METHODS FOR GENOME INTERPRETATION: CAUSAL GENE DISCOVERY AND PERSONAL PHENOTYPE PREDICTION

by

Yun-Ching Chen

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland
September, 2014

© Yun-Ching Chen 2014
All rights reserved

Abstract

Genome interpretation – illustrating how genomic variation affects phenotypic variation – is one of the central questions of the early 21st century. Deciphering the mapping between genotypes and phenotypes requires the collection of a large amount of data, both genetic and phenotypic. Phenotypic profiles, for example, have been systematically recorded and archived in hospitals and national health-related organizations for years. Human genome sequences, however, had not been sequenced in a high throughput manner until next-generation sequencing technologies became available in 2005. Since then, vast amounts of genotype-phenotype data have been collected, allowing for the unprecedented opportunity for genome interpretation.

Genome interpretation is an ambitious, poorly understood goal that may require collaboration between many disciplines. In this dissertation, I focus on the development of computational methods for genome interpretation. Based on recent interest in relating genotypes and phenotypes, the task is divided into two stages: discovery (Chapters 2-6) and prediction (Chapters 7-10). In the discovery stage, the location of genomic loci associated with a phenotype of interest is identified based on sequence-based case-control studies. In the prediction stage, I propose a probabilistic model to predict personal phenotypes given an individual's genome by integrating many sources of information, including the phenotype-associated loci found in the discovery stage.

Advisor: Dr. Rachel Karchin

Reader: Dr. Joel Bader

Acknowledgement

I would like to acknowledge the people who have helped me during my graduate studies. First, I would like to thank my PhD advisor Dr. Rachel Karchin. I feel grateful for her encouragement and full support of my ideas for my graduate studies. She works very hard bringing together collaborations and connections that help her students in the lab. I admire her enthusiasm for science and the constant effort she puts forth towards research. From her, I learned a lot about how to be a great scientist and pursue impactful and rigorous research. She not only gave me insightful advice regarding my research but also shared her personal experiences with me as guidance for my future directions. I truly feel lucky that I had the opportunity to work for my PhD with her support.

I would like to thank my thesis committee, which consisted of Dr. Donald Geman, Dr. Joel Bader and Dr. Dan Arking. I thank them for their generosity and willingness to take time from their busy schedules for my thesis meeting and defense. They all gave me many great ideas and advice to improve my work.

I would like to thank my collaborators in the Bipolar disorder project at Hopkins: Dr. James Potash, Dr. Peter Zandi, Dr. Mehdi Pirooznia and Dr. Fernando Goes. They closely collaborated with us to help me on my first project – the development of a causal gene discovery method. Mehdi helped me with the upstream data analysis and discussed every technical detail with me with endless patience. Peter gave great insights on method development. Jimmy and Fernando offered important input from the clinical perspective.

I would like to thank the people who have worked with me in the Karchin lab, including Dr. Hannah Carter, Dr. David Masica, Andy Wong, Noushin Niknafs, Christopher Douville, Violeta Beleva Guthrie, Dewey Kim, Jean Fan, Grace Yeo, Cheng Wang, Xinyuan Wang and

Dr. Josue Samayoa. Hannah was the first Ph.D. from the Karchin lab and I am the second. She is always motivated and helpful and was a great leader in the lab until she left in October 2012. Hannah and Violeta occasionally held the wine tasting event for the lab, which is an unforgettable memory. Andy and Dewey are great colleagues and friends. We spent a lot of time together after work and shared all news good and bad with each other. Noushin and Chris are smart and diligent. We discussed, shared ideas and solved problems together in the past few years. It was my pleasure to work with them. Grace and Cheng helped me a lot on my second project – personal phenotype prediction. Jean and Xinyuan helped me with coding and software implementation during my first and second year in the Karchin lab. Dave and Josue gave me much great advice based on their Ph.D. research experiences.

I would like to thank Institute of Computational Medicine (ICM) staffs, especially Christina Anger and Joanne Hall, the system administrators, Kyle Reynolds, Paul Tatarsky and Dr. Anping Liu, and Biomedical Engineering program PhD manager, Hong Lan. Hong, Joanne and Chris helped me a lot with administration stuff. Anping was the system administrator of Karchin lab in my first year and occasionally joined the lab meeting offering some opinions. Kyle and Paul were the system administrators in the past four years. They helped me solving all kinds of system issues. It is hard to believe my research could have been done without their support.

I would like to thank those people who have been my (222) roommates: Alan Huang, Lisa Huang, Xindong Song, Yunke Song, Cun-Ting (Amy) Chao, Dan Jiang, Henan Xu, Ai-Cheng (Ariel) Yung, Wen-Ting (Brook) Sung, Wei-Lun (Angela) Chang, Kailun Wan, Xi Chen, Bryce Chiang, Pauline Che and Lin Kuo. They gave me wonderful memories of my Ph.D. life. We cooked and ate dinner together every day. We climbed to the summit of Mt. Whitney, went sea kayaking in Alaska and had fun in Las Vegas. We celebrated every

Chinese festival, discussed research ideas, cleaned the house, fixed the washing machine, played board games and experienced laughs and tears together. I really appreciate every aspect of their support. They greatly enriched my life as a PhD student.

I would like to thank all my friends at Hopkins, especially Nan Li, Ceci Ng, Lingyun Zhao, Tao Yu, Lingje Sang, Alice Ho, Yu-Ja (Kevin) Huang, Wei-Chang Chen, An-Chi Wei, Jackie Li, Jixin Li, Changji Shi and Muzi Na. Their support and friendship made my 6 years in Baltimore wonderful.

I would like to thank my Kong-Fu mentor Liang-Wei Chang in Taiwan. He brought me into the essence of Chinese Kong-Fu, and wisdom from ancient China. Kong-Fu training not only brings me a healthy body but also makes me confident in my heart regardless of the circumstances I confront outside.

Last, but not least, I am indebted to my family. Their full support is always the best cure for the frustration I sometimes face in my research. Although they know very little about the details of my research, they are always excited to hear about any progress I have made during the video chats we share once every two weeks. They are always with me. I really appreciate the confidence they place in me, as always. Finally and most importantly, I want to thank my wife, Pai-Ling Lo, for her support over the past 6 years, day and night.

To my family

Table of contents

Chapter 1 Overview 1
  1.1 Causal gene discovery 1
  1.2 Personal phenotype prediction 2
Chapter 2 Complex disease and sequence-based association studies 3
  2.1 History of association studies 3
    2.1.1 Family-based linkage study 3
    2.1.2 Population-based association study 4
  2.2 Genome-wide association studies 5
    2.2.1 Common disease common variant (CDCV) hypothesis 5
    2.2.2 Genome-wide association study (GWAS) 6
    2.2.3 Missing heritability 7
  2.3 Sequence-based association study 8
    2.3.1 Challenge in statistical analysis 8
    2.3.2 Related work 10
Chapter 3 Burden Or Mutation Position (BOMP) 13
  3.1 A hybrid likelihood ratio test 13
  3.2 Mutation burden statistic 14
    3.2.1 Individual burden 14
    3.2.2 Individual burden thresholds 15
    3.2.3 Aggregated burdens 15
    3.2.4 Burden likelihood ratio statistic 17
  3.3 Mutation position distribution statistic 18
    3.3.1 Window mutation counts 18
    3.3.2 Position distribution likelihood ratio statistic 19
    3.3.3 Windows and sequence segmentation 22
  3.4 Statistical significance 24
  3.5 Extensions to the basic method 25
    3.5.1 Genetic models 25
    3.5.2 Variant scores 25
    3.5.3 Bioinformatics score 25
    3.5.4 Allele-frequency-based scores 26
Chapter 4 BOMP Performance Evaluation 27
  4.1 Simulation framework 27
    4.1.1 Demographic models 28
    4.1.2 Generating genomic populations 30
    4.1.3 Generating phenotypic traits for individuals with a single causal gene 30
    4.1.4 Case-control study generation 33
    4.1.5 Null case-control study 33
    4.1.6 Generalization to multiple genes 34
    4.1.7 Heterogeneity in the simulation 34
  4.2 Power analysis using simulated data 36

4.2.1 Power analysis of simulated case-control studies 38
4.2.2 Relative contributions of mutation burden and mutation positional distribution in simulated case-control studies 50
4.3 Performance summary for each method based on simulations 56
4.4 Dallas Heart Study 59
Chapter 5 Application on bipolar case-control study 66
5.1 Genetic studies in Bipolar disorder 66
5.1.1 Contribution of genetic components in Bipolar disorder 66
5.1.2 Linkage analysis in Bipolar disorder 67
5.1.3 GWA Studies in Bipolar disorder 68
5.1.4 Rare variant search in Bipolar disorder 68
5.2 Bipolar case-control study 69
5.2.1 Sequencing platform, variant calling and quality control 70
5.2.2 BOMP analysis 70
5.2.3 BOMP single gene results 71
5.2.4 BOMP gene set results 78
Chapter 6 BOMP: Future Work and Conclusion 83
6.1 Power increase by removing non-causal variants 83
6.2 Extension to quantitative trait study 83
6.3 Network analysis 85
6.4 Conclusion 86
Chapter 7 Personalized Genome Interpretation 88
7.1 State-of-art personalized genome interpretation 88
7.2 The ultimate goal of personalized genome interpretation – personal phenotype prediction 89
Chapter 8 Probabilistic Model for Personal Phenotype Prediction 90
8.1 Overview of Bayesian network model 90
8.2 Topology of the probabilistic model 94
8.3 Inference of phenotype status 95
8.3.1 Posterior probabilities of random variables S_{VH}, S_{VL}, S_{GH}, S_{GL} 95
8.3.2 Penetrances of random variables (S_{VH}, S_{VL}, S_{GH}, S_{GL}) 97
8.3.3 Penetrance of unknown factors 102
8.3.4 Phenotype specific weights 103
8.4 Penetrance of GWAS hits 105
8.5 Functional impact of variants on phenotype-associated genes 106
8.5.1 Predicted functional impact of variants on gene products 106
8.5.2 Estimating the probability that a gene is functionally altered 107
Chapter 9 Model Performance 109
9.1 Data sources 109
9.1.1 Individual genomes 109
9.1.2 Individual phenotypes 110
9.1.3 Phenotype prevalence 110
9.1.4 Gene and variant annotations 110
9.2 Performance evaluation 111

    - 9.2.1 Evaluation metric 111
    - 9.2.2 Statistical significance 111
  - 9.3 Result 112
    - 9.3.1 Overall performance 112
    - 9.3.2 Contributions from prevalence and genome sequences 116
    - 9.3.3 Contributions from S_{VH}, S_{VL}, S_{GH}, S_{GL} 119
  - 9.4 Critical Assessment for Genome Interpretation (CAGI) 2012-13 122
    - 9.4.1 PGP challenge 122
    - 9.4.2 Matching algorithm 123
    - 9.4.3 Assessment of phenotype-genotype matching algorithms in CAGI 2012-13 124
    - 9.4.4 CAGI 2012-13 PGP result 124
- Chapter 10 Discussion and Future Work for Personal Phenotype Prediction 126
  - 10.1 Strengths of the model 126
  - 10.2 Limitations of the model 129
  - 10.3 Future work 132
- Chapter 11 Concluding Remark 134
- APPENDX A: Glossary of terms 136
- APPENDX B: Supplementary tables 139
- Bibliography 153
- Curriculum Vitae 161

List of figures

Figure 2.1: Changes in the statistical significance of a phenotype-associated SNP with allele frequency (AF) and study size (SS). 9
Figure 3.1: Aggregated burden calculation in BOMP mutation burden statistic. 16
Figure 3.2: Aggregated window mutation count calculation for BOMP mutation position distribution statistic. 21
Figure 3.3: Window and sequence segmentations. 23
Figure 4.1: Demographic models of European-American and African-American populations. 29
Figure 4.2: Example of how our simulations capture genetic heterogeneity in complex phenotype. 35
Figure 4.3: Single gene methods power comparison. 37
Figure 4.4: Distributions of allele frequencies and raw allele counts in simulated European-American and African-American populations. 39
Figure 4.5: Power estimates for multiple gene case-control studies with causal variants equally likely to be from any phenotype etiology dominated by rare variants. 42
Figure 4.6: Nine multinomial distributions used to construct sets of multiple candidate genes for case-control studies. 45
Figure 4.7: Power estimates for multiple genes case-control studies with causal variants from phenotype etiologies randomly sampled from nine multinomial distributions (Figure 4.6). 47
Figure 4.8: BOMP Power estimates for multiple genes (24) case-control studies. 49
Figure 4.9: BOMP burden and position statistics complement each other. 51
Figure 4.10: Example variation pattern in which positional distribution outperforms burden tests. 53
Figure 4.11: Power of position distribution statistics compared to burden methods and SKAT. 55
Figure 8.1: Topology of the model to predict phenotype from an individual's genome sequence. 92
Figure 9.1: Prediction results of the model on 38 dichotomous phenotypes. 114
Figure 9.2: Contribution of population prevalence and genome sequence to prediction results in Fig 9.1. 117
Figure 9.3: Contribution of GWAS hits, low penetrance genes, high penetrance genes, and high penetrance variants to prediction results. 120
Figure 10.1: Prediction results of simple mutation burden model: Six phenotypes predicted with AUC>0.7 are shown. 128
Figure 10.2: Distribution of annotated genes, GWAS hits and high penetrance variants for phenotypes analyzed in this study. 131

List of tables

Table 4.1: Eight phenotypic etiologies used in simulation experiments.................................. 31
Table 4.2: Dallas Heart Study... 62
Table 5.1: Top 30 genes of the BOMP results using disruptive (DIS) variants only. 73
Table 5.2: Top 30 genes of the BOMP results using non-synonymous strict (NSS) variants only. 74
Table 5.3: Top 30 genes of the BOMP result using non-synonymous broad (NSB) variants only. 76
Table 5.4: Top 10 gene sets of the BOMP results using disruptive (DIS) variants only. 80
Table 5.5: Top 10 gene sets of the BOMP results using non-synonymous strict (NSS) variants only.. 81
Table 5.6: Top 10 gene sets of the BOMP results using non-synonymous broad (NSB) variants only.. 82
Table 8.1: Penetrance estimates for blood. 101
Table B.1: GWAS hits correctly identified that PGP-48 has Graves' disease.......................139
Table B.2: GWAS hits correctly identified that PGP-69 has alopecia areata.........................140
Table B.3: GWAS hits correctly identified that PGP-39 has Crohn's disease. 143
Table B.4: GWAS hits correctly identified that PGP-142 and PGP-72 have deep vein thrombosis. 144
Table B.5: GWAS hits predicted that PGP-158 has aortic aneurism. 145
Table B.6: GWAS hits correctly identified that PGP-39 and PGP-38 have chronic obstructive pulmonary disease (COPD). 146
Table B.7: Phenotype model prediction performance (AUC) for 130 PGP participants, using genome sequence only, prevalence only, and both...150
Table B.8: Phenotypes that were predicted no better than random by our models. 152

Chapter 1 Overview

In the following sections, I provide a brief introduction to the methods of genome interpretation that I propose for discovery and prediction of genotype-phenotype relationships. Note that terminologies used in this dissertation are explained in Appendix A Glossary.

1.1 Causal gene discovery

In the past few years, case-control studies, aiming to identify the associations between genomic loci and common diseases, have shifted their focus from single genes to whole exomes. New sequencing technologies now routinely detect hundreds of thousands of sequence variants in a single study, many of which are rare or even novel. The limitation of classical single-marker association analysis for rare variants has been a challenge in such studies. A new generation of statistical methods for case-control association studies has been developed to meet this challenge. A common approach to association analysis of rare variants is the burden-style collapsing methods to combine rare variant data within individuals across or within genes. Here, I propose a new hybrid likelihood model that combines a burden test with a test of the position distribution of variants. In extensive simulations and on empirical data from the Dallas Heart Study, the new model demonstrates consistently good power, in particular when applied to a gene set (e.g., multiple candidate genes with shared biological function or pathway), when rare variants cluster in key functional regions of a gene, and when protective variants are present. When applied to data from an ongoing sequencing study of bipolar disorder (1,135 cases, 1,142 controls) on >12,000 genes, the model identifies the microtubule cytoskeleton gene set and the Golgi apparatus gene set significantly associated with bipolar disorder but is unable to detect any statistically significant genes after correcting for multiple testing.

1.2 Personal phenotype prediction

Genetic screening is becoming possible on an unprecedented scale. However, its utility remains controversial. Although most variant genotypes cannot be easily interpreted, many individuals nevertheless attempt to interpret their genetic information. Initiatives such as the Personal Genome Project (PGP) and Illumina's Understand Your Genome are sequencing thousands of adults, collecting phenotypic information and developing computational pipelines to identify the most important variant genotypes harbored by each individual. These pipelines consider database and allele frequency annotations and bioinformatics classifications. I propose that the next step will be to integrate these different sources of information to estimate the probability that a given individual has specific phenotypes of clinical interest. To this end, a Bayesian probabilistic model has been designed to predict the probability of dichotomous phenotypes. When applied to a cohort from PGP, predictions of Gilbert syndrome, Graves' disease, non-Hodgkin lymphoma, and various blood groups were accurate, as individuals manifesting the phenotype in question exhibited the highest, or among the highest, predicted probabilities. Thirty-eight PGP phenotypes (26%) were predicted with area-under-the-ROC curve (AUC) > 0.7, and 23 (15.8%) of these were statistically significant, based on permutation tests. Moreover, in a Critical Assessment of Genome Interpretation (CAGI) blinded prediction experiment, the models were used to match 77 PGP genomes to phenotypic profiles, generating the most accurate prediction of 16 submissions, according to an independent assessor. Although the models are currently insufficiently accurate for diagnostic utility, I expect their performance to improve with growth of publicly available genomics data and model refinement by domain experts.

Chapter 2 Complex disease and sequence-based association studies

2.1 History of association studies

Compared with Mendelian diseases – whose familial trait patterns are controlled by a single genomic locus – complex diseases (a group that includes bipolar disorder, many cancers and coronary heart disease) tend to be clustered within families but their patterns are not segregated by a single allele. In order to test whether a strong genetic component exists in complex diseases, others have estimated disease heritability – the fraction of variance in a population contributed by genetic components – based on twin and adoption studies. Diseases that show a strong genetic component are studied further in order to identify the causal genetic component. To further research this genetic component, two research designs have been developed: family-based linkage studies and population-based associated studies.

2.1.1 Family-based linkage study

The family-based linkage study was developed based on the co-segregation of marker variants and affected relatives within families. Co-segregated marker variants may be located far from the causal variants found on the same chromosome. The classical linkage analysis assumes that a single major locus (SML) dominates the disease status in a family and the calculation is carried-out using the lod-score method, which is parameterized by disease allele frequency and penetrance. Loci responsible for several complex diseases were successfully identified by linkage analysis, including early-onset familial breast cancer [1] and early-onset Alzheimer's disease [2,3]. One limitation of linkage analysis is that the identified loci only contribute to a small fraction of affected cases in the population. Also, when multiple loci affect a person's disease status within a family, co-segregation of marker variants and disease yields insignificant results.

Another type of linkage analysis, non-parametric linkage analysis, identifies marker variants whose occurrences in affected relatives in families are higher than expected. This method does not assume SML domination, but it usually requires relatively large family size and higher penetrance of the responsible loci to have strong statistical power.

2.1.2 Population-based association study

Population-based association studies identify disease susceptibility loci based on a collection of unrelated individuals in a population. Because of the large sample size, association studies are more effective in identifying multiple loci with modest effect size compared to family-based linkage analyses. A key feature of the human genome which supports this approach is called linkage disequilibrium (LD). LD is helpful in that it shows that the human genome can be divided into many blocks in which variants are likely to co-occur in a population. As a result, if a marker variant is associated with a disease, the actual functional variant is also likely found in the LD block that contains the marker variant. Until recently, there was no high-quality and high-density set of marker variants that could effectively cover LD blocks within the human genome. Due to this limitation, researchers performed population-based association studies on candidate genes pre-selected based on prior knowledge about the disease in question. Marker variants on candidate genes relied on variants discovered in previous studies. Unfortunately, results generated from candidate gene studies were difficult to accurately replicate [4], suggesting that most were false positives. This high false positive rate implied that our prior knowledge about complex diseases used to pre-select candidate genes was limited and ineffective.

To overcome the lack of high-density marker variants on human genomes, the International HapMap project was initiated in 2002 in order to identify common SNPs with minor allele frequencies (MAF) greater than 5% across several continental populations, including Yoruba (YRI), European (CEU), Han Chinese (CHB) and Japanese (JPT). In 2005, the HapMap project verified that the LD feature is found across the entire human genome and reported more than one million unique common SNPs [5]. The reported common SNPs, which cover most of the human genome, were further chosen as tags for LD blocks and the resulting tag SNPs opened the era of genome-wide association (GWA) studies.

2.2 Genome-wide association studies

The discovery of millions of SNPs and verification of LD features in human genomes by the International HapMap project have enabled unbiased association analysis across the whole genome. Further advances in SNP array technology have driven genotyping costs down and made it affordable to perform large-scale association studies. With a large sample size and an unbiased search across the whole genome, genome-wide association (GWA) studies were expected to uncover the genetic origin of complex diseases.

2.2.1 Common disease common variant (CDCV) hypothesis

The anticipated success of GWA studies was underpinned by the common disease common variant (CDCV) hypothesis, which states that a few common allelic variants could account for the genetic variance in complex disease susceptibility and contribute to disease risk additively or multiplicatively with modest effect [6]. With the support of the CDCV hypothesis and because common variants are likely to co-occur in the same LD blocks, common SNPs (the most abundant type of common variants in genomes) would serve as proxies of the disease susceptibility variants nearby. Thus, if the CDCV hypothesis holds, large-scale GWA studies would identify statistically significant tag SNPs whose corresponding LD blocks contain disease susceptibility variants.

2.2.2 Genome-wide association study (GWAS)

A common approach for GWA studies is a case-control study design, where two groups of people are selected based on their disease status – one healthy control group and one case group affected by the disease. Every individual in the study is genotyped for pre-selected SNPs, followed by calculations that identify any SNPs with significant allele frequency differences between the case group and the control group. The effect size of these SNP groups is measured using the odds ratio statistic. Larger allele frequency of a SNP in the case group compared to the control group would yield an odds ratio significantly greater than 1, but the odds ratio would be less than 1 if the opposite were true. A p-value can also be calculated for the odds ratio in order to further quantify the significance. The more an odds ratio deviates from 1, the more significant the p-value will become.

Another common approach of GWA studies is quantitative trait study design, which is designed to find a quantitative measurement of phenotypes for each individual rather than a dichotomous status. In this study design, the effect size of each SNP can be evaluated by other methods, including the beta coefficients of linear regression.

The first successful GWA study was conducted in 2005, comparing 96 cases and 50 healthy controls for age-related macular degeneration [7]. It was then followed by seven large-scale GWA studies (~2000 cases and 3000 shared controls for each disease) conducted by the Wellcome Trust Case Control Consortium – WTCCC [8]. Twenty-three significant SNPs were reported and twenty-two were replicated in independent studies. Many similar successes later on showed the effectiveness of GWA studies. Now, hundreds of diseases and traits have been examined and the results are summarized in a continuously updated online catalog: https://www.genome.gov/gwastudies/ [9].

2.2.3 Missing heritability

Although many successful GWA studies have been conducted in past few years, most of the reported SNPs (GWAS hits) have small effect size and only can explain a small fraction of heritability, the phenotypic variance caused by genetic components [10]. For example, only 5% of phenotypic variance is explained by more than 40 associated loci identified in GWA studies for human height, a classic complex trait with estimated heritability of about 80% [11]. With the ultimate goal of uncovering genetic sources that cause phenotypic variance, finding missing heritability (the fraction of heritability not explained by GWAS hits) is the next emergent issue.

Several potential sources of missing heritability have been proposed [10]: (1) the estimated heritability was inflated; (2) a large number of disease susceptibility common SNPs with very small effect size could not be identified based on the current study size; (3) rare variants or structural variants poorly captured by SNP chips play an important role in phenotypes; and (4) gene-gene interactions, or any interactive effect, among variants not considered in analyses have a significant contribution on phenotypic variation. Despite many proposed explanations, identifying the major source of missing heritability is an ongoing research topic. Fortunately, advances in biotechnologies have accelerated the entire research field in recent years. Since 2005, the development of next-generation sequencing technology has made genome sequencing more time-efficient and affordable for large-scale association studies, allowing for effective examination of the role of rare variants in complex diseases.

2.3 Sequence-based association study

Inexpensive, high-throughput sequencing has transformed the field of case-control association studies. Research efforts over the past few years led to an explosion of exome sequencing studies and exomic variation data (reviewed in [12,13]). One surprising result has been the discovery of hundreds of thousands of novel and rare non-silent variants in protein coding genes, some of which may have functional consequences related to human health. Common diseases, once hypothesized to be primarily due to common variants [6], are now believed to have heterogeneous genetic causes, due to both common and rare variants [14-16].

2.3.1 Challenge in statistical analysis

The challenge of association tests for both common and rare variants comes from two aspects that reduce the test's statistical power: the large number of total SNPs being tested and the low frequency of rare variants. Under the CDCV hypothesis, common variants that are hypothesized to be responsible for disease susceptibility are linked to tag SNPs within LD blocks and the total number of tag SNPs being tested in any study is around one million. A significant hit requires a p-value less than 1E-8 after the Bonferroni correction is applied for multiple testing when a conventional significance level of 0.05 is used. If the CDCV hypothesis is not true, the total number of both common and rare variants considered in a study is expected to be much larger than one million. Multiple testing correction for such a large number of tested SNPs creates a high barrier for reporting any statistically significant hit.

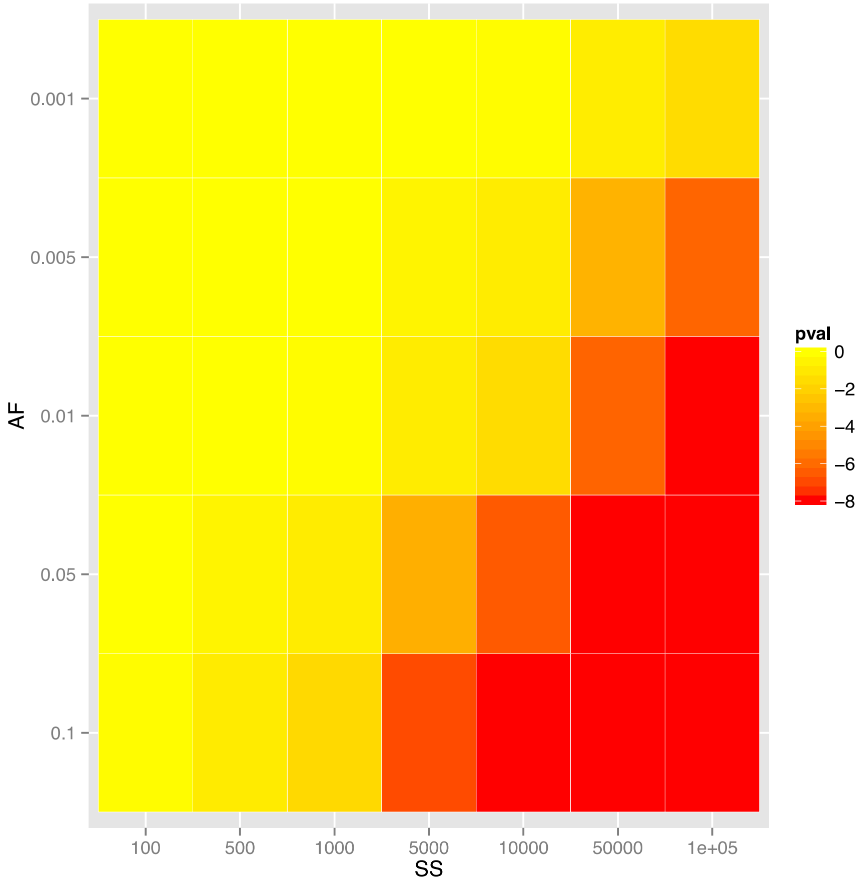


Figure 2.1: Changes in the statistical significance of a phenotype-associated SNP with allele frequency (AF) and study size (SS). The statistical significance of a phenotype-associated SNP is calculated by varying the allele frequency of the SNP and the case-control study size, assuming that the effect size of the SNP is fixed to a relative risk of 1.5 and the chi-square test is used for testing the association.

Compared with common variants, the low frequency of rare variants requires a larger study size to achieve the same statistical significance. Figure 2.1 shows how the p-value varies with study size and allele frequency when testing association for a phenotype-associated SNP. For example, a genome-wide significance level (1E-8) for a SNP with minor allele frequency (MAF) of 10% requires a study size of 10,000, but to reach the same significance level for a SNP with MAF of 1%, a study size of 100,000 is needed (See Figure 2.1).

Regarding the issue of multiple testing corrections, statistical methods exist to test the disease association of a genomic region (ex: a gene), rather than a single SNP, to avoid the large number of tests needed to analyze for rare SNPs. For example, testing association by genes requires 20,000 tests for genome-wide analysis. Moreover, testing a genomic unit gains statistical power by integrating several phenotype-associated SNPs into the same unit. This allows for the same statistical significance level with a smaller study size, compared with the study size needed to test a single rare SNP.

2.3.2 Related work

Increasingly powerful analysis methods exist to detect association between phenotypes and variants with small to moderate effect sizes (reviewed in [14]). Rather than testing each variant individually, variants can be collapsed or summed with a "burden" approach, in which the strength of the phenotypic association is considered with respect to a group of variants occurring at a common region or allelic frequency threshold [17-20]. The contribution of each variant to the association is weighted by frequency or bioinformatically predicted impact [20]. Burden strategies yield a power gain compared to independent tests of single variants but lose power when variants with a neutral or protective effect are included. Regression models [21] and overdispersion tests [22] detect variants that affect phenotype, regardless of the direction of the effect (deleterious or protective). New approaches continue to be introduced, including a mixture model that incorporates gene-gene interactions and an adaptive weighting procedure [23]. Furthermore, a recent study suggests that single-variant test statistics may be more powerful than collapsing strategies on real data [24]. Importantly, no single method appears to be best for all phenotypes, genomic regions, disease models and populations [14,25,26].

In this dissertation, I propose a new method of detecting the association between phenotypes and variants within a genomic region, and compare its performance with three existing methods: VT, SKAT and KBAC. VT (Variable Threshold) is a burden test. It selects an optimal minor allele frequency (MAF) threshold, by trying many thresholds and identifying the one that maximizes the z-score difference between cases and controls. To attempt to remove neutral or protective variants from analysis with more focus on rare and potentially deleterious variants, all variants whose MAF exceed the threshold are filtered out and not considered for analysis. The remaining variants are summed in cases and controls and a z-score statistic is used to quantify the difference between the two groups. P-values are estimated by permutation [20].

SKAT (Sequence Kernel Association Test) uses a regression model (logistic regression for dichotomous traits and linear regression for quantitative traits) where phenotypes are the response and genotypes are the predictors. A variance-component score statistic Q is computed by comparing the full model with a null model, using a kernel function that incorporates allele frequency weights. The statistic follows a mixture of χ^2 distributions, enabling analytical computation of P-values. The regression model enables SKAT to incorporate covariates such as weight, age and gender into analysis and, unlike a burden test, the variance-component score statistic Q examines the variance of each genotype across cases and controls, allowing SKAT to identify both deleterious and protective effects. The default weighting scheme in its kernel function up-weighs rare variants in analysis [21].

The Kernel Based Adaptive Cluster (KBAC) was developed to overcome the problem of detecting rare variant associations in the presence of misclassification and interaction. A vector of genotypes within a genomic region of interests is modeled for each sample using a mixture distribution. The calculation adaptively up-weighs the vectors with genotype patterns that are more frequently in cases than controls. Distributions of genotype vector counts are compared between cases and controls to evaluate the phenotype association of the genomic region. The statistical significance of the KBAC can be assessed using either permutation or a Monte Carlo approximation. Uniquely, considering multiple genotypes enables KBAC identifying interactions among variants [23].

Chapter 3 Burden Or Mutation Position (BOMP)

3.1 A hybrid likelihood ratio test

Here I describe a new hybrid likelihood test BOMP (Burden Or Mutation Position test), designed for case-control exome sequencing studies, to detect the presence of causal variants in a functional group. The functional group can be defined as a gene, genomic region, or gene set (multiple genes involved in a pathway or biological process). The test can incorporate variant weighting by bioinformatically-predicted functional impact. I combine, into a single statistic, a directional burden test in which low frequency variants have increased weight and a non-directional positional distribution test that does not consider allele frequency. My burden test uses a collapsing strategy and metrics of variant functional importance, which are similar to previously published burden tests. An advantage of the test is that its formulation into a likelihood ratio uniquely allows us to combine it with the positional distribution test. The two tests complement each other and together yield increased power to detect biologically important variants, particularly when applied to a gene set containing genes with different kinds of variants (e.g., rare, low frequency, common, protective).

The hybrid likelihood model consists of two likelihood ratio tests (mutation burden and mutation position distribution statistics) with the same general form,

$$\Lambda_j = log\left(\frac{L(FG_j|H_A)}{L(FG_j|H_0)}\right)$$

Eq 3.1

and tests the evidence for the alternative hypothesis H_A that a functional group (FG) of interest is associated with a dichotomous phenotype, compared to the null hypothesis H_0

that they are not associated. Higher values of Λ_j indicate stronger association between unit *j* and the phenotype. In this work, the functional groups of interest are either single genes or sets of multiple genes.

3.2 Mutation burden statistic

The first likelihood ratio test is based on comparing mutation burden in cases and controls. Each individual is represented by a Bernoulli random variable, which is 1 if the individual's burden exceeds a burden threshold, and 0 otherwise. To model the likelihood, I assume that individual burden status is independent and identically distributed (IID). The ratio compares an alternative hypothesis that the probability of exceeding the burden threshold is higher in cases than in controls and the null hypothesis (that probabilities are equal or lower in cases than in controls). Biologically, the IID assumption is not necessarily true. I control for such violations by assessing the statistical significance of the likelihood ratio by permuting case and control labels.

3.2.1 Individual burden

For individual *k* the gene burden of g_j is

$$S_{g_j,k} = \sum_{i=1}^{n_{j,k}} x_{i,k} \qquad \text{Eq 3.2}$$

where $n_{j,k}$ is the number of variants carried by individual *k* in gene *j* and $x_{i,k}$ is the genotype of variant v_i (0, 1 and 2 representing homozygous reference allele, heterozygous allele and homozygous alternative allele respectively).

3.2.2 Individual burden thresholds

A binary variable is used to label individuals whose mutation burden in a gene of interest exceeds a critical threshold. If the burden of gene g_j in individual k is greater than or equal to the threshold t, then it is considered to be phenotype-associated for that individual and $Y_{g_j,k} = 1$ (0 otherwise). Because genes are heterogeneous in size, functional importance, mutation rate, and tolerance to variation, each gene may have a different value of t. For each gene j, this cutoff t_j is computed by iterating over all cut-offs and selecting the one that maximizes its burden statistic (Equation 3.4).

3.2.3 Aggregated burdens

The $Y_{g_j,k}$ values are then aggregated by summing over cases and controls:

$$T_j^A = \sum_{k \in cases} Y_{g_j,k}$$

$$T_j^U = \sum_{k \in controls} Y_{g_j,k}$$

The maximum likelihood estimate of the probability that the mutation burden of gene j exceeds the threshold in cases is then $\hat{p}_j^A = \frac{1}{m+2}(T_j^A + 1)$, the estimate for controls is $\hat{p}_j^U = \frac{1}{l+2}(T_j^U + 1)$ and the estimate for both cases and controls is $\hat{p}_j = \frac{1}{m+l+4}(T_j^A + T_j^U + 2)$, where m is the number of cases and l the number of controls. The probability estimates $\hat{p}_j^A$, $\hat{p}_j^U$ and $\hat{p}_j$ are used as the parameters of three Bernoulli distributions (one for cases, one for controls, and one for cases and controls together). Pseudocounts are added to avoid zero counts. The aggregated burden calculation (without pseudocounts) is illustrated in Figure 3.1.

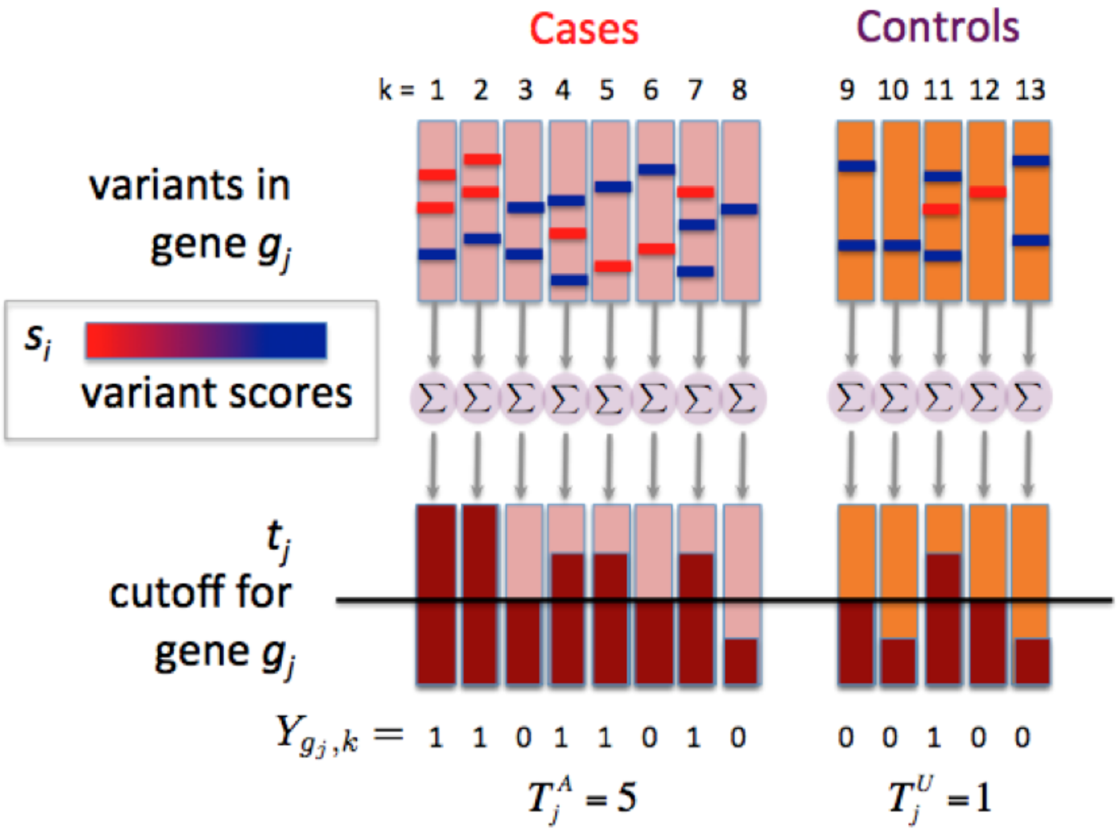


Figure 3.1: Aggregated burden calculation in BOMP mutation burden statistic. Vertical bars are samples (8 cases and 5 controls). Horizontal bars on each sample are variants colored with weights. The individual burden is the weighted sum of variants calculated for each sample. The indicator variable $Y_{gj,k}$ is set depending on whether the individual burden exceeds the individual burden threshold. The mutation burden statistic uses the aggregated burden for cases, T_j^A, and controls T_j^U, which are the sums of indicator variables across cases and controls respectively.

3.2.4 Burden likelihood ratio statistic

For a gene g_j the mutation burden statistic is defined as a ratio of Bernoulli likelihoods:

$$\Lambda_B(g_j) = log\left(\frac{L_B^A \times L_B^U}{L_B^{A+U}}\right) \qquad \text{Eq 3.3}$$

$$= \log\left(\frac{(\hat{p}_j^A)^{T_j^A}(1-\hat{p}_j^A)^{(m-T_j^A)}(\hat{p}_j^U)^{T_j^U}(1-\hat{p}_j^U)^{(l-T_j^U)}}{(\hat{p}_j)^{(T_j^A+T_j^U)}(1-\hat{p}_j)^{(m+l-T_j^A-T_j^U)}}\right) \qquad \text{Eq 3.4}$$

where *m* is the number of cases, *l* the number of controls; T_j^A is the number of cases whose mutation burden in g_j exceeds an optimized threshold (3.2.2 Individual Burden Thresholds); T_j^U is the number of controls exceeds the threshold; $\hat{p}_j^A$ is the maximum likelihood estimate of the probability that the burden of gene *j* exceeds the threshold in cases; $\hat{p}_j^U$ is the estimate that gene *j* exceeds the threshold in controls, $\hat{p}_j$ is the estimate that gene *j* exceeds the threshold in both cases and controls. First I consider only genes with higher burden in the cases, for which $\hat{p}_j^A \geq \hat{p}_j^U$. Next, for the remaining genes, for which $\hat{p}_j^A < \hat{p}_j^U$, Equation 3.4 is modified,

$$= \log\left(\frac{(\hat{p}_j^U)^{T_j^A}(1-\hat{p}_j^U)^{(m-T_j^A)}(\hat{p}_j^A)^{T_j^U}(1-\hat{p}_j^A)^{(l-T_j^U)}}{(\hat{p}_j)^{(T_j^A+T_j^U)}(1-\hat{p}_j)^{(m+l-T_j^A-T_j^U)}}\right) \qquad \text{Eq 3.5}$$

The modification in Equation 3.5 follows the burden hypothesis that the mutation burden in cases is higher than that in controls. Formally, on average, under H_A, the number of cases in which the burden of gene *j* exceeds the threshold will be larger than that in controls, and otherwise under H_0. Thus genes with higher burdens in cases than controls (calculated using Equation 3.4) get a high value and those with higher burdens in controls than cases (calculated using Equation 3.5) get a low value due to the violation of the burden hypothesis.

If a gene set rather than a single gene is used as the functional group, the burden is aggregated across all genes in the set, and the procedure is otherwise identical.

3.3 Mutation position distribution statistic

The second likelihood ratio test is based on comparing the positional distribution of mutations in cases and controls. The codons of a gene are partitioned into windows and mutation count (burden score) is computed for each window in cases only, controls only, and in cases and controls together. To model the likelihood, each window mutation count is considered to be a random variable in a multinomial distribution. If the partition contains *d* windows, there are *d* possible outcomes for each mutation. There are also *d* multinomial parameters for the partition.

3.3.1 Window mutation counts

Let the window mutation counts in the multinomial distributions be $W_{x,j}^{A}$, $W_{x,j}^{U}$ and $W_{x,j}^{A+U}$ (cases only, controls only and in cases and in controls together)

for cases $W_{x,j}^{A} = \sum_{k \in cases} S_{x,j,k}$,

for controls $W_{x,j}^{U} = \sum_{k \in controls} S_{x,j,k}$, for cases and controls $W_{x,j}^{A+U} = \sum_{k \in cases\ and\ controls} S_{x,j,k}$.

where $S_{x,j,k}$ (computed as in Equation 3.2) is the score for individual k in window x.

The maximum likelihood estimate of the multinomial parameters (including pseudocounts) is then

$$\hat{p}_{x,j}^{A} = \frac{W_{x,j}^{A}}{\sum_x \left(W_{x,j}^{A} + 1\right)}$$

Eq 3.6

$$\hat{p}_{x,j}^{U} = \frac{W_{x,j}^{U}}{\sum_x \left(W_{x,j}^{U} + 1\right)}$$

Eq 3.7

$$\hat{p}_{x,j}^{A+U} = \frac{W_{x,j}^{A+U}}{\sum_x \left(W_{x,j}^{A+U} + 2\right)}$$

Eq 3.8

3.3.2 Position distribution likelihood ratio statistic

For a gene g_j, the statistic is defined as a ratio of multinomial likelihoods:

$$\Lambda_P(g_j) = log\left(\frac{L_P^A \times L_P^U}{L_P^{A+U}}\right)$$

Eq 3.9

where

for cases $\mathrm{L}_P^A = \prod_x \left(\hat{p}_{x,j}^{A}\right)^{W_{x,j}^{A}+1}$

Eq 3.10

for controls $\mathrm{L}_P^U = \prod_x \left(\hat{p}_{x,j}^U\right)^{W_{x,j}^U+1}$

Eq 3.11

for cases and controls $\mathrm{L}_P^{A+U} = \prod_x \left(\hat{p}_{x,j}\right)^{W_{x,j}^{A+U}+2}$

Eq 3.12

It follows that under H_A, the likelihood for cases will be different than for controls and that under H_0, they are not different. In contrast to the mutation burden statistic, there is no directionality in the mutation position distribution statistic, because $\Lambda_P(g_j)$ will be large when either $\hat{p}_{x,j}^A$ or $\hat{p}_{x,j}^U$ is large.

A toy example of aggregated window mutation count calculation is illustrated in Figure 3.2.

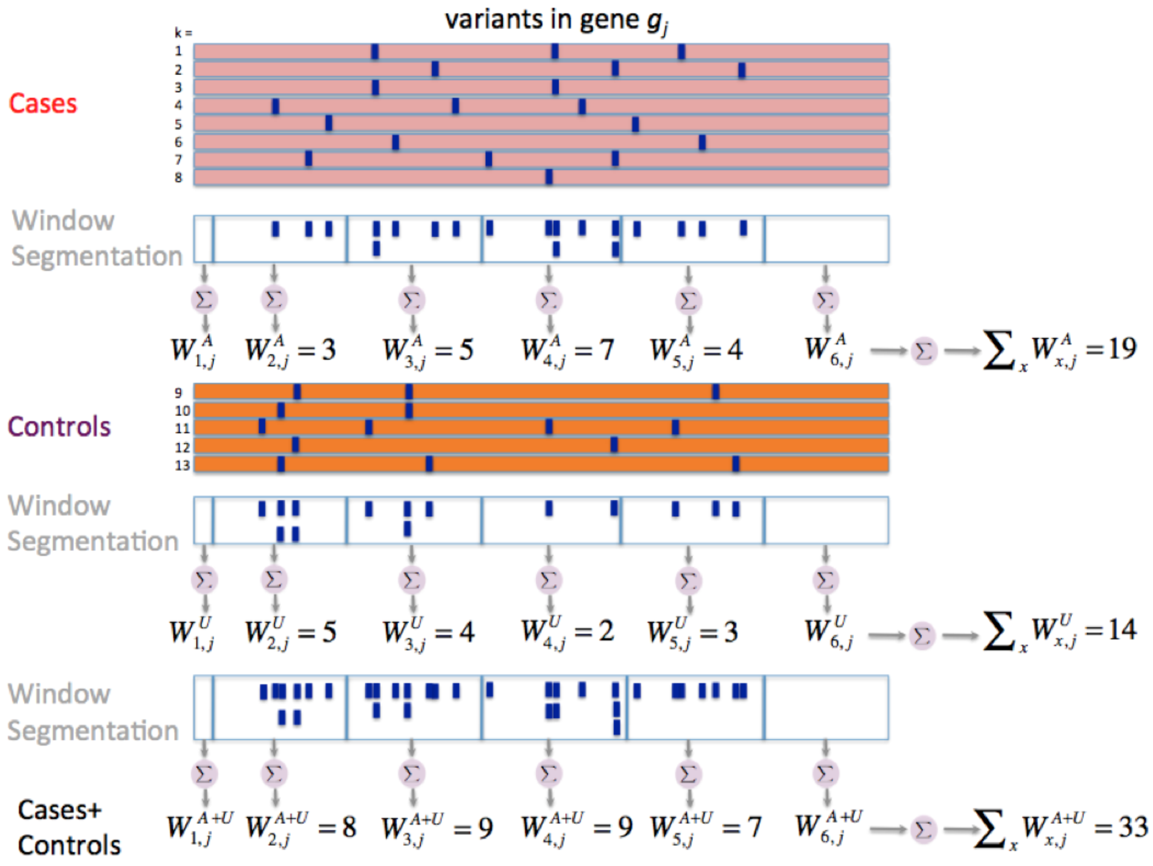


Figure 3.2: Aggregated window mutation count calculation for BOMP mutation position distribution statistic. Aggregated window mutation counts are calculated for cases, W^A, controls, W^U, and cases and controls combined, W^{A+U}, across x windows for gene j.

3.3.3 Windows and sequence segmentation

Each gene has many possible window partitions, and I don't know in advance which is the most informative for the position distribution statistic. One way to create candidate window partitions (*i.e.*, sequence segmentation) for a gene of length L is to select a window size s and a series of possible offsets, based on a selected shift increment *t*. Each offset generates a new segmentation (Figure 3.3). For example, if the window size is 8 and the shift increment is 1, the first offset begins at the first position of the gene and generates a segmentation of $\left\lceil \frac{L}{8} \right\rceil$ windows. The second offset will begin at position 2 of the gene and generate a new segmentation of $\left\lceil \frac{L-1}{8} \right\rceil + 1$ windows, *etc*. In this work I used four combinations of window size *s* and shift increment *t*: (8,1), (16,2), (32,4) and (64,8), yielding 32 candidate segmentations for a gene. These choices were not optimized and can be adjusted, according to user preference and/or prior knowledge. The best segmentation is selected by computing the likelihood ratio $\Lambda_P(g_j)$ (Equation 3.9) for each segmentation and picking the segmentation with the largest $\Lambda_P(g_j)$. Alternatively, this likelihood ratio can be modified by computing $W_{x,j}^{A}$, $W_{x,j}^{U}$ and $W_{x,j}^{A+U}$ with respect to total positions mutated, rather than total number of mutations.

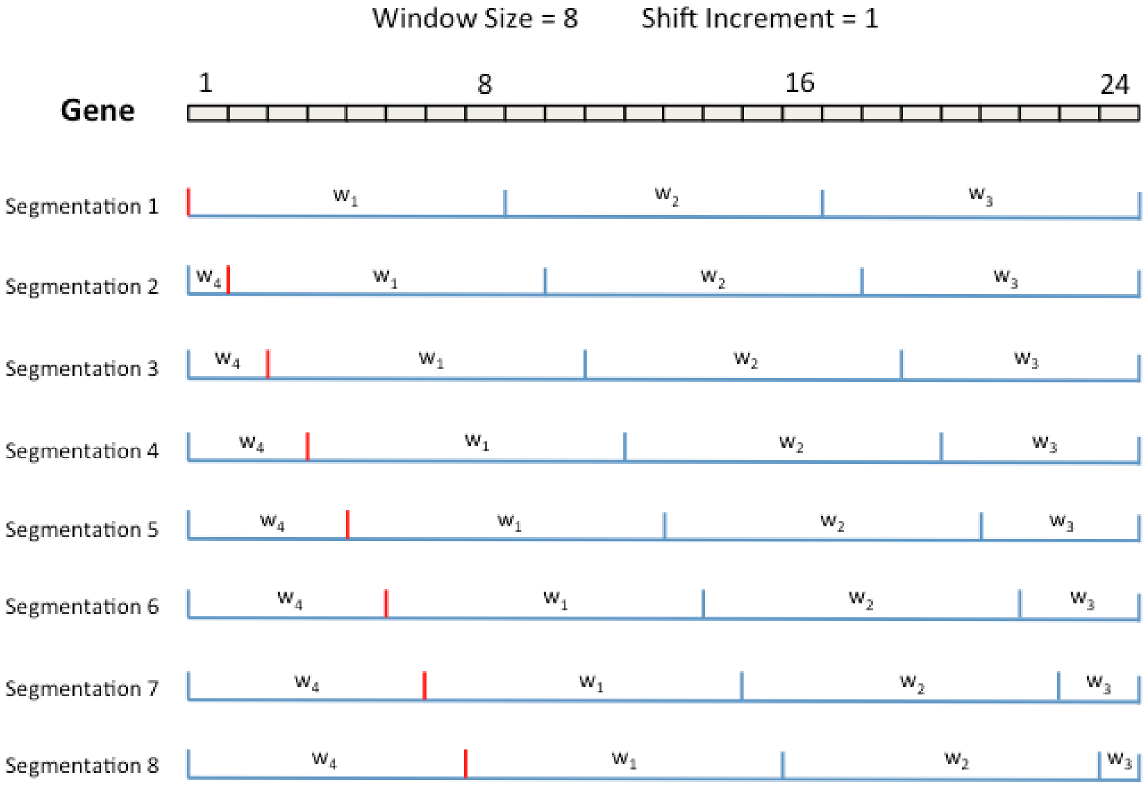


Figure 3.3: Window and sequence segmentations. The mutation position distribution statistic requires a segmentation for a sequence of interest (e.g., a gene). I generate candidate segmentations by selecting a window size s and allowing a series of possible offsets, based on a selected shift increment t. In this example, I illustrate the eight possible window segmentations of a gene with 24 codons (represented by rectangles), using a window size of 8 and a shift increment of 1.

For the position distribution statistic, if a gene set rather than a single gene is used as a functional group, the best window segmentation is computed for each gene, and the calculation of the position distribution statistic is otherwise identical.

The mutation burden and mutation position burden statistics are combined into a single log likelihood ratio,

$$\Lambda_{g_j} = \Lambda_B(g_j) + \Lambda_P(g_j)$$

Eq 3.13

3.4 Statistical significance

P-values for each Λ_{g_j} are computed with a null distribution, generated by repeated permutation of case and control labels. All parameters of $\Lambda_B(g_j)$ and $\Lambda_P(g_j)$, including the maximum likelihood burden threshold and segmentation pattern, are calculated initially for empirical data and then re-calculated for each iteration of the permutation. Thus, *N* iterations yield *N* null $\Lambda_{g_j}^{(n)}$ where *n* ranges from 1 to *N* (see also 5.1.4 computation complexity). The permutation controls for confounding effects, such as properties that characterize a particular gene or region of interest (*i.e.*, nucleotide diversity, GC content and recombination rate), which are the same when used to estimate Λ_{g_j} and each null $\Lambda_{g_j}^{(n)}$.

After *N* iterations (*e.g.*, $N = 10^6$)

$$\mathrm{P}_{g_j} = \frac{\#\left(null\ \Lambda_{g_j}^{(n)} \geq \Lambda_{g_j}\right) + 1}{N + 1}$$

Eq 3.14

While $\Lambda_B(g_j)$ and $\Lambda_P(g_j)$ are not independent, using the permutation test yields an accurate P-value estimate, because any dependencies in Λ_{g_j} are reproduced in each null $\Lambda_{g_j}^{(n)}$.

3.5 Extensions to the basic method

3.5.1 Genetic models

Either *dominant* or *additive* genetic models can be specified. Under the dominant genetic model, both homozygous and heterozygous variants have $x_i = 1$; under the additive model, homozygous variants have $x_i = 2$ and heterozygous variants $x_i = 1$. For all experiments in this work, additive models were used.

3.5.2 Variant scores

The individual burden (see 3.2.1 Individual burden) can be modified by incorporation of score coefficients so that $S_{g_j,k} = \sum_{i=1}^{n_{j,k}} x_{i,k}\, s_i$. The score for a variant v_i can be either a bioinformatics-based score $s_i = f_i$, an allele-frequency-based score $s_i = \frac{1}{\sqrt{\theta_i(1-\theta_i)}}$, following [18,20]; or the product of both $s_i = \frac{f_i}{\sqrt{\theta_i(1-\theta_i)}}$. Next I explain how these scores are calculated. In this work, variant scores were used only in the burden statistic. The allele-frequency-based score was used for all simulations and the product of allele-frequency and bioinformatics scores was used on all empirical data.

3.5.3 Bioinformatics score

Each nonsilent variant v_i can be assigned a score $f_i \in [0,1]$ to represent its contribution to a disease phenotype of interest, where $f_i = 0$ indicates no contribution and $f_i = 1$ indicates a strong contribution. These scores are estimated with *Variant Effect Scoring Tool* (VEST) [27].

Variants causing nonsense, nonstop or frameshift alterations to a gene's protein product receive $f_i = 1$. Variants causing missense alterations are scored with a Random Forest classifier [28,29]; the score is the fraction of decision trees in the forest that classified the variant as deleterious. Alternatively, other bioinformatics methods that score missense variants can be used to generate f_i values, if scaled to range from 0 to 1.

The Variant Effect Scoring Tool is a Random Forest classifier, trained with the CHASM software suite's Classifier Pack and SNVBox [30]. The Forest contains 1000 decision trees. The positive class of 45,000+ missense variants is taken from the Human Gene Mutation Database (HGMD) [31]. The negative class of variants is taken from a set of common variants (MAF>0.01 validated by the 1000genomes project [32], compiled in the SNP135 table of the UCSC Genome Browser database[33]. Each missense variant is represented by 86 features in SNVBox, including conservation scores, amino acid residue substitution scores, UniProtKB annotations [34], and predicted local protein structure [35].

3.5.4 Allele-frequency-based scores

For each variant v_i, I estimate its mean population allele frequency θ_i as follows:

$$\mathrm{E}[\theta_i] = \frac{x_i^A + x_i^U + 1}{2N + 2}$$

Eq 3.15

where x_i^A and x_i^A are allele counts of variant v_i in cases and controls, respectively; N is the number of individuals in both cases and controls; and the constants are pseudocounts from a beta prior.

Chapter 4 BOMP Performance Evaluation

I evaluated the power of the BOMP hybrid likelihood model with both simulations and empirical data from the Dallas Heart Study [36]. All results were compared with several leading statistical methods to detect causal variation in case-control association studies. I attempted to select representative methods for burden (VT), regression (SKAT), and mixture modeling approaches (KBAC). The permutation test was used to compute p-values. Briefly, the null statistic was generated by repeating the whole calculation described in Chapter 3 with randomly permuted case-control labels, while holding fixed the choices of the individual burden threshold and the window segmentation, used on the original data. The generation of the null statistic only assumed the exchangeability among samples, which holds true because it is a population-based case-control study. Thus, a permutation test yields a correct null distribution and type I error should be well controlled. Section 4.1 introduces how the simulated data were generated; 4.2 shows the statistical power analysis using simulated data; and 4.3 describes the experimental results for the empirical benchmark set, the Dallas Heart Study.

4.1 Simulation framework

Simulated case-control studies were generated using two demographic growth models, eight disease etiologies, and a stochastic model of genotype-phenotype association. The true disease etiology for most complex diseases is still unknown and many factors, such as population structure and causal allele frequencies, can potentially affect the performance of association tests. Varying the parameter combination allowed me not only to simulate the complex diseases as realistically as possible but also to provide a fair and comprehensive comparison among different methods.

4.1.1 Demographic models

Kryukov et al. [37] used sequence data for 58 genes from 757 European-American individuals to fit parameters of a demographic model, with Wright-Fisher diffusion approximation, assuming long-term constant size succeeded by a bottleneck and then exponential growth (Figure 4.1). The best fitting distribution of fitness effects/selection coefficients (DFE) was a two-component mixture of gamma distributions:

$$S \sim 0.2*\mathrm{Gamma}(1, 106) + 0.8*\mathrm{Gamma}(0.56231, 0.01) \tag*{Eq 4.1}$$

Mixture parameters and software to estimate the mixture were generously provided by G. Kryukov of the Broad Institute (Boston, MA USA). However, because his study focused on the simulation of variants that are deleterious to the phenotype, protective variants were not modeled in the mixture of gamma distributions.

Boyko et al. [38] used genome-wide polymorphism data and fixed differences between the human and chimp genomes, to estimate demography and DFE for 19 African American samples, with Wright-Fisher and Poisson Random Field theory [39-41]. The best-fitting demographic model was a two-epoch instantaneous growth model for the African Americans (Figure 4.1). A best fitting DFE model, by maximum likelihood estimate, had three parameters: proportion of positively selected sites (1.86%); shape (0.228) and rate (16.54) of a gamma distribution for sites with deleterious fitness effects. Positive selection was fixed at $s = 9.7 \times 10-5$.

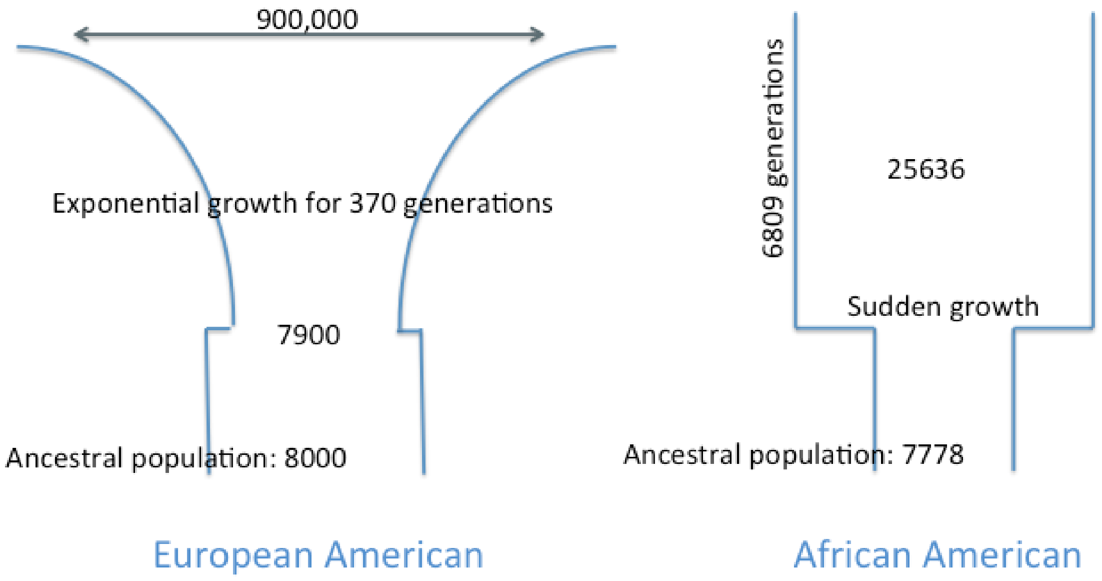

900,000
Exponential growth for 370 generations
7900
Ancestral population: 8000
European American
6809 generations
25636
Sudden growth
Ancestral population: 7778
African American

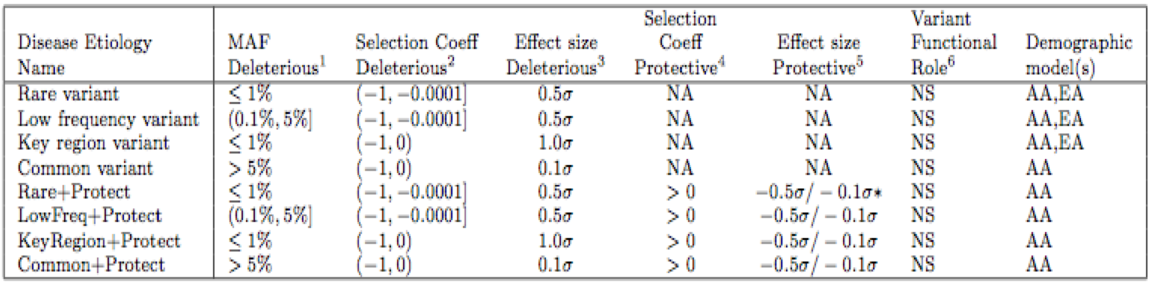
Figure 4.1: Demographic models of European-American and African-American populations. The models were fit to European-American [37] and African-American sequencing data [38].

4.1.2 Generating genomic populations

The general Wright-Fisher model/forward population genetic simulation tool SFS code [42] was used to generate 100 effective genomic populations and to sample 1,000,000 haplotypes per population. As in [23], the simulated haplotypes in each population were randomly paired to generate 500,000 diploid individuals. Two demographic growth structures were used: an exponential growth model fitted to deep resequencing data from European Americans [37] and a simple bottleneck model fitted to whole-genome polymorphism data from African-Americans [38] (4.1.1 Demographic models). For the exponential growth demographic model, DFE was modeled with a two-component gamma mixture (similar to [37]). For the simple bottleneck model, DFE was modeled as described in [38]. Following [37], the mutation rate was set to 1.8 × 1E-8 per generation for all simulations.

4.1.3 Generating phenotypic traits for individuals with a single causal gene

The individuals in a population were then associated with a quantitative phenotypic trait, which I assumed to drive a disease or other dichotomous phenotype, so that individuals with high values of the trait would have the phenotype and those with low values of the trait would not. Eight possible phenotypic etiologies were considered (Table 4.1). Each etiology was defined by properties of its causal variants. Variants could be rare, low frequency, or common. They could occur only in key functional regions. They could have small or large effects (value of k in Equation 4.2). Protective modifier variants might or might not be present. In this work, only coding, non-synonymous variants were considered as causal or protective for all etiologies. (However, etiologies that consider silent variants, which impact gene regulation, could also be defined.) For etiologies with key region variants, I used haplotypes that contained multiple coding segments (100 segments, each 30 bases long). Otherwise haplotypes contained a single coding segment of 1500 bases.



Disease Etiology Name	MAF Deleterious[1]	Selection Coeff Deleterious[2]	Effect size Deleterious[3]	Selection Coeff Protective[4]	Effect size Protective[5]	Variant Functional Role[6]	Demographic model(s)
Rare variant	$\leq 1\%$	$(-1, -0.0001]$	0.5σ	NA	NA	NS	AA,EA
Low frequency variant	$(0.1\%, 5\%]$	$(-1, -0.0001]$	0.5σ	NA	NA	NS	AA,EA
Key region variant	$\leq 1\%$	$(-1, 0)$	1.0σ	NA	NA	NS	AA,EA
Common variant	$> 5\%$	$(-1, 0)$	0.1σ	NA	NA	NS	AA
Rare+Protect	$\leq 1\%$	$(-1, -0.0001]$	0.5σ	> 0	$-0.5\sigma/-0.1\sigma*$	NS	AA
LowFreq+Protect	$(0.1\%, 5\%]$	$(-1, -0.0001]$	0.5σ	> 0	$-0.5\sigma/-0.1\sigma$	NS	AA
KeyRegion+Protect	$\leq 1\%$	$(-1, 0)$	1.0σ	> 0	$-0.5\sigma/-0.1\sigma$	NS	AA
Common+Protect	$> 5\%$	$(-1, 0)$	0.1σ	> 0	$-0.5\sigma/-0.1\sigma$	NS	AA

Table 4.1: Eight phenotypic etiologies used in simulation experiments. *Rare variant*=phenotype caused by multiple rare deleterious variants. *Low frequency variant*=phenotype caused by multiple low frequency deleterious variants. *Key Region variant*=rare deleterious variants are localized to key regions. *Common variant*=phenotype caused by a single deleterious common variant. The etiologies *Rare+Protect*, *LowFreq+Protect, KeyRegion+Protect* and *Common+Protect* were identical to the first four except that they include protective variants. [1]Minor allele frequency of deleterious causal variants, [2]Selection coefficients of deleterious causal variants, [3]Effect size of deleterious causal variants, [4]Selection coefficient of protective causal variants, [5]Effect size of protective modifier variants, [6]Required functional role of causal and protective variants, NS=coding non-synonymous, AA=African-American simple bottleneck demographic model [38], EA=European-American exponential growth demographic model [37]). $*-0.5\sigma$ for protective modifier variants with $AF<5\%$,-0.1σ for protective modifier variants with $AF>5\%$.

To generate the phenotypic traits for the genomic populations, I selected a phenotype etiology and a population that contains variants meeting the criteria for causality in that etiology (Table 4.1). Next the quantitative phenotype trait QT was generated for each individual in the population, using an approach based on [37]. Trait values were drawn from Gaussian distributions, such that individuals with no causal variants had

$$\mathrm{QT} \sim \mathrm{N}(\mu, \sigma), (\mu, \sigma) = (0,1)$$

Eq 4.2

Individuals with n causal variants had

$$\mathrm{QT} \sim \mathrm{N}\left(\mu + \sum_{i=1}^{n} k_i \sigma\,, \sigma\right), (\mu, \sigma) = (0,1)$$

Eq 4.3

where $\sum_{i=1}^{n} k_i \sigma$ is the mean shift in trait value (the shift per variant) for an individual and k_i is the effect size of causal variant *i.* To match the expected effect size of significant common and rare variants in GWAS, effect sizes of 0.1σ for common variants and a range of $0.5\sigma - 1.0\sigma$ for rare variants were used.

Most significant common variants in GWAS have odds ratios < 2.0 [15]. For case-control simulations, an odds ratio was analytically equated with mean shift in a quantitative phenotypic trait, given the assumed phenotype prevalence of 1%. Effect size in phenotype etiology “common” (Table 4.1) was 0.1σ, which can be shown equivalent to an odds ratio of 1.41 assuming the MAF of 10% for the common variant.

In the simulations, the strongest effects were 1.0σ for rare variants in Key Regions. I assumed that these variants occur at functionally important positions (Key Regions) in a gene of interest, and that they were unlikely to occur more than once in a single individual because very few variants were located in Key Regions and each of them was rare in the population. This choice of effect size is somewhat larger than that used by Kryukov et al. [37], who used a range of $0.25\sigma - 0.5\sigma$ for rare variants, since these variants were not only rare but also located in Key Regions assumed to be functionally important to the gene function. To account for heterogeneity within a particular phenotype etiology, effect size was allowed to be 0 for a designated fraction of causal variants.

4.1.4 Case-control study generation

At this stage, each population consists of a set of "genomic individuals", each with a real-valued quantitative trait. To construct the case-control studies, an extreme phenotype model was used. Phenotype prevalence was set at 1%, *i.e.*, the 1% of individuals in the selected population with the highest values of QT were considered *affected* and the 25% of individuals with the lowest values of QT *unaffected*.

Case-control studies were generated by sampling without replacement from affected and unaffected groups in a population. Individuals with intermediate phenotype values were not included in case-control studies. The random process used to generate QT (Phenotypic trait generation) ensures varied penetrance and phenocopy rates in each case-control study, *e.g.*, some individuals carrying deleterious variants were not affected, while some with no deleterious variants were affected.

4.1.5 Null case-control study

A null case-control sample was also generated, with no phenotype etiology, in which the phenotypic trait was drawn from a standard normal distribution for every individual in the sample.

4.1.6 Generalization to multiple genes

For a scenario in which the functional group of interest was a gene set *e.g.*, involved in a pathway or biological process, I constructed a new population of 500,000 individuals, in which each individual had multiple genes. This population was created by sampling genes from the diploid gene populations generated previously (4.1.2 Generating genomic populations). Next, I specified a gene set size and fraction of genes in the gene set that contain causal variants. A phenotype etiology was then randomly selected for each gene that contains causal variants. Finally, the phenotypic trait for each individual in the population was generated using Equations 4.2 and 4.3. Gene set case-control studies were generated with the same protocol as for single causal genes.

4.1.7 Heterogeneity in the simulation

Complex phenotypes are expected to have considerable genetic heterogeneity i.e., they may be the consequence of alterations in hundreds of potentially causal genes, and affected individuals may have causal and/or protective variants in different subsets of these genes. The simulations done in this work reflected this heterogeneity (example in Figure 4.2).

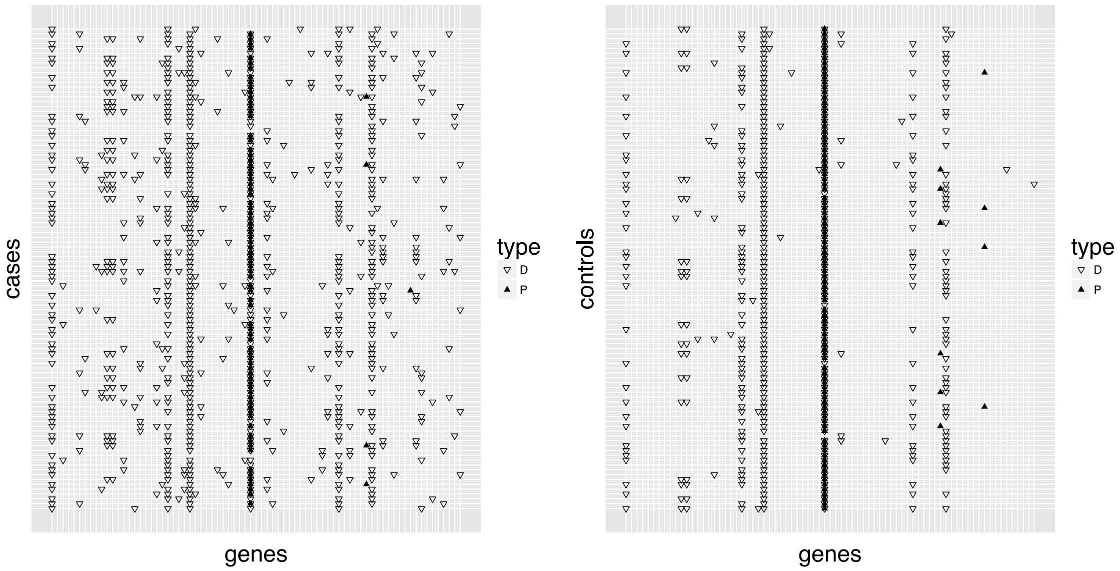


Figure 4.2: Example of how our simulations capture genetic heterogeneity in complex phenotype. Each horizontal grid line represents a genomic individual. (Cases and controls shown separately.) Each vertical gridline represents a gene. Causal variants (both deleterious and protective) are shown as triangles. Different case individuals have different patterns of causal variants and the allele frequencies of the variants range from rare (1 allele) to common (190 alleles). Causal variants are also observed in the control individuals. Type=Downward pointing triangles are deleterious variants, upward pointing triangles are protective variants. (200 genomic individuals from African-American demographic model are shown).

4.2 Power analysis using simulated data

First, I assessed the power of BOMP to detect genes with causal variants in an extreme phenotype case-control study, for a phenotype with 1% population prevalence, and significance level $\alpha = 0.05$. I considered that deleterious causal variants might either be rare, low frequency or common and that modifying protective variants might be present. Power to detect causal variants was assessed initially with respect to a single candidate gene and then for candidate gene sets, ranging in size from 2 to 24 genes. I studied gene sets in which all genes contained causal variants and those in which only a fraction of genes contained causal variants. Both African-American and European-American demographic models were considered. For each combination of attributes (phenotype etiology, population demographic, case-control study size), 250 case-control studies were simulated to assess power.

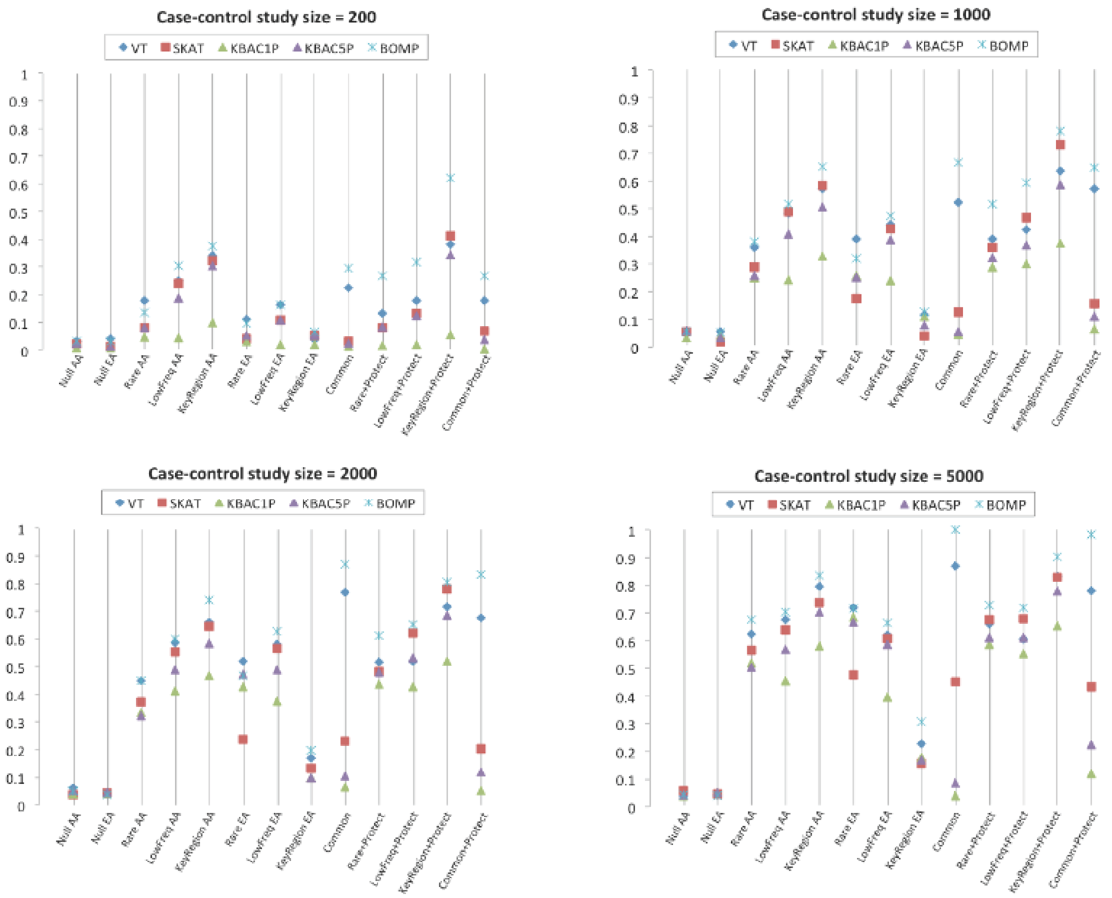


Figure 4.3: Single gene methods power comparison. Power estimates for BOMP, VT, SKAT, KBAC (KBAC1P= minor allele frequency defined as < 1%, KBAC5P= minor allele frequency defined as < 5%). Each vertical line represents power estimates for each method, based on 250 simulated case-control studies. AA=the case-control studies were drawn from gene populations generated with an African-American simple bottleneck demographic model. EA=the case-control studies were drawn from gene populations generated with a European-American exponential growth demographic model. The eight variant causality (phenotype etiology) models are defined in Table 4.1. Since the European-American demographic model does not account for common or protective variants, etiologies involving common or protective variants were only considered for the African-American demographic model.

4.2.1 Power analysis of simulated case-control studies

In single-gene case-control study simulations, a study size of 2000 (1000 cases, 1000 controls) was required for any of the methods to achieve at least 80% power to detect causal variants. BOMP had > 80% power for three of the tested phenotype etiologies (Common variant, KeyRegion+Protect, and Common+Protect). When the study size was increased to 5000, several of the methods (BOMP, SKAT, VT, and KBAC5P (MAF< 5%)) had > 80% power for selected etiologies (Figure 4.3). BOMP was consistently more powerful than other methods and appeared to be particularly useful for certain phenotype etiologies (Key region variant, Common variant, and all etiologies involving protective variants (Table 4.1)). All methods were less powerful when applied to case-control studies using the European-American demographic model (in which variants are either rare or singletons) (Figure 4.4).

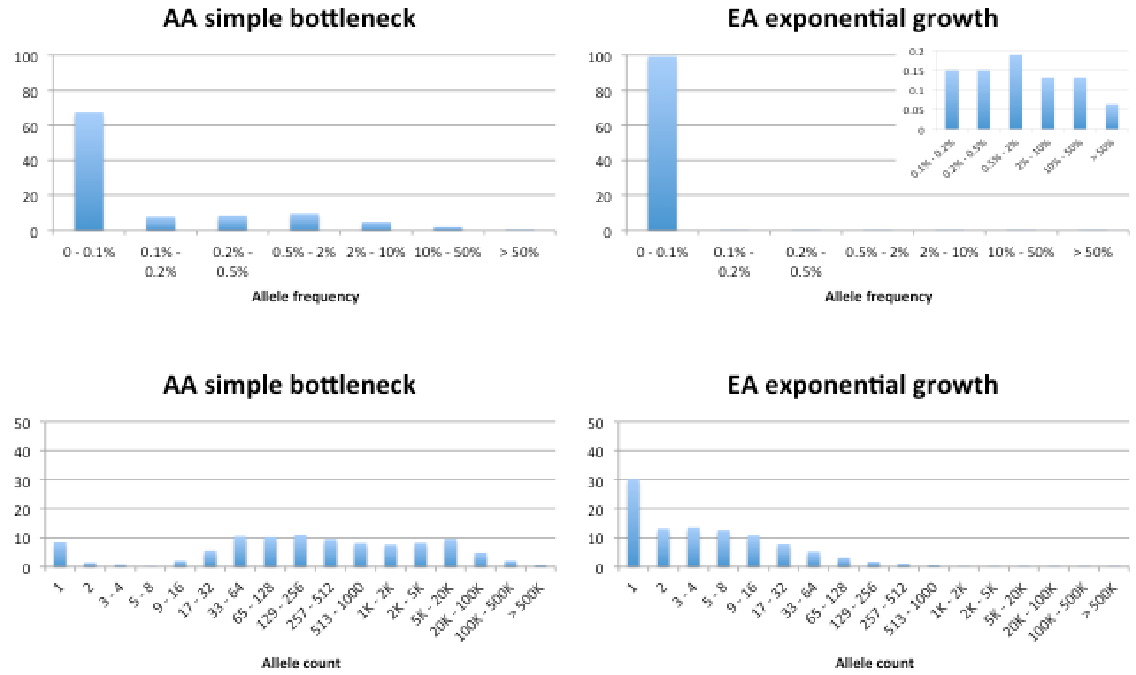


Figure 4.4: Distributions of allele frequencies and raw allele counts in simulated European-American and African-American populations. The European-American population consists almost entirely of rare variants, while the African-American population contains a wider range of rare, low-frequency, and common variants. Percentage of variants with allele frequencies and raw allele counts in the designated ranges are shown. Because > 99% of European-American allele frequencies are < 0.1%, I include a blow-up of frequencies > 0.1%, which range from 0.05% to 0.2%. Demographic models shown in Figure 4.1.

Next, I explored how the power of the tested methods could be improved by application to a candidate gene set rather than a single candidate gene. I simulated case-control studies, in which each genomic individual had multiple genes, all or some of which contained causal variants. The gene sets in which all genes contained causal variants ranged from 2 to 5 genes. Gene sets with mixtures of casual and non-causal genes ranged from 4 to 15 genes (ratios of causal to non-causal 3:1, 3:3, 3:6, 3:9, and 3:12). Causal variants were equally likely to be from any of the phenotype etiologies dominated by rare variants. The assumption that even 25% of genes in a set contain causal variants is certainly optimistic, but this experiment allowed us to compare the extent to which each method was affected by the fraction of causal genes in a set.

When all genes in a gene set contained causal variants, power increased for all methods as gene set size increased. When the gene sets contained a mixture of genes, both with and without causal variants, the power decreased with the causal to non-causal ratio. For the African-American demographic model, BOMP and SKAT were the most robust to gene sets with low causal to non-causal ratio. As in the single gene experiment, all methods had less power in the European-American demographic than in the African-American. For the European-American, none of the methods had power > 80% for any of the gene sets. For the African-American demographic, BOMP, SKAT, and VT had power > 80% when gene sets of sizes 4 and 5 contained all causal variants. BOMP was the only method with power > 80% for any of the mixed gene sets tested (gene set sizes 3,6, and 9, with ratio of causal to non-causal 3:0, 3:3 and 3:6) (Figure 4.5).

Biologically, I didn't expect that every gene in a real gene set would contain causal variants. Thus our simulated gene sets were designed to contain a mix of genes with causal variants and those without. The burden tests (VT) were not able to effectively capture the difference between the two and lost power as the number of non-causal variants in the simulations increased (Figure 4.5). The genotype vectors computed by KBAC become larger and more heterogeneous when applied to a gene set, rather than a single gene. Thus, the KBAC strategy of leveraging the number of shared genotype vectors among cases and/or controls was less effective when applied to gene sets than to single genes. The BOMP hybrid likelihood statistic (with strong contributions from the BOMP position distribution statistic), and SKAT were the most powerful when applied to gene sets, rather than single genes. I attributed this result to the increase in the number of significant localized units in a gene set that contained more than one causal gene.

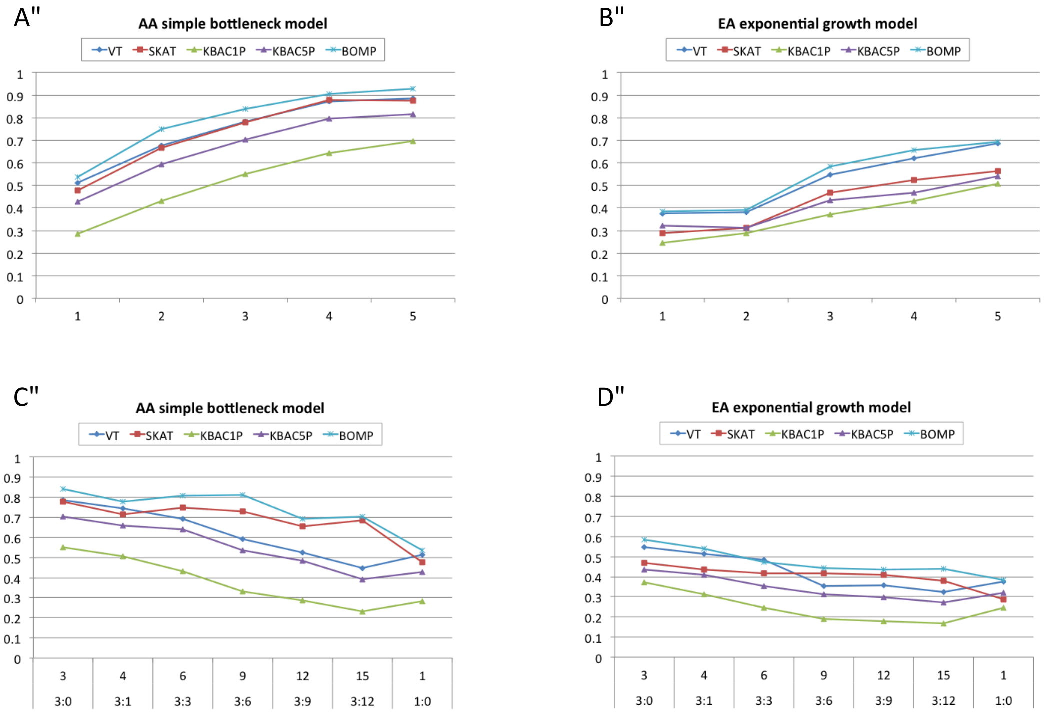


Figure 4.5: Power estimates for multiple gene case-control studies with causal variants equally likely to be from any phenotype etiology dominated by rare variants. A,B. X-axis shows number of candidate genes in 250 simulated case-control studies (approximately one-third each from phenotype etiologies Rare, LowFreq and KeyRegion). All genes contain causal variants. For each method, average power is shown. Power increases for all methods as the number of candidate genes with causal variants increases. C,D. X-axis shows the number of candidate genes and the ratio of genes containing causal variants to those that do not contain causal variants. As the ratio decreases, the power of the tested methods also decreases. (Tested methods are BOMP, VT, SKAT and KBAC1P= minor allele frequency defined as < 1%, KBAC5P= minor allele frequency defined as < 5%). AA=the case-control studies were drawn from gene populations generated with an African-American simple

bottleneck demographic model. EA=the case-control studies were drawn from gene populations generated with a European-American exponential growth demographic model.)

Next, because the underlying etiologies could be heterogeneous, I reconsidered the assumption that casual variants in a gene set were equally likely to come from a few phenotype etiologies that are dominated by rare causal variants. Instead, I sampled phenotype etiologies from nine multinomial distributions (Figure 4.6). For these experiments, the number of candidate genes was fixed at nine and the ratio of causal to non-causal genes was 3:6. BOMP's power advantage over the other tested methods was larger in this experiment than in the single candidate gene experiment. For case-control study size of 1000, BOMP power was > 80% for the multinomial distributions dominated by the key region variant etiology (African-American) and etiologies involving protective variants. For case-control study size of 2000, BOMP power was > 80% for all six multinomial distributions possible for the African-American model (Figure 4.7), and SKAT power was > 80% for the multinomial distributions dominated by the key region variant etiology (African-American) and etiologies involving protective variants.

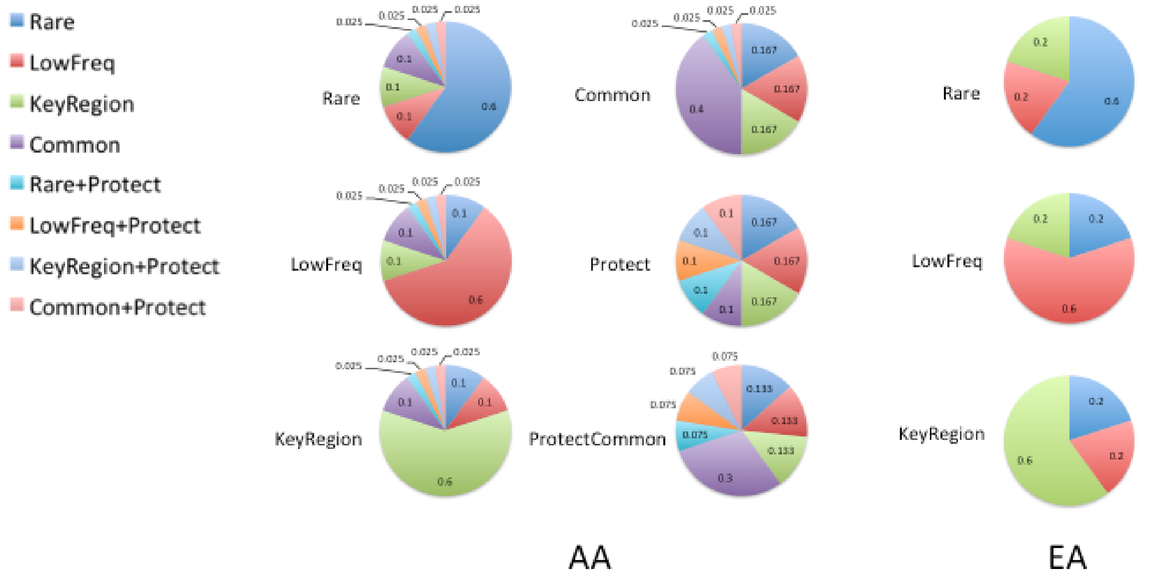


Figure 4.6: Nine multinomial distributions used to construct sets of multiple candidate genes for case-control studies. Each multinomial distribution is named for its dominant phenotype etiology.

In Figure 4.3, each of the eight designed etiologies was simulated for a single gene. In Figure 4.7, etiologies of causal genes in the gene set were dominated by one of the eight designed etiologies with the possibility of mixing other etiologies. Thus, comparing the single gene experiments in Figure 4.3 with the gene set experiments in Figure 4.7, performance of each method in general showed a similar trend across corresponding etiologies. However, BOMP was slightly powerful than other methods when testing on a gene set than when testing on a single gene. The power gain was attributed to the combination of mutation burden statistic and mutation position distribution statistic, expanding BOMP's capability of identifying causal genes with more than one etiology in a gene set.

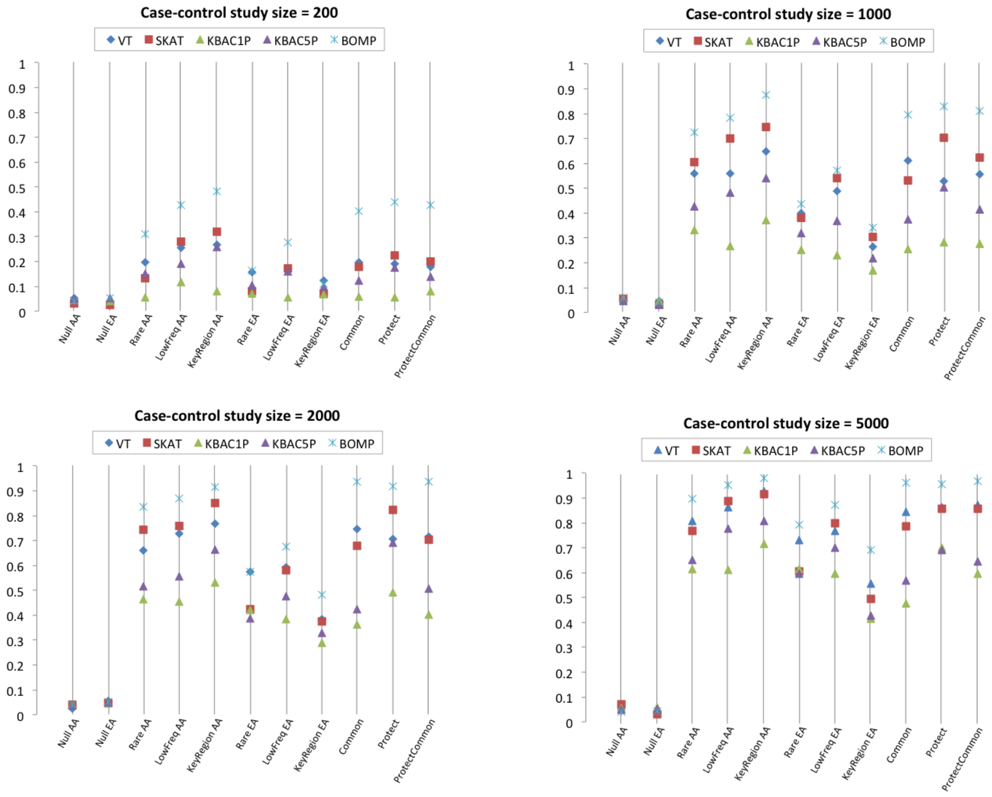


Figure 4.7: Power estimates for multiple genes case-control studies with causal variants from phenotype etiologies randomly sampled from nine multinomial distributions (Figure 4.6). Power estimates for BOMP, VT, SKAT, KBAC (KBAC1P= minor allele frequency defined as < 1%, KBAC5P= minor allele frequency defined as < 5%). Each vertical line represents power estimates for each method, based on 250 simulated case-control studies. The genomic individuals each had nine genes, of which three contained causal variants and six did not. The phenotype etiologies for the three genes with causal variants were randomly sampled from nine multinomial distributions (Figure 4.5). AA=African-American simple bottleneck demographic model. EA=European-American exponential growth demographic model.

Genetic causes of a complex trait can be heterogeneous. Many genes are causal but having variants in few of them is sufficient to develop the phenotype. Testing on single genes or a small gene set will not have sufficient power to identify the association due to the sparse signals. Instead, testing on a carefully chosen larger gene set, which contains a significant part of causal genes, may work. To examine this, I explored the power of BOMP with respect to case-control study size, using a set of 24 candidate genes as the functional group. I varied the ratio of casual to non-causal genes from 1:3, 1:1, and 3:1. Here, causal variants were again equally likely to be from any of the phenotype etiologies dominated by rare variants. For a case-control study size of 1000, BOMP's power exceeded 0.8, regardless of the causal-to-non-causal gene ratio (African-American only), and for the 1:1 and 3:1 causal-to-non-causal gene ratios for European-American. A study size of 200 was sufficient for power > 0.8 for 1:1 and 3:1 ratios (African-American only) (Figure 4.8).

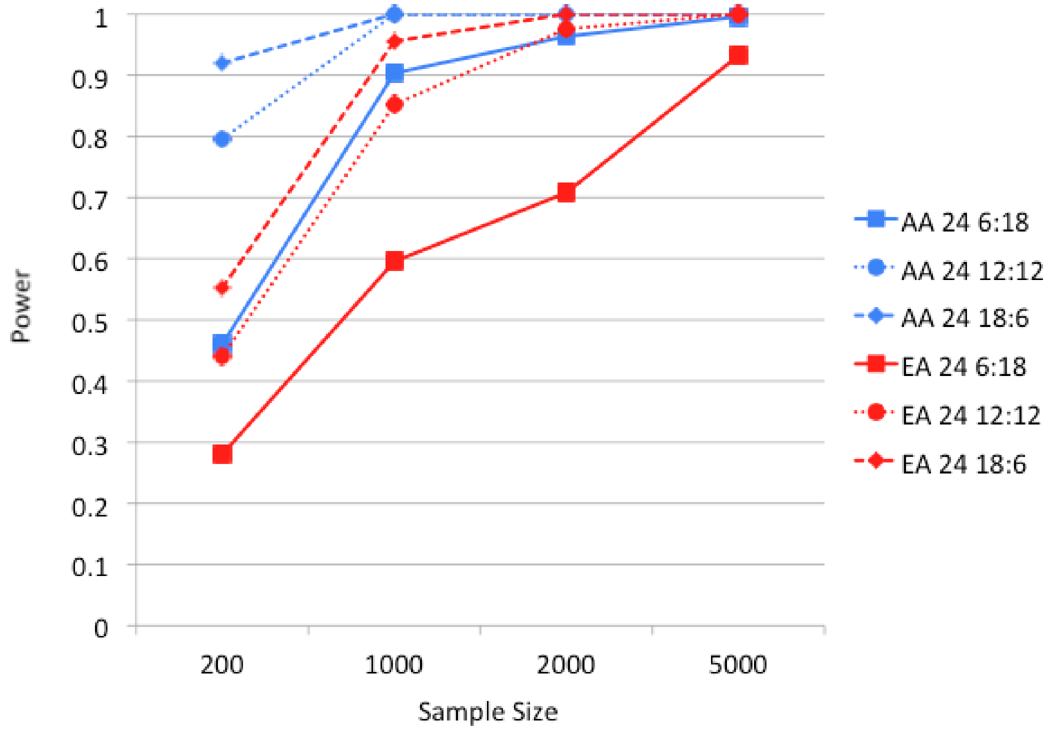


Figure 4.8: BOMP Power estimates for multiple genes (24) case-control studies. Power estimates for BOMP; each estimate is based on 250 simulated case-control studies ((approximately one-third each from phenotype etiologies Rare, LowFreq and KeyRegion). The genomic individuals each had 24 genes, the tatio of genes with causal variants to those without causal variants was either 1:3 (6 causal, 18 non-causal), 1:1 (12 causal, 12 non-causal), or 3:1 (18 causal, 6 non-causal). AA=African-American simple bottleneck demographic model. EA=European-American exponential growth demographic model.

I reasoned from these results that, for a population whose allele frequency spectrum is similar to our European-American demographic model simulations, current whole-exome case-control studies were not sufficiently powered. These studies lacked power to find causal variants both at the single gene level (as proposed by [37]) and for modestly-sized gene sets. However, if the allele frequency spectrum is more similar to the African-American demographic model, BOMP may be able to detect causal variation in larger gene sets, given the size of current whole-exome studies.

4.2.2 Relative contributions of mutation burden and mutation positional distribution in simulated case-control studies

BOMP is composed of the mutation burden statistic and the mutation position distribution statistic, each of which tests the association using different approaches. The mutation burden statistic collapses all (weighted) variants within a functional group, performing a unidirectional burden test, while the mutation position distribution statistic collapses variants within small windows that segment a functional group, performing a bidirectional over-dispersion test over the windows. It is interesting to find whether the combined one is better than either and when one is better than the other.

I computed average power for single candidate gene case-control studies and multiple candidate gene case-control studies (nine genes, 3:6 causal to non-causal ratio), with respect to both demographic models, all phenotype etiologies (Table 4.1) for single genes, and all combinations of phenotype etiologies for gene sets (Figure 4.6). BOMP's hybrid likelihood model had better power than either of its components: the mutation burden and mutation position distribution statistics (Figure 4.9). The burden statistic had more power in single-gene studies, while the position statistic had more power in the gene set studies.

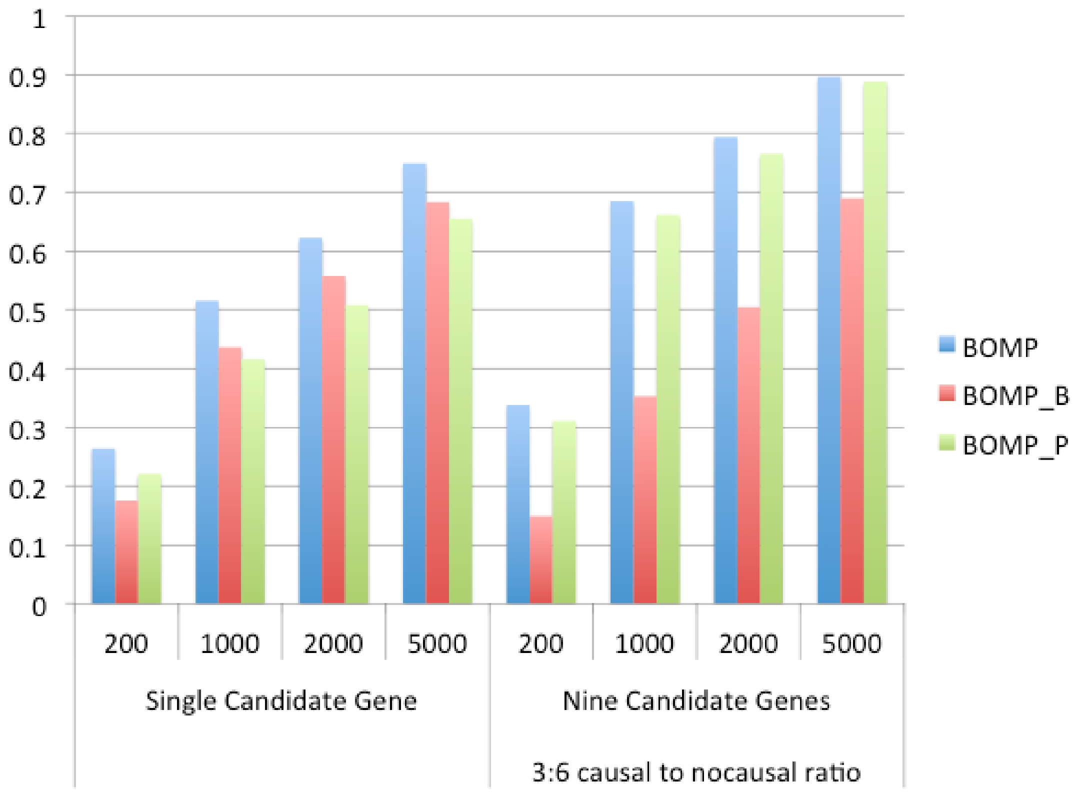


Figure 4.9: BOMP burden and position statistics complement each other. Breakdown of contribution of BOMP mutation burden (BOMP B) and BOMP positional distribution (BOMP P) statistics averaged over single candidate gene power estimates (Figure 4.3) and multiple candidate gene power estimates (nine genes, 3 with causal variants and 6 with no causal variants) (Figure 4.7) for case-control study sizes of 200, 1000, 2000, and 5000. Combining the two statistics consistently yielded improved power with respect to each statistic on its own. The BOMP burden statistic had more power than BOMP position for the simulations based on a single candidate gene, and vice versa in the simulations with nine candidate genes and 3:6 causal to non-causal ratio.

In general, mutation burden tests outperformed the position distribution statistic when causal variants were rare and were not clustered. The position distribution test outperformed burden tests when the number of rare variants was similar in cases and controls, but where cases and controls differed with respect to the positional distribution of the variants. To illustrate this point, I show a case in which burden tests would miss such a difference (Figure 4.10). In the genomic region shown, cases and controls each have 9 total variants, but an informative window segmentation yields distinct regions in which the number of variants seen in cases and controls is substantially different. The difference between cases and controls is also missed by SKAT, which consider variants one at a time, because at each position the number of variants in cases and controls is similar.

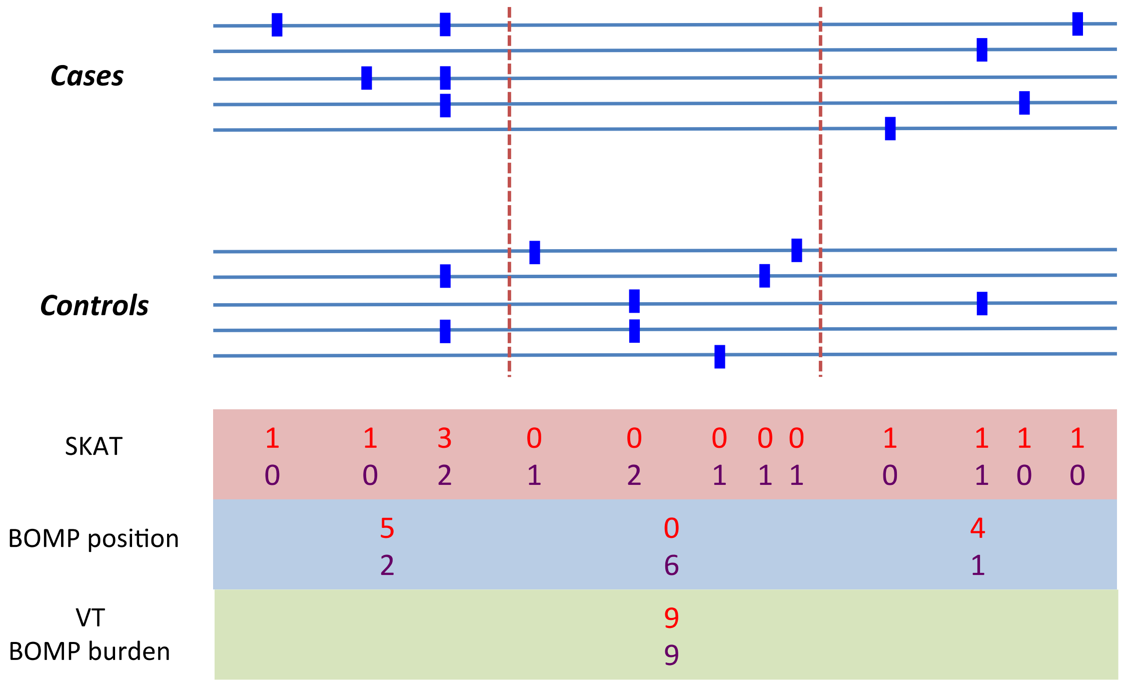


Figure 4.10: Example variation pattern in which positional distribution outperforms burden tests. A toy example of a genomic region containing variants (blue squares) in cases and controls. I assume that the region is important for phenotype. Variant counts in casees (red). Variant counts in controls (purple). Cases and controls each have a total of 9 variants in this region, so burden statistics (e.g., VT or BOMP burden) are not able to detect that the region is important for phenotype. BOMP's position distribution statistic collapses variants into short, localized windows (red dashed lines) and detects that the number of variants seen in cases and controls is different within the windows. I note that a method that does not collapse variants, such as SKAT, does not have much power to detect the difference between cases and controls, because at each position the number of variants in cases and controls is similar.

Both collapsing burden and position distribution tests outperform SKAT when causal variants are very rare. Figure 4.11 shows the power of VT, SKAT, BOMP burden, and BOMP position distribution in 10,000 simulated European-American individuals, using our Rare and Key Region Variant phenotype etiologies. The European-American populations contain a large fraction of rare variants (Figure 4.4 shows exact allele frequencies and raw counts). Our simulations of the Rare Variant etiology in this population generate rare variants that are not clustered, and the methods with highest power are VT burden and BOMP burden tests. In Key Region Variant simulations where rare variants are positioned differently in cases and controls, the position distribution statistic has higher power than either of these burden tests.

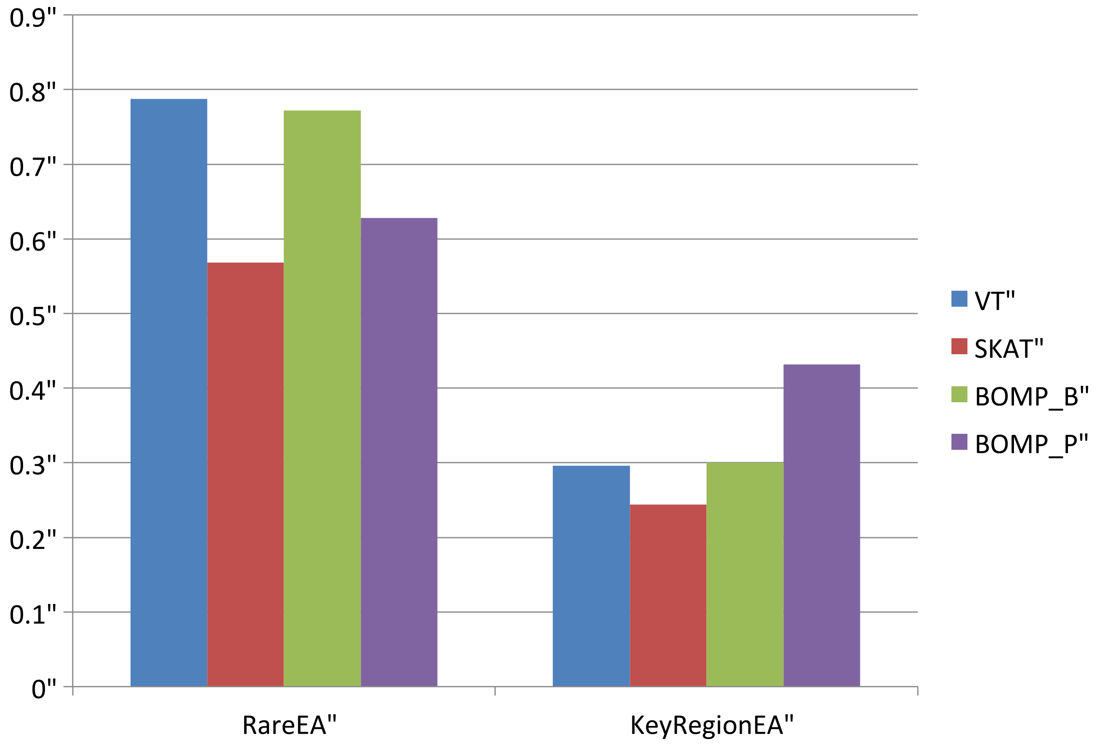


Figure 4.11: Power of position distribution statistics compared to burden methods and SKAT. Burden tests outperform the position distribution statistic when causal variants are rare and are not clustered, as in our simulations of Rare Variant phenotype etiology and European-American demographic. The position distribution test outperforms burden tests when the number of rare variants is similar in cases and controls but the positional distribution of the variants differs in cases and controls like simulations of Key Region Variant phenotype etiology and European-American demographic. Both collapsing burden and position distribution tests outperform SKAT when causal variants are very rare. RareEA = rare variant phenotype etiology (Table 4.1) and European-American demographic model. KeyRegionEA = key region variant phenotype etiology (Table 4.1) and European-American demographic model. Power shown on Y-axis. Simulations with 10,000 samples are shown.

In summary, collapsing unidirectional causal variants increases power while collapsing neutral (non-causal) variants or causal variants in the opposite direction dilutes (or cancels out) the mutation burden difference between cases and controls, and thus causes power loss. In the gene set simulation, the functional group was mixed with both causal genes and non-causal genes, which contain many neutral (non-causal) variants. Burden tests, which collapse all variants within the functional group, lost power while the position statistic, which does not collapse all variants across the functional group, was relatively invariant to the mix of neutral variants compared with burden tests. In the simulations of Rare Variant etiology, burden tests gained power by collapsing unidirectional causal variants. In the simulations of Key Region Variant etiology where rare causal variants occurred only in small domains in the protein sequences, burden tests still obtained some power gain by collapsing all variants when compared to SKAT, which does not collapse any variants. The position statistic, which precisely collapses causal variants in the domains without mixing other neutral variants, gained the most power.

4.3 Performance summary for each method based on simulations

In the simulations, I found that the performance of all tested methods depended on choice of phenotype etiology and demographic model.

The VT method is a burden test, a class of methods in which it is assumed that all variants lower than a MAF of interest are deleterious. Burden tests accumulate signal by collapsing variants across a genomic region. They are valuable in detecting associations between rare variants and phenotypes, when case-control studies are not large enough to detect association between a single rare variant and the phenotype. An important advance of the VT burden test is that it adaptively learns a MAF threshold from the data, efficiently filtering out a large number of non-causal variants. In our simulations, it always has the most power for rare variant etiology and European-American demographic, in which the population is enriched for very rare variants. However, in other etiologies, it does not do quite as well. For example, in the low frequency etiology, in which causal variants must have at least 0.1% MAF, a single MAF threshold will not well separate causal and non-causal variants. VT also loses some power when common or protective variants are included in an etiology and does not gain power when variant positions are distributed differently in cases and controls.

The SKAT linear kernel regression model circumvents an important limitation of burden tests, which is that they can only detect causal variants that are deleterious. In contrast, SKAT belongs to a class of methods in which emphasis is placed on the variance of genotype frequencies in cases and controls. It does no collapsing and each variant (genotype) is treated as an independent covariate. Thus, it is sensitive to both deleterious and protective effects, and it has good power in the presence of protective variants. SKAT weights variants according to their MAF, by treating the MAF as a random variable from a Beta density. Beta parameters are set so as to give increased weight to those MAFs considered most likely to be causal a priori. Using default settings, common variants get decreased weight, which is reflected in power loss in our simulations based on the common variant etiology (MAF>5%). Variants in simulations based on the low frequency etiology (0.1%<MAF≤5%) are well captured by this Beta density, and I see that SKAT has good power for this etiology. While rare variants also get high SKAT weights, I found that SKAT loses power in simulations based on the rare variant etiology in the European-American demographic model. Because there are many extremely rare variants present, the SKAT

model will estimate a very small marginal effect size for each variant and fail to reject the null hypothesis of zero marginal effect size.

KBAC differs from SKAT, VT, and BOMP in that the phenotype association of a genotype vector, rather than single variants, is calculated. Each individual is represented by a pattern of 0's, 1's, and 2's for M sites in a candidate genomic region. Only sites where rare variants have been seen in a case or control are included. An adaptive weight is computed for each vector using a mixture model, and the KBAC statistic for a genomic region of interest is the weighted sum of vector frequency differences in cases vs. controls. A strength of this strategy is that it implicitly considers interactions among variants, which are not incorporated into the other tested methods. A weakness is that the information in a vector frequency difference depends on the number of individuals sharing a common vector. In sequencing studies with many rare variants, the probability of seeing the same vector more than once is low. In our simulations, there are no interactions among variants and an important strength of KBAC is not utilized. Particularly in simulations involving the European-American exponential demographic model, which is enriched for very rare variants and thus has few shared vectors, KBAC has low power. Like SKAT, KBAC can handle protective variation, and in our simulations, it has relatively good power for etiologies that include protective variants. KBAC only considers variants with less than a pre-specified MAF to be causal. I selected MAF thresholds of (recommended) 1% and 5%, which resulted in KBAC having low power for our common variant etiology. However, KBAC can be extended to include common variants as co-variates in a logistic regression model, which could improve power for this etiology.

BOMP is a hybrid likelihood model, which conceptually (but not mathematically) incorporates the general approaches represented by VT and SKAT. VT assumes that cases have more variants in phenotype-associated genomic regions, with MAF below an optimized threshold, than controls. SKAT assumes that the subset of variants that are truly phenotype-associated are over-represented in cases or controls. The BOMP burden statistic captures the scenario in which cases have more variants than controls (allele frequencies are incorporated via coefficient scores), while the position distribution statistic captures the scenario in which cases have more variants than controls (or vice versa) in highly localized, functionally important genomic regions, from 8 to 64 codons in length. Essentially, SKAT does the same thing, except that the localization unit is defined as a single variant (codon or nucleotide). When SKAT or the BOMP position distribution statistic detects significant differences between cases and controls within a localization unit, no further collapsing takes place. Thus, both do well at detecting protective vari- ants. However, using localized collapsing, the BOMP position distribution statistic gains power to detect differences between cases and controls (Figure 4.10) that would be missed by either SKAT or a burden statistic.

BOMP's combination statistic is effective for most of our tested etiologies. It is particularly effective for the key region etiology, but relatively less effective for the rare variant etiology, particularly with the European-American demographic. For this etiology, by construction, the causal rare variants are distributed randomly across the simulated genomic region, and the BOMP position distribution statistic gains no power by collapsing within a localized region. A burden statistic using an optimized MAF threshold is very effective for this etiology.

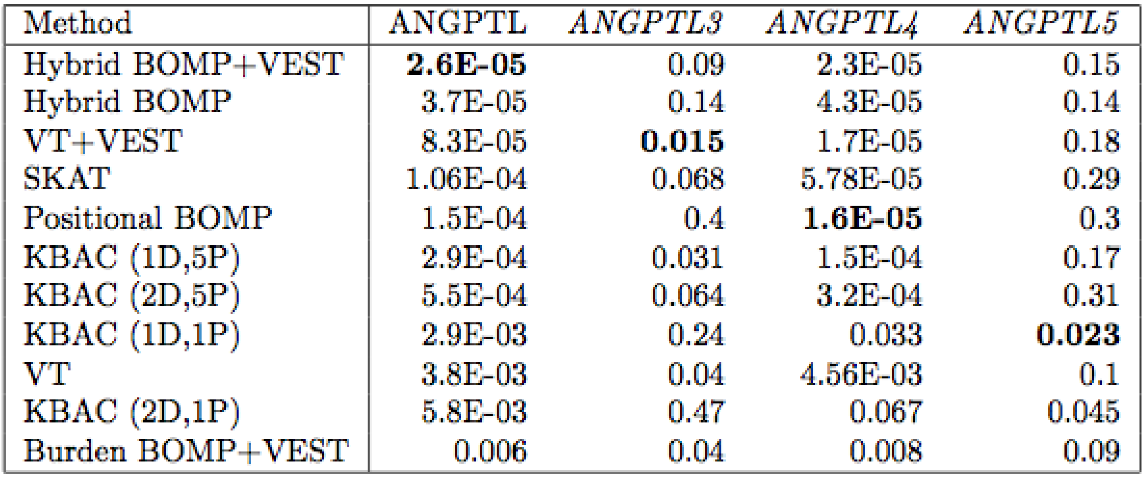
4.4 Dallas Heart Study

Extensive simulation investigated the performance of each method being applied on various phenotype etiologies. However, the performance of each method being tested on simulated data may differ from that on real data due to the complex nature of population genetics. For example, the strength of linkage disequilibrium varies across the human genome. Real genomes may contain hidden population structures that are likely to produce false associations. Moreover, when dealing with real data, technical problems such as batch effects and sequencing errors may be introduced. Therefore, in addition to extensive simulation, an empirical benchmark is required for a fair comparison.

I applied the BOMP hybrid likelihood model to the analysis of data from the Dallas Heart Study (DHS) [36]. Romeo *et al.* explored genetic contributions to plasma triglyceride (TG) levels in ~ 3500 individuals in the DHS, by resequencing the coding regions of angiopoietin-like (ANGPTL) family genes, which were hypothesized to play key roles in TG metabolisms in humans. The ANGPTL genes regulate the activity of a key enzyme in TG metabolism, lipoprotein lipase (LPL), via post-transcriptional modifications and were jointly associated with low triglyceride levels by [36]. Specifically, ANGPTL3, ANGPTL4, and ANGPTL5 were functionally validated as causal genes, playing non-redundant roles and underlying TG levels as a functional group [36]. The ANGPTL gene set has been analyzed in several computational papers and used as a benchmark to compare methods that predict the impact of rare and common variants from sequencing data [20,21,23,43-46].

I stratified the DHS samples by ethnicity (Hispanic, non-Hispanic white, non-Hispanic black) and gender. Because BOMP was designed for dichotomous phenotypes, I selected the lower and upper quartiles from each group, by TG level (totaling 1775 individuals, with 897 cases and 878 controls). Sixty mutations in ANGPTL3, ANGPTL4, and ANGPTL5 occurred in these individuals.

I computed a P-value for each of the three ANGPTL genes and for the ANGPTL gene set, using BOMP (with and without bioinformatics scores), the burden statistic VT (with and without bioinformatics variant weighting), the overdispersion statistic SKAT, and the mixture-model KBAC statistic (with four parameter settings) (Table 4.2).



Method	ANGPTL	*ANGPTL3*	*ANGPTL4*	*ANGPTL5*
Hybrid BOMP+VEST	**2.6E-05**	0.09	2.3E-05	0.15
Hybrid BOMP	3.7E-05	0.14	4.3E-05	0.14
VT+VEST	8.3E-05	**0.015**	1.7E-05	0.18
SKAT	1.06E-04	0.068	5.78E-05	0.29
Positional BOMP	1.5E-04	0.4	**1.6E-05**	0.3
KBAC (1D,5P)	2.9E-04	0.031	1.5E-04	0.17
KBAC (2D,5P)	5.5E-04	0.064	3.2E-04	0.31
KBAC (1D,1P)	2.9E-03	0.24	0.033	**0.023**
VT	3.8E-03	0.04	4.56E-03	0.1
KBAC (2D,1P)	5.8E-03	0.47	0.067	0.045
Burden BOMP+VEST	0.006	0.04	0.008	0.09

Table 4.2: Dallas Heart Study. P-values of association between dichotimized trygliceride levels and variation in three ANGPTL family genes sequenced in Dallas Heart Study. ANGPTL - multiple gene set including ANGPTL3, ANGPTL4, and ANGPTL5. The most significant P-value for each is highlighted in bold. BOMP= combined Burden and Position statistics VT = variable threshold burden test [20] SKAT = sequence kernel association test (linear weighting version) [21], KBAC = Kernel-based adaptive cluster [44] (1D = single direction, 2D = two direction, 1P=rare variants defined as < 1% MAF, 5P=rare variants defined as < 5% MAF). VEST = BOMP and VT with VEST score variant weighting.

The hybrid BOMP test, with bioinformatics scores and allele frequency variant weighting, had the most significant P-value for the ANGPTL gene set (P = 2.6E − 05), which should be sufficient to detect ANGPTL-phenotype association, using a gene set based analysis in a whole-exome study, after multiple testing correction (Table 4.2). The hybrid BOMP P-value was more significant than either of its components (the BOMP burden and position distribution scores). This result was consistent with the average behavior of BOMP in our simulation-based analysis of power (Figure 4.8). However, the two component scores did not yield an improved hybrid score on every gene. For ANGPTL3, ANGPTL4, and ANGPTL5, the hybrid score P-value was not as significant as the P-values of the most significant component score. The burden-based VT score (with bioinformatics score variant weighting) had the most significant P-values for ANGPTL3 (P = 0.015); the BOMP position distribution score for ANGPTL4 (P = 1.6E − 05), closely followed by VT (with bioinformatics scores) (P = 1.7E − 05), and the overdispersion test SKAT (P = 5.78E−05). KBAC, with single directional scoring (only deleterious variants counted) and threshold for rare variation set at MAF < 1%, had the most significant P-value for ANGPTL5 (P = 0.023).

These results confirm previous reports that the performance of current methods to detect causal variants depends on which genes are selected for benchmarking [14,24]. While the dataset is small, it is interesting to note that P-values of association between variant ANGPTL family genes and dichotomized serum triglyceride levels from the Dallas Heart Study were most significant for the BOMP hybrid model, when the genes were considered together as a gene set. However, the burden statistic VT had the most significant P-value for ANGPTL3, and the KBAC P-value was the most significant for ANGPTL5 (specifically with single directional scoring and threshold for rare variation set at MAF < 1%). For

ANGPTL4, the most significant P-values were from the BOMP positional distribution score, VT, and SKAT. Each of these genes had a different pattern of variation frequencies in cases and controls, which presented advantages and obstacles for each method. For example, ANGPTL3 had a high frequency variant (M259T) that occurred more often in cases than controls and many "noisy" rare/singleton variants that occurred either in cases or controls. VT took advantage of the signal in M259T because its threshold adapted to maximize the burden increase in cases versus controls, and thus M259T was included in its burden calculation. BOMPs burden statistic did not give as much importance to M259T, because it down-weighs high frequency variants. KBAC included M259T only when its allele frequency threshold parameter was set to 5% but was penalized when it was set to 1%. ANGPTL4 had two high frequency variants (T266M with AF=0.27 and R278Q with AF=0.03). T266M occurred more often in controls while R278Q occurred more often in cases. VT took advantage of signal in R278Q and other rare variants that occurred more often in cases and adaptively learned the allele frequency threshold to filter out T266M from analysis. BOMP position distribution statistic and SKAT took advantage of signals in both T266M and R278Q while BOMP burden statistic was penalized by T266M. ANGPTL5 had a high frequency variant (T268M), which occurred more often in controls. Other variants are very rare although they occurred more often in cases. Because of the sparse signals, BOMP, VT and SKAT did not perform well on ANGPTL5.

BOMP is not designed to be adjusted for additional covariates, which are often available in phenotype studies. For example, it is not designed to explicitly deal with different ancestries in a structured population. However, if the true population structure is known and the number of subpopulations is not too large, I can run analyses with stratification to get around this problem, as I (and the authors of the VT and SKAT papers) did for ANGPTL family genes in the Dallas Heart Study [20,21]. Using this strategy, one begins with a quantitative trait (serum triglyceride levels), stratifies individuals into groups, then identifies extreme phenotype individuals from each group. Cases are then those individuals from all groups at one extreme and controls are those individuals from all groups at the other extreme. An alternative strategy is to permute case-control labels only within each group to generate a correct null distribution.

Incorporating bioinformatics scoring of variants (by VEST) yielded improved P-values for both BOMP and VT on the Dallas Heart Study data. While it has been suggested that bioinformatics misclassification of variants might be more of a liability than a benefit, our results (albeit on a small gene set) suggest the opposite. Functional classification of variants in both coding and non-coding regions of the genome is an active research area in bioinformatics, and as methods improve, it is likely that they will increasingly contribute to statistical analysis of causal variation.

Chapter 5 Application on bipolar case-control study

5.1 Genetic studies in Bipolar disorder

5.1.1 Contribution of genetic components in Bipolar disorder

Bipolar disorder (BP) is among the most important public health problems in the world. According to the WHO Global Burden of Disease Study, BP is one of the top ten leading causes of lifelong disability. The illness is characterized by manias and depressions, which are syndromes of abnormal mood, thinking, and behavior. Originally called manic-depressive illness, BP was one of the first psychiatric disorders to be targeted for genetic study; however, unraveling the specific genetic causes of BP has proven to be a formidable challenge. Studies in genetic epidemiology have revealed a substantial genetic component to this illness, supporting the rationale for screening the genome for variants associated with BP. Evidence for the importance of genetic factors in BP etiology is well supplied by family, twin, and adoption studies. BP probands were studied in 12 family studies, in which the majority of subjects were directly interviewed [47]. The results showed that the combined rate for BP in relatives of ill probands was 10.7%, while the comparable rate in relatives of control probands was 1.0%. Moreover, compared to schizophrenia, the sibling recurrence risk for BP is roughly the same, but far below that for single-gene diseases such as phenylketonuria where the relative risk to siblings is many fold higher. The magnitude of the relative risk to siblings suggests that genes associated with BP should be discoverable. Twin studies have attempted to distinguish the impact of shared environment from that of shared genes. In the three studies that assessed BP, the differential concordance rate for MZ twins (63%) and for DZ twins (13%) yielded a heritability figure of 0.78 [47]. Adoption studies have also attempted to separate genetic from environmental effects. Mendlewicz and Rainier, who conducted the most methodologically rigorous adoption study of BP, found that the biological parents of BP adoptees had a 31% rate of mood disorders, which was significantly higher than the 12% rate of mood disorders in the adoptive parents of these adoptees and the 2% rate in the biological parents of the control adoptees [48]. Several segregation analyses in family studies supported a major locus contributing to BP inheritance [49-51], but others have not found such evidence [52,53].

5.1.2 Linkage analysis in Bipolar disorder

Many BP linkage analyses have been conducted to identify BP susceptibility loci. A reliable finding requires genome-wide statistical significance and replication of loci in more than one study. Many BP susceptibility loci have been repeatedly implicated, but not with genome-wide significance. Several findings have reached genome-wide statistical significance in multi-family samples: 8q24 [54], 15q14 [55], 18q12 [56], 21q22 [57], and 22q12 [58]. Unfortunately, these regions have not been consistently replicated across studies. Three meta-analyses of linkage studies have been conducted for BP. Badner and colleagues performed a meta-analysis of 11 BP genome scans, replicating two significant regions, 13q32 and 22q 12-13, at a genome-wide level [59]. However, a second meta-analysis, which used data from 18 genome scans and obtained unpublished data from investigators, found that no region reached genome-wide significance across the combined studies. The strongest regions were: 9p21-22, 10q11-22, 14q24-32, 18p-18q21, and to a lesser extent 8q24 [60]. The third meta-analysis, examining original genotype data from genome-wide scans including 5,179 subjects in 1,067 families, provided strong support for linkage on 6q and 8q as well as suggestive evidence for loci on 9p and 20p [61]. Three meta-analyses did not demonstrate a clear agreement.

5.1.3 GWA Studies in Bipolar disorder

Recently, genome-wide association studies (GWAS) have been performed for BP based on the CDCV (Common Disease Common Variant) hypothesis. The first four independent studies did not identify any genome-wide significant signals [8,62-64]. Signals in ANK3 gene and CACNA1C gene, however, showed statistical significance at a genome-wide level in a meta-analysis, which combined 3 GWAS samples [65]. Interestingly, both of these genes encode proteins that play a role in synaptic function, as ANK3 is an adaptor protein found at axon initial segments that has been shown to regulate the assembly of voltage-gated sodium channels, and CACNA1C is a calcium channel subunit. Later, a meta-analysis combining 11 GWAS samples identified another genome-wide significant gene, SYNE1, associated with BP [66].

5.1.4 Rare variant search in Bipolar disorder

Both CDCV (Common Disease Common Variant) and CDRV (Common Disease Rare Variant) were theoretically supported to explain the underlying genetics for complex diseases [6,67]. GWA studies, which are based on the CDCV hypothesis, have had a number of notable successes in many diseases like diabetes, coronary heart disease, and Crohn's disease [8]. On the other hand, rare variants (the genetic causes hypothesized by CDRV) have been found to play a role in lipid abnormalities [68] and in severe childhood onset obesity [69].

The two approaches are not incompatible as all human diseases have an allelic series of mutations that can range from the very rare to the common. Classical diseases such as hemoglobinopathies have long been known to have extremely rare (<0.01%) mutations in the beta globin gene leading to thalassemias and very common (>10%) mutations such as the sickle mutation. These variants have different mutation rates but are likely maintained by different population genetic processes such as natural selection: mutation-selection balance for the rare variants and overdominant selection for the common ones. Broad screens of a variety of cases reveal both rare and common variants. Studies of RET mutations in Hirschsprung disease [70] and those by Jeff Murray on IRF6 in cleft lip and palate [71] are good examples of the likely mutation spectrum. In Hirschsprung disease, the vast majority of mutational types are individually rare (<1/1,000) with large effects on penetrance, but the most common mutation is a polymorphic (24%) enhancer variant with a smaller effect on penetrance.

Because GWA studies are only designed to assay variants with at least 1-5% minor allele frequencies, rare variants are not captured. Only deep resequencing can assess this potentially critical portion of disease alleles. Several recent studies have uncovered rare variants related to common diseases. One high- throughput sequence study of the ANGPTL4 gene found a 3.8% prevalence of rare variants in Caucasian subjects with triglycerides in the lower quartile compared with 0.5% in subjects with triglycerides in the highest quartile [72]. Another study of MCR4 and obesity found a rate for rare variants of 2.6% vs. 0.6% in controls [73]. These studies demonstrate that differences in the gene-wide rates of coding rare variants in the 2-3% range can be detected in a deep resequencing study. Considering limited success in BP GWA studies and the evidence of both common and rare variants contributing to complex diseases, performing a deep resequencing study to uncover rare variants that are associated with BP is required.

5.2 Bipolar case-control study

A sequence-based case-control study for BP was conducted in the psychiatry department at Johns Hopkins University since 2010. The project originally targeted the coding regions of genes expressed in synapses, which were hypothesized to be highly related to BP, and later extended to a whole-exome sequencing case-control study. BOMP was used in the study for identifying BP causal genes.

5.2.1 Sequencing platform, variant calling and quality control

Groups in the psychiatry department at Johns Hopkins examined whole-exome sequencing data on the first 1,177 cases and 1,155 controls from this study. These samples were sequenced in four rounds, over a four-year period. In the first round, Nimblegen v1.0 arrays were used for exome capture and the Illumina GAII platform for next-generation sequencing. In the subsequent rounds, Nimblegen v2.0 arrays and the Illumina HiSeq2000 platform were used, with promoter regions and some extra genes added in the last two rounds. Only samples with target sequencing coverage of at least 80% at 20X sequencing depth were included for further analysis. Sequence readouts from the samples were aligned to the human reference genome sequence database using BWA [74]. Variants were then called after realignment around indels and recalibration of base quality scores with GATK [75] in target regions. Quality control for variant calling required coverage of at least 6X depth with a SNP quality score of 30 or higher to eliminate false-positives. Variants were annotated to dbSNP135 and collected in VCF files. The quality controls include missingness (missing entries) per position, missingness per subject, Hardy-Weinberg Equilibrium (HWE) and principle component analysis (PCA) for ancestry background checks. Finally, 1,135 cases and 1,142 controls (1,076 males and 1,201 females) were analyzed using BOMP.

5.2.2 BOMP analysis

Acknowledging the results of genome-wide association studies, we hypothesized that BP should be largely attributed to rare variants in the population. We defined variants with minor allele frequencies less than 5% to be rare. Any position with a missing rate- the fraction of missing calls across subjects- higher than 0.15 is removed in the analysis. For the remaining positions, missing calls are filled with major alleles and only non-synonymous variants (including missense, stop loss and stop gain variants), small indels causing frameshits and exonic splicing variants are examined.

To predict whether a missense variant is damaging to the protein function or not, we used several bioinformatics tools, including Polyphen2 (both the HumDiv model and the HumVar model) [76], SIFT [77], Mutationtaster [78] and VEST [27]. Based on the predictions from these tools and the mutation types, variants are sorted into three groups – a disruptive group (DIS), a non-synonymous strict group (NSS) and a non-synonymous broad group (NSB). DIS included stop loss, stop gain, exonic splicing and frameshift variants. NSS included variants in DIS plus missense variants predicted as damaging by all of the bioinformatics tools listed above. Finally, NSB included variants in DIS plus missense variants predicted as damaging by any of the bioinformatics tools listed above.

BOMP was run on each group. The bioinformatics weights for variants in DIS were set to 1.0 and the bioinformatics weights for the remaining missense variants were set to VEST scores in the BOMP burden statistic calculation. A maximum of 1,000,000 permutations were used to evaluate the statistical significance for each phenotype-gene association. Both Bonferroni and Benjamini-Hochberg procedures [79] were used for correction of testing multiple genes.

5.2.3 BOMP single gene results

The top 30 genes, in order of ascending BOMP p-values, from the DIS, NSS and NSB groups are shown in Table 5.1, 5.2 and 5.3 respectively. Unfortunately, no gene was statistically significant after multiple testing corrections. TYRO3 was the only gene with false discovery rates less than 1.0 in all three analyses, and may be an interesting candidate gene for further study.

TYRO3 produces a receptor protein tyrosine kinase (RPTK) whose function is to transduce signals from the extracellular matrix into the cell by binding to several ligands. It regulates many physiological processes including cell survival, migration and differentiation. TYRO3 signaling is also involved in processes such as neuron protection from excitotoxic injury, platelet aggregation and cytoskeleton reorganization. Additionally, RPTKs have been shown to modulate signaling cascades that influence synaptic function in the central nervous system (CNS) [80,81]. A recent study found prominent expression of TYRO3 in dendrites might suggest the capability to modulate signaling pathways triggered by synaptic transmission [82]. This finding aligned well with our original hypothesis of a causal relationship between synaptic genes and BP.

Name	BOMP	BOMP_B	BOMP_P	Bonferroni	BH-FDR
TYRO3	0.000046	0.000058	0.000281	0.410044	0.410044
FAM81B	0.000370	0.001990	0.000790	1.000000	1.000000
ZNF677	0.000610	0.000610	0.001300	1.000000	1.000000
PKHD1	0.000880	0.010050	0.000950	1.000000	1.000000
RBMX	0.001100	0.001100	0.003800	1.000000	1.000000
MUC6	0.001900	0.010699	0.003700	1.000000	1.000000
PFAS	0.002000	0.016998	0.001200	1.000000	1.000000
ETFB	0.002400	0.002500	0.009799	1.000000	1.000000
LY75-CD302	0.003300	0.011799	0.003400	1.000000	1.000000
DEPDC1	0.003900	0.019398	0.003800	1.000000	1.000000
VPS13B	0.004500	0.006199	0.008799	1.000000	1.000000
TMCO4	0.004600	0.025697	0.006299	1.000000	1.000000
BRCA2	0.005899	0.029697	0.003500	1.000000	1.000000
CLEC7A	0.007099	0.007099	0.015298	1.000000	1.000000
PTH2R	0.007099	0.024198	0.009299	1.000000	1.000000
CBX8	0.007199	0.010299	0.010499	1.000000	1.000000
LARP7	0.008099	0.007999	0.028497	1.000000	1.000000
PCDHAC1	0.008499	0.008499	0.063894	1.000000	1.000000
OR5H2	0.009299	0.009299	0.017198	1.000000	1.000000
NUMA1	0.009399	0.009399	0.018898	1.000000	1.000000
RAB18	0.009999	0.009999	0.034097	1.000000	1.000000
LY75	0.011099	0.011099	0.037496	1.000000	1.000000
GPRC6A	0.011199	0.012899	0.058994	1.000000	1.000000
LRIG1	0.013199	0.013199	0.030597	1.000000	1.000000
NWD1	0.013299	0.037796	0.116888	1.000000	1.000000
TIGD4	0.014599	0.014599	0.040496	1.000000	1.000000
ABCA2	0.015698	0.015698	0.034397	1.000000	1.000000
GPATCH2L	0.015998	0.015998	0.034597	1.000000	1.000000
CCDC59	0.016598	0.018398	0.035196	1.000000	1.000000
ATP11A	0.017698	0.017698	0.033697	1.000000	1.000000

Table 5.1: Top 30 genes of the BOMP results using disruptive (DIS) variants only. Disruptive variants include stop loss, stop gain, exonic splicing and frameshift variants. NAME=gene name, BOMP=p-value of BOMP statistic, BOMP_B=p-value of BOMP burden statistic, BOMP_P=p-value of BOMP position distribution statistic, Bonferroni=corrected family-wise p-value of BOMP statistic and BH-FDR=Benjamini-Hochberg false discovery rate at p-value of BOMP statistic.

Name	BOMP	BOMP_B	BOMP_P	Bonferroni	BH-FDR
TYRO3	0.000028	0.000030	0.000283	0.347648	0.347648
ZNF677	0.000660	0.000660	0.001230	1.000000	1.000000
IFT81	0.001100	0.001400	0.002300	1.000000	1.000000
RBMX	0.001100	0.001100	0.003700	1.000000	1.000000
PKHD1	0.001100	0.043196	0.000300	1.000000	1.000000
ACADS	0.001100	0.001100	0.047995	1.000000	1.000000
DDOST	0.002100	0.005899	0.008299	1.000000	1.000000
LY75-CD302	0.002500	0.010999	0.003200	1.000000	1.000000
MUC6	0.002700	0.009699	0.003400	1.000000	1.000000
TRPV1	0.002800	0.002800	0.423658	1.000000	1.000000
ETFB	0.002900	0.003000	0.011199	1.000000	1.000000
FHL2	0.003400	0.003400	0.297170	1.000000	1.000000
FAM81B	0.003500	0.112689	0.001700	1.000000	1.000000
XYLB	0.004400	0.013299	0.014499	1.000000	1.000000
MCTP2	0.004700	0.002600	0.603340	1.000000	1.000000
BRCA2	0.005799	0.036596	0.006999	1.000000	1.000000
ACOX3	0.005999	0.004200	0.402060	1.000000	1.000000
GEMIN5	0.005999	0.441156	0.001100	1.000000	1.000000
FBXW5	0.006799	0.032097	0.025097	1.000000	1.000000
GLB1L	0.006799	0.009099	0.066593	1.000000	1.000000
CBX8	0.007199	0.010099	0.014199	1.000000	1.000000
CLEC7A	0.007299	0.007299	0.014399	1.000000	1.000000
PCDHAC1	0.007299	0.007299	0.062194	1.000000	1.000000
TMCO4	0.007299	0.040896	0.004300	1.000000	1.000000
PFAS	0.007499	0.021098	0.018098	1.000000	1.000000
OR5H2	0.007599	0.007599	0.015498	1.000000	1.000000
LARP7	0.007899	0.016198	0.013499	1.000000	1.000000
RASAL2	0.007999	0.033997	0.066593	1.000000	1.000000
TRMT44	0.008099	0.008099	0.277172	1.000000	1.000000
DEPDC1	0.008199	0.033597	0.008199	1.000000	1.000000

Table 5.2: Top 30 genes of the BOMP results using non-synonymous strict (NSS) variants only. Non-synonymous strict variants include stop loss, stop gain, exonic splicing and frameshift variants, and missense variants predicted as damaging by all of the bioinformatics tools used in the analysis: PolyPhen 2, SIFT, Mutationtaster and VEST. NAME=gene name, BOMP=p-value of BOMP statistic, BOMP_B=p-value of BOMP burden statistic, BOMP_P=p-value of BOMP position distribution statistic, Bonferroni=corrected family-

wise p-value of BOMP statistic and BH-FDR=Benjamini-Hochberg false discovery rate at p-value of BOMP statistic.

Name	BOMP	BOMP_B	BOMP_P	Bonferroni	BH-FDR
GALNT15	0.000120	0.001140	0.000540	1.000000	0.752866
C9orf9	0.000130	0.001140	0.017340	1.000000	0.752866
OR1B1	0.000140	0.000410	0.005910	1.000000	0.752866
TYRO3	0.000280	0.005080	0.000390	1.000000	0.981910
GMPS	0.000380	0.001770	0.000670	1.000000	0.981910
FAM81B	0.000490	0.190548	0.000260	1.000000	0.981910
ZNF677	0.000590	0.000590	0.001320	1.000000	0.981910
SEMA3G	0.000630	0.047150	0.000570	1.000000	0.981910
OR6K3	0.000740	0.001180	0.029130	1.000000	0.981910
SH3RF1	0.000810	0.000600	0.032330	1.000000	0.981910
MEGF11	0.000930	0.001740	0.031300	1.000000	0.981910
GFOD1	0.001100	0.001500	0.061394	1.000000	0.981910
ANKHD1-EIF4EBP3	0.001100	0.004500	0.011499	1.000000	0.981910
ABCC11	0.001100	0.102990	0.000500	1.000000	0.981910
ANKRD26	0.001100	0.000800	0.039996	1.000000	0.981910
EIF2AK4	0.001190	0.002500	0.053689	1.000000	0.981910
CACHD1	0.001190	0.145459	0.000380	1.000000	0.981910
GEMIN5	0.001200	0.153285	0.000500	1.000000	0.981910
CDC42BPG	0.001300	0.003200	0.014999	1.000000	0.981910
ZZZ3	0.001320	0.001990	0.014840	1.000000	0.981910
TMEM115	0.001400	0.003300	0.041296	1.000000	0.981910
DUOXA2	0.001400	0.001900	0.188981	1.000000	0.981910
ZNF776	0.001400	0.001400	0.007999	1.000000	0.981910
FAM212B	0.001500	0.001500	0.875212	1.000000	1.000000
PYHIN1	0.001600	0.087491	0.001000	1.000000	1.000000
USP6NL	0.001700	0.010799	0.002200	1.000000	1.000000
RBMX	0.001700	0.001700	0.004400	1.000000	1.000000
KCTD9	0.001800	0.006499	0.002600	1.000000	1.000000
CYTL1	0.001820	0.002930	0.010620	1.000000	1.000000
MAML1	0.001900	0.002100	0.039496	1.000000	1.000000

Table 5.3: Top 30 genes of the BOMP result using non-synonymous broad (NSB) variants only. Non-synonymous broad variants include stop loss, stop gain, exonic splicing and frameshift variants, and missense variants predicted as damaging by any of the bioinformatics tools used in the analysis: PolyPhen 2, SIFT, Mutationtaster and VEST. NAME=gene name, BOMP=p-value of BOMP statistic, BOMP_B=p-value of BOMP burden statistic, BOMP_P=p-value of BOMP position distribution statistic,

Bonferroni=corrected family-wise p-value of BOMP statistic and BH-FDR=Benjamini-Hochberg false discovery rate at p-value of BOMP statistic.

5.2.4 BOMP gene set results

Although it lacks a rigorous definition, a gene set refers to a combination of genes. With about 20,000 genes for consideration, the number of possible gene sets for analysis is vast. To constrain our search space, our collaborators suggested that I focus on 306 gene sets collected in SynaptomDB where each gene set contains a significant fraction of genes involved in synaptic functions [83]. Gene sets in SynaptomeDB were originally created by three pathway curation organizations: KEGG [84], BioCarta [85] and Gene Ontology [86]. The top 10 gene sets, in order of ascending BOMP p-values, from the DIS, NSS and NSB groups are shown in Table 5.4, 5.5 and 5.6 respectively.

In general, the results aligned with the findings in simulations that the BOMP position distribution statistic was much more effective for detecting the association for a gene set than the BOMP burden statistic. In the DIS group, 2 out of 306 gene sets had a p-value < 0.05 after Bonferroni correction and 7 had a Benjamini-Hochberg false discovery rate (BH-FDR) < 0.05. In the NSS group, only 1 (1 for NSB group) out of 306 gene sets had a Bonferroni-corrected p-value < 0.05 and 1 (5 for NSB group) had a BH-FDR < 0.05.

Interestingly, the set of genes associated with the microtubule cytoskeleton was the most significant gene set with Bonferroni corrected p-value and BH-FDR < 0.05 in both the DIS group and the NSS group, in which only highly damaging variants were considered. This gene set was created by Gene Ontology containing genes involved in the part of the cytoskeleton composed of microtubules and associated proteins. Microtubules are well known to play a key role in the trafficking of neurotransmitters to the synapse and recent evidence further illustrated the mechanism of neurotransmission regulated by the microtubule cytoskeleton [87]. The relationship between mental disorders such as schizophrenia and defects in the microtubule has been wildly discussed [88-90]. BOMP showed that the genes involved in the microtubule cytoskeleton are significantly associated with BP. Furthermore, four genes, PKHD1, BRCA2, NUMA1 and ABCA2, in the gene set had p-values less than 0.05. Among them, a SNP in BRCA2 was reported to be associated with BP in a case-control study [91] and ABCA2 interacts with the gene GPR50, which has been identified as a genetic risk factor for BP [92].

The genes associated with the Golgi apparatus, defined by Gene Ontology, was the only gene set with both Bonferroni corrected p-value and BH-FDR less than 0.05 when all possible damaging variants were considered. Genes in the gene set form a compound membranous cytoplasmic organelle in eukaryotic cells where proteins produced on the ribosomes of the rough endoplasmic reticulum are further processed for glycol-modification and sorting and packaging to a variety of cellular locations, which plays a key role in neurotransmitter synthesis. The Golgi apparatus was suspected to cause neurodegenerative diseases but the mechanisms remain to be clarified [93]. One recent study illustrated that mitochondria alterations in the fragmented (falling apart) Golgi apparatus in the neurons of patients with Alzheimer's disease causes the accumulation of amyloid deposits, demonstrating the synaptic pathology [94]. Interestingly, the alteration of the Golgi apparatus may be associated with alterations of microtubules, the cellular component that had a significant p-value in the DIS and NSS groups [95]. In the Golgi apparatus gene set, twelve genes had a p-value < 0.05: ATXN2, PCSK5, CHST2, TYR, DOPEY1, SLC30A5, CLASP1, CLCN3, AFTPH, SYNE1, RHOT2 and SI. Some of them have been implicated as being responsible for mental disorders including BP: ATXN2 schizophrenia [96], PCSK5 mental retardation [97], TYR and SYNE1 bipolar disorder [66,98].

#Name	BOMP	BOMP_B	BOMP_P	Bonferroni	BH-FDR
MICROTUBULE_CYTOSKELETON	0.000039	0.050447	0.000011	0.011934	0.011934
INTRINSIC_TO_PLASMA_MEMBRANE	0.000160	0.949771	0.000070	0.048960	0.017340
INTRINSIC_TO_MEMBRANE	0.000170	0.920241	0.000080	0.052019	0.017340
ORGANELLE_ORGANIZATION_AND_BIOGENESIS	0.000660	0.026710	0.002640	0.201958	0.041310
INTEGRAL_TO_PLASMA_MEMBRANE	0.000680	0.948111	0.000380	0.208078	0.041310
INTEGRAL_TO_MEMBRANE	0.000810	0.957040	0.000470	0.247858	0.041310
CYTOSKELETON	0.001090	0.176098	0.000990	0.333537	0.047648
TRANSPORT	0.003400	0.653635	0.001800	1.000000	0.113209
LIGASE_ACTIVITY	0.003600	0.011999	0.023798	1.000000	0.113209
ESTABLISHMENT_OF_LOCALIZATION	0.003700	0.384462	0.003200	1.000000	0.113209

Table 5.4: Top 10 gene sets of the BOMP results using disruptive (DIS) variants only. In consideration of disruptive variants only, BOMP gene set analysis was performed on candidate gene sets in which synaptic genes are significantly enriched. Disruptive variants include stop loss, stop gain, exonic splicing and frameshift variants. NAME=gene name, BOMP=p-value of BOMP statistic, BOMP_B=p-value of BOMP burden statistic, BOMP_P=p-value of BOMP position distribution statistic, Bonferroni=corrected family-wise p-value of BOMP statistic and BH-FDR=Benjamini-Hochberg false discovery rate at p-value of BOMP statistic.

#Name	BOMP	BOMP_B	BOMP_P	Bonferroni	BH-FDR
MICROTUBULE_CYTOSKELETON	0.000032	0.722813	0.000005	0.009792	0.009792
REGULATION_OF_GENE_EXPRESSION	0.000780	0.256907	0.000840	0.238678	0.119339
INTRINSIC_TO_PLASMA_MEMBRANE	0.002600	0.914209	0.001700	0.795520	0.206529
SECRETORY_PATHWAY	0.002700	0.286171	0.001900	0.826117	0.206529
ORGANELLE_ORGANIZATION_AND_BIOGENESIS	0.004400	0.458854	0.003800	1.000000	0.209471
IMMUNE_SYSTEM_PROCESS	0.005699	0.741026	0.003300	1.000000	0.209471
NUCLEUS	0.005699	0.937506	0.003600	1.000000	0.209471
REGULATION_OF_CELLULAR_METABOLIC_PROCESS	0.005999	0.195180	0.008099	1.000000	0.209471
CYTOSKELETON	0.006199	0.369063	0.006299	1.000000	0.209471
INTRINSIC_TO_MEMBRANE	0.007099	0.904110	0.005000	1.000000	0.209471

Table 5.5: Top 10 gene sets of the BOMP results using non-synonymous strict (NSS) variants only. In consideration of non-synonymous strict variants only, BOMP gene set analysis was performed on candidate gene sets in which synaptic genes are significantly enriched. Non-synonymous strict variants include stop loss, stop gain, exonic splicing and frameshift variants and missense variants predicted as damaging for all of the bioinformatics tools used in the analysis: PolyPhen 2, SIFT, Mutationtaster and VEST. NAME=gene name, BOMP=p-value of BOMP statistic, BOMP_B=p-value of BOMP burden statistic, BOMP_P=p-value of BOMP position distribution statistic, Bonferroni=corrected family-wise p-value of BOMP statistic and BH-FDR=Benjamini-Hochberg false discovery rate at p-value of BOMP statistic.

#Name	BOMP	BOMP_B	BOMP_P	Bonferroni	BH-FDR
GOLGI_APPARATUS	0.000050	0.827392	0.000020	0.015300	0.015300
INTRINSIC_TO_MEMBRANE	0.000250	0.776272	0.000200	0.076499	0.038250
INTEGRAL_TO_MEMBRANE	0.000490	0.918181	0.000420	0.149939	0.039015
NERVOUS_SYSTEM_DEVELOPMENT	0.000510	0.735773	0.000430	0.156058	0.039015
GOLGI_APPARATUS_PART	0.000710	0.066999	0.001170	0.217258	0.043452
PERINUCLEAR_REGION_OF_CYTOPLASM	0.001100	0.563744	0.000600	0.336566	0.056094
MICROTUBULE_CYTOSKELETON	0.002300	0.938106	0.001600	0.703730	0.100533
SECRETORY_PATHWAY	0.002900	0.735626	0.001700	0.887311	0.110914
MICROTUBULE_BASED_PROCESS	0.003900	0.626537	0.002800	1.000000	0.132587
CELL_PROJECTION	0.004800	0.307869	0.005000	1.000000	0.146865

Table 5.6: Top 10 gene sets of the BOMP results using non-synonymous broad (NSB) variants only. In consideration of non-synonymous broad variants only, BOMP gene set analysis was performed on candidate gene sets in which synaptic genes are significantly enriched. Non-synonymous broad variants include stop loss, stop gain, exonic splicing and frameshift variants, and missense variants predicted as damaging by all of the bioinformatics tools used in the analysis: PolyPhen 2, SIFT, Mutationtaster and VEST. NAME=gene name, BOMP=p-value of BOMP statistic, BOMP_B=p-value of BOMP burden statistic, BOMP_P=p-value of BOMP position distribution statistic, Bonferroni=corrected family-wise p-value of BOMP statistic and BH-FDR=Benjamini-Hochberg false discovery rate at p-value of BOMP statistic.

Chapter 6 BOMP: Future Work and Conclusion

6.1 Power increase by removing non-causal variants

A causal gene may contain both causal and non-causal variants. Including non-causal variants in the analysis often decreases statistical power. Case-control labeling is the most informative feature in terms of removing or lowering the weight of non-causal variants. Existing methods utilize this feature in different ways. The BOMP burden statistic uses it to pick the burden threshold (see 3.2.2 individual burden threshold) that maximizes the likelihood ratio. VT uses it to choose the allele frequency threshold that maximizes the z-scores. SKAT regresses the case-control label on variants, putting higher weights on variants that are likely to be causal. Many more undiscovered methods exist to calculate the gene statistic, and better utilization of the case-control labels may better separate causal variants from non-causal variants in the analysis, yielding a statistical power increase.

In addition to the case-control label, other variant features may help filter out or lower the weight of non-causal variants in association tests. Allele frequency and bioinformatics scores have been used for this purpose in several existing methods. The BOMP position distribution statistic attempts to capture the hypothesis that causal variants may be locally clustered in functional domains. Any intuition that can distinguish causal from non-causal variants would be useful to increase the statistical power of association tests in the future.

6.2 Extension to quantitative trait study

Currently, the BOMP statistic can only handle dichotomous phenotypes in a case-control study design. A natural progression is to extend the framework to identify causal genes for quantitative trait studies. The simplest approach is to transform a quantitative trait study into a case-control study: for example, taking the extreme phenotype where samples with extreme traits in both directions are labeled as either cases or controls. In the extreme phenotype study design, an individual is either a case (with a case membership of 1 and a control membership of 0) or a control (with a case membership of 0 and a control membership of 1) or neither (with a case membership of 0 and a control membership of 0). A more generalized transformation from the quantitative trait study to a case-control study is to give a soft assignment to the case-control membership for each individual. Each individual can have nonzero values for both case and control memberships. An individual's two memberships must sum to 1 and the fraction belonging to the case group and the control group depends on how extreme the individual's trait is. If the trait of a higher value is defined to be more severe, the individual with a higher value trait will have a greater percentage of the case membership and a lesser percentage of the control membership. The BOMP statistic can be calculated in the same way with few modifications listed as follows.

Equations in 3.2.3:

$$T_j^A = \sum_{k \in samples} m_k^A * Y_{g_j,k}$$

$$T_j^U = \sum_{k \in samples} m_k^U * Y_{g_j,k}$$

$$T_j^{A+U} = \sum_{k \in samples} (m_k^A + m_k^U) * Y_{g_j,k}$$

Equations in 3.3.1:

$$W_{x,j}^A = \sum_{k \in samples} m_k^A * S_{x,j,k},$$

$$W_{x,j}^U = \sum_{k \in samples} m_k^U * S_{x,j,k},$$

$$W_{x,j}^{A+U} = \sum_{k \in samples} (m_k^A + m_k^U) * S_{x,j,k}.$$

where m_k^A and m_k^U are the case and the control memberships for individual *k* respectively ($m_k^A = 0.7$ and $m_k^U = 0.3$ if the individual is 70% of the case group and 30% of the control group, for example). With these modifications, the extended version of BOMP for quantitative traits can also be evaluated using the simulated and empirical data from the Dallas Heart Study.

The assignment of the case-control memberships for each individual actually weights each sample in the association analysis. A sample with a case/control membership of 1 has the highest weight while a sample with equal case/control memberships (a membership of 0.5 for each) does not contribute to the association analysis. The choice of the assignment function depends on the distribution of quantitative traits in the study and the phenotype that is being analyzed. It is an interesting topic, but beyond the scope of this dissertation.

6.3 Network analysis

Sparsity is the major difficulty of association analyses for identifying genetic causes, greatly decreasing the power of association tests. In addition to improving the power of a method in the statistical sense, effective approaches to aggregate the sparse signals may strengthen the association between the phenotype and the genomic unit that contains the aggregated signals. Adding more samples aggregates signals at every causal variant but increases cost. Testing the association based on single genes rather than single variants, as do the methods discussed in this dissertation, aggregate variant-level signals by genes, with no additional cost or sacrifice of genomic resolution. Network analysis, which has been used commonly in cancer research, follow the same concept, aggregating gene-level signals onto a set of genes that interact with each other. I believe this type of analysis will be the next significant effort, if genetic heterogeneity behind complex phenotypes is currently underestimated. Given that no significant genes were identified in the BP project (See Chapter 5), genetic heterogeneity for certain phenotypes probably goes beyond the capability of existing gene association tests at the sample size of ~ 2000. Thus, network analysis for identifying a set of associated genes may be emphasized as more and more sequence-based association studies are conducted in the near future.

Here I propose a framework of network analysis for sequence-based association studies, identifying a set of functionally relevant genes with phenotypic associations quantified by the phenotype-gene association test. Suppose that I have a network where nodes are genes or gene products and edges are interactions between two nodes. Many kinds of networks have been constructed [99]. For example, a PPI network is built by taking proteins as nodes and physical interactions between proteins as edges. The construction of the network depends on the biological mechanisms underlying the phenotype. Additionally, by running the phenotype-gene association test, each gene has a quantitative measurement, a p-value for example, indicating the strength of its phenotypic association. To combine association strengths and biological interactions, I will first project the quantitative measurement associated with each gene onto the corresponding node in the network, and then identify a set of highly interconnected genes with significantly strong phenotypic associations. The algorithm of identifying this gene set requires further development.

6.4 Conclusion

In summary, I have developed a new method for identifying causal variants in high-throughput sequencing data from case-control studies. It is shown to have good power relative to other leading methods and can be flexibly used in a variety of realistic scenarios. The genetic architecture of most common human phenotypes is likely complex, involving variants with a wide spectrum of frequencies from rare to common. The emergence of whole-exome and genome sequencing studies promises to accelerate our ability to interrogate the genetic architecture of these phenotypes. However, a major challenge remains: how to make sense of the enormous amounts of data generated by such studies. This new method provides another useful tool in a growing toolbox for analyzing the data from such studies.

Chapter 7 Personalized Genome Interpretation

A central question in modern human genetics is how inter-individual variation impacts human phenotypes. Unprecedented technological advances will soon make whole genome DNA sequencing services available to a large number of people. However, interpreting the variant genotypes found in an individual's genome remains challenging, and is the focus of many academic, government, and commercial efforts. Here I address some limitations of state-of-the-art biomedical informatics tools to interpret genomic data, and I propose a Bayesian probabilistic model that begins to address these limitations.

7.1 State-of-art personalized genome interpretation

An individual's whole genome sequence yields 3.2 million variant genotypes on average [100]. Genome interpretation requires reducing this very large number to a more tractable list. Current informatics tools prioritize variant genotypes, using database annotations, bioinformatics function prediction, and allele frequencies. For example, the PGP's GET-Evidence pipeline [100] prioritizes non-synonymous substitution variant calls over other alterations and ranks variant calls with a heuristic point system incorporating PolyPhen-2 classifications [76], and variant allele frequencies, variant and gene annotations in multiple public databases. The "Disease Risk of Volunteers Project" informatics pipeline identifies disease-causing mutations (DMs) in the Human Gene Mutation Database [101], eliminates any variants with minor allele frequency (MAF) > 0.01, those predicted to be benign by two out of three bioinformatics classifiers, and those seen more than three times in their cohort. In both projects, short lists of putatively important risk variant genotypes identified by the pipelines are reviewed by researchers and shared with participants.

7.2 The ultimate goal of personalized genome interpretation – personal phenotype prediction

The purpose of personal genome interpretation is to understand how variant genotypes impact upon an individual's lifetime risk of specific diseases or traits. Annotating single variant genotypes is just the first step. Most human phenotypes result from a constellation of variant genotypes and non-genetic contributions. Here I shift the focus from interpretation of single variant genotypes to identifying genes and genotypes that impact the phenotype and estimating their penetrance. To my knowledge, the only previous comparable approach to this problem considered each variant genotype as an independent medical test with an associated likelihood ratio [102]. A "pre-test" probability of phenotype, based on age- and gender-based prevalence, was multiplied by a chain of likelihood ratios for each common variant, yielding a post-test probability of phenotype. In a pioneering study of the genome of a single individual, this method was used to predict the probability of 55 disease phenotypes [103]. The likelihood ratios were derived from extensive database annotations and 480 publications of cohort and case-control studies.

I present a formal Bayesian probabilistic model that for the first time integrates annotations of phenotype prevalence, both rare and common variant genotypes and disease-associated genes, and yields a single posterior probability for a phenotype of interest. I use self-reported phenotypes and medical information shared by participants in the PGP to quantitatively assess the performance of the model on a cohort of individuals. Notably, our models do not use information from the 130 members of the PGP cohort to fit or optimize parameters. However, eventually the availability of information from thousands of individuals could enable learning these parameters directly from individuals' genomes and reported phenotypes, enabling significantly better phenotype predictions.

Chapter 8 Probabilistic Model for Personal Phenotype Prediction

8.1 Overview of Bayesian network model

I designed a Bayesian model to predict whether an individual possessed a phenotype of interest, based on genome sequence and estimated prevalence (Figure 8.1). Three categories of variables are included in the model. Categorical variables (0, 1, or 2) in the first layer represent observed genotypes, limited to those with phenotype-associated variants and predicted functional variants. Real-valued variables [0,1] in the second layer represent the probability that phenotype-associated genes are functionally altered. To estimate the aggregated penetrance of the genotypes, functional alterations are grouped into four abstract categories in the third layer. The probability that each of these categories is altered depends either on high penetrance variants (Bernoulli variable S_{VH}), low penetrance variants (Bernoulli variable S_{VL}), high penetrance genes (Bernoulli variable S_{GH}), or low penetrance genes (Bernoulli variable S_{GL}). The joint distribution of S_{VH}, S_{VL}, S_{GH}, S_{GL} is used to infer the state of Bernoulli variable Y, which represents phenotype status. All equations and derivations are reported in the following sections.

The model was designed to compute the probability by integrating the most common types of identified genetic annotations – phenotype-associated variants/genes. However, most of phenotype-associated variants/genes do not have quantitative measurements for the strength of the phenotypic association, which is essential for parameter estimation and inference calculation in the probabilistic model. To overcome this practical problem, I make assumptions to reduce the number of parameters based on qualitative statements. For example, one may suggest that the nodes in the 3rd layer (S_{VH}, S_{VL},

S_{GH}, S_{GL}) can be removed and the links to the 3rd layer can be directly linked to the phenotype Y in the 4th layer, meaning that phenotype-associated variants/genes affect the phenotypic status, which is more intuitive than adding the 3rd layer. The model without the 3rd layer, however, requires the calculation of the joint distribution of all phenotype-associated variants/genes and the phenotype Y, which may require a huge number of parameters depending on the number of phenotype-associated variants/genes. Adding the 3rd layer helps to group the variants and genes into sets and to apply different inference techniques based on their qualitative properties and constraints on quantitative measurements. Other assumptions that are made to overcome the sparsity of quantitative measurement will be discussed when describing the inference calculation.

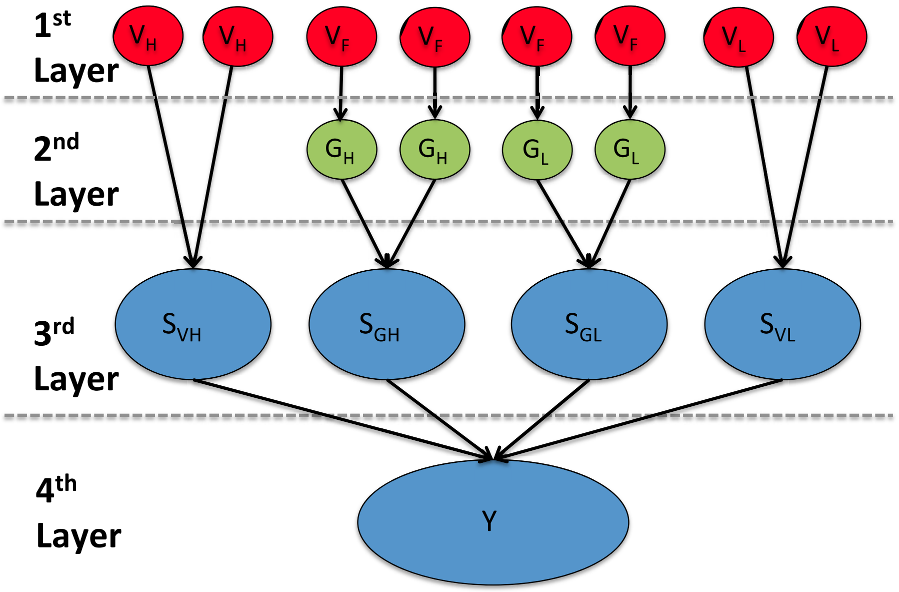


Figure 8.1: Topology of the model to predict phenotype from an individual's genome sequence. Red nodes in the first layer of the model represent the individual's genotype calls at genomic positions associated with the phenotype of interest. They are sorted into three categories: V_H (HGMD DM variants), V_L (NHGRI GWAS hits), and V_F (<0.01 MAF in any population reported in ESP6500 (http://evs.gs.washington.edu/EVS/) or 1000 Genomes [32]), found in genes annotated as associated with the phenotype. Green nodes in the second layer represent genes split into high penetrance G_H or low penetrance G_L based on database annotations. Blue nodes in the third layer are Bernoulli random variables, abstractly representing mechanisms that explain the phenotype, sorted into those altered by high penetrance variants S_{VH}, low penetrance variants S_{VL}, high penetrance genes S_{GH}, or low penetrance genes S_{GL}. The blue node Y is a Bernoulli random variable representing

individual phenotypic status. Directed edges show the dependencies between nodes. A set of model parameters is estimated for each phenotype and each individual.

8.2 Topology of the probabilistic model

The model has the same overall topology, irrespective of the phenotype predicted and the individual being assessed (Figure 8.1).

First layer. The nodes in the first layer represent observed genotypes (0, 1 or 2) from an individual's genome (homozygous reference allele, heterozygous allele, or alternate homozygous allele). Only genotypes annotated as directly associated with the phenotype are included. Genotypes are sorted into the following categories: high penetrance V_H (HGMD DM variants); low penetrance V_L (NHGRI GWAS hits); and rare (putatively functional) genotypes V_F (<0.01 MAF in any population reported in ESP6500 [104] or the 1000 Genomes Project [32]). Putatively functional genotypes are only counted if they occur in genes annotated as being associated with the phenotype.

Second layer. These nodes represent genes, split into those annotated as high penetrance G_H or low penetrance G_L. Their values depend on links to nodes in the first layer. Only genes whose translated products were bioinformatically predicted to be functionally altered by V_F genotypes are included (8.5 Functional impact of variants on phenotype-associated genes).

Third layer. These nodes are Bernoulli random variables, which represent sets of hidden mechanisms that account for the clinical phenotype. Conditional independence given an individual's genomic data is assumed. The probability that each of the nodes is set to 1 depends on the high penetrance variants (Bernoulli variable S_{VH}); the low penetrance variants (Bernoulli variable S_{VL}); the high penetrance genes (Bernoulli variable S_{GH}); and the low penetrance genes (Bernoulli variable S_{GL}), respectively. The joint distribution of S_{VH}, S_{VL}, S_{GH}, S_{GL} is used to infer the state of Bernoulli variable Y.

Fourth layer. The Bernoulli variable Y represents the phenotypic status of the individual, and the posterior probability of Y is the final output of the model.

8.3 Inference of phenotype status

The topology of the model yields the following equation for the posterior probability of an individual's phenotypic status, given genome sequence data.

$$\begin{aligned} P(Y=1 \mid Data) &= \sum P(Y=1, S_{VH}, S_{GH}, S_{GL}, S_{VL} \mid Data) \\ &= \sum P(Y=1 \mid S_{VH}, S_{GH}, S_{GL}, S_{VL}) P(S_{VH}, S_{GH}, S_{GL}, S_{VL} \mid Data) \\ &= \sum P(Y=1 \mid S_{VH}, S_{GH}, S_{GL}, S_{VL}) P(S_{VH} \mid Data) P(S_{GH} \mid Data) P(S_{GL} \mid Data) P(S_{VL} \mid Data) \end{aligned}$$

Eq 8.1

where the summation is over all possible configurations of S_{VH}, S_{GH}, S_{GL} and S_{VL}. I reduce the number of penetrance parameters by assuming that I can disregard lower penetrance genotypes if higher penetrance genotypes are present, as follows:

$$\begin{aligned} &P(Y=1 \mid S_{VH}=1, S_{GH}, S_{GL}, S_{VL}) = P(Y=1 \mid S_{VH}=1) \\ &P(Y=1 \mid S_{VH}=0, S_{GH}=1, S_{GL}, S_{VL}) = P(Y=1 \mid S_{VH}=0, S_{GH}=1) \\ &P(Y=1 \mid S_{VH}=S_{GH}=0, S_{GL}=1, S_{VL}) = P(Y=1 \mid S_{VH}=S_{GH}=0, S_{GL}=1) \end{aligned}$$

Eq 8.2

Using (Eq 8.2), (Eq 8.1) can be rewritten so that the joint distribution of S_{VH}, S_{VL}, S_{GH}, S_{GL} depends only on nine parameters.

$$\begin{aligned} &P(Y=1 \mid Data) = \\ &P(Y=1 \mid S_{VH}=1) P(S_{VH}=1 \mid Data) \\ &+P(Y=1 \mid S_{VH}=0, S_{GH}=1) P(S_{VH}=0, S_{GH}=1 \mid Data) \\ &+P(Y=1 \mid S_{VH}=S_{GH}=0, S_{GL}=1) P(S_{VH}=S_{GH}=0, S_{GL}=1 \mid Data) \\ &+P(Y=1 \mid S_{VH}=S_{GH}=S_{GL}=0, S_{VL}=1) P(S_{VH}=S_{GH}=S_{GL}=0, S_{VL}=1 \mid Data) \\ &+P(Y=1 \mid S_{VH}=S_{GH}=S_{GL}=S_{VL}=0) P(S_{VH}=S_{GH}=S_{GL}=S_{VL}=0 \mid Data) \end{aligned}$$

Eq 8.3

8.3.1 Posterior probabilities of random variables S_{VH}, S_{VL}, S_{GH}, S_{GL}

The model contains four random variables that represent sets of "hidden" biological processes impacted by four categories of variants: high penetrance variants (S_{VH}), low penetrance variants (S_{VL}), high penetrance genes (S_{GH}) and low penetrance genes (S_{GL}). I designed equations to estimate the posterior probabilities of these random variables based on individuals' variants, assuming that very little is known about the actual mechanisms by which each set of categorized variants impact the representative set of biological processes. The random variable (S_{VH}, S_{VL}, S_{GH}, S_{GL}) is set to 1, indicating that the representative set of biological processes is functionally damaged by the associated variants.

The posterior probabilities of S_{VH}, S_{VL}, S_{GH}, S_{GL} are computed as:

$$P(S_{VH}=1 \mid Data)=\begin{cases} 1, & \text{annotated variant genotype match} \\ 0, & \text{otherwise} \end{cases} \qquad \text{Eq 8.4}$$

S_{VH} represents the functional status of the set of biological processes that are affected by high penetrance variants. In Equation 8.4, it is assumed that any high penetrance variant genotype is able to cause functional damage ($P(S_{VH} = 1|Data) = 1$) due to its high penetrance.

$$P(S_{GH}=1 \mid Data)=\max\begin{cases} \max_i\{P(G_{H_i}=1 \mid Data)\}-\max_i\{E[P(G_{H_i}=1)]\} \\ 0 \end{cases} \qquad \text{Eq 8.5}$$

where $P(G_{H_i}=1 \mid Data)$ is calculated as in (Eq. 8.28) and *Data* is the VEST gene level statistic (Eq. 8.27). S_{GH} represents the functional status of the set of biological processes that are affected by high penetrance genes. I make the assumption that if there are multiple high penetrance genes, the one with the most severe damage ($max_i\{P(G_{H_i} = 1|Data)\}$) dominates the impact to the functional status. Considering the tolerance to variants, the function is impacted only when the most severe gene damage exceeds a baseline.

$$P(S_{GL}=1\,|\,Data)=1-\left[\prod_i (1-P(G_{L_i}=1\,|\,Data))\right]^{\alpha_{S_{GL}}}$$

Eq 8.6

where $P(G_{L_i}=1\,|\,Data)$ is calculated as in (Eq. 8.28) and *Data* is the VEST gene level statistic (Eq. 8.27). S_{GL} represents the functional status of the set of biological processes that are affected by low penetrance genes. If there are multiple low penetrance genes, the combined impact of $P(G_{L_i}=1\,|\,Data)$ is estimated with a noisy-OR model, exponentiated by a phenotype-specific weight ($\alpha_{S_{GL}}$), which controls for ascertainment bias (some phenotypes have hundreds of annotated low-penetrance genes while others have very few annotated low-penetrance genes) (Eq. 8.17). The noisy-or model assumes that low penetrance genes impact S_{GL} independently.

$$P(S_{VL}=1\,|\,Data)=1-\left[\prod_i OR_i^{V_{L_i}}\right]^{-\alpha_{S_{VL}}}$$

Eq 8.7

where OR_i is the odds ratio of genotype $V_{L_i} \square \{0,1,2\}$. S_{VL} represents the functional status of the set of biological processes that are affected by low penetrance variants, GWAS hits in my current implementation. I assume that the impact of low penetrance variants follow the multiplicative model, adjusted by a phenotype-specific weight ($\alpha_{S_{VL}}$), which controls for ascertainment bias (Eq. 8.19).

8.3.2 Penetrances of random variables (S_{VH}, S_{VL}, S_{GH}, S_{GL})

Penetrance of S_{VH}.

$$P(Y=1 \mid S_{VH}=1)=\begin{cases} 0.90, \text{ Homozygous variant genotype or dominant heterozygous genotype} \\ 0.45, \text{ Heterozygous genotype with unknown genetic model} \end{cases}$$

Eq 8.8

In the absence of quantitative annotations (effect size) I estimate that a homozygous variant genotype or heterozygous variant genotype (if the genetic model is dominant) has penetrance of 0.9 and a heterozygous variant genotype has penetrance of 0.45, when the genetic model is unknown to us. The penetrance estimation is somewhat arbitrary due to the absence of quantitative annotations. The estimation can be customized for the phenotypes where the genetic causality is well studied (See penetrance of ABO blood SNPs as an example at the end of this section).

Penetrance of S_{VL}:

$$P(Y=1 \mid S_{VH}=S_{GH}=S_{GL}=0, S_{VL}=1)=\frac{1}{n}\sum_{i=1}^{n} P(Y=1 \mid V_i=1)$$

Eq 8.9

where $P(Y=1 \mid V_i=1)$ is computed by (Eq. 8.25) and n is the total number of low penetrance variants associated with the phenotype.

Penetrance of S_{GH}:

$$P(Y=1 \mid S_{VH}=0, S_{GH}=1)=\frac{q \cdot P(Y=1)}{P(V=1)}$$

Eq 8.10

where q is a variable related to $k_1 \square \frac{1}{n} \square_{i=1}^{n} OR_i$ (Eqs. 8.22-25), P(Y=1) is the prevalence of the phenotype for the individual, and P(V=1) is the frequency of a rare variant, estimated as 0.01. $k_1 = 5$ based on estimates by [15] about the higher penetrance of rare vs. common variants.

Penetrance of S_{GL}:

$$P(Y=1 \mid S_{VH} = S_{GH} = 0, S_{GL} = 1) = \frac{q \square P(Y=1)}{P(V=1)}$$

Eq 8.11

where q is a variable related to $k_2 \square \frac{1}{n} \square_{i=1}^{n} OR_i$ (Eqs. 8.22-25), P(Y=1) is the prevalence of the phenotype for the individual, and P(V=1) is the frequency of a rare variant, estimated as 0.01. $k_2 = 2$ based on estimates by [15] about the higher penetrance of rare vs. common variants.

Penetrance of ABO blood SNPs.

ABO blood group is an example whose genetic components have been well studied, enabling customized penetrance estimation with better accuracy. Differences in the presence (or absence) of certain antigens on the exterior of human red blood cells define our blood type system, which is an important consideration in human blood transfusion. The ABO blood type system, the most important blood type system, was initiated from discoveries in early 1900s [105,106]. Individuals, depending on the appearance of A and B antigens on their red blood cells, were classified into four blood types, A, B, O and AB. ABO blood type is determined by a single gene, ABO, so the parents' alleles at the ABO gene determine one individual's blood type. In 1990, Yamamoto *et al.* shows that three SNPs at the ABO gene could segregate individuals with blood types A, B, O and AB [107]. Later, more polymorphisms were found in the ABO gene that might affect the presence of A and B antigens [108,109]. The combinations of 3 key SNPs (rs8176719, rs8176746 and rs8176747) with their linked ABO blood types was summarized in SNPedia to predict individuals' ABO blood types [110]. I incorporated these 3 SNPs into S_{VH} (high penetrant variant) for ABO blood type prediction and the penetrance of each pattern of the 3 SNPs regarding to ABO blood types was set based on the previous studies summarized in SNPedia (Table 8.1).

rs8176719	rs8176746	rs8176747	P(O)	P(A)	P(B)	P(AB)
-/-	T/T	G/G	0.95	0	0	0
-/-	T/T	C/G	0.95	0	0	0
-/-	T/T	CC	0.95	0	0	0
-/-	G/T	G/G	0.95	0	0	0
-/-	G/T	C/G	0.95	0	0	0
-/-	G/T	CC	0.95	0	0	0
-/-	G/G	G/G	0.95	0	0	0
-/-	G/G	C/G	0.95	0	0	0
-/-	G/G	CC	0.95	0	0	0
C/-	G/G	CC	0	0.95	0	0
C/C	G/G	CC	0	0.95	0	0
C/-	G/T	C/G	0	0.25	0.75	0
C/-	T/T	G/G	0	0	0.95	0
C/C	T/T	G/G	0	0	0.95	0
C/C	G/T	C/G	0	0	0	0.95

Table 8.1: Penetrance estimates for blood. The penetrance of each blood group is assigned based on genotype of SNPs rs8176719, rs8176746, and rs8176747. The assignment is determined according to the qualitative description of blood group determination in SNPedia [110].

8.3.3 Penetrance of unknown factors

In 8.3.2, I described the penetrance estimate when any of S_{VH}, S_{VL}, S_{GH}, or S_{GL} equals 1. The penetrance when S_{VH}, S_{VL}, S_{GH}, S_{GL} are all zeros, called the penetrance of unknown factors, is missing and will be discussed in this section. S_{VH}, S_{VL}, S_{GH}, S_{GL} are all zeros when none of the variants that would cause the phenotype occur in the individual's genome based on the incorporated genetic annotations. Unknown factors, including non-genetic components (such as environmental factors) and those genetic components that are phenotype-associated but not incorporated in the model, may contribute to the individual's phenotype susceptibility. Unless the genetic annotations used in the model can fully determine the phenotypic outcome (such as for the ABO blood group), the penetrance of unknown factors, estimated in the following, should not be zero. The estimate of the penetrance of unknown factors is derived as follows:

$$P(Y=1 \mid S_{VH}=S_{GH}=S_{GL}=S_{VL}=0)=\frac{[5]}{E[P(S_{VH}=S_{GH}=S_{GL}=S_{VL}=0)]} \tag{Eq 8.12}$$

Derivation of Eq. 8.12

$$\begin{array}{lll}
 & \text{Prevalence} & = E[P(Y=1)] = \sum P(Y=1 \mid Data)P(Data) \\
 & & = \sum P(Y=1 \mid S_{VH},S_{GH},S_{GL},S_{VL}) \cdot E[P(S_{VH},S_{GH},S_{GL},S_{VL})] \\
[1] & & = P(Y=1 \mid S_{VH}=1) \cdot E[P(S_{VH}=1)] \\
[2] & & +P(Y=1 \mid S_{VH}=0,S_{GH}=1) \cdot E[P(S_{VH}=0,S_{GH}=1)] \\
[3] & & +P(Y=1 \mid S_{VH}=S_{GH}=0,S_{GL}=1) \cdot E[P(S_{VH}=S_{GH}=0,S_{GL}=1)] \\
[4] & & +P(Y=1 \mid S_{VH}=S_{GH}=S_{GL}=0,S_{VL}=1) \cdot E[P(S_{VH}=S_{GH}=S_{GL}=0,S_{VL}=1)] \\
[5] & & +P(Y=1 \mid S_{VH}=S_{GH}=S_{GL}=S_{VL}=0) \cdot E[P(S_{VH}=S_{GH}=S_{GL}=S_{VL}=0)]
\end{array} \tag{Eq 8.13}$$

[1]+[2]+[3]+[4] is the fraction of prevalence from genetic contributions and [5] is the fraction of prevalence from other contributions (unknown factors). The ratio between

[1]+[2]+[3]+[4] and [5] can be determined by heritability if available. Otherwise, a ratio of 1 is used in this work.

I assume [1]=[2]=0 ($E[P(S_{VH}=1)]=E[P(S_{GH}=1)]=0$) and [3]=[4]. Thus, Eq. 8.14 and Eq. 8.15 can be derived. The assumption, [1]=[2]=0, indicates that the fraction of phenotype carriers who have high penetrance genetic causes is negligible in the population.

$$E[P(S_{GL}=1)]=\frac{[3]}{P(Y=1\,|\,S_{VH}=S_{GH}=0,S_{GL}=1)} \tag{Eq 8.14}$$

$$E[P(S_{GL}=0,S_{VL}=1)]=\frac{[4]}{P(Y=1\,|\,S_{VH}=S_{GH}=S_{GL}=0,S_{VL}=1)} \tag{Eq 8.15}$$

Assuming that S_{GL} and S_{VL} are independent,

$$\begin{aligned}&E[P(S_{GL}=0,S_{VL}=0)]=E[P(S_{GL}=0)]E[P(S_{VL}=0)]\\&=(1-E[P(S_{GL}=1)])(1-\frac{E[P(S_{GL}=0,S_{VL}=1)]}{1-E[P(S_{GL}=1)]})\end{aligned} \tag{Eq 8.16}$$

8.3.4 Phenotype specific weights

Posterior probabilities of S_{VL} (Eq. 8.6) and S_{GL} (Eq. 8.7) are likely to be confounded by ascertainment bias, given the wide range of annotated variants and genes available for different phenotypes (Figure 10.2). I incorporate two weights $\alpha_{S_{GL}}$ and $\alpha_{S_{VL}}$, computed with numerical optimization, to control this bias.

Derivation:

$$E[P(S_{GL}=1)] = 1 - \prod_i E[(1-P(G_{L_i}=1))^{\alpha_{S_{GL}}}]$$

Eq 8.17

Equate (Eq. 8.14) and (Eq. 8.17)

Solve for $\alpha_{S_{GL}}$.

According to (Eq. 8.15) and (Eq. 8.16)

$$E[P(S_{VL}=1)] = \frac{E[P(S_{GL}=0, S_{VL}=1)]}{1-E[P(S_{GL}=1)]}$$

Eq 8.18

and

$$\begin{aligned} E[P(S_{VL}=1)] &= 1 - \prod_i E[((OR_i)^{V_{L_i}})^{-\alpha_{S_{VL}}}] \\ &= 1 - \prod_i \sum_{j \in \{0,1,2\}} (OR_i)^{-j\alpha_{S_{VL}}} P(V_{L_i}=j) \end{aligned}$$

Eq 8.19

Equate (Eq. 8.18) and (Eq. 8.19)

Solve for $\alpha_{S_{VL}}$.

Optimization requires the following constraints for numerical stability:

$$0 \le \alpha_{S_{VL}} \le 1$$

$$0 \le \alpha_{S_{GL}} \le 1$$

Eq 8.20

To compute $\alpha_{S_{GL}}$ and $\alpha_{S_{VL}}$ in (Eq. 8.17) and (Eq. 8.19) requires estimates of expected values for the frequency of functionally impacted low penetrance genes and the odds ratios of GWAS hits associated with the phenotype. I estimated these expected values using databases of variants in general populations, the Exome Variant Server ESP6500 [104] and 1000 Genomes Project data [32]. Using the ESP6500, I find all rare variants (<1% MAF) in the selected genes and their population frequencies and compute functional impact scores (Eq. 8.28). Next, for each gene I simulate a population of 10,000 individuals, to match the frequency spectrum of rare variants in ESP6500. I assume that rare variants within a gene are not in linkage disequilibrium. I calculate $P(G_{L_i} = 1 | Data)$ for each simulated individual to estimate (Eq. 8.17). I calculate the allele frequency of each selected GWAS hit in the ESP6500 (for coding variants) and 1000 Genomes (for non-coding variants). I use the allele frequencies and the assumption of Hardy-Weinberg equilibrium, to compute (Eq. 8.19).

8.4 Penetrance of GWAS hits

For the great majority of variant genotypes, I was unable to find literature or database annotations that estimated penetrance, with respect to the associated phenotypes in my study. However a quantitative measure related to penetrance, the odds ratio, was available for most GWAS hits. I converted odds ratio to penetrance, using estimates of genotype population frequencies and phenotype prevalence, as follows:

The binary random variables V and Y represent a variant genotype and a phenotype of interest. By definition,

$$OR = \frac{P(V=1|Y=1)/(1-P(V=1|Y=1))}{P(V=1|Y=1)/(1-P(V=1|Y=0))}$$

Eq 8.21

which I rewrite by setting the numerator to q/(1-q) and the denominator to p/(1-p)

$$OR = \frac{q/(1-q)}{p/(1-p)} = \frac{q-qp}{p-qp}$$

Eq 8.22

then

$$OR \cdot p - q = (OR-1) \cdot qp$$

Eq 8.23

$$\begin{aligned} P(V=1) &= P(V=1, Y=0) + P(V=1, Y=1) \\ &= P(V=1 \mid Y=0)P(Y=0) + P(V=1 \mid Y=1)P(Y=1) \\ &= p \cdot P(Y=0) + q \cdot P(Y=1) \end{aligned} \quad \text{Eq 8.24}$$

The term $P(V = 1)$ represents the population frequency of the variant genotype V. I estimate this term by counting how often it occurs in the 1000 Genomes database of human variation. In this work, I used frequencies from the 1000 Genomes European-American population, but estimates could be improved by using a population matched to a particular individual. The term $P(Y = 1)$ represents the frequency of the phenotype, or its *prevalence*. Wherever possible, I estimated phenotype prevalence for each individual, considering her/his self-reported age, gender, and ancestry if the data are available.

Finally, solving for q, the penetrance can be computed with Bayes' rule:

$$P(Y=1 \mid V=1) = \frac{q \cdot P(Y=1)}{P(V=1)} \quad \text{Eq 8.25}$$

8.5 Functional impact of variants on phenotype-associated genes

8.5.1 Predicted functional impact of variants on gene products

I predicted the impact of variants on the protein product of a gene in a particular individual for all genes annotated as associated with the phenotype. Only rare variants were considered (MAF < 1% in ESP6500 and 1000 Genomes) because their low frequency may be the result of selection assuming that the phenotype that they are associated with decreases an individual's fitness. Then, each rare variant that caused an amino acid substitution was scored with the Variant Effect Scoring Tool (VEST) [27], yielding a score m_i. VEST is a bioinformatics tool predicting the impact of a missense mutation on the gene's function (See also 3.5.3). The logic behind it is that a missense rare variant with a high VEST score is predicted to functionally alter the phenotype-associated gene, so it is likely to cause the phenotype. Rare truncating (nonsense, nonstop, frameshift) and splice site variants d_j were assumed to have a larger impact on average than rare missense variants. These events were given a score proportional to the highest scoring amino acid substitution variant in the gene and their allele frequency AF_{dj} in the 130 PGP genomes as

$$d_j = \max_i \{m_i\} \cdot (1 - AF_{d_j})$$

Eq 8.26

I made the simplifying assumption that rare variants in a gene were not in linkage disequilibrium and were therefore independent. I used Fisher's method to combine their VEST p-values, yielding a gene-level VEST statistic

$$T_{GENE} = -2 \cdot \sum_{i=1}^{N} \ln(p_i)$$

Eq 8.27

8.5.2 Estimating the probability that a gene is functionally altered

T_{GENE}, derived in Eq. 8.27, quantifies the functional impact of N rare variants within the gene. Thus, it can be used to estimate the probability that the gene is functionally altered, by computing $P(G = 1|T_{GENE})$ where G is a Bernoulli random variable and set to 1 when the gene G is functionally altered. The probability can be calculated using Bayes' rule (Eq. 8.28) if the null distribution (the distribution of T_{GENE} given G=0), the alternative distribution (the distribution of T_{GENE} given G=1) and the prior, $P(G = 0)$, are available. Formally, the probability that a gene is functionally altered in an individual is:

$$P(G=1 \mid T_{GENE}) = \frac{P(T_{GENE} \mid G=1)P(G=1)}{P(T_{GENE} \mid G=1)P(G=1) + P(T_{GENE} \mid G=0)P(G=0)}$$

Eq 8.28

where $T_{GENE} = -2\sum_{i=1}^{N}\ln(p_i)$ and p_i is the VEST P-value of each variant *i* in the gene.

$P(T_{GENE} \mid G=1)$ and $P(T_{GENE} \mid G=0)$ are estimated with simulation, based on empirical data. I assume that a single rare functional variant in a gene is sufficient for the function of that gene's translated product to be altered. I simulate the distribution of T_{GENE} in a sample of genes having one rare functional variant and *N-1* benign variants, varying *N* from to 1 to 50. $P(T_{GENE} \mid G=1)$ is estimated by generating 10,000 functionally altered genes, each of which contains one rare functional variant randomly drawn from the HGMD DM class and *N-1* variants randomly drawn from 1000 Genomes (MAF > 0.01). $P(T_{GENE} \mid G=0)$ is estimated by generating 10,000 genes that are not functionally altered, by randomly drawn *N* variants (MAF > 0.01) from 1000 genomes. I assume a uniform prior.

$$P(G=1) = P(G=0) = 0.5$$

Eq 8.29

Chapter 9 Model Performance

9.1 Data sources

To evaluate the model performance, I need (1) individual genomes as input data, (2) phenotypic profiles of the individuals in (1) for evaluation, (3) phenotype prevalence for assigning priors, and (4) phenotype-associated variant/gene annotations for model construction. Many individual genomes are available in public databases such as 1000 genome project, dbGAP and PGP but only individual genomes in PGP have corresponding phenotypic profiles. Thus, I used the genome-phenotype data from PGP and collected priors and genetic annotations for the phenotypes listed in PGP phenotypic profiles.

9.1.1 Individual genomes

PGP collected tissue samples from the participants and created cell lines, which are cells with the ability to divide for indefinite periods, for the purpose of DNA sequencing [100]. The whole genome sequencing was done by Complete Genomics, a biotechnology company developing human genome sequencing platform [111], with periodically updated sequencing pipelines from v1.0 to v2.5 by the time this dissertation is written. I downloaded variant genotypes from 174 genomes sequenced by Complete Genomics with the 2.0 Standard pipeline, from the PGP website (http://my.pgp-hms.org) (as of 02/10/2014). Variant genotypes were obtained from the GFF format (General Feature Format, a format wildly used as a protocol for the transfer of genomic feature information) file produced by PGP's Genome-Environment-Trait-Evidence (GET-Evidence) pipeline [100]. Only variant position, reference, and alternative allele calls from the GFF file were employed. 44 genomes were excluded from consideration because they were missing a trait survey, associated age, gender or ancestry, or did not have GET-Evidence GFF files, yielding 130 genomes to be analyzed.

9.1.2 Individual phenotypes

PGP participants have the option of filling out a "traits questionnaire", consisting of 239 dichotomous phenotypes. Blood groups were also provided in "Personal Health Records" of the participants, yielding a total of 243 phenotypes. Results of the questionnaire and blood groups were downloaded from the PGP website and considered to be accurate. Of the 243 phenotypes, only 153 were reported by at least one PGP participant, and 146 also had available prevalence information.

9.1.3 Phenotype prevalence

Internet searches for information about the prevalence and heritability of each trait were performed manually. Wherever possible, I found the most relevant prevalence for an individual, considering her/his age, gender, and self-reported ancestry. Data sources included SEER (NCI), websites for CDC (http://www.cdc.gov) and HHS (http://www.hhs.gov/), and the published literature.

9.1.4 Gene and variant annotations

Variant annotations were collected from NHGRI-GWAS (https://www.genome.gov/26525384) (downloaded 09/11/2013), HGMD Professional (HGMD Pro) v.2013,2 [101] (downloaded 06/26/2013), and SNPedia [110]. Gene annotations were collected from OMIM [112] (downloaded 09/09/2013), disease-gene associations were mined from the literature (http://diseases.jensenlab.org downloaded 07/25/2013), and HGMD Pro v.2013.2 [101]. NHGRI GWAS variants were included if they had an odds-ratio (OR) or beta regression coefficient > 1 and <= 20. HGMD Pro variants were included if and only if they were in the most confident disease mutation class (DM). SNPedia was used as to identify SNPs associated with blood groups, known to be high penetrance (http://snpedia.com/index.php?title=ABO_blood_group&oldid=560223) [110]. Disease-gene associations mined from literature were included only if they were rated as high confidence by the mining algorithm. For associations from Jensen's database [113], which computes a Z-score for each disease-gene association, I required Z-score > 4.0 or ranking in the top-5 associated genes for the disease, according to Z-score. HGMD Pro genes were considered to be in the DM class if they contained at least one mutation in the DM class.

9.2 Performance evaluation

9.2.1 Evaluation metric

Each phenotypic model was assessed by its ability to correctly rank individuals in the PGP cohort, as area under the ROC curve (AUC). No cross-validation was performed because neither model topology nor parameters were estimated or optimized with information from the PGP cohort. P-values and FDR were estimated with permutation.

9.2.2 Statistical significance

I assessed models by their classification performance, as area under the ROC curve (AUC). I computed the statistical significance of AUC with permutation tests as follows. Let $\boldsymbol{Y}_{ij}$ and $\boldsymbol{M}_{ij}$ be two 130x146 matrices, where each row *i* indexes a PGP participant and each column *j* indexes a phenotype. $\boldsymbol{Y}_{ij}$ is a matrix of posterior probabilities, with respect to each PGP participant *i* having phenotype *j*. $\boldsymbol{M}_{ij}$ is a binary matrix, and each component shows the true status of PGP participant *i* with respect to phenotype *j* (0 or 1). I calculated the actual AUC for each phenotype *j* by comparing columns $\boldsymbol{Y}_{\cdot j}$ and $\boldsymbol{M}_{\cdot j}$. Next, I generated matrices $\boldsymbol{M}_{ij}^{1}$,

$\boldsymbol{M}_{ij}^{2}, \ldots, \boldsymbol{M}_{ij}^{K}$ (K=10,000), where each matrix was a random permutation of the rows of $\boldsymbol{M}_{ij}$. I constructed a null distribution of AUC statistics by calculating the AUC for each phenotype *j* using columns $\boldsymbol{Y}_{\cdot j}$ and $\boldsymbol{M}_{\cdot j}^{1}, \boldsymbol{M}_{\cdot j}^{2}, \ldots, \boldsymbol{M}_{\cdot j}^{K}$. The estimated p-value for phenotype *j* AUC is

$$\text{p-val}_{j} = \frac{\#(\text{nullAUC}_{\text{j}} \square \text{AUC}_{\text{j}})+1}{K+1}$$

Eq 9.1

The null distribution of AUC statistics was also used to compute p-values for each null AUC *k*

$$\text{Null p-val}_{j}^{(k)} = \frac{\#(\text{nullAUC}_{\text{j}} \square \text{nullAUC}_{\text{j}}^{(k)})}{K}$$

Eq 9.2

Let $\{\text{p-val}^{(1)}, \text{p-val}^{(2)}, \cdots, \text{p-val}^{(L)}\}$ be a list of p-values (Eq. 9.1) for all $L=146$ phenotypes, sorted in ascending order. Then for each p-value cutoff (at rank *l*).

$$FDR^{(l)} = \text{E}[\#\,\text{FD}(\text{p-val}^{(\mathrm{l})})]/l$$

Eq 9.3

and

$$\text{q-val}^{(l)} = min\{FDR^{(l)}, \cdots, FDR^{(L)}\}$$

Eq 9.4

9.3 Result

9.3.1 Overall performance

For each phenotype, I used the model to compute the posterior probability of each individual having that phenotype (Eq. 8.1) and ranked the 130 PGP participants accordingly.

The individual with the largest posterior probability was assigned Rank #1, the second largest Rank #2, and so forth. Then I assessed the ranking for each phenotype using area under the ROC curve (AUC) and computed the statistical significance of the AUC according to nominal p-value and false discovery rate (FDR) (9.2 Performance evaluation). Thirty-eight PGP phenotypes (26%) were predicted with area-under-the-ROC curve (AUC) > 0.7, and 23 (15.8%) of these were statistically significant ($p\text{-value} < 0.05$ and $FDR < 0.1$) (Figure 9.1). Sixty-four phenotypes were predicted as random or worse ($AUC \leq 0.5$).

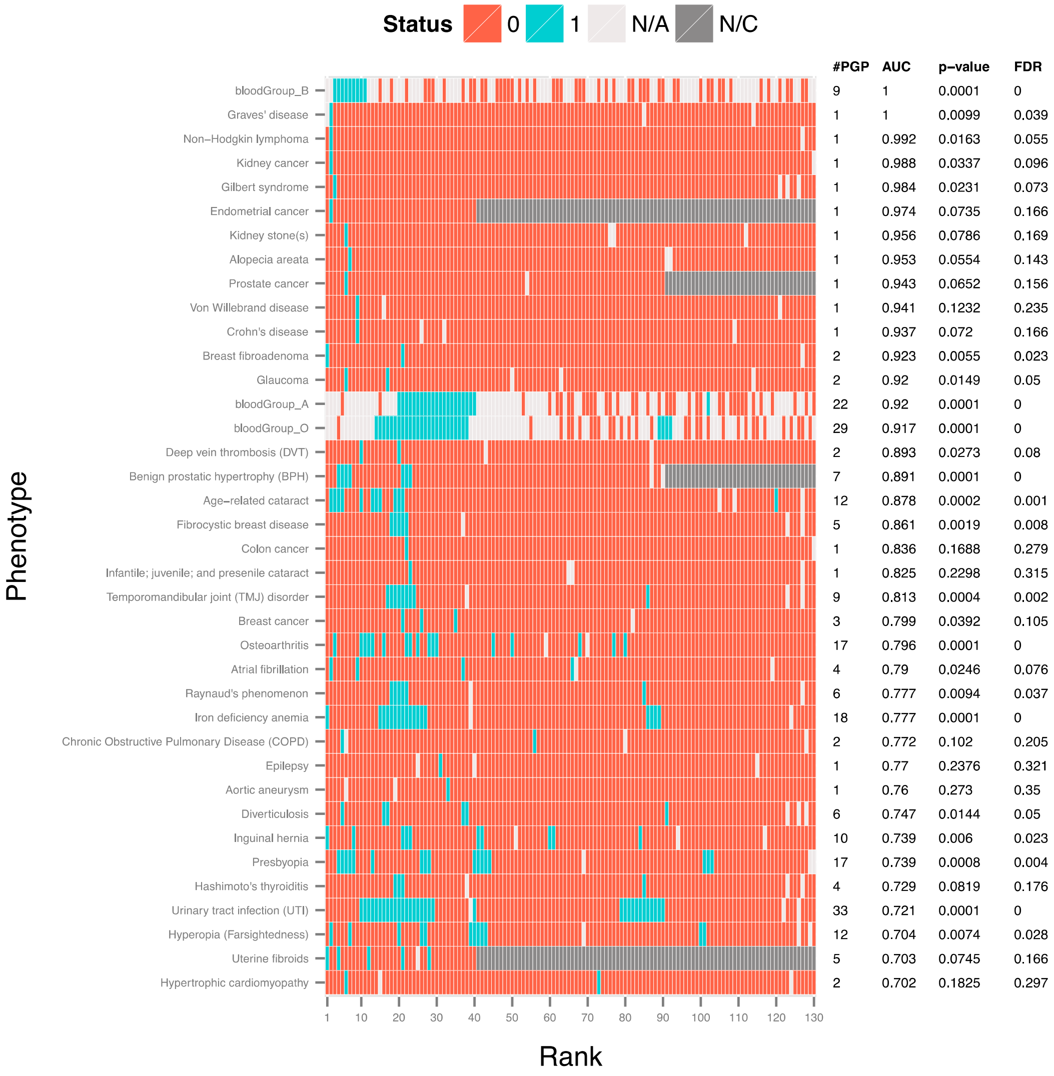


Figure 9.1: Prediction results of the model on 38 dichotomous phenotypes. Each row represents a clinical phenotype and consists of 130 cells, each of which represents a Personal Genome Project (PGP) participant. Cells in each row are ranked by the posterior probability that the participant has the phenotype. Cells are colored by true phenotypic status. Blue cells indicate that a participant has the phenotype, and red cells that a participant does not have the phenotype. If a cell is colored light grey, the true phenotypic status is unknown. If a cell is colored dark grey, the PGP participant is not considered in the evaluation because the phenotype is gender-specific. #PGP=number of participants in each row having the true

phenotypic status. AUC = area under the receiver operating characteristic curve, a threshold-free metric of classifier performance. p-value and FDR = statistical significance of the AUC value, based on permutation testing.

9.3.2 Contributions from prevalence and genome sequences

The model incorporated both genome sequence and population phenotype prevalence, and I measured the contributions of each to prediction performance. First, AUC, p-values and FDR for the top predicted 38 phenotypes were computed using each individual's estimated phenotype prevalence instead of a posterior probability. Next, I repeated these computations using the genome sequences and assigning each phenotype the same baseline prevalence, set to be the average prevalence across all phenotypes. Comparison of genome-only, prevalence-only, and combined results showed that 14 phenotypes had higher genome-only than prevalence-only AUCs (Figure 9.2). Thus, these phenotypes likely have a strong genetic component, and at least some of the underlying genes and variant genotypes are represented in the annotation databases.

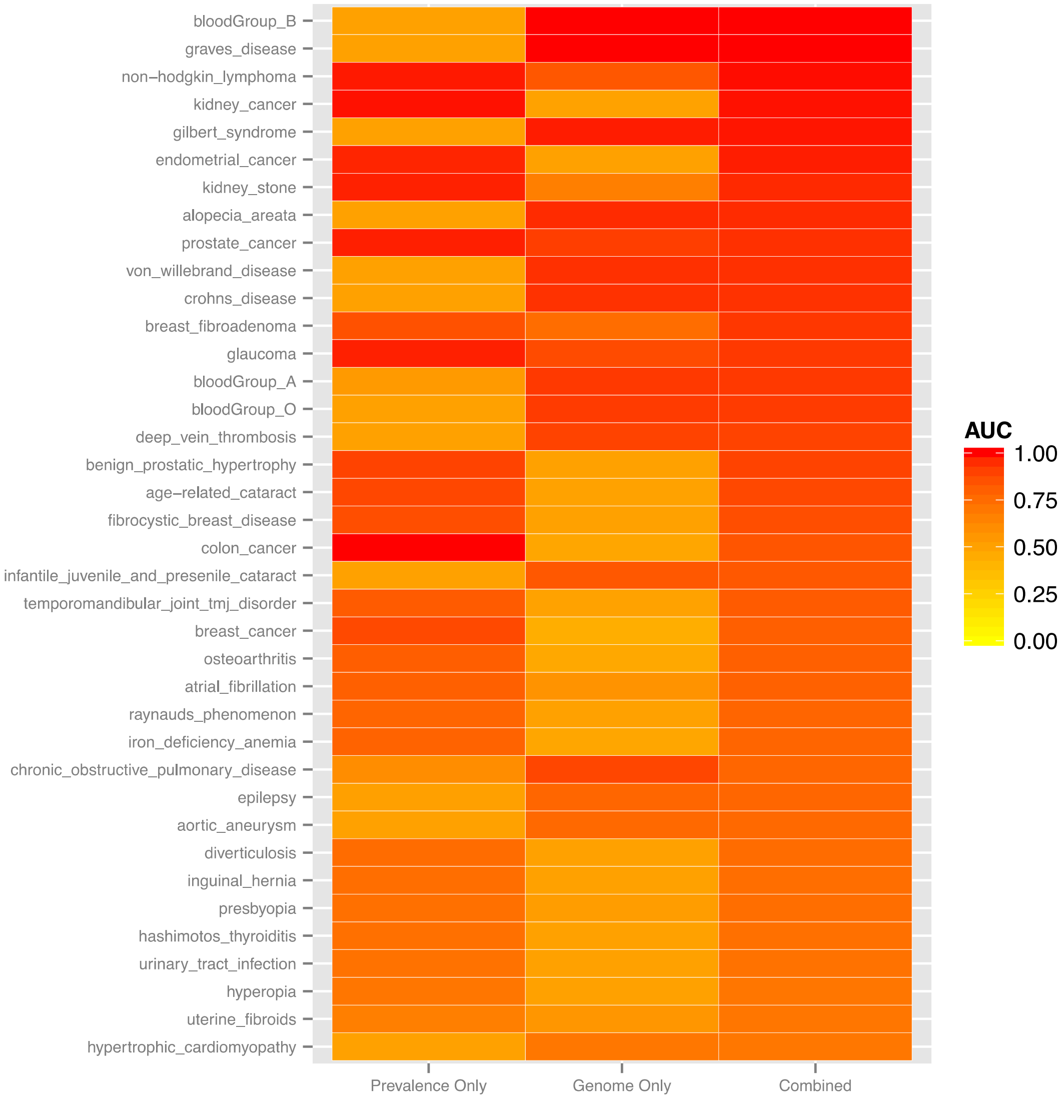


Figure 9.2: Contribution of population prevalence and genome sequence to prediction results in Fig 9.1. Each row represents a phenotype and consists of three cells, representing (a) model predictions based only on phenotype-specific population prevalence (Prevalence Only), (b) model predictions based on genome sequence (with assumption that every phenotype and every individual has the same prevalence), and (c) model predictions that combine genome sequence and phenotype-specific population prevalence. Cells are colored by the area under the ROC curve (AUC) yielded by each model. Contributions vary among

phenotypes due to differences in quality of available information with respect to prevalence and database annotations of variant genotypes.

9.3.3 Contributions from S_{VH}, S_{VL}, S_{GH}, S_{GL}

Finally, I explored whether all categories of genomic annotations -- GWAS hits, variant genotypes in disease-associated genes, and high-penetrance variant genotypes -- were useful in predicting each phenotype, by calculating the prediction performance if only one of these had been used.

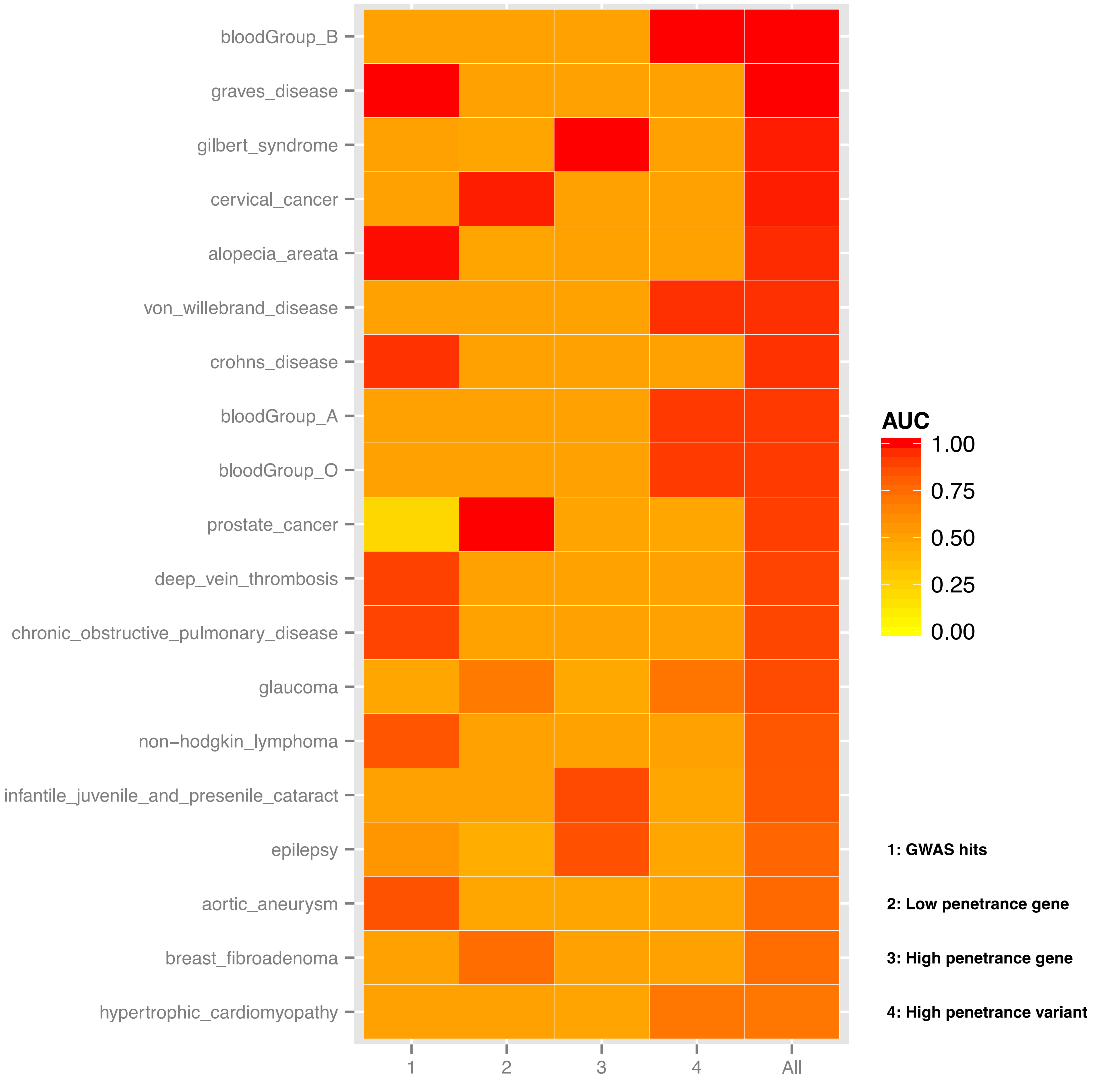


Figure 9.3: Contribution of GWAS hits, low penetrance genes, high penetrance genes, and high penetrance variants to prediction results. Each row represents a phenotype predicted with AUC>0.7 by genome sequence (Figure 9.2 (b)) and contains five cells. Cells are colored by the area under the ROC curve (AUC) yielded by a model that contains only 1:GWAS hits, 2:Low penetrance genes, 3:High penetrance genes, 4:High penetrance variants. The fifth cell shows AUC of the combination model used to assess results in this work that considers all of 1,2,3, and 4. The combination model generally yields the best performance: however, for most phenotypes, only one or two of 1, 2, 3 or 4 appears to contribute.

Six of the phenotypes were predicted best by GWAS hits -- the autoimmune disorders Graves' disease, alopecia areata and Crohn's disease; the cardiovascular disorders, deep vein thrombosis and aortic aneurism; and chronic obstructive pulmonary disease (Figure 9.3). Only one PGP participant (PGP-48) had Graves' disease, and she was ranked second out of 130 (AUC=1.0) (Figure 9.1). Her genome harbored numerous risk alleles at the sites of 16 GWAS hits (9 homozygous and 7 heterozygous risk alleles). One PGP participant (PGP-69) had alopecia areata (autoimmune-related hair loss), and he was ranked seventh out of 130 (AUC=0.953). His genome harbored 7 GWAS hits (4 homozygous and 3 heterozygous risk alleles). A complete list of PGP participants with these six phenotypes and the underlying GWAS hits are in Tables A.1-A.6.

Three phenotypes were predicted best by non-synonymous coding variants in annotated disease genes and bioinformatics variant classifications -- the common, hereditary liver disease Gilbert's syndrome, epilepsy and non-age-related cataracts (Figure 9.3). Only one of these predictions was statistically significant (Gilbert's syndrome, P=0.023 and FDR=0.073) (Figure 9.1). One PGP participant (PGP-125) reported having Gilbert's syndrome, and he was ranked third out of 130 (AUC=0.984). He had a rare, heterozygous missense mutation P229L in the Gilbert's syndrome-associated gene *UGT1A1*. Of note, with only 130 samples, if only one PGP participant had a particular phenotype, statistical significance according to our permutation test required that the model allocate them rank 1 – 4 within the cohort.

Five phenotypes were predicted best by high penetrance variant genotypes -- von Willebrand disease, hypertrophic cardiomyopathy, and three blood groups (Figure 9.3). Only the blood groups were statistically significant (Figure 9.1). The A, B and O blood groups were well represented in the 130 PGP participants, and known variant genotypes [107] ranked individuals with AUC=0.92 for group A, AUC=1 for group B, and AUC=0.917 for group O. In addition, 27 phenotypes had combined results -- genome sequence plus prevalence – better than or equal to prevalence-only AUC (Figure 9.2, Table A.7).

With a few exceptions (blood groups and Gilbert's syndrome), all of our best predicted phenotypes were complex and multi-genic. Common variants, likely involved in transcriptional regulation, and rare variants causing protein defects, both played important roles in these predictions. However, for each phenotype, the best predictions were generated by only a single category of annotations and were either GWAS hits, high penetrance variants, low penetrance genes containing rare variants, or high penetrance genes containing rare variants.

Of all the best predicted phenotypes, only glaucoma benefited from more than one category of annotations -- high penetrance variants and low penetrance genes. For this phenotype, the two PGP participants with glaucoma (PGP-15 and PGP-88) were ranked as 6 and 17 out of 130 (AUC=0.92) (Figure 9.1). PGP-15 had a glaucoma-associated high-penetrance variant in the gene *WDR36* (A449T), and PGP-88 had a rare variant (N286T) in the glaucoma-associated gene *PCMTD1*.

9.4 Critical Assessment for Genome Interpretation (CAGI) 2012-13

9.4.1 PGP challenge

In 2012-13, the Critical Assessment of Genome Interpretation (CAGI) blinded prediction experiment included a challenge based on prediction of PGP phenotypes. A total of 291 PGP participants provided phenotypic profiles, reporting their status with respect to 243 dichotomous clinical traits to the experiment organizers. Lab members and I were one of several prediction teams, who were provided both genomic data for 77 PGP participants and 291 phenotypic profiles, of which 214 were decoys. The challenge was to identify the 77 PGP participants by matching their genomes and profiles. We used the posterior probabilities of our phenotypic models to provide a rank order matching of the PGP participants and their profiles. Briefly, for each participant, the phenotypic profiles were ranked from most probable to least probable for that individual. Prediction teams were evaluated by an independent assessor based on count of correct top-ranked profiles and also by mean rank of the correct profiles for all participants.

9.4.2 Matching algorithm

We calculated a weighted Bernoulli likelihood for each pair of PGP genome *i* and phenotypic profile *k* as

$$L_{i,k} = \prod_x \left[P(D_x = 1|data)^{PS_x} P(D_x = 0|data)^{(1-PS_x)}\right]^{w_x}$$

Eq 9.5

where *x* indexes phenotypes, D_x is the predicted status of phenotype *x* for PGP genome *i*, PS_x is the phenotypic status reported in the phenotypic profile *k*, and w_x is the weight of our prediction for phenotype *x*. If any information is not available for a phenotype, for example if the phenotypic status is not reported or there is no prediction for the phenotype, we assigned 0.5^{w_x} to that phenotype in Eq. 9.1. The probability of matching the pair was calculated by normalizing the likelihoods over phenotypic profiles for each PGP genome as

$$P(i,k)_{Bern} = \frac{L_{i,k}}{\sum_k L_{i,k}}$$

Eq 9.6

w_x was estimated using a training set where possible, and was guessed otherwise. In this competition, 20 profile-matched genomes were given. We trained the weights on genomes in the training set, by maximizing the multinomial log-likelihood with regularization as

$$-\, q(D) = \left[\sum_{(i,k)\in\{matched\ genome-profile\ pairs\}} P(i,k)_{Bern}\right] - \|w\|$$

Eq 9.7

where $\|w\|$ is the norm of the weight vector. The likelihood was maximized using a greedy optimization algorithm.

9.4.3 Assessment of phenotype-genotype matching algorithms in CAGI 2012-13

Prediction accuracy was measured by an independent assessor with the following criteria. First, the number of correctly top-ranked phenotypic profiles was computed. To assess the significance of that finding, benchmark or null prediction used uniformly random matches between phenotypic profiles and genomes, i.e., for a given genome, each phenotypic profile being equally possible. The simulation was repeated 10^4 times and the number of correctly top-ranked profiles was recorded each time. In this setting, none of the simulations yielded five or more correctly top-ranked phenotypic profiles to the corresponding genomes, and hence the significance level for observing five or more correct matches is $< 10^{-4}$.

9.4.4 CAGI 2012-13 PGP result

For 27 of the 77 PGP participants, genotypic data from 23andMe was also available to the prediction teams on the PGP website, and identification of these participants was considered to be trivial. Furthermore, the website contained the critical information that no blood or saliva samples had been collected for 108 of the profile decoys, thereby making it possible to exclude these profiles as potential matches. According to the independent assessor, after elimination of the 27 participants with genotypic data and the 108 profile decoys, our team correctly predicted the largest number (six) of top-ranked participants and had the lowest mean rank for correct profiles (25.4), of the 16 submissions to the challenge. Based on an empirical null distribution, our prediction had p-value<10^{-4}.

Chapter 10 Discussion and Future Work for Personal Phenotype Prediction

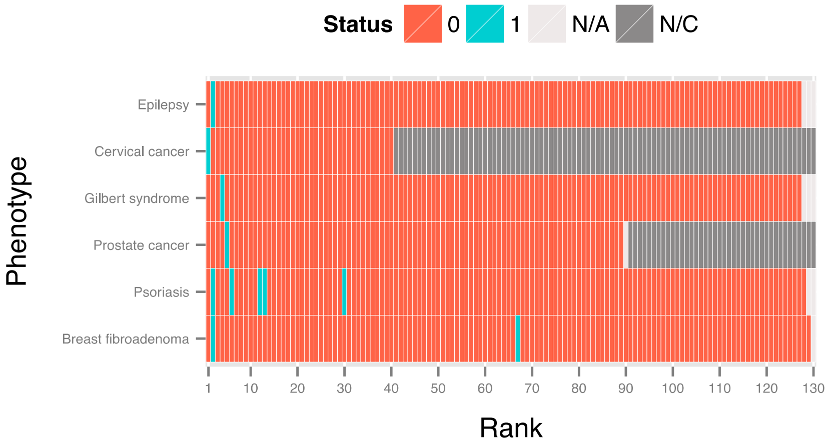
I introduce a Bayesian probabilistic model that allows individuals to estimate their risk of having a dichotomous phenotype. The models could be useful as an extension to existing pipelines for genome interpretation, such as those currently used by PGP (GET-Evidence) [100], DRV [114], UYG, and the Interpretome [115]. These pipelines rely on database annotations of variant genotypes and genes, allele frequencies and bioinformatics methods for variant function prediction. The PGP, DRV, and UYG pipelines yield lists of prioritized variant genotypes and associated evidence to support the hypothesis as to whether a single variant genotype is involved in a given disease/trait of interest. The Interpretome provides prioritized lists for rare variants and phenotype predictions based on common variants. My Bayesian probabilistic model could use any of these prioritized lists and provide phenotype predictions, which consider the contributions of both rare and common variants.

10.1 Strengths of the model

The model presented here could be used in the setting of an adult volunteer cohort. Within this setting, it provides interpretable results to help individuals understand their risk of a phenotype of interest. To our knowledge, it is the first such model to use population-level prevalence as a prior, integrate the contribution of rare and common variant genotypes harbored by an individual, and consider the modulating effects of incomplete penetrance, environmental, and unknown factors. In addition to a final posterior estimate of an individual's phenotypic risk, the model provides information about the separate contributions of population-level prevalence and personal genome sequence. Each individual can also learn their rank probability within the cohort, a number that may be easier to understand than a raw posterior probability. I can further inform individuals as to how much each prediction can be trusted, based on the model's previous performance.

The model is flexible, and it can be reasonably applied to predict any individual's probability of having any dichotomous phenotype with a genetic component. The key elements are: estimating prior probabilities that the individual has the phenotype, ideally considering age, gender and ancestry; identifying annotated genes and variant genotypes associated with the phenotype; finding the subset of those present in the individual's genome sequence and estimating their aggregate penetrance; and finally computing the posterior probability that the individual has the phenotype. Genes and variant genotypes are sorted into four categories: low penetrance variants, low penetrance genes, high penetrance variants, and high penetrance genes. The aggregate penetrance of each category is estimated with a mathematical model (Eqs. 8.8-8.11). Bioinformatics variant function predictions are also incorporated. Variant genotypes in all phenotype-associated genes are scored with VEST [27], a bioinformatics classifier that estimates a significance level (p-value) for each variant score. The p-values are aggregated into a gene-level score using Fisher's method, then used to estimate the posterior probability that the gene was affected, with empirical data. Any variant function prediction method that yields p-values and/or any of a number of gene-level variant aggregation methods can be used.

The advantage of integrating the impact of both rare and common variants can be quantified by comparing our model with a model based only on the burden of putatively damaging alleles ($MAF<0.01$) in our sets of phenotype-associated genes. When applied to the same PGP cohort, this simple burden model yielded only one predicted phenotype that was statistically significant after multiple testing correction (in contrast to my model's 23 statistically significant predicted phenotypes) (Figure 10.1).



	#PGP	AUC	p–value	FDR
Epilepsy	1	0.988	0.0304	0.258
Cervical cancer	1	0.987	0.0514	0.258
Gilbert syndrome	1	0.976	0.0552	0.258
Prostate cancer	1	0.955	0.0652	0.258
Psoriasis	5	0.92	0.0003	0.002
Breast fibroadenoma	2	0.738	0.063	0.258

Figure 10.1: Prediction results of simple mutation burden model: Six phenotypes predicted with AUC>0.7 are shown. Each row represents a clinical phenotype and consists of 130 cells, each of which represents a Personal Genome Project (PGP) participant. Cells in each row are ranked by the burden of putatively damaging alleles (MAF<0.01) in the same sets of phenotype-associated genes used in Figure 9.1. Cell coloring has the same meaning as in Figure 9.1. #PGP=number of participants in each row having the true phenotypic status. AUC = area under the receiver operating characteristic curve, a threshold-free metric of classifier performance. p-value and FDR = statistical significance of the AUC value, based on permutation testing.

10.2 Limitations of the model

Incomplete and inaccurate information about genes, variant genotypes, and phenotypes in current databases limit the model's utility. As an example, for 42 PGP phenotypes, I was unable to find any associated genes or variants. Furthermore, the association of a particular gene or variant to a phenotype may not be quantitative, with respect to effect size. Thus, I make simplifying quantitative assumptions about their aggregate penetrance, as follows: 1) GWAS hits and any disease-gene associations lacking careful curation are assigned low penetrance; 2) curated (DM) disease variants and genes in HGMD and OMIM are assigned high penetrance; 3) presence of a highly penetrant variant dominates the posterior (Eq. 8.2); 4) penetrance of a GWAS hit is estimated by its reported odds ratio, allele frequency and phenotype prevalence (Eqs. 8.21-8.25); 5) the effect sizes of rare non-silent variants are assigned to be greater than the effect sizes of GWAS hits associated with the same phenotype [10]; 6) changes in gene product function are computed using only rare (MAF<0.01) non-silent variants; 7) interactions among genes and variant genotypes are not considered; 8) low prevalence is assigned to variants and genes with high penetrance; 9) Only small-scale, non-silent variants are considered, although some phenotypes may be better predicted by other genetic or epigenetic alterations.

The phenotypes predicted in our study include those known to have strong genetic components, such as Gilbert's syndrome, von Willebrand disease and epilepsy [116-118] and others lacking evidence of strong genetic contribution, such as hiatal hernia and dental cavities [119,120]. Of 146 phenotypes, I identified associated genes or variant genotypes for 104. If I consider the raw count of annotated genes, GWAS hits, and high penetrance variants per phenotype, the range is large, with some phenotypes having thousands and others fewer than ten annotations (Figure 10.2). These differences affected our ability to predict phenotypes. For example, Gilbert's syndrome and von Willebrand disease were among our best predicted phenotypes. Both have been studied for many years and causal genes (*UGT1A1* for Gilbert's syndrome and *VWF* for von Willebrand disease) are known [116,118]. By contrast, ulcerative colitis is believed to have a genetic component but the causal genes are still largely unknown [121], and no genetic components for dandruff have been identified [122].

Out of seven cancer phenotypes, only kidney cancer and non-Hodgkins lymphoma were predicted with AUC>0.7 and high statistical significance (P-value<0.05, FDR<0.1), and these predictions were driven by population-based priors (Figure 9-2). Because most cancers have strong environmental contributions, improved predictions would require more information about carcinogen exposures and resulting patterns of somatic mutations. Our current models rely on germline variants, which may be useful for predicting familial cancers, but are less relevant for the more common sporadic cancers.

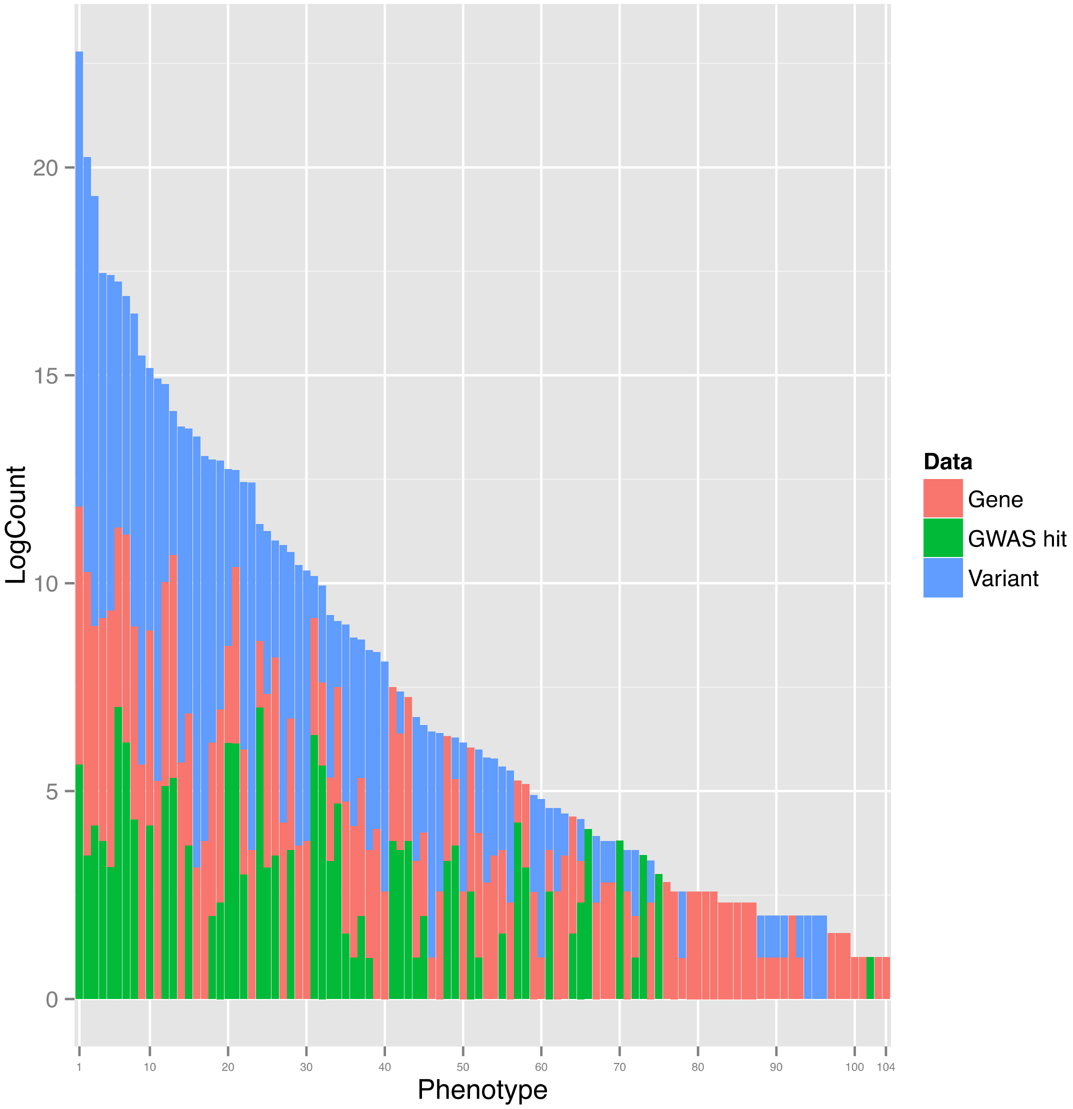


Figure 10.2: Distribution of annotated genes, GWAS hits and high penetrance variants for phenotypes analyzed in this study. Phenotypes are ordered by total counts of annotated genes and variants that I found. Counts are shown on a log scale for easier visualization. For each phenotype the (log) count of annotated genes is colored red, GWAS hits green, and high penetrance variants blue. Some phenotypes have a very large number of annotations and others have very few. For 42 phenotypes, I did not find any annotated genes or variants.

The model predictions for 64 PGP phenotypes were no better than random (Table A.8). For 38 phenotypes, either I was unable to find evidence of phenotypic association with genes or variant genotypes, or I found such evidence but was unable to match it with variants in any PGP genomes. For the remaining 26 phenotypes, I suspect that errors in annotations and in our model assumptions about penetrance are responsible. For example, my predictions of hereditary neuropathies in PGP (including Charcot-Marie-Tooth disease) yielded an AUC of 0.405. I identified 813 mutations in 37 genes associated with this phenotype in HGMD's high confidence of disease association (DM) class. Although six of these were found in the genomes of nine PGP participants, none of them reported this clinical phenotype. It appears that in assuming that HGMD DM mutations had high penetrance, I overestimated the probability that these nine individuals had a hereditary neuropathy. In addition, one PGP participant reported the phenotype but did not have any of the mutations, which could be due to our omission of the most common causes of hereditary neuropathy -- duplications or deletions of the *PMP22* (peripheral myelin protein) gene [123].

10.3 Future work

Improvements in the infrastructure of disease gene annotations and the growing communities of adult volunteers, such as the PGP, have the potential to significantly improve the utility of the model proposed here. I have discussed the many simplifying assumptions about penetrance parameters that were used in the current work. However, if a resource that provided the genomes and phenotypic profiles of a large number of people were available, I could use it for maximum likelihood estimation of the penetrance parameters in our model. Such a resource would also allow us to generate reference panels for adult genetic testing. I could use the model to compute the posterior probability of each sequenced individual for each phenotype of interest and generate ranked lists consisting of thousands of individuals. As the lists grow larger, they would also grow in utility for individuals who learn their ranking within the lists. The model could also be extended to include genomic copy number variations and even data from microbiomes.

I expect that as a larger number of individuals become interested in personal genomics, members of communities such as the PGP will have access to family pedigree information and/or genotype or sequencing data from family members. The availability of pedigree information would allow me to estimate a personalized phenotype prior for each individual, rather than estimating these priors only by population prevalence. Numerous methods have been developed for this purpose [124-126]. Genotype or sequencing data from family members could be used to improve both imputation of missing genotypes and phasing [127,128]. While phasing is not currently considered in our models, knowledge of whether multiple variants are in the same haplotype or simply on the same chromosome could be informative with respect to their phenotypic impact [129-131].

I am optimistic that integrated models such as the one presented here will contribute to increasingly accurate and interpretable predictions of clinical phenotype from genome sequence in the near future.

Chapter 11 Concluding Remark

Leveraging the recent advances in next-generation sequencing technologies, in this dissertation I proposed two probabilistic models that further facilitate genome interpretation. The first model was developed to discover the causal genes for a complex disease by contrasting the genetic background between a group of disease-affected cases and a group of healthy controls. The second model predicts personal phenotypes based on an individual's genome by integrating population prevalence, bioinformatics variant functional predictions, population allele frequency and annotated phenotype-associated variants/genes.

In the first part (chapters 2-6), a hybrid likelihood model, BOMP, was proposed to identify disease-associated genes based on sequence-based case-control studies. BOMP was developed to combine the advantages of burden tests and over-dispersion tests with the additional strength of detecting locally clustered causal variants. The model was evaluated using both simulated and empirical data. In simulation, it shows consistently good power under various disease etiologies, and even outperforms other leading methods when a set of putatively causal genes is tested, when rare causal variants are locally clustered in a gene, or when protective variants exist. In an empirical benchmark set from Dallas Heart Study, BOMP successfully identifies the set of three causal genes, with a p-value slightly more significant than other leading methods. When applied to a bipolar disorder case-control study, BOMP identified 2 significant gene sets, the microtubule cytoskeleton and the Golgi apparatus, from 306 preselected synaptic gene sets. However, it did not report any significant gene after correcting for multiple testing. This disappointing result is probably due to genetic heterogeneity. The development of higher-level integrative analysis such as gene set analysis or network analysis may be required to overcome the genetic heterogeneity behind complex diseases.

In the second part (chapter 7-10), a Bayesian network was proposed for predicting personal phenotypes based on individuals' genome sequences. It is the first probabilistic model predicting personal phenotypes using both common and rare variants on the genome sequence. To compute the posterior probability for a phenotype of interest, it uses age, gender and ancestry-specific prevalence as priors, incorporates population allele frequency and bioinformatics predictions to estimate the functional impact on disease-associated genes and integrates various types of phenotype-associated genes/variants annotated in databases. Phenotypic profiles and whole genomes of 130 individuals were downloaded from the Personal Genome Project (PGP) and evaluated the model performance on 146 clinical phenotypes. Thirty-eight PGP phenotypes (26%) were predicted with area-under-the-ROC curve (AUC) > 0.7, and 23 (15.8%) of these were statistically significant, based on permutation tests. Although currently the model does not have strong predictive power, and is far from use as a diagnostic for most of the phenotypes, there is significant improvement possible, as more and more genotype-phenotype data become available in the near future.

These two models are independent of each other, but tightly linked in achieving genome interpretation. From the perspective of a predictive framework, the first model identifies the effective predictors (disease-associated genes) in population level by searching over a huge set of possible predictors (genes on the entire genome), while the second model performs a personalized prediction by integrating the effective predictors identified in association methods like the first model. Thus, in the future, these two parts could be either improved separately for any phenotype, or better integrated and customized for a particular phenotype of interest to advance and improve genome interpretation.

APPENDX A: Glossary of terms

Association test (genetics): a statistical test that measures the strength of the association between a genomic locus and a phenotype of interest.

Case control study: a type of observational study in which two existing groups differing in outcome are identified and compared on the basis of some supposed causal attribute.

Common disease common variant (CDCV): a hypothesis, which predicts that common disease-causing alleles, or variants, will be found in all human populations that manifest a given disease.

Effect size: a quantitative measure of the strength of a phenomenon. For example, effect size of a disease susceptibility SNP can be measured in several ways such as penetrance, relative risk or odds ratio.

Etiology: the manner of causation of a disease or a condition.

Genome-wide association (GWA) study: an examination of many common genetic variants (typically SNPs) in different individuals to see if any variant is associated with a trait.

Genotype: the genetic makeup of a cell, an organism or an individual. The genotype of a single nucleotide polymorphism (SNP) in human DNA is usually expressed in the form of 0 (homozygous reference alleles), 1 (heterozygous alleles) and 2 (homozygous alternative alleles) by comparing it with the human reference genome.

Heritability: the fraction of the phenotypic variance in the population that is explained by a genetic component.

Linkage disequilibrium (LD): the non-random association of alleles at two or more loci that descend from single, ancestral chromosomes.

Odds ratio (OR): a measurement to quantify how strongly the presence (A) or absence ($\bar{A}$) of property A is associated with the presence (B) or absence ($\bar{B}$) of property B in a given population. Formally,

$$OR = \frac{P(A|B)/P(A|\bar{B})}{P(\bar{A}|B)/P(\bar{A}|\bar{B})}$$

Minor allele frequency: the frequency at which the least common allele occurs in a given population.

Multiple testing correction: a measurement to quantify the statistical significance of seeing a true positive when multiple hypotheses are tested.

P-value: a quantitative measurement of the statistical significance of an element X, formally, the probability of seeing an element equal to or more extreme than X from the null distribution.

Penetrance: a measurement to quantify how strongly the presence (A) or absence ($\bar{A}$) of property A is associated with the presence (B) or absence ($\bar{B}$) of property B in a given population. Formally,

$$penetrance = P(A|B)$$

Phenocopy: the variation in phenotype that is caused by non-genetic factors, such that the individual's phenotype matches the phenotype that is determined by genetic factors.

Quantitative trait study: a type of observational study in which phenotypes that vary in degree and can be attributed to polygenic effects, i.e., product of two or more genes, and their environment.

Relative risk (RR): a measurement to quantify how strongly the presence (A) or absence ($\bar{A}$) of property A is associated with the presence (B) or absence ($\bar{B}$) of property B in a given population. Formally,

$$RR = \frac{P(A|B)}{P(A|\bar{B})}$$

Tag SNP: a representative single nucleotide polymorphism (SNP) in a region of the genome with high linkage disequilibrium that represents a group of SNPs called a haplotype.

Single nucleotide polymorphism (SNP): a DNA sequence variation occurring commonly within a population in which a Single Nucleotide – A, T, C and G – in the genome differs between members of biological species or paired chromosomes.

APPENDX A: Supplementary tables

ID	Risk allele	PGP48 Zygosity
rs3761959	C	1
rs1024161	C	1
rs6832151	T	1
rs370409	C	2
rs4313034	C	2
rs3893464	A	2
rs9355610	A	1
rs1521	C	2
rs4947296	T	1
rs4248154	T	2
rs4713693	C	1
rs3132613	G	1
rs2273017	G	2
rs9394159	A	2
rs6457617	C	2
rs505922	C	2

Table B.1: GWAS hits correctly identified that PGP-48 has Graves' disease. A single PGP participant had Graves' disease and she was ranked second out of 130, according to the posterior probability of having this phenotype (AUC=1.0, P-value=0.01, FDR=0.039). Listed are the rsIDs of 16 GWAS hits, the risk alleles harbored by PGP-48, and the zygosity of each GWAS hit.

ID	Risk allele	PGP69 Zygosity
rs694739	G	2
rs1701704	T	1
rs1024161	C	1
rs7682241	G	1
rs9479482	C	2
rs9275572	A	2
rs10760706	T	2

Table B.2: GWAS hits correctly identified that PGP-69 has alopecia areata. A single PGP participant had alopecia areata, and he was ranked seventh out of 130, according to posterior probability (AUC=0.953, P-value=0.055, FDR=0.143). Listed are the rsIDs of 7 GWAS hits, the risk alleles harbored by PGP-69, and the zygosity of each GWAS hit.

ID	Risk allele	PGP39 Zygosity
rs1998598	A	1
rs11584383	C	1
rs1142287	C	1
rs7517810	C	1
rs9286879	A	1
rs7554511	A	1
rs6601764	T	1
rs4409764	G	1
rs1398024	G	1
rs12722489	T	1
rs11190140	C	1
rs11190141	T	1
rs12242110	A	1
rs1250550	A	1
rs1250544	A	1
rs10883365	A	1
rs7076156	A	1
rs17582416	T	1
rs7927894	C	1
rs11229030	T	1
rs7927997	C	1
rs11564258	G	1
rs11175593	C	1
rs17221417	C	1
rs151181	T	1
rs2076756	A	1
rs11871801	C	1
rs2542151	T	1
rs1893217	A	1
rs740495	A	1
rs10495903	C	1
rs13003464	A	1
rs7423615	C	1
rs13428812	A	1
rs10188217	T	1
rs762421	A	1
rs1736020	A	1
rs2838519	A	1
rs1736135	C	1

rs713875	G	1
rs2413583	T	1
rs3197999	G	1
rs9858542	G	1
rs7702331	G	1
rs3091338	C	1
rs2188962	C	1
rs11742570	T	1
rs12521868	G	1
rs6596075	G	1
rs9292777	C	1
rs2549794	T	1
rs7746082	G	1
rs17309827	G	1
rs212388	T	1
rs1847472	A	1
rs9469220	G	1
rs2301436	C	1
rs6568421	A	1
rs415890	G	1
rs7807268	C	1
rs1551398	G	1
rs4871611	G	1
rs12677663	G	1
rs4263839	A	1
rs3810936	T	1
rs4077515	C	1
rs11209026	A	2
rs4656940	G	2
rs11465804	G	2
rs2274910	T	2
rs2797685	C	2
rs1819658	T	2
rs694739	G	2
rs4902642	A	2
rs4780355	C	2
rs3091315	G	2
rs744166	G	2
rs3091316	A	2
rs2872507	G	2
rs736289	C	2
rs6545946	T	2
rs3792109	G	2

rs10210302	C	2
rs2241880	A	2
rs3828309	A	2
rs4809330	A	2
rs4820425	C	2
rs1386478	G	2
rs10045431	A	2
rs6556412	G	2
rs11167764	A	2
rs359457	C	2
rs6908425	T	2
rs1456896	C	2
rs1456893	G	2
rs6651252	C	2
rs10758669	A	2

Table B.3: GWAS hits correctly identified that PGP-39 has Crohn's disease. A single PGP participant had Crohn's disease, and she was ranked ninth out of 130, according to posterior probability (AUC=0.937, P-value=0.072, FDR=0.166). Listed are the rsIDs of 97 GWAS hits, the risk alleles harbored by PGP-39, and the zygosity of each GWAS hit.

ID	Risk allele	PGP142 Zygosity	PGP72 Zygosity
rs1018827	G	0	1
rs7659024	G	1	1
rs2519093	C	2	1
rs495828	G	2	1
rs505922	T	2	1

Table B.4: GWAS hits correctly identified that PGP-142 and PGP-72 have deep vein thrombosis. Two PGP participants had deep vein thrombosis, and they were ranked tenth and twentieth out of 130, according to posterior probability (AUC=0.893, P-value=0.027, FDR=0.08). Listed are the rsIDs of 9 GWAS hits, the risk alleles harbored by PGP-142 and PGP-72, and the zygosity of each GWAS hit.

ID	Risk allele	PGP158 Zygosity
rs1466535	A	2
rs2383207	A	2

Table B.5: GWAS hits predicted that PGP-158 has aortic aneurism. One PGP participant had aortic aneurism, and she was ranked 34 out of 130, according to posterior probability (AUC=0.76, P-value=0.273, FDR=0.35). Listed are the rsIDs of 2 GWAS hits, the risk alleles harbored by PGP-158, and the zygosity of each GWAS hit.

ID	Risk allele	PGP38 Zygosity	PGP39 Zygosity
rs114216682	C	2	2
rs117607728	T	2	2
rs8034191	T	1	1
rs8050136	C	1	2
rs76351433	C	2	2
rs10928927	T	1	1

Table B.6: GWAS hits correctly identified that PGP-39 and PGP-38 have chronic obstructive pulmonary disease (COPD). Two PGP participants had COPD, and they were ranked fifth and 56th out of 130, according to posterior probability (AUC=0.772, P-value=0.102, FDR=0.205). Listed are the rsIDs of 6 GWAS hits, the risk alleles harbored by PGP-39 and PGP-38, and the zygosity of each GWAS hit.

Phenotype	Prevalence Only	Genome Only	Combined
bloodGroup_B	0.5	1	1
graves_disease	0.5	1	1
non-hodgkin_lymphoma	0.98	0.828	0.992
kidney_cancer	0.988	0.496	0.988
gilbert_syndrome	0.5	0.976	0.984
endometrial_cancer	0.962	0.5	0.974
kidney_stone	0.968	0.663	0.956
alopecia_areata	0.5	0.953	0.953
prostate_cancer	0.972	0.909	0.943
von_willebrand_disease	0.5	0.941	0.941
crohns_disease	0.5	0.937	0.937
breast_fibroadenoma	0.85	0.742	0.923
glaucoma	0.97	0.868	0.92
bloodGroup_A	0.535	0.919	0.92
bloodGroup_O	0.499	0.915	0.917
deep_vein_thrombosis	0.5	0.893	0.893
benign_prostatic_hypertrophy	0.891	0.5	0.891
age-related_cataract	0.88	0.496	0.878
fibrocystic_breast_disease	0.861	0.5	0.861
colon_cancer	1	0.477	0.836
infantile_juvenile_and_presenile_cataract	0.5	0.825	0.825
temporomandibular_joint_tmj_disorder	0.813	0.496	0.813
breast_cancer	0.874	0.431	0.799
osteoarthritis	0.803	0.466	0.796
atrial_fibrillation	0.795	0.571	0.79
raynauds_phenomenon	0.777	0.5	0.777
iron_deficiency_anemia	0.782	0.477	0.777
chronic_obstructive_pulmonary_disease	0.598	0.884	0.772
epilepsy	0.508	0.77	0.77
aortic_aneurysm	0.5	0.76	0.76
diverticulosis	0.747	0.5	0.747
inguinal_hernia	0.739	0.5	0.739
presbyopia	0.728	0.521	0.739
hashimotos_thyroiditis	0.729	0.5	0.729
urinary_tract_infection	0.721	0.5	0.721
hyperopia	0.704	0.5	0.704
uterine_fibroids	0.662	0.556	0.703
hypertrophic_cardiomyopathy	0.5	0.702	0.702
endometriosis	0.5	0.694	0.694

hair_loss	0.65	0.649	0.685
myocardial_infarction	0.724	0.283	0.685
dupuytrens_contracture	0.5	0.685	0.685
hypertension	0.69	0.509	0.672
colon_polyps	0.684	0.517	0.67
central_serous_retinopathy	0.655	0.5	0.655
chronic_bronchitis	0.648	0.5	0.648
dry_eye_syndrome	0.653	0.496	0.646
age-related_hearing_loss	0.65	0.496	0.644
bundle_branch_block	0.654	0.472	0.642
chronic_tension_headaches	0.633	0.5	0.633
migraine_with_aura	0.564	0.552	0.632
peptic_ulcer	0.5	0.627	0.627
diabetes_mellitus_type_2	0.696	0.461	0.626
tinnitus	0.624	0.5	0.624
rheumatoid_arthritis	0.607	0.68	0.615
tennis_elbow	0.612	0.5	0.612
cardiac_arrhythmia	0.5	0.61	0.61
carpal_tunnel_syndrome	0.5	0.609	0.609
osteoporosis	0.685	0.361	0.607
migraine_without_aura	0.55	0.57	0.604
barretts_esophagus	0.5	0.595	0.595
non-melanoma_skin_cancer	0.594	0.5	0.594
hemorrhoids	0.591	0.5	0.591
scoliosis	0.608	0.435	0.591
high_triglycerides	0.579	0.565	0.585
premature_ventricular_contractions	0.572	0.5	0.572
impacted_tooth	0.5	0.57	0.57
asthma	0.572	0.522	0.566
allergic_contact_dermatitis	0.562	0.495	0.56
rotator_cuff_tear	0.559	0.5	0.559
ovarian_cysts	0.526	0.528	0.548
myopia	0.552	0.499	0.547
astigmatism	0.588	0.398	0.547
hypothyroidism	0.509	0.537	0.541
varicose_veins	0.502	0.522	0.539
gallstones	0.5	0.539	0.539
gingivitis	0.529	0.5	0.529
pilonidal_cyst	0.528	0.5	0.528
varicocele	0.465	0.667	0.527
gout	0.514	0.5	0.514
essential_tremor	0.633	0.315	0.509
geographic_tongue	0.508	0.5	0.508

trigger_finger	0.5	0.5	0.5
dermatographia	0.5	0.5	0.5
frozen_shoulder	0.5	0.5	0.5
hyperhidrosis	0.5	0.5	0.5
strabismus	0.5	0.5	0.5
tongue_tie	0.5	0.5	0.5
cafe_au_lait_spots	0.5	0.5	0.5
osgood-schlatter_disease	0.5	0.5	0.5
chronic_sinusitis	0.5	0.5	0.5
sciatica	0.5	0.5	0.5
floaters	0.5	0.5	0.5
scheuermanns_kyphosis	0.5	0.5	0.5
other_thrombophilia	0.5	0.5	0.5
dandruff	0.5	0.5	0.5
folate_deficiency_anemia	0.5	0.5	0.5
spermatocele	0.5	0.5	0.5
irritable_bowel_syndrome	0.5	0.5	0.5
rectal_prolapse	0.5	0.5	0.5
skin_tags	0.5	0.5	0.5
hiatal_hernia	0.5	0.5	0.5
plantar_fasciitis	0.5	0.5	0.5
menieres_disease	0.5	0.5	0.5
canker_sores	0.5	0.5	0.5
keloids	0.5	0.5	0.5
chronic_recurrent_tonsillitis	0.5	0.5	0.5
cluster_headaches	0.5	0.5	0.5
fissured_tongue	0.5	0.5	0.5
female_infertility	0.5	0.5	0.5
urethral_diverticulum	0.5	0.5	0.5
uterine_prolapse	0.5	0.5	0.5
nasal_polyps	0.5	0.5	0.5
deviated_septum	0.5	0.5	0.5
lichen_planus	0.5	0.496	0.496
thyroid_nodule	0.5	0.496	0.496
mitral_valve_prolapse	0.5	0.496	0.496
appendicitis	0.5	0.496	0.496
fibromyalgia	0.5	0.492	0.492
retinal_detachment	0.5	0.492	0.492
lactose_intolerance	0.492	0.5	0.492
rosacea	0.5	0.492	0.492
dental_cavities	0.491	0.5	0.491
spinal_stenosis	0.5	0.483	0.483
narcolepsy	0.5	0.482	0.482

porphyria	0.5	0.48	0.48
color_blindness	0.5	0.48	0.48
male_infertility	0.5	0.476	0.476
psoriasis	0.5	0.47	0.468
acne	0.463	0.5	0.463
bunions	0.463	0.5	0.463
lipoma	0.5	0.459	0.459
high_cholesterol	0.417	0.556	0.455
congenital_heart_defect	0.5	0.448	0.448
idiopathic_thrombocytopenic_purpura	0.444	0.5	0.444
cleft_uvula	0.5	0.437	0.437
growth_hormone_deficiency	0.487	0.407	0.435
gastroesophageal_reflux_disease	0.5	0.431	0.431
eczema	0.486	0.435	0.431
hereditary_motor_and_sensory_neuropathy	0.5	0.405	0.405
allergic_rhinitis	0.5	0.397	0.397
ulcerative_colitis	0.5	0.382	0.382
cervical_cancer	0.295	0.974	0.333
sjogrens_syndrome	0.295	0.5	0.295
polycystic_ovary_syndrome	0.5	0.284	0.284
nonalcoholic_fatty_liver_disease	0.492	0.111	0.111

Table B.7: Phenotype model prediction performance (AUC) for 130 PGP participants, using genome sequence only, prevalence only, and both. Phenotypes are sorted by the difference between AUC of the prevalence only model and the AUC of the model that uses both genome sequence and prevalence. Only phenotypes reported in at least one PGP are listed.

Phenotype	#PGP	AUC	p-value	FDR
Strabismus	1	0.500	1.0000	1.000
Osgood-Schlatter disease	1	0.500	1.0000	1.000
Other thrombophilia (includes antiphospholipid syndrome)	1	0.500	1.0000	1.000
Folate deficiency anemia	1	0.500	1.0000	1.000
Scheuermann's kyphosis	1	0.500	1.0000	1.000
Meniere's disease	1	0.500	1.0000	1.000
Chronic/recurrent tonsillitis	1	0.500	1.0000	1.000
Urethral diverticulum	1	0.500	1.0000	1.000
Dermatographia	2	0.500	1.0000	1.000
Uterine prolapse	2	0.500	1.0000	1.000
Tongue tie (ankyloglossia)	2	0.500	1.0000	1.000
Rectal prolapse	2	0.500	1.0000	1.000
Fissured tongue	2	0.500	1.0000	1.000
Trigger finger	3	0.500	1.0000	1.000
Spermatocele	3	0.500	1.0000	1.000
Frozen shoulder	3	0.500	1.0000	1.000
Female infertility	3	0.500	1.0000	1.000
Hyperhidrosis (excessive sweating)	4	0.500	1.0000	1.000
Cluster headaches	5	0.500	1.0000	1.000
Hiatal hernia	8	0.500	1.0000	1.000
Keloids	8	0.500	1.0000	1.000
Cafe au lait spots	8	0.500	1.0000	1.000
Nasal polyps	10	0.500	1.0000	1.000
Sciatica	11	0.500	1.0000	1.000
Chronic sinusitis	13	0.500	1.0000	1.000
Irritable bowel syndrome (IBS)	15	0.500	1.0000	1.000
Deviated septum	16	0.500	1.0000	1.000
Plantar fasciitis	17	0.500	1.0000	1.000
Skin tags	33	0.500	1.0000	1.000
Floaters	34	0.500	1.0000	1.000
Dandruff	44	0.500	1.0000	1.000
Canker sores (oral ulcers)	48	0.500	1.0000	1.000
Lichen planus	1	0.496	1.0000	1.000
Thyroid nodule(s)	3	0.496	1.0000	1.000
Mitral valve prolapse	6	0.496	1.0000	1.000
Appendicitis	11	0.496	1.0000	1.000
Retinal detachment	1	0.492	1.0000	1.000
Fibromyalgia	2	0.492	1.0000	1.000
Lactose intolerance	5	0.492	1.0000	1.000

Rosacea	9	0.492	1.0000	1.000
Dental cavities	109	0.491	1.0000	1.000
Spinal stenosis	7	0.483	1.0000	1.000
Narcolepsy	2	0.482	0.6223	0.669
Porphyria	1	0.480	1.0000	1.000
Color blindness	5	0.480	1.0000	1.000
Male infertility	3	0.476	1.0000	1.000
Psoriasis	5	0.468	0.5940	0.658
Acne	50	0.463	0.8593	0.831
Bunions	8	0.463	1.0000	1.000
Lipoma	8	0.459	1.0000	1.000
High cholesterol (hypercholesterolemia)	36	0.455	0.7911	0.787
Congenital heart defect	3	0.448	1.0000	1.000
Idiopathic thrombocytopenic purpura (ITP)	2	0.444	0.6555	0.684
Cleft uvula	1	0.437	0.5607	0.631
Growth hormone deficiency	3	0.435	0.6511	0.684
Gastroesophageal Reflux Disease (GERD)	3	0.431	1.0000	1.000
Eczema	20	0.431	0.8391	0.817
Hereditary motor and sensory neuropathy (includes Charcot-Marie-Tooth disease and HNPP)	1	0.405	1.0000	1.000
Allergic rhinitis (includes hay fever)	1	0.397	0.6100	0.663
Ulcerative colitis	4	0.382	0.7820	0.785
Cervical cancer	1	0.333	0.6714	0.686
Sjogren's syndrome (Sicca syndrome)	1	0.295	1.0000	1.000
Polycystic ovary syndrome (PCOS)	2	0.284	0.8360	0.817
Nonalcoholic fatty liver disease (NAFLD)	1	0.111	0.8915	0.847

Table B.8: Phenotypes that were predicted no better than random by our models. Phenotype: the poorly predicted clinical phenotypes. #PGP: the number of PGP participants who reported having the phenotype. AUC: the area under the receiver operating characteristic curve of our model predictions. p-value: the statistical significance of each AUC. FDR: the false discovery rate (FDR). None of the predictions are significant.

Bibliography

1. Hall JM, Lee MK, Newman B, Morrow JE, Anderson LA, et al. (1990) Linkage of early-onset familial breast cancer to chromosome 17q21. Science 250: 1684-1689.
2. (1995) The structure of the presenilin 1 (S182) gene and identification of six novel mutations in early onset AD families. Alzheimer's Disease Collaborative Group. Nat Genet 11: 219-222.
3. Sherrington R, Rogaev EI, Liang Y, Rogaeva EA, Levesque G, et al. (1995) Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease. Nature 375: 754-760.
4. Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K (2002) A comprehensive review of genetic association studies. Genet Med 4: 45-61.
5. International HapMap C (2005) A haplotype map of the human genome. Nature 437: 1299-1320.
6. Reich DE, Lander ES (2001) On the allelic spectrum of human disease. Trends Genet 17: 502-510.
7. Haines JL, Hauser MA, Schmidt S, Scott WK, Olson LM, et al. (2005) Complement factor H variant increases the risk of age-related macular degeneration. Science 308: 419-421.
8. Wellcome Trust Case Control C (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447: 661-678.
9. Welter D, MacArthur J, Morales J, Burdett T, Hall P, et al. (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res 42: D1001-1006.
10. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. (2009) Finding the missing heritability of complex diseases. Nature 461: 747-753.
11. Visscher PM (2008) Sizing up human height variation. Nat Genet 40: 489-490.
12. Majewski J, Schwartzentruber J, Lalonde E, Montpetit A, Jabado N (2011) What can exome sequencing do for you? J Med Genet 48: 580-589.
13. Stitziel NO, Kiezun A, Sunyaev S (2011) Computational and statistical approaches to analyzing variants identified by exome sequencing. Genome Biol 12: 227.
14. Bansal V, Libiger O, Torkamani A, Schork NJ (2010) Statistical analysis strategies for association studies involving rare variants. Nat Rev Genet 11: 773-785.
15. Bodmer W, Bonilla C (2008) Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet 40: 695-701.
16. Witte JS (2010) Genome-wide association studies and beyond. Annu Rev Public Health 31: 9-20 24 p following 20.
17. Li B, Leal SM (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet 83: 311-321.
18. Madsen BE, Browning SR (2009) A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet 5: e1000384.
19. Morgenthaler S, Thilly WG (2007) A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). Mutat Res 615: 28-56.
20. Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, et al. (2010) Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet 86: 832-838.

21. Wu MC, Lee S, Cai T, Li Y, Boehnke M, et al. (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet 89: 82-93.
22. Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, et al. (2011) Testing for an unusual distribution of rare variants. PLoS Genet 7: e1001322.
23. Liu DJ, Leal SM (2012) A flexible likelihood framework for detecting associations with secondary phenotypes in genetic studies using selected samples: application to sequence data. Eur J Hum Genet 20: 449-456.
24. Kinnamon DD, Hershberger RE, Martin ER (2012) Reconsidering association testing methods using single-variant test statistics as alternatives to pooling tests for sequence data with rare variants. PLoS One 7: e30238.
25. Bansal V, Libiger O, Torkamani A, Schork NJ (2011) An application and empirical comparison of statistical analysis methods for associating rare variants to a complex phenotype. Pac Symp Biocomput: 76-87.
26. Ladouceur M, Dastani Z, Aulchenko YS, Greenwood CM, Richards JB (2012) The empirical power of rare variant association methods: results from sanger sequencing in 1,998 individuals. PLoS Genet 8: e1002496.
27. Carter H, Douville C, Stenson PD, Cooper DN, Karchin R (2013) Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genomics 14 Suppl 3: S3.
28. Amit Y, Geman D (1997) Shape quantization and recognition with randomized trees. Neural Computation 9: 1545-1588.
29. Breiman L (2001) Random forest. Machine Learning 45: 5-32.
30. Wong WC, Kim D, Carter H, Diekhans M, Ryan MC, et al. (2011) CHASM and SNVBox: toolkit for detecting biologically important single nucleotide mutations in cancer. Bioinformatics 27: 2147-2148.
31. Stenson PD, Ball EV, Howells K, Phillips AD, Mort M, et al. (2009) The Human Gene Mutation Database: providing a comprehensive central mutation database for molecular diagnostics and personalized genomics. Hum Genomics 4: 69-72.
32. Genomes Project C, Abecasis GR, Altshuler D, Auton A, Brooks LD, et al. (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061-1073.
33. Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, et al. (2014) The UCSC Genome Browser database: 2014 update. Nucleic Acids Res 42: D764-770.
34. UniProt C (2011) Ongoing and future developments at the Universal Protein Resource. Nucleic Acids Res 39: D214-219.
35. Katzman S, Barrett C, Thiltgen G, Karchin R, Karplus K (2008) PREDICT-2ND: a tool for generalized protein local structure prediction. Bioinformatics 24: 2453-2459.
36. Romeo S, Yin W, Kozlitina J, Pennacchio LA, Boerwinkle E, et al. (2009) Rare loss-of-function mutations in ANGPTL family members contribute to plasma triglyceride levels in humans. J Clin Invest 119: 70-79.
37. Kryukov GV, Shpunt A, Stamatoyannopoulos JA, Sunyaev SR (2009) Power of deep, all-exon resequencing for discovery of human trait genes. Proc Natl Acad Sci U S A 106: 3871-3876.
38. Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, et al. (2008) Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet 4: e1000083.

39. Bustamante CD, Nielsen R, Hartl DL (2003) Maximum likelihood and Bayesian methods for estimating the distribution of selective effects among classes of mutations using DNA polymorphism data. Theor Popul Biol 63: 91-103.
40. Bustamante CD, Wakeley J, Sawyer S, Hartl DL (2001) Directional selection and the site-frequency spectrum. Genetics 159: 1779-1788.
41. Sawyer SA, Hartl DL (1992) Population genetics of polymorphism and divergence. Genetics 132: 1161-1176.
42. Hernandez RD (2008) A flexible forward simulator for populations subject to selection and demography. Bioinformatics 24: 2786-2787.
43. King CR, Rathouz PJ, Nicolae DL (2010) An evolutionary framework for association testing in resequencing studies. PLoS Genet 6: e1001202.
44. Liu DJ, Leal SM (2010) A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. PLoS Genet 6: e1001156.
45. Liu DJ, Leal SM (2010) Replication strategies for rare variant complex trait association studies via next-generation sequencing. Am J Hum Genet 87: 790-801.
46. Yi N, Liu N, Zhi D, Li J (2011) Hierarchical generalized linear models for multiple groups of rare and common variants: jointly estimating group and individual-variant effects. PLoS Genet 7: e1002382.
47. Goodwin FK JK, Akiskal H, Fawcett J, Ghaemi N, Hammen C, Ketter T, Manji H, Mondimore FM, Potash JB, Sackheim H, Weissman MM (2005) Manic-Depressive Illness: New York, Oxford University Press.
48. Mendlewicz J, Rainer JD (1977) Adoption study supporting genetic transmission in manic--depressive illness. Nature 268: 327-329.
49. Pauls DL, Bailey JN, Carter AS, Allen CR, Egeland JA (1995) Complex segregation analyses of old order Amish families ascertained through bipolar I individuals. Am J Med Genet 60: 290-297.
50. Rice J, Reich T, Andreasen NC, Endicott J, Van Eerdewegh M, et al. (1987) The familial transmission of bipolar illness. Arch Gen Psychiatry 44: 441-447.
51. Spence MA, Flodman PL, Sadovnick AD, Bailey-Wilson JE, Ameli H, et al. (1995) Bipolar disorder: evidence for a major locus. Am J Med Genet 60: 370-376.
52. Bucher KD, Elston RC, Green R, Whybrow P, Helzer J, et al. (1981) The transmission of manic depressive illness--II. Segregation analysis of three sets of family data. J Psychiatr Res 16: 65-78.
53. Goldin LR, Gershon ES, Targum SD, Sparkes RS, McGinniss M (1983) Segregation and linkage analyses in families of patients with bipolar, unipolar, and schizoaffective mood disorders. Am J Hum Genet 35: 274-287.
54. Cichon S, Schumacher J, Muller DJ, Hurter M, Windemuth C, et al. (2001) A genome screen for genes predisposing to bipolar affective disorder detects a new susceptibility locus on 8q. Hum Mol Genet 10: 2933-2944.
55. Turecki G, Grof P, Grof E, D'Souza V, Lebuis L, et al. (2001) Mapping susceptibility genes for bipolar disorder: a pharmacogenetic approach based on excellent response to lithium. Mol Psychiatry 6: 570-578.
56. Maziade M, Roy MA, Rouillard E, Bissonnette L, Fournier JP, et al. (2001) A search for specific and common susceptibility loci for schizophrenia and bipolar disorder: a linkage study in 13 target chromosomes. Mol Psychiatry 6: 684-693.

57. Liu J, Juo SH, Terwilliger JD, Grunn A, Tong X, et al. (2001) A follow-up linkage study supports evidence for a bipolar affective disorder locus on chromosome 21q22. Am J Med Genet 105: 189-194.
58. Kelsoe JR, Spence MA, Loetscher E, Foguet M, Sadovnick AD, et al. (2001) A genome survey indicates a possible susceptibility locus for bipolar disorder on chromosome 22. Proc Natl Acad Sci U S A 98: 585-590.
59. Badner JA, Gershon ES (2002) Meta-analysis of whole-genome linkage scans of bipolar disorder and schizophrenia. Mol Psychiatry 7: 405-411.
60. Segurado R, Detera-Wadleigh SD, Levinson DF, Lewis CM, Gill M, et al. (2003) Genome scan meta-analysis of schizophrenia and bipolar disorder, part III: Bipolar disorder. Am J Hum Genet 73: 49-62.
61. McQueen MB, Devlin B, Faraone SV, Nimgaonkar VL, Sklar P, et al. (2005) Combined analysis from eleven linkage studies of bipolar disorder provides strong evidence of susceptibility loci on chromosomes 6q and 8q. Am J Hum Genet 77: 582-595.
62. Baum AE, Akula N, Cabanero M, Cardona I, Corona W, et al. (2008) A genome-wide association study implicates diacylglycerol kinase eta (DGKH) and several other genes in the etiology of bipolar disorder. Mol Psychiatry 13: 197-207.
63. Sklar P, Smoller JW, Fan J, Ferreira MA, Perlis RH, et al. (2008) Whole-genome association study of bipolar disorder. Mol Psychiatry 13: 558-569.
64. Smith EN, Bloss CS, Badner JA, Barrett T, Belmonte PL, et al. (2009) Genome-wide association study of bipolar disorder in European American and African American individuals. Mol Psychiatry 14: 755-763.
65. Ferreira MA, O'Donovan MC, Meng YA, Jones IR, Ruderfer DM, et al. (2008) Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. Nat Genet 40: 1056-1058.
66. Green EK, Grozeva D, Forty L, Gordon-Smith K, Russell E, et al. (2013) Association at SYNE1 in both bipolar disorder and recurrent major depression. Mol Psychiatry 18: 614-617.
67. Pritchard JK (2001) Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet 69: 124-137.
68. Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, et al. (2004) Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science 305: 869-872.
69. Farooqi IS, Wangensteen T, Collins S, Kimber W, Matarese G, et al. (2007) Clinical and molecular genetic spectrum of congenital deficiency of the leptin receptor. N Engl J Med 356: 237-247.
70. Emison ES, McCallion AS, Kashuk CS, Bush RT, Grice E, et al. (2005) A common sex-dependent mutation in a RET enhancer underlies Hirschsprung disease risk. Nature 434: 857-863.
71. Rahimov F, Marazita ML, Visel A, Cooper ME, Hitchler MJ, et al. (2008) Disruption of an AP-2alpha binding site in an IRF6 enhancer is associated with cleft lip. Nat Genet 40: 1341-1347.
72. Romeo S, Pennacchio LA, Fu Y, Boerwinkle E, Tybjaerg-Hansen A, et al. (2007) Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. Nat Genet 39: 513-516.
73. Calton MA, Ersoy BA, Zhang S, Kane JP, Malloy MJ, et al. (2009) Association of functionally significant Melanocortin-4 but not Melanocortin-3 receptor mutations with severe adult obesity in a large North American case-control study. Hum Mol Genet 18: 1140-1147.

74. Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26: 589-595.
75. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, et al. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20: 1297-1303.
76. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, et al. (2010) A method and server for predicting damaging missense mutations. Nat Methods 7: 248-249.
77. Ng PC, Henikoff S (2003) SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res 31: 3812-3814.
78. Schwarz JM, Rodelsperger C, Schuelke M, Seelow D (2010) MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods 7: 575-576.
79. Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B-Methodological 57: 289-300.
80. Blum R, Konnerth A (2005) Neurotrophin-mediated rapid signaling in the central nervous system: mechanisms and functions. Physiology (Bethesda) 20: 70-78.
81. Pasquale EB (2005) Eph receptor signalling casts a wide net on cell behaviour. Nat Rev Mol Cell Biol 6: 462-475.
82. Prieto AL, O'Dell S, Varnum B, Lai C (2007) Localization and signaling of the receptor protein tyrosine kinase Tyro3 in cortical and hippocampal neurons. Neuroscience 150: 319-334.
83. Pirooznia M, Wang T, Avramopoulos D, Valle D, Thomas G, et al. (2012) SynaptomeDB: an ontology-based knowledgebase for synaptic genes. Bioinformatics 28: 897-899.
84. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, et al. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 27: 29-34.
85. Nishimura D (2001) BioCarta. Biotech Software & Internet Report 2: 117-120.
86. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25-29.
87. Gardiner J, Overall R, Marc J (2011) The microtubule cytoskeleton acts as a key downstream effector of neurotransmitter signaling. Synapse 65: 249-256.
88. Andrieux A, Salin PA, Job D (2004) [A role for microtubules in mental diseases?]. Pathol Biol (Paris) 52: 89-92.
89. Benitez-King G, Ramirez-Rodriguez G, Ortiz L, Meza I (2004) The neuronal cytoskeleton as a potential therapeutical target in neurodegenerative diseases and schizophrenia. Curr Drug Targets CNS Neurol Disord 3: 515-533.
90. Wong GT, Chang RC, Law AC (2013) A breach in the scaffold: the possible role of cytoskeleton dysfunction in the pathogenesis of major depression. Ageing Res Rev 12: 67-75.
91. Tesli M, Athanasiu L, Mattingsdal M, Kahler AK, Gustafsson O, et al. (2010) Association analysis of PALB2 and BRCA2 in bipolar disorder and schizophrenia in a scandinavian case-control sample. Am J Med Genet B Neuropsychiatr Genet 153B: 1276-1282.
92. Grunewald E, Tew KD, Porteous DJ, Thomson PA (2012) Developmental expression of orphan G protein-coupled receptor 50 in the mouse brain. ACS Chem Neurosci 3: 459-472.

93. Fan J, Hu Z, Zeng L, Lu W, Tang X, et al. (2008) Golgi apparatus and neurodegenerative diseases. Int J Dev Neurosci 26: 523-534.
94. Baloyannis SJ (2011) Mitochondria are related to synaptic pathology in Alzheimer's disease. Int J Alzheimers Dis 2011: 305395.
95. Baloyannis SJ (2014) Golgi apparatus and protein trafficking in Alzheimer's disease. J Alzheimers Dis 42: S153-162.
96. Zhang F, Wang G, Shugart YY, Xu Y, Liu C, et al. (2014) Association analysis of a functional variant in ATXN2 with schizophrenia. Neurosci Lett 562: 24-27.
97. Boudry-Labis E, Demeer B, Le Caignec C, Isidor B, Mathieu-Dramard M, et al. (2013) A novel microdeletion syndrome at 9q21.13 characterised by mental retardation, speech delay, epilepsy and characteristic facial features. Eur J Med Genet 56: 163-170.
98. Kumar HB, Purushottam M, Kubendran S, Gayathri P, Mukherjee O, et al. (2007) Serotonergic candidate genes and puerperal psychosis: an association study. Psychiatr Genet 17: 253-260.
99. Yu D, Kim M, Xiao G, Hwang TH (2013) Review of biological network data and its applications. Genomics Inform 11: 200-210.
100. Ball MP, Thakuria JV, Zaranek AW, Clegg T, Rosenbaum AM, et al. (2012) A public resource facilitating clinical use of genomes. Proc Natl Acad Sci U S A 109: 11920-11927.
101. Stenson PD, Mort M, Ball EV, Shaw K, Phillips AD, et al. (2013) The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum Genet.
102. Morgan AA, Chen R, Butte AJ (2010) Likelihood ratios for genome medicine. Genome Med 2: 30.
103. Ashley EA, Butte AJ, Wheeler MT, Chen R, Klein TE, et al. (2010) Clinical assessment incorporating a personal genome. Lancet 375: 1525-1535.
104. Anonymous NHLBI exome sequencing project (ESP) exome variant server.
105. Decastello Av, Sturli A (1902) Ueber die Isoagglutinine im Serum gesunder und kranker Menschen. Mfinch med Wschr 49: 1090-1095.
106. Landsteiner K (1900) Zur Kenntnis der antifermentativen, lytischen und agglutinierenden Wirkungen des Blutserums und der Lymphe. Zentralblatt Bakteriologie 27: 357-362.
107. Yamamoto F, Clausen H, White T, Marken J, Hakomori S (1990) Molecular genetic basis of the histo-blood group ABO system. Nature 345: 229-233.
108. Ogasawara K, Bannai M, Saitou N, Yabe R, Nakata K, et al. (1996) Extensive polymorphism of ABO blood group gene: three major lineages of the alleles for the common ABO phenotypes. Hum Genet 97: 777-783.
109. Seltsam A, Hallensleben M, Kollmann A, Blasczyk R (2003) The nature of diversity and diversification at the ABO locus. Blood 102: 3035-3042.
110. Cariaso M, Lennon G (2012) SNPedia: a wiki supporting personal genome annotation, interpretation and analysis. Nucleic Acids Res 40: D1308-1312.
111. Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, et al. (2010) Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science 327: 78-81.
112. McKusick VA (2007) Mendelian Inheritance in Man and its online version, OMIM. Am J Hum Genet 80: 588-604.
113. Frankild S, Jensen LJ DISEASES: Disease-gene associations mined from literature.

114. Gonzalez-Garay ML, McGuire AL, Pereira S, Caskey CT (2013) Personalized genomic disease risk of volunteers. Proc Natl Acad Sci U S A 110: 16957-16962.

115. Karczewski KJ, Tirrell RP, Cordero P, Tatonetti NP, Dudley JT, et al. (2012) Interpretome: a freely available, modular, and secure personal genome interpretation engine. Pac Symp Biocomput: 339-350.

116. Koiwai O, Nishizawa M, Hasada K, Aono S, Adachi Y, et al. (1995) Gilbert's syndrome is caused by a heterozygous missense mutation in the gene for bilirubin UDP-glucuronosyltransferase. Hum Mol Genet 4: 1183-1186.

117. Pandolfo M (2011) Genetics of epilepsy. Semin Neurol 31: 506-518.

118. Sadler JE, Ginsburg D (1993) A database of polymorphisms in the von Willebrand factor gene and pseudogene. For the Consortium on von Willebrand Factor Mutations and Polymorphisms and the Subcommittee on von Willebrand Factor of the Scientific and Standardization Committee of the International Society on Thrombosis and Haemostasis. Thromb Haemost 69: 185-191.

119. Carre IJ, Johnston BT, Thomas PS, Morrison PJ (1999) Familial hiatal hernia in a large five generation family confirming true autosomal dominant inheritance. Gut 45: 649-652.

120. Shuler CF (2001) Inherited risks for susceptibility to dental caries. J Dent Educ 65: 1038-1045.

121. Danese S, Fiocchi C (2011) Ulcerative Colitis. The New England Journal of Medicine 365: 1713-1725.

122. DeAngelis YM, Gemmer CM, Kaczvinsky JR, Kenneally DC, Schwartz JR, et al. (2005) Three etiologic facets of dandruff and seborrheic dermatitis: Malassezia fungi, sebaceous lipids, and individual sensitivity. J Investig Dermatol Symp Proc 10: 295-297.

123. Young P, Suter U (2003) The causes of Charcot-Marie-Tooth disease. Cell Mol Life Sci 60: 2547-2560.

124. Hunt SC, Williams RR, Barlow GK (1986) A comparison of positive family history definitions for defining risk of future disease. J Chronic Dis 39: 809-821.

125. Scheuner MT, Wang SJ, Raffel LJ, Larabell SK, Rotter JI (1997) Family history: a comprehensive genetic risk assessment method for the chronic conditions of adulthood. Am J Med Genet 71: 315-324.

126. Silberberg J, Fryer J, Wlodarczyk J, Robertson R, Dear K (1999) Comparison of family history measures used to identify high risk of coronary heart disease. Genet Epidemiol 16: 344-355.

127. Chen MH, Huang J, Chen WM, Larson MG, Fox CS, et al. (2012) Using family-based imputation in genome-wide association studies with large complex pedigrees: the Framingham Heart Study. PLoS One 7: e51589.

128. Meuwissen T, Goddard M (2010) The use of family relationships and linkage disequilibrium to impute phase and missing genotypes in up to whole-genome sequence density genotypic data. Genetics 185: 1441-1449.

129. Vormfelde SV, Brockmoller J (2007) On the value of haplotype-based genotype-phenotype analysis and on data transformation in pharmacogenetics and -genomics. Nat Rev Genet 8.

130. Kong A, Steinthorsdottir V, Masson G, Thorleifsson G, Sulem P, et al. (2009) Parental origin of sequence variants associated with complex diseases. Nature 462: 868-874.

131. Moskowitz SM, Chmiel JF, Sternen DL, Cheng E, Gibson RL, et al. (2008) Clinical practice and genetic counseling for cystic fibrosis and CFTR-related disorders. Genet Med 10: 851-868.

Curriculum Vitae

Yun-Ching Chen

222 E. University Parkway, Baltimore, MD, 21218

EDUCATION

2008 - present	**Ph.D.** in Biomedical Engineering, Johns Hopkins University, MD Dissertation: Methods for Genome Interpretation: Causal Gene Discovery and Personal Phenotype Prediction. Advisor: Dr. Rachel Karchin
2001 - 2003	**M.S.** in Computer Science, National Chiao Tung University, Taiwan Thesis: A parsimony-spaced suffix tree for DNA sequences. Advisor: Dr. Suh-Yin Lee
1997 - 2001	**B.S.** in Computer Science, National Chiao Tung University, Taiwan

RESEARCH EXPERIENCE

Ph.D. Candidate ***2008-Present***

Department of Biomedical Engineering and Institute of Computational Medicine, Johns Hopkins University, MD

- **Bipolar disorder sequence-based case-control study**:
    - Analyzed whole exome sequencing data for 2332 samples in collaboration with physicians and faculties in psychiatry department at Johns Hopkins.
    - Performed variant function impact analysis, disease-gene association analysis, gene set analysis and mutation burden analysis.
- **Probabilistic model for personal phenotype prediction**:
    - Developed a probabilistic model (*Bayesian network*) for predicting clinical phenotypes by integrating individual genomes, bioinformatics functional scores, phenotype prevalence, population allele frequencies and annotated disease-associated genes/variants from heterogeneous databases.
    - Evaluated model performance using 130 whole genome sequenced volunteers with their profiles of 146 clinical phenotypes in Personal Genome Project (PGP) based on Area Under ROC Curve (AUC), p-value and false discovery rate.
- **Personal Genome Project (PGP) competition**:
    - Led a team for 5 months from scratch to develop the model (see probabilistic model for personal phenotype prediction), which yielded the best prediction in Critical

  Assessment for Genome Interpretation (CAGI) 2012-13 competition (https://genomeinterpretation.org/content/PGP2012).
  - Organized teamwork for brainstorming, model design/derivation and processing/searching/parsing/integrating various types of data including 97 genomes, 4 phenotypic annotation databases and online health reports for 243 human phenotypes, Exome Variant Server (EVS) and 1000 Genomes project.
- **Hybrid likelihood ratio model for causal gene discovery**:
  - Leveraged conceptually combining two mainstreams of association tests, burden tests and overdispersion tests, to develop a *hybrid likelihood ratio model* for causal gene identification based on sequence-based association studies.
  - Illustrated the model outperforms leading methods under various conditions based on both simulated and empirical data.
  - Published the result on a high impact peer review journal (*PLoS Genetics 2013, Impact factor: 8.6*), link: http://karchinlab.org/apps/appBomp.html

Research Assistant ***2003-2008***
Institute of Information Science, Academia Sinica, Taiwan

- A web-based Expressed Sequence Tag (EST) annotation pipeline: developed/implemented a web service that manages lab users, data submission, EST analysis pipeline and result visualization (*ISRN Bioinformatics 2013*).
- ZooDDD -- a cross-species database for digital differential display analysis: developed/implemented a web service for visualizing the tissue-specific gene expression level across species using EST data (*Bioinformatics 2006*)
- *De novo* repeat identification in genomes: developed/implemented an algorithm to identify repetitive sequences in genomes without knowing repeat patterns *a priori.*

Research Assistant ***2001-2003***
National Chiao-Tung University (NCTU), Taiwan

- Developed/implemented an algorithm to build a suffix tree (a powerful data structure for string matching) for DNA sequences with less memory used (*ISMSE 03*).

JOURNAL PUBLICATIONS

- **Chen YC**[*], Parla J[*], Kramer M, Goodwin S, Deshpande P, Ethe-Sayers S, Marchica J, Ghiban E, Muller S, Karchin R and McCombie WR Microtargeting of small genomic regions in very large cohorts to overcome genetic heterogeneity in complex disorders (submitted)
- **Chen YC**, Douville C, Wang C, Niknafs N, Yeo G, Beleva-Guthrie V, Carter H, Stenson PD, Cooper DN, Li B, Mooney S, Karchin R (2014) A probabilistic model to predict clinical phenotype traits from genome sequencing PLoS Computational Biology. Sep 4. 10(9):e1003825
- **Chen YC**, Carter H, Parla J, Kramer M, Goes FS, Pirooznia M, Zandi PP, McCombie WR, Potash JB, Karchin R (2013) A hybrid likelihood model for sequence-based disease association studies PLoS Genetics. 9(1): e1003224

- Chen YC, **Chen YC**, Lin WD, Hsiao CD, Chiu HW and Ho JM (2012) Bio301: A Web-Based EST Annotation Pipeline That Facilitates Functional Comparison Studies ISRN Bioinformatics Volume 2012, Article ID 139842
- Goes FS, Rongione M, **Chen YC**, Karchin R, Elhalk E, and Potash JB (2011) Exonic DNA Sequencing of ERBB4 in Bipolar Disorder PLoS One. 6(5):e20242
- **Chen YC**, Hsiao CD, Lin WD, Hu CM, Hwang PP, Ho JM, ZooDDD: a cross-species database for digital differential display analysis Bioinformatics Advance Access, July, 2006.
- Lin WD, **Chen YC**, Ho JM, Hsiao CD, GOBU: toward to an integration interface for biological objects Journal of Information Science and Engineering, volume 22, number 1, pages 19-30, January, 2006.
- Chu SL, Weng CF, Hsiao CD, Hwang PP, **Chen YC**, Ho JM, Lee SJ, Profile analysis of expressed sequence tags derived from the ovary of tilapia, Oreochromis mossambicus Aquaculture, volume 251, pages 537-548, May, 2005.

*co-first authors

CONFERENCE PRESENTATIONS

- **Chen YC**, Carter H, Parla J, Kramer M, Goes FS, Pirooznia M, Zandi PP, McCombie WR, Potash JB, Karchin R (2013) A hybrid likelihood model for sequence-based disease association studies, ASHG meeting, 2012
- **Chen YC**, Hsiao CD, Ho JM, Huang PP, Construction of ZOO-DDD Database for Mining Kidney-Specific Markers Conserved in Diverse Vertebrate Species, 2005 West Coast Zebrafish Meeting, September, 2005.
- **Chen YC**, Hsiao CD, Ho JM, Huang PP, Annotation of Zebrafish Expressed Sequence Tags by using Enhanced EST-Ferret Pipeline, 2005 Conference on Developmental Biology, 2005.
- Hu CM, Hsiao CD, **Chen YC**, Lin WD, Ho JM, Huang PP, Establishment of Zebrafish as a Model Animal for Studying Human Disease: Construction of Diseaase - Digital Differential Display (D4) Database, 2005 Conference on Developmental Biology 2005.
- Hsiao CD, **Chen YC**, Lin WD, Shieh YE, Wang YC, Ho JM, Huang PP, Construction of gill and skin EST database (ZGSED)in zebrafish, 6th International Conference on Zebrafish Development and Genetics, July, 2004.
- **Chen YC** and Lee SY, Parsimony-spaced suffix trees for DNA sequences, ISMSE'03, November, 2003.

HONORS AND AWARDS

- Leader of the team considered as the BEST prediction at PGP challenge (*phenotype prediction based on individuals' genomes*) in Critical Assessment of Genome Interpretation (CAGI) competition, *2012-2013*
- Molly award from Critical Assessment of Genome Interpretation (CAGI) competition, *2010*
- 10 Presidential Awards, National Chiao Tung University, *1997-03*

- Member of the Phi Tau Phi Scholastic Honor Society, *2001*
- 49th and 50th Lin Hsiung-Chen Scholarship, *1999 – 2000*

INVENTIONS

- Presentation for commercial partnership: "a next-generation biomarker detection algorithm" at JHU Alliance for Science & Technology Development and UMB Commercial Advisory Board, *2014*
- Technology inventions at JHU Tech Transfer, "*A Bayesian Inference Model to Predict Phenotype from Personal Variation*", 2013
- Technology invention at JHU Tech Transfer, "*BOMP - Burden or Mutation Position. A Hybrid Likelihood Model for Sequence-based disease association studies*", 2013

TEACHING EXPERIENCE

Teaching Assistant **2012 Spring**
Course: Foundations of Computational Biology and Bioinformatics II, Department of Biomedical Engineering, Johns Hopkins University, MD
Held practical session every week (lecture & practice) and office hours, graded homework

Teaching Assistant **2010 Fall**
Course: Bioengineering III – Computational Biology, Department of Biomedical Engineering, Johns Hopkins University, MD
Held TA sessions and office hours, provided homework answer key and graded exams

Teaching Assistant **2002 Fall**
Course: Algorithm, Department of Computer Science, National Chiao Tung University, Taiwan
Held TA sessions and office hours, graded homework and exams

Teaching Assistant **2001 Fall**
Course: Data Structure, Department of Computer Science, National Chiao Tung University, Taiwan
Held TA sessions and office hours, graded homework and exams