RARE AND COMMON GENETIC VARIANT ASSOCIATIONS WITH QUANTITATIVE HUMAN PHENOTYPES

A Dissertation
Presented to
The Academic Faculty

by

Jing Zhao

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Biology

Georgia Institute of Technology
August 2015

RARE AND COMMON GENETIC VARIANT ASSOCIATIONS

WITH QUANTITATIVE HUMAN PHENOTYPES

Approved by:

Dr. Greg Gibson, Advisor
School of Biology
Georgia Institute of Technology

Dr. Joseph Lachance
School of Biology
Georgia Institute of Technology

Dr. Fredrik Vannberg
School of Biology
Georgia Institute of Technology

Dr. Eva Lee
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Patrick McGrath
School of Biology
Georgia Institute of Technology

Date Approved: June 25, 2015

To my family and friends

ACKNOWLEDGEMENTS

The path to a PhD is full of laughter and tears. During the long journey of pursuing the degree of PhD, many people have provided me help and encouragement. Among those people, thesis advisor plays a very important role in my success of graduate study and research thesis. I always feel grateful to have Dr. Greg Gibson as my PhD advisor for almost five years. I remember a lot of moments when I was frustrated with the disappointing research results, Dr. Gibson encouraged me and helped me find the right direction to solve the problems. His encouragement and support has boosted my strong belief to overcome the challenges and keep my enthusiasm for science. What is more important, his attitude about science will have long-term influence on my future career.

I appreciate the guidance and advice from my thesis committee members: Dr. Fredrik Vannberg, Dr. Patrick McGrath, Dr. Eva Lee and Dr. Joseph Lachance. I would also like to thank my previous thesis committee member Dr. Chong Shin. My PhD thesis cannot be successful without their kind guidance, advice and feedback.

I would like to thank lab manager Dalia Arafat and all previous and present lab members. Everyone in the lab has made the whole lab as a big warm family. We discuss research questions, exchange ideas, encourage and help each other. We share jokes, hang out together in spare time, as good friends. The help and friendship from them has made my PhD study life full of fun.

I would like to thank my parents for their unconditional love and support. Their love and support has made me believe in myself and have the courage to overcome all the problems I met here, a place ten thousand miles away from home. I want to thank my

husband Jinyong, whose love and patience has made me strong. I would like to thank all my dearest friends, especially Wanxue Lu, Fei He, Gena Tang, Jin Xu, Haozheng Tian, Wen Xu and Mianbo Huang, who are always accompanying me in my ups and downs. My graduate life is more colorful and warmer because of the love from my family and friends.

In addition to those individuals, I also appreciate all the help and collaborations provided by collaborators and many other people during my graduate study. I acknowledge their scientific advice and kind help.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

GWAS	Genome-wide Association Study
CAD	Coronary Artery Disease
T2D	Type 2 Diabetes
SNP	Single Nucleotide Polymorphism
TNF-α	Tumor Necrosis Factor Alpha
BMI	Body Mass Index
CRP	C-reactive Protein
eQTL	Expression Quantitative Trait Loci
eSNP	Expression Single Nucleotide Polymorphism
CHDWB	Center for Health Discovery and Well Being
ACTSI	Atlanta Clinical and Translational Science Institute
TSS	Transcription Start Site
CNV	Copy Number Variation
indel	Insertion or Deletion
NGS	Next Generation Sequencing
ChIP-sequencing	Chromatin Immunoprecipitation combined with DNA Sequencing
CAST	Cohort Allelic Sum Test
CMC	Combined Multivariate and Collapsing
WSRRT	Weighted Sum Rare allele Rank Test
SKAT	Sequence Kernel Association Test
dbGaP	the database of Genotypes and Phenotypes
CDC	Centers for Disease Control and Prevention
LR	Likelihood Ratio

FHS	Framingham risk Score
TG	Triglyceride
GxE	Gene-environment interaction
QC	Quality Control
EUR	European
CAU	Caucasian
ASN	Asian
AFR	African
MAF	Minor Allele Frequency
SV	Structure Variant
HWE	Hardy-Weinberg Equilibrium
DHS	DNase Hypersensitive Site
RVIS	Residual Variation Intolerance Score
CEU	Utah Residents with Northern and Western European Ancestry
YRI	Yoruba in Ibadan, Nigeria
LWK	Luhya in Webuye, Kenya
ASW	Americans of African Ancestry in SW USA
PCR	Polymerase Chain Reaction
hg19	Human genome Reference version 19
dbSNP	Single Nucleotide Polymorphism Database
CI	Confidence Interval
VQSR	Variant Quality Score Recalibration
RIN	RNA Integrity Number
GEO	Gene Expression Omnibus
SNM	Supervised Normalization of Microarrays

SUMMARY

Maintaining health and eliminating diseases are always intriguing topics that everyone cares about. Along with medical biology and physiology which cover knowledge of human tissues and biochemistry, the study of human genomes now also attracts our attention. Gene mutations can cause diseases, exemplified by mutations in the genes *BRCA1* and *BRCA2* which increase a woman's risk of breast cancer. However, most disease has a more complex genetic basis that is just beginning to be explained. The completion of the Human Genome Project has led to the development of genotyping and sequencing techniques in the past two decades that now allow human geneticists to better understand the relationship between genotypes and disease.

Discovery of associations between genetic variants and disease status or quantitative traits is shedding light on disease mechanisms and promoting improved prediction of risk. For many years the prevalent model of disease risk has been the "common disease, common variants" hypothesis. It postulates that common diseases are mostly caused by common variants of quite large effect. However, the introduction of Genome Wide Association Studies (GWAS) in the mid 2000's has shown that such large effect common variants only explain a small proportion of heritability, leading to the "missing heritability" problem. Attention has switched to deep DNA sequencing studies of rare variants that may contribute to individual cases. However, exome sequencing has dominated and the focus has been on rare coding variants. Rare regulatory variants have not been well studied.

This dissertation aims at investigating the association between genotypes and phenotypes in human. Both common and rare regulatory variants have been studied. The phenotypes include disease risk, clinical traits and gene expression levels. This dissertation describes three different types of association study.

The first study investigated the relationship between common variants and three sub-clinical traits as well as three complex diseases in the Center for Health Discovery and Well Being (CHDWB) study. A post test disease probability for the three diseases - coronary artery disease (CAD), type 2 diabetes (T2D) and asthma - was calculated incorporating prevalence of the disease and the likelihood ratio for all significantly associated common SNPs identified by GWAS. The polygenic risk scores for three traits – height, body mass index and triglycerides -were calculated based on the total counts of alleles which increase the trait levels. Although I studied a small cohort of ~200 people, statistically significant relationships were found between polygenic scores and quantitative traits, and also between combined likelihood ratio and Framingham risk scores for CAD. The explanatory power of the top-ranked SNPs was compared with that of all significantly associated SNPs by adding SNPs stepwise to the regression models. The result shows that the top-ranked SNPs could explain as much of the variance as is explained by the top few hundred SNPs. While the detection of positive genotype-phenotype associations in a small cohort is encouraging, the results of this study also highlighted the limited clinical potential of risk scores based on common variants identified in genome-wide association studies.

The second study is GWAS analysis of TNF-α and BMI/CRP conducted as a contribution to a meta-GWAS analysis of these traits with investigators at the University of Groningen in the Netherlands, and the 1000 Genomes Consortium. The TNF-α analysis was performed as a replication study, based on 44 SNPs previously discovered, but incorporating a linear mixed effect model to account for longitudinal measures obtained over three visits. The top-ranked SNPs all replicated despite the small sample size of the CHDWB cohort, and the longitudinal model showed essentially the same association significance as the baseline-only model. The GWAS on BMI/CRP was performed as a contribution to meta-GWAS discovery analysis after imputing SNPs genome-wide. SNPTEST was implemented for the association test with expected and

threshold models that provided alternative candidate SNPs for the meta-analysis. This study shows that small cohort is nevertheless capable of serving as replication study for meta-GWAS analysis. In addition, it confirms that imputation is an effective way to detect GWAS signals even with a low resolution genotyping array.

The third study was the most original contribution of my thesis as it assessed the association between rare regulatory variants in promoter regions and gene expression levels. By targeted sequencing of the promoter regions of 480 genes in 410 individuals, I was able to develop a novel burden test for rare variants at the extreme of expression. Burden tests were performed by calculating the summed rare allele counts in ranked expression level bins for all individuals, and separately for European-ancestry individuals for more targeted analyses avoiding possible influences of population structure. The results clearly show an enrichment of rare variants at both extremes of gene expression. The rare regulatory variant effects were also partitioned into subsets of genes based on their regulatory functions, positions relative to transcription start sites, disease relatedness, with some intriguing biases. The enrichment of rare regulatory variants in extremely expressed genes was replicated using another cohort and different sequencing and gene expression profiling technologies. The effects of three of four rare variants with large effect sizes were experimentally validated by CRISPR/Cas9 mediated genome editing.

This dissertation provides insight into how common and rare variants associate with broadly-defined quantitative phenotypes. The demonstration that rare regulatory variants make a substantial contribution to gene expression variation has important implications for personalized medicine as it implies that de novo and other rare alleles need to be considered as candidate effectors of rare disease risk.

CHAPTER 1

INTRODUCTION

Description of the Dissertation

My dissertation aims to investigate the association between genetic variants and broadly-defined phenotypes. The broadly-defined phenotypes here include gene expression levels, clinical measures and disease risks. The genetic variants include common variants and rare regulatory variants. Association analysis is expected to provide a better understanding of the genetic contribution to variation of phenotypes among individuals, leading to better insight into genetic prediction of diseases.

The studies were based on the CHDWB (Center for Health Discovery and Well Being) cohort, which is a collaborative center between Emory University, Georgia Tech and the ACTSI. The CHDWB cohort includes in total of 697 healthy adults, among whom 651 are Emory employees and the remaining small proportion is from the general public and Georgia Tech. The individuals had blood withdrawn every six months, with at least 3 time points drawn from most of them. The Center provided health partners who provided enrolled individuals with health – related advice including exercise, diet recommendations intended to maintain or improve their health status.

The dissertation is composed of five studies, each with a chapter. The object of the first study was to investigate how well common variants explain three quantitative phenotypes and 3 common disease risks (Chapter 2) in the small cohort. The quantitative phenotypes were height, body mass index and triglyceride; the 3 common diseases were

coronary artery disease, type 2 diabetes, and asthma. This study was published in Genetics Research and the chapter is slightly edited from the version co-written with my supervisor, Professor Gibson. In Chapter 3, GWAS analysis of TNF-α and BMI/CRP was performed by participation in the meta-GWAS analysis led by group at the University of Groningen in the Netherlands and the 1000 Genomes group, with imputed genotypes from microarray data.

The object of the third study was to extract the detailed genotype information from targeted sequencing (Chapter 4). The targeted sequencing was designed to discover rare variant within 1kb of each side of the transcription start site (TSS) of 472 genes. Different variant calling algorithms were compared. Rare variants distributions were then compared among different populations, and among subgroups of genes, conditioned on their regulatory functions, disease relatedness, and other attributes.

The fourth study aimed for the first time to investigate the hypothesis that there is enrichment of rare regulatory variants in promoter regions in the context of extreme gene expression levels (Chapter 5). Rare variants in coding regions are now being investigated as a source of congenital abnormalities. My thesis asks whether rare regulatory variants might contribute to aberrant gene expression, which might in turn promote disease. These associations were also analyzed within subgroups of genes conditioned on disease relatedness and regulatory functions. Parts of Chapter 5 have been co-written with my supervisor as we prepare to submit the paper for publication, whereas Chapter 4 reports my more detailed analyses that will be incorporated into the paper.

Genetic Variation in Human Genome

Genetic variation refers to the difference in DNA sequences among samples or between populations. Usually it refers to the differences when comparing the genome of one person with a reference genome. Genetic variation is caused by mutation and is mainly composed of single nucleotide polymorphism (SNP), copy number variation (CNV), insertions and deletions (indels) (Figure 1.1) [1, 2]. SNPs, which are polymorphic sites affecting a single nucleotide, are the most commonly occurring type of genetic variation. Copy number variation is the variation in the number of copies of longer sections of DNA between tens and thousands of kilobases. Indels are most often the insertion or deletion of up to 100 nucleotide bases in the DNA.

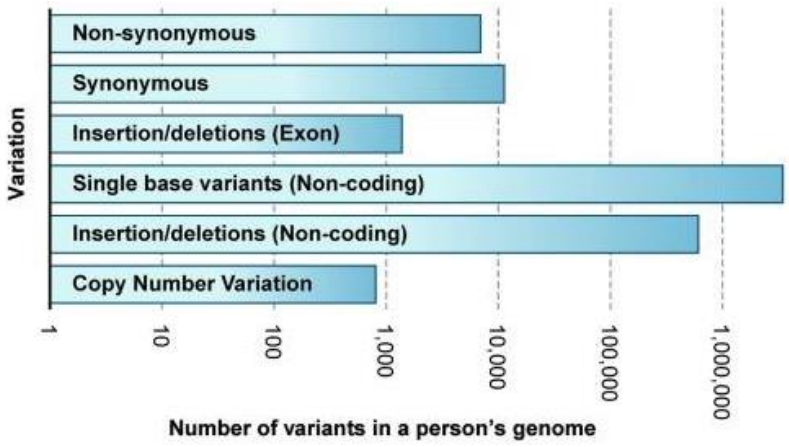

Figure 1.1 Number of different types of variants in a person's genome. Taken from [1]

A series of techniques have been designed in order to detect genetic variation. Among those techniques, a commonly used one when I started this thesis was genotyping

arrays which were initially designed for SNP detection [3, 4]. The basic principle depends on hybridization of fragmented single-stranded DNA to complementary nucleotide probes in arrays. Each array contains hundreds of thousands of unique probes. After hybridization, the signal intensity of fluorescence at each probe is measured. The raw signal intensity is then converted to genotypes via computational algorithms provided by the manufacturer, in our case Illumina. These SNP arrays have been widely used by many project groups such as HapMap consortium. However, SNP arrays have the drawback that there is ascertainment bias due to the non-random distribution of SNP probes throughout the genome.

Starting around 5 years ago, there was a major shift in variant detection with the arrival of next generation sequencing (NGS) methods [5, 6]. Compared with automated Sanger sequencing [7] which is considered to be a first generation technology, next generation sequencing has revolutionized human genetic analysis. It includes DNA sequencing [8, 9], RNA sequencing [10, 11], ChIP-sequencing [12] and others, enabling the assessment of a broad range of biological phenomena such as genetic variation, RNA expression, chromatin conformation, and DNA-protein interactions. The NGS technologies differ from the Sanger method in aspects of massively parallel analysis, high throughput, and have reduced cost. With the maturation of sequencing technologies, even more sequencings reads can be produced within a shorter time and with much lower costs.

For my research I have used the Illumina Truseq System protocol [13] on Hiseq platform. This is a versatile technology that now dominates the market and has well

validated performance with relatively low cost. Typically tens of millions of paired reads each 100 nucleotides long are generated from either end of DNA fragments. These are then computationally aligned to the reference genome to detect sequence variants. The process is called re-sequencing since each individual's DNA is compared with a previously known standard, hg19.

Bioinformatics Tools for Sequencing Analysis

Raw sequencing reads require bioinformatics analysis before they can reveal scientific insight. From alignment to variant calling to functional annotation, there are many types of analysis tools and software. The most popular sequence aligners are Bowtie, Bowtie2 and BWA. Bowtie is a short read aligner based on indexes built with the Burrows-Wheeler algorithm [14]. Bowtie 2 uses an FM Index (based on the Burrows-Wheeler Transform (BWT)) to index the genome [15]. Compared to Bowtie 1, Bowtie 2 performs faster when aligning reads longer than 50bp and deals more flexibly with paired-end alignment. The BWA (Burrows-Wheeler Aligner) is based on backward search with the BWT [16]. There are also a number of open source variant calling tools available. GATK [17] calls SNPs and indels using Bayesian model with Java code. GATK applies machine learning methods to base quality recalibration and variant quality recalibration. SAMtools [18] uses HMM & MAQ model with C code. It calls the variants that maximize the posterior probability with the highest Phred quality score using a general Bayesian framework. There are also several other commonly used variant calling tools such as VarScan [19] and SNVer [20]. Varscan is compatible with several aligners like Bowtie and BLAT [21] and can process SAMtools pileup files. VarScan calls variants

according to the coverage, quality, as well as the number of supporting reads. In my rare variant association study, I used BWA to align my targeted sequence data, and chose GATK to perform variant calling after comparing different calling methods.

Genome Wide Association Studies

Understanding the genetic basis of diseases helps scientists understand the cause of disease, predict disease risk in individuals and develop clinical interventions. Prior to 2007, most research on mapping genetic loci that have effects on diseases or other complex traits was performed by linkage analysis [22, 23]. Linkage analysis is based on the co-segregation of causal variants with disease status within pedigrees. It has been most successfully applied in discovery of genetic variants which are related to Mendelian diseases and traits [24]. Since the development of genotyping techniques, especially the development of the HapMap project [25], genome-wide association studies (GWAS) have been more popular. These are based on the existence of linkage disequilibrium between causal SNPs and tagged SNPs in unrelated individuals. Following theory first proposed by Risch and Marikengas [26], in 2005, the first GWAS reported two SNPs found to be associated with age-related macular degeneration [27]. Since then, GWAS has been widely used to find thousands of disease- and trait- associated genetic variants for hundreds of clinical phenotypes.

At the early stage, GWAS was focused on common variants, based on expectation of the common disease - common variant hypothesis [28, 29]. However, it has been found that only a small proportion of variance can be explained by common variants, leading to discussion of “missing heritability” problem [30, 31]. Much more variance was expected

to be discovered, but we now think that most of it is hidden, rather than missing. This is because effects of individual SNPs are very small and there is not enough statistical power unless hundreds of thousands of people are studied. Subsequently, the hypothesis was proposed that rare variants with moderate to high effect size may also contribute to the missing heritability. Evidence that rare variants associate with complex diseases and traits has been found for schizophrenia [32], HDL [33] and T1D [34]. The frequency and effect size in genetic association analysis is shown in Figure 1.2.

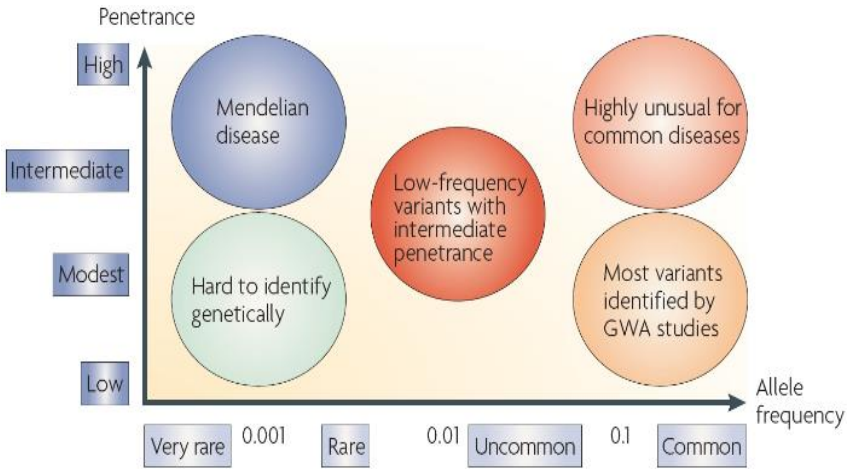

Figure 1.2 The frequency and penetrance in genetic association analysis. Taken from [35]

Genetic Risk Prediction

People care about their health and wellness on a day to day basis. Knowing their disease risk may help them to make relevant adjustments to lifestyle, and take precaution against diseases enabling them to maintain wellness. Knowing the classification of high

or low disease risk could help people adjust for their lifestyles including eating and exercise habits to maintain wellbeing. Disease risk prediction previously was based merely on clinical measures such as body mass index, smoking status, lipid levels, and so on. With the advent of GWAS, genetic variants have been found to be associated with the disease onset and disease status. Incorporating genetic information with clinical measures and family history may provide a more informative prediction of the disease risk. While most genetic risk prediction is focused on cancers, here I presented a genetic risk prediction for 3 quantitative phenotypes height, triglyceride, body mass index, and 3 common diseases coronary artery disease, type 2 diabetes and asthma, and compared the variance explained by genetic information for phenotypes and clinical risk for those 3 common diseases. However, I concluded that the technology is not yet ready for clinical evaluation since too little variance is explained by known SNPs from GWAS. This may change in the next 10 years.

eQTL Analysis

Gene expression, alongside protein function, is the major source of variation in cellular function. The patterns and properties of gene expression influence protein activity levels, which then determine cell states. Gene expression also plays an important role in disease status by influencing the phenotype. For example, researchers have found that the aberrant expression of surfactant protein C may lead to lung disease [36]. As a result, gene expression has been attracting scientists' interest as an intermediate phenotype between genetic variants and phenotypic traits. Mutations in the coding region may lead to abnormally translated proteins, thus causing change of signaling pathways or

other biological processes that the proteins participate in. Abnormally expressed genes due to mutations in non-coding regions that for example change affinity of transcription factor binding sites, may lead to abnormal genetic regulation and gene expression. Therefore, finding the genetic variants that are associated with gene expression levels, i.e., expression quantitative trait loci (eQTL), can help reveal the mechanisms of how genetic variants influence phenotypes and uncover the genetic basis of many complex traits at the molecular level. Recent studies have revealed that there are substantial overlaps between eQTLs and genetic variants identified in genome-wide association studies which are associated with diseases [37, 38]. It has been shown that trait-associated QTLs are more likely to be associated with expression levels (Figure 1.3) [38]. Therefore, researching the genetic basis for variation in gene expression is not only of academic interest, but also provides information relevant to mechanisms of disease.

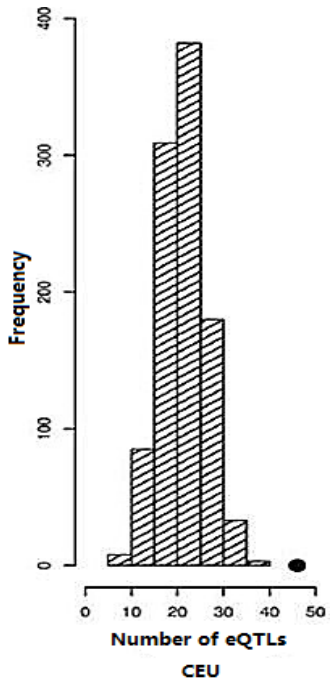

Figure 1.3 The distribution of the number of eQTLs defined with p-value<10^{-6} observed for each of 1000 draws of random 1,598 SNPs (bar graphs), with the actual number of eQTLs observed in the 1,598 SNPs from the NHGRI GWAS catalog (solid circle). Referred and adjusted from [38].

Rare Variant Association Analysis

Although the improvement of experimental strategies and statistical analysis have enlarged the proportion of gene expression variation explained by common variants, the vast majority of gene expression variation still needs to be accounted for. Much of this variation is likely due to factors not located within the gene that encodes the transcript. This can be trans-eQTL (compared with cis-eQTL, located in the gene itself), systemic influences like hormones and metabolites, and the environment. There is also the potential importance of rare variants contributing to the regulation of gene expression levels. Compared to the rare coding variants whose genetic functions have been studied

extensively [39-41], rare regulatory variants have not been systematically studied for their effects on gene expression regulation.

I assume that the rare regulatory variants affect gene expression is based on the following reasons. First, it is possible that a common variant association is actually attributed to linkage with multiple rare or less common variants which contribute to the major effects, a situation known as synthetic association [42]. A well characterized example is the Hepatitis C virus-anemia associated locus ITPA, where the minor allele frequency of the causal SNP is much lower than the most significantly associated one [43]. Second, analogous to the well known fact that rare variants in coding regions affect lipid levels (for example, rare NPC1L1 variants affect plasma LDL lipoprotein levels [44]) and gene expression levels (for example, a rare synonymous CRP2 variant affects serum CRP level [45]), rare polymorphisms in regulatory regions could also influence differential transcript abundance, by affecting the regulatory effects of transcription of the nearby gene. Promoter regions, which play an important role in regulating the transcription process, are an ideal choice for initial regulatory rare variant studies, since enhancers vary greatly in their complexity and location relative to TSS. Third, my hypothesis is that the observed enrichment of common eSNPs in promoter regions (Figure 1.4) [46], may also imply that rare variants in promoter regions also are likely to affect transcription. Recent work such as ENCODE project [47, 48] has established the complexity of regulatory regions, leading to the expectation that more than one causal variant is often present where there is any functional genetic variant at a locuss.

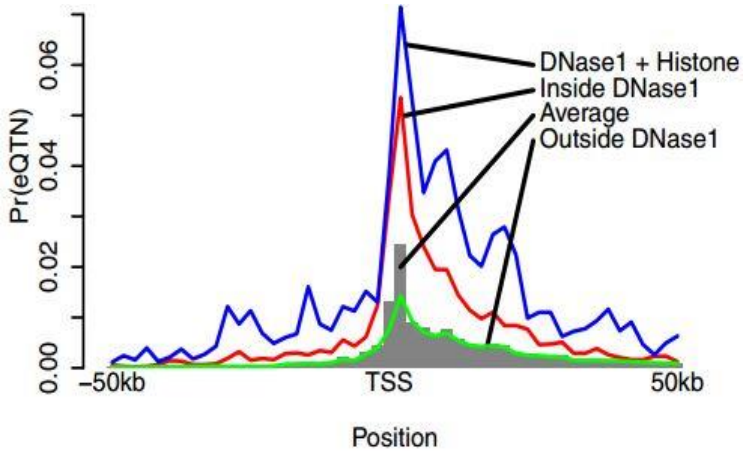

Figure 1.4 The probability of a random SNP (or the SNP conditioned on different functional sites) to be eSNP as a function of distance to TSS. Taken from [43]

In standard GWAS, individual variant tests are employed to test common variant association. However, considering the small individual effect size and the low allele frequency of rare variants, single variant tests will be underpowered to detect rare variant effects. In consideration of this, collapsing the rare variants in a region is the most commonly used approach for analyzing rare variant association. Among the region-based rare variant tests, the most popular ones include burden tests such as the cohort allelic sum test (CAST) [49], the combined multivariate and collapsing (CMC) method [50], and the weighted sum rare allele rank test (WSRRT) [51]. CAST compares the number of individuals with at least one rare variant between affected cases and unaffected controls. CMC collapses all the rare variants as a common variant and performs a multivariate regression together with other common variants. WSRRT weights the rare variant score by the rare allele frequency among unaffected individuals, add ups the ranks of individuals based on the weighted score, then performs a permutation to compare the

rank sum difference between affected and unaffected groups. There are some other rare variant association tests not based on the simple total burden, which take the possible different directions of effect into consideration. This is important for regulation of transcription, since rare variants might either increase or decrease the activity of the promoter. Among these tests are the well-known C-alpha test [52] and Sequence Kernel Association Test (SKAT) [53]. The C-alpha test contrasts the expected variance with the actual variance of allele frequency distributions, testing for a mixture of effects across a set of rare variants. SKAT is a regression based approach which tests for association of rare (and common) variants with a dichotomous or continuous phenotype while adjusting for covariates.

However, these tests require sample sizes of at least tens of thousands of individuals, and we do not have gene expression datasets that large. Consequently I proposed a pooling strategy, which required the development of a novel test based on the significance of the quadratic component of a regression of rare allele burden on rank of expression. This is described in Chapter 5 and shown to provide strong and replicated evidence that rare variants are enriched in the tails of transcript distributions.

CHAPTER 2

GENETIC RISK PREDICTION

Introduction

Despite high heritability, most complex traits and diseases in humans have such a polygenic inheritance pattern that prediction of phenotype or liability on the basis of genetic risk profile has proven elusive. The possible benefits of genetic risk evaluation were recognized more than a decade ago [54-56], but only recently has the application of high-density genotyping technology [57] brought us closer to the goal. Polygenic scores, which represent the summed effects of multiple trait-associated genetic variants, contain more information than single markers and explain more of the variance in phenotype or disease risk. This type of analysis has been applied to several complex traits including height [58], body mass index [59], and rheumatoid arthritis [60], in each case evaluating the joint effects of polymorphisms identified in samples of tens or even hundreds of thousands of individuals, in large validation cohorts of several thousand.

Most cohort studies are focused on a single trait or condition, so do not allow the evaluation of genetic risk across multiple phenotypes. Here I report on common variant contributions to three traits and three diseases in the Emory-Georgia Tech Center for Health Discovery and Wellbeing study of clinically deeply profiled adults, 182 of whom have SNP array genotype information available. Despite the relatively small size of the cohort, I nevertheless detect significant association with quantitative traits, and take the opportunity to compare methods that do or do not weight allelic effects, while also

comparing genetic risk scores with Framingham risk scores for coronary artery disease and type 2 diabetes. Examination of the correlations highlights the strong influence of outliers on risk prediction, and raises the hope that in a well-characterized cohort it may be possible to identify the hidden variables that are shared by such outliers, which may in turn suggest strategies for conditional analysis to uncover more of the hidden heritability.

Some studies have also reported that most of the variation explained by polygenic risk scores can actually be explained by the top-ranked markers [61, 62]. The rationale is that the top-ranked markers tend to have the largest genetic effects, so explain more of the disease or trait than markers that only emerge once large-scale meta-analyses have been performed. The proportion of variance that they explain will be a function of the distribution of effect sizes, which is itself difficult to estimate due to the high noise level in GWAS data. Therefore, the risk attributable to top-ranked SNPs is an empirical question, and my dataset allowed me to address performance across phenotypes in the same individuals. I applied forward-step regression adding alleles sequentially to investigate the influence of adding more marker information to the regression on explaining the genetic variance. It turns out that in the small cohort of 182 people considered in this Chapter, the inclusion of covariates such as gender and ethnicity, influences performance of genetic risk scores, presumably because of residual correlation between genotype and those covariates. While the detection of positive genotype-phenotype associations in a small study is encouraging, the results also highlight the limited clinical potential of risk scores based on common variants identified in genome-wide association studies [63].

Materials and Methods

Participants

The CHDWB is a longitudinal study of health measures in over 600 employees of Emory University. I describe data for 182 participants for whom genotypic data was available, consisting of 136 Caucasians, 34 Africans, 11 Asians and 1 American Indian. Two thirds of the individuals were women (120 females and 62 males), and the ages ranged from 26 to 79. The data of interest for this study is height (in cm), BMI (weight /height^2 in kg/m^2), serum triglyceride levels (mg/dL), serum cholesterol levels (mg/dL), and various measures of blood flow and arterial stiffness. I also computed Framingham risk scores for type 2 diabetes and for coronary artery disease as described in [64, 65].

Genotypes

Whole genome genotypes were measured using Illumina OmniQuad arrays which contains 733,202 probes. Identities of 169 Height, 49 BMI, 48 triglyceride, 34 coronary artery disease, 66 type 2 diabetes, and 31 asthma related SNPs were collected from the dbGaP database hosted by the US NIH, in March 2012. All of the selected SNPs were previously reported to be significantly associated with the respective traits or diseases at the significance level of $p < 10^{-7}$. Individual genotypes for each of these SNPs were extracted, or if missing from the Illumina genotype data files, were imputed using IMPUTE2 [66]. Accuracy of the imputation was estimated to be 98% by comparison with 9 individual whole genome sequences (2 African Americans, 7 Caucasians).

Genetic Risk Score Analyses

Three approaches to calculating the proportion of phenotypic variance explained by the common genetic variants were considered. All computations were performed using R scripts. For the continuous quantitative traits height, BMI and triglyceride level, I first calculated the sum of increasing alleles for each individual. Second, I calculated a weighted sum of allelic effects according to the effect size of each SNP reported in dbGaP. Each of these allelic sum and weighted allelic effect scores was then linearly regressed on the relevant phenotype(s), with or without adjustment for gender and ethnicity. In the latter case, the genotypic contribution was estimated from the difference in the variance explained (R-squared on the Pearson correlation coefficient) by the models including the genetic risk (allelic sum, or weighted allelic effect) score, and without it. Furthermore, the influences of gender and ethnicity were estimated directly from our cohort by including these terms as covariates; or by incorporating reported population averages for each gender and ethnicity from the CDC website as "pre-height".

For disease risk variants, the third approach was to compute the multiallelic odds ratio essentially as described in Ashley *et al*.[67]. I computed an adjusted relative genetic risk by setting each individual's prior odds as that corresponding to the prevalence for their gender and ethnicity as reported by the CDC. In order to obtain the genetic contribution to the post-test odds, a slight adjustment to the Ashley *et al*. [67] method was performed as follows. According to those authors,

pre-test odds = pre-test probability/(1-pre-test probability);

post-test odds = pre-test odds×LR;

post-test probability = post-test odds/(1+post-test odds).

Rearranging their equations,

post-test probability = pre-test probability×LR/(1+pre-test probability×(LR-1)).

As the reported 95% confidence intervals for genotypic contribution LRs lie between 1.6 and 0.7, pre-test probabilities range between 4.2% and 14.3%, then post-test probabilities range from (0.93~1.04)×pre-test probability×LR. It follows that I can approximate the post-test probability as the pre-test probability×LR. That is,

log10(post-test probability)=log10(pre-test probability)+log10(LR)).

I confirmed this relationship by observing that this approximated post-test probability is highly significantly linearly associated with the post-test probability using Ashley *et al*'s method ($P < 2\times10^{-16}$). The advantage of my portioning is that it allows estimation of the genetic contribution to the post-test probability independent of the pre-test probability.

In order to ask whether addition of more SNPs always improves risk prediction, the SNPs were sorted by previously reported effect sizes from larger to smaller. Each SNP was added sequentially to the regression model, taking the negative $\log_{10}$ p-value and percent variance explained by each successive model. For comparison with random SNP selection, I randomly added those SNPs sequentially to the regression model, and then averaged the percent of variance explained by each successive model, averaging the results over 100 permutations.

Results

Regression of Genotypic Risk Scores on Phenotypes

Significant and positive regression of genotype on phenotype was observed, as

expected, for each of the continuous traits (height, BMI, and serum triglycerides) as shown in Table 2.1. In each case, the estimated variance explained by the SNPs was in the range of 3% to 5%, which is lower than that reported in the respective discovery samples [58, 60, 68]. The inclusion of an estimated effect size in weighted sum scores did not significantly improve the model fitting. For each of these traits, gender and ethnicity explains considerably more of the variance than the genotypes, and fitting these covariates slightly improved the estimate of the genetic contribution (with the exception of the weighted sum for BMI). The weighted sum was not calculated for triglycerides since the effect sizes were not fully reported in dbGaP. I also fit multiple regression models based on jointly fitting all of the SNPs for each trait, and although the variance explained reached 16% for height the estimates were not significant after adjustment for the number of SNPs included.

Regressions were also computed for disease-associated risk scores, namely: T2D risk with the Framingham T2D risk score, and with serum triglyceride, cholesterol, fasting glucose, and insulin levels; CAD risk with the Framingham CAD risk score, blood pressure, arterial stiffness, and serum metabolites; and asthma risk with estimated VO2-max from treadmill performance. Only two of these analyses (CAD SNPs with Framingham CAD risk score, and with cholesterol) yielded nominally significant correlations as reported in Table 2.1, and these would not formally survive adjustment for multiple comparisons. Nevertheless, for CAD, the total number of increasing alleles showed a surprising positive relationship with total cholesterol levels, even though there is little overlap between these SNPs and those associated with cholesterol by GWAS.

Table 2.1 Variance explained by genetic risk scores.

SNPs for	Trait	Score type	Covariate	Adj. Rsq	P-value	σ^2 expl.	Adj. Rsq	P-value	σ^2 expl.	Panel in Figure 2.1
					All samples			CAU only		
Height	Height	Allele sum	-	0.027	0.026	2.70%	0.025	0.066	2.50%	a
Height	Height	Allele sum	Gender/Ethn	0.467	4×10^{-4}	4.00%	0.45	0.002	4.00%	
Height	Height	Weighted sum	-	0.037	0.009	3.70%	0.036	0.027	3.60%	b
Height	Height	Weighted sum	Gender/Ethn	0.466	5×10^{-4}	3.80%	0.498	0.001	4.20%	
Height	Height	Weighted sum	Pre-height	0.027	0.026	2.70%	0.498	0.001	4.20%	
BMI	BMI	Allele sum	-	0.052	0.002	5.20%	0.065	0.003	6.50%	c
BMI	BMI	Allele sum	Gender/Ethn	0.18	0.001	5.30%	0.087	0.003	6.10%	
BMI	BMI	Weighted sum	-	0.051	0.002	5.10%	0.029	0.049	2.90%	d
BMI	BMI	Weighted sum	Gender/Ethn	0.169	0.003	4.20%	0.054	0.047	2.90%	
BMI	BMI	Weighted sum	Pre-height	0.101	0.005	4.10%	0.054	0.047	2.90%	
Triglycerides	Triglycerides	Allele sum	-	0.042	0.005	4.20%	0.067	0.002	6.70%	e
Triglycerides	Triglycerides	Allele sum	Gender/Ethn	0.115	0.003	4.50%	0.115	0.001	7.60%	
CAD	Cholesterol	Allele sum	-	0.044	0.005	4.40%	0.028	0.052	2.80%	f
CAD	Cholesterol	Allele sum	Gender/Ethn	0.005	0.06	4.30%	0.569	0.04	1.40%	
CAD	Log_{10}(FHS+1)	Log_{10}(LR)	-	0.055	0.02	2.00%	0.053	0.007	5.30%	g
CAD	Log_{10}(FHS+1)	Log_{10}(LR)	Log_{10}(pre-test)	0.008	0.418	2.30%	0.062	0.006	5.40%	
CAD	Log_{10}(Chol)	Log_{10}(LR)	-	0.015	0.032	3.20%	0.036	0.028	3.60%	h
CAD	Log_{10}(Chol)	Log_{10}(LR)	Log_{10}(pre-test)	0.016	0.042	3.20%	0.046	0.023	3.80%	

Table 2.1 shows that regression of genotypic risk scores on phenotypes was little affected by considering only the Caucasians. The proportion of variance explained by SNPs for triglycerides and CAD was slightly increased relative to the full cohort, likely due to better capture of LD between tagging and causal SNPs in Caucasians, but this effect is offset by the smaller sample size for other traits.

Effect of Number of Alleles on Risk Prediction

Step forward regression, sequentially adding SNPs in the order of previously reported effect size, was performed to address whether the addition of more SNPs to the model continuously improves the prediction. Figure 2.1 shows the results for height, BMI, and cholesterol on the left hand panels, compared with average effects for 100 randomly permuted orders of SNP addition on the right panels. In each case, explanatory power of the SNPs increases at least for the first 30 SNPs included in the model.

For height (Figure. 2.1 *a* and *b*) it is also clear that most of the variance is explained by the top 30 SNPs and that sequential addition up to 169 SNPs does not improve the fit. Models without gender and ethnicity covariates (blue and brown curves) actually explained the most variance when an intermediate number of SNPs were selected. However, since inclusion of more SNPs reduced the estimates to levels more consistent with those obtained when gender and ethnicity are included, the scores with intermediate numbers of SNPs are likely to be over-estimates. For BMI (Figure 2.1 *c* and *d*), there is a suggestion of a plateau effect after 10 SNPs, without a clear further increase until 40 SNPs are included in the model. In this case, fitting gender and ethnicity does not affect the genetic estimates. For CAD (Figure 2.1 *e* and *f*), significant explanation is not observed until 30 SNPs are included, but there may be a plateau thereafter, and the estimates are not obviously influenced by inclusion of the covariates. For BMI and cholesterol, the weighted sum (or likelihood ratio) scores performed slightly less well than the simple allelic sum predictors.

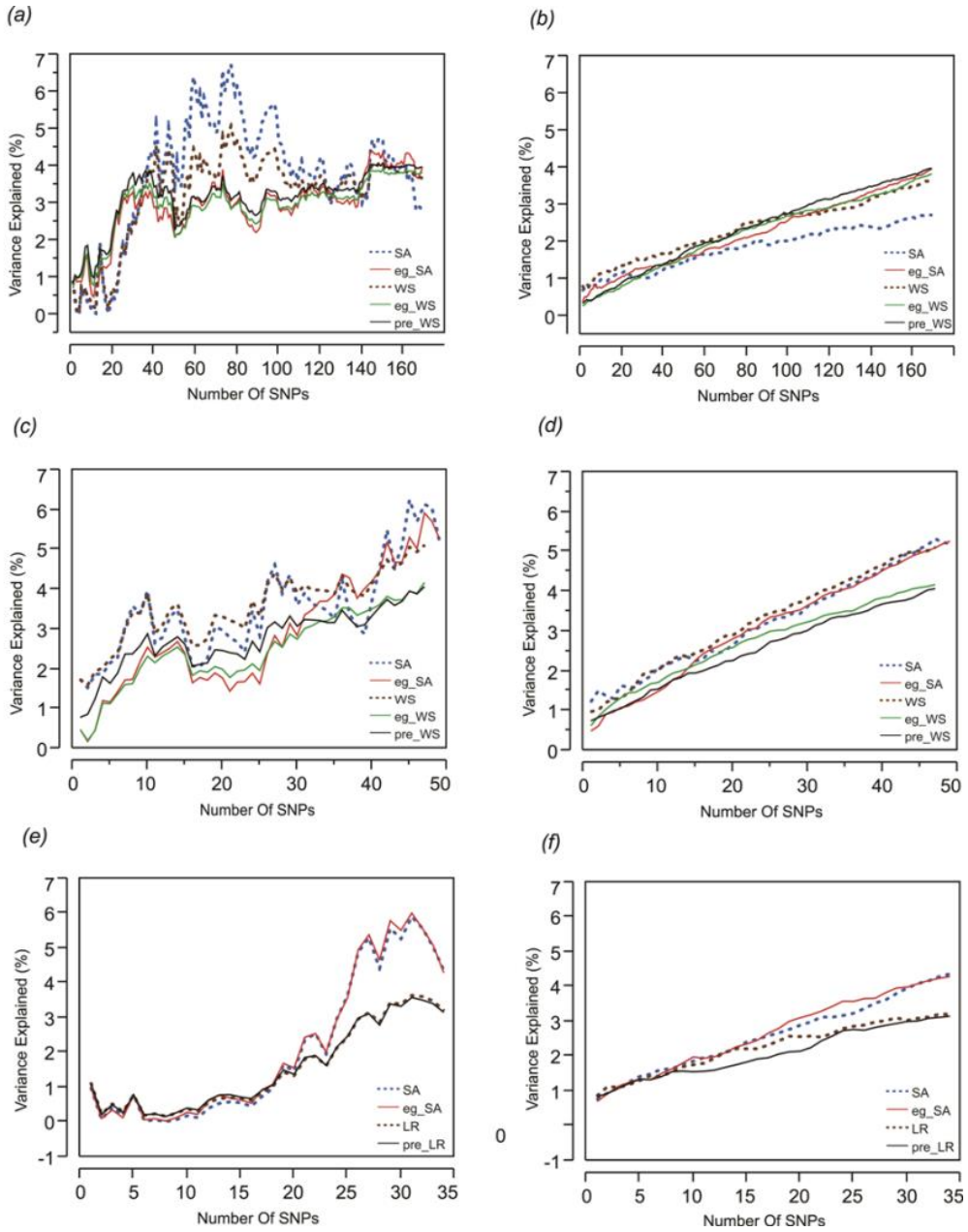

Figure 2.1 The percentage of variance explained by the models with sequentially adding SNPs in the order of their effect sizes and compared with random orders. The percentage of variance explained by the model by sequentially adding SNPs (Height, BMI and cholesterol-CAD SNPs from top to bottom, *(a)*, *(c)* and *(e)*). Right: The percentage of variance explained by the models randomly adding SNPs (Height, BMI and cholesterol-CAD SNPs from top to bottom, *(b)*, *(d)* and *(f)*) averaged over 100 permutations. SA refers to models with just the sum of alleles score, while eg_SA refers to models additionally fitting ethnicity and gender as covariates with the sum of alleles score. WS refers to models with sum of weighted allelic effects, while eg_WS and pre_WS refer to weighted allelic sum includng ethnicity and gender in the CHDWB cohort, or taken as the population averages, as covariates respectively. LR refers to likelihood ratio models, with or without pre-test probability as a covariate.

Effect of Outliers on Explanatory Power

Inspection of the regression plots in Figure 2.2 suggests the estimated variance explained can be strongly influenced by outliers. Thus, the sum of alleles test for height in males shows several men who grow taller than their genetic information predicts (asterisks in Figure 2.2). Except for one Asian, all of these men are Caucasians. Removal of them improves the percent variance explained from 8% to 37%.

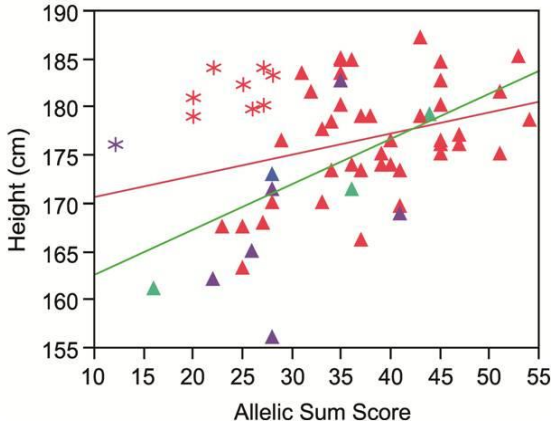

Figure 2.2 Linear regression plot fitting real height by sum of increasing alleles in males. Red dots Caucasians; blue American Indian; green African Americans; purple Asians. Asterisks: the individuals who are taller than their genetic information would indicate. Red line: regression fitting line for all men. Green line: regression fitting line for males without those taller than expected men.

The sum of alleles test plot for triglycerides (Figure 2.3) appears to differ between higher and lower triglyceride levels. If the analysis is restricted to individuals with TG more than 100 mg/dL, the genotypes explain a trivial 0.5% of the variance, while regression on the remaining individuals with lower TG has a similar slope but explains 5.3% of the variance. The increased phenotypic variance in the high triglyceride range reduces the significance of the overall regression even though the slope is greater than in either the low or high TG ranges. Moreover, the association is more significant in

Caucasians than other ethnicities and more significant in males than females.

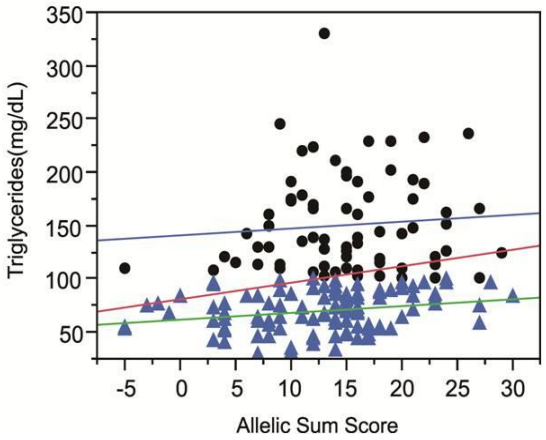

Figure 2.3 Linear regression plot fitting total triglyceride levels by sum of increasing alleles. Dots: TG > 100 mg/ml. Triangles: TG < 100 mg/ml. Red line: linear regression for all the individuals (P=0.0053, R^2=0.042). Blue line: linear regression for individuals with TG > 100 mg/ml (P=0.5453, R^2=0.005). Green line: linear regression for individuals with lower TG levels (P=0.0169, R^2=0.053).

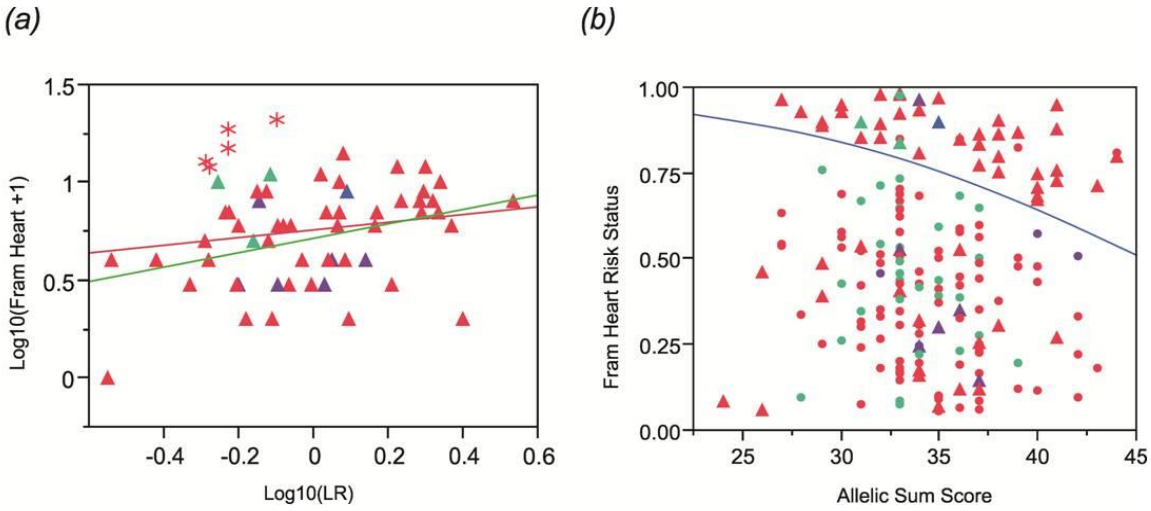

Figure 2.4 Regression of Framingham risk score for heart against likelihood ratio and allelic sum score. *(a)* Regression of $\log_{10}$(FHS for heart disease) against genotypic log likelihood in males. Exclusion of five older Caucasian males (asterisks) elevates the regression from P=0.18, R^2=0.03 to P=0.0065, R^2=0.13. *(b)* Logistic regression of Framingham risk status on sum of CAD risk alleles in all study participants shows a significant association (P=0.0221, R^2=0.027). Red dots: Caucasians, blue: American Indian, green: African Americans, purple: Asians. Circles females, Triangles males.

The regression plot for logarithm transformation of Framingham heart disease risk score and likelihood ratio for CAD SNPs in males shows that there are five Caucasian males (asterisks in Figure 2.4 *a*) who have higher Framingham risk scores than expected. They are all older than the average male ages. Exclusion of those five males results in a more significant association (Figure 2.4 *a*). In addition, I set Framingham risk status as 0 and 1 based on the Framingham risk scores (0 when FRS <4, 1 when FRS >=4). The logistic regression shows a significant association between Framingham risk status and sum of CAD risk alleles (Figure 2.4 *b*), but the area under the ROC curve is just 0.60, indicating that this is not a clinically useful predictor [69].

For BMI (Figure 2.5 *c*), the weighted sum regression plot suggests that the genotypes are more strongly associated with the BMI in African Americans than Europeans. While weighted sum of effects account for 17% of BMI variation in African Americans, the effects only explain 3% of variation in Caucasians. The high variance explained in African Americans is plausibly an overestimate due to the small sample.

Stability of Height Predictions to Number of Included SNPs

I also re-estimated each person's height predictors (allelic sum score, and weighted allelic effect score) from the average of 100 bootstrap samples of 50, 75, and 100 SNPs. The estimates were all highly correlated (Spearman's rank correlation coefficient $\rho>0.99$) and explained almost the same percentage of variance for height. Similarly, the BMI predictors from the average of 100 bootstrap samples of 30, 40 and 49 SNPs were also highly correlated and contributed similarly to the BMI variance. This suggests the increased estimated variance explained for intermediate number of SNPs in height and for more than 40 SNPs in BMI is probably just noise contributing false positive signal. A corollary is that an attempt to include all available SNPs in a model is not necessarily guaranteed to yield the most accurate predictor, since, for example, had only the top 80 SNPs for height been available, more variance would have been

explained than is reasonable given the stepwise increments expected for each additional SNP. On the other hand, the addition of the last 10 SNPs markedly reduced the proportion of variance explained using just the allelic sum score for height, suggesting more stable predictors might sometimes be obtained by considering a range of numbers of included SNPs.

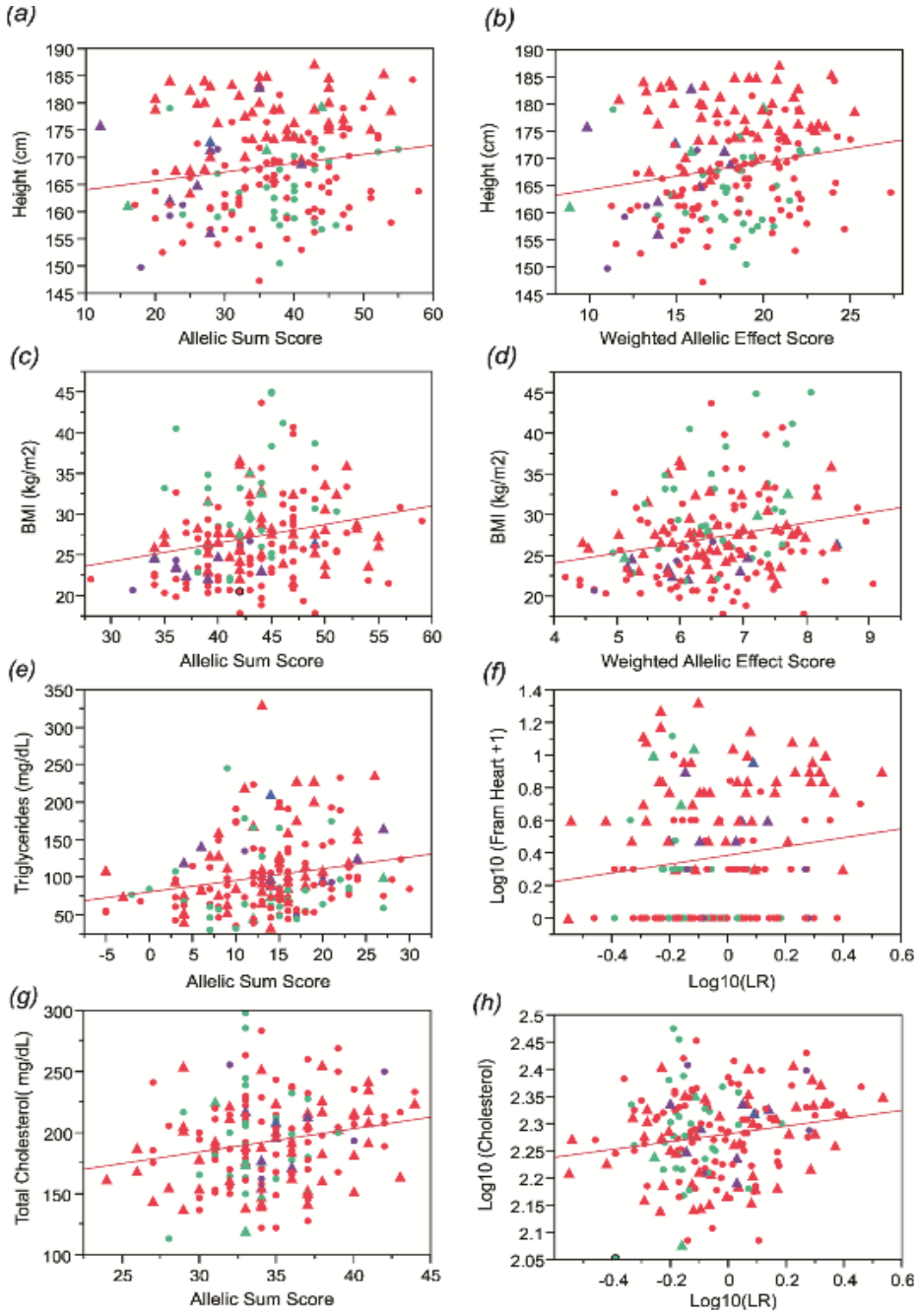

Figure 2.5 Linear regression between observed traits and predicted traits (or predicted disease probabilities). The graphs on the left hand side show the relationship between *(a)* height, *(c)* BMI and *(e)* serum triglyceride concentration with sum of increasing alleles, while those on the right show the relationships with the weighted sum of allelic effects *(b)* height and *(d)* BMI. The bottom plots show the regressions with CAD risk allele scores, namely between $\log_{10}$(Framingham risk score for heart disease +1) and $\log_{10}$(likelihood ratio) *(f)*, between total cholesterol levels and sum of CAD-risk alleles *(g)*, and between $\log_{10}$(cholesterol) and $\log_{10}$(likelihood ratio) for the CAD risk alleles *(h)*. Circle markers and triangle markers represent females and males. Green African Americans; Blue American Indian; Purple Asians; Red Caucasians.

Discussion

According to my results, genotypes ascertained for the most part in large GWAS metanalyses are somewhat predictive of the relevant traits in our small study cohort of typical residents of Atlanta, Georgia. In general, however, the amount of variance explained was smaller than expected, and for CAD and type 2 diabetes the genotypes were not significant predictors of individual disease status. Approximately 4% of height variance was explained by the 169 SNPs whether using sum of increasing alleles, or using weighted sum of effects. This contrasts with 10.5% of adult height variance (using sum of effects method) being explained by 180 SNPs in the analysis of 133,653 individuals [58]. The 700 fold difference in sample size may contribute to the halving of the variance explained, since inspection of the data suggests that a small fraction of outliers (taller than expected men) strongly influence the regression. Additionally, the Atlanta cohort is ethnically diverse, and covers three generations that would have experienced very different socio-economic conditions during growth. On the other hand, it is surprising that the amount of variance in BMI explained by our 49 SNPs is similar to the 4.1% of variance in BMI that is accounted for by 56 variants in a 3,600 sample discovery cohort [61]. The heritability of BMI is considerably lower than that of height, and gender and ethnic differences are strong, yet the genotypic risk score was more consistent than that for height.

In comparison with the recorded 10% triglyceride variance explained by common SNPs [70], the 48 SNPs in my study explained just 4.5% of triglyceride variation. An unexpected finding was that approximately 4.3% and 3.2% of the variation of total cholesterol levels could be attributed to the 34 CAD-related SNPs, performing sum of increasing alleles and multiplication of likelihood ratios respectively. This is the same order of magnitude of explanation as height and BMI, both of which are due to SNPs discovered for the respective trait. The CAD SNPs are related to all forms of coronary artery disease, including atherosclerosis, which is certainly related to cholesterol levels,

but there is no obvious enrichment in the SNPs for cholesterol metabolism. Correspondingly, variation in the Framingham risk score was also partially explained by the likelihood ratio score from the CAD-related SNPs, at a level only slightly less than the 4% explained in [71].

Weighting the allelic effects by the effect sizes reported on dbGaP did not notably improve the prediction of height, BMI, triglycerides or cholesterol. This is perhaps not surprising since there is large variance in the estimation of effect sizes, and to some extent including them in the model adds as much noise as it does signal. In addition, the effect sizes recorded in dbGaP were obtained from the studies which are usually composed of one specific ethnicity (generally European). Even in a few studies whose samples contain more than one ethnicity, the compositions are different from ours. For example, Gudbjartsson et al. [72] recorded effect sizes for 35 of our height SNPs in a study composed of 25,174 Icelanders, 2,876 Dutch, 1,770 European Americans and 1,148 African Americans, which is obviously different from the ethnicity composition in our study, which has 19% African American and 6% Asian. These differences in ethnicity composition could result in the different effect sizes of the SNPs, further reducing the accuracy of the weighted allelic effect scores. It is also the case that reduced linkage disequilibrium in Africans should decrease the proportion of causal effects captured by tagging SNPs, which should reduce the variance explained in the full model.

The identification of subsets of outlier individuals who do not fit the general correlation between genotype and phenotype has implications both for improved estimation of individual genetic effects, and also for prediction. To the extent that shared properties of such individuals can be identified, those properties can be considered as covariates in statistical models, either as regular environmental effects or sources of genotype-by-environment interaction. This conclusion is at odds with arguments that G×E is unlikely to contribute strongly to explained genetic variance [73] or prediction [74]. I think it is relevant that interactions such as those in Figure 2.2 are between the

environmental property shared by the outliers, and the genotypic risk score, rather than with single genotypes. Since the risk score is the sum of 30 or more effects, individual genotype-by-environment interactions can be small, and if they only affect a few individuals, they will not make a substantive contribution to risk averaged across the population. Marigorta and Gibson [75] on our group explored this possibility by simulation and confirmed that Genetic Risk Score-by-Environment effects are much easier to detect than Genotype-by-Environment effects.

In predictive health genetics it may be a problem that variants that exceed GWAS thresholds only explain a small fraction of the heritability [63, 76], and yet there is widespread intention to use these variants to classify individuals with respect to disease risk. A possible rationale for this can be seen in the result that most of the genetic signal is actually due to the SNPs with the strongest effect sizes. This is clearly the case for height, and to some extent triglycerides and cholesterol, though I do not yet have data on whether the addition of a further 100 SNPs would improve the BMI prediction. If effect sizes are Poisson-distributed, then the contributions of the top 30 SNPs are likely to be much greater than those of the next 100 SNPs, which may just tend to cancel one another out and contribute noise. Whole genome regression methods show that inclusion of undiscovered variants can improve genetic prediction [77, 78], but my results suggest that for individually ascertained SNPs, the top few dozen variants will often be as good as the top few hundred. Although they only explain a small fraction of the variance, in keeping with individually small effect sizes, it is notable that the effects are significant across multiple traits even in a cohort of fewer than 200 people.

CHAPTER 3

META-GWAS ANALYSIS ON TNF-α and CRP/BMI

Introduction

GWAS aims to detect variants in particular single-nucleotide polymorphisms that are associated with complex traits such as common diseases and clinical quantitative traits. In the last decade, hundreds of diseases and traits were investigated by GWA studies and thousands of SNP associations have been found. The number of loci found to be associated with diseases and traits continues to expand with the development of improved genotyping arrays, methods for imputation, next generation sequencing, and advanced statistical methods.

Tumor necrosis factor alpha (TNF-α) is a key mediator of inflammatory disease [79]. It is a cytokine involved in inflammation and acute phase reaction stimulation. While it can be produced by many cell types such as CD4+ lymphotypes, NK cells and neutrophils etc, it is produced mainly by activated macrophages. TNF-α plays an important role in cell signaling by activating NF-κB, MAPK, and the apoptosis signaling pathway after it binds with TNF receptors. Here I performed a replication study for TNF-α with 44 lead SNPs as the replication component of a meta-GWAS for TNF-α, in which 16 cohorts were participating with in total > 23,000 individuals.

C-reactive protein is also an acute-phase protein [80]. The level of CRP rises in response to inflammation and tissue damage. Body mass index, defined as body mass divided by the square of height, is one of the commonly used criteria for classification of underweight, normal, overweight, and obesity. It is also a useful predictor of health status. Many studies have shown that there is an association between CRP levels and BMI [81, 82]. In order to evaluate whether the correlation is due to causation or independent correlation between two phenotypes and another causal variable, an

approach known as Mendelian randomization [83] has been proposed. The genotypes are considered to be the possible alternative variable, and their influence on BMI is evaluated with and without CRP in the model. Here, I participated in the GWAS meta-analysis of CRP/BMI as a proof of the principle of bidirectional Mendelian randomization model.

Materials and Methods

Imputation

a. Samples

The CHDWB cohort was genotyped in three batches. A total of 156 samples were genotyped for batch I, and 144 samples for batch II. The first step of the TNF-α GWAS was to impute 8.5 million SNP genotypes based on the 1000 Genomes project data. Prior to imputation, PLINK QC procedures [84] for genotype data cleaning were performed following instructions provided by the consortium to ensure comparison across contributing studies. The quality for samples was checked including sex concordance between annotation and genotypic implication; missing call rate (MCR, set at 0.05 for both batches); sample relatedness; and population structure. After the QC process, a cleaned dataset of 153 unique samples for batch I and 144 unique samples for batch II was yielded.

b. SNPs

All samples in CHDWB were genotyped with Illumina genotyping array. 209 samples (in phase I) were genotyped using Illumina OminiQuad array which contains 733,202 SNPs, another 156 and 144 samples (in phase II) were genotyped separately in two batches using the Illumina CoreExome array. The genotypes were imputed by using IMPUTE2. The phase I data was imputed by UW-Seattle group (Cathie Laurie and Sarah Nelson). The phase II batches were imputed by me using the same protocol as with phase I. Here I only describe the methods and results for the 2 batches in phase II.

Both batches were genotyped on the Illumina Human CoreExome array. The array version of batch I was HumanCoreExome-12v1-0_B, which was based on human genome build 37 and contains 542,882 SNPs. The array version of batch II was HumanCoreExome-12v1-1_B, which was also based on human genome build 37 but contains 542,585 SNPs. The arrays differ from the OmniQuad array used in Phase I by having fewer SNPs overall, but a much higher density of all coding region SNPs including most known variants down to a minor allele frequency of 0.01 in Europeans. Consequently, the common intergenic variants are sparse and imputation was not expected to be as comprehensive as in Phase 1. PLINK QC procedures were used to identify poor quality or otherwise questionable SNPs. The QC checks in SNP level included MCR; Hardy-Weinberg equilibrium; and sex differences according to allelic frequency or heterozygosity rate. Table 3.1 shows the number of SNPs lost and SNPs left after each filter step. Where an observed study SNP had sporadic missing data, the missing genotypes were imputed by the pre-phasing software.

Table 3.1 Summary of SNP quality filters from genotype data cleaning. Top: batch I. Bottom: batch II.

Filter	SNPs lost	SNPs kept
SNP probes	NA	542,882
missing call rate > 0.05	47,250	495,632
HWE p-value < E-4	6,335	489,297
sex difference in allelic frequency >= 0.2	10	489,287
sex difference in heterozygosity rate > 0.3	0	489,287

Filter	SNPs lost	SNPs kept
SNP probes	NA	542,585
missing call rate > 0.05	3,599	538,986
HWE p-value < E-4	34	538,952
sex difference in allelic frequency >= 0.2	384	538,568
sex difference in heterozygosity rate > 0.3	519	538,049

c. Data formatting

The raw Illumina text files were first re-formatted to long-format fileset (LGEN), then converted to PLINK binary file with the samples passing the QC.

Before imputation, bed files were made from PLINK binary files for each chromosome. The haploid genotypes in chromosome X in male which were called as heterozygotes were set as missing. Only the SNPs and samples which have passed the QC filtering were included in the output bed files. In addition, the strands of the variants were flipped if they were not "+" strands according to Illumina annotation. An example of the command line to create the bed files is shown below.

*plink --bfile Coreexome_genotype *
*--extract snp_passquality.txt --flip fliplist.txt *
*--keep sample_keep.txt --set-hh-missing --chr 1 *
--make-bed -–out Coreexome_chr1

d. Pre-phasing

The bed file creation is followed by pre-phasing with SHAPEIT2 haplotype estimation tool [85]. SHAPEIT2 could get the best guess haplotypes based on the input bed files. Then the best guess haplotypes were used by IMPUTE2 [86] to perform imputation. An example of the command line to run pre-phasing is shown below.

*shapeit2 -B Coreexome_chr1 *
*-M genetic_map_chr1_combined_b37.txt *
*-O Coreexome_chr1.haps.gz Coreexome_chr1.sample.gz *
-S 200 –T 3 -L shapeit_chr1.log

e. Reference panel

The reference panel data was downloaded from the September 2013 release of 1000 Genomes on the IMPUTE2 website and the previous March 2012 version was used for the X chromosome. As 1000 Genomes sequencing data generally has low coverage [87], the variants with very low frequency especially many singletons are likely to be

genotyping errors. To avoid imputation errors caused by very low frequency variants, imputation was only performed on the variants with at least two copies of the minor allele in EUR and AFR samples in 1000 Genomes project. According to the number of samples in EUR and AFR group, the filtering cutoff of minor allele frequency of EUR was 0.0026 and of AFR was 0.004. All three variant types (SNPs, indels, and SVs) were included in the imputation. EUR and AFR refer to European and African ancestry individuals.

f. Imputation

An example of the command line to run imputation is shown below.

*impute2-use_prephased_g -m genetic_map_chr1_combined_b37.txt *
-h
*ALL.chr1.integrated_phase1_v3.20101123.snps_indels_svs.genotypes.nosing.haplotypes.gz *
-l
*ALL.chr1.integrated_phase1_v3.20101123.snps_indels_svs.genotypes.nosing.legend.gz *
*-int 0 5000000 -buffer 500 -allow_large_regions *
*-known_haps_g Coreexome_chr1.haps.gz *
*-filt_rules_l 'eur.maf<0.0026' 'afr.maf<0.004' *
-o Coeexome_imputed_chr1_set1.gprobs -os 0 2 –o_gz
-i Coreexome_imputed_chr1_set1.metrics –verbose

TNF-α Study

TNF-α was measured from buffy coat samples isolated from peripheral blood. Buffy coats contain most of the white blood cells and platelets after density gradient centrifugation of blood. The unit of TNF-α was pg/ml. There were 266 Caucasians (Europeans) in the CHDWB cohort whose TNF-α levels were available. The TNF-α level was first transformed to the natural logarithm. Nine samples whose ln(TNF-α) level was larger than 4 standard deviation from mean levels were excluded. Therefore, there were

257 samples included in the study which were comprised of 98 males and 159 females. The age ranged between 19 and 82 with a mean of 50.1 and standard deviation of 10.9.

The first visit (baseline) TNF-α ranged from 0.1 pg/ml to 36.2 pg/ml with mean of 4.1 pg/ml and standard deviation of 3.3. The covariates were sex, age, age^2, and BMI. BMI has a mean of 26.8 kg/m^2 and a standard deviation of 5.0. The first step was to run a linear regression on ln(TNF-α) while adjusting for covariates. The regression was performed on both sexes combined, and each sex separately. The regression model for combined men and women was: ln(TNF-α) = age+age^2+BMI+sex. The modes for each sex were: ln(TNF-α) = age+age^2+BMI. The residuals for each model were saved as the phenotype to be used in the association analysis.

44 candidate SNPs identified in the discovery phase of the TNF-α meta-analysis were provided by Bram Prin's group at the University of Groningen. There were 6 SNPs already genotyped on the arrays, while the remaining 38 SNPs were imputed using IMPUTE2 as described above. Association tests were performed by SNPTEST (v2.5b) [88] using the frequentist association test with additive genotype dosages assumed.

Subsequently, I performed a longitudinal analysis, using 815 total measures of TNF-α. Excluding samples with no covariate information, 778 total measures remained, but of these, 19 ln(TNF-α) measures deviated by more than 4 standard deviation units of the mean and were removed. The individuals who only had a single visit were also removed. If an individual had more than 3 visits, I used the first 3 visit measures. In the end, there were 583 measures from 218 Caucasians (84 men and 134 women). The additive genotype dosage was calculated using the formula: dosage=$1*p_{ij1}+2*p_{ij2}$. P_{ij1} and p_{ij2} are the probability of genotypes AB and BB corresponding to IMPUTE2 format with coded allele (the same coded allele as provided by Bram Prin's group) as allele B. A linear mixed effect model in R was performed as

$$\ln(\text{TNF-}\alpha) = \mu + \text{genodosage} + \text{visit} + \text{indivID} + \text{age} + \text{age}^2 + \text{BMI} + \text{sex} + \varepsilon$$

where visit is an ordinal variable , indivID is a random effect, μ the grand mean, ε the residual error assumed to be normally distributed with a mean of zero, and sex was excluded for analyses of women and men separately.

1000G BMI/CRP Study

There were 266 samples whose BMI and CRP and covariates age, sex, BMI were all available. They were composed of 204 Caucasians and 62 Africans, age ranging between 22 and 76, including 77 males and 189 females. Imputation was performed using IMPUTE2 as above. In total, there were 8,970,590 variants (including SNPs and Indels) included in the study.

CRP ranged from 0.1 to 5.5 mg/L with a mean of 0.39 and a standard deviation of 0.54. The first step was to perform regression models on ln(CRP) while adjusting for covariates. Two regression models were used. The first one was fitting ln(CRP) with covariates age, sex, and the first ancestry principle component assessed from the genotype matrix. The second model was to fit ln(CRP) with the covariates age, sex, the first ancestry principle component as well as BMI. Then association tests were performed using the residuals from the two models separately, with imputed genotypes in each chromosome. SNPTEST v2 was used to perform the association test on autosomes and the X chromosome in females. The study was performed with two methods – “threshold” and “expected”, both under additive models. The “threshold” method uses intensity data for genotype determinations and was performed with 0.8 as the genotype threshold. The “expected” method was performed with expected genotype counts. For the X chromosome in males, GWAS was performed in R using a linear model with "threshold" and "expected" methods similar to SNPTEST, but with the difference that the "threshold" method was performed with 0.5 as genotype threshold (to simply avoid NA in the analysis), and for SNPs on the X chromosome, one copy of the coded allele was coded as 2.

BMI ranged from 18.4 to 50.9 kg/m^2 with a mean of 28.7 and a standard deviation of 5.7. BMI was inverse normally transformed ("inBMI") using the "rntransform" function in GenABEL. The regressions on inBMI were performed by adjusting covariates, also with two models. The first model was to regress inBMI on covariates age, sex, and the first ancestry principle component. The second model was to regress inBMI on covariates age, sex, the first ancestry principle component, and lnCRP. The association between residuals and genotypes was then performed as for ln(CRP).

Results

Imputation

Table 3.2 lists the number of imputed SNPs by chromosome based on the indicated number of SNPs on the array and number used for imputation after filtering. As seen by contrasting the top and bottom panels, almost identical numbers of imputed SNPs were obtained, even though the Batch 1 genotyping quality was lower than Batch 2 (Table 3.3), due to a problem with reagents.

The qualities of imputation were assessed based on quality data given by IMPUTE2 such as "info" which represents the imputation certainty and "concordance" which shows the concordance between imputed genotypes and original genotypes for one SNP. Table 3.3 summarizes the quality metrics, based on contrast of imputed genotypes with a set of masked SNPs, indicating overall concordance over 95% for SNPs with MAF<0.05 and 92% for common variants. Figure 3.1 shows that the confidence in imputation increases as MAF increases, as expected, with a plateau after MAF ~ 0.2. Figure 3.2 summarizes confidence scores by chromosome showing slight reduction in scores for the shorter chromosomes, possibly due to reduced efficiency of long range phasing.

Table 3.2 Imputation variant summary. Top: batch I. Bottom: batch II.

Chromosome	Study	Imputation	Imputation
1	44,777	18,704	668,341
2	37,164	18,693	720,978
3	31,241	15,568	618,134
4	24,671	13,772	637,124
5	25,776	13,919	558,284
6	28,390	14,648	593,657
7	24,235	12,529	511,581
8	21,294	12,082	476,431
9	20,558	10,193	377,977
10	21,897	11,883	441,860
11	27,634	11,681	426,805
12	24,172	11,027	416,516
13	12,851	8,202	320,765
14	15,591	7,471	286,850
15	15,689	6,970	250,060
16	18,281	7,610	268,571
17	20,249	6,926	235,730
18	10,519	6,540	250,195
19	21,046	5,891	206,783
20	12,531	6,015	193,171
21	6,178	3,278	124,885
22	8,826	3,674	123,724
X	13,026	7,227	272,090
Total	486,596	234,503	8,980,512

Chromosome	Study SNPs	Imputation	Imputation
1	48,430	21,420	668,341
2	41,224	22,017	720,978
3	34,656	18,469	618,134
4	28,550	17,112	637,124
5	28,894	16,506	558,284
6	33,678	19,093	593,657
7	27,039	14,849	511,581
8	23,840	14,231	476,431
9	22,681	11,963	377,977
10	24,059	13,689	441,860
11	30,020	13,506	426,805
12	26,540	12,908	416,516
13	14,900	10,016	320,765
14	17,156	8,722	286,982
15	17,020	8,052	250,060
16	19,434	8,501	268,571
17	21,346	7,622	235,730
18	11,951	7,804	250,195
19	21,910	6,379	206,783
20	13,380	6,688	193,171
21	6,887	3,893	124,872
22	9,215	3,956	123,724
X	12,675	7,028	267,913
Total	535,485	274,424	8,976,454

Table 3.3 Quality metrics for all masked SNPs, dichotomized into groups of MAF < 0.05 vs. MAF ≥ 0.05. Top: batch I. Bottom: batch II.

MAF (in study samples)	Number of SNPs	Mean (Median) of Overall Concordance	Mean (Median) of empirical dosage r2
<0.05	9,695	0.96 (0.987)	0.667 (0.759)
>=0.05	224,808	0.925 (0.954)	0.839 (0.899)

MAF (in study samples)	Number of SNPs	Mean(Median) of Overall Concordance	Mean(Median) of empirical dosage r2
<0.05	8,935	0.976 (0.987)	0.729 (0.823)
>=0.05	265,489	0.934 (0.965)	0.866 (0.925)

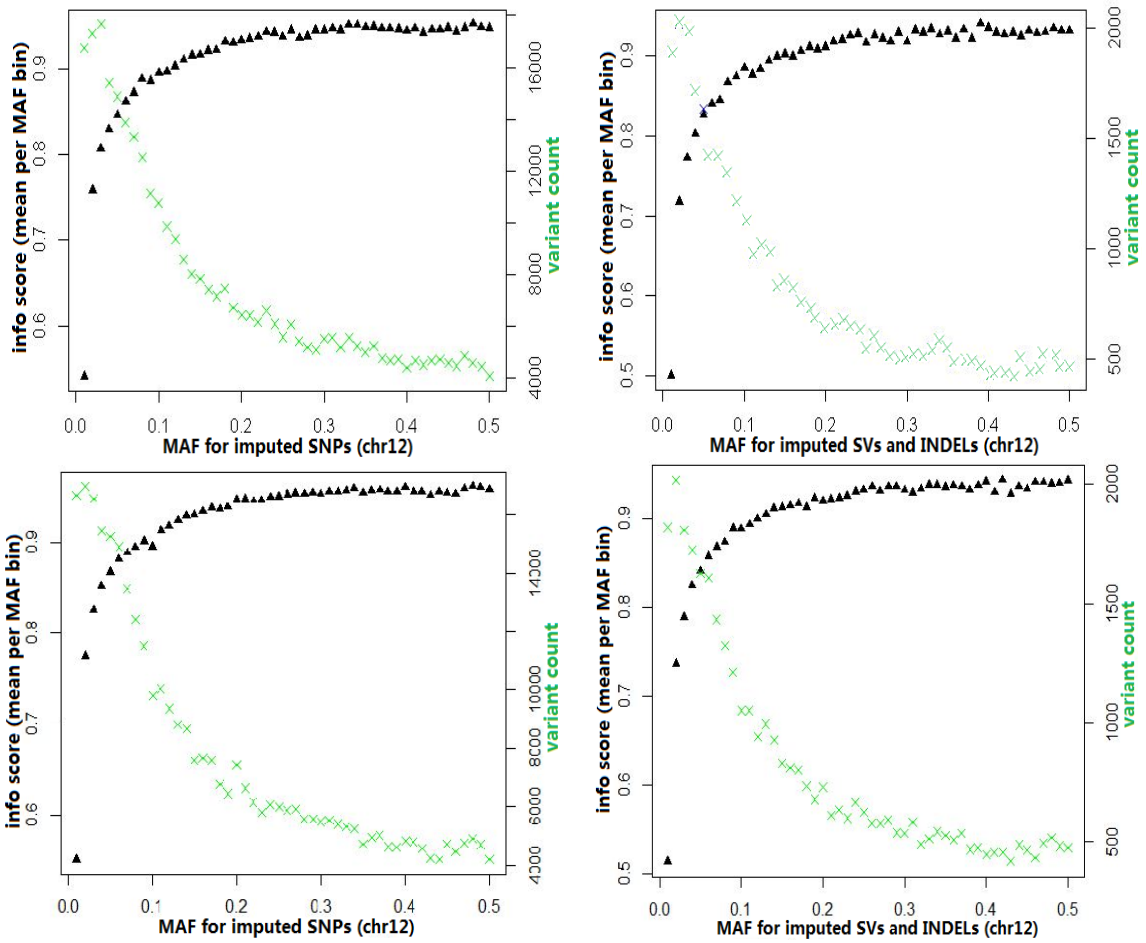

Figure 3.1 Summaries of quality metrics at imputed variants: SNPs, SVs, and indels in chromosome 12 (chromosome randomly picked, for simplicity). In each plot, imputed variants are binned by MAF (0.01 intervals) along the x-axis and then mean "info" score per bin is plotted on the y-axis. Left panel is for SNPs and right panel for indels and SVs. The secondary y-axes indicate the count of variants in each MAF bin. Top: batch I. Bottom: batch II.

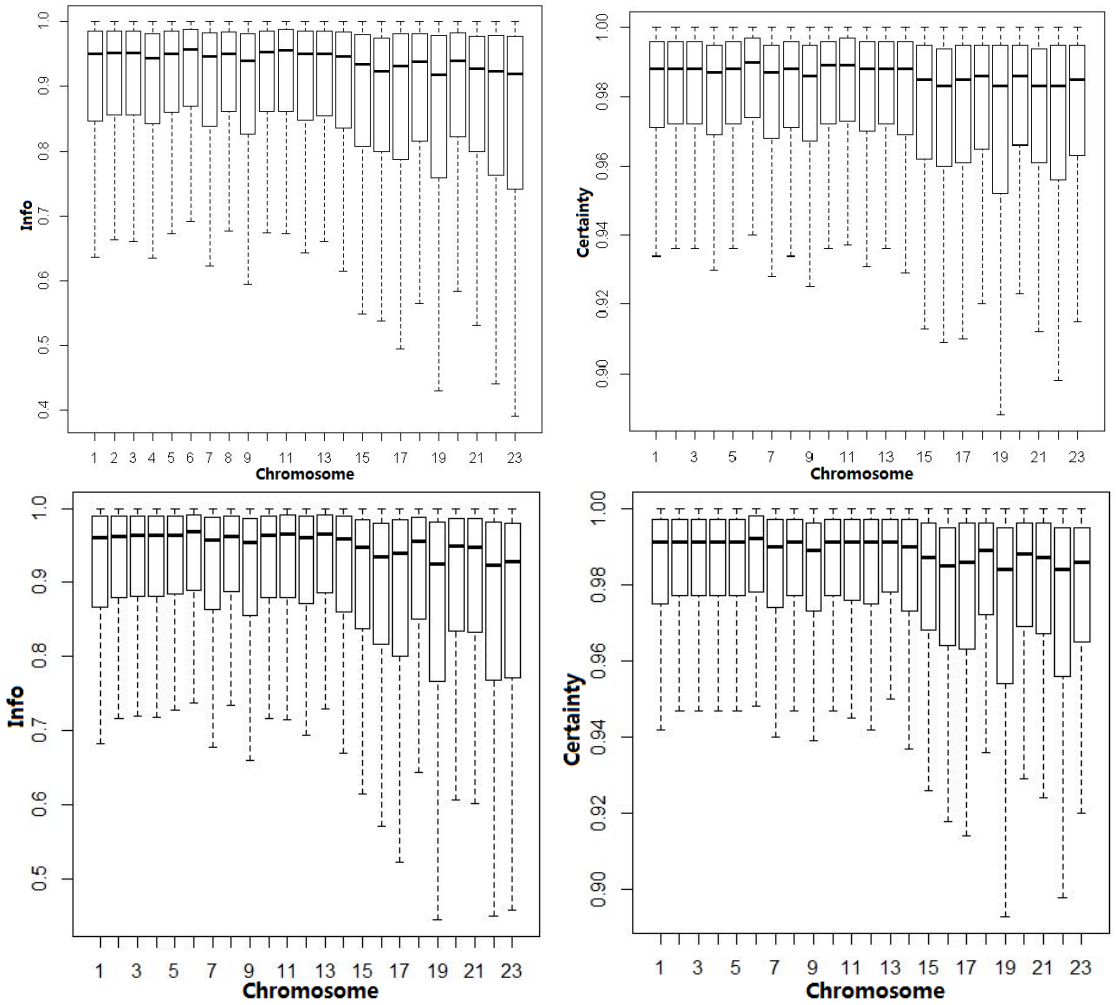

Figure 3.2 A comparison of imputation quality metrics by chromosome for all imputed SNPs. "info" in left panel and "certainty" in right panel for SNPs. Outlier values are not displayed in these box plots. On the x-axis, "23" denotes the X chromosome. Top: batch I. Bottom: batch II.

TNF-α

Overall, compared to the effect directions reported in the discovery phase meta-analysis, 30, 25 and 26 out of the 44 SNPs had the same effect direction for baseline concentrations of TNF-α in the combined sample, and in men and women, separately. By contrast, 27, 27 and 25 out of the 44 SNPs had the same effect direction for longitudinal analyses of the combined sample, and men and women analyzed separately.

According to the association significance in the discovery phase, a priority list for the replication study was provided by Bram Prin's group with 11 SNPs in Tier 1 and 9 SNPs in Tier 2. Table 3.4 shows that all of the significant SNPs in the Tier 1 and Tier 2 groups (the highest confidence SNPs) had the same effect direction with the meta analysis overall direction. However, rs13112532, which showed the strongest association in our cohort, presented in the opposite direction. The association between TNF-α and genotype appears to be quite robust in the baseline and longitudinal TNF-α analyses, as shown in Figure 3.3. The top panels show that the relationship between the effect size estimates longitudinally and at baseline are always linear, although there is more variability in the negative log p-value estimates.

Table 3.4 TNF-a significant association p-values. (-) means opposite direction in comparison with discover phase.

SNP	Baseline_ total	Baseline_ men	Baseline _women	Longitudinal _total	Longitudinal _men	Longitudinal _women	
rs1089208	0.0231	0.0249		0.0233	0.0341		Tier 1
rs1311253	7.31E-5(-)		2.07E-	2.13E-4(-)		9.35E-4(-)	
rs644234	0.0254	0.0301		0.0215	0.0225		Tier 1
rs9940180	0.0321		0.0361	0.0321		0.0357	
rs1077462		0.0421			0.0463		
rs3184504		0.0460			0.0401		Tier 2
rs1320731	0.0231		0.0132	0.0302		0.0144	Tier 1
rs4779129	0.0404	0.0321	0.0483	0.0487	0.0354	0.0464	
rs7182229				0.0459		0.0477	Tier 1
rs2131355					0.0384(-)		

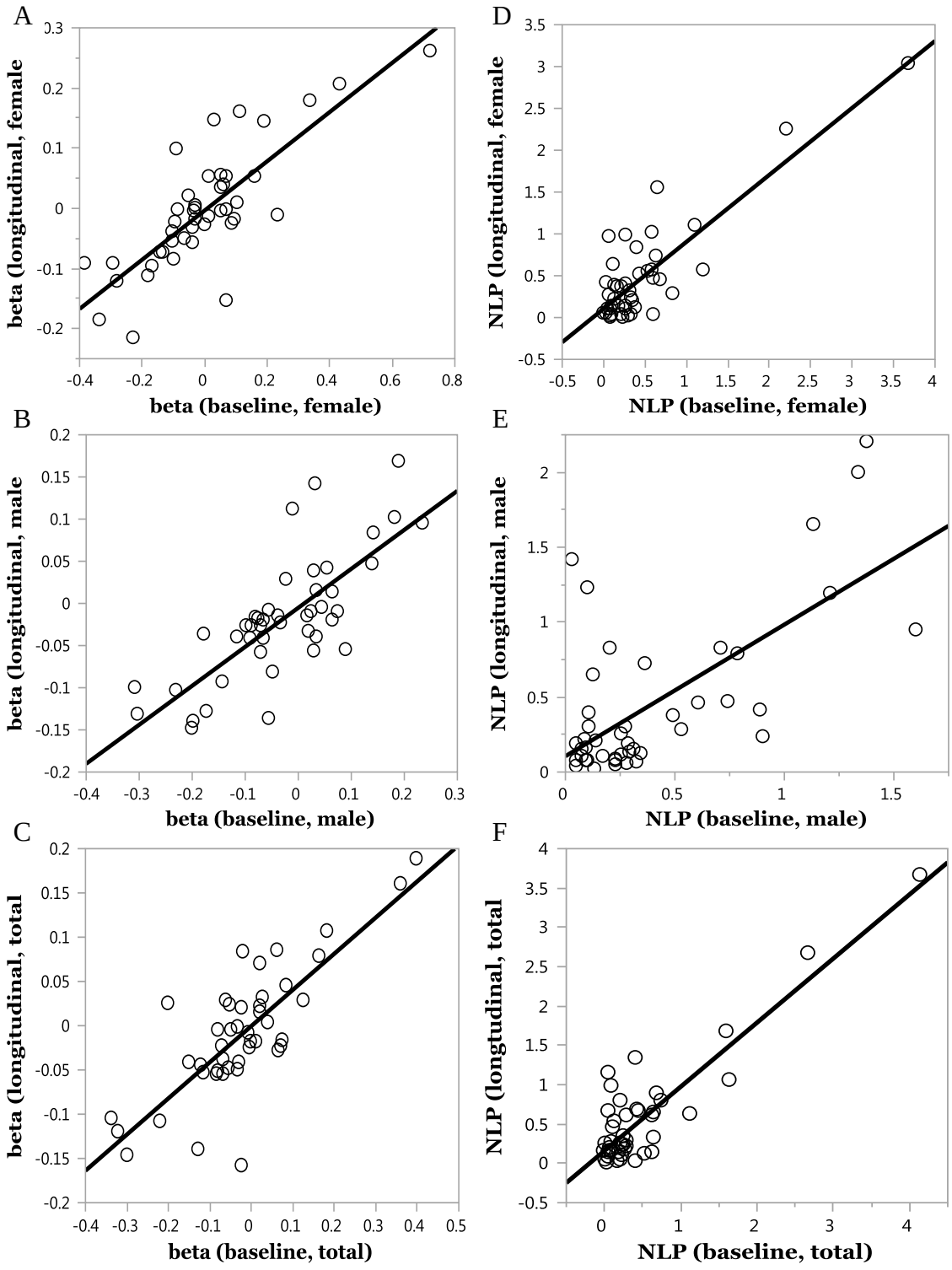

Figure 3.3 beta and p-value comparison between TNF-a baseline and longitudinal study. Beta comparison in female (A), male (B) and all samples (C). NLP ($-\log_{10}p$) comparison in female (D), male (E) and all samples (F).

1000G BMI/CRP

From Table 3.5, BMI has approximately 500 significant associated SNPs (defined as p-value < 0.0001) by using either expected or threshold method. CRP has approximately 1000 significant associated SNPs by those two methods. Those two methods show around 60% overlap in both BMI and CRP studies. In addition, 216 SNPs were overlapped between unadjusted expected method and adjusted expected method for BMI. 217 SNP were overlapped between unadjusted threshold method and adjusted expected method for BMI. As to CRP, the two overlapped SNP numbers were 421 and 455 respectively. Therefore, the overlapping proportion was approximately 40% for unadjusting and adjusting CRP/BMI. There is less than 10 SNPs overlap between SNPs of BMI-associated and CRP-associated no matter what the adjustment is.

Table 3.5 Number of significant SNPs (p-value < 0.0001) for different association methods and adjustment models.

	BMI		CRP	
	adjust CRP	**unadjust CRP**	**adjust BMI**	**unadjust BMI**
expected	510	473	980	1091
threshold	476	546	1115	1161
in common	320	309	592	683

Discussion

There were 10 SNPs found to be associated significantly with TNF-α at p<0.05 using 257 Caucasians in the CHDWB cohort. The lack of significant association for most of the other 34 SNPs is almost certainly due to the small sample size. However, I calculated the HWE_Pvalue for all 44 SNPs with best guess genotype from the IMPUTE2 result and observed that rs107744774 (HWE $p=3\times10^{-7}$), rs10892063 ($p=1\times10^{-5}$), rs17377218 ($p=5\times10^{-9}$) all failed the HWE test, which may have contributed to their lack of association detected with TNF-α.

In general, the threshold method detected more significant SNP associations. Compared to the expected method, which uses the expected dosage of coded alleles (any continuous value between 0 and 2) according to the likelihood of each genotype, the threshold method uses the best guess genotype (only 0, 1 or 2 possible). In this way, the expected method is more stringent when fitting the models with phenotype. My view is that the expected method is more reliable as it takes the likelihood of each genotype into consideration and avoids the bias from arbitrary setting of the best guess genotype likelihood threshold.

Preliminary clinical data have suggested TNF-α may be involved in the pathogenesis of a variety of human diseases including autoimmune diseases and septic shock and inhibition of TNF-α may take a role in disease prevention and treatment [89, 90]. Anti-TNF-α therapy has been studied for treatment of diseases such as rheumatoid arthritis [91]. GWAS on TNF-α would make a better understanding of the genetic basis of TNF-α response, and may provide candidate gene targets for pharmacogenetic considerations in TNF-α related diseases. Imputation on array genome data has been proved to provide a lot of information for GWAS on TNF-α. However, as with the development of sequencing techniques, more genetic sites associated with TNF-α including rare variants would be likely to be found in future.

In addition, there are studies showing relationship between the levels of CRP and BMI. However, the causal relationship is unclear. The knowledge of genetic variants associated with CRP and BMI could play a role in elucidating the causal relationship between those two traits with Mendelian randomization, but this is only possible in the very large meta-analysis dataset. In addition, CRP is an acute-phase protein, whose level rises in response to inflammation and tissue damage [92, 93]. BMI is also a useful predictor of health status. It has been shown that many metabolic and disease outcomes are related to elevated BMI. For example, coronary artery disease, and type 2 diabetes [94, 95] have been found to be more prevalent in people with high BMI. The GWAS

analysis of these two traits could therefor provide candidate variants that may have implications for development of possible genetic treatment to their related diseases.

While GWAS analysis on disease-associated traits is very helpful in finding genetic positions that are relevant to disease, endophenotype studies such as the ones reported in this chapter can also shed light on disease mechanisms and influence drug design. This work has shown that genotyping data, in together with imputation of missing genotypes, is an efficient way to perform GWAS analysis. However, given the bias and incompleteness of genotyping probes, it would be more accurate and more informative if sequencing data were used for fine mapping in GWAS.

According to the results, CRP and BMI have several hundred associated common SNPs. The TNF-α discovery phase study found a few hundred candidate SNPs associated with the cytokine levels. However, previous studies [96, 97] have shown that hundreds of GWAS common SNPs usually only explain less than 20% of the variance. This has also been pointed out in Chapter 2. It implies that other factors also contribute to the trait variance, for example gene-environment interaction, epigenetic effects, and also rare variants, and common variants of very small effect. In the next chapters, I will focus on rare regulatory variants and investigate their association with gene expression levels.

CHAPTER 4

DETAILED SEQUENCING OF 472 PROMOTER REGIONS

Introduction

Along with the development of sequencing technologies, genomic variant detection has become more and more accessible. The large human genome sequencing Consortium projects, such as the 1000 Genomes project, have particularly helped researchers gain a deeper knowledge of human genome variation. Improved understanding of genomic variant structure has also benefitted our insight into disease susceptibility and causation. GWAS have successfully identified thousands of genetic variants which show association with common diseases.

Common variants that have been detected as being associated with complex diseases are known to be considerably more enriched for regulatory than coding variants [37, 98-100]. Many analyses utilizing data generated by the ENCODE project have shown that specific classes of non-coding variants can have complex impacts on cell function and phenotype. According to GWAS analysis, 11% of GWAS hits lie in coding regions [101] while 57% of GWAS hits lie in broadly-defined DHS regions which span 42% of the genome [99]. This implies that non-coding regions have a large impact on gene functions and in turn phenotypes, and motivates better understanding of non-coding regions..

Here, taking the advantage of targeted resequencing technique, I describe an investigation of the distribution of rare variants in promoter regions of 472 genes. In the following chapter, I will discuss investigation of the association between cis-acting regulatory rare variants and quantitative gene expression traits. In this chapter, the focus is on the distribution of rare variants with the aim of generating a more explicit picture of genomic variants construction. The analysis includes discussion of the impact of variant

calling methods on detection, and comparisons of variant counts between different subsets of genes as well as between human populations. I also include a comparison of sequence diversity in regulatory regions versus in coding regions, taking advantage of public available 1000 Genomes data, and between genes which are thought to be disease-related (or potentially disease-related) and genes which do not harbor disease SNPs.

A genome-wide scoring system called (Residual Variation Intolerance Score) that ranks human genes in terms of their intolerance to standing functional genetic variation in the human population was introduced by Petrovski et al [102]. They used empirical single nucleotide variant data from the NHLBI Exome Sequencing Project and ranked the genes based on whether they have more or less functional genetic variation relative to the genome wide expectation. Their work has shown that the genes which harbor fewer or more common functional variants may be more or less prone to cause certain kinds of diseases in healthy individuals. Here, I utilized the targeted sequencing data to detect the intolerance of rare variants in promoter regions and compared the results with Petrovski et al's result with respect to disease-related genes.

Benefitting from the accessible genotyping and expression data of the CHDWB study cohort, eQTL analysis was performed. I investigated the rare variant distribution in genes with strong common eQTL and genes without common eQTL. As studies have shown that trait – associated SNPs are more likely to be eQTLs, the genes that show a strong signal of eQTL are more likely to be correlated with diseases. In this way, comparing the rare variant counts in promoter regions shows the intolerance of rare variants between different potential disease – associated gene status.

I also used Metabochip and Immunochip identities to investigate the difference of intolerance to rare variants in promoter regions in different gene sets. The "Metabochip", a custom Illumina iSelect genotyping array, was designed for the genome-wide association tests of diseases and traits relating to metabolic, atherosclerotic and

cardiovascular traits [103]. It contains approximately 200,000 SNPs including variants with low frequency identified by the 1000 Genomes project. Similarly, the "Immunochip" is an Illumina Infinium SNP array which was designed for association studies of autoimmune and inflammatory diseases [104]. The genes which harbor SNPs represented on the Metabochip are with high probability related to metabolic disease, atherosclerosis, and cardiovascular diseases, while, the genes which harbor SNPs on the Immunochip are potentially related to autoimmune and inflammatory diseases.

While different sequencing techniques and platforms will lead to different sequencing qualities, affecting variant calling accuracy, different variant calling algorithms could also affect the variant calling result enormously. The commonly used software packages for variant calling include the Genome Analysis Toolkit (GATK) [17], SOAPsnp [105], VarScan [19], and ATLAS [106]. Here I compared GATK and VarScan. Developed by the Broad Institute, GATK is one of the most popular methods for variant calling using aligned reads. It is designed in a modular way and is based on the MapReduce functional programming approach. Developed by the Genome Institute at Washington University in St. Louis, VarScan is an open source tool for short read variant detection of SNPs and indels that is compatible with multiple sequencing platforms and aligner algorithms such as Bowtie [14] and Novoalign (http://novocraft.com). Some variants were then verified by the "gold standard" method of Sanger sequencing.

Accurate detection of rare variants also requires that the read depth should be high enough to avoid false positives caused by sequencing errors and false negatives caused by insufficient coverage to detect heterozygotes. The ideal alternate allele proportion is 0.5 for heterozygous site, and 1 for homozygous site. Figure 4.1 shows the alternate allele proportion over read depth of the bases. It shows that the alternate allele proportion is consistently approximately 0.5 or 1 when the read depth is larger than 100. This empirically demonstrates that high read depth is also appropriate and necessary in a study focused on rare SNPs.

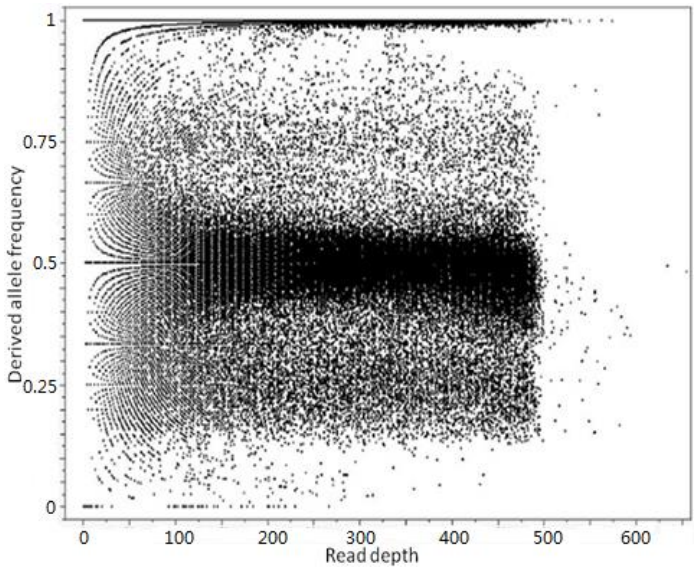

Figure 4.1 alternate allele proportion versus read depth for each base.

The variant profile was expected to be somewhat different among different genes and populations. Analysis of the 1000 Genomes project phase I based on 1092 sample whole genome sequencing data, common and rare variants all show different distributions among populations. More than 50% of rare variants (MAF less than 5%) were observed in just a single population. Furthermore, rare variants detected in individuals with African ancestry were three times as prevalent as those in individuals of European or East Asian origin. In addition, individuals from all populations showed an enrichment of rare variants relative to some classes of neutral evolutionary model, which has been attributed to population size explosion and accompanying geographic differentiation [107, 108].

Materials and Methods

410 Samples

The 410 Samples included in this study are listed in Table 4.1. They were composed of 297 Caucasian Americans, 85 African Americans and 18 Asian Americans. There were 274 females and 136 males. Age ranged from 19 to 83 with an average of 50.

Table 4.1 Anthropomorphic information for the 410 samples.

ID	AGE	GENDER	ANCESTRY	ID	AGE	GENDER	ANCESTRY	ID	AGE	GENDER	ANCESTRY
Sample1	51	F	AFR	Sample138	63	F	CAU	Sample275	48	M	CAU
Sample2	52	M	CAU	Sample139	48	F	CAU	Sample276	46	M	CAU
Sample3	50	M	CAU	Sample140	63	M	CAU	Sample277	52	M	CAU
Sample4	60	F	CAU	Sample141	48	F	CAU	Sample278	61	F	CAU
Sample5	59	F	CAU	Sample142	47	M	ASN	Sample279	39	M	CAU
Sample6	69	F	CAU	Sample143	40	F	CAU	Sample280	45	F	AFR
Sample7	46	M	CAU	Sample144	59	F	CAU	Sample281	41	F	CAU
Sample8	55	F	AFR	Sample145	48	F	AFR	Sample282	61	F	CAU
Sample9	57	F	AFR	Sample146	59	F	CAU	Sample283	49	F	AFR
Sample10	37	M	CAU	Sample147	58	M	AFR	Sample284	55	F	CAU
Sample11	59	M	CAU	Sample148	33	M	CAU	Sample285	46	F	AFR
Sample12	44	F	CAU	Sample149	36	F	CAU	Sample286	29	F	CAU
Sample13	42	F	AFR	Sample150	66	M	CAU	Sample287	45	F	AFR
Sample14	61	F	CAU	Sample151	53	F	ASN	Sample288	47	M	AFR
Sample15	49	F	AFR	Sample152	63	F	AFR	Sample289	47	F	AFR
Sample16	50	F	CAU	Sample153	65	M	CAU	Sample290	40	M	CAU
Sample17	56	M	CAU	Sample154	52	F	AFR	Sample291	52	M	CAU
Sample18	59	F	AFR	Sample155	37	F	CAU	Sample292	49	F	AFR
Sample19	52	F	CAU	Sample156	43	M	AFR	Sample293	53	F	CAU
Sample20	61	M	CAU	Sample157	34	F	AFR	Sample294	53	F	CAU
Sample21	54	M	CAU	Sample158	40	F	CAU	Sample295	35	F	CAU
Sample22	54	F	AFR	Sample159	79	M	CAU	Sample296	54	F	CAU
Sample23	53	F	AFR	Sample160	70	M	CAU	Sample297	47	M	CAU
Sample24	55	F	CAU	Sample161	30	F	CAU	Sample298	50	F	AFR
Sample25	54	F	CAU	Sample162	43	F	AFR	Sample299	51	F	CAU
Sample26	54	F	AFR	Sample163	59	M	CAU	Sample300	31	F	CAU
Sample27	60	F	CAU	Sample164	47	M	CAU	Sample301	44	F	AFR
Sample28	47	F	CAU	Sample165	58	M	CAU	Sample302	68	F	CAU
Sample29	56	M	CAU	Sample166	51	M	ASN	Sample303	57	M	CAU
Sample30	53	F	CAU	Sample167	50	F	AFR	Sample304	56	M	ASN
Sample31	59	F	CAU	Sample168	50	F	ASN	Sample305	66	M	CAU
Sample32	57	M	CAU	Sample169	54	F	AFR	Sample306	48	F	AFR
Sample33	60	M	CAU	Sample170	49	F	AFR	Sample307	56	F	AFR
Sample34	48	M	CAU	Sample171	40	F	CAU	Sample308	59	F	CAU
Sample35	36	F	AFR	Sample172	46	F	CAU	Sample309	51	F	AFR
Sample36	54	F	CAU	Sample173	64	F	CAU	Sample310	54	F	CAU
Sample37	59	F	AFR	Sample174	44	F	CAU	Sample311	51	M	CAU
Sample38	48	F	CAU	Sample175	54	F	CAU	Sample312	62	F	AFR
Sample39	55	M	CAU	Sample176	60	M	CAU	Sample313	59	M	CAU
Sample40	40	F	CAU	Sample177	55	M	CAU	Sample314	57	M	CAU
Sample41	31	F	CAU	Sample178	52	F	CAU	Sample315	58	M	AFR
Sample42	35	F	AFR	Sample179	75	M	CAU	Sample316	35	F	CAU
Sample43	36	F	CAU	Sample180	57	F	CAU	Sample317	48	F	CAU
Sample44	50	F	AFR	Sample181	53	F	AFR	Sample318	55	F	CAU
Sample45	50	M	CAU	Sample182	49	F	AFR	Sample319	56	F	CAU
Sample46	36	F	CAU	Sample183	64	M	CAU	Sample320	46	M	CAU
Sample47	52	F	ASN	Sample184	41	M	CAU	Sample321	55	F	CAU
Sample48	58	F	CAU	Sample185	57	M	CAU	Sample322	55	F	ASN
Sample49	33	F	CAU	Sample186	41	F	CAU	Sample323	40	F	CAU
Sample50	37	F	AFR	Sample187	43	M	AFR	Sample324	57	F	CAU
Sample51	55	M	CAU	Sample188	48	F	AFR	Sample325	63	M	CAU
Sample52	59	M	CAU	Sample189	55	F	AFR	Sample326	41	F	AFR
Sample53	52	F	CAU	Sample190	45	M	CAU	Sample327	60	M	CAU
Sample54	57	M	CAU	Sample191	58	M	CAU	Sample328	26	M	CAU
Sample55	41	M	CAU	Sample192	57	M	CAU	Sample329	56	F	CAU
Sample56	59	F	CAU	Sample193	66	F	CAU	Sample330	19	M	ASN
Sample57	50	F	AFR	Sample194	57	M	CAU	Sample331	60	M	CAU
Sample58	60	F	CAU	Sample195	41	F	CAU	Sample332	53	F	CAU
Sample59	60	F	CAU	Sample196	64	F	CAU	Sample333	61	F	CAU
Sample60	29	F	AFR	Sample197	39	F	CAU	Sample334	59	F	CAU
Sample61	56	F	AFR	Sample198	39	F	CAU	Sample335	44	M	CAU
Sample62	74	M	CAU	Sample199	29	M	CAU	Sample336	66	M	CAU
Sample63	55	F	CAU	Sample200	52	M	CAU	Sample337	49	M	CAU
Sample64	37	F	CAU	Sample201	47	M	ASN	Sample338	71	F	CAU
Sample65	50	F	CAU	Sample202	37	M	CAU	Sample339	41	M	CAU
Sample66	59	F	CAU	Sample203	56	F	CAU	Sample340	50	F	CAU
Sample67	46	F	CAU	Sample204	66	M	CAU	Sample341	42	M	CAU
Sample68	58	M	CAU	Sample205	40	F	AFR	Sample342	44	F	CAU
Sample69	31	F	CAU	Sample206	35	F	CAU	Sample343	41	F	CAU
Sample70	58	M	CAU	Sample207	83	M	CAU	Sample344	36	F	AFR
Sample71	57	M	CAU	Sample208	38	F	CAU	Sample345	50	F	CAU

Table 4.1 (continued)

Sample72	68	F	CAU	Sample209	67	M	CAU	Sample346	59	M	ASN
Sample73	69	M	CAU	Sample210	45	F	AFR	Sample347	45	F	CAU
Sample74	47	F	CAU	Sample211	56	F	AFR	Sample348	57	M	CAU
Sample75	79	F	CAU	Sample212	51	F	CAU	Sample349	41	F	CAU
Sample76	56	F	AFR	Sample213	34	F	CAU	Sample350	46	F	CAU
Sample77	42	M	CAU	Sample214	31	F	CAU	Sample351	46	F	AFR
Sample78	56	F	CAU	Sample215	54	F	CAU	Sample352	44	M	CAU
Sample79	44	F	CAU	Sample216	59	F	AFR	Sample353	37	F	AFR
Sample80	53	M	CAU	Sample217	63	F	CAU	Sample354	37	F	CAU
Sample81	57	F	CAU	Sample218	53	F	CAU	Sample355	38	F	CAU
Sample82	57	F	CAU	Sample219	44	F	AFR	Sample356	37	M	CAU
Sample83	50	F	CAU	Sample220	50	F	CAU	Sample357	36	F	CAU
Sample84	46	M	CAU	Sample221	40	F	CAU	Sample358	51	M	CAU
Sample85	35	F	CAU	Sample222	54	M	CAU	Sample359	37	F	CAU
Sample86	48	F	CAU	Sample223	61	M	CAU	Sample360	45	F	AFR
Sample87	45	F	AFR	Sample224	55	F	CAU	Sample361	35	F	CAU
Sample88	39	F	CAU	Sample225	55	M	AFR	Sample362	61	F	CAU
Sample89	54	F	AFR	Sample226	54	F	CAU	Sample363	57	M	CAU
Sample90	42	M	ASN	Sample227	20	F	AFR	Sample364	62	M	CAU
Sample91	37	F	CAU	Sample228	65	F	CAU	Sample365	42	F	CAU
Sample92	49	F	CAU	Sample229	53	M	AFR	Sample366	59	M	CAU
Sample93	40	M	CAU	Sample230	57	F	CAU	Sample367	44	F	CAU
Sample94	55	F	ASN	Sample231	43	F	CAU	Sample368	53	M	CAU
Sample95	57	M	AFR	Sample232	43	M	CAU	Sample369	43	F	CAU
Sample96	49	F	AFR	Sample233	57	F	CAU	Sample370	28	F	CAU
Sample97	35	F	AFR	Sample234	55	F	AFR	Sample371	58	M	CAU
Sample98	35	F	AFR	Sample235	37	F	CAU	Sample372	55	F	CAU
Sample99	46	M	ASN	Sample236	54	M	CAU	Sample373	39	F	CAU
Sample100	47	F	AFR	Sample237	49	F	CAU	Sample374	66	M	CAU
Sample101	36	M	CAU	Sample238	74	F	CAU	Sample375	55	M	CAU
Sample102	61	F	CAU	Sample239	50	F	CAU	Sample376	54	F	CAU
Sample103	62	F	CAU	Sample240	49	F	CAU	Sample377	52	F	CAU
Sample104	53	F	CAU	Sample241	54	F	CAU	Sample378	42	F	CAU
Sample105	50	M	ASN	Sample242	29	M	CAU	Sample379	53	M	CAU
Sample106	60	M	CAU	Sample243	26	F	AFR	Sample380	58	F	CAU
Sample107	63	M	CAU	Sample244	30	F	AFR	Sample381	55	M	CAU
Sample108	68	F	CAU	Sample245	36	F	CAU	Sample382	39	F	CAU
Sample109	31	F	AFR	Sample246	26	M	ASN	Sample383	63	F	CAU
Sample110	38	F	AFR	Sample247	54	F	CAU	Sample384	53	F	AFR
Sample111	55	M	CAU	Sample248	56	M	CAU	Sample385	63	M	CAU
Sample112	32	F	CAU	Sample249	49	M	CAU	Sample386	19	F	AFR
Sample113	56	F	CAU	Sample250	43	M	ASN	Sample387	50	F	CAU
Sample114	53	F	AFR	Sample251	47	F	AFR	Sample388	55	F	CAU
Sample115	40	F	CAU	Sample252	58	M	CAU	Sample389	57	F	CAU
Sample116	58	F	CAU	Sample253	37	M	AFR	Sample390	58	F	CAU
Sample117	39	M	CAU	Sample254	55	F	AFR	Sample391	49	F	CAU
Sample118	37	F	CAU	Sample255	57	M	CAU	Sample392	45	F	CAU
Sample119	64	M	CAU	Sample256	62	F	CAU	Sample393	36	M	CAU
Sample120	42	F	AFR	Sample257	60	F	CAU	Sample394	45	F	AFR
Sample121	43	M	CAU	Sample258	65	M	CAU	Sample395	44	F	CAU
Sample122	50	F	CAU	Sample259	48	F	AFR	Sample396	48	F	AFR
Sample123	35	M	CAU	Sample260	46	F	CAU	Sample397	43	F	CAU
Sample124	47	F	AFR	Sample261	35	F	CAU	Sample398	64	F	CAU
Sample125	53	M	CAU	Sample262	59	F	CAU	Sample399	60	M	CAU
Sample126	59	F	AFR	Sample263	61	F	AFR	Sample400	36	F	CAU
Sample127	34	F	CAU	Sample264	22	M	CAU	Sample401	54	F	CAU
Sample128	61	F	AFR	Sample265	58	M	CAU	Sample402	57	F	CAU
Sample129	28	M	CAU	Sample266	59	F	AFR	Sample403	52	F	AFR
Sample130	66	F	CAU	Sample267	30	F	AFR	Sample404	56	M	CAU
Sample131	41	M	AFR	Sample268	65	F	CAU	Sample405	54	F	AFR
Sample132	47	F	CAU	Sample269	62	F	CAU	Sample406	36	F	AFR
Sample133	54	F	CAU	Sample270	55	F	CAU	Sample407	55	M	CAU
Sample134	48	M	CAU	Sample271	39	F	CAU	Sample408	40	F	CAU
Sample135	49	F	CAU	Sample272	41	F	AFR	Sample409	56	F	ASN
Sample136	58	F	AFR	Sample273	39	F	ASN	Sample410	59	F	CAU
Sample137	46	F	AFR	Sample274	26	F	CAU				

Gene Selection

Table 4.2 lists the genes, which were selected based on the idea of having genes harboring common cis-eSNPs (and hence established to be genetically regulated), a subset of these also having a high probability of being related with disease traits, with the remainder being control genes. The common cis-eSNP results were based on imputed genotyping results from analyses described in the previous chapter. I also took advantage of Metabochip and Immunochip arrays to choose the gene that have Metabochip or Immunochip SNPs in the vicinity so that the genes had a high probability to be involved in some diseases or traits. The gene selection was performed in August 2012, but cis-eSNP status was updated as my dissertation studies proceeded.

Table 4.2 The 472 genes selected in the study.

ACADM	ACER3	ADCK3	ADK	ANXA11	AOAH	B4GALT4	BTN3A2	CARD8	CARD9
CDA	CDK10	CFDP1	CTSH	CWF19L1	DNAJC15	DR1	FCER1G	FDFT1	GAPT
GATAD2A	GATS	GNA12	HIST1H2BD	HPS1	HSD17B12	IQCB1	ITGAX	KCTD10	KIAA0368
KLHDC4	MBNL1	MED16	MFN2	MSRA	MTMR3	NT5C3	ORMDL3	PASK	PDCD4
PTGS2	RFWD3	RPA1	SF3A3	SIDT2	SIRPB1	SIRPG	SLC39A8	SLC40A1	SMAP1
SNX29	SP110	SPNS1	SUMF1	TATDN2	TBC1D15	TOMM7	TRAK1	TRPC4AP	TUFM
UBE2L3	UGDH	USP48	XRCC6BP1	ADSS	ALDH2	ALPL	APOL3	ARID5B	ARPP19
C1orf38	CD96	CDAN1	CLN3	CPPED1	CRISPLD2	CTDSP1	CTNNAL1	CTSO	DDX52
DEF6	DHRS3	DHX38	DIAPH2	DLAT	DOCK10	DOCK11	E2F6	ECHS1	EIF4G3
EPB41	EXOC4	F13A1	FAM49A	FAR2	GAB3	GMCL1	GNPAT	HBP1	HSPA4
IL18R1	IQGAP1	ITGAM	ITK	KIAA0319L	LILRB2	LMAN2L	LMNA	LRPPRC	M6PR
MAN2A2	MBNL2	MSI2	NEDD9	NUDT5	OXR1	PDK3	POR	RBBP4	RNF130
RNPS1	SERTAD2	SH2D1B	SNX14	SRP54	STAT3	STAT4	SUSD3	TAF1C	TAF1L
TBC1D2B	TMEM175	TPM1	TSSC1	UBR2	UQCC	USO1	USP4	VIM	WDFY2
ZAK	ZDHHC17	ZMIZ1	ZNF185	ZNF407	ABCC5	ACOX1	ACP2	ALS2	ATP13A1
C12orf35	C14orf102	C1orf86	C2orf28	C7orf25	CEP63	CRLS1	DHRS1	DUS2L	EEF1G
EMR3	ERP27	FEZ2	FN3KRP	FXYD5	HDDC2	KIAA1598	LACTB	LDLRAP1	LINS
MAU2	MEFV	MGST3	MRPS21	NSFL1C	PCMTD2	PIGQ	PYGB	RNF181	RPL13
RPS6KB2	RUFY1	SAMM50	SENP2	SF3A1	STAT6	STYXL1	SURF6	TAGLN	TMEM140
VAMP1	VAMP8	VASP	VEZT	ZMIZ2	IPP	FCRL3	AHSA2	CAPG	ASNSD1
UBA7	HCLS1	SRI	KRIT1	TRIM4	PILRB	ZYX	TNFSF8	RRP12	ZRANB1
PRDX5	ALDH3B1	FAM76B	TRAPPC4	PEX5	PWP1	POLR1D	ACTR10	GALC	CLCN7
DCTN5	GDPD3	ZFP90	CXCL16	SMARCE1	ZNF266	LRRC25	PLAUR	SYS1	LTN1
XBP1	ACTR6	ALKBH1	ASXL2	ATMIN	BSDC1	C14orf179	C2orf44	C6orf129	DNAJC8
E2F2	EIF2AK4	ERCC3	FBXO11	HADH	HNRNPC	HSPA9	ID3	ILF3	LARP4
LASS5	LFNG	MED22	MRPL17	MSH6	MYC	NDUFB10	NEDD8	NUFIP2	OGFOD1
OGFRL1	PACSIN2	PFDN1	PHF21A	PIP4K2B	PRRC1	RAB11FIP2	RAB8B	RHOT1	RNF135
SERPINB8	SESN3	SHROOM4	SLC35A3	STARD3NL	TRAF5	TSPAN33	UBN1	ZNF787	ZNF839
CPSF3L	MOBKL2C	SLC27A3	SLAMF7	TRAF3IP3	MKI67IP	ICOS	FRG1	DHX29	BRD8
DKFZp686I15217	TDP2	KATNA1	STOM	C9orf114	COBRA1	RPP38	HNRNPH3	SLC3A2	HSPA8
C12orf32	TUBA1B	ADPGK	SNRNP25	EIF4A1	STAT5A	C18orf21	VPS4B	RANBP3	GTF2F1
TMEM149	ALDH16A1	ZNF613	ADRM1	HSCB	TLR7	GSPT2	ATG4A	NAA10	ABHD10
ARL16	ATP5J2	ATP5S	ATPIF1	C14orf129	C17orf90	C1orf85	C8orf40	C9orf78	CALR
CAT	CCDC23	CCDC88B	CIB1	CKS2	CRIPT	CXCL5	DCXR	DSTYK	EIF2B2
EIF5	EPHX2	ETS2	EVI2A	GPATCH4	GPX7	H1F0	HEBP2	IKBIP	IL8
KIAA1737	LIPT1	MAD2L1BP	MR1	MRPL34	MRPL43	MRPL52	MRPL53	NOP10	NUDT18
NUDT2	NUP43	ORMDL1	PIGN	PPIL3	PRKAR1A	PTGDR	RPL14	RPL36AL	RPUSD4
RRM2B	SNAP29	TAPBPL	TFG	TRAPPC5	TRAPPC6B	ABHD8	ACAA2	AKR7A3	ALPP
AP1G2	AP1S1	ARAF	ARGLU1	AZI2	BEX2	BFAR	BIRC3	BRF2	BRMS1
BTK	C11orf17	C14orf142	C15orf63	C18orf10	C1orf123	CD160	CHRNB1	CIDECP	COMMD4
COMMD7	COMMD9	CUL4B	CXXC5	DDOST	DDX41	DGUOK	DRAM1	EBP	ECH1
EIF1AX	EIF2S3	FAM193B	GPAA1	GRK6	HARS2	HBQ1	HDHD1	HIF1AN	HSD17B11
HSP90AB1	KIAA1191	KLF4	LAPTM5	LONP2	LRRCC1	MAF1	MAGEH1	MAPK1IP1L	MPHOSPH10
MPL	NFE2L1	NONO	PCNA	PDCD2	PEA15	PFN1	PIGH	PPCS	RAB10
RAB24	RBMX2	RCE1	RPL10A	RPL4	RPL9	RPUSD3	SCAND1	SDAD1	SERTAD1
SETD3	SF3B4	SHCBP1	SNORA70	TCEAL8	TFE3	THAP7	TMEM199	TNFRSF4	TUBA1A
TXNIP	UBL4A	UTP18	UXT	WAS	WDR45	YIF1A	YIPF3	ZNF439	ZNF549
ZNF671	ZNF75D								

The 472 genes have been classified into 8 groups, color coded in Table 4.2, according to the following criteria:

1) Genes which have both Metabochip and Immunochip SNPs, and have cis-eSNPs inside the gene body and in the promoter regions. These 64 genes are marked as red in Table 4.1.

2) Genes which have both Metabochip SNPs and Immunochip SNPs inside the gene body and in the promoter regions, but don't have common cis-eSNPs. These 81 genes are marked as purple.

3) Genes which have Metabochip SNPs and common cis-eSNPs inside the gene body and in the promoter regions, but don't have Immunochip SNPs in the vicinity. These 50 genes are marked as green.

4) Genes which have Immunochip SNPs and common cis-eSNPs inside the gene body and in the promoter regions, but don't have Metabochip SNPs in the vicinity. These 36 genes are marked as blue.

5) Genes which have Metabochip SNPs inside the gene body and in the promoter regions, but don't have either Immunochip SNPs or common cis-eSNPs in the vicinity. These 49 genes are marked as orange.

6) Genes which have Immunochip SNPs inside the gene body and in the promoter regions, but don't have either Metabochip SNPs or common cis-eSNPs in the vicinity. These 39 genes are marked as dark blue.

7) Genes which have common cis-eSNPs inside the gene body and in the promoter regions, but don't have either Metabochip SNPs or Immunochip SNPs in the vicinity. These 57 genes are marked as yellow.

8) 96 randomly chosen genes from which without harboring common cis-SNPs or having Metabochip or Immunochip SNPs in the vicinity. They are marked as black.

In summary, 207 genes have common cis-eSNPs in the gene body or in the promoter regions. 244 genes have Metabochip SNPs in the vicinity while 220 genes have

Immunochip SNPs nearby.

Targeted Sequencing

Whole genomic DNA was isolated from buffy coats of 410 CHDWB samples using Flexigene DNA kits (QIAGEN, Valencia, CA). The major transcription start site (TSS) of each gene was extracted from the UCSC Genome Browser and oligonucleotide probes were designed using the Illumina Design Studio so as to pull down 1kb upstream and 1kb downstream of the major TSS for each of the 472 genes. 5 oligonuceotide probes were designed per each gene to ensure that the percentage of the total length of all regions targeted for enrichment was not less than 90%. Sequence capture libraries were generated and pooled using Illumina TruSeq DNA Sample Preparation Kits and TruSeq Custom Enrichment Kits. The sample preparation steps included DNA shearing with Covaris to an average size of 300bp, conversion of the overhangs after fragmentation to produce blunts ends, adenylation of 3' ends to prevent fragments ligating to each other, ligation of indexing adapters to the ends of DNA fragments, and amplification of the adapter-ligated DNA fragments by PCR. The quality of each library was assessed on an Agilent Bioanalyzer 2100, and the DNA concentration was quantified with a Qubit. Samples were pooled in groups of 12 samples, and TruSeq Enrichment was performed including two rounds of hybridization with capture probes of the targeted regions on streptavidin beads. Subsequently, the pooled DNA libraries were amplified with PCR, and 24 samples (two of the 12-plex pull-downs) were pooled together and quantified by real time-PCR. Paired end 100bp sequencing was performed on an Illumina HiSeq 2500 at Georgia Tech.

FastQC Information

Sequence quality is illustrated taking all of the reads for one sample as an example.

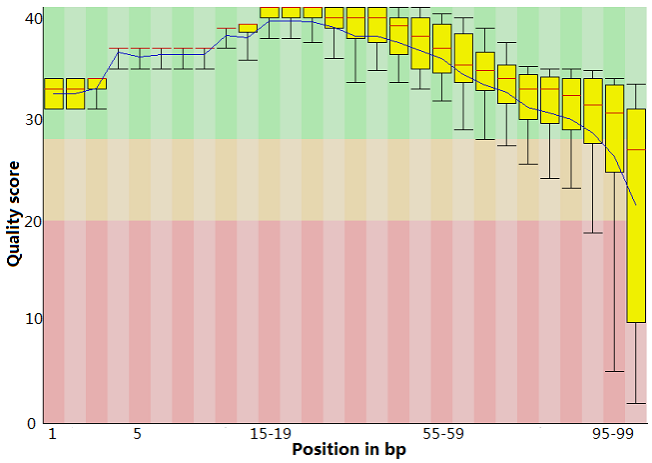

Figure 4.2 Quality score across all bases. The background divides the quality into three bins by colors. Green: very good quality calls. Orange: reasonable quality calls. Red: poor quality calls. The central red line represents the median value. The yellow box shows the inter-quartile range with the upper and lower whiskers representing respectively 10% and 90% range. Blue line: mean quality.

Figure 4.2 shows that for most of the reads except for the last few bases, the quality scores are very high, representing good calling quality. It is common to see base calls falling into the reasonable quality orange area, or even the poor quality red area, towards the end of a read, as the quality of calls on most platforms degrade as the run progresses.

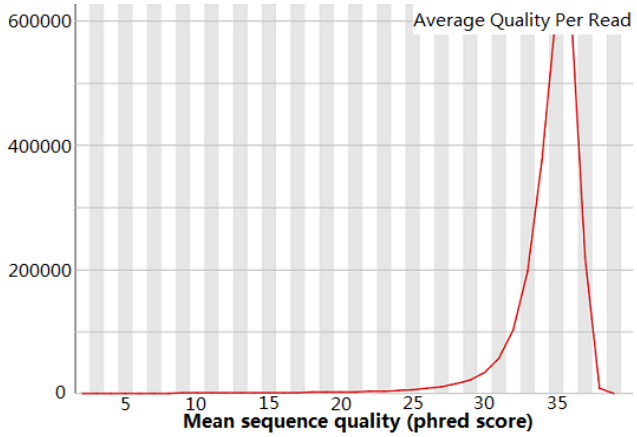

Figure 4.3 Qualtity score distribution over all sequences. The average quality scores per read uses Phred score.

From Figure 4.3, it can be seen that over 95% of the reads have high quality scores which are larger than 28. The Phred quality score = $-10\log_{10}P$ where P represents the error rate, shows that most of the reads have an error rate less than 0.16%. Furthermore, in Figure 4.4, the lines run parallel with and very close to each other, indicating that there is little difference between the proportions of A, G, T and C bases. The lines are flat across different positions in read, also showing the base proportion does not differ between different positions, as expected of high quality and unbiased data.

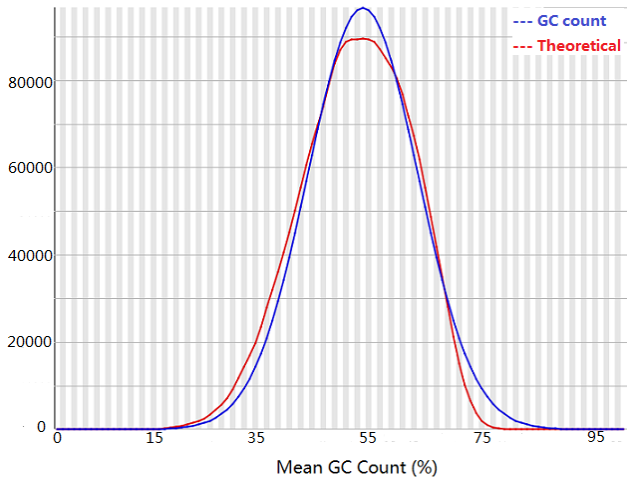

Figure 4.4 GC distribution over all sequences. Red: GC count per sequence. Blue: theoretical distribution

Finally, Figure 4.4 shows that the GC counts per reads do not deviate from the normal distribution, indicating that the library is normal random. The blue line is the theoretical GC content distribution. The red line represents the real GC content across the whole length of each sequence.

Variant Calling Method Comparison

Variant calling method comparison was implemented in the first batch of sequences, which had 12 samples.

Following short read alignment with BWA, three different algorithms for variant calling were evaluated. First, the BWA aligner was used to align fastq files to hg19. Next, Samtools was used to pileup .bam files, with a minimum mapping quality of 10, coverage of 20, and per base quality of 15, and then VarScan was used to call all variants with a minimum alternate allele proportion of 0.25. Alternatively, GATK version 2.7 was used to call variants relative to hg19, marking duplicates with Picard, realigning around known indels and recalibrating around known SNPs according to recommendations for users. UnifiedGenotyper and HaplotypeCaller were run separately, both with emission confidence of 10 and calling confidence of 50. I then applied a hard filter to call variants with QD>=2, FS<=60, MQ>=40, HaplotypeScore<=13, MQRankSum>=-12.5, ReadPosRankSum>=-8 based on GATK best practice.

Given differences in polymorphism call rates as much as 30% between the algorithms, with a notable deficiency of variants observed using VarScan (Figure 4.7), I decided to adopt the more theoretically validated Bayesian strategy implemented in GATK for rare variant inclusion. The rare variants were identified in this study with minor allele frequency (MAF) less than 0.05 in those 410 samples.

Rare Variant Comparison Among Sample Subgroups And Gene Subsets

The number of rare variants were compared between different sample subgroups based on ancestries and genders, and also between different gene subsets based on whether the genes harbor common cis-eQTLs, whether they were potentially related with diseases, and also between different positions with respect to TSS and RegulomeDB classes. Metabochip and Immunochip were used to classify the genes with respect to disease relatedness. There were 196,293 SNPs in the Metabochip array and 196,450 SNPs in the Immunochip array, of which 11,621 SNPs occurring on both the Metabochip array and the Immunochip array. There were 2,936 SNPs in the 244 of our 472 genes that are on the Metabochip array that are in the gene body and 1kb upstream of the TSS.

Similarly, there were 3,778 SNPs on the Immunochip array that were in the gene body and 1kb upstream of TSS of the 220 of 472 genes. Of these, 153 SNPs occurred in the gene body and upstream region of 472 genes for both arrays. There were 146 genes that have both Metabochip SNPs and Immunochip SNPs around the genes, and 319 genes that are represented on at least one of the Metabochip or Immunochip SNPs.

Nucleotide Diversity Calculation and Comparison

Nucleotide diversity (π) in promoter region was calculated based on common variants and rare variants called by targeted resequencing with CHDWB Caucasian data. Nucleotide diversity in coding region was calculated based on 1000 Genomes release version 20130502 European data. The coding regions of 472 genes were obtained from UCSC Genome Browser. Nucleotide diversity were calculated using VCFtools [109] with each variant position. And then summed π's were divided by the region size to generate averaged per site π. As coding regions sometimes overlap with the 2kb promoter regions and hence these diversity estimates are not independent, I also calculated averaged nucleotide diversity for the regions 1kb upstream of each TSS based on Caucasians in the CHDWB cohort.

Results

Overall Summary of Sequence Distribution

The overall sequencing information was obtained after aligning sequencing fastQ files with BWA (with BWA mem), and then running SAMtools flagstat on the BAM files after fastQ. For the 410 samples, the total reads ranged from approximately 2 million to 40 million, with a mean of 14 million and a standard deviation of 5 million. The mapped reads proportion was within 94.08% and 99.92% with a mean of 99.33% and a standard deviation of 1.05. The properly paired reads ranged from 92.95% to 98.95% with a mean of 97.92% and a standard deviation of 1.18 (Figure 4.5**)**.

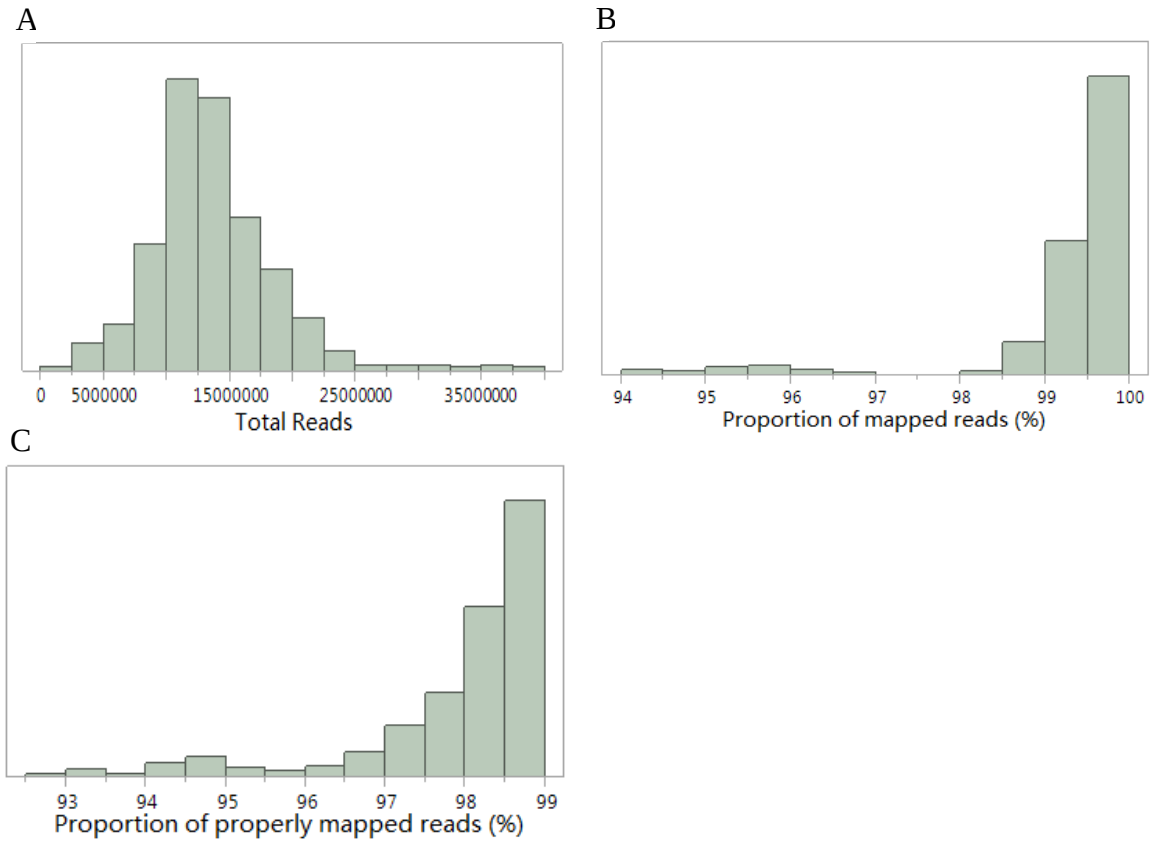

Figure 4.5 The distribution of A. total reads B. mapped reads proportion (%) C. properly paired reads proportion (%) for 410 samples.

Various data quality checks were performed. Approximately 75% of the aligned reads were mapped to the 2kb promoter regions of the 472 genes, indicating good enrichment to the targeted regions. The average read depth across the dataset was more than 600X, with over 90% of the genes having more than 80% reads of the 2kb promoter regions having more than 20X read depth (Table A.1). Due to the nature of the pull-downs, in many cases an extra 500 bp was recovered, but for consistency I restricted our analysis to 1kb upstream and 1kb downstream of each major TSS.

Figure 4.6 shows the read depth distribution across the 2kb promoter regions in TNFRSF4. The 5 peaks clearly show 5 probes covering the gene. Over 90% of the positions have more than 20 read depth. Inverted triangles below the x-axis also show the

positions where common (blue) and rare (green) variants were discovered.

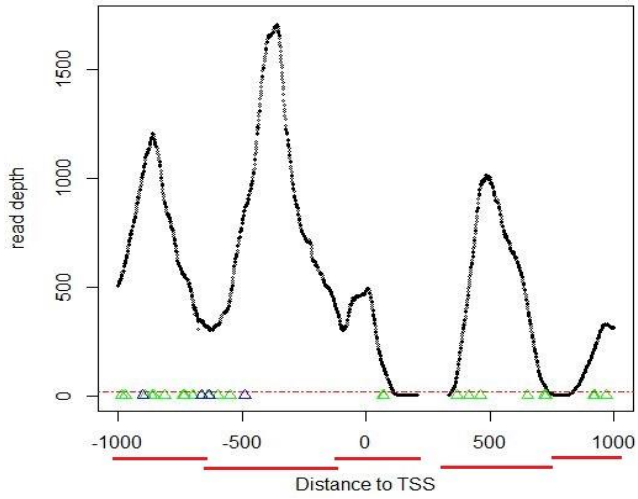

Figure 4.6 The read depth distribution across the promoter regions of TNFRSF4. Green triangle: rare SNPs. Blue triangle: common SNPs. Red horizontal dashed line: 20x read depth. Red lines: The possible probe regions.

Comparison of Different Variant Calling Algorithms

For comparison of the different calling algorithms, consider one sample for whom the number of SNPs within promoter regions called by different variant calling methods was 679 SNPs called by the Samtools/Varscan method and 1007 SNPs and 903 SNPs respectively called by HaplotypeCaller and UnifiedGenotyper. The Venn diagrams below show the overlap among methods for this one individual. HaplotypeCaller is reported to have a higher sensitivity and lower false positive rate than UnifiedGenotyper, and clearly shows a stronger overlap with Varscan.

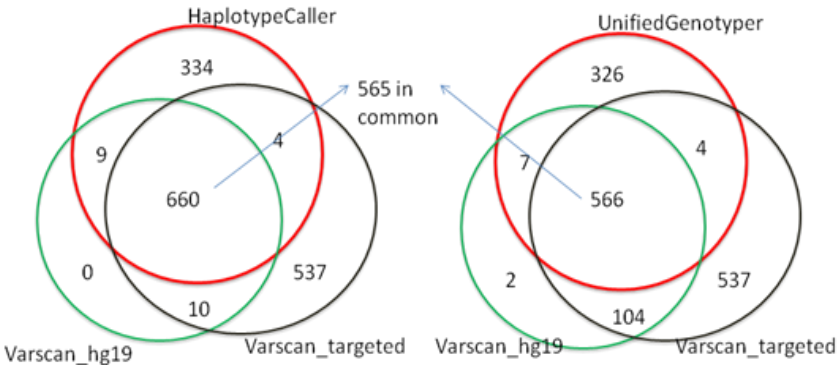

Figure 4.7 Venn diagram of the number of SNPs called by different variant calling methods.

To quantify the effect of the nature of the reference template on the variant calling, I compare the SNPs called by Varscan with hg19 or with just the targeted 2kb promoter regions from hg19 as the references, for 12 samples. From Table 4.3, approximately twice as many SNPs were called with the targeted regions as the reference, compared to ones called with the whole hg19 genome as the reference. The excess SNPs had relatively high read depth, and only a few SNPs that were called with hg19 whole genome as the reference were not called with targeted regions as the reference, in which case those SNPs showed relatively low read depth. A probable explanation of this phenomenon is that some reads mapped better elsewhere in the genome, but were forced to map to similar sequences in the targeted regions. Consequently, when only using the targeted regions as the reference, those reads were incorrectly mapped to the targeted regions, resulting in false positive calls.

Table 4.3 Comparison of Varscan called SNP counts when aligning with hg19 and targeted regions respectively as reference for 18 samples.

	SNP_hg19*		SNP_Targeted*		SNPs in common*		Targeted-hg19 SNP*		hg19-Targeted SNP*	
sample_ID	# of SNPs	avg rd*	# of SNPs	avg rd	# of SNPs	avg rd	# of SNPs	avg rd	# of SNPs	avg rd
Sample 1	681	570	1155	605	671	584	485	635	10	252
Sample 2	751	616	1303	612	743	630	561	586	8	285
Sample 3	702	623	1234	635	692	636	544	631	10	329
Sample 4	704	740	1442	643	695	751	747	544	9	340
Sample 5	551	690	1169	650	545	705	624	602	6	196
Sample 6	683	531	1399	497	676	535	723	456	7	201
Sample 7	741	764	1569	660	733	766	836	556	8	573
Sample 8	677	497	1186	515	666	501	521	523	11	240
Sample 9	751	662	1273	668	732	665	541	653	19	562
Sample 10	823	627	1318	659	810	633	508	695	13	234
Sample 11	730	739	1295	746	719	744	575	728	11	366
Sample 12	683	598	1200	602	671	604	530	592	12	265
Sample 13	709	819	1702	960	701	825	1003	1045	8	304
Sample 14	685	613	1606	817	665	623	942	932	20	286
Sample 15	750	658	1647	862	731	669	918	1000	19	199
Sample 16	722	575	1582	757	709	581	875	878	13	213
Sample 17	755	748	1676	963	737	760	941	1106	18	252
Sample 18	753	795	1733	991	743	803	991	1116	10	247

* “SNP_hg19” means SNPs called with hg19 whole genome as the reference genome.
* “SNP_Targeted” mean SNPs called with targeted regions as reference.
* “SNPs in common” means the SNPs which are called with hg19 as reference are also called with targeted regions as reference.
* “Targeted-hg19 SNP” means the SNPs are called with the reference of targeted regions, but not called with hg19 as the reference.
* “hg19-Targeted SNP” is the other way around.
* “avg rd” means average read depth.

Since the UnifiedGenotyper algorithm is more suited to joint mapping of a large dataset with GATK version 2.7, the release I used in 2013, namely where all reads are aligned together as opposed to combining the results of individual mapping, I decided to use it for the final rare variant calling. Pooling is expected to control the false positive rate, and it was infeasible to run HaplotypeCaller on the 472 samples on our cluster. The UnifiedGenotyper final processing on 24 nodes used 17Gb of memory and ran for over 37 hours. After variant calling, the GATK VQSR tool was used for further variant filtering using the Illumina Omni chip array based on the 1000 Genomes Project as the training data with the highest confidence SNPs from the 1000 Genomes Project's callset used to validate the SNPs. The filtering criteria included QD (quality by depth ratio), DP (depth), FS (Fisher's exact test of strand bias) , ReadPosRankSum (Mann-Whitney Rank Sum Test for the distance of reads with the alternate allele to the end of the read), MQRankSum (Mann-Whitney Rank Sum Test for mapping qualities) and InbreedingCoeff (likelihood-based test for the inbreeding among samples). The Venn diagram in Figure 4.8 shows the number of SNPs in one sample called by UnifiedGenotyper batch calling followed by VQSR filtering, in comparison with HyplotypeCaller, and with UnifiedGenotyper on the sample alone followed by hard filtering. Both UnifiedGenotyper and HyplotypeCaller have good overlap with UnifiedGenotyper batch calling (with 93.4%, 97.6% overlap respectively). However, UnifiedGenotyper batch calling has 95.2% overlap with HyplotypeCaller while 54.6% overlap with UnifiedGenotyper alone. The smaller proportion of overlap for UnifiedGenotyper also occurs when compared with Varscan (hg19 as reference) as shown in Figure 4.7. The raw SNPs were then filtered using ts_filter_level of 99.0, corresponding to sensitivity that would allow retrieve 99% of true variants from the known truth training sets of HapMap and Omni SNPs.

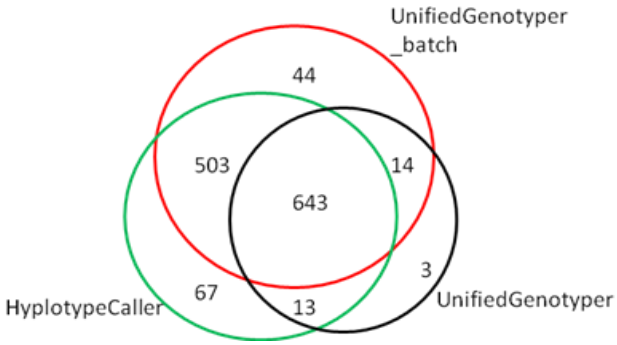

Figure 4.8 Venn diagram of the number of SNPs called by HyplotypeCaller, UnifiedGenotypeCaller, and UnifiedGenotyper batch calling.

After variant calling, I detected 17,584 raw SNPs in total, but these were reduced to 10,451 total SNPs passing the filters that lie within the 2kb promoter regions of 472 genes. 8,833 of the SNPs are rare (defined as MAF < 0.05) and 1,618 are common (with MAF >= 0.05), which averages 1.5 rare variants per promoter per individual. An average of 60% of these rare variants are private, meaning that they were only observed in a single individual. Table A.2 lists the number of rare, common, and private SNPs per gene, along with an estimate of the polymorphism rate (π) per gene.

Validation with Sanger Sequencing

To verify the accuracy of the high throughput sequencing, I Sanger sequenced 500bp segments of two genes, *TRAF3IP3* and *HSPA8*. The sequenced region of TRAF3IP3 was chr1: 209929132 – 209929708, in which 2 rare SNPs and 4 common SNPs were observed in the 410 samples. All of these were included in 96 samples that were Sanger sequenced, and all were validated. The sequenced region of *HSPA8* was chr11:122932665 – 122933158, which contained 18 rare and 5 common SNPs in the 96 sequenced samples, which were again verified by Sanger sequencing. A handful of other individuals were nominally positive at some of the rare sites, but manual inspection of the traces revealed poor quality sequence toward the ends of the reads in those individuals

suggesting a false positive rate that in any case would be less than 0.5%. Five variants that were not present in the GATK analysis but were called by Sanger sequencing were all with low confidence, whereas all common variants were also validated by the Sanger sequencing.

I also have whole genome genotypes either from Illumina Omni or CoreExome arrays, imputed onto 1000G with Impute2, for the majority of individuals. 99.8% concordance was observed. These genotypes were thus used for common variant eQTL analysis, which will be described in detail elsewhere.

Rare Variant Distribution Comparison

1) different ethnicities/genders

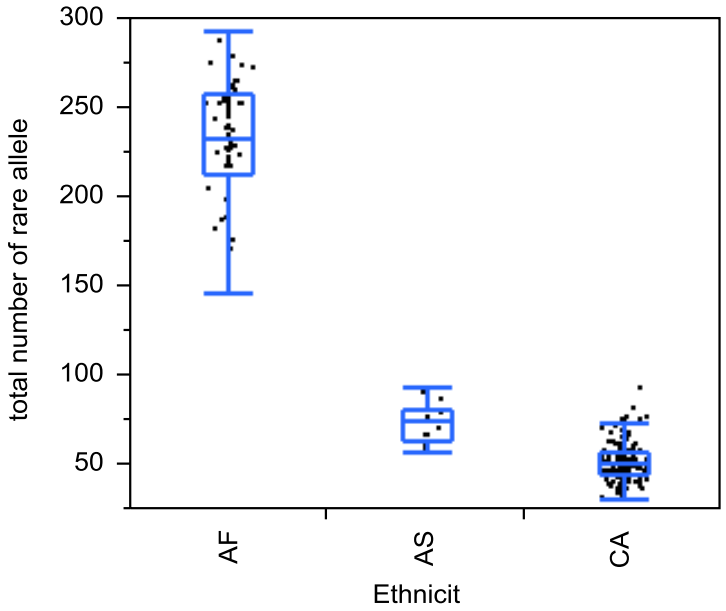

Figure 4.9 Number of rare alleles in promoter regions for 410 samples from different ethnicities. AFR: African Americans, ASN: Asian Americans. CAU: Caucasian Americans.

As expected, the number of rare variants per individual varied by ethnicity.

Shown in Figure 4.9, on average, there were 233, 74, and 52 variants with MAF<5% in each of the African, Asian and Caucasian ancestry samples respectively, representing a 3-fold excess in Africans relative to Asians, and 4-fold excess relative to Caucasians. The difference is highly significant (ANOVA, $p = 3x10^{-263}$). An excess of rare variants in African ancestry samples compared to Asian and Caucasian ancestry samples was also observed in the independently analyzed whole genome sequence replication sample, where the number of rare variants per promoter of the 472 genes in African ancestry individuals was not statistically different from that observed in the CHDWB sample Africans, although almost twice as many variants were called per individual in the Caucasians. [1] and [110] also report a 3-fold excess of rare variants genome-wide in the 1000G genome sequence data, averaged over regulatory, coding and intergenic regions.

There was no significant difference of rare allele count between females and males, in all three ethnicities (data not shown; t-test p-values 0.54, 0.34, and 0.68 in Africans, Asians and Caucasians separately).

2) eQTL genes vs non-eQTL genes

Common eQTL analysis were performed with PLINK using imputed microarray genotyping data and microarray expression data. The detailed information will be described in the next chapter. 207 genes out of 410 total genes have significant common eQTL. There was no significant difference of rare SNP numbers between genes with common eQTL and genes without common eQTL (Figure 4.10, P-value of t-test was 0.13).

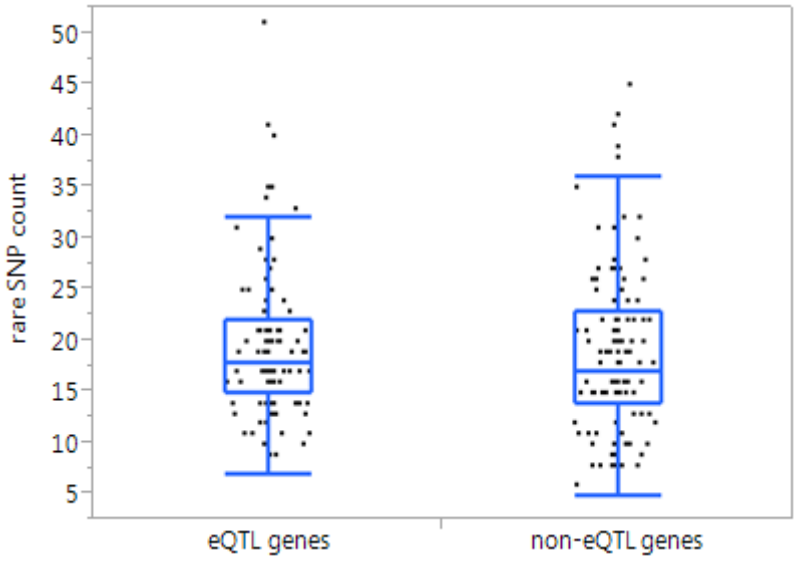

Figure 4.10 The rare allele count in genes with common eQTL SNPs versus genes without common eQTL SNPs.

3) disease related genes vs non-disease related genes

From Figure 4.11, the mean of rare allele count in genes with Metabochip SNPs inside the gene body or within the promoter regions was 19.11, while the mean of rare allele count in genes without Metabochip SNPs was 18.29. The difference was not significant (t-test P-value 0.10). Similarly, the mean rare allele count in genes with Immunochip SNPs inside the gene body or within the promoter regions was 19.03, while the mean of rare allele count in genes without Immunochip SNPs was 18.44 (t-test P-value 0.17). The mean of rare allele count in genes with Metabochip SNPs or Immunochip SNPs inside the gene body or within the promoter regions was 19.12, while the mean of rare allele count in genes without Metabochip or Immunochip SNPs was 17.86, which is a marginally significant difference (t-test P-value=0.023).

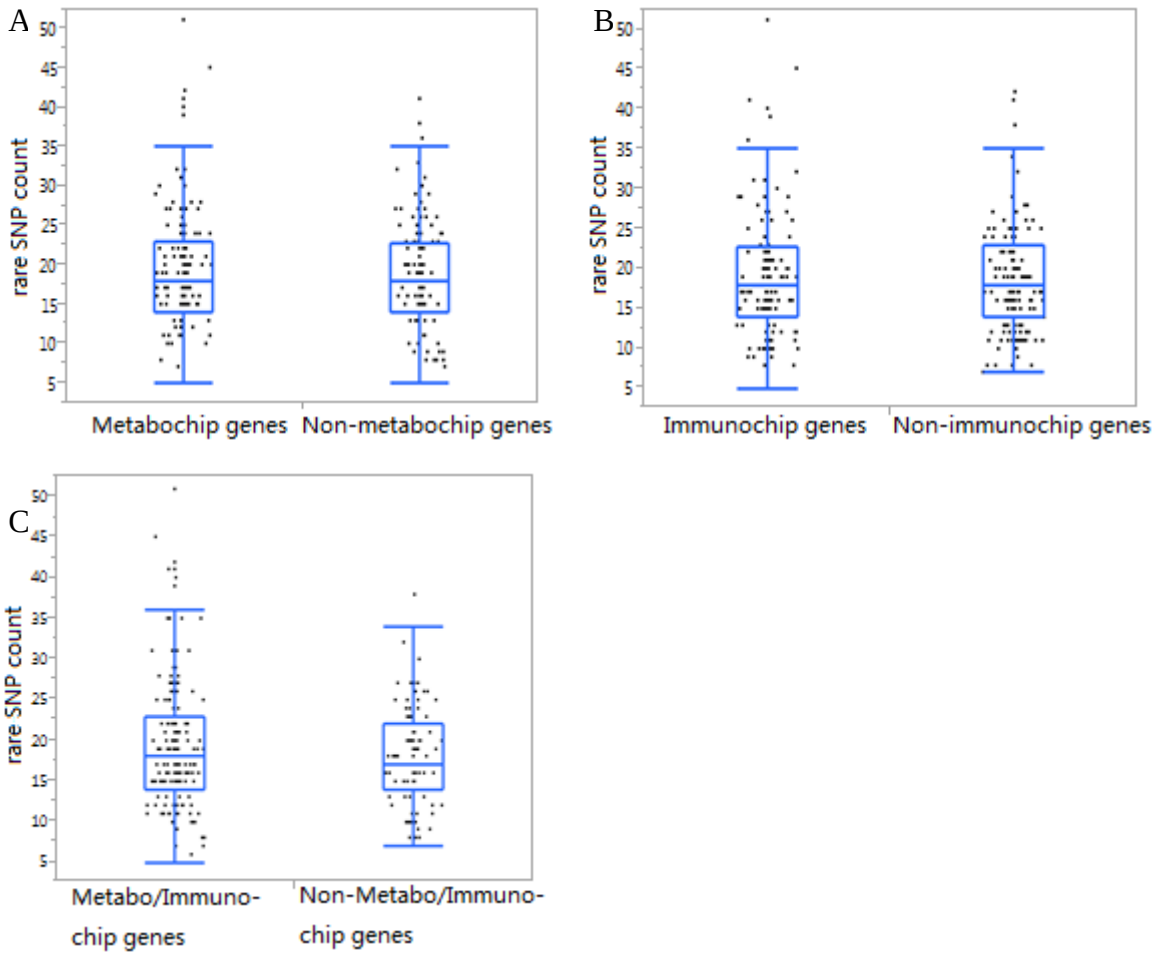

Figure 4.11 The rare allele count in gene subsets with respect to Metabochip and Immunochip. A. genes with Metabochip SNPs versus genes without Metabochip SNPs. B: with Immunochip SNPs versus genes without Immunochip SNPs. C. in genes with Metabochip SNPs or Immunochip SNPs versus genes without Metabochip SNPs or Immunochip SNPs, in the promoter regions or gene body.

4) upstream vs downstream

Within 1kb upstream region of TSS of 472 genes, there were 4,543 rare SNPs, compared with 4,290 rare SNPs within 1kb TSS downstream region of 472 genes (Figure 4.12). These distributions of rare allele count were not significantly different between upstream and downstream, with 95% CI of [2.8, 20] and [3, 19] in upstream and downstream separately.

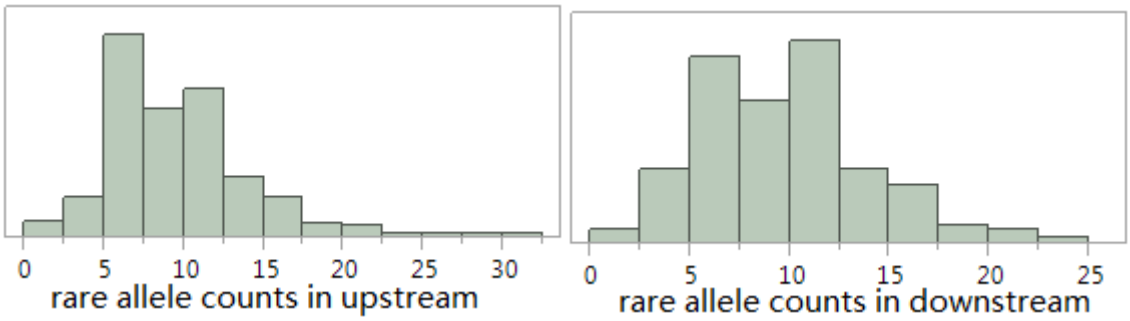

Figure 4.12 The distribution of rare allele counts in upstream (left) and downstream (right) of TSS for all 472 genes in all 410 samples. There is no difference between those two sets (paired t-test p-value = 0.38)

5) RegulomeDB classifications

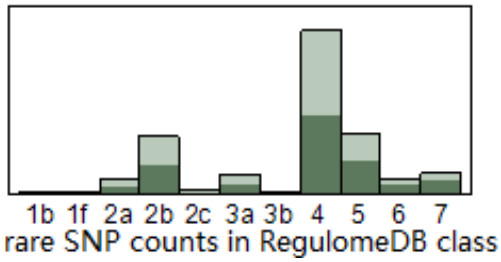

Figure 4.13 The distribution of the number of rare SNPs in each RegulomeDB class. Light green is the downstream region, and dark green is the upstream region.

Figure 4.13 shows that the RegulomeDB scores don't differ significantly between upstream region and downstream region. There are 5 genes that harbor the rare variants in RegulomeDB 1b and 1f classes. 432 genes harbor the rare variants in RegulomeDB 2a-2c classes. 264 genes have rare variants in RegulomeDB 3a and 3b classes. 461 genes have rare variants in RegulomeDB 4 class. In addition, 464 genes have rare variant which are in RegulomeDB classes 1 through 4, which are the classes with strong evidence that those regions encompass functional regulatory elements. So in total, more than 98% of those 472 genes have promoter proximal rare variants in regulatory sites. There is thus a lot of opportunity for rare variants to have regulatory functions.

Nucleotide Diversity Comparison

1) Between promoter region and coding region

The average nucleotide diversity 1kb upstream of TSS is highly correlated with that observed in the full 2kb promoter region ($p=3.5x10^{-119}$, $r^2=0.68$, Figure 4.14A), indicating that just the upstream region is a good proxy for the full promoter. Next I observed that there is also a significant association between nucleotide diversity in upstream promoter regions and in coding regions (slope=0.6, $p=1.3x10^{-22}$, $r^2=0.19$, Figure 4.14B). This indicates that upstream promoter regions have relatively lower sequence diversity than coding regions. However the high correlation is likely due to a combination of linkage disequilibrium affecting background purifying selection on function, and similar constraints on coding and regulatory regions.

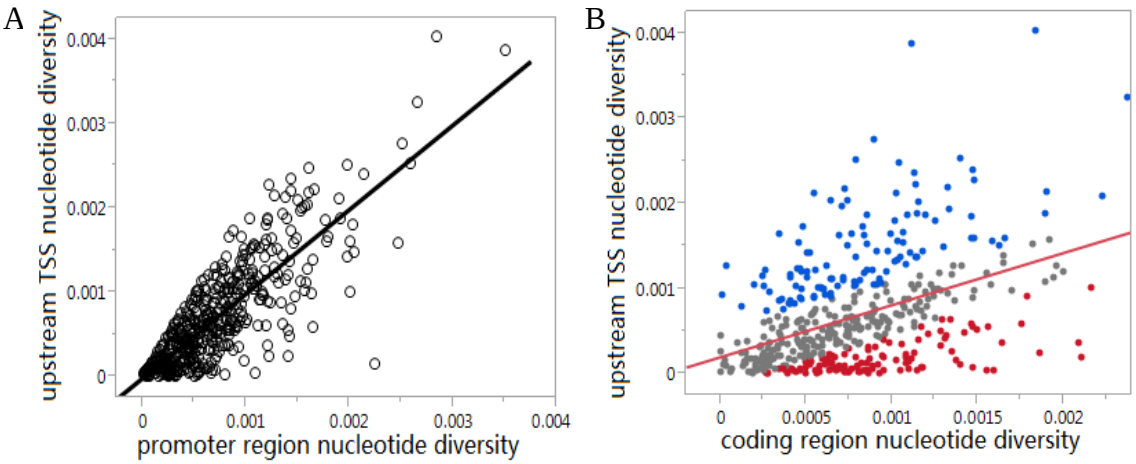

Figure 4.14 Linear regression fitting nucleotide diversity in upstream region on the full promoter (A), and on the coding region (B). Red dots in B represent the bottom 25% genes, with negative relative promoter polymorphism compared to coding regions. Blue dots represent the top 25% genes with positive relative promoter polymorphism.

Despite this high correlation, there are nevertheless genes that have more promoter polymorphism than expected, and genes with less, relative to the coding region. In order to better interpret the deviation of promoter polymorphism from the expected value, I

calculated the studentized residuals after regression of per-base nucleotide diversity (π) in upstream promoter regions on π in coding regions, generating a parameter I call "relative promoter polymorphism". A residual of 0 means the gene has average level of promoter polymorphism. A positive value means the gene has more promoter polymorphism and negative value means less promoter polymorphism, with respect to the coding region of the same gene.

Next, I compared the relative promoter polymorphism with the "residual variation intolerance scores" (RVIS) as reported by [92]. RVIS is a measure of intolerance to functional mutations, and is computed as the ratio of common missense or truncation variants (MAF>0.01) to all variants in a gene region. In coding regions, RVIS is highly significantly associated with functional measures such as whether a gene is known to harbor Mendelian mutations, and has been proposed as a means of scaling the likelihood that a rare or de novo coding variant causes disease. There is however no significant linear relationship between my relative promoter polymorphism measure and coding RVIS. This implies that tolerance of functional variation in coding regions is not strongly correlated with levels of promoter polymorphism.

2) Between gene subsets with disease related status

The genes with Immunochip SNPs show significantly higher nucleotide diversity in upstream regions, compared with the genes not harboring Immunochip SNPs (ANOVA p=0.036, Figure 4.15A). The genes harboring Metabochip SNPs also show higher nucleotide diversity in upstream regions in comparison with the genes without Metabochip SNPs (ANOVA p=0.009, Figure 4.15B).

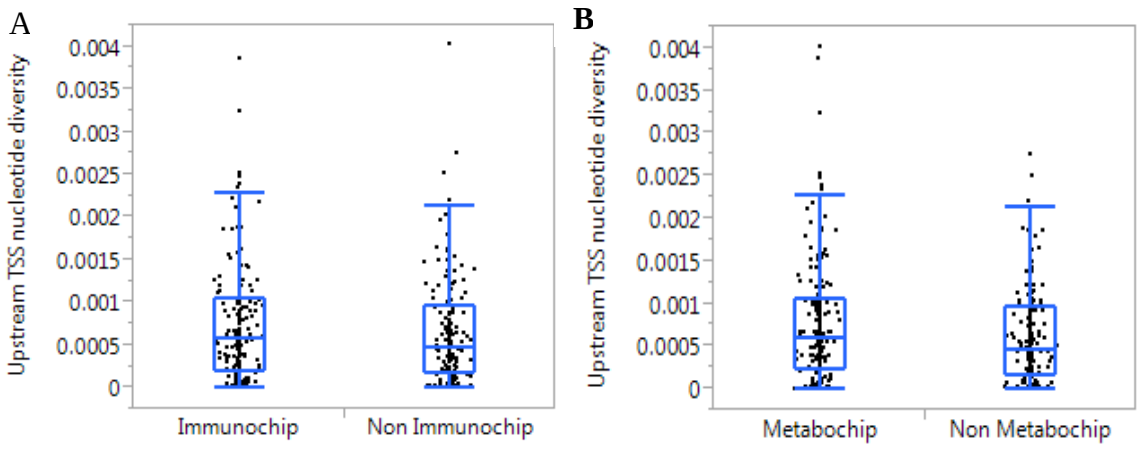

Figure 4.15 Comparison of nucleotide diversity in upstream region for genes with or without Immunochip SNPs, and for genes with or without Metabochip SNPs.

Discussion

From the comparison of SNP calling by GATK and VarScan, the calling algorithm introduces considerably more bias in SNP detection than does read depth. Other researchers have reported similar results: O'Rawe et al [111] reported that different variant calling pipelines (including SOAP, BWA-GATK, BWA-SNVer, GNUMAP [112], and BWA-SAMtools) only had a 57.4% concordance rate. The reference template also causes a big difference in targeted sequencing alignment. As shown in Varscan, using the targeted region as template tends to call 4 folds more variants than using hg19 reference. When choosing the algorithms for calling the variants, decisions should be guided by the nature of the sequencing dataset and the specific questions being asked, and appropriate parameters need to be selected for each algorithm. For example, the whole genome reference such as hg19 is usually better than targeted sequence as reference, as using whole genome reference will get the best mapping position for a read to decrease the likelihood that reads are mapped to wrong places.

The sequence diversity of rare variants in promoter regions is very high in African ancestry samples compared to rare variants in Caucasian and Asian ancestry samples. It shows that African ancestry has more diversity compared to Caucasian and Asian

ancestries. The population bottleneck caused by the out-of-Africa migration may serve as an explanation of the lower diversity in Caucasian and Asian ancestry samples. The hypothesis has been raised [113, 114] that due to the out-of-Africa migration reducing the effective population size for modern humans, non-African populations tend to experience stronger genetic drift in addition to the founder effect leading to loss of diversity. In addition, adaptation to new environments possibly leads to some selection against common genotypes or selective sweeps. Further, in comparison with the finding of three times the number of rare variants in Africans than in Caucasians from 1000 Genomes, my result showed more enrichment of rare variants in Africans. That is possibly because the read depth is much higher in my sequencing data compared to 1000 Genomes project, leading to higher power to detect rare variants. Another explanation is that the filtering method VQSR after GATK variant calling is more stringent than what is used in 1000 Genomes project, leading to calling of fewer rare variants found in Caucasians.

The rare SNP count is not obviously different between different regions and classes of genes. The only comparison that showed a statistically significant difference was that the genes represented on either Metabochip SNPs or Immunochip SNPs have a slight tendency to harbor more rare variants. The significant association between nucleotide diversity of promoter region and coding region implies that the genes with evolutionary constraint on the coding region also show constraint on the promoter region. Petrovski et al [102] reported that the genes that are related to immune disease tend to be tolerant to more common functional mutations. My result supports their finding to the extent that the genes which are potentially related with metabolic disease or immune disease tend to have more highly polymorphic promoters. However, the RVIS score is not likely to be a good measure of the intolerance of promoter regions to functional variation. Development of a promoter specific intolerance score will require incorporation of an estimate of which variants influence gene expression, which is the subject of Chapter 5.

CHAPTER 5

RARE REGULATORY VARIANTS ASSOCIATION WITH TRANSCRIPT ABUNDANCE

Introduction

In recent years, whole exome sequencing has been used effectively to demonstrate that there is a burden of rare coding variants in individuals with a variety of neurological and developmental conditions [115-118]. Considering estimates that as many as 90% of disease associated common variants are regulatory rather than structural [37, 98-100], it is reasonable to assume that rare regulatory variants influencing the expression of causal genes might also be enriched in individuals with congenital abnormalities. Whole genome sequencing may address this issue, but will need to confront multiple comparison issues given the enormous size and complexity of regulatory sequences.

Gene expression, whose pattern and properties play a fundamental role in affecting the functions of genes, cells and also phenotypes [119], has attracted considerable attention. As an important intermediate phenotype between gene variants and phenotypic traits, gene expression is a key to uncovering the underlying genetic mechanisms responsible for variation for complex traits and diseases. While many studies have explored the association of genetic variants with gene expression traits, common eSNPs can only explain a medium proportion of gene expression variation [120], which implies the rare variants may also make a contribution. Whereas the genetic functions of rare variants in coding regions are well studied, association of regulatory rare variants with expression traits has yet to be described. Here I test for a burden of rare variants with gene expression itself, focusing on just the promoter regions of a targeted set of genes whose expression was measured by microarray analysis of peripheral blood samples.

Power issue is one the most important considerations in rare variant association tests. Previous studies have reported that at least 20,000 samples were needed to reach the enough power to perform rare variant association test [121, 122]. In this study, I utilized a novel pooled burden test to overcome the difficulties when using much fewer samples. As the promoter regions of each gene could be viewed as independent and only cis-acting variants were considered in this study, it is reasonable to view the promoter regions of each gene in each individual as independent measures. The overall test approach I used was to combine all the promoter regions across all individuals to enlarge the number of independent tests.

The burden test was designed to test whether rare variants are enriched in the promoters of genes that are at the extreme of transcript abundance. Robust association tests were also performed by evaluating a series of rare variant association tests, conditioning on common-eSNPs, and regulatory features. The results of this study showed an overall signature of association enrichment across subsets of genes, as well as provided an estimate the magnitude of rare variant effects. This study provides evidence for rare variant association with transcript abundance, strongly supporting a possible general and pervasive contribution of cis-regulatory effects of rare variants as a source of genetic variance for rare clinical traits.

Materials and Methods

Samples

The 410 samples were from the Atlanta CHDWB cohort under approval of the Emory University and Georgia Tech IRBs for genetic studies. The 410 samples were comprised of 274 females and 136 males. The samples were a mixture of 297 Caucasians, 95 Africans and 18 Asians. The age at entry into the program and initial sampling of blood spanned from 19 to 83 with a mean of 50 and standard deviation of 10.6.

DNA Re-sequencing and Variant Calling

The targeted capture sequencing of promoter regions was described in detail in the previous chapter and is briefly summarized here. Whole blood was extracted from 410 samples and DNA was extracted by QIAGEN FlexiGene DNA Kit. 472 genes were selected (list and criteria as in Chapter 4). The major transcript start sites (TSS) were obtained from UCSC Genome Browser. A capture array was designed using Illumina design studio to pull down 1kb upstream and 1kb downstream of the major TSS for each of the 472 genes. Sequence capture libraries were generated and pooled using Illumina TruSeq DNA Sample Preparation Kits and TruSeq Custom Enrichment Kits. Sequencing was performed on an Illumina HiSeq 2500 with the assistance of Ms. Shweta Biliya in Dr. Fredrik Vannberg's group at Georgia Tech, providing an average coverage of 600 X.

The sequences were aligned by BWA and variants were called using the GATK UnifiedGenotyper in one batch for all 410 samples. Variants were defined as rare in this study if the minor allele frequency (MAF) was less than 0.05 in the 410 samples. After quality filtering using VQSR, there was a total of 10,451 SNPs, including 1,618 common SNPs and 8,833 rare SNPs.

Gene Expression Profiling

Transcript abundance measures were generated in two batches using Illumina-HT12 human gene expression arrays. RNA was prepared from whole blood samples collected and stored in Tempus tubes (Life Technologies), following manufacturer-recommended protocols, and quality was confirmed using an Agilent Bioanalyzer such that all samples had RIN numbers greater than 8. The first batch of samples was processed for hybridization and bead intensity extraction by ExpressionAnalysis (Durham, NC) and the second by HudsonAlpha (Huntsville, AL). The raw data is available at the Gene Expression Omnibus (GEO) as accession GSE61672, but additional data processing steps were employed for this study to account for batch effects that

skewed the rare variant association statistics.

Gene Expression Normalization

The objective of this study was to evaluate whether or not there is enrichment for rare variants at the extremes of the transcript abundance distribution across a population. The question is not whether genes that have high or low abundance tend to have more rare variants. It is whether individuals with extreme abundance of single genes have an excess of rare variants in those genes. If we could measure the abundance of a single transcript in 100,000 people, we may expect that promoter regions of genes in the lower and/or upper percentiles of the distribution tend to have more rare variants than genes in the middle of the distribution. Since we did not have the resources to perform such a comparison, instead I reasoned that we could pool the results of almost 500 genes, each measured in more than 400 people, and evaluate whether there is signal in the tails of the distributions of all of the genes considered jointly. The signal for any one gene would not likely be significant, but with 472 genes there are over 200,000 gene expression measures, providing power.

For this strategy to be informative, it is essential that there is no relationship between polymorphism levels in an individual, and the tendency of individuals to have more genes with extreme expression. For example, if a subset of genes tend to be more highly expressed in African Americans, then those genes would be toward the upper tail and they would tend to have more polymorphic variants, and hence a regression of rare variant count against percentile of gene expression would have a positive slope. Vice versa for genes that tend to have lower expression in the group with higher polymorphism, while the combination of both biases would tend to generate concave "smile" plots. It is also relevant to note that biases other than ethnicity could result in a significant relationship, including technical batch effects. In the presence of such biases, permutation of the genotype and gene expression matrices across individuals would result

in significant linear or quadratic terms in our test statistic just as observed for the real data. Actually, such effects were observed in our initial evaluation of the raw data, which led us to adopt the normalization strategy outlined here.

Raw expression data in the form of average bead intensities from the Illumina Genome Studio were first transformed to $\log_2$ values, and then processed with the Supervised Normalization of Microarrays (SNM) [123] algorithm in R. I fit Age as the biological variable, and removed effects of Batch and Ethnicity by including these as adjustment variables with the rm=True option. Individual effects were accounted for as the intensity-dependent variable, resulting in overall gene expression profiles with the following distribution, colored by Batch.

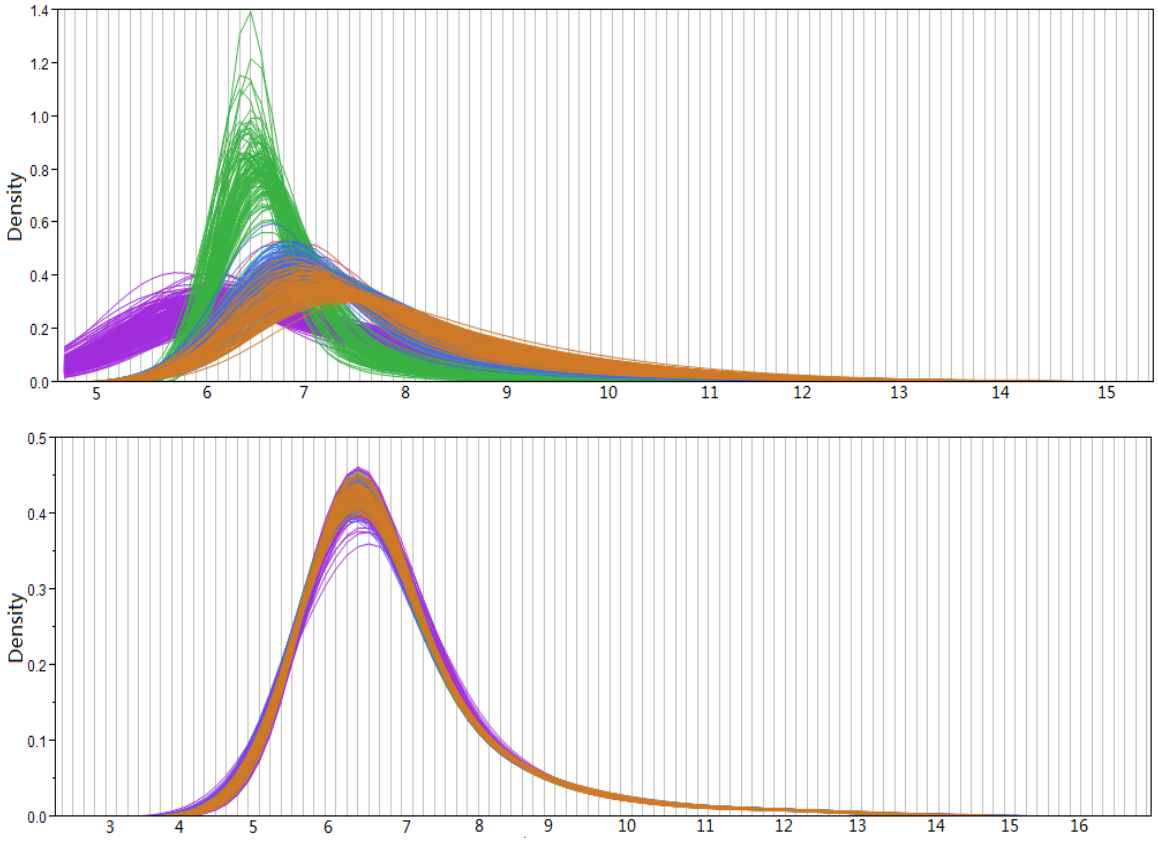

Figure 5.1 Distribution of raw log2 transformed expression (top), and SNM normalized expression (bottom). Four colors represent four batches of samples.

Figure 5.1 shows the distribution of raw $\log_2$ transformed expression and SNM normalized expression. The raw expressions have a strong batch effect which is adjusted well by SNM normalization. After performing variance component analysis (Figure 5.2), batch effect reduced from 74.6% to 0, making 90.8% of the variance explained by residuals after SNM normalization.

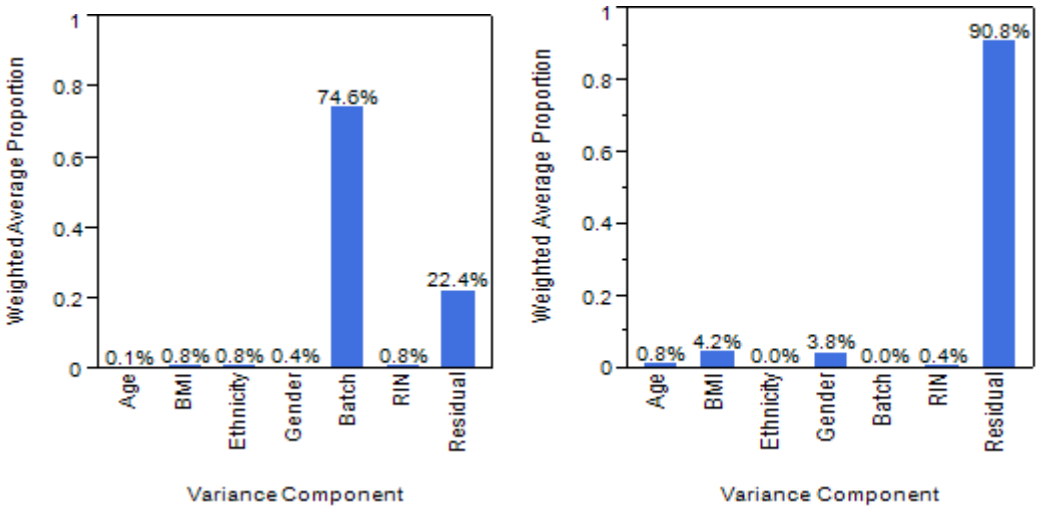

Figure 5.2 Variance component analysis of raw log2 transformed expression (left), and SNM normalized expression (right).

I next extracted the 472 genes for which we have promoter genotypes, and averaged the estimates for 177 genes that are represented by 2 or more probes in the Illumina-HT12 arrays. In order to combine the genes, I next sought to convert each gene expression distribution to the same scale, namely to z-scores, which are standard normal distributions with a mean of 0 and standard deviation of 1. To ensure that there was no overall batch effect on the variances (namely, that individuals from one batch are not, for technical reasons, more likely to have extreme values), I fit the z-scores by batch and combined them.

Rare Variant Burden test

In order to evaluate whether there was a relationship between transcript abundance and number of rare variants in the promoter, each gene was placed in one of 82 equal-sized bins of five individuals based on the rank of the batch-adjusted z-scores. For each gene separately, the lowest 5 individuals are in bin 1, the next lowest 5 are bin 2, and so forth until the top 5 are bin 82. The number of rare variants, defined as a variant with a MAF < 0.05, in the promoter region of each gene in each bin was summed, and subsequently these values were summed across all 472 genes. The plots in Figure 5.3 show the strategy obtaining the sum total of rare variants in 472 promoters in bins of 5 individuals, starting with bin 1 at the left and ending with bin 82 at the right. I then evaluated deviation of the distribution from the null hypothesis of no relationship by fitting a quadratic model where the linear term captures bias toward enrichment for either higher or lower expression, and the quadratic term captures the effect of bias at both extremes simultaneously.

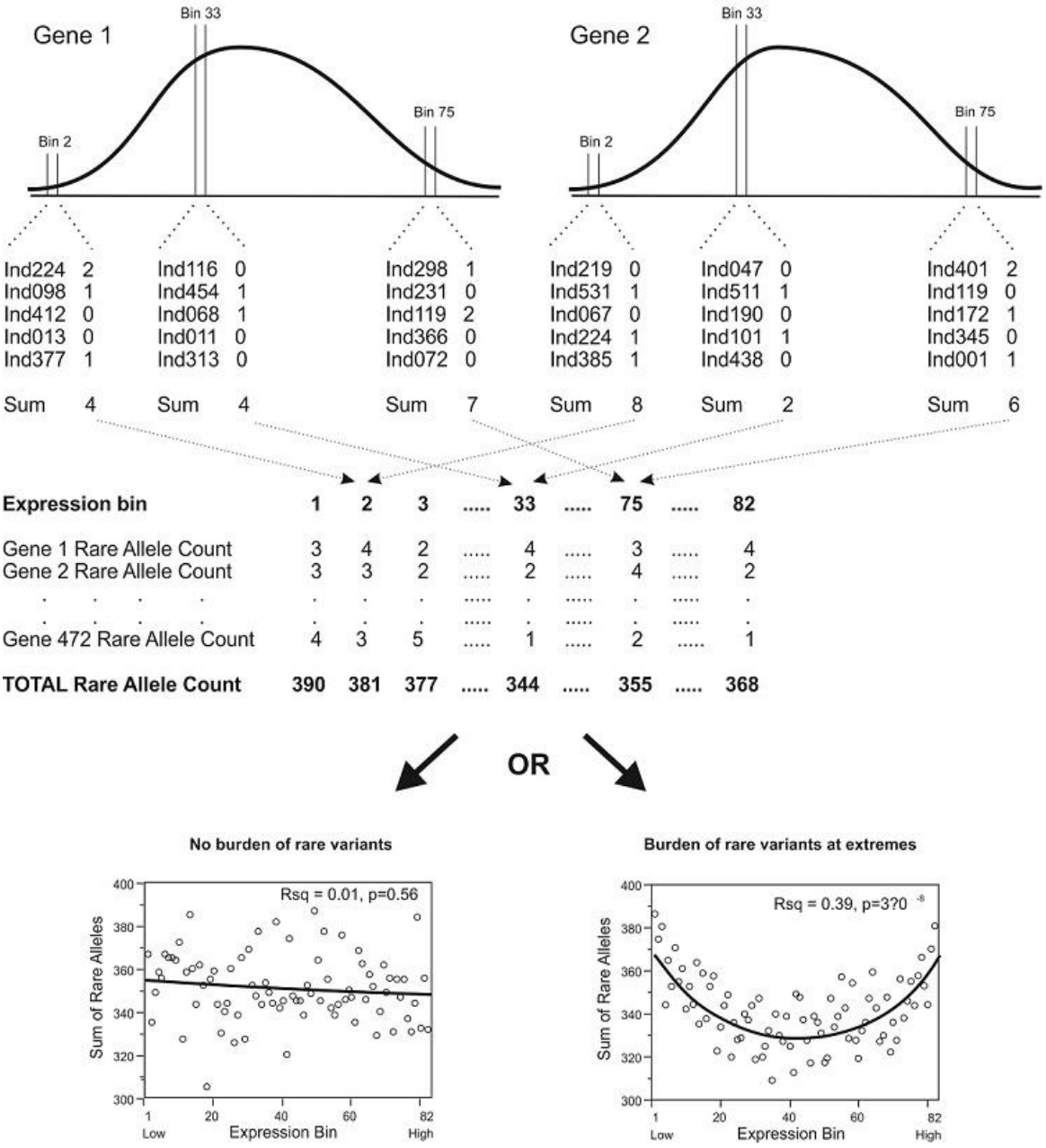

Figure 5.3 Schema showing the pooling strategy to evaluate rare variant enrichment.

The significance values of the two terms were observed to be very similar to the empirical p-values obtained by permuting the sum counts against the bin number. However, a more robust permutation is to shuffle the genotype and gene expression vectors, keeping the full vector of promoter counts within each individual, and the full vector of expression ranks, constant so as to preserve any biological covariance. With the appropriately normalized gene expression data, such permutations generally resulted in flat regressions of allele count on expression bin, with non-significant linear and quadratic terms. I then evaluated the significance of the actual data by documenting how many permutations out of 10,000 have a more significant overall model fit, which turns out to be just a few cases, strengthening support for the inference (i) that the normalization has removed systematic biases, and (ii) that there is a true burden of rare variants at either extreme of the transcript distribution, averaging across 472 transcripts.

A further adjustment was made to account for unequal total read counts among individuals or in specific genes. In particular, the up to 4-fold difference in rare variant counts between African and Caucasian ancestry individuals could bias the analysis if certain genes are more variably expressed between ethnicities; and I noticed that not infrequently, individuals harbored "extreme genotypes" with three or more private or rare polymorphisms in the same gene. For the analyses involving mixed races, I thus performed a haplotype burden analysis by collapsing all multi-SNP promoters down to a count of 1, instead of the actual number of rare variants. This should be conservative since it will tend to underestimate the contributions of two or more variants in a single promoter. Application of this haplotype burden test to just the Caucasian data reduced the significance of the quadratic fit in smile plots, but the general trend remained the same as with the full SNP count data.

A variety of biological factors could mask the effect of rare variants by increasing the variance of gene expression. Two obvious effects are the contribution of common eQTL, which will tend to cause individuals with the less active polymorphism

to be in lower expression bins, and trans-acting sources of gene expression covariance. Since peripheral blood preserved in Tempus tubes is a complex mixture of leukocytes (residual red blood cell and platelet gene expression is not thought to contribute strongly to observed transcript abundance), an obvious source of covariance is cell counts: individuals with low T cell counts for example may tend to have coordinate low abundance of thousands of transcripts that are enriched in T cells. Cell counts explain a little over a third of the transcriptional variance in the dataset. This is actually half the amount explained by seven empirically determined common axes of covariance that likely reflect a mixture of contributions of cell counts and coordinate gene regulation for example by interferon or other systemic factors. These seven axes are defined by the first principal component of the expression of 10 "blood informative transcripts (BIT)" [124] per axis, where the BIT have been defined by comparison of multiple blood gene expression datasets. Fitting PC1 to the 5 or 100 most correlated transcripts in each axis results in almost identical scores.

Simple linear regression was used to fit either eQTL or co-expression Axes or both together. For the eQTL adjustment, I first performed whole genome cis-eQTL analysis on the full CHDWB dataset and identified significant eQTL located within 5kb of the TSS or in the gene body for 207 of the 472 genes at $p<10^{-4}$, observing more than 70% overlap with the Blood eQTL browser variants derived from meta-analysis of over 5,000 samples. For 112 genes, multiple additional cis-eQTL were observed conditioned on the primary eQTL. With the assistance of another graduate student in my lab, Biao Zeng, stepwise linear regression was useed to fit these empirical eQTL in our dataset, also including Age and Sex as covariates in the model (although neither age nor sex account for more than a few percent of the variance of any of the 472 genes). The residuals from the eQTL fit were then ranked and placed in bins for the burden test. For the Axis adjustment, I computed the 7 PC1 scores from the full SNM normalized gene expression matrix, and then identified which axis was most strongly correlated with the

expression of each gene (448 were influenced by an axis at $p<10^{-4}$). Univariate linear regression was then used to fit the relevant axis for each gene, and again the residuals were taken forward to the adjusted burden test. For the joint fitting of eQTL and Axis scores, I performed the regressions in a sequential manner.

Two other versions of the analysis were performed to confirm the robustness of the core result. First, I performed quantile normalization of the $\log_2$ transformed data [125]. This procedure results in identical overall transcript abundance distributions by ranking gene expression within individuals and assigning each rank the mean value of transcript abundance for that rank. It does not however remove systematic sources of variance at the level of individual genes, such as batch or ethnicity effects. Second, rather than ranking each gene separately, I also performed an overall percentile method where all of the z-scores of the 472 genes were combined and assigned to 82 bins of 2,360 transcripts. This analysis allows the same gene to be present in the same bin in multiple individuals, so does not assume that each gene is approximately normally distributed. Clusters of individuals who share a rare variant and extreme expression might consequently be further enriched toward the extreme.

Partitioning the Sources of Rare Variant Burden on Gene Expression

As described in the results section, several potential modifiers of the rare variant contribution were evaluated by dividing the total Caucasian dataset into subsets and comparing the model fit. For example, to evaluate whether suspected regulatory sites are more likely to harbor rare variants, I downloaded the RegulomeDB assignments for each SNP and contrasted sites with scores in the ranges 1-4 (likely regulatory) or 5-7 (weak or no evidence for regulatory potential). Similar analyses reported in Table 5.1 contrast SNPs upstream and downstream of the TSS; SNPs in genes in the upper of lower halves of the overall average transcript abundance spectrum; SNPs in genes in the upper or lower halves of the average promoter polymorphism distribution; SNPs in genes with or

without common eQTL; and SNPs in genes represented on the Metabochip, Immunochip, or neither.

Replication Dataset

In order to replicate the rare variant enrichment on a completely independent dataset generated with different gene expression and genotyping technologies, Professor David Goldstein provided a small cohort of 75 individuals with whole genome sequence and whole blood RNA-Seq at the Duke Center for Human Genome Variation. Most of these individuals are from a Schizophrenia study. Permission to perform genetic analysis was obtained under IRB approval of Duke University, affirmed by the Georgia Tech IRB. Analysis of the principal components of the genotypes indicated that the sample includes approximately 49 Africans and 26 European Ancestry individuals.

Whole genome sequences were obtained on Illumina HiSeq2000 automated DNA sequencers and genotypes were called individually with the GATK algorithm. RNA-Seq of whole blood preserved in Tempus tubes was performed also by paired end 100bp sequencing on the Illumina platform. Raw read counts were log2 transformed, and mean centered, and linear regression fitting each of the 7 axes of variation (represented by PC1 of the blood informative transcripts) as well as the overall PC1 of gene expression variation (which is correlated with genetically determined ancestry). Subsequently, I assigned the rank of each gene in each individual, and performed quadratic regression of the total rare allele counts (MAF<0.05) for each of 75 ranks. That is, rather than pooling 5 individuals per bin, the analysis was essentially on bin sizes of 1, necessitated by the small sample of individuals.

Experimental validation of SNP effects by genome editing

I chose four SNPs for experimental validation by CRISPR/Cas9 mediated genome editing. Two (rs in *TDP1* and rs in *COMMD4*) were associated with loss of gene

expression and two (rs in *DHX29* and rs in *UQCC*) with gain of gene expression in the CHDWB targeted sequencing analysis. For each gene Dr Idowu Akinsanmi in the Gibson lab and Dr Ciaran Lee in the laboratory of Professor Gang Bao generated 11 or 12 independent approximately 10-cell K562 clones targeted by guide RNAs as follows (need paragraph from Idy). Although K562 cells are erythroleukemic, rather than lymphoid or myeloid [126], since the variants are promoter proximal, we reasoned that they may have effects generally on transcript abundance and this cell line is well established for CRISPR experiments.

After confirming disruption of the relevant SNP by the T7E1 assay [127], 8 clones for each of the 4 genes with average heterozygosity between 16% and 23% were selected. Assuming a 2-fold modification of transcript abundance in one fifth of the cells affected by the CRISPR, we require a sensitive gene expression assay capable of detecting a 20% modulation of gene expression. Droplet PCR is a new quantitative PCR approach that does not rely on thresholds of product generation in real time, but rather consists of 10,000 dilute droplets of RNA [128]. Quantitative transcript abundance estimation is generated by scaling the counts of droplets (typically between 500 and 1000 out of 10,000 for the transcripts considered here) to the number of positive droplets for two control genes measured in aliquots of the same RNA sample. HPRT was a common housekeeping gene control, while UQCC was used as the second control for TDP2 and DHX29, and TDP2 as the second control for UQCC and COMMD4 disrupted clones.

Results

The overall burden test on the full cohort (Figure 5.4A) showed that the rare variants were enriched in both increased and decreased gene expression (model R^2 = 0.19, p=0.0003, permutation p=0.0002). In addition, the enrichment of rare variants in decreased gene expression is more obvious than the enrichment in increased gene expression. The full cohort was composed of 297 Caucasians, 18 East Asians, and 95

Africans. Because of the limitation of imputation quality, after fitting common variants the cohort included 384 individuals, composed of 279 Caucasian, 18 East Asian, and 87 African American individuals. As African Americans have differential expression for one third of the genes, and almost three times more rare variant counts than Caucasians in the promoter regions (consistent with HapMap estimates), I also considered each of the two larger population groups separately. In addition, adjusting for covariates which are known to affect gene expression improves the rare variant test significance. The covariates include common eQTLs and principle component scores for 7 common axes of peripheral blood gene expression covariance. These 7 axes collectively explain 26% of the total expression variance of the 472 genes. Fitting those two covariates improve the model R^2 from 0.05 to 0.17, p=0.07 to 0.0003 for Caucasians (Figure 5.4B). That is the further evidence that the regression models truly capture enrichment for rare variants. Africans separate test (Figure 5.4C) also shows the same trend with enrichment of rare variants at the two extremes of gene expression. The less significance in Africans may be due to smaller samples size compared to Caucasians. It was verified by randomly selecting the same number of Caucasians as Africans to perform the test. Besides rare variant count, I also used rare haplotype counts as some individuals carry more than one rare SNP. Both show the same trend of enrichment in high and low gene expression. I also applied different normalization methods for gene expression, including quantile normalization and an approach that firstly pooled z-score for all the gene expression measures together and then assigned bins. Both normalization approaches show the enrichment of rare variants at two extremes of gene expression, meaning the enrichment test is robust with different normalization methods (Figure 5.5). In addition, grouping gene expression into 41 bins with 10 individuals in each bin (Figure 5.6A) or setting the rare variants as MAF less than 0.01 (Figure 5.6B) also shows significant enrichment of rare variants at the two extremes of gene expression.

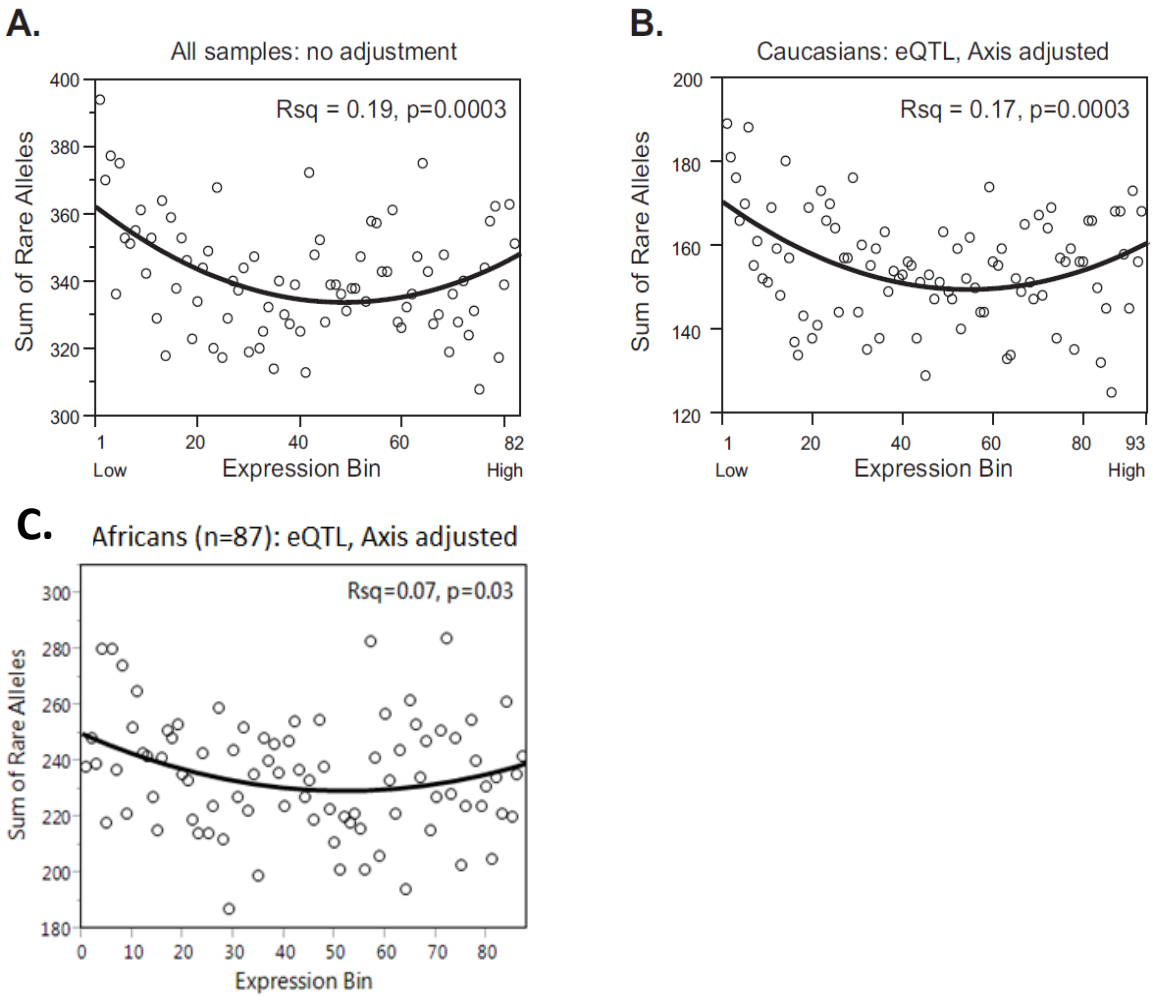

Figure 5.4 The rare variant burden in sorted expression (SNM normalized) bins. A. 410 samples, no adjustment B. 279 Caucasian Americans, after fitting SNM expression levels by significant common eQTLs and 7 blood transcript axes. C. 87 African Americans, after fitting SNM expression levels by significant common eQTLs and 7 blood transcript axes.

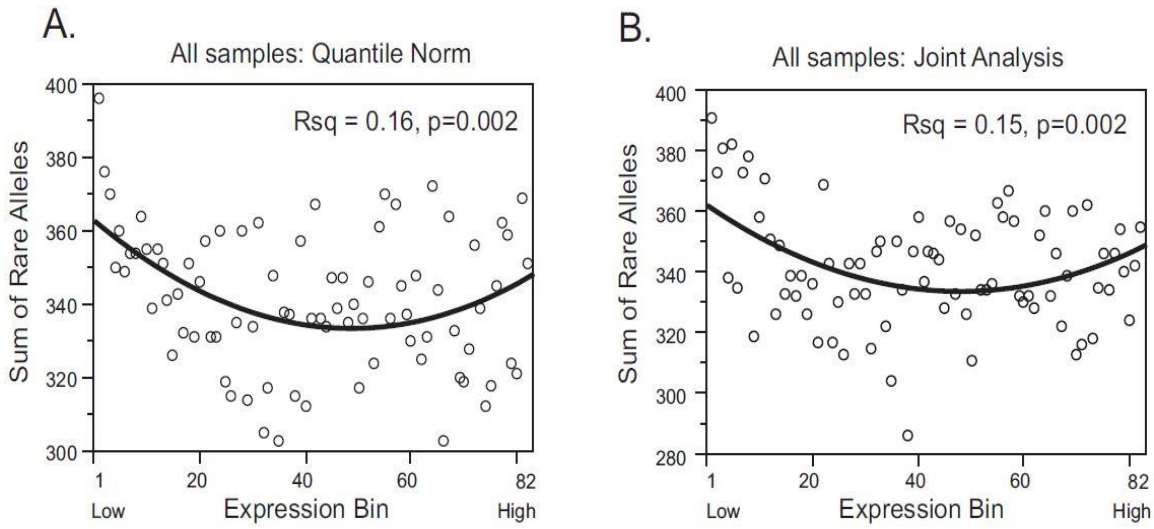

Figure 5.5 The rare variant burden using different normalization method of gene expression. A. The rare variant burden in sorted expression (QNM normalized) bins for 410 samples. (R^2=0.16, p-value=0.002) B. The rare variant burden in expression percentiles with a pooled z-score method. (R^2=0.15, p-value=0.002)

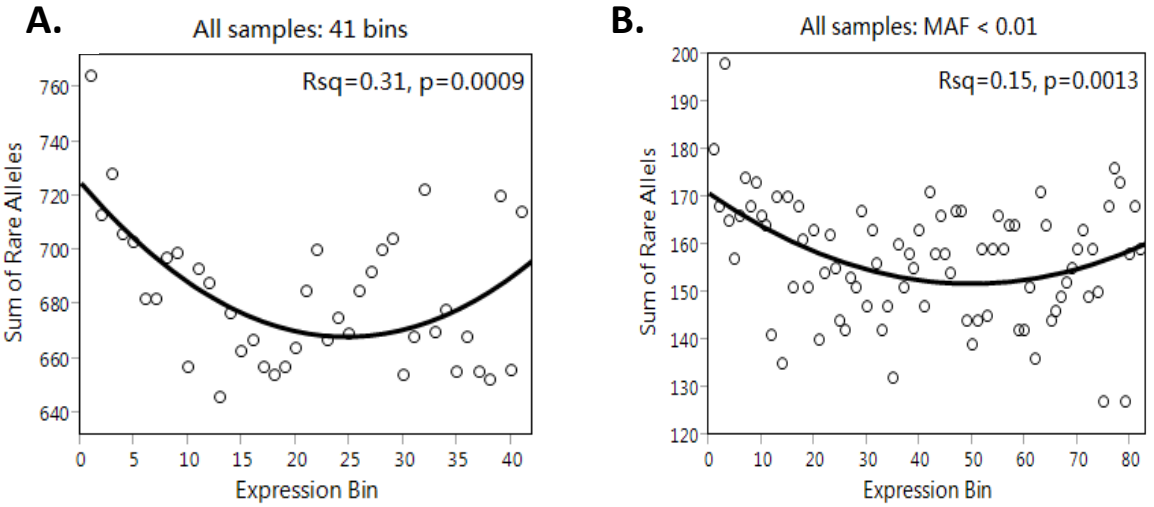

Figure 5.6. The rare variant burden using different bin sizes and different minor allele frequency cutoff. A. separating gene expression to 41 bins with 10 individuals in one bin B. using rare variants with MAF < 0.01

I also asked if the rare variant enrichment was due to certain subsets of genes or gene regions. The result was shown in Table 5.1 and Figure 5.7. I first categorized genes by RegulomeDB classifications. Lower classification of RegulomeDB (classes 1 to 4) means higher confidence that the variants lie within features such as DNAse Hypersensitive Sites or Transcription Factor Binding Sites. It showed that the rare variants which lie in class 1 to 4 of RegulomeDB were more enriched at the extremes than the variants in classes 5 to 7 ($p=0.002$ versus 0.08). Second, the variants were grouped into two subsets according to if they are in upstream or downstream of TSS. It showed that the rare variants in the downstream of TSS were more enriched at the extremes of gene expression in comparison of upstream ($p=0.00067$ vs 0.094). Third, I also asked if high or low abundance transcripts and if high or low polymorphic genes tend to show different patterns of rare variant enrichment. Result showed that the enrichment of rare variants occur in both high abundance transcripts and low abundance transcripts ($p=0.0062$, $p=0.0009$ separately). Similar result showed that the enrichment occur in high polymorphic genes in comparison with low polymorphic genes ($p=0.0009$ vs 0.23).

Table 5.1 Summary of quadratic model coefficients in different gene subsets. Model significance: * 0.001<p<0.05; ** 0.0001 <p<0.001; *** p<0.0001

	First Set			Second Set		
	Linear p	**Quad p**	**Model R^2**	**Linear p**	**Quad p**	**Model R^2**
Caucasian vs African	0.029	0.00046	0.17**	0.16	0.046	0.068*
RegulomeDB class 1-4 vs 5-7	0.16	0.0013	0.13*	0.08	0.15	0.055
Upstream vs Downstream	0.12	0.12	0.051	0.12	0.00042	0.15**
Low vs High Expression	0.00079	0.079	0.14**	0.35	0.0023	0.11*
Low vs High Polymorphism	0.50	0.12	0.032	0.032	0.0018	0.14**
with common eQTL vs not	0.027	0.0012	0.15**	0.33	0.058	0.049
with Metabochip vs not	0.81	0.45	0.007	0.0057	0.000046	0.23***
with Immunochip vs not	0.72	0.038	0.048	0.021	0.017	0.11*

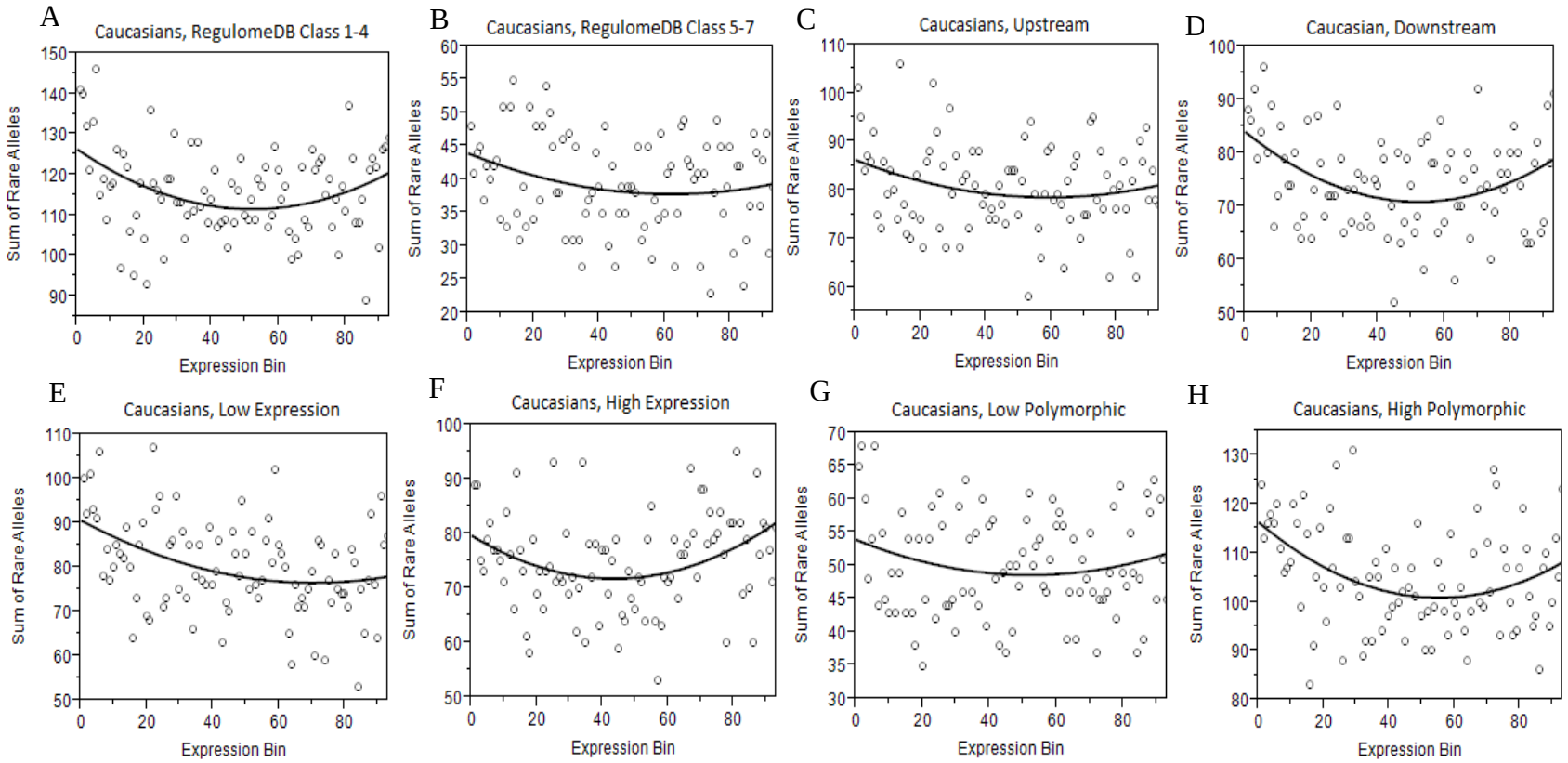

Figure 5.7 The rare variant burden in sorted expression (SNM normalized) bins for 279 Caucasians in different subsets. A, B. rare variants in regulomeDB classes 1-4, and 5-7 respectively C, D. rare variants in upstream and downstream of TSS E, F. genes with high and low transcript abundance G-H genes with low and high polymorphic levels.

Next, I also investigated the enrichment of rare variants with respect to gene function. It showed that rare variants were more enriched in the genes with common eQTLs (Figure 5.8A, p=0.0006) compared to the genes without eQTLs (p=0.11). Another remarkable finding was that rare variants tend to show more significant enrichment in genes that do not harbor SNPs in Metabochip or Immunochip. ($p=9\times10^{-6}$, Figure 5.8B, versus 0.73 for metabolic disease-related genes, and p=0.0047 versus 0.11 for immune-disease-related genes). This result might be explained by relaxation of purifying selection on genes not associated with disease.

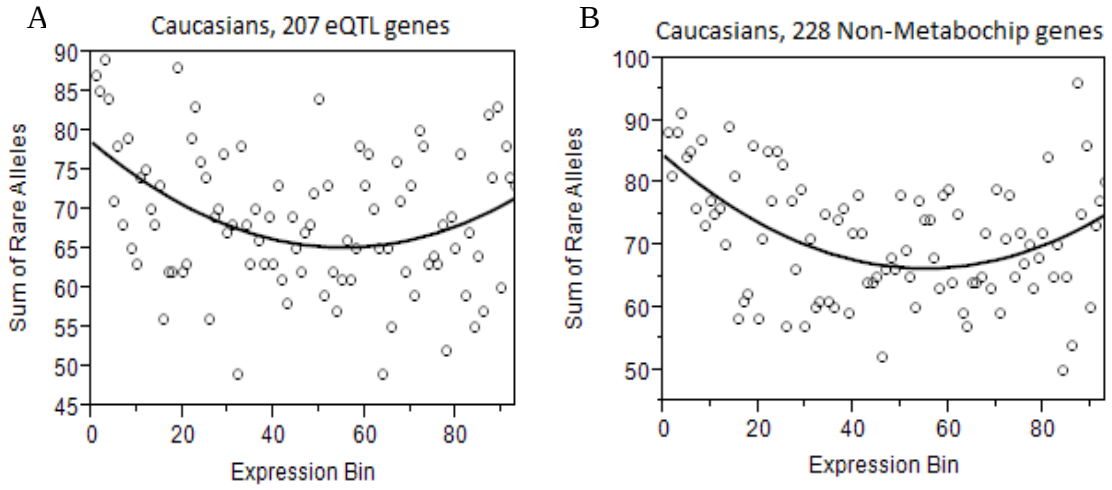

Figure 5.8 The rare variant burden in sorted expression (SNM normalized) bins for 279 Caucasians in different gene sets with respect to gene function. A. in genes with common eQTLs B. in genes without metabochip SNPs inside and within 5kb promoter regions.

The effect size calculation was based on the difference between the average expression levels of the samples with rare variants and the average expression levels of the rest samples in each SNP position (Figure 5.9A). When simulate with assigning effect size as gamma(1.5, 0.12) distributed and expression levels as normally distributed for each gene, while keeping the empirical SNP matrix, it got the enrichment results (Figure

5.9C) comparable to the true results. However the simulated distribution has smaller variance when compared to the empirical effect size (Figure 5.9B). This is likely due to the absence of technical noise in the simulated data. Accordingly, the observed effect sizes have only a modestly greater variance than those estimated in 100 random permutations (Figure 5.9D).

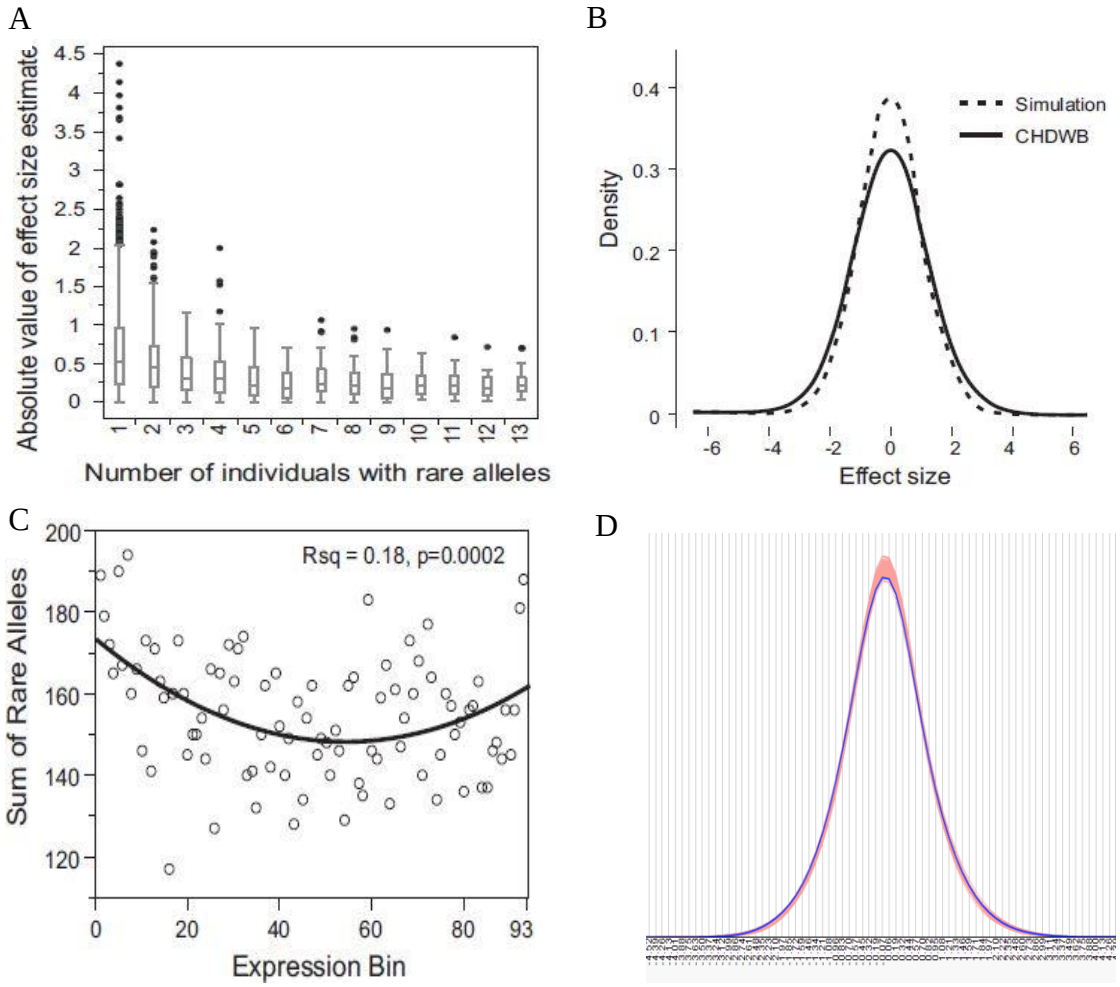

Figure 5.9 Effect size analysis. A. The distribution of absolute effect size. B. The distribution comparison of rare SNP effect size in CHDWB Caucasians and simulation. C. The rare variant burden with simulated effect size. D. The distribution comparison of rare SNP effect size in CHDWB Caucasians with 100 permutations.

I performed a replication test with 75 whole genome sequencing and corresponding whole blood RNA sequencing. When using the same 472 genes, the rare variants showed a significant enrichment at the extremes of gene expression (p=0.008, Figure 5.10A), in which the smaller p-value compared to 410 samples may be caused by smaller sample size. However, when enlarging the number of genes to 5000 genes and also adjusting the axes of variation, the results showed very significant enrichment of rare variants in decreased and increased gene expression ($p=2.3\times10^{-11}$, Figure 5.10B).

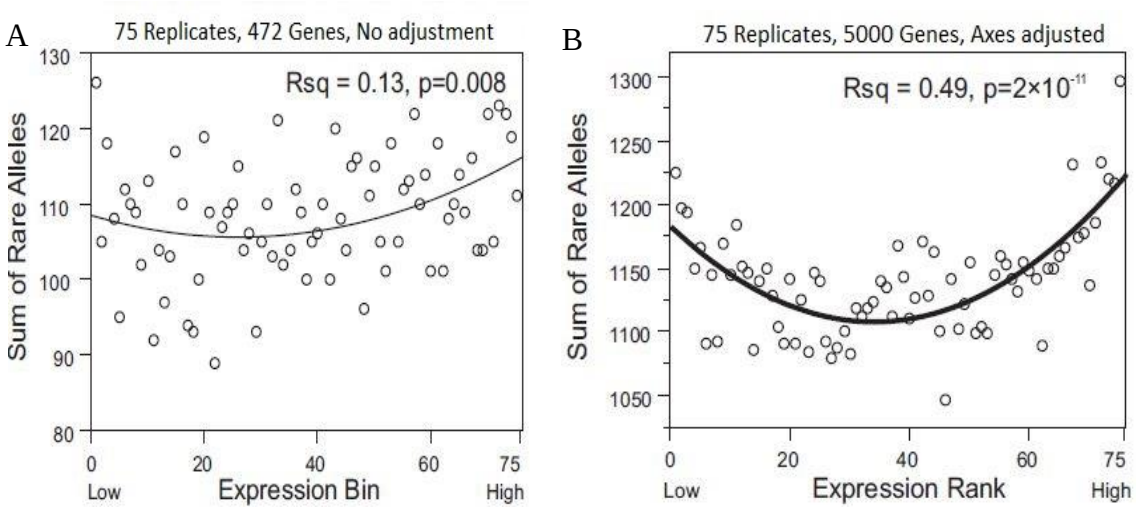

Figure 5.10 The rare variant association test with replicates. A. The rare variant burden in sorted expression (SNM normalized) bins with 75 individuals and 472 genes. B. The rare variant burden in sorted expression (SNM normalized) bins with 75 individuals and 4633 randomly selected genes.

In order to experimentally validate rare variant regulatory effects predicted from the statistical analysis, we used CRISPR/Cas9 to mutagenize four sites that had estimated effect sizes greater than 2.5 standard deviation units, in K562 erythroleukemia cells. Five individual clones were grown for each disruption and cleavage was confirmed using the T7E1 assay indicating 16 to 23% average heterozygosity. Quantitative RT-PCR confirmed that disruption of both sites where the rare variant associates with decreased

expression (in genes UQCC and TDP2), resulted in reduced transcript abundance, as did disruption of one of the two sites associated with increased expression (in genes COMMD4 and DHX29, Figure 5.11). Since this protocol causes small deletions rather than targeted replacement of polymorphisms, it is possible that the disruption removes a binding site for an activator in each case, leading to loss of gene expression whereas the alleles that associate with increased expression interact more strongly with the relevant transcription factor.

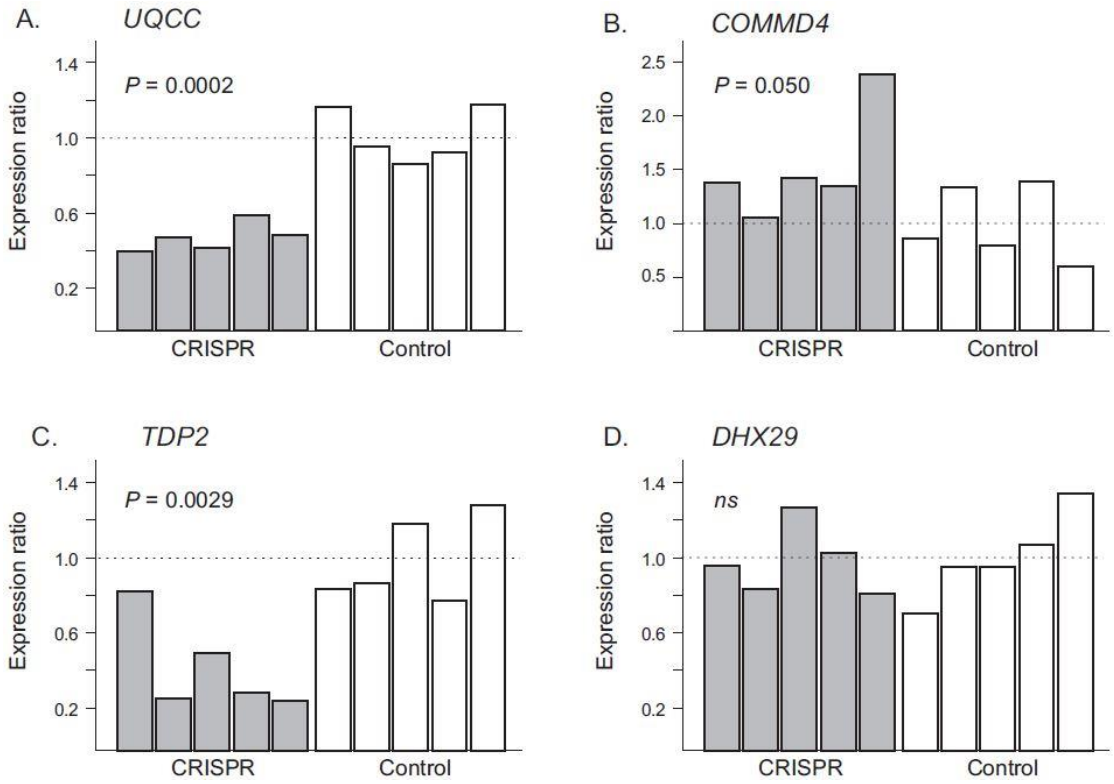

Figure 5.11 CRISPR / Cas9 mutagenesis validation of rare SNP regulatory effects. A. All 5 clones with disruptions in *UQCC* reduced expression almost by half. B. 3 clones with disruptions in *COMMD4* weakly increased the expression while 1 clone greatly increased the expression. C. 4 clones with disruptions in *TDP2* reduced expression to varying degrees. D. Only 1 clone out of 5 clones with disruptions in *DHX29* weakly increased expression.

Discussion

Whereas many studies working on gene expression regulation focus on common variants and rare coding variants, this study presents an analysis of rare variants in gene promoter regions with 410 samples. While it has been shown that large samples sizes are needed to detect rare variant association with enough power, here I use a novel burden test to perform the association test by considering each promoter region as independent and summing effects over all the genes. This method has shown the potential of using burden test as a method of detecting rare regulatory variants. It also provides candidate rare variants significantly associated with gene expression that can be tested in further gene-based analysis. Different gene subset regions and gene functions were analyzed which provides insight into the sources of the rare variant effects. These demonstrate that the position relative to the TSS, RegulomeDB classes that significantly associated rare variants belong to, and common eQTL and disease relatedness features, all influence whether genes are likely to harbor rare regulatory variant effects. This implies that further studies in different tissues and larger samples may pinpoint the sources of rare regulatory variants that may also impact gene expression. This study also demonstrates the good potential of CRISPR as a screening approach to validate rare regulatory variants that would affect gene expression.

Gene expression, whose levels could affect cellular functions and cellular states [129, 130], is an important intermediate phenotype between genotype and disease status. The extreme expression of genes may lead to abnormal phenotypes and even lead to diseases [131]. Many common SNPs have been found to be associated with gene expression levels [132-134]. Those common eSNPs would affect the average level of disease risks for the individual groups. For example, the individual groups carrying common risk alleles which increase or decrease the gene expression levels may in aggregate have higher disease risk than the groups which don’t carry risk alleles. In comparison with the overall effect of common eSNPs, the rare eSNPs in regulatory

regions will influence only less than 5% individuals who carry the risk alleles but may push their expression to extreme levels that lead to the change of their health status.

People may argue that rare variants do not deserve much attention because they only exist in a few numbers of individuals and also is likely to be selected against. However, Keinan et al [107] have stated that human population is recently under explosive growth, which is expanding by more than three orders of magnitude over 400 generations. This explosive population growth will contribute a larger number of rare variants and is likely to increase the individual genetic burden of complex disease risk. My result has shown that the rare variants tend to affect gene expression with a similar magnitude as common eSNPs. Considering the expanding number of rare variants and the three to four fold greater numbers of rare variants than common variants, the potential contribution cannot be ignored. In addition, rare coding variants which may directly cause change of protein structure and function tend to have strong effects on fitness, which would be filtered out by evolution. In contrast, rare eQTLs could have more easily escaped purifying selection. So exome sequencing based studies of rare coding variants is not sufficient and focus should shift to rare regulatory variants.

This study has limitations in some aspects and could be developed in further analysis with more accessible resources and funding. One of the limitations is that this study only focused on 2 kb regions in the vicinity of TSS. Although enrichment of cis-eSNPs were found within the regions of 1kb each side of TSS [46, 135], a large number of cis-eSNPs are located in more distal regions. This matches with the knowledge that many regions with regulatory functions such as DNase hypersensitive sites and transcription factor binding sites lie in distal positions relative to TSS [136, 137]. Expanding the study regions to tens or hundreds upstream of TSS will enable discovery of a more comprehensive list of rare variants that affect gene expression.

Here the tests were based on gene sequencing and transcript profiling in peripheral blood. It can be expected that if the test could be applied to single cell or tissue

types, it will reduce the noises caused by variable environments and cell counts. Especially, if the test could be implemented in cell or tissue types that are directly relevant to certain diseases, such as cardiac cells for coronary heart disease, it will provide more information about the rare variants effect on disease-related gene expression. To validate the rare variants effect on gene expression, more experimental validations are required such as experiments on knockout/knockdown mouse models. To further investigate whether there is similar effect of rare variants in promoter-proximal regions on diseases, the approaches applied here are required to implement on individuals who have congenital disorders.

The most important implication of this study is that the potential contribution of de novo and very rare variants to congenital disease cannot be ignored. Currently whole exome sequencing is demonstrating that there is a burden of rare variants in relevant subsets of genes in children born with developmental disorders and neurological conditions [138]. However, by no means all cases are explained by exome sequencing. I propose that rare regulatory variants that either greatly reduce or overexpress the gene product are likely to be an important source of non-syndromic conditions.

CHAPTER 6

CONCLUSION

This dissertation reveals the association between genetic variants and phenotypes in overall. Chapter 2 utilizes the common genetic variants recorded in GWAS database to calculate the polygenic risk score for 3 common diseases and 3 quantitative phenotypes. The result shows that even in a small cohort, polygenic score could explain 5% variation of disease risk and quantitative phenotypes, which is comparable to other studies. Chapter 3 is GWAS analysis on TNF-α and BMI/CRP with imputed genotypes, by participating in GWAS replication and meta-GWA study. The result reveals a replication of the top few SNPs being discovered as the most significant variants associated with TNF-α in the discovery phase. In addition, it shows that the effect size and effect direction in longitudinal model matches with the baseline model. The study also provided candidate variants to a meta BMI/CRP study throughout the genome. Chapter 4 makes use of the targeted resequencing to make a population genetics comparison focused on rare variants in promoter regions. It reveals that the rare variants in promoter regions in Africans are 3 times more than in Caucasians and 2 times more than Asians. Chapter 5 investigates the association between rare variants in promoter regions and gene expression levels. A significant enrichment of rare variants for both increased and decreased gene expression was observed in the study.

The studies were based on small cohorts – approximately 400 samples in CHDWB. The small samples would limit the power to detect the significant signals in genomic association studies. Spencer et al [139] show that thousands and more samples are needed to reach enough power to detect common variants with medium effect sizes in GWAS, whereas tens of thousands of samples are needed to detect common variants with small effect sizes in GWAS. Furthermore, for variants with smaller minor allele

frequency, much more samples compared to common variants are required to detect the associated SNPs (Figure 6.1). Their simulation study was based on single SNP analysis. More powerful statistical methods such as SKAT [53], C-alpha test [52] etc have been used for rare variants analysis, but large sample sizes are still required for association analysis.

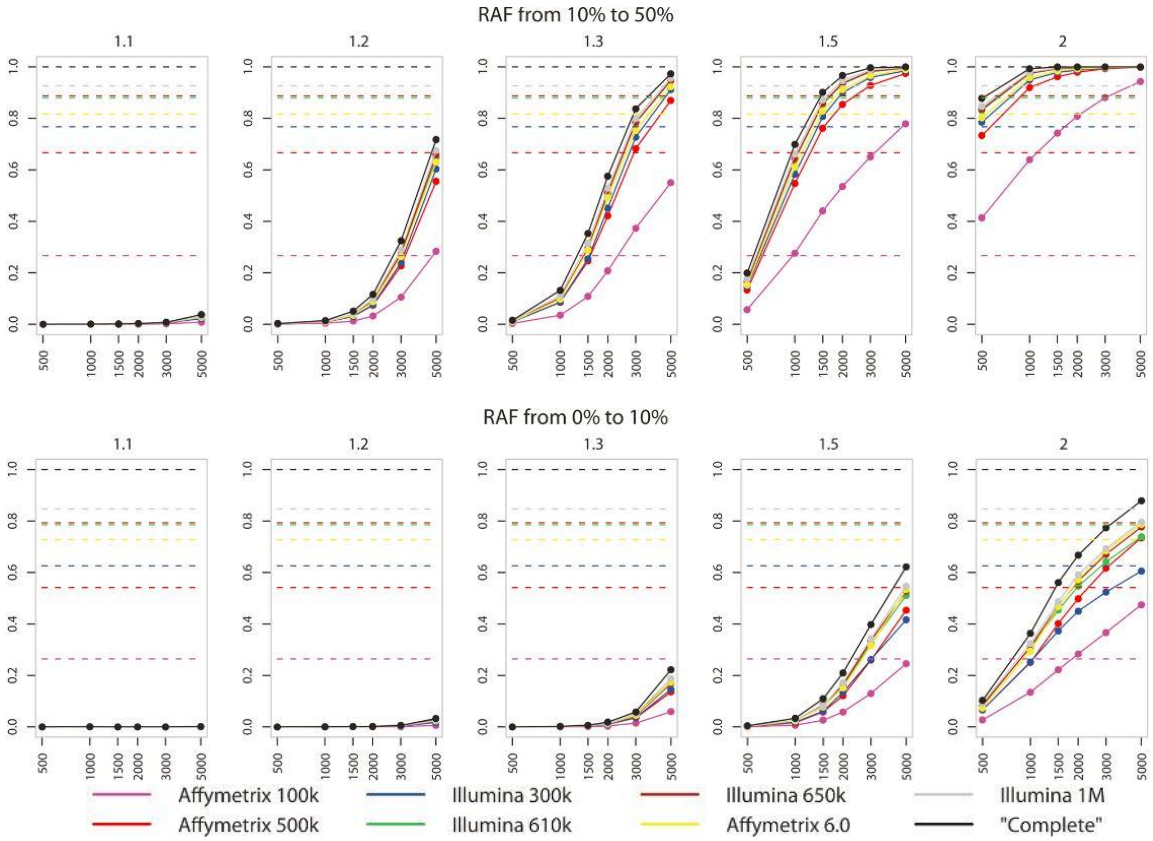

Figure 6.1 Power for Common versus Rare alleles based on case-control studies simulated in Caucasian population based on CEU HapMap panel. Power is shown as solid lines and coverage is represented as dotted lines with y-axis as the value. Increasing sample sizes are shown in x-axis. From left to right plots, effect sizes are increasing from 1.1 to 2. (taken from [139])

The rare variant association test in this dissertation was based on expression levels measured in healthy individuals. It would also be interesting to detect how the rare variants associate with disease status if a case control cohort were available. A trio study comprised of children with diseases and their parents with or without diseases could also be a powerful experimental design for the discovery of genetic factors including rare variants contributing to diseases.

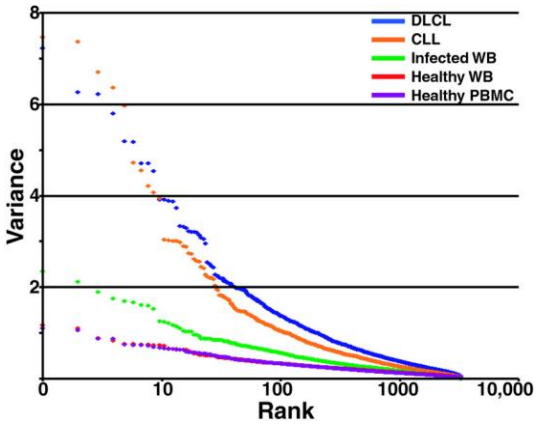

Figure 6.2 Variation of gene expression in different specimen groups for 45 samples. Ranks are based on 3,826 microarray measured genes. DLCL, diffuse large B-cell lymphomas. CLL, chronic lymphocytic leukemia. WB, whole blood. PBMC, peripheral blood mononuclear cell. Taken from [118]

The genetic information, traits and expression levels were all extracted from whole blood from the samples. There are several cell types such as lymphocytes, neutrophils, monocytes in whole blood. Each cell type may harbor its own expression characteristic and different genetic association pattern. Studies based on whole blood tends to overlook the underlying patterns in different cell types. Importantly, [140] has shown that gene expression variation in the whole blood of healthy samples is smaller than the variation observed from samples with bacterial infection, and is much smaller than that seen in certain constituent cell types (Figure 6.2). It would be interesting if the

study could be performed on isolated cell types, or specific types of stimulated cells.

Experimental validation is an important approach to validate the rare variant association. The CRISPR/Cas9 results performed in our lab have validated 3 of 4 rare variants with large effect sizes, at least with respect to the requirement for the transcription factor binding site that is affected by the rare SNP. The ideal situation would be that all rare variants with large or medium effect sizes could be validated by CRISPR/Cas9 assays. Future work may evaluate whether predictions from evolutionary or ENCODE data can help to narrow down causal variants in LD blocks..

The association between genetic variants and gene expression or clinical traits will shed light on genetic association with diseases, in consideration of how aberrant gene expression levels lead to abnormal phenotypes and how clinical traits imply the status of health. It helps understand the molecular causes of diseases and identifies genetic contribution to variability in persons' response to treatments. Each person has a unique genome. Thus the information on personal genome and personal genetic variants based on the knowledge of genetic variant association will result in personalized medicine [141], where genetic information together with clinical information could be used to predict health risk. When a certain disease is diagnosed, personal treatment could be advised based on the patient's genetic information, using methods such as that developed in this thesis to interpret the function of genomic sequence variants.

APPENDIX A

SUPPLEMENTARY TABLES FOR CHAPTER 4

Table A.1 Read depth summary for 2kb promoter regions of 472 genes

GENE	AVERAGE READ DEPTH	PROPORTION >= 20X	PROPORTION >= 100X	GENE	AVERAGE READ DEPTH	PROPORTION >= 20X	PROPORTION >= 100X
ABCC5	967.61	0.89	0.81	**LMAN2L**	640.11	1	0.90
ABHD10	232.04	1	0.54	**LMNA**	374.88	0.97	0.84
ABHD8	342.62	0.85	0.73	**LONP2**	613.94	0.97	0.94
ACAA2	648.05	0.95	0.88	**LRPPRC**	202.78	0.86	0.54
ACADM	531.89	0.92	0.77	**LRRC25**	1202.00	1	0.98
ACER3	591.86	0.86	0.62	**LRRCC1**	700.69	0.97	0.93
ACOX1	732.98	0.99	0.96	**LTN1**	522.29	0.68	0.59
ACP2	919.33	1	0.94	**M6PR**	942.01	1	1
ACTR10	529.78	0.99	0.96	**MAD2L1BP**	614.31	0.97	0.81
ACTR6	800.19	0.97	0.83	**MAF1**	583.77	0.70	0.56
ADCK3	791.69	0.92	0.84	**MAGEH1**	877.61	1	1
ADK	615.06	0.69	0.56	**MAN2A2**	879.03	1	0.98
ADPGK	896.71	0.91	0.89	**MAPK1IP1L**	612.44	0.90	0.84
ADRM1	493.36	0.76	0.63	**MAU2**	790.05	0.96	0.90
ADSS	659.47	0.85	0.73	**MBNL1**	857.20	0.96	0.94
AHSA2	457.19	0.85	0.71	**MBNL2**	853.06	1	1
AKR7A3	821.00	0.94	0.91	**MED16**	594.32	0.93	0.74
ALDH16A1	760.05	1	0.97	**MED22**	467.59	0.88	0.75
ALDH2	794.47	0.85	0.79	**MEFV**	811.08	0.94	0.85
ALDH3B1	923.90	1	1	**MFN2**	802.88	0.89	0.76
ALKBH1	690.88	0.84	0.73	**MGST3**	1003.18	0.98	0.95
ALPL	696.66	0.93	0.84	**MKI67IP**	571.78	0.99	0.97
ALPP	740.51	1	0.82	**MOBKL2C**	378.27	0.90	0.70
ALS2	815.91	0.92	0.90	**MPHOSPH10**	679.22	1	1
ANXA11	695.83	0.94	0.91	**MPL**	1216.80	1	0.98
AOAH	945.49	1	1	**MR1**	783.25	1	1
AP1G2	621.99	1	0.88	**MRPL17**	1000.30	1	0.87
AP1S1	636.69	0.97	0.86	**MRPL34**	806.25	1	0.94
APOL3	696.60	1	1	**MRPL43**	1159.66	1	0.99
ARAF	425.30	0.97	0.93	**MRPL52**	859.84	0.87	0.72
ARGLU1	523.46	0.82	0.75	**MRPL53**	734.97	1	1
ARID5B	430.77	0.99	0.90	**MRPS21**	962.16	0.95	0.88

ARL16	565.28	0.92	0.73	**MSH6**	382.81	0.93	0.80
ARPP19	564.50	0.95	0.80	**MSI2**	319.81	0.74	0.65
ASNSD1	944.69	1	1	**MSRA**	550.34	0.97	0.92
ASXL2	627.28	0.92	0.77	**MTMR3**	564.05	0.78	0.59
ATG4A	568.33	1	0.99	**MYC**	732.29	0.92	0.81
ATMIN	602.45	0.88	0.79	**NAA10**	865.92	0.83	0.71
ATP13A1	342.72	0.80	0.49	**NDUFB10**	740.15	0.89	0.81
ATP5J2	819.30	0.86	0.77	**NEDD8**	627.31	0.94	0.87
ATP5S	597.54	0.90	0.84	**NEDD9**	925.46	1	1
ATPIF1	764.53	1	0.97	**NFE2L1**	791.68	0.94	0.82
AZI2	670.69	0.92	0.90	**NONO**	1029.18	1	0.96
B4GALT4	496.83	0.95	0.75	**NOP10**	529.59	0.97	0.77
BEX2	381.95	0.90	0.79	**NSFL1C**	1296.91	0.99	0.96
BFAR	970.20	1	1	**NT5C3**	733.05	0.90	0.77
BIRC3	401.99	1	0.92	**NUDT18**	680.49	0.95	0.86
BRD8	982.92	1	1	**NUDT2**	619.01	1	0.98
BRF2	896.46	1	0.99	**NUDT5**	1049.42	1	1
BRMS1	1006.70	1	1	**NUFIP2**	580.26	1	0.93
BSDC1	789.71	1	1	**NUP43**	874.54	0.95	0.88
BTK	808.35	1	1	**OGFOD1**	782.13	1	1
BTN3A2	573.32	1	0.87	**OGFRL1**	474.91	0.81	0.75
C11ORF17	679.51	0.87	0.64	**ORMDL1**	497.88	0.85	0.79
C12ORF32	778.41	0.93	0.90	**ORMDL3**	845.41	0.95	0.89
C12ORF35	838.63	0.98	0.92	**OXR1**	484.16	0.75	0.52
C14ORF102	520.24	0.99	0.88	**PACSIN2**	565.73	0.87	0.77
C14ORF129	565.70	1	0.81	**PASK**	499.80	0.94	0.78
C14ORF142	589.99	0.93	0.90	**PCMTD2**	256.23	0.78	0.69
C14ORF179	795.58	1	0.97	**PCNA**	558.55	0.86	0.52
C15ORF63	1278.31	1	1	**PDCD2**	493.50	0.88	0.66
C17ORF90	636.12	0.87	0.76	**PDCD4**	761.35	0.92	0.87
C18ORF10	468.48	0.79	0.72	**PDK3**	912.91	0.95	0.92
C18ORF21	769.53	0.91	0.88	**PEA15**	754.35	0.73	0.57
C1ORF123	572.22	0.85	0.70	**PEX5**	581.60	0.96	0.84
C1ORF38	1382.12	0.96	0.94	**PFDN1**	737.82	1	1
C1ORF85	494.49	0.89	0.65	**PFN1**	318.03	0.85	0.64
C1ORF86	959.24	1	1	**PHF21A**	269.38	0.81	0.64
C2ORF28	495.29	0.76	0.59	**PIGH**	583.53	0.97	0.79
C2ORF44	1093.54	0.99	0.94	**PIGN**	876.91	0.87	0.85
C6ORF129	728.78	1	0.93	**PIGQ**	614.29	0.82	0.73
C7ORF25	634.51	0.97	0.82	**PILRB**	455.11	0.79	0.63
C8ORF40	838.46	0.97	0.92	**PIP4K2B**	169.65	0.64	0.48

C9ORF114	825.45	0.89	0.82	**PLAUR**	719.26	1	1
C9ORF78	665.00	0.96	0.87	**POLR1D**	385.44	0.95	0.74
CALR	500.59	0.78	0.58	**POR**	703.65	0.85	0.82
CAPG	707.90	1	1	**PPCS**	980.48	0.94	0.93
CARD8	513.31	1	0.89	**PPIL3**	1198.30	1	1
CARD9	775.91	1	1	**PRDX5**	1004.16	0.95	0.88
CAT	589.23	1	0.91	**PRKAR1A**	611.21	0.94	0.86
CCDC23	873.38	0.96	0.94	**PRRC1**	868.20	0.96	0.91
CCDC88B	970.39	0.87	0.81	**PTGDR**	643.27	0.97	0.88
CD160	592.79	0.92	0.87	**PTGS2**	711.44	1	0.80
CD96	573.27	1	1	**PWP1**	478.61	0.92	0.86
CDA	1353.78	1	1	**PYGB**	968.42	0.90	0.76
CDAN1	647.05	0.96	0.78	**RAB10**	1095.53	0.84	0.78
CDK10	130.28	0.68	0.45	**RAB11FIP2**	993.41	0.93	0.91
CEP63	858.12	0.91	0.89	**RAB24**	763.83	0.90	0.79
CFDP1	865.78	0.97	0.91	**RAB8B**	651.53	0.97	0.88
CHRNB1	849.92	0.93	0.92	**RANBP3**	519.86	0.84	0.76
CIB1	777.10	0.76	0.74	**RBBP4**	633.70	0.87	0.82
CIDECP	967.82	1	1	**RBMX2**	550.77	0.91	0.53
CKS2	700.91	0.94	0.86	**RCE1**	958.76	0.92	0.77
CLCN7	582.31	0.83	0.81	**RFWD3**	553.91	0.95	0.83
CLN3	716.00	1	0.97	**RHOT1**	628.72	0.93	0.91
COBRA1	899.51	0.81	0.65	**RNF130**	589.05	0.73	0.64
COMMD4	1013.46	1	1	**RNF135**	543.23	0.95	0.85
COMMD7	565.03	0.99	0.71	**RNF181**	706.38	1	1
COMMD9	776.21	1	1	**RNPS1**	412.63	0.70	0.60
CPPED1	600.82	0.93	0.76	**RPA1**	559.61	0.89	0.73
CPSF3L	449.59	0.93	0.76	**RPL10A**	770.14	0.89	0.86
CRIPT	518.45	1	0.92	**RPL13**	852.95	0.95	0.88
CRISPLD2	1034.78	0.97	0.96	**RPL14**	671.29	0.99	0.97
CRLS1	589.05	0.94	0.83	**RPL36AL**	574.53	0.84	0.74
CTDSP1	245.44	0.53	0.42	**RPL4**	722.99	1	1
CTNNAL1	645.51	0.90	0.82	**RPL9**	661.41	0.92	0.88
CTSH	371.70	0.94	0.86	**RPP38**	794.70	0.96	0.88
CTSO	608.93	0.98	0.92	**RPS6KB2**	973.00	0.96	0.89
CUL4B	817.25	1	1	**RPUSD3**	916.80	1	1
CWF19L1	640.57	1	0.93	**RPUSD4**	1101.65	1	0.99
CXCL16	899.39	0.88	0.72	**RRM2B**	677.43	0.93	0.78
CXCL5	794.46	1	1	**RRP12**	673.79	1	0.85
CXXC5	452.06	0.80	0.62	**RUFY1**	626.22	0.74	0.66
DCTN5	674.91	0.86	0.78	**SAMM50**	642.72	0.96	0.93

DCXR	324.40	0.75	0.52	**SCAND1**	496.43	0.87	0.67
DDOST	1132.84	1	1	**SDAD1**	1010.52	1	1
DDX41	828.38	0.96	0.92	**SENP2**	838.82	0.95	0.92
DDX52	873.73	1	1	**SERPINB8**	803.85	0.92	0.87
DEF6	561.85	1	1	**SERTAD1**	1102.76	0.84	0.78
DGUOK	632.13	1	1	**SERTAD2**	582.79	0.79	0.70
DHRS1	705.05	0.99	0.88	**SESN3**	742.60	0.83	0.80
DHRS3	334.45	0.94	0.78	**SETD3**	1150.20	0.86	0.83
DHX29	767.86	0.87	0.79	**SF3A1**	530.72	0.85	0.75
DHX38	1022.45	1	1	**SF3A3**	1036.68	1	0.97
DIAPH2	518.43	0.96	0.80	**SF3B4**	948.31	1	1
DKFZP686I15217	980.48	1	1	**SH2D1B**	722.32	1	1
DLAT	623.66	1	0.90	**SHCBP1**	828.27	0.96	0.94
DNAJC15	835.13	1	1	**SHROOM4**	830.56	0.99	0.93
DNAJC8	375.05	0.98	0.89	**SIDT2**	895.54	0.87	0.80
DOCK10	186.84	0.88	0.56	**SIRPB1**	894.23	1	1
DOCK11	416.21	0.77	0.53	**SIRPG**	979.11	1	0.97
DR1	473.86	0.95	0.76	**SLAMF7**	434.43	0.96	0.77
DRAM1	230.57	0.71	0.49	**SLC27A3**	640.91	0.84	0.65
DSTYK	883.87	0.91	0.86	**SLC35A3**	865.36	0.98	0.92
DUS2L	739.71	0.97	0.81	**SLC39A8**	433.18	0.95	0.84
E2F2	460.43	0.84	0.77	**SLC3A2**	888.93	0.94	0.90
E2F6	561.66	0.94	0.78	**SLC40A1**	781.17	0.94	0.92
EBP	696.64	0.91	0.82	**SMAP1**	591.71	0.95	0.87
ECH1	928.06	1	1	**SMARCE1**	721.27	1	0.95
ECHS1	195.40	0.59	0.45	**SNAP29**	507.51	0.93	0.71
EEF1G	1240.69	1	1	**SNORA70**	805.53	1	0.96
EIF1AX	552.87	0.88	0.70	**SNRNP25**	430.19	0.88	0.71
EIF2AK4	781.47	0.95	0.90	**SNX14**	771.09	0.98	0.91
EIF2B2	1080.54	1	1	**SNX29**	873.04	1	0.91
EIF2S3	601.99	0.90	0.73	**SP110**	963.83	1	0.87
EIF4A1	542.22	0.97	0.85	**SPNS1**	731.89	0.98	0.92
EIF4G3	294.18	0.78	0.57	**SRI**	532.23	1	0.85
EIF5	319.81	0.82	0.68	**SRP54**	614.96	1	1
EMR3	649.13	1	0.96	**STARD3NL**	769.57	0.98	0.94
EPB41	933.50	0.92	0.86	**STAT3**	637.01	0.96	0.88
EPHX2	567.37	0.85	0.78	**STAT4**	720.58	1	0.95
ERCC3	442.15	0.98	0.78	**STAT5A**	908.67	0.96	0.88
ERP27	658.69	1	0.83	**STAT6**	1238.77	1	1
ETS2	303.17	0.88	0.73	**STOM**	691.30	1	0.95
EVI2A	735.79	0.99	0.85	**STYXL1**	870.17	0.93	0.81

EXOC4	636.98	1	1	**SUMF1**	561.63	0.98	0.87
F13A1	899.97	1	0.97	**SURF6**	585.70	0.97	0.91
FAM193B	520.18	0.77	0.62	**SUSD3**	645.17	0.98	0.91
FAM49A	528.11	0.80	0.73	**SYS1**	612.65	1	1
FAM76B	679.80	0.91	0.88	**TAF1C**	513.54	0.86	0.67
FAR2	576.17	1	1	**TAF1L**	396.71	0.81	0.67
FBXO11	153.56	0.71	0.49	**TAGLN**	1038.40	1	1
FCER1G	941.32	1	0.98	**TAPBPL**	1231.22	1	0.86
FCRL3	775.73	0.96	0.80	**TATDN2**	552.59	0.93	0.79
FDFT1	732.41	0.96	0.89	**TBC1D15**	1166.57	1	1
FEZ2	731.32	0.86	0.64	**TBC1D2B**	794.52	0.94	0.74
FN3KRP	654.05	0.90	0.82	**TCEAL8**	850.76	1	1
FRG1	559.90	0.98	0.89	**TDP2**	1093.58	0.99	0.96
FXYD5	660.71	0.91	0.82	**TFE3**	712.39	0.86	0.80
GAB3	835.89	0.87	0.83	**TFG**	464.63	0.96	0.92
GALC	787.67	0.97	0.95	**THAP7**	339.32	0.95	0.80
GAPT	528.83	1	0.77	**TLR7**	761.71	1	1
GATAD2A	505.88	0.75	0.63	**TMEM140**	564.34	1	1
GATS	119.20	0.65	0.46	**TMEM149**	858.12	1	0.90
GDPD3	1063.49	1	0.96	**TMEM175**	657.46	0.83	0.73
GMCL1	658.06	0.86	0.81	**TMEM199**	706.46	0.98	0.94
GNA12	722.36	0.90	0.80	**TNFRSF4**	545.47	0.89	0.82
GNPAT	947.19	1	0.99	**TNFSF8**	948.24	1	0.89
GPAA1	556.60	0.71	0.57	**TOMM7**	1088.64	1	1
GPATCH4	1133.92	1	1	**TPM1**	378.53	0.78	0.75
GPX7	588.48	0.93	0.81	**TRAF3IP3**	1277.39	1	1
GRK6	539.47	0.95	0.79	**TRAF5**	978.67	0.95	0.89
GSPT2	703.47	1	1	**TRAK1**	1273.42	1	1
GTF2F1	420.48	0.95	0.84	**TRAPPC4**	834.08	1	1
H1F0	758.59	0.87	0.66	**TRAPPC5**	701.23	0.90	0.73
HADH	795.50	0.94	0.85	**TRAPPC6B**	798.97	1	1
HARS2	593.39	1	1	**TRIM4**	546.64	0.98	0.94
HBP1	499.89	0.75	0.66	**TRPC4AP**	909.87	0.97	0.92
HBQ1	648.99	0.65	0.63	**TSPAN33**	585.97	0.99	0.98
HCLS1	1193.39	1	1	**TSSC1**	685.42	1	0.96
HDDC2	753.18	0.83	0.75	**TUBA1A**	862.11	0.98	0.95
HDHD1	781.23	0.89	0.82	**TUBA1B**	715.79	0.98	0.96
HEBP2	544.33	0.89	0.78	**TUFM**	741.53	1	0.95
HIF1AN	705.80	1	0.88	**TXNIP**	969.72	1	1
HIST1H2BD	701.43	1	0.99	**UBA7**	720.04	1	0.90
HNRNPC	932.01	1	0.98	**UBE2L3**	737.06	0.92	0.88

HNRNPH3	453.81	0.97	0.82	**UBL4A**	866.24	0.85	0.83
HPS1	524.99	0.97	0.85	**UBN1**	694.43	0.64	0.42
HSCB	1029.53	1	1	**UBR2**	870.48	0.98	0.91
HSD17B11	543.85	1	0.75	**UGDH**	564.06	0.90	0.83
HSD17B12	813.02	1	0.99	**UQCC**	275.76	1	0.59
HSP90AB1	663.04	0.96	0.91	**USO1**	683.82	1	0.91
HSPA4	534.87	0.88	0.82	**USP4**	652.99	0.97	0.92
HSPA8	806.84	1	0.96	**USP48**	764.11	0.90	0.85
HSPA9	803.88	0.94	0.92	**UTP18**	423.66	0.87	0.71
ICOS	687.29	0.94	0.76	**UXT**	923.69	1	1
ID3	614.86	1	0.93	**VAMP1**	429.91	0.90	0.80
IKBIP	476.12	0.83	0.67	**VAMP8**	1193.93	1	1
IL18R1	660.24	1	1	**VASP**	559.59	0.87	0.81
IL8	585.82	1	1	**VEZT**	686.02	0.98	0.96
ILF3	621.12	0.90	0.78	**VIM**	553.86	0.87	0.76
IPP	802.33	0.98	0.78	**VPS4B**	746.41	1	0.91
IQCB1	698.57	1	0.89	**WAS**	1086.37	1	1
IQGAP1	865.04	0.95	0.89	**WDFY2**	463.45	0.86	0.67
ITGAM	849.98	0.98	0.91	**WDR45**	469.64	0.91	0.68
ITGAX	959.37	1	0.99	**XBP1**	515.17	0.87	0.76
ITK	588.87	1	0.99	**XRCC6BP1**	667.70	0.98	0.92
KATNA1	365.26	0.86	0.76	**YIF1A**	1041.26	0.97	0.95
KCTD10	917.67	1	0.83	**YIPF3**	604.21	1	1
KIAA0319L	400.98	0.70	0.63	**ZAK**	826.41	0.94	0.83
KIAA0368	238.50	0.83	0.69	**ZDHHC17**	949.15	0.93	0.91
KIAA1191	676.71	1	0.98	**ZFP90**	633.00	0.88	0.68
KIAA1598	828.53	0.81	0.67	**ZMIZ1**	180.24	0.72	0.38
KIAA1737	601.89	0.83	0.75	**ZMIZ2**	1107.48	0.91	0.90
KLF4	378.59	0.80	0.60	**ZNF185**	1046.58	1	1
KLHDC4	303.59	0.78	0.64	**ZNF266**	958.87	1	1
KRIT1	733.81	0.95	0.93	**ZNF407**	764.85	1	1
LACTB	456.72	0.74	0.65	**ZNF439**	687.04	1	1
LAPTM5	1147.57	1	1	**ZNF549**	815.31	1	0.78
LARP4	948.51	1	1	**ZNF613**	913.90	1	1
LASS5	717.42	0.90	0.76	**ZNF671**	743.34	1	1
LDLRAP1	598.40	0.93	0.86	**ZNF75D**	818.75	1	1
LFNG	749.90	1	0.94	**ZNF787**	633.16	0.67	0.66
LILRB2	1260.17	1	0.99	**ZNF839**	801.56	0.98	0.95
LINS	545.28	0.96	0.90	**ZRANB1**	601.90	0.99	0.72
LIPT1	614.28	1	0.85	**ZYX**	184.94	0.67	0.51

Table A.2 The number of rare, common, and private SNPs and an estimate of the polymorphism rate (Pi) in promoter region for 472 genes.

GENE	COMMON SNPS	RARE SNPS	PRIVATE SNPS	PI PER BASE	GENE	COMMON SNPS	RARE SNPS	PRIVATE SNPS	PI PER BASE
ABCC5	2	26	11	0.00044	**LMAN2L**	6	13	7	0.00056
ABHD10	9	19	11	0.00144	**LMNA**	2	14	7	0.00044
ABHD8	5	16	10	0.00165	**LONP2**	5	23	13	0.00065
ACAA2	2	17	7	0.00042	**LRPPRC**	5	45	16	0.00097
ACADM	11	28	17	0.00214	**LRRC25**	0	11	5	0.00002
ACER3	2	10	3	0.00080	**LRRCC1**	1	24	12	0.00026
ACOX1	3	18	7	0.00034	**LTN1**	5	18	12	0.00066
ACP2	4	11	6	0.00059	**M6PR**	3	19	9	0.00065
ACTR10	1	20	12	0.00027	**MAD2L1BP**	0	15	6	0.00012
ACTR6	1	25	16	0.00018	**MAF1**	3	24	11	0.00059
ADCK3	2	13	6	0.00037	**MAGEH1**	1	9	5	0.00025
ADK	2	10	3	0.00072	**MAN2A2**	4	24	10	0.00089
ADPGK	1	22	7	0.00037	**MAPK1IP1L**	3	24	15	0.00081
ADRM1	0	18	11	0.00011	**MAU2**	3	16	10	0.00039
ADSS	1	14	7	0.00029	**MBNL1**	1	12	10	0.00016
AHSA2	3	25	15	0.00074	**MBNL2**	1	11	5	0.00010
AKR7A3	2	18	10	0.00036	**MED16**	7	27	21	0.00177
ALDH16A1	6	26	14	0.00128	**MED22**	7	21	10	0.00097
ALDH2	0	8	4	0.00021	**MEFV**	2	22	13	0.00057
ALDH3B1	3	35	17	0.00081	**MFN2**	2	29	17	0.00083
ALKBH1	4	17	4	0.00044	**MGST3**	13	26	13	0.00285
ALPL	2	24	11	0.00031	**MKI67IP**	1	16	8	0.00023
ALPP	7	25	11	0.00120	**MOBKL2C**	5	21	7	0.00061
ALS2	3	14	9	0.00049	**MPHOSPH10**	4	18	8	0.00092
ANXA11	3	12	6	0.00052	**MPL**	2	18	9	0.00052
AOAH	8	18	11	0.00191	**MR1**	2	26	12	0.00043
AP1G2	0	14	8	0.00003	**MRPL17**	6	35	15	0.00108
AP1S1	5	20	12	0.00093	**MRPL34**	3	13	11	0.00076
APOL3	5	16	7	0.00109	**MRPL43**	2	13	6	0.00044
ARAF	3	13	10	0.00040	**MRPL52**	5	11	2	0.00089
ARGLU1	3	18	13	0.00067	**MRPL53**	1	22	9	0.00029
ARID5B	3	12	6	0.00047	**MRPS21**	4	22	12	0.00097
ARL16	5	16	8	0.00084	**MSH6**	7	15	12	0.00134
ARPP19	1	23	8	0.00034	**MSI2**	1	14	8	0.00048
ASNSD1	4	24	11	0.00067	**MSRA**	6	20	14	0.00092
ASXL2	3	15	8	0.00053	**MTMR3**	2	14	5	0.00023
ATG4A	2	5	3	0.00050	**MYC**	4	20	8	0.00038

ATMIN	4	19	12	0.00050	**NAA10**	3	10	1	0.00050
ATP13A1	1	25	15	0.00029	**NDUFB10**	8	24	10	0.00123
ATP5J2	1	27	9	0.00036	**NEDD8**	2	17	10	0.00027
ATP5S	2	24	10	0.00084	**NEDD9**	3	20	10	0.00041
ATPIF1	7	19	9	0.00139	**NFE2L1**	4	14	7	0.00083
AZI2	2	20	11	0.00054	**NONO**	0	8	4	0.00002
B4GALT4	6	15	4	0.00077	**NOP10**	6	16	4	0.00104
BEX2	6	9	3	0.00086	**NSFL1C**	6	19	10	0.00130
BFAR	1	21	19	0.00007	**NT5C3**	7	14	6	0.00142
BIRC3	1	22	11	0.00011	**NUDT18**	5	17	9	0.00179
BRD8	3	16	9	0.00071	**NUDT2**	6	19	10	0.00091
BRF2	7	19	9	0.00119	**NUDT5**	2	16	11	0.00025
BRMS1	0	13	6	0.00003	**NUFIP2**	1	18	8	0.00016
BSDC1	0	12	5	0.00004	**NUP43**	1	13	6	0.00030
BTK	1	9	5	0.00012	**OGFOD1**	4	18	8	0.00085
BTN3A2	9	18	9	0.00119	**OGFRL1**	5	11	6	0.00068
C11ORF17	0	16	7	0.00066	**ORMDL1**	4	30	16	0.00084
C12ORF32	6	17	7	0.00121	**ORMDL3**	1	19	7	0.00026
C12ORF35	4	28	16	0.00078	**OXR1**	1	12	8	0.00009
C14ORF102	2	18	10	0.00035	**PACSIN2**	4	17	4	0.00086
C14ORF129	3	16	9	0.00045	**PASK**	4	30	11	0.00115
C14ORF142	4	15	8	0.00080	**PCMTD2**	5	10	6	0.00097
C14ORF179	0	27	18	0.00007	**PCNA**	3	14	11	0.00026
C15ORF63	0	23	8	0.00007	**PDCD2**	5	20	7	0.00094
C17ORF90	2	14	9	0.00027	**PDCD4**	4	15	10	0.00045
C18ORF10	2	12	8	0.00051	**PDK3**	6	10	5	0.00119
C18ORF21	3	23	13	0.00065	**PEA15**	1	10	4	0.00010
C1ORF123	1	17	11	0.00023	**PEX5**	3	19	10	0.00080
C1ORF38	5	17	7	0.00050	**PFDN1**	2	11	8	0.00044
C1ORF85	1	15	7	0.00013	**PFN1**	1	25	14	0.00054
C1ORF86	4	24	10	0.00091	**PHF21A**	1	11	2	0.00022
C2ORF28	3	22	15	0.00046	**PIGH**	4	30	9	0.00136
C2ORF44	4	11	5	0.00055	**PIGN**	3	19	10	0.00033
C6ORF129	1	17	12	0.00018	**PIGQ**	1	14	10	0.00029
C7ORF25	5	12	6	0.00086	**PILRB**	1	13	6	0.00053
C8ORF40	1	20	11	0.00056	**PIP4K2B**	1	16	7	0.00034
C9ORF114	4	11	4	0.00071	**PLAUR**	2	15	5	0.00050
C9ORF78	7	14	8	0.00158	**POLR1D**	8	30	15	0.00093
CALR	2	19	8	0.00036	**POR**	3	14	10	0.00050
CAPG	3	20	10	0.00055	**PPCS**	1	18	10	0.00032
CARD8	18	15	9	0.00351	**PPIL3**	0	13	6	0.00030

CARD9	1	29	20	0.00035	**PRDX5**	4	28	16	0.00060
CAT	7	13	4	0.00143	**PRKAR1A**	1	20	10	0.00025
CCDC23	3	11	5	0.00072	**PRRC1**	7	15	7	0.00153
CCDC88B	6	18	11	0.00110	**PTGDR**	7	23	12	0.00104
CD160	2	8	4	0.00020	**PTGS2**	2	20	9	0.00036
CD96	9	19	7	0.00190	**PWP1**	7	16	7	0.00084
CDA	13	15	6	0.00201	**PYGB**	5	20	7	0.00119
CDAN1	4	12	5	0.00070	**RAB10**	1	19	15	0.00053
CDK10	5	16	6	0.00121	**RAB11FIP2**	0	8	5	0.00039
CEP63	3	23	12	0.00095	**RAB24**	2	10	6	0.00065
CFDP1	4	35	20	0.00122	**RAB8B**	1	25	15	0.00032
CHRNB1	0	12	10	0.00025	**RANBP3**	0	6	3	0.00007
CIB1	5	11	6	0.00150	**RBBP4**	3	22	11	0.00034
CIDECP	7	17	14	0.00105	**RBMX2**	0	14	3	0.00007
CKS2	9	27	13	0.00104	**RCE1**	1	10	6	0.00020
CLCN7	8	12	6	0.00247	**RFWD3**	8	26	14	0.00166
CLN3	1	24	14	0.00033	**RHOT1**	2	20	12	0.00047
COBRA1	4	18	6	0.00060	**RNF130**	7	14	7	0.00126
COMMD4	2	27	14	0.00050	**RNF135**	0	27	10	0.00081
COMMD7	1	17	8	0.00020	**RNF181**	1	16	10	0.00017
COMMD9	7	15	11	0.00105	**RNPS1**	2	14	9	0.00088
CPPED1	4	39	18	0.00096	**RPA1**	3	20	11	0.00106
CPSF3L	1	10	5	0.00019	**RPL10A**	5	27	16	0.00074
CRIPT	3	34	18	0.00069	**RPL13**	0	41	18	0.00049
CRISPLD2	11	21	13	0.00146	**RPL14**	5	33	15	0.00088
CRLS1	4	19	11	0.00111	**RPL36AL**	3	20	8	0.00086
CTDSP1	4	27	17	0.00063	**RPL4**	3	26	16	0.00047
CTNNAL1	2	9	6	0.00086	**RPL9**	6	27	19	0.00098
CTSH	8	15	10	0.00161	**RPP38**	4	36	18	0.00091
CTSO	13	21	10	0.00203	**RPS6KB2**	1	11	11	0.00060
CUL4B	1	8	2	0.00025	**RPUSD3**	5	24	11	0.00044
CWF19L1	6	23	15	0.00080	**RPUSD4**	5	25	14	0.00087
CXCL16	3	14	6	0.00078	**RRM2B**	7	22	13	0.00102
CXCL5	4	17	9	0.00057	**RRP12**	5	17	9	0.00094
CXXC5	2	8	3	0.00032	**RUFY1**	2	27	12	0.00073
DCTN5	3	23	16	0.00042	**SAMM50**	2	12	5	0.00039
DCXR	4	14	5	0.00080	**SCAND1**	2	15	6	0.00022
DDOST	2	21	10	0.00035	**SDAD1**	2	20	9	0.00048
DDX41	1	17	10	0.00014	**SENP2**	1	15	8	0.00025
DDX52	2	22	6	0.00072	**SERPINB8**	6	23	19	0.00085
DEF6	2	23	11	0.00055	**SERTAD1**	3	25	16	0.00062

DGUOK	2	17	9	0.00042	**SERTAD2**	1	16	5	0.00062
DHRS1	3	18	9	0.00074	**SESN3**	2	11	2	0.00058
DHRS3	1	19	11	0.00027	**SETD3**	2	17	12	0.00055
DHX29	3	29	15	0.00051	**SF3A1**	6	10	6	0.00093
DHX38	4	17	9	0.00088	**SF3A3**	4	17	9	0.00069
DIAPH2	4	15	7	0.00031	**SF3B4**	6	15	9	0.00042
DKFZP686I15217	2	22	13	0.00037	**SH2D1B**	3	16	8	0.00078
DLAT	3	27	12	0.00083	**SHCBP1**	1	21	9	0.00012
DNAJC15	11	16	7	0.00225	**SHROOM4**	1	23	11	0.00034
DNAJC8	5	16	11	0.00101	**SIDT2**	1	24	12	0.00021
DOCK10	3	27	14	0.00100	**SIRPB1**	4	22	14	0.00083
DOCK11	2	11	5	0.00015	**SIRPG**	10	21	12	0.00136
DR1	5	15	4	0.00084	**SLAMF7**	0	19	12	0.00004
DRAM1	9	14	5	0.00128	**SLC27A3**	3	11	7	0.00033
DSTYK	3	14	10	0.00062	**SLC35A3**	3	17	9	0.00051
DUS2L	2	15	8	0.00021	**SLC39A8**	1	17	10	0.00018
E2F2	2	8	7	0.00033	**SLC3A2**	4	20	12	0.00028
E2F6	3	22	9	0.00072	**SLC40A1**	6	13	8	0.00130
EBP	2	10	5	0.00023	**SMAP1**	6	16	7	0.00109
ECH1	4	23	11	0.00056	**SMARCE1**	1	20	15	0.00025
ECHS1	1	5	1	0.00045	**SNAP29**	5	18	7	0.00111
EEF1G	0	20	14	0.00003	**SNORA70**	1	10	3	0.00007
EIF1AX	2	17	12	0.00025	**SNRNP25**	3	24	12	0.00078
EIF2AK4	5	27	13	0.00095	**SNX14**	3	18	9	0.00063
EIF2B2	2	16	11	0.00053	**SNX29**	2	11	6	0.00068
EIF2S3	4	15	9	0.00063	**SP110**	5	17	10	0.00068
EIF4A1	5	31	20	0.00083	**SPNS1**	2	25	11	0.00018
EIF4G3	1	14	6	0.00040	**SRI**	4	24	13	0.00070
EIF5	5	18	10	0.00096	**SRP54**	8	32	21	0.00163
EMR3	0	24	13	0.00019	**STARD3NL**	7	19	9	0.00144
EPB41	2	14	7	0.00042	**STAT3**	1	15	9	0.00028
EPHX2	4	15	3	0.00060	**STAT4**	4	15	7	0.00063
ERCC3	0	24	15	0.00007	**STAT5A**	0	11	6	0.00003
ERP27	7	25	9	0.00178	**STAT6**	1	12	4	0.00015
ETS2	5	11	4	0.00072	**STOM**	3	23	13	0.00076
EVI2A	5	16	8	0.00125	**STYXL1**	5	21	13	0.00157
EXOC4	0	20	11	0.00045	**SUMF1**	5	19	10	0.00069
F13A1	5	29	16	0.00090	**SURF6**	1	15	10	0.00013
FAM193B	1	14	7	0.00036	**SUSD3**	7	16	10	0.00160
FAM49A	5	11	8	0.00075	**SYS1**	5	19	10	0.00096
FAM76B	5	31	19	0.00136	**TAF1C**	12	31	19	0.00260

FAR2	7	14	4	0.00116	**TAF1L**	1	20	7	0.00022
FBXO11	2	22	13	0.00038	**TAGLN**	6	35	14	0.00100
FCER1G	3	25	20	0.00054	**TAPBPL**	1	15	4	0.00047
FCRL3	5	16	7	0.00113	**TATDN2**	5	21	7	0.00060
FDFT1	7	51	28	0.00142	**TBC1D15**	3	9	3	0.00047
FEZ2	6	25	15	0.00205	**TBC1D2B**	10	12	6	0.00132
FN3KRP	2	21	7	0.00044	**TCEAL8**	1	11	4	0.00024
FRG1	4	22	10	0.00165	**TDP2**	4	23	10	0.00064
FXYD5	4	21	15	0.00068	**TFE3**	4	9	4	0.00026
GAB3	0	10	7	0.00001	**TFG**	4	29	9	0.00106
GALC	8	19	13	0.00122	**THAP7**	8	26	11	0.00140
GAPT	5	32	16	0.00106	**TLR7**	1	18	8	0.00015
GATAD2A	2	21	14	0.00046	**TMEM140**	7	16	12	0.00154
GATS	0	7	3	0.00019	**TMEM149**	6	17	11	0.00070
GDPD3	2	20	11	0.00040	**TMEM175**	1	10	5	0.00017
GMCL1	4	22	10	0.00098	**TMEM199**	1	24	10	0.00014
GNA12	7	22	16	0.00138	**TNFRSF4**	4	27	18	0.00046
GNPAT	5	30	13	0.00121	**TNFSF8**	5	22	15	0.00141
GPAA1	0	12	4	0.00025	**TOMM7**	12	17	5	0.00265
GPATCH4	0	22	7	0.00052	**TPM1**	1	14	4	0.00084
GPX7	4	17	10	0.00106	**TRAF3IP3**	1	13	10	0.00042
GRK6	1	11	6	0.00013	**TRAF5**	1	28	15	0.00025
GSPT2	2	9	7	0.00033	**TRAK1**	2	29	18	0.00080
GTF2F1	1	16	7	0.00048	**TRAPPC4**	10	30	9	0.00155
H1F0	1	11	8	0.00023	**TRAPPC5**	4	17	9	0.00092
HADH	1	14	7	0.00018	**TRAPPC6B**	4	19	9	0.00045
HARS2	3	16	9	0.00061	**TRIM4**	2	16	15	0.00033
HBP1	6	21	12	0.00124	**TRPC4AP**	1	15	7	0.00025
HBQ1	0	16	6	0.00034	**TSPAN33**	2	19	12	0.00044
HCLS1	1	9	5	0.00009	**TSSC1**	8	20	9	0.00148
HDDC2	6	17	9	0.00149	**TUBA1A**	1	18	10	0.00029
HDHD1	3	13	3	0.00072	**TUBA1B**	5	27	15	0.00107
HEBP2	1	20	13	0.00077	**TUFM**	2	15	8	0.00048
HIF1AN	6	19	12	0.00107	**TXNIP**	1	20	13	0.00019
HIST1H2BD	5	40	21	0.00072	**UBA7**	0	10	6	0.00002
HNRNPC	1	19	10	0.00027	**UBE2L3**	3	19	9	0.00082
HNRNPH3	0	29	21	0.00026	**UBL4A**	3	8	3	0.00047
HPS1	2	21	13	0.00034	**UBN1**	7	23	13	0.00198
HSCB	2	18	9	0.00034	**UBR2**	3	17	13	0.00055
HSD17B11	5	19	8	0.00096	**UGDH**	8	21	9	0.00144
HSD17B12	4	26	12	0.00073	**UQCC**	2	17	5	0.00054

HSP90AB1	12	38	15	0.00199	**USO1**	4	15	8	0.00056
HSPA4	3	19	12	0.00062	**USP4**	1	13	8	0.00023
HSPA8	15	41	23	0.00139	**USP48**	4	21	7	0.00066
HSPA9	0	31	13	0.00017	**UTP18**	2	27	13	0.00045
ICOS	6	11	8	0.00068	**UXT**	0	7	4	0.00002
ID3	6	42	28	0.00113	**VAMP1**	11	15	8	0.00201
IKBIP	2	25	14	0.00043	**VAMP8**	3	19	7	0.00067
IL18R1	4	12	9	0.00084	**VASP**	0	14	4	0.00005
IL8	3	14	8	0.00081	**VEZT**	5	15	10	0.00073
ILF3	4	30	12	0.00109	**VIM**	1	21	10	0.00050
IPP	5	21	9	0.00120	**VPS4B**	0	23	14	0.00010
IQCB1	2	18	10	0.00036	**WAS**	0	14	8	0.00008
IQGAP1	2	17	9	0.00107	**WDFY2**	0	28	17	0.00007
ITGAM	2	17	7	0.00056	**WDR45**	0	10	5	0.00005
ITGAX	3	20	10	0.00050	**XBP1**	2	9	7	0.00049
ITK	2	19	9	0.00028	**XRCC6BP1**	6	15	10	0.00161
KATNA1	4	17	6	0.00087	**YIF1A**	1	12	8	0.00037
KCTD10	5	13	9	0.00062	**YIPF3**	2	25	18	0.00033
KIAA0319L	0	8	6	0.00017	**ZAK**	0	25	9	0.00099
KIAA0368	0	17	9	0.00008	**ZDHHC17**	3	18	7	0.00049
KIAA1191	4	15	10	0.00059	**ZFP90**	8	19	8	0.00159
KIAA1598	0	18	11	0.00013	**ZMIZ1**	2	17	6	0.00075
KIAA1737	1	11	3	0.00028	**ZMIZ2**	0	7	3	0.00031
KLF4	0	26	16	0.00030	**ZNF185**	8	14	8	0.00159
KLHDC4	3	27	13	0.00054	**ZNF266**	6	21	10	0.00116
KRIT1	3	17	10	0.00064	**ZNF407**	5	16	9	0.00070
LACTB	6	15	8	0.00143	**ZNF439**	3	32	16	0.00068
LAPTM5	12	26	22	0.00252	**ZNF549**	2	25	13	0.00062
LARP4	0	17	12	0.00003	**ZNF613**	1	18	8	0.00057
LASS5	2	22	6	0.00055	**ZNF671**	1	15	8	0.00029
LDLRAP1	1	11	5	0.00022	**ZNF75D**	0	8	6	0.00005
LFNG	5	32	17	0.00088	**ZNF787**	0	14	8	0.00019
LILRB2	5	27	10	0.00092	**ZNF839**	1	16	7	0.00010
LINS	7	25	15	0.00128	**ZRANB1**	3	16	9	0.00058
LIPT1	5	21	12	0.00117	**ZYX**	0	12	3	0.00004

REFERENCES

1. The 1000 Genomes Project Consortium, T.G.P., et al., *An integrated map of genetic variation from 1,092 human genomes.* Nature, 2012. **491**(7422): p. 56-65.

2. Kumar, S., et al., *Phylomedicine: an evolutionary telescope to explore and diagnose the universe of disease mutations.* Trends in Genetics, 2011. **27**(9): p. 377-386.

3. Gershon, D., *Microarray technology - An array of opportunities.* Nature, 2002. **416**(6883): p. 885-+.

4. Wang, D.G., et al., *Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome.* Science, 1998. **280**(5366): p. 1077-1082.

5. Metzker, M.L., *Sequencing technologies - the next generation.* Nat Rev Genet, 2010. **11**(1): p. 31-46.

6. Mardis, E.R., *The impact of next-generation sequencing technology on genetics.* Trends Genet, 2008. **24**(3): p. 133-41.

7. Sanger, F., S. Nicklen, and A.R. Coulson, *DNA sequencing with chain-terminating inhibitors.* Proc Natl Acad Sci U S A, 1977. **74**(12): p. 5463-7.

8. Shendure, J. and H. Ji, *Next-generation DNA sequencing.* Nat Biotechnol, 2008. **26**(10): p. 1135-45.

9. Mardis, E.R., *Next-generation DNA sequencing methods.* Annu Rev Genomics Hum Genet, 2008. **9**: p. 387-402.

10. Nagalakshmi, U., et al., *The transcriptional landscape of the yeast genome defined by RNA sequencing.* Science, 2008. **320**(5881): p. 1344-9.

11. Ozsolak, F. and P.M. Milos, *RNA sequencing: advances, challenges and opportunities.* Nat Rev Genet, 2011. **12**(2): p. 87-98.

12. Park, P.J., *ChIP-seq: advantages and challenges of a maturing technology.* Nat Rev Genet, 2009. **10**(10): p. 669-80.

13. Chandrasekharappa, S.C., et al., *Massively parallel sequencing, aCGH, and RNA-Seq technologies provide a comprehensive molecular diagnosis of Fanconi anemia.* Blood, 2013. **121**(22): p. E138-E148.

14. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol, 2009. **10**(3): p. R25.

15. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nat Methods, 2012. **9**(4): p. 357-9.

16. Li, H. and R. Durbin, *Fast and accurate long-read alignment with Burrows-Wheeler transform.* Bioinformatics, 2010. **26**(5): p. 589-95.

17. McKenna, A., et al., *The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data.* Genome Research, 2010. **20**(9): p. 1297-1303.

18. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.

19. Koboldt, D.C., et al., *VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing.* Genome Res, 2012. **22**(3): p. 568-76.

20. Wei, Z., et al., *SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data.* Nucleic Acids Res, 2011. **39**(19): p. e132.

21. Kent, W.J., *BLAT--the BLAST-like alignment tool.* Genome Res, 2002. **12**(4): p. 656-64.

22. Lathrop, G.M., et al., *Strategies for multilocus linkage analysis in humans.* Proc Natl Acad Sci U S A, 1984. **81**(11): p. 3443-6.

23. Inoue, T.K., et al., *Linkage analysis of moyamoya disease on chromosome 6.* J Child Neurol, 2000. **15**(3): p. 179-82.

24. Botstein, D. and N. Risch, *Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease.* Nat Genet, 2003. **33 Suppl**: p. 228-37.

25. The International Hapmap Consortium, T.I.H., et al., *The International HapMap Project.* Nature, 2003. **426**(6968): p. 789-796.

26. Risch, N. and K. Merikangas, *The future of genetic studies of complex human diseases.* Science, 1996. **273**(5281): p. 1516-1517.

27. Klein, R.J., et al., *Complement factor H polymorphism in age-related macular degeneration.* Science, 2005. **308**(5720): p. 385-389.

28. Reich, D.E. and E.S. Lander, *On the allelic spectrum of human disease.* Trends in Genetics, 2001. **17**(9): p. 502-510.

29. Pritchard, J.K. and N.J. Cox, *The allelic architecture of human disease genes: common disease - common variant ... or not?* Human Molecular Genetics, 2002. **11**(20): p. 2417-2423.

30. Maher, B., *Personal genomes: The case of the missing heritability.* Nature, 2008. **456**(7218): p. 18-21.

31. Manolio, T.A., et al., *Finding the missing heritability of complex diseases.* Nature, 2009. **461**(7265): p. 747-753.

32. Walsh, T., et al., *Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia.* Science, 2008. **320**(5875): p. 539-543.

33. Cohen, J.C., et al., *Multiple rare Alleles contribute to low plasma levels of HDL cholesterol.* Science, 2004. **305**(5685): p. 869-872.

34. Nejentsev, S., et al., *Rare Variants of IFIH1, a Gene Implicated in Antiviral Responses, Protect Against Type 1 Diabetes.* Science, 2009. **324**(5925): p. 387-389.

35. McCarthy, M.I., et al., *Genome-wide association studies for complex traits: consensus, uncertainty and challenges.* Nature Reviews Genetics, 2008. **9**(5): p. 356-369.

36. Nogee, L.M., *Abnormal expression of surfactant protein C and lung disease.* American Journal of Respiratory Cell and Molecular Biology, 2002. **26**(6): p. 641-644.

37. Nica, A.C., et al., *Candidate Causal Regulatory Effects by Integration of Expression QTLs with Complex Trait Genetic Associations.* Plos Genetics, 2010. **6**(4).

38. Nicolae, D.L., et al., *Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS.* Plos Genetics, 2010. **6**(4).

39. Tennessen, J.A., et al., *Evolution and functional impact of rare coding variation from deep sequencing of human exomes.* Science, 2012. **337**(6090): p. 64-9.

40. Cruchaga, C., et al., *Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer's disease.* Nature, 2014. **505**(7484): p. 550-4.

41. Lohmueller, K.E., et al., *Whole-exome sequencing of 2,000 Danish individuals and the role of rare coding variants in type 2 diabetes.* Am J Hum Genet, 2013. **93**(6): p. 1072-86.

42. Dickson, S.P., et al., *Rare variants create synthetic genome-wide associations.* PLoS Biol, 2010. **8**(1): p. e1000294.

43. Fellay, J., et al., *ITPA gene variants protect against anaemia in patients treated for chronic hepatitis C.* Nature, 2010. **464**(7287): p. 405-8.

44. Cohen, J.C., et al., *Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma low-density lipoprotein levels.* Proc Natl Acad Sci U S A, 2006. **103**(6): p. 1810-5.

45. Russell, A.I., et al., *Polymorphism at the C-reactive protein locus influences gene expression and predisposes to systemic lupus erythematosus.* Human Molecular Genetics, 2004. **13**(1): p. 137-147.

46. Gaffney, D.J., et al., *Dissecting the regulatory architecture of gene expression QTLs.* Genome Biology, 2012. **13**(1).

47. Neph, S., et al., *An expansive human regulatory lexicon encoded in transcription factor footprints.* Nature, 2012. **489**(7414): p. 83-90.

48. Gerstein, M.B., et al., *Architecture of the human regulatory network derived from ENCODE data.* Nature, 2012. **489**(7414): p. 91-100.

49. Morgenthaler, S. and W.G. Thilly, *A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST).* Mutation Research-Fundamental and Molecular Mechanisms of Mutagenesis, 2007. **615**(1-2): p. 28-56.

50. Li, B.S. and S.M. Leal, *Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data.* American Journal of Human Genetics, 2008. **83**(3): p. 311-321.

51. Madsen, B.E. and S.R. Browning, *A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic.* Plos Genetics, 2009. **5**(2).

52. Neale, B.M., et al., *Testing for an unusual distribution of rare variants.* PLoS Genet, 2011. **7**(3): p. e1001322.

53. Wu, M.C., et al., *Rare-variant association testing for sequencing data with the sequence kernel association test.* Am J Hum Genet, 2011. **89**(1): p. 82-93.

54. Collins, F.S., et al., *A vision for the future of genomics research.* Nature, 2003. **422**(6934): p. 835-47.

55. Bell, J., *Predicting disease using genomics.* Nature, 2004. **429**(6990): p. 453-456.

56. Wray, N.R., M.E. Goddard, and P.M. Visscher, *Prediction of individual genetic risk of complex disease.* Current Opinion in Genetics & Development, 2008. **18**(3): p. 257-263.

57. Khoury, M.J., K. Jones, and S.D. Grosse, *Quantifying the health benefits of genetic tests: The importance of a population perspective.* Genetics in Medicine, 2006. **8**(3): p. 191-195.

58. Allen, H.L., et al., *Hundreds of variants clustered in genomic loci and biological pathways affect human height.* Nature, 2010. **467**(7317): p. 832-838.

59. Stahl, E.A., et al., *Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis.* Nature Genetics, 2012. **44**(5): p. 483-+.

60. Speliotes, E.K., et al., *Genome-wide association analysis identifies variants associated with nonalcoholic fatty liver disease that have distinct effects on metabolic traits.* PLoS Genet, 2011. **7**(3): p. e1001324.

61. Peterson, R.E., et al., *Genetic risk sum score comprised of common polygenic variation is associated with body mass index.* Hum Genet, 2011. **129**(2): p. 221-30.

62. Kang, J., et al., *Improved risk prediction for Crohn's disease with a multi-locus approach.* Hum Mol Genet, 2011. **20**(12): p. 2435-42.

63. Kraft, P. and D.J. Hunter, *Genetic risk prediction--are we there yet?* N Engl J Med, 2009. **360**(17): p. 1701-3.

64. Wilson, P.W., et al., *Prediction of incident diabetes mellitus in middle-aged adults: the Framingham Offspring Study.* Arch Intern Med, 2007. **167**(10): p. 1068-74.

65. D'Agostino, R.B., et al., *General cardiovascular risk profile for use in primary care: The Framingham heart study.* Circulation, 2008. **118**(4): p. E86-E86.

66. Howie, B., J. Marchini, and M. Stephens, *Genotype Imputation with Thousands of Genomes.* G3-Genes Genomes Genetics, 2011. **1**(6): p. 457-469.

67. Ashley, E.A., et al., *Clinical assessment incorporating a personal genome.* Lancet, 2010. **375**(9725): p. 1525-1535.

68. Teslovich, T.M., et al., *Biological, clinical and population relevance of 95 loci for blood lipids.* Nature, 2010. **466**(7307): p. 707-13.

69. Fan, J., S. Upadhye, and A. Worster, *Understanding receiver operating characteristic (ROC) curves.* CJEM, 2006. **8**(1): p. 19-20.

70. Johansen, C.T. and R.A. Hegele, *The complex genetic basis of plasma triglycerides.* Curr Atheroscler Rep, 2012. **14**(3): p. 227-34.

71. Peden, J.F. and M. Farrall, *Thirty-five common variants for coronary artery disease: the fruits of much collaborative labour.* Hum Mol Genet, 2011. **20**(R2): p. R198-205.

72. Gudbjartsson, D.F., et al., *Many sequence variants affecting diversity of adult human height.* Nat Genet, 2008. **40**(5): p. 609-15.

73. Hill, W.G., M.E. Goddard, and P.M. Visscher, *Data and theory point to mainly additive genetic variance for complex traits.* PLoS Genet, 2008. **4**(2): p. e1000008.

74. Aschard, H., et al., *Inclusion of gene-gene and gene-environment interactions unlikely to dramatically improve risk prediction for complex diseases.* Am J Hum Genet, 2012. **90**(6): p. 962-72.

75. Marigorta, U.M. and G. Gibson, *A simulation study of gene-by-environment interactions in GWAS implies ample hidden effects.* Front Genet, 2014. **5**: p. 225.

76. Manolio, T.A., et al., *Finding the missing heritability of complex diseases.* Nature, 2009. **461**(7265): p. 747-53.

77. Yang, J., et al., *Common SNPs explain a large proportion of the heritability for human height.* Nat Genet, 2010. **42**(7): p. 565-9.

78. Makowsky, R., et al., *Beyond missing heritability: prediction of complex traits.* PLoS Genet, 2011. **7**(4): p. e1002051.

79. Xu, N., X. Li, and Y. Zhong, *Inflammatory cytokines: potential biomarkers of immunologic dysfunction in autism spectrum disorders.* Mediators Inflamm, 2015. **2015**: p. 531518.

80. Yudkin, J.S., et al., *C-reactive protein in healthy subjects: associations with obesity, insulin resistance, and endothelial dysfunction: a potential role for cytokines originating from adipose tissue?* Arterioscler Thromb Vasc Biol, 1999. **19**(4): p. 972-8.

81. Visser, M., et al., *Elevated C-reactive protein levels in overweight and obese adults.* JAMA, 1999. **282**(22): p. 2131-5.

82. Rawson, E.S., et al., *Body mass index, but not physical activity, is associated with C-reactive protein.* Medicine and Science in Sports and Exercise, 2003. **35**(7): p. 1160-1166.

83. Smith, G.D. and S. Ebrahim, *Mendelian randomization: prospects, potentials, and limitations.* International Journal of Epidemiology, 2004. **33**(1): p. 30-42.

84. Purcell, S., et al., *PLINK: A tool set for whole-genome association and population-based linkage analyses.* American Journal of Human Genetics, 2007. **81**(3): p. 559-575.

85. Delaneau, O., J. Marchini, and J.F. Zagury, *A linear complexity phasing method for thousands of genomes.* Nature Methods, 2012. **9**(2): p. 179-181.

86. Howie, B.N., P. Donnelly, and J. Marchini, *A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies.* Plos Genetics, 2009. **5**(6).

87. Genomes Project, C., et al., *A map of human genome variation from population-scale sequencing.* Nature, 2010. **467**(7319): p. 1061-73.

88. Marchini, J., et al., *A new multipoint method for genome-wide association studies by imputation of genotypes.* Nature Genetics, 2007. **39**(7): p. 906-913.

89. Chung, E.S., et al., *Randomized, double-blind, placebo-controlled, pilot trial of infliximab, a chimeric monoclonal antibody to tumor necrosis factor-alpha, in patients with moderate-to-severe heart failure: results of the anti-TNF Therapy Against Congestive Heart Failure (ATTACH) trial.* Circulation, 2003. **107**(25): p. 3133-40.

90. Fisher, C.J., Jr., et al., *Influence of an anti-tumor necrosis factor monoclonal antibody on cytokine levels in patients with sepsis. The CB0006 Sepsis Syndrome Study Group.* Crit Care Med, 1993. **21**(3): p. 318-27.

91. Feldmann, M., *Development of anti-TNF therapy for rheumatoid arthritis.* Nat Rev Immunol, 2002. **2**(5): p. 364-71.

92. Ridker, P.M., et al., *C-reactive protein and other markers of inflammation in the prediction of cardiovascular disease in women.* N Engl J Med, 2000. **342**(12): p. 836-43.

93. Griselli, M., et al., *C-reactive protein and complement are important mediators of tissue damage in acute myocardial infarction.* J Exp Med, 1999. **190**(12): p. 1733-40.

94. Romero-Corral, A., et al., *Association of bodyweight with total mortality and with cardiovascular events in coronary artery disease: a systematic review of cohort studies.* Lancet, 2006. **368**(9536): p. 666-78.

95. Wang, Y., et al., *Comparison of abdominal adiposity and overall obesity in predicting risk of type 2 diabetes among men.* Am J Clin Nutr, 2005. **81**(3): p. 555-63.

96. Visscher, P.M., et al., *Five years of GWAS discovery.* Am J Hum Genet, 2012. **90**(1): p. 7-24.

97. Yang, J., et al., *Genome partitioning of genetic variation for complex traits using common SNPs.* Nat Genet, 2011. **43**(6): p. 519-25.

98. Dunham, I., et al., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.

99. Maurano, M.T., et al., *Systematic Localization of Common Disease-Associated Variation in Regulatory DNA.* Science, 2012. **337**(6099): p. 1190-1195.

100. Schaub, M.A., et al., *Linking disease associations with regulatory information in the human genome.* Genome Research, 2012. **22**(9): p. 1748-1759.

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

102. Petrovski, S., et al., *Genic intolerance to functional variation and the interpretation of personal genomes.* PLoS Genet, 2013. **9**(8): p. e1003709.

103. Voight, B.F., et al., *The Metabochip, a Custom Genotyping Array for Genetic Studies of Metabolic, Cardiovascular, and Anthropometric Traits.* Plos Genetics, 2012. **8**(8).

104. Cortes, A. and M.A. Brown, *Promise and pitfalls of the Immunochip.* Arthritis Research & Therapy, 2011. **13**(1).

105. Li, R., et al., *SNP detection for massively parallel whole-genome resequencing.* Genome Res, 2009. **19**(6): p. 1124-32.

106. Challis, D., et al., *An integrative variant analysis suite for whole exome next-generation sequencing data.* BMC Bioinformatics, 2012. **13**: p. 8.

107. Keinan, A. and A.G. Clark, *Recent explosive human population growth has resulted in an excess of rare genetic variants.* Science, 2012. **336**(6082): p. 740-3.

108. Genomes Project, C., et al., *An integrated map of genetic variation from 1,092 human genomes.* Nature, 2012. **491**(7422): p. 56-65.

109. Danecek, P., et al., *The variant call format and VCFtools.* Bioinformatics, 2011. **27**(15): p. 2156-8.

110. Altshuler, D.M., et al., *Integrating common and rare genetic variation in diverse human populations.* Nature, 2010. **467**(7311): p. 52-58.

111. O'Rawe, J., et al., *Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing.* Genome Med, 2013. **5**(3): p. 28.

112. Clement, N.L., et al., *The GNUMAP algorithm: unbiased probabilistic mapping of oligonucleotides from next-generation sequencing.* Bioinformatics, 2010. **26**(1): p. 38-45.

113. Campbell, M.C. and S.A. Tishkoff, *African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping.* Annu Rev Genomics Hum Genet, 2008. **9**: p. 403-33.

114. Amos, W. and J.I. Hoffman, *Evidence that two main bottleneck events shaped modern human genetic diversity.* Proc. R. Soc. B, 2010(277): p. 6.

115. Nuytemans, K., et al., *Whole exome sequencing of rare variants in EIF4G1 and VPS35 in Parkinson disease.* Neurology, 2013. **80**(11): p. 982-989.

116. Purcell, S.M., et al., *A polygenic burden of rare disruptive mutations in schizophrenia.* Nature, 2014. **506**(7487): p. 185-190.

117. Sanders, S.J., et al., *De novo mutations revealed by whole-exome sequencing are strongly associated with autism.* Nature, 2012. **485**(7397): p. 237-U124.

118. Tennessen, J.A., et al., *Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes.* Science, 2012. **337**(6090): p. 64-69.

119. Montgomery, S.B. and E.T. Dermitzakis, *The resolution of the genetics of gene expression.* Hum Mol Genet, 2009. **18**(R2): p. R211-5.

120. Grundberg, E., et al., *Mapping cis- and trans-regulatory effects across multiple tissues in twins.* Nat Genet, 2012. **44**(10): p. 1084-9.

121. McCarthy, M.I., et al., *Genome-wide association studies for complex traits: consensus, uncertainty and challenges.* Nat Rev Genet, 2008. **9**(5): p. 356-69.

122. Zondervan, K.T. and L.R. Cardon, *The complex interplay among factors that influence allelic association (vol 5, pg 89, 2004).* Nature Reviews Genetics, 2004. **5**(3): p. 238-238.

123. Mecham, B.H., P.S. Nelson, and J.D. Storey, *Supervised normalization of microarrays.* Bioinformatics, 2010. **26**(10): p. 1308-1315.

124. Preininger, M., et al., *Blood-Informative Transcripts Define Nine Common Axes of Peripheral Blood Gene Expression.* Plos Genetics, 2013. **9**(3).

125. Irizarry, R.A., et al., *Exploration, normalization, and summaries of high density oligonucleotide array probe level data.* Biostatistics, 2003. **4**(2): p. 249-264.

126. Lozzio, B.B. and C.B. Lozzio, *Properties of the K562 cell line derived from a patient with chronic myeloid leukemia.* Int J Cancer, 1977. **19**(1): p. 136.

127. Kim, H.J., et al., *Targeted genome editing in human cells with zinc finger nucleases constructed via modular assembly.* Genome Res, 2009. **19**(7): p. 1279-88.

128. Kiss, M.M., et al., *High-throughput quantitative polymerase chain reaction in picoliter droplets.* Anal Chem, 2008. **80**(23): p. 8975-81.

129. Eisen, M.B., et al., *Cluster analysis and display of genome-wide expression patterns.* Proc Natl Acad Sci U S A, 1998. **95**(25): p. 14863-8.

130. McHugh, R.S., et al., *CD4(+)CD25(+) immunoregulatory T cells: gene expression analysis reveals a functional role for the glucocorticoid-induced TNF receptor.* Immunity, 2002. **16**(2): p. 311-23.

131. Cookson, W., et al., *Mapping complex disease traits with global gene expression.* Nat Rev Genet, 2009. **10**(3): p. 184-94.

132. Stranger, B.E., et al., *Population genomics of human gene expression.* Nat Genet, 2007. **39**(10): p. 1217-24.

133. Dimas, A.S., et al., *Common regulatory variation impacts gene expression in a cell type-dependent manner.* Science, 2009. **325**(5945): p. 1246-50.

134. Spielman, R.S., et al., *Common genetic variants account for differences in gene expression among ethnic groups.* Nat Genet, 2007. **39**(2): p. 226-31.

135. Veyrieras, J.B., et al., *High-resolution mapping of expression-QTLs yields insight into human gene regulation.* PLoS Genet, 2008. **4**(10): p. e1000214.

136. Crawford, G.E., et al., *Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS).* Genome Res, 2006. **16**(1): p. 123-31.

137. Boyle, A.P., et al., *High-resolution mapping and characterization of open chromatin across the genome.* Cell, 2008. **132**(2): p. 311-22.

138. Deciphering Developmental Disorders, S., *Large-scale discovery of novel genetic causes of developmental disorders.* Nature, 2015. **519**(7542): p. 223-8.

139. Spencer, C.C., et al., *Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip.* PLoS Genet, 2009. **5**(5): p. e1000477.

140. Whitney, A.R., et al., *Individuality and variation in gene expression patterns in human blood.* Proc Natl Acad Sci U S A, 2003. **100**(4): p. 1896-901.

141. Hamburg, M.A. and F.S. Collins, *The path to personalized medicine.* N Engl J Med, 2010. **363**(4): p. 301-4.