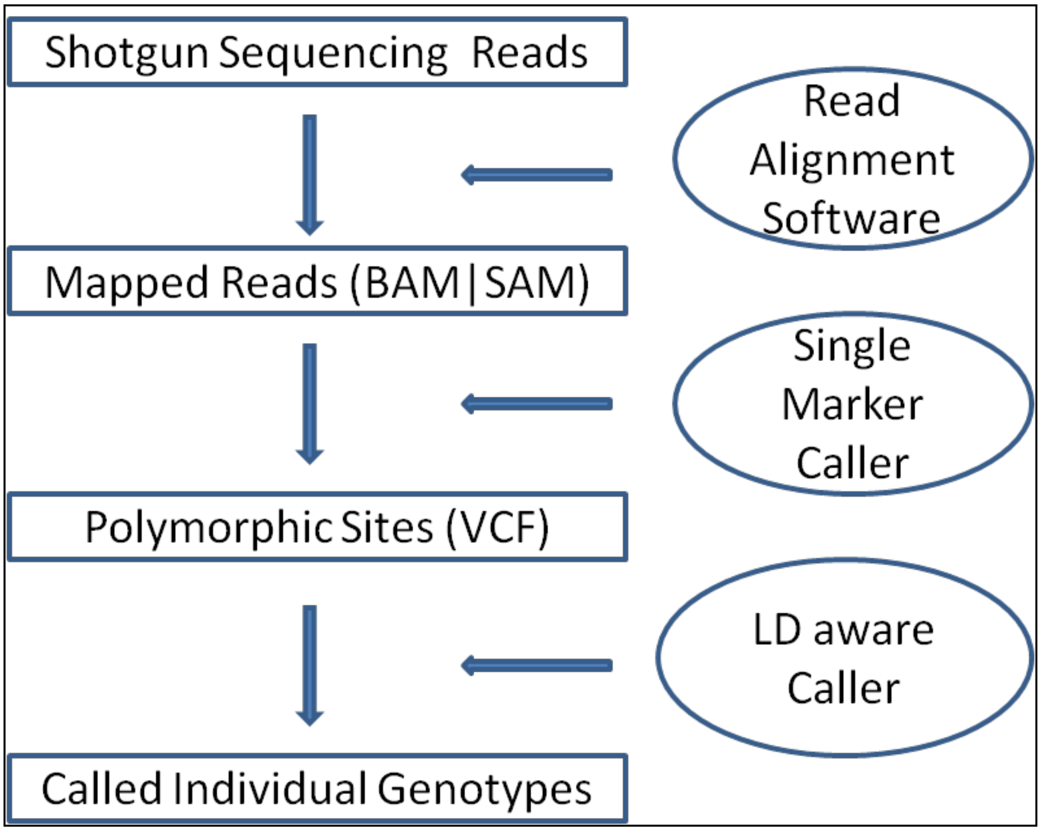
Reference Allele	0-.1	.1-.2	.2-.3	.3-.4	.4-.5	.5-.6	.6-.7	.7-.8	.8-.9	.9-1
1X										
60 unrelated	.3076	.2005	.1444	.1226	.1137	.1187	.1259	.1551	.2164	.2630
90 unrelated	.2814	.1597	.1036	.0913	.0874	.0903	.0915	.1115	.1754	.2569
30 trios	.1492	.0484	.0368	.0373	.0378	.0394	.0411	.0461	.0677	.1160
2X										
60 unrelated	.1299	.0657	.0479	.0423	.0394	.0407	.0406	.0484	.0704	.1122
90 unrelated	.0988	.0452	.0339	.0302	.0290	.0287	.0285	.0327	.0480	.0979
30 trios	.0556	.0160	.0130	.0130	.0130	.0130	.0128	.0144	.0187	.0368
4X										
60 unrelated	.0435	.0131	.0108	.0105	.0099	.0098	.0100	.0114	.0155	.0329
90 unrelated	.0298	.0103	.0078	.0073	.0076	.0075	.0072	.0081	.0106	.0272
30 trios	.0175	.0036	.0035	.0035	.0035	.0035	.0036	.0038	.0044	.0102
8X										
60 unrelated	.0090	.0028	.0024	.0023	.0022	.0021	.0023	.0024	.0031	.0072
90 unrelated	.0072	.0018	.0014	.0016	.0016	.0016	.0016	.0017	.0020	.0058
30 trios	.0035	.0008	.0007	.0008	.0008	.0008	.0008	.0008	.0009	.0026

In samples of trios and unrelated individuals at the sequencing depth of 1X, 2X, 4X, 8X and base error rate of 0.01

Table 4.8 Comparisons of mismatch rate between the child and the parents for the simulation of 30 trios

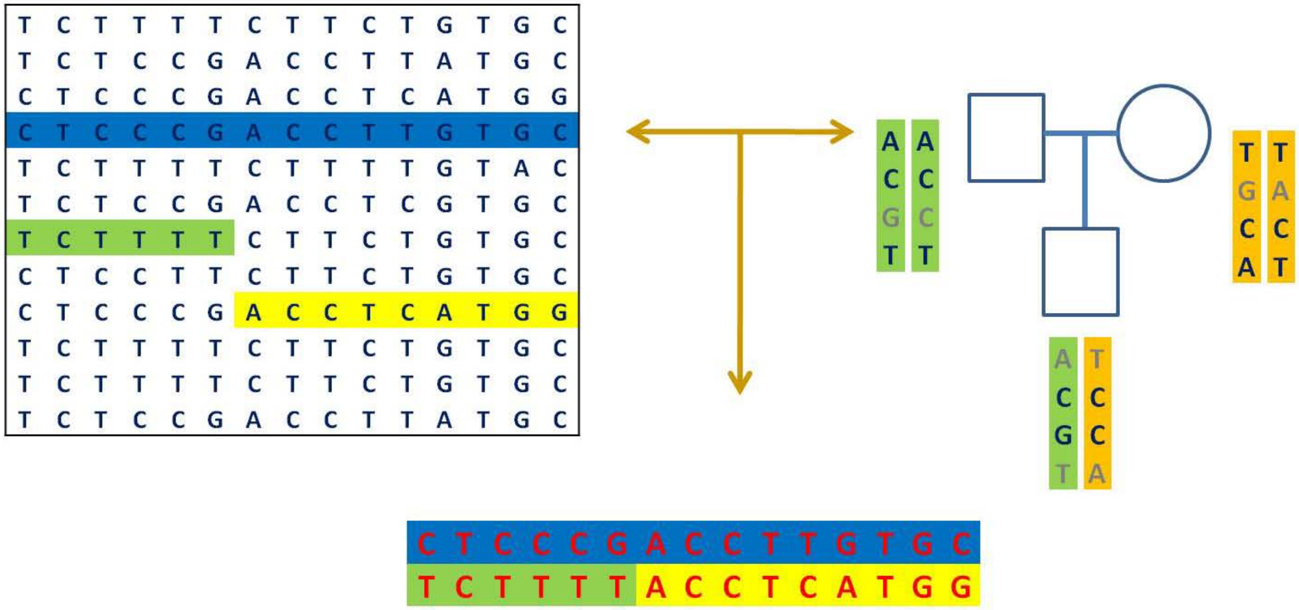
		All Sites				Heterozygous Sites			
Error	Sample	1X	2X	4X	8X	1X	2X	4X	8X
0.01	Parents	.0395	.0160	.0045	.0010	.0523	.0175	.0050	.0013
	Child	.0348	.0134	.0030	.0003	.0524	.0173	.0045	.0008
0.001	Parents	.0333	.0113	.0035	.0008	.0617	.0173	.0047	.0012
	Child	.0293	.0093	.0022	.0002	.0587	.0161	.0043	.0008

Figure 4.1 Workflow of SNP discovery and genotype calling



This figure describes a typical pipeline currently used in next generation sequencing studies. This paper focuses on the last step of refining genotypes and haplotype inference.

Figure 4.2 Cartoon view of LD-aware method for unrelated samples and parent-offspring trios



T	C	T	T	T	T	C	T	T	C	T	G	T	G	C
T	C	T	C	C	G	A	C	C	T	T	A	T	G	C
C	T	C	C	C	G	A	C	C	T	C	A	T	G	G
C	T	C	C	C	G	A	C	C	T	T	G	T	G	C
T	C	T	T	T	T	C	T	T	T	T	G	T	A	C
T	C	T	C	C	G	A	C	C	T	C	G	T	G	C
T	C	T	T	T	T	C	T	T	C	T	G	T	G	C
C	T	C	C	T	T	C	T	T	C	T	G	T	G	C
C	T	C	C	C	G	A	C	C	T	C	A	T	G	G
T	C	T	T	T	T	C	T	T	C	T	G	T	G	C
T	C	T	T	T	T	C	T	T	C	T	G	T	G	C
T	C	T	C	C	G	A	C	C	T	T	A	T	G	C

C	T	C	C	C	G	A	C	C	T	T	G	T	G	C
T	C	T	T	T	T	A	C	C	T	C	A	T	G	G

This cartoon sketches our method. The top left panel is the unrelated reference haplotypes. The top right figure is the current updating trio. The bottom figure is one of the parents in the trio awaiting for updating. In the trio, current configuration of the two haplotypes is shown next the each individual and the grey letters indicate uncertainty of the genotypes inferred from individual sequence data.

Figure 4.3 Frequency stratified mismatch rate at heterozygote sites at different depths for 30 trios, 60 unrelated and 90 unrelated samples at base error rate 0.01

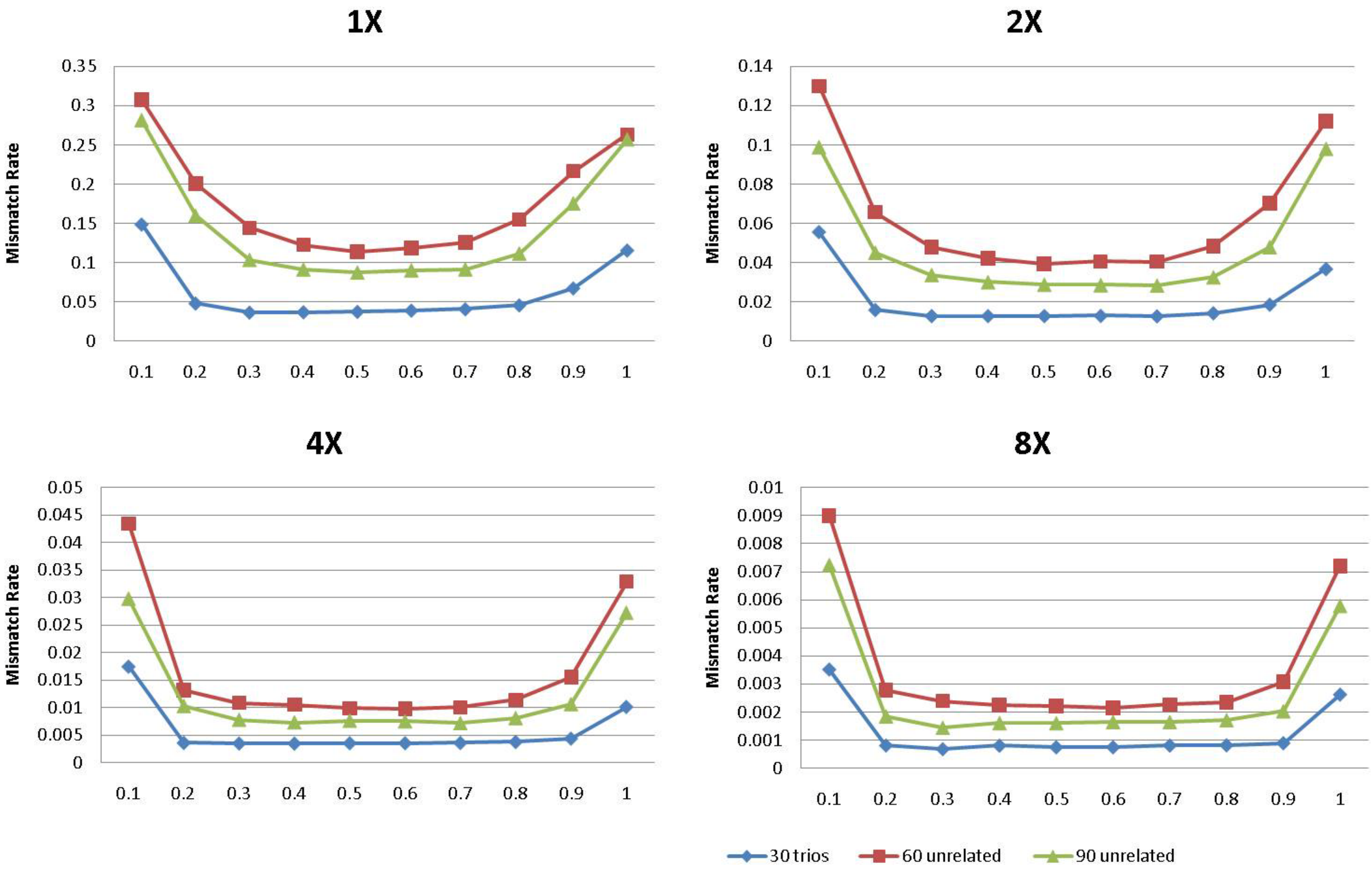
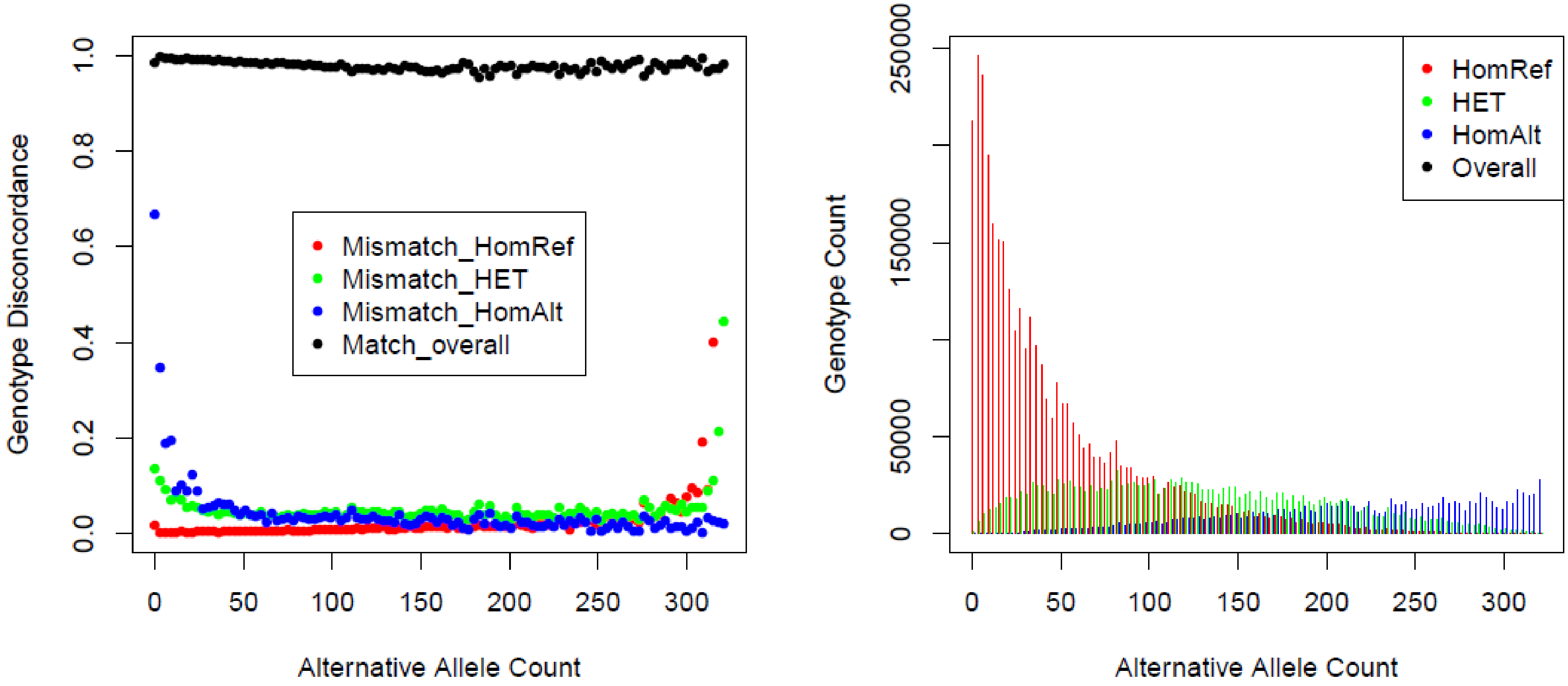


Figure 4.4 Genotype distributions and disconcordance for heterozygotes, reference homozygotes and alternative homozygotes



APPENDIX

Sampling an Initial Haplotype Set

To start the hidden Markov chain, I need to make an initial guess of genotypes and haplotypes. There are several ways to obtain the initial genotypes. I proposed two as follows.

1. Single marker genotype call and random haplotype inference

For each unrelated sample, the individual genotype is usually inferred by calculating the posterior probabilities *P(G|R) = P(R|G)×P(G)/P(R)* based on the estimated population frequency *P(G)*. The genotype is unordered and no phase information is available from this initial guess. For parent-offspring trios, the accuracy of the initial guess will be greatly improved by calculating posterior probabilities conditional on the whole trio. For example,

$$P(G_f \mid R_f, R_m, R_C) = \sum_{G_m, G_c} P(R_f \mid G_f) \times P(R_m \mid G_m) \times P(R_c \mid G_c) \times P(G_f) \times P(G_m) \times P(G_c \mid G_f, G_m)$$

More importantly, if I order the genotypes as mentioned above - that the allele of each parent always passes to the offspring - the initial inferred haplotypes are much closer to the true haplotypes than randomly filled haplotypes. The reason is similar to phasing trios with known genotypes, where uncertainty only occurs at sites that are heterozygous in all trio members. The initial inference of the haplotypes at highly covered sites facilitates the follow-up phasing procedure and improves the convergence of the algorithm. The benefit becomes even larger as sequencing depth of coverage increases.

2. External genotypes and haplotypes from other software

The alternative way to have an initial haplotype configuration is through other software (e.g. BEAGLE). It is usually better than random guess and can speed up our method by reducing iterations.

Chapter 5
State Space Reduction Model for Haplotyping and Genotype Calling

5.1 Introduction

Genome-wide association studies (GWAS) have recently been a powerful method to discover the genetic basis of human disease in the area of human genetics. GWAS try to identify causal single nucleotide polymorphisms (SNP) which contribute to complex disease. Most GWAS use commercial DNA chips to genotype typically hundreds of thousands of SNPs to serve as proxies of causal SNPs in the whole genome. A key problem of such studies is the imputation of missing genotypes and haplotype inference, which help detect additional signals and combine multiple data sets across multiple platforms. Hidden Markov model (HMM) is a commonly used tool to describe the special features of the sampled haplotypes [8, 11-13, 81]. This approach has been widely used in many GWAS and meta-analysis across multiple platforms [4, 86]. As the rapid development of next generation sequencing technologies, this model has been extended to deal with shotgun sequencing data. As the scale of the data increases, the computational challenge requires more attention. Table 5.1 describes the typical running time and memory requirement in haplotype inference using widely used software MaCH. When the number of samples reaches thousand, using all template haplotyes as state space becomes infeasible in practice. Motivated by the practical needs, I proposed a fast but mathematically and numerically equivalent algorithm to partially overcome the computational burden. I organize our paper as follows. First, I describe the underlying model used in haplotype inference and highlight the computational difficulty. Second, I propose a state space reduction algorithm to reduce the computational burden, but retain the same accuracy. Finally, I evaluate our method through both simulated and real data sets and compare the performance with standard methods.

5.2 Method

In this section, I will illustrate how the space state reduction method works efficiently in the framework of standard HMM approaches used in haplotype inference. This framework is consistent with Li et al. [14].

Describing Chromosomes as Imperfect Mosaics

Hidden Markov models can be used to describe the haplotypes of each individual as imperfect mosaics of other haplotypes in the sample[81]. The approach is commonly used for genotype imputation [10, 12, 14] and can be extended to the analysis of short read sequence data [75]. In this section, I briefly review how these models can be used to model sequence data in unrelated individuals. First, haplotypes for each individual are initialized randomly – sampling a genotype consistent with observed read data at each position. Then, I iterate over individuals, updating the haplotypes of each individual using a hidden Markov model that describes the pair of haplotypes as an imperfect mosaic of other haplotypes in the sample.

To describe the model, it is sufficient to specify how haplotypes for one individual can be updated conditional on current haplotype estimates for all other individuals. In the first step, is to define $P(G_i|S_i)$, which is the probability of an underlying true genotype G_i given mosaic state S_i. To calculate this, I use the function $T(S_i)$ which returns the number of variant alleles in the template haplotypes indexed by S_i. Consistent with Yun et al [14], I define:

$$P(G_i \mid S_i) = \begin{cases} (1-\varepsilon_i)^2 & \{T(S_i)=0 \text{ or } T(S_i)=2\} \text{ and } T(S_i)=G_i \\ \varepsilon_i(1\text{-}\varepsilon_i) & \{T(S_i)=0 \text{ or } T(S_i)=2\} \text{ and } \left|T(S_i)-G_i\right|=1 \\ \varepsilon_i^{\,2} & \{T(S_i)=0 \text{ or } T(S_i)=2\} \text{ and } \left|T(S_i)-G_i\right|=2 \\ (1-\varepsilon_i)^2+\varepsilon_i^{\,2} & T(S_i)=1 \text{ and } T(S_i)=G_i \\ 2\varepsilon_i(1\text{-}\varepsilon_i) & T(S_i)=1 \text{ and } T(S_i)\neq G_i \end{cases}$$

Then, the second ingredient in the definition of our hidden Markov model is to define the transition probabilities $P(S_{i+1}|S_i)$.

$$\mathrm{P}(S_{i+1}=(w,v) \mid S_i=(x,y)) = \begin{cases} \theta_i^{\,2}/N^2 & x \neq \mathrm{w} \text{ and } \mathrm{y} \neq \mathrm{v} \\ (1\text{-}\theta_i)\theta_i/N+\theta_i^{\,2}/N^2 & \text{Either } x \neq \mathrm{w} \text{ or } \mathrm{y} \neq \mathrm{v} \\ (1\text{-}\theta_i)^2+2(1\text{-}\theta_i)\theta_i/N+\theta_i^{\,2}/N^2 & \mathrm{x}=\mathrm{w} \text{ and } \mathrm{y}=\mathrm{v} \end{cases}$$

Here, (x,y) and (w,v) denote indexes for the template haplotypes at position i and $i+1$. With all elements discussed above, I can calculate posterior probability $P(S_i|\boldsymbol{G})$, which can be used to infer genotypes for *ith* marker through function $T(S_i)$. Recall S_i is the underlying pair of haplotypes, $\boldsymbol{G}$ is the all read information for current updating sample. This nontrivial calculation is proportional to $P(\boldsymbol{G},S_i)$, which can be simplified by Baum's forward-backward algorithm through a recursive formula[83] as follows.

Left probability for *i+1th* marker is denoted as L_{i+1}

$$
\begin{aligned}
L_{i+1}(w,v) &= P(G_1,...,G_{i+1},S_{i+1}=(w,v)) \\
&= \sum_{x,y} P(G_1,...,G_i,S_i=(x,y)) \times P(S_{i+1}=(w,v) \mid S_i=(x,y)) \times P(R_{i+1} \mid S_{i+1}=(w,v)) \\
&= \sum_{x,y} L_i(x,y) \times P(S_{i+1}=(w,v) \mid S_i=(x,y)) \times P(R_{i+1} \mid S_{i+1}=(w,v)) \\
&= [L_i(w,v) \times (1-\theta_i)^2 + \sum_y L_i(w,y) \times (1-\theta)\theta / N + \sum_x L_i(x,v) \times (1-\theta)\theta / N \\
&\quad + \sum_{x,y} L_i(x,y) \times \theta^2 / N^2] \times P(G_{i+1} \mid S_{i+1}=(w,v)
\end{aligned}
$$

I start with equal probabilities for first marker and loop through all markers for possible pairs *(w,v)*. At the *Mth* marker, I obtain the joint probability $L_M(w,v) = P(\mathbf{G},S_M=(w,v))$.

The right probability $Q_{i+1}(w,v)$ function is defined as:

$$
\begin{aligned}
Q_{i+1}(w,v) &= P(G_{i+2},...,G_M \mid S_{i+1}=(w,v)) \\
&= \sum_{x,y} P(G_{i+3},...,G_M \mid S_{i+2}=(x,y)) \times P(S_{i+2}=(x,y) \mid S_{i+1}=(w,v)) \times P(G_{i+2} \mid S_{i+2}=(x,y)) \\
&= \sum_{x,y} Q_i(x,y) \times P(S_{i+2}=(x,y) \mid S_{i+1}=(w,v)) \times P(G_{i+1} \mid S_{i+1}=(x,y))
\end{aligned}
$$

At the last variant site *M*, the function is defined as $Q_M(w,v)=1$ for convenience.

Finally, I have $P(S_i=(w,v) \mid G) \propto P(S_i=(w,v),G) = L_i(w,v) \times Q_i(w,v)$.

Again, *(x,y)* and *(w,v)* denote indexes for the template haplotypes.

I can obtain the posterior probabilities by a simple product of left and right probabilities at the same position. The genotype calls can be inferred from the probabilities conveniently.

$$
P(S_i=(w,v) \mid G) \propto P(S_i=(w,v),G) = L_i(w,v) \times Q_i(w,v)
$$

Another asymptotically equivalent approach is through a sampling procedure. For each updating sample, I sample states from the *Mth* marker from $L_M = P(G_1,G_2,\ldots,G_M, S_M)$. Then I continue to sample haplotypes reversely from $L_{i+1} = P(G_1,G_2,\ldots G_i, S_i|S_{i+1})$, where S_{i+1} is the sampled states at *i+1* th marker. I summarize all sampled states after pre-defined rounds for each sample.

The initial inference of the haplotypes can be randomly assigned or obtained from external software (e.g. BEAGLE[11]), which has been described in the last chapter appendix.

For each round, I can update the haplotypes for each sample and a pair of sampled haplotypes is stored for each sequenced individual. After a pre-defined number of rounds (e.g. M) are finished, a set of consensus haplotypes for each individual is generated from M sampled haplotype pairs. More specifically, for each individual, I need to specify a reference haplotype for each unordered haplotype pair by checking the first heterozygous site. Then, I can find the most frequent haplotype configuration for each subsequent SNP until a heterozygous state occurs. I flip the reference haplotype of the pair where the configuration is the other heterozygous state. I repeat this procedure until the last marker.

State Space Reduction Method

The key step of the calculation described above is the recursive formula in forward calculation. This calculation requires looping from first to last maker (M markers). At each marker, I need to calculate the probabilities *L(x,y)* for all possible states and store them for future use, which requires N^2 memory space in the unit of float or double allocation. Therefore, the computational time and the memory cost are approximately proportional to MN^2. The computational cost increases quadratically as the number of samples increases for each updating step. In the recent studies, a typical used reference panel is about 120 haplotypes for CEU samples from HapMap 2. However, the number will increase quickly to thousands in next two years when the 1000 Genomes Projects (www.1000genomes.org) and other large sequence projects are completed. The quadratic increase of computational cost will make current algorithm impractical. Motivated by this practical challenge, I propose a state space reduction method to speed up the calculation and preserve the high accuracy of the HMM. The main idea is to take advantage of similarity among segments of chromosomes in short region and reduces the number of states in the forward calculation. The idea is inspired by the fact that the number of unique haplotypes increases slowly in a short region of the genome as the number of haplotypes increases, given the nature of inheritance according to coalescent theory. Typically, for current density of the reference panel, H different reference haplotypes can share a few identical short fragments in a narrow window, which is due to the tight linkage within a short region (Figure 5.2). I can illustrate this observation through a pool of 10,000 simulated chromosomes. The simulation details will be given in the simulation section. In Figure 5.2(a) and (b), a 10kb window is randomly chosen and I counted the average unique haplotypes and the number of the SNPs as the number of individual increases. In Figure 5.2(c), I plotted the number of unique haplotypes against the number of SNPs in the haplotypes in all 10,000 samples.

Based on this process, I can segment the chromosomes into a series of short windows and reduce the number of reference haplotypes in each window, resulting in potential calculation savings. Figure 5.3 illustrates the idea. Assume that there are N template haplotypes with M markers. I divide the chromosomes into T windows and denote t_i as the number of markers and h_i as the number of unique haplotypes in the *ith* window. The strategy to allocate the windows will be discussed in next section. I summarize the algorithm in forward calculation within one window below:

1. At the start SNP of the window, calculate and store the full probability L(x,y) for each state.

2. Within each window, fold the full probability into groups defined by the identical haplotypes segments in the window.

3. At the last SNP of the window, unfold the group probability into full probability.

4. Repeat 1-3 for all windows.

This strategy reduces the state space with N^2 to a reduced space of size $N*h_i$ within each window, hence it saves memory cost for storage of full probability and also reduces the computational cost.

I describe the details in step 2 and 3. The main idea is illustrated in Figure 5.3. Assume the current window is from pth to qth marker and the full probability at *pth* marker is $L_p(x,y)$, where x or y is from 1 to N, indicating each reference haplotype (Figure 5.3(a)). The standard forward calculation requires us to calculate N^2 probabilities pth marker and continue to qth marker regardless of the haplotype configuration within the window (Figure 5.2(b)). Let the number of unique haplotypes within this window be h. Denote x^* as the group *xth* haplotype belongs to, so the x^* ranges from 1 to h and x ranges from1 to N. Then, I define four types of quantities in a reduced space at each marker in the window. Within the window, $L_i^{NR}(x^*,y^*)$, $L_i^{1RL}(x^*,y)$, $L_i^{1RR}(x,y^*)$ and $L_i^{2R}(x^*,y^*)$ are the grouped probabilities with no recombination at both chromosomes, recombination only in first chromosome, recombination only in second chromosome and recombination in both chromosomes between first and ith marker respectively.

First, I initialize the four quantities

$$L_p^{NR}(x^*, y^*) = \sum_{x \in x^*} \sum_{y \in y^*} L_i(x, y) \qquad L_p^{1RL}(x^*, y) = L_p^{1RR}(y^*, x) = L_p^{2R}(x^*, y^*) = 0$$

Second, I calculate the four quantities along the window recursively

$$L_{i+1}^{NR}(x^*, y^*) = L_i^{NR}(x^*, y^*) \times (1-\theta_i)^2 \times P(G_{i+1} / S_{i+1} = (x^*, y^*))$$

$$\begin{aligned} L_{i+1}^{1RL}(x^*, y) = {} & [L_i^{1RL}(x^*, y) \times (1-\theta_i)^2 + \sum_{a^*} L_i^{1RL}(a^*, y) \times m_{x^*}(1-\theta)\theta / N \\ & \sum_{a^*} L_i^{NR}(a^*, y^*) \times L_p(a^*, y) / L_p(a^*, y^*) \times m_{x^*}(1-\theta)\theta / N] \times P(G_{i+1} / S_{i+1} = (x^*, y^*)) \end{aligned}$$

$$L_{i+1}^{1RR}(x, y^*) = L_{i+1}^{1RL}(y^*, x)$$

$$\begin{aligned} L_{i+1}^{2R}(x^*, y^*) = {} & [L_i^{2R}(x^*, y^*) \times (1-\theta_i)^2 + \sum_{b^*} L_i^{2R}(x^*, b^*) \times m_{y^*}(1-\theta_i)\theta_i / N + \\ & \sum_{a^*} L_i^{2R}(a^*, y^*) \times m_{x^*}(1-\theta_i)\theta_i / N \\ & + \sum_{a^*, b^*} (L_i^{NR}(a^*, b^*) + L_i^{2R}(a^*, b^*)) \times m_{x^*} m_{y^*} \theta_i^2 / N^2] + \\ & + \sum_{b} L_i^{1RL}(x^*, b) \times m_{y^*}(1-\theta_i)\theta_i / N + \sum_{a} L_i^{1RR}(a, y^*) \times m_{x^*}(1-\theta_i)\theta_i / N \\ & + (\sum_{a^*, b} L_i^{1RL}(a^*, b) + \sum_{a, b^*} L_i^{1RR}(a, b^*)) \times m_{x^*} m_{y^*} \theta_i^2 / N^2] \\ & \times P(G_{i+1} / S_{i+1} = (x^*, y^*)) \end{aligned}$$

At the end of the window (*qth* marker), I can recover the full probability in full state space based on the four grouped quantities in reduced state space (Figure 5.3(c)).

$$\begin{aligned} L_q(x, y) &= L_q^{NR}(x^*, y^*) \times \frac{L_p(x, y)}{L_p^{NR}(x^*, y^*)} + \frac{L_q^{1RL}(x^*, y)}{m_{x^*}} + \frac{L_q^{1RR}(x, y^*)}{m_{y^*}} + \frac{L_q^{2R}(x^*, y^*)}{m_{x^*} \times m_{y^*}} \\ &= L_q^{NR}(x^*, y^*) \times \frac{L_p(x, y)}{L_p^{NR}(x^*, y^*)} + \frac{L_q^{1RL}(x^*, y)}{m_{x^*}} + \frac{+ L_q^{1RL}(y^*, x)}{m_{y^*}} + \frac{L_q^{2R}(x^*, y^*)}{m_{x^*} \times m_{y^*}} \end{aligned}$$

I need to emphasize that the covered full space quantities $L_q(x,y)$ is exactly same as the calculation from standard forward calculation in full state space. Hence, it is an exact method without any approximation to original HMM. I can move forward and loop through all windows. At each marker, only a set of grouped quantities are stored.

Optimization of Window Allocation

An inevitable question concerns the allocation of the windows. A very long window will yield many unique haplotypes while a very short window will yield many overhead costs on the boundaries of each window. An optimal strategy could be explored. Since the total computing cost is the sum of the calculation on the two boundaries of each window in full states and within the window in reduced states, an optimized window allocation based on the panel density will affect the final performance greatly. I am motivated by the dynamic programming for the shortest path problem. To illustrate the idea, I first define some cost (memory or CPU) functions:

C(i) : cost for ith marker in full space, *C(1,i)* : total cost from first to ith marker,

C(g(i,j)): cost to calculate from ith marker to jth marker in reduced space

The goal is to find a path *Path(1,M),* which minimizes *C(1,M)*, where M is the total number of markers. Since the number of all possible paths increase exponentially with M, looping through all path space is infeasible. Here, I present a dynamic programming strategy to find the optimal path. Assume there is an optimal path *Path(1,i)* minimizing the cost from first marker to *ith* marker *C(1,i)* for all $i < k$, the optimal path *Path(1,k)* is min{ *C(1,k)* + *C(g(k-1,k)), C(1,k-2)* + *C(g(k-2,k)), …, C(1,2)* + *C(g(2,k))* }. The computational complexity of looping from first marker to *Mth* markers with this dynamic programming is O(M^2), which is trivial compared to the forward calculation O(N^2M). Here N is the number of template haplotypes.

Simulated Data

I simulated a pool of 10,000 chromosomes in 1Mb region which mimics the degree of Linkage Disequilibrium (LD) in CEU samples [84]. Then, a subset of 50, 100, 200, 400 and 800 unphased individuals was randomly drawn from the reference pool. Three sets of the markers with 200 SNPs each were included in our simulation, representing different density panels similar to practice: a) 1 SNP per 5kb; b) 1SNP per 1kb; c) 1 SNP per 200b. I repeated above simulations for 100 times.

Real Dataset

I applied our method to three published real data sets: 1) 2000 samples genotyped on 500K chip from psoriasis GWAS; 2) 1094 samples from 1000 Genomes Project genotyped on Illumina Omni platform; 3) 1094 samples from 1000 Genomes Project Phase I data (www.1000genomes.org). The densities of the three data sets are approximately 1 SNP per 5000 bases, 1000 bases and 200 bases respectively. I randomly picked 50, 100, 200, 400 and 800 samples from the data sets with a 500-SNP window. Both the state space reduction method and the standard method were applied to infer the haplotypes.

5.3 Results and Discussion

As I discussed about the motivation of our method, the number of unique haplotypes increases slowly as the number of reference panel increases or the number of SNPs increases. I examined on a large pool of simulated haplotypes. In Figure 5.2 (a), for a fixed 10kb window, the number of unique haplotypes from 40 to 120 as the number of individuals increases from 200 to 1500. In Figure 5.2(b), the number of SNPs increases from 60 to 120 as the number of individuals increase from 200 to 1500. In Figure 5.2 (c), for a fixed sample size (N = 10,000), the number of unique haplotypes increases from 40 to 180 as the window size increases from 50 to 200. The pattern is consistent with the coalescent theory that the number of unique haplotypes increases in a log scale as the number of sequences increases, but increases linearly as the region expands [87].

Next, I will show the performance of our method in terms of CPU time and memory, compared to standard HMM algorithm. Table 5.2 presents the comparisons of the standard approach and our state space reduction method on three sets of simulated data. The general conclusion is that our method can reduce savings substantially in memory and computing time. The actual savings vary on sample size and marker density. The standard method performs consistently in different densities as expected and increases cubically as the sample size increases while our state space method increases more slowly. More specifically, for the density of 1 SNP per kb, the memory savings increase from 3 folds to 8 folds and the computing time savings increase from 2 folds to 6 folds as the sample size increases from 50 samples to 800 samples. The performance is better for the density of 1 SNP per 200b and slightly worse for the density of 1 SNP per 5 kb.

Despite the encouraging performance of our method in simulation data sets, real data sets are more useful to estimate the actual savings in practice. Table 5.3 presents the comparisons of computational cost in real data sets. The performance is similar to what I observed in simulated data sets. With 50 samples in the lowest density of 1 SNP per 5 kb, memory saving is about 3 fold and cpu time is comparable. With 800 samples in the highest density of 1 SNP per 200 b, memory saving is about 20 fold and CPU saving is about 6 fold.

Given the consistent results for both simulated and real data sets, I can conclude that the proposed state space reduction method can have a substantial saving in memory cost and a modest saving in computing time, comparing to the standard method, especially with a large number of samples. As the state space expands and density of the reference panel increases, I expect to have more gains.

From a pure algebraic point of view, the standard forward calculation in HMM here is just a series of matrix and vector operations. The calculation will go through each element of the matrix ignoring the specific properties of the matrix pattern. The state space reduction method is reorganizing the probabilities into a group of vectors with same properties and performing the calculation in a condensed way. Despite of the numerically equivalence of the calculation, the mathematical proof for haploid case is provided in the supplemental materials. The diploid case is very similar. This idea is potentially useful in more general scenarios where the transition and emission probability matrices are in specific formats.

Accuracy is not shown here because our method is numerical equivalent to the standard HMM implemented in MaCH. The accuracy of MaCH and other software has been comprehensively discussed [8, 88]. MaCH implemented the standard HMM efficiently using symmetric matrix for probabilities storage. The actual computing time and memory are halved compared to the standard approach in Table 5.2 and Table 5.3. Our method still outperforms MaCH in both computing time and memory, which suggests a potential replacement of MaCH for the purpose of haplotype inference for future application.

This algorithm can be potentially incorporated into some existing program such as MaCH [14]. A separate pre-released version for haplotype inference can be requested individually from the author at weich@umich.edu .

Table 5.1 Estimate of running time and memory for haplotype inference using MaCH

Sample	100	500	1000	2000
Time (hours)	0.47	59	472	3778
Memory (Gb)	0.08	1.99	7.98	31.97

The estimate is based on a single 2.8 GHz AMD Opterons CPU. For each data set, all samples are used to perform a full state space calculation. The number of iterations is set to 50.

Table 5.2 Comparisons of the computing time and memory in simulations

	Samples	50	100	200	400	800
5Kbp						
Memory(M)	Standard	7.7	31.5	127.4	512.0	2,053.1
	Grouping	2.9	9.2	29.3	93.9	316.3
CPU(s)	Standard	28	226	1,796	15,014	124,448
	Grouping	18	113	827	7,393	63,696
1Kbp						
Memory(M)	Standard	7.7	31.5	127.4	512.0	2053.1
	Grouping	2.3	7.3	24.0	79.2	269.4
CPU(s)	Standard	27	222	1,813	15,542	132,422
	Grouping	14.2	83.8	519.7	3,622	28,148.7
200bp						
Memory(M)	Standard	7.7	31.5	127.4	512.0	2,053.1
	Grouping	1.6	5.0	16.1	53.5	179.0
CPU(s)	Standard	28	222	1,785	15,313	124,895
	Grouping	11	59	345	2,233	15,937.6

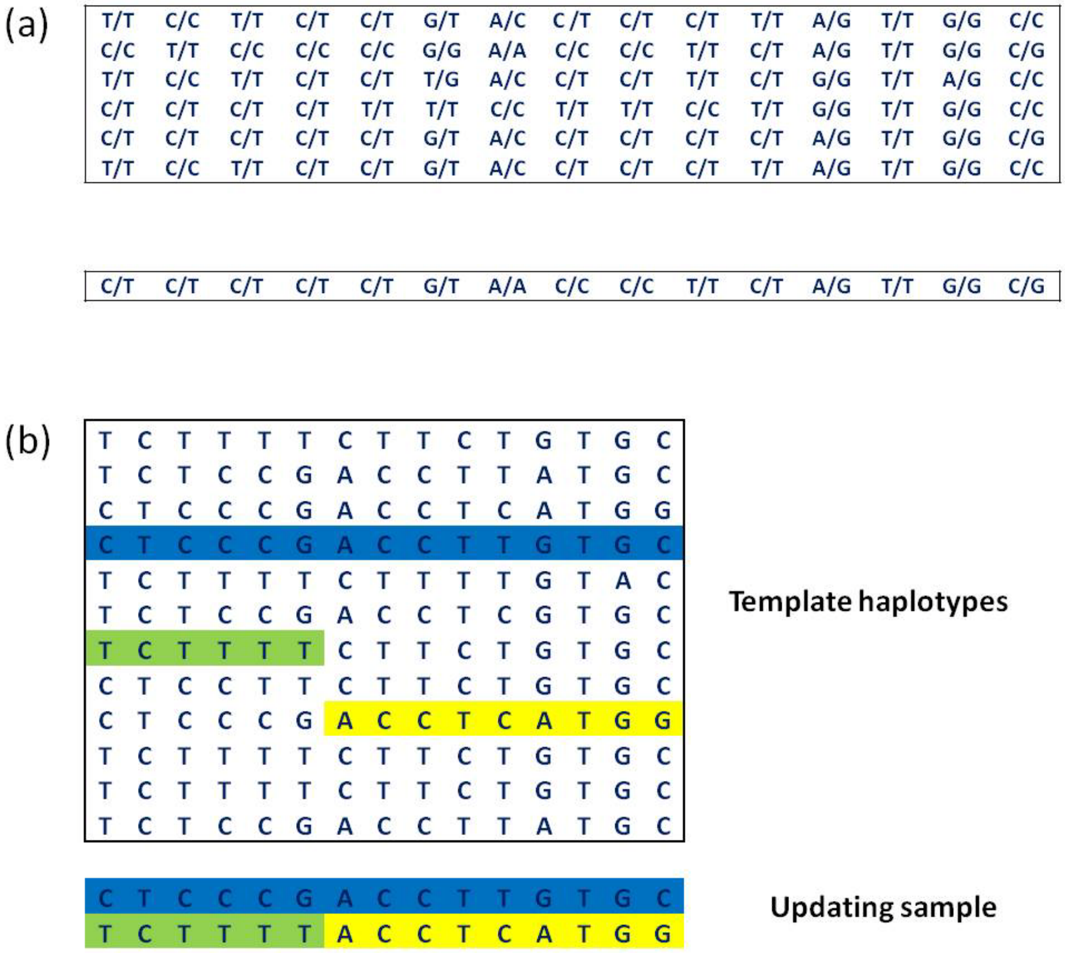
Comparisons of the computing time and memory between standard approach and state space reduction method for simulated data sets with 200-SNP window in three different densities (1 SNP per 5Kb/1Kb/200b)

Table 5.3 Comparisons of the computing time and memory in real data

	samples	50	100	200	400	800
GAIN_500K						
Memory(M)	Standard	19.2	78.4	316.8	1,273.6	5,107.2
	Grouping	6.2	18.6	58.1	188.5	629.4
CPU(s)	Standard	72	590	4,931	42,733	354,089
	Grouping	49	346	2,642	20,361	156,212
1KG_OMNI						
Memory(M)	Standard	19.2	78.4	316.8	1,273.6	5,107.2
	Grouping	3.7	11.0	33.7	127.1	464.0
CPU(s)	Standard	71	576	4965	42,342	353,343
	Grouping	38.2	256.4	1947	15,111.4	115,713.9
1KG						
Memory(M)	Standard	19.2	78.4	316.8	1,273.6	5107.2
	Grouping	2.7	8.3	26.9	100.4	361.4
CPU(s)	Standard	72	577	4,851	42,882	354,920
	Grouping	22	121	756	5664	46,285.8

Comparisons of the computing time and memory between standard approach and state space reduction method for real data with 500-SNP window

Figure 5.1 Cartoon view of the standard hidden Markov model for imputation and haplotype inference



T/T	C/C	T/T	C/T	C/T	G/T	A/C	C /T	C/T	C/T	T/T	A/G	T/T	G/G	C/C
C/C	T/T	C/C	C/C	C/C	G/G	A/A	C/C	C/C	T/T	C/T	A/G	T/T	G/G	C/G
T/T	C/C	T/T	C/T	C/T	T/G	A/C	C/T	C/T	T/T	C/T	G/G	T/T	A/G	C/C
C/T	C/T	C/T	C/T	T/T	T/T	C/C	T/T	T/T	C/C	T/T	G/G	T/T	G/G	C/C
C/T	C/T	C/T	C/T	C/T	G/T	A/C	C/T	C/T	C/T	C/T	A/G	T/T	G/G	C/G
T/T	C/C	T/T	C/T	C/T	G/T	A/C	C/T	C/T	C/T	T/T	A/G	T/T	G/G	C/C

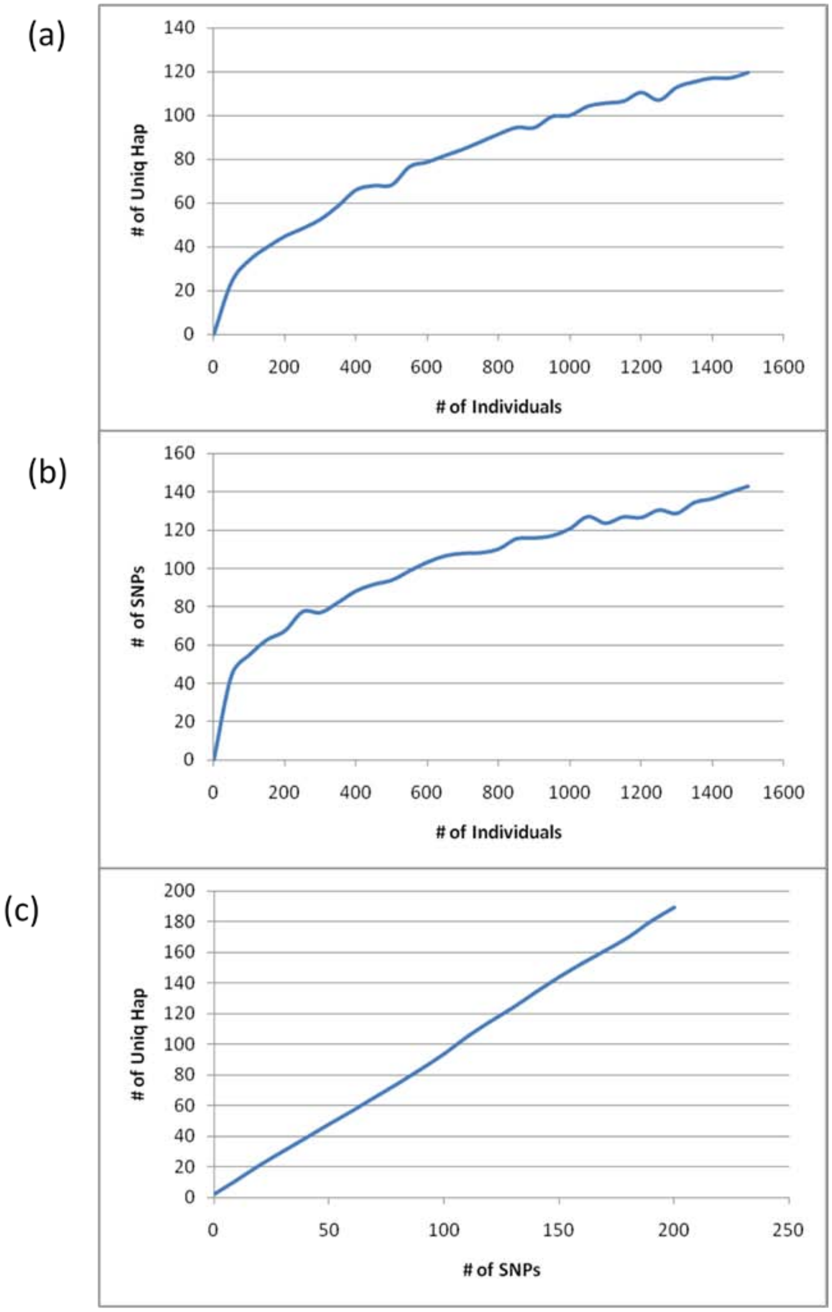
C/T	C/T	C/T	C/T	C/T	G/T	A/A	C/C	C/C	T/T	C/T	A/G	T/T	G/G	C/G

T	C	T	T	T	T	C	T	T	C	T	G	T	G	C
T	C	T	C	C	G	A	C	C	T	T	A	T	G	C
C	T	C	C	C	G	A	C	C	T	C	A	T	G	G
C	T	C	C	C	G	A	C	C	T	T	G	T	G	C
T	C	T	T	T	T	C	T	T	T	T	G	T	A	C
T	C	T	C	C	G	A	C	C	T	C	G	T	G	C
T	C	T	T	T	T	C	T	T	C	T	G	T	G	C
C	T	C	C	T	T	C	T	T	C	T	G	T	G	C
C	T	C	C	C	G	A	C	C	T	C	A	T	G	G
T	C	T	T	T	T	C	T	T	C	T	G	T	G	C
T	C	T	T	T	T	C	T	T	C	T	G	T	G	C
T	C	T	C	C	G	A	C	C	T	T	A	T	G	C

C	T	C	C	C	G	A	C	C	T	T	G	T	G	C
T	C	T	T	T	T	A	C	C	T	C	A	T	G	G

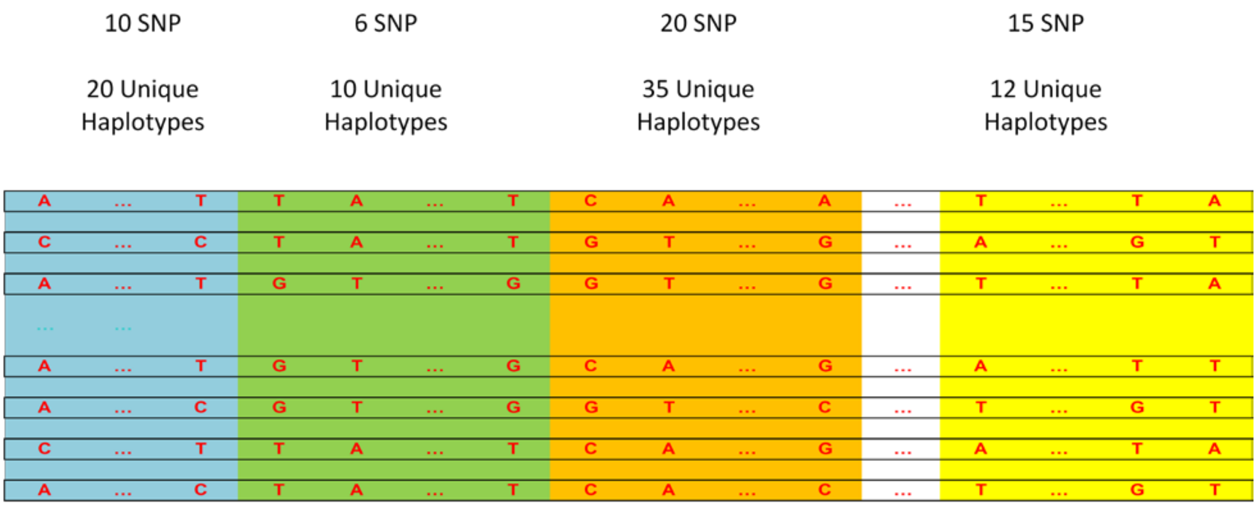
Haplotype inference of current updating sample. (a) A number of samples with observed genotypes. The bottom sample is being updated. The rest of the samples are treated as template haplotypes with random haplotypes assigned initially. (b) The phasing haplotypes of current sample are identified by mosaics of short stretches of the template haplotypes.

Figure 5.2 The pattern of the number of unique haplotypes (for a fixed window of 10kb)



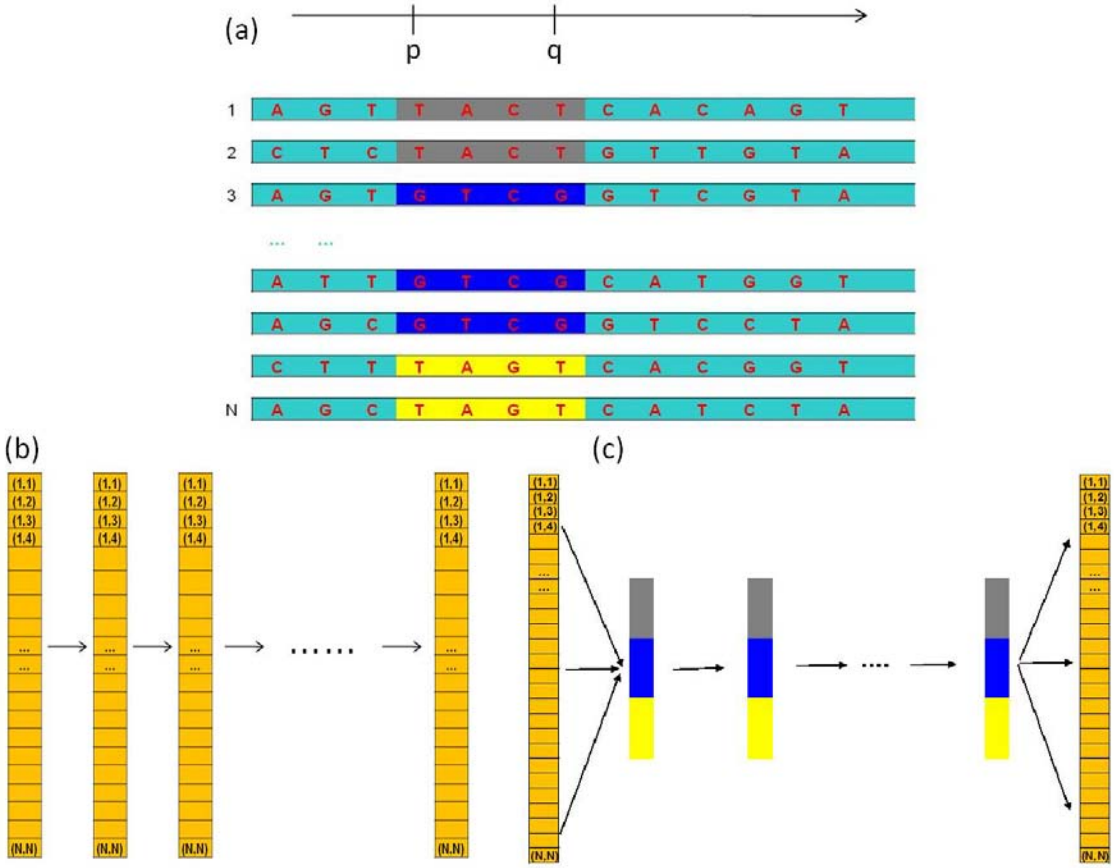
(a) The increasing pattern of unique haplotypes as the number of individuals increases.
(b) The increasing pattern of SNPs as the number of individuals increases.
(c) The increasing pattern of unique haplotypes as the number of individuals increases based on a pool of 50,000 samples

Figure 5.3 Decomposition of template haplotypes


10 SNP
20 Unique Haplotypes
6 SNP
10 Unique Haplotypes
20 SNP
35 Unique Haplotypes
15 SNP
12 Unique Haplotypes
A … T T A … T C A … A … T … T A
C … C T A … T G T … G … A … G T
A … T G T … G G T … G … T … T A
… …
A … T G T … G C A … G … A … T T
A … C G T … G G T … C … T … G T
C … T T A … T C A … G … A … T A
A … C T A … T C A … C … T … G T

This example consists of 200 haplotypes with 100 SNPs. 100 SNPs were chunked into difference windows. Within each window, the number of unique haplotypes is counted. E.g. The first windows consists of 10 SNPs and 20 unique haplotypes.

Figure 5.4 Cartoon view of the state space reduction method in an exampled window


(a)
p
q
1 A G T T A C T C A C A G T
2 C T C T A C T G T T G T A
3 A G T G T C G G T C G T A
... ...
A T T G T C G C A T G G T
A G C G T C G G T C C T A
C T T T A G T C A C G G T
N A G C T A G T C A T C T A
(b)
(1,1)
(1,2)
(1,3)
(1,4)
...
(N,N)
(c)
(1,1)
(1,2)
(1,3)
(1,4)
...
(N,N)

(a) Grouping of unique haplotypes in colored window. Color indicates the different groups in the window.
(b) The standard forward calculation from pth marker to qth marker.
(c) The state space reduction method from pth marker to qth marker

Supplemental Materials

Proof of the equivalence in probability of the two approaches for haploid case

$$F_p(j) = P(H_1,...,H_p, S_p = j) \qquad \mathrm{j=1,2,...}N$$

$$initiation:$$

$$F_p^R(j^*) = 0 \qquad\qquad F_p^{NR}(j^*) = \sum_{k \in j^*} F_p(k)$$

$$F_{i+1}^{NR}(j^*) = F_i^{NR}(j^*) \times (1-\theta_i) \times P(H_{i+1} / S_{i+1} = j^*)$$

$$F_{i+1}^R(j^*) = \left\{ \sum_{k^*} [F_i^R(k^*) + F_i^{NR}(k^*)] \times m_{j^*}\theta_i / N + F_i^R(j^*)[1-\theta_i] \right\} \times P(H_{i+1} / S_{i+1} = j^*)$$

$$\vdots$$

$$F_q^R(j^*) \quad \text{and} \quad F_q^{NR}(j^*)$$

$$F_q(j) = \frac{F_q^R(j^*)}{m_{j^*}} + F_q^{NR}(j^*) \times \frac{F_p(j)}{F_p^{NR}(j^*)} \qquad m_{j^*} \text{ is the number of haplotypes in group } j^* \quad (*)$$

$$\mathrm{P(S_i} = j^* / S_{i-1} = k^*) = \begin{cases} 1-\theta_{i-1} + m_{k^*}\theta_{i-1} / N & \mathrm{j^* = k^*} \\ m_{j^*}\theta_{i-1} / N & \mathrm{j^* \neq k^*} \end{cases}$$

To prove in (*), $F_q(j) = \frac{F_q^R(j^*)}{m_{j^*}} + F_q^{NR}(j^*) \times \frac{F_p(j)}{F_p^{NR}(j^*)} = P(H_1,...,H_q, S_q = j)$

Let Rec(i, j) denotes that there are recombinations from i to j and

Nonrec(i, j) denotes that there are no recombination from i to j, i < j

$$F_q^{NR}(j^*) = F_p^{NR}(j^*) \prod_{i=p}^{q-1} (1-\theta_i) P(H_{i+1} / S_{i+1}) = F_p^{NR}(j^*) \prod_{i=p}^{q-1} P(S_{i+1} = j, \mathrm{Nonrec(i,i+1)} / S_i = j) P(H_{i+1} / S_{i+1})$$

$$F_q^{NR}(j^*) \times \frac{F_p(j)}{F_p^{NR}(j^*)} = F_p(j) \prod_{i=p}^{q-1} P(S_{i+1} = j, \mathrm{Nonrec(i,i+1)} / S_i = j) P(H_{i+1} / S_{i+1})$$

$$= P(H_1, H_2,...,H_q, S_p = S_{p+1} = \cdots = S_q = j, \mathrm{Nonrec(p,q)})$$

$$= P(H_1, H_2,...,H_q, S_q = j, \mathrm{Nonrec(p,q)}) \quad (1)$$

$$\frac{F_q^R(j^*)}{m_{j^*}} = \left\{ \sum_{k^*} [F_{q-1}^R(k^*) + F_{q-1}^{NR}(k^*)] \times m_{j^*}\theta_i / N + F_{q-1}^R(j^*)[1-\theta_{q-1}] \right\} \times P(H_q / S_q = j^*)$$

$$= P(H_1, H_2,...,H_q, S_q = j, \mathrm{Rec(p,q)}) \qquad (2)$$

$$F_q(j) = \frac{F_q^R(j^*)}{m_{j^*}} + F_q^{NR}(j^*) \times \frac{F_p(j)}{F_p^{NR}(j^*)} = P(H_1,...,H_q, S_q = j)$$

Proof for diploid case is very similar.

Chapter 6
Summary and Discussion

So far, I have shown several topics I have been working on in the past few years. These novel methods and analysis results are all motivated from the experimental genetic data and have practical importance. I will briefly summarize each chapter and then discuss about their limitations and possible future work. Furthermore, I will also discussion ongoing work will also be discussed.

6.1 Summary and Future Work of Chapter 2-5

In Chapter 2, I presented a genome-wide association study on AMD and proposed a prediction model to investigate the cumulative risk of individuals. A novel gene *TIMP3* was identified and a potential interesting HDL-related pathway was proposed to be associated the AMD. Those findings broaden our knowledge of etiology of the disease and improve our understanding of biology in the retina. In addition, integrating all known and novel loci, a prediction model based on logistic regression can cluster high and low risk patients very well and will potentially be useful in clinical practice.

In Chapter 3, motivated by the numerous requests for GWAS results from collaborators, I developed an efficient user-friendly program to store and visualize the results, especially for multiple dimensional phenotypes. A successful application example is used in a project of expression Quantitative Traits Loci (eQTL). It has been downloaded over four hundred times.

In Chapter 4, I proposed a novel method to infer genotypes and phasing for sequencing parent-offspring trios. This method was motivated by the ongoing Sardinia low pass whole-genome sequencing project. I combined the family constraints of parent-offspring trios and LD information into a unified framework. The results from simulation and real data sets suggest that my method can improve the genotype accuracy significantly. In addition, the haplotype inference is also more accurate, which is often very crucial in follow-up imputation in existing GWAS data.

In Chapter 5, I proposed a state space reduction method to reduce the computational complexity of existing programs for haplotype inference and imputation. The idea is to group the haplotypes, which have the same local sequences, into a reduced space and reduce computation and storage. With the same accuracy, this method can reduce the computational cost in both memory and CPU time greatly in the process of the forward calculation in the hidden Markov model. The gain is growing as the sample size increases.

Although a lot of progresses have been done, none of the above topics reach the end. Those works have potential applications in the new context. I will brief describe some ongoing and future works.

Our GWAS in Chapter 2 identified a novel gene and a potentially interesting pathway. To detect more genetic variants with small effect size or low minor allele frequency, more samples are needed. A common approach is meta-analysis: combining a number of studies with GWAS data and estimating pooled pvalue. To achieve this goal, an international consortium “AMDGene Consortium” was organized with over 15 groups with a total of ~8000 cases and ~50,000 controls. As one of the core members, my work involves designing the analysis protocol, checking data quality and conducting the analysis. I applied both z-statistic and inverse variance method to the data [89]. The preliminary results are very promising and confirmed my findings in Chapter 2 and also brought the two proposed interesting hits to the genome-wide significance. Besides that, several new signals have been identified and under replication in several independent studies with a total of ~10,000 cases and ~8,000 controls. All those findings will greatly improve our knowledge of genetic basis of AMD and provide more clues about biology of the disease. As sequencing cost drops, sequencing in known genes, exome and whole genome will become feasible. This will further our understanding of AMD and biology of retina.

With the rapid advances of the next generation sequencing technology, I will have more computational and statistical challenges with huge amount of sequencing data. The method I have developed for SNP calling and haplotype inference for parent-offspring trios will have more applications in the real data analysis. However, as I can expect the sequencing cost will reduce drastically in the next several years, more family-based sequencing projects will be carried out. My method could be extended to nuclear and general family in a few ways, from heuristic to more complicated approaches. I will explore more in this direction to assistant existing and future sequencing projects. With this tool in hand, together with the variant calling method my colleague Bingshan and I developed, I plan to study the optimal design at different scenarios. For example, people would like to know if sequencing 100 parent-offspring trios at the coverage of 8X has more power to detect association signals than sequencing 100 nuclear families with two children at the coverage of 6X given the same sequencing cost.

6.2 Ongoing Work and Future plan

The next crucial step after genotype calling is association test. The fast increasing number of variants poses the difficulty in multiple testing problems. I would like to briefly describe an ongoing work to control the false discovery rate (FDR) accounting for correlations, which is potentially useful in genetic association studies. Methods of FDR control are widely used in genetics and genomics study [90]. However, the traditional procedures usually ignore the dependency of the tests, resulting loss of power. Genome-wide association study (GWAS) examines the association between phenotype and hundreds of thousands SNPs. The highly correlated structure of the SNP array requires some novel method to account for the dependency information.

It is motivated by the paper Sun and Cai (2007), which shows z statistic based method is more powerful than p-value based methods. In particular, their procedure is very useful when the underlying dependence structure forms a markov chain with time-independent transition matrix [91, 92]. However, the assumption of constant transition matrix does not hold in practice and may result in loss of efficiency. I aim to extend their approach under more general dependence structure, (e.g. time-dependent transition matrix) and incorporate some prior information from raw data to increase the power while controlling the nominal FDR level. First, I present a HMM approach described in Sun's paper.

Assume there are m tests, with observed a test statistic x_i i = 1,2,…,m in some specific order. The hidden state is denoted as $\theta_i = 0$ (null) or 1 (signal). The transition matrix between adjacent tests is defined as $A = \begin{pmatrix} a & 1-a \\ 1-b & b \end{pmatrix}$.

Emission probability is defined as $x_i \mid \theta_i = 0 \sim N(0,1)$, $x_i \mid \theta_i = 1 \sim N(\mu,1)$ (or mixture normal). Local index of significance (LIS) is defined as $LIS_i = P_\vartheta(\theta_i = 0 \mid \vec{x})$, where $P_\vartheta(\theta_i = 0 \mid \vec{x})$ is the posterior probability calculated by forward backward algorithm. $Let\ k = \max\left\{i: \frac{1}{i}\sum_{j=1}^{i} \widetilde{LIS}_{(j)}(x) \le \alpha\right\}$, rejecting all $H_{(i)}, i = 1, \dots, k$, controls FDR level at α.

As a demonstration example, I assume initial state distribution $\pi = (1,0)$, Transition matrix $A_i = \begin{pmatrix} 0.8 & 0.2 \\ 1-a_i & a_i \end{pmatrix}$

Underlying state $\theta_i = 0$ or 1. 0 refers non-significant, 1 refers significant.

The observations x_i , i = 1,2,…,m are generated by following distribution:

$x_i \mid \theta_i = 0 \sim N(0,1)$, $x_i \mid \theta_i = 1 \sim N(\mu,1)$

I applied conventional method BH procedure, adaptive p-value procedure, homogeneous HMM approach (OR) and nonhomogeneous HMM (OR.1) approach to the data and compare FDR and FNR.

For nonhomogenous transition matrix, the parameters are difficult to estimate and often not identifiable. If the transition matrices share the same set of parameters, I can estimate them through an EM algorithm. Although there is no theory to guarantee the convergence, it converges for most cases in practice from our experience.

In the demonstration example, the simulation results (Figure 6.1) show that BH is always conservative as expected, while the other three procedures control FDR at desired level even for non-homogenous case. OR.1 has more power (less FNR) than other three since it considers the true nonhomogeneous matrix especially in modest size of signals. However, in reality, it is difficult to estimate all parameters with the assumption that each transition matrix has its own parameters. Instead, I can assume some pattern of the transition matrices such that they share same set of parameters.

GWAS examine tens of thousands SNPs placed along the genome. SNPs tend to be correlated with each other in short region due to linkage disequilibrium, often measured in r^2. Figure 6.2 shows the distribution of r^2 on chromosome 22 from HapMap project. LD between SNPs can extend as far as a few hundred thousand bases, which results to clusters of signals. In next generation sequencing, multiple rare variants not easily tagged by common SNPs in small region may cause similar patterns. I aim to propose some realistic transition matrices to account for LD information between adjacent SNPs and increase the power to detect true variants.

A simple starting example is defined below:

$$A_i = \begin{pmatrix} p_0 + r_i^2(1-p_0) & (1-p_0)(1-r_i^2) \\ p_0(1-r_i^2) & 1-p_0+p_0r_i^2 \end{pmatrix}$$

where A_i and r_i are the transition matrix and correlation for two adjacent SNPs. P_0 is the proportion of false signals and requires be to estimated. Given this specific transition matrix, the E-M algorithm described previously can be used to estimate p_o. The p-value for each marker can be calculated through this process. I will explore it further in near future.

6.3 Conclusion

The research of human genetics will reach to another era with the breaking-through and remarkable technologies. The GWAS approach has proven to be successful to identify common variants associated with various disorders or complex traits. Next generation sequencing provides a deep catalog of the human genetic variations and enables us to look for missing heritability through rare variants and mutations. However, we are still far away from the ultimate goal: the cure of human disease. The intermediate steps require many novel statistical and computational methods to deal with prodigious accounts of data from different levels: DNA, RNA and protein. Careful analysis and appropriate interpretation will help us to understand the fundamental mechanisms of the biology. I believe what I have done and what I plan to do will continue to facilitate the discovery of diseases-associated genes/pathways and understanding of the etiology of the disease, particularly in the context of large-scale sequencing.

Figure 6.1 A comparison of some conventional methods, homogeneous HMM approach and non-homogeneous HMM approach

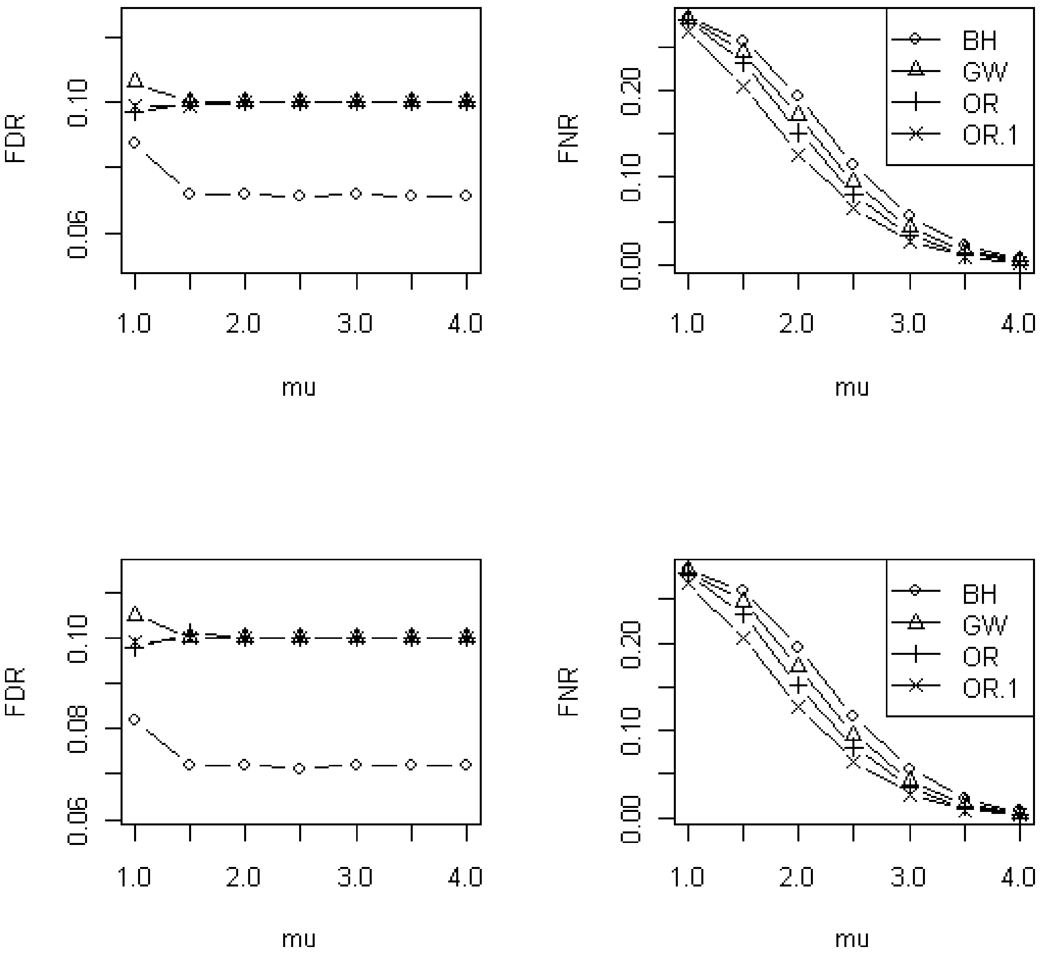
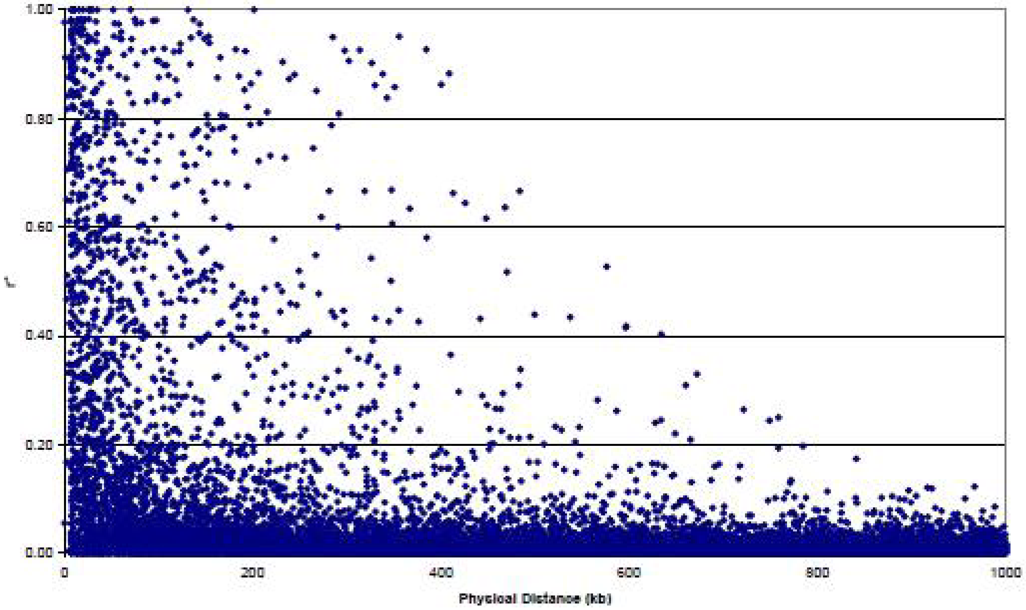


Figure 6.2 Raw R^2 distribution from chromosome 22



REFERENCES

1. McCarthy, M.I., et al., *Genome-wide association studies for complex traits: consensus, uncertainty and challenges.* Nat Rev Genet, 2008. **9**(5): p. 356-69.
2. Lander, E.S. and N.J. Schork, *Genetic dissection of complex traits.* Science, 1994. **265**(5181): p. 2037-48.
3. Wellcome Trust Case Control Consortium, *Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.* Nature, 2007. **447**(7145): p. 661-78.
4. Chen, W., et al., *Genetic variants near TIMP3 and high-density lipoprotein-associated loci influence susceptibility to age-related macular degeneration.* Proc Natl Acad Sci U S A, 2010. **107**(16): p. 7401-6.
5. Kathiresan, S., et al., *Common variants at 30 loci contribute to polygenic dyslipidemia.* Nature Genetics, 2009. **41**(1): p. 56-65.
6. Manolio, T.A., et al., *Finding the missing heritability of complex diseases.* Nature, 2009. **461**(7265): p. 747-53.
7. Risch, N.J., *Searching for genetic determinants in the new millennium.* Nature, 2000. **405**(6788): p. 847-56.
8. Li, Y., et al., *Genotype imputation.* Annu Rev Genomics Hum Genet, 2009. **10**: p. 387-406.
9. Willer, C.J., et al., *Six new loci associated with body mass index highlight a neuronal influence on body weight regulation.* Nat Genet, 2009. **41**(1): p. 25-34.
10. Scheet, P. and M. Stephens, *A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase.* Am J Hum Genet, 2006. **78**(4): p. 629-44.
11. Browning, B.L. and S.R. Browning, *A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals.* Am J Hum Genet, 2009. **84**(2): p. 210-23.
12. Marchini, J., et al., *A new multipoint method for genome-wide association studies by imputation of genotypes.* Nat Genet, 2007. **39**(7): p. 906-13.
13. Li, Y. and G. Abecasis, *Mach 1.0: Rapid Haplotype Reconstruction and Missing Genotype Inference.* Am J Hum Genet, 2006. **S79**: p. 2290.
14. Li, Y., et al., *MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes.* Genet Epidemiol, 2010. **34**(8): p. 816-34.
15. Swaroop, A., et al., *Unravelling a Late-Onset Multifactorial Disease: From Genetic Susceptibility to Disease Mechanisms for Age-related Macular Degeneration.* Annu Rev Genomics Hum Genet, 2009. **10**: p. (in press).
16. Congdon, N., et al., *Causes and prevalence of visual impairment among adults in the United States.* Arch Ophthalmol, 2004. **122**(4): p. 477-85.

17. Jager, R.D., W.F. Mieler, and J.W. Miller, *Age-related macular degeneration.* N Engl J Med, 2008. **358**(24): p. 2606-17.

18. Jackson, G.R., C. Owsley, and C.A. Curcio, *Photoreceptor degeneration and dysfunction in aging and age-related maculopathy.* Ageing Res Rev, 2002. **1**(3): p. 381-96.

19. Fisher, S.A., et al., *Meta-analysis of genome scans of age-related macular degeneration.* Hum Mol Genet, 2005. **14**(15): p. 2257-64.

20. Weeks, D.E., et al., *Age-related maculopathy: an expanded genome-wide scan with evidence of susceptibility loci within the 1q31 and 17q25 regions.* Am J Ophthalmol, 2001. **132**(5): p. 682-92.

21. Weeks, D.E., et al., *Age-related maculopathy: a genomewide scan with continued evidence of susceptibility loci within the 1q31, 10q26, and 17q25 regions.* Am J Hum Genet, 2004. **75**(2): p. 174-89.

22. Abecasis, G.R., et al., *Age-related macular degeneration: a high-resolution genome scan for susceptibility loci in a population enriched for late-stage disease.* Am J Hum Genet, 2004. **74**(3): p. 482-94.

23. Seddon, J.M., et al., *A genomewide scan for age-related macular degeneration provides evidence for linkage to several chromosomal regions.* Am J Hum Genet, 2003. **73**(4): p. 780-90.

24. Klein, R.J., et al., *Complement factor H polymorphism in age-related macular degeneration.* Science, 2005. **308**(5720): p. 385-9.

25. Dewan, A., et al., *HTRA1 promoter polymorphism in wet age-related macular degeneration.* Science, 2006. **314**(5801): p. 989-92.

26. Rivera, A., et al., *Hypothetical LOC387715 is a second major susceptibility gene for age-related macular degeneration, contributing independently of complement factor H to disease risk.* Hum Mol Genet, 2005. **14**(21): p. 3227-36.

27. Jakobsdottir, J., et al., *Susceptibility genes for age-related maculopathy on chromosome 10q26.* Am J Hum Genet, 2005. **77**(3): p. 389-407.

28. Haines, J.L., et al., *Complement factor H variant increases the risk of age-related macular degeneration.* Science, 2005. **308**(5720): p. 419-21.

29. Edwards, A.O., et al., *Complement factor H polymorphism and age-related macular degeneration.* Science, 2005. **308**(5720): p. 421-4.

30. Fagerness, J.A., et al., *Variation near complement factor I is associated with risk of advanced AMD.* Eur J Hum Genet, 2009. **17**(1): p. 100-4.

31. Gold, B., et al., *Variation in factor B (BF) and complement component 2 (C2) genes is associated with age-related macular degeneration.* Nat Genet, 2006. **38**(4): p. 458-62.

32. Maller, J.B., et al., *Variation in complement factor 3 is associated with risk of age-related macular degeneration.* Nat Genet, 2007. **39**(10): p. 1200-1.

33. Yates, J.R., et al., *Complement C3 variant and the risk of age-related macular degeneration.* N Engl J Med, 2007. **357**(6): p. 553-61.

34. Gunderson, K.L., et al., *Whole genome genotyping.* Methods in Enzymology, 2006. **410**: p. 359-376.

35. Thornton, T. and M.S. McPeek, *Case-control association testing with related individuals: a more powerful quasi-likelihood score test.* Am J Hum Genet, 2007. **81**(2): p. 321-37.

36. Barrett, J.C., et al., *Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease.* Nat Genet, 2008. **40**(8): p. 955-62.

37. Zeggini, E., et al., *Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes.* Nat Genet, 2008. **40**(5): p. 638-45.

38. Willer, C.J., et al., *Six New Loci Associated with Body Mass Index Highlight a Neuronal Influence on Body Weight Regulation.* Nature Genetics, 2009. **41**: p. 25-34.

39. Willer, C.J., et al., *Genome-Wide Association Scans Identify Novel Loci That Influence Lipid Levels and Risk of Coronary Artery Disease.* Nature Genetics, 2008. **40**: p. 161-9.

40. Price, A.L., et al., *Principal components analysis corrects for stratification in genome-wide association studies.* Nat Genet, 2006. **38**(8): p. 904-9.

41. The International HapMap Consortium, *A second generation human haplotype map of over 3.1 million SNPs.* Nature, 2007. **449**: p. 851-61.

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

43. Tian, C., et al., *Analysis and application of European genetic substructure using 300 K SNP information.* PLoS Genet, 2008. **4**(1): p. e4.

44. Li, M., et al., *CFH haplotypes without the Y402H coding variant show strong association with susceptibility to age-related macular degeneration.* Nat Genet, 2006. **38**(9): p. 1049-54.

45. Maller, J., et al., *Common variation in three genes, including a noncoding variant in CFH, strongly influences risk of age-related macular degeneration.* Nat Genet, 2006. **38**(9): p. 1055-9.

46. Goring, H.H., J.D. Terwilliger, and J. Blangero, *Large upward bias in estimation of locus-specific effects from genomewide scans.* Am J Hum Genet, 2001. **69**(6): p. 1357-69.

47. Kanda, A., et al., *A variant of mitochondrial protein LOC387715/ARMS2, not HTRA1, is strongly associated with age-related macular degeneration.* Proc Natl Acad Sci U S A, 2007. **104**(41): p. 16227-32.

48. Fritsche, L.G., et al., *Age-related macular degeneration is associated with an unstable ARMS2 (LOC387715) mRNA.* Nat Genet, 2008. **40**(7): p. 892-6.

49. Feng, J., et al., *Regulation of neurotransmitter release by synapsin III.* J Neurosci, 2002. **22**(11): p. 4372-80.

50. Weber, B.H., et al., *Mutations in the tissue inhibitor of metalloproteinases-3 (TIMP3) in patients with Sorsby's fundus dystrophy.* Nat Genet, 1994. **8**(4): p. 352-6.

51. Malek, G., et al., *Apolipoprotein B in cholesterol-containing drusen and basal deposits of human eyes with age-related maculopathy.* Am J Pathol, 2003. **162**(2): p. 413-25.

52. Mullins, R.F., et al., *Drusen associated with aging and age-related macular degeneration contain proteins common to extracellular deposits associated with atherosclerosis, elastosis, amyloidosis, and dense deposit disease.* FASEB J, 2000. **14**(7): p. 835-46.

53. Curcio, C.A., et al., *Esterified and unesterified cholesterol in drusen and basal deposits of eyes with age-related maculopathy.* Exp Eye Res, 2005. **81**(6): p. 731-41.

54. Klein, R., B.E. Klein, and T. Franke, *The relationship of cardiovascular disease and its risk factors to age-related maculopathy. The Beaver Dam Eye Study.* Ophthalmology, 1993. **100**(3): p. 406-14.

55. Tomany, S.C., et al., *Risk factors for incident age-related macular degeneration: pooled findings from 3 continents.* Ophthalmology, 2004. **111**(7): p. 1280-7.

56. Nair, R.P., et al., *Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways.* Nat Genet, 2009. **41**(2): p. 199-204.

57. Seddon, J.M., et al., *Prediction model for prevalence and incidence of advanced age-related macular degeneration based on genetic, demographic and environmental variables.* Investigative Ophthalmology and Visual Science, 2009. **50**: p. (in print).

58. Jakobsdottir, J., et al., *Interpretation of genetic association studies: markers with replicated highly significant odds ratios may be poor classifiers.* PLoS Genet, 2009. **5**(2): p. e1000337.

59. Chen, W., L. Liang, and G.R. Abecasis, *GWAS GUI: graphical browser for the results of whole-genome association studies with high-dimensional phenotypes.* Bioinformatics, 2009. **25**(2): p. 284-5.

60. Scott, L.J., et al., *A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants.* Science, 2007. **316**(5829): p. 1341-5.

61. WTCCC, *Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.* Nature, 2007. **447**(7145): p. 661-78.

62. Moffatt, M.F., et al., *Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma.* Nature, 2007. **448**(7152): p. 470-3.

63. Dixon, A.L., et al., *A genome-wide association study of global gene expression.* Nat Genet, 2007. **39**(10): p. 1202-7.

64. Sanna, S., et al., *Common variants in the GDF5-UQCC region are associated with variation in human height.* Nat Genet, 2008. **40**(2): p. 198-203.

65. Cheung, V.G., et al., *Mapping determinants of human gene expression by regional and genome-wide association.* Nature, 2005. **437**(7063): p. 1365-9.

66. 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, 2007. **3**(4): p. e58.

67. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses.* Am J Hum Genet, 2007. **81**(3): p. 559-75.

68. Chen, W.M. and G.R. Abecasis, *Family-based association tests for genomewide association scans.* Am J Hum Genet, 2007. **81**(5): p. 913-26.

69. Hindorff, L.A., et al., *Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.* Proc Natl Acad Sci U S A, 2009. **106**(23): p. 9362-7.

70. Cirulli, E.T. and D.B. Goldstein, *Uncovering the roles of rare variants in common disease through whole-genome sequencing.* Nat Rev Genet, 2010. **11**(6): p. 415-25.

71. Li, B. and S.M. Leal, *Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data.* Am J Hum Genet, 2008. **83**(3): p. 311-21.
72. Eichler, E.E., et al., *Missing heritability and strategies for finding the underlying causes of complex disease.* Nat Rev Genet, 2010. **11**(6): p. 446-50.
73. Le, S.Q. and R. Durbin, *SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples.* Genome Res, 2010.
74. Roach, J.C., et al., *Analysis of genetic inheritance in a family quartet by whole-genome sequencing.* Science, 2010. **328**(5978): p. 636-9.
75. Li, Y., et al., *Low-coverage sequencing: Implications for design of complex trait association studies.* Genome Res, 2011. **21**(6): p. 940-51.
76. Le, S.Q. and R. Durbin, *SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples.* Genome Res, 2010. **21**(6): p. 952-60.
77. Li, H., J. Ruan, and R. Durbin, *Mapping short DNA sequencing reads and calling variants using mapping quality scores.* Genome Res, 2008. **18**(11): p. 1851-8.
78. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.
79. McKenna, A., et al., *The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.* Genome Res, 2010. **20**(9): p. 1297-303.
80. Li, B., W. Chen, and G. Abecasis, *A likelihood based framework for variant calling and de novo mutation detection in families for sequencing data.* Submitted, 2011.
81. Li, N. and M. Stephens, *Modelling Linkage Disequilibrium, and identifying recombination hotspots using SNP data.* Genetics, 2003. **165**: p. 2213-2233.
82. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.
83. Rabiner, L.R., *A Tutorial on Hidden Markov-Models and Selected Applications in Speech Recognition.* Proceedings of the Ieee, 1989. **77**(2): p. 257-286.
84. Schaffner, S.F., et al., *Calibrating a coalescent simulation of human genome sequence variation.* Genome Res, 2005. **15**(11): p. 1576-83.
85. Nothnagel, M., et al., *A comprehensive evaluation of SNP genotype imputation.* Hum Genet, 2009. **125**(2): p. 163-71.
86. Kathiresan, S., et al., *Common DNA Sequence Variants at Thirty Genetic Loci Contribute to Polygenic Dyslipidemia.* Nature Genetics, 2009. **41**: p. 56-65.
87. Hein, J.S., M. H., and Wiuf, C, *Gene Genealogies, Variation and Evolution – A Primer in Coalescent Theory*. 2005: Oxford University Press.
88. Pei, Y.F., et al., *Analyses and comparison of imputation-based association methods.* PLoS One, 2010. **5**(5): p. e10827.
89. Willer, C.J., Y. Li, and G.R. Abecasis, *METAL: fast and efficient meta-analysis of genomewide association scans.* Bioinformatics. **26**(17): p. 2190-1.
90. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society Series B-Methodological, 1995. **57**(1): p. 289-300.

91. Sun, W.G. and T.T. Cai, *Large-scale multiple testing under dependence*. Journal of the Royal Statistical Society Series B-Statistical Methodology, 2009. **71**: p. 393-424.

92. Wei, Z., et al., *Multiple testing in genome-wide association studies via hidden Markov models*. Bioinformatics, 2009. **25**(21): p. 2802-8.