Leveraging genetic association data to investigate the polygenic architecture of human traits and diseases

Leveraging genetic association data to investigate the polygenic architecture of human traits and diseases

A dissertation presented

by

Ying Leong Chan

to

The Division of Medical Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Genetics and Genomics

Harvard University

Cambridge, Massachusetts

April 2014

Dissertation Advisor: Professor Joel N. Hirschhorn Ying Leong Chan

Leveraging genetic association data to investigate the polygenic architecture of human traits and diseases

ABSTRACT

Many human traits and diseases have a polygenic architecture, where phenotype is partially determined by variation in many genes. These complex traits or diseases can be highly heritable and genome-wide association studies (GWAS) have been relatively successful in the identification of associated variants. However, these variants typically do not account for most of the heritability and thus, the genetic architecture remains uncertain.

This dissertation describes analytical approaches to look for evidence of models of genetic architecture that could explain the remaining heritability. We develop methods to make predictions under various models, and compare the expected results from these predictions against the observed data for several traits and diseases. First, in studies of height (a classical polygenic trait), we modeled the expected cumulative effect of common variants identified from GWAS and compared the model with empirical data in individuals from the tails of the height distribution. We found that these common variants are predictive of stature, but have less than expected effects specifically at the short end of the height distribution. This result is consistent with models where rare variants with moderate effect, influence stature only in the shortest individuals. Second, we showed that under genetic models where low frequency variants make

polygenic contributions to disease, there will be an excess of low frequency risk-increasing variants detected in GWAS. As such, by comparing the number of detected risk-increasing to risk-decreasing variants, one can detect a signal of the contribution to polygenic inheritance from low frequency variants. Finally, we examine the genetic architecture of sitting height ratio (SHR), a measure of body proportion that varies dramatically between individuals of African and European ancestry. We find that the SHR difference between populations is largely due to polygenic architecture; there is no evidence for any major locus accounting for most of this difference.

These results show that, with the appropriate computational and genetic models, one can use empirical results of genetics studies to make inferences regarding genetic architecture of human traits and diseases. Doing so can help investigators prioritize strategies for uncovering the remaining unexplained heritability.

TABLE OF CONTENTS

DEDICATION

I dedicate this thesis to my loving wife, Teng Ting (Elaine) Lim. You are the person that has always been there during difficult and trying times. We share and do almost everything together. I would not be the person I am today without your love and support. Elaine, I dedicate this thesis to you.

ACKNOWLEGEMENTS

When I first arrived on the Harvard Medical School campus, I was captivated by the breath and majesty of just being there. Besides the Victorian building design of the Quad buildings, the place was surrounded by many hospitals and people plus that the sound of sirens wailing from ambulances made the whole area a really busy place. It was recruitment weekend for new students and part of the program was to have a couple of faculty members give talks to the potential incoming students and there was where I met my eventual dissertation advisor, Joel Hirschhorn. It was March of 2009.

I rotated in the Hirschhorn lab for the summer of that year and eventually joined the lab as a student the next year. Joel was extremely helpful and encouraging mentor throughout graduate school. We (members of the lab) meet regularly with him, at least once in two weeks (30 to 60 minutes) despite his extremely busy schedule and we will always get his full undivided attention during each session. Also, during lab meetings, he will always be there to give critical feedback and suggestions when you present your work and ideas, whether it is about possible experiments to perform or just feedback on giving the presentation itself. Furthermore, he sometimes performs his own analysis, contrary to the long held belief that "PIs don't do experiments themselves". To me, it is a privilege to have the opportunity to be his student as well as a member of his lab.

I remember when I first joined the lab, I was given the task of "coming up with 10 ideas" by Joel. I only did 8 and after long discussions about each of the aims, 1 of them eventually became one of my thesis aims with the help of another member of the lab, Andrew Dauber. My time in graduate school would not be the same without Andrew. Andrew started as a fellow in the lab the same time when I started my rotation so he has been in the lab for about a year when I

eventually joined. It was during one of the lab meetings that Andrew presented some genotyping data on height extremes (very short and tall individuals) that could answer one of the 8 ideas that I had initially. Andrew was very kind and helpful and we worked together to answer the question. I will always be grateful to Andrew for getting me started as well as his mentorship throughout my time in the lab.

Rany Salem, a post doctoral fellow in the lab, is someone I would also like to specially mention. Rany started as a post doctoral fellow the same day I started my rotation in the lab. Although, he is based at the Broad Institute, he comes to Children's (Boston Children's Hospital) frequently and we would always get coffee and exchange ideas. His work on diabetic nephropathy allowed me to develop my next idea, using the summary statistics generated from that project. Besides that, his efforts to obtain genotype data from many different cohorts allowed me and others to explore other ideas with regards to complex traits.

In general, members of the Hirschhorn lab are a very sociable bunch which is odd, considering that most of us are 'computational' people that do not have a reputation for being sociable. To illustrate this, a former student of the lab, who is now a post doctoral fellow, Charleston Chiang, frequently organizes "games-night" at his place (about once a month) where he invites fellow members of the lab to hang out and play board games. One of our favorite games is called "Betrayal", which is a game about a group of adventurers exploring a haunted house and one of the members will become the "traitor" midway in the game. That was fun and I will always remember those good times. Another example worth mentioning is our regular sashimi buffets. Somehow, there is a sizeable number of people in our lab that just love gorging on raw-fish, me included. We started out at a reasonably priced restaurant called Yamato somewhere in Allston where they have an all-you-can-eat buffet for just thirty dollars. However,

when Tonu Esko joined us midway during my time in the lab, he “discovered” a new place, called Takusan where it is cheaper and serves oysters as well. Takusan is now our regular hangout until we find a new place.

Therefore, I would like to take the opportunity to thank other members of the lab, past and present for creating the wonderful lab environment that is conducive for sharing ideas and establishing collaboration. Thanks to Tune Pers for providing opportunity to work together on DEPICT. Thanks to Sophie Wang for the help in performing the forward simulations for our 2nd project as well as for free ice-cream. Thanks to Sailaja Vedamtam for pointing me to necessary files when I need them as well as for the wonderful vegetarian dinner that you make. Thanks to Tonu Esko for having the opportunity to work with the Estonian data as well as introducing Takusan Sushi. Thanks to Michael Guo for coffee in exchange for performing LD calculations. Thanks to Yan Meng for discussions about finance and investments. Thanks to Jon Swartz for wonderful story about seaweed soup. Thanks to Vidhu Thaker for sharing her knowledge of Obesity and New York. Thanks to Jennifer Moon for lunchtime discussions about biological techniques as well as all things Korean. Thanks to Yu-Han Hsu for accompanying us when we have our regular coffee breaks and thanks to Frances Lopez for suggesting Waterville valley for a summer getaway. Finally, I would also like to say a big thank you to Meghan Foster for your patience in scheduling the weekly meetings with Joel despite his very busy schedule.

I would also thank members of my dissertation advisory committee: Jonathon Seidman, Matthew Warman and Mark Daly for the thoughtful advice and generous comments during graduate school. I would also like to thank my dissertation examiners: Souyma Raychaudhuri, David Page and Shaun Purcell for agreeing to take time off from your busy schedule to be my examiners as well as for going through this dissertation.

A huge thank you to fellow graduate students for the enjoyable time and company spent during graduate school. I really enjoyed the regular games night, playing “Singaporean bridge” and Mario party. Donkey Kong (Rigel) would like to thank Mario (Elaine), Luigi (Laura) and Yoshi (Palak) for the company.

I would also like to thank my life-long mentors, Guna Rajagopal and Arnold Levine for getting me started in my scientific career and providing me with adequate opportunities and training prior to graduate school.

Also, I would like to thank my wife, Elaine Lim, for the support and guidance. Life in America can be lonely as our families are back in Singapore. We could not have what we have now without you being here with me.

Finally, I would like to thank my family for their support in my graduate career, especially for the support for us traveling more than 10,000 miles away from home for graduate school. To my parents, Ngai Kong and Hwee Koon, older brother, Ying Soon, sister-in-law, Grace, thank you so much.

ATTRIBUTIONS

Chapter 2

Yingleong Chan: Together with J.N.H and A. D., conceived and performed the single SNP ORs comparisons for FINRISK. Conceived and performed WAS analysis for FINRISK, WAS simulations for HUNT, FINRISK and modeling of additional variants for HUNT, FINRISK. Wrote the initial draft of the manuscript and together with O.L.H., A.D., T.M.F., J.N.H and M.N.W edited the later drafts.

Oddgeir L Holmen: Conceived and performed the single SNP ORs comparisons for HUNT. Edited parts of the manuscript and provided helpful comments.

Andrew Dauber: Conceived and performed the single SNP ORs comparisons for FINRISK. Directed the genotyping of the height associated SNPs for the FINRISK samples. Substantially edited parts of the manuscript and provided many helpful comments.

Lars Vatten: Provided data for the HUNT cohort.

Aki S Havulinna: Provided data for the FINRISK cohort.

Frank Skorpen: Provided data for the HUNT cohort.

Kirsti Kvaløy: Provided data for the HUNT cohort.

Kaisa Silander: Provided data for the FINRISK cohort.

Thutrang T Nguyen: Perform the genotyping of the height associated SNPs for the FINRISK samples.

Cristen Willer: Provided data for the HUNT cohort.

Michael Boehnke: Provided data for the HUNT cohort. Suggested edits to manuscript.

Markus Perola: Provided data for the FINRISK cohort.

Aarno Palotie: Provided data for the FINRISK cohort.

Veikko Salomaa: Provided data for the FINRISK cohort.

Kristian Hveem: Provided data for the HUNT cohort.

Timothy M Frayling: Conceived of single SNP ORs comparisons for HUNT. Edited parts of the manuscript and provided helpful comments and direction.

Joel N Hirschhorn: Conceived of single SNP ORs comparisons for FINRISK. Calculated the expected single SNP ORs. Edited most of the manuscript and provided a lot of helpful comments and direction.

Michael N Weedon: Conceived of single SNP ORs comparisons for HUNT. Calculated the expected single SNP ORs. Calculated the combined HUNT and FINRISK single SNP ORs. Calculated the WAS for HUNT. Performed sibling analysis for HUNT. Wrote the initial parts of the manuscript, edited most of the manuscript and provided helpful comments and direction.

Chapter 3

Yingleong Chan: Conceived and performed R/P ratio analysis. Performed power calculations with varying parameters. Performed the calculations and simulations to obtain phenotypes for the selection model. Performed the R/P ratio simulations for uneven sample sizes and population stratification. Assessed the R/P ratio on published GWAS results. Wrote the manuscript and together with E.T.L and J.N.H edited the later drafts with revisions.

Elaine T Lim: Conceived and directed the acquisition of GWAS summary statistics of

Schizophrenia, Bipolar and Major depressive disorder, Crohn's disease and ulcerative colitis for R/P ratio analysis. Edited the manuscript and provided helpful suggestions.

Niina Sandholm: Performed GWAS for cohorts relevant to diabetic nephropathy.

Sophie R Wang: Performed the forward simulation to obtain allele frequencies and effect sizes.

Amy Jayne McKnight: Performed GWAS for cohorts relevant to diabetic nephropathy.

Stephan Ripke: Performed GWAS for cohorts relevant to inflammatory bowel disease.

DIAGRAM Consortium: Provided GWAS summary statistic for type 2 diabetes.

GENIE Consortium: Provided GWAS summary statistics for diabetic nephropathy.

GIANT Consortium: Provided GWAS summary statistics for obesity.

IIBDGC Consortium: Provided GWAS summary statistics for inflammatory bowel disease.

PGC Consortium: Provided GWAS summary statistics for Schizophrenia, Bipolar and Major depressive disorder.

Mark J Daly: Provided suggestions on power calculations.

Benjamin M Neale: Provided suggestions on power calculations.

Rany M Salem: Performed GWAS for cohorts relevant to diabetic nephropathy.

Joel N Hirschhorn: Provided guidance and suggestions on R/P ratio analysis. Conceived the negative selection analysis. Conceived the population stratification analysis. Provided much of the NCP ratio proof. Heavily edited the manuscript. Provided extensive feedback for revisions.

Chapter 4

Yingleong Chan: Conceived and performed sitting height ratio (SHR) comparisons between European and African Americans. Performed principal component analysis for determining admixture percentages. Performed association of admixture percentages with SHR. Performed GWAS on African and European cohorts with SHR. Performed comparisons with height associated variants. Wrote the manuscript.

Rany M Salem: Downloaded the data from dbGAP, set up the pipeline for quality control on the data.

Joel N Hirschhorn: Conceived the idea of studying sitting height as a phenotype. Suggested various analysis of studying sitting height. Provided extensive feedback on manuscript.

Chapter 1

Introduction

A PREAMBLE

Mendelian inheritance

It has long been recognized that physical traits are more likely to be shared by parents and their offspring, between siblings or close relatives as well as between individuals of similar ethnic ancestry [1]. Such a phenomenon is known as heritability and the modern explanation of heritability was first broadly described by Gregor Mendel more than two centuries ago, where he showed that some traits of pea plants follow a specific pattern of inheritance [2]. Mendel theorized that each individual possesses a pair of alleles for each trait and will randomly pass on one of the alleles to its offspring. The offspring would then inherit two alleles, one from its father and one from its mother and the pair of alleles would determine the trait of the offspring. Such a pattern is now popularly known as Mendelian inheritance.

Polygenic inheritance

In the beginning of the 20th century, there were anthropologists and biologists who argued that since Mendelian inheritance predicts that traits would be discrete in nature, it cannot account for the number of continuous or quantitative traits (e.g. height) observed in humans and thus the theory cannot be applied to humans. However, in 1918, R. A. Fisher demonstrated that if there were multiple allele pairs, that each pair is responsible for only a fraction of the trait and each of these pairs observed the same pattern of Mendelian inheritance, it could account for most of the continuous or quantitative traits observed in humans [3]. This proposed model of Fisher is what we now call polygenic inheritance.

Disease mapping

Today, we know that the source of heritability is largely from within the variants contained

in the DNA of genes of our diploid chromosomes although there is some indication that epigenetics, molecular factors that attach to DNA, can have a role as well [4]. With the invention of methods like molecular cloning and the subsequent typing of genetic markers, it became possible to map heritable diseases to their respective genetic locus. Doing so allowed researchers to pinpoint the exact genetic variants that are responsible for causing the disease. Studying the genes underlying these variants can potentially inform us about the disease etiology and thereby be informative for developing therapeutics. Therefore to map a disease to a genetic locus, one must be able to determine if a genetic marker is associated with disease status.

Linkage analysis

There many types of genetic markers that can be used for this purpose. One of the earliest markers that were used for this purpose were microsatellite markers or short tandem repeats (STRs) [5]. These STR alleles can be genotyped in a variety of ways, from performing gel-electrophoresis to parallel sequencing [6]. However, more recently, since the completion of the human genome project [7] and the international hapmap project [8], single nucleotide polymorphisms (SNPs) have become the dominant marker of choice as it is more abundant and covers more of the human genome than any of the other known markers [9]. Having determined the marker, determining if the genetic marker is associated with disease status is the next problem. One of the first methodology used for determining this is linkage analysis [10,11]. Linkage analysis is a process by which researcher use genetic markers to determine if disease status co-segregates with any of these markers more so than by random chance by studying the inheritance pattern of these markers in families that have the trait or disease. The degree of co-segregation is measured by the LOD (logarithm of odds) score and a LOD score of 3.0 or greater is usually taken as evidence that the genetic locus represented by the marker harbors the variant

that causes the disease. This approach has been very successful at identifying Mendelian diseased genes [12] but fall short when trying to identify genes for complex disease [13].

Linkage analysis not amenable for complex diseases

There are a number of reasons why linkage analysis is not amenable to identifying genes associated with complex diseases. First, complex diseases are thought to be genetically influenced by multiple genes rather than a single gene. This could mean that an affected individual could be genetically predisposed to having the disease because of variants from many genes, each of which causes a small increase to the risk of obtaining the disease (polygenic inheritance). This could also mean that while for each family, only mutations in a single gene is responsible, that gene is different for different families (locus heterogeneity). For example, an autosomal recessive disease like Fanconi Anemia has about 16 different genes [14]. If the number of genes were to be much more, for example 160 instead of 16, then there would be a good chance that every family analyzed for the disease will have a different causal gene and thus no overlapping genes. Whichever the case maybe, be it polygenic inheritance of locus heterogeneity, linkage analysis will be less powered for complex diseases as the genetic basis for each affected child within each family or across families is different.

Genetic Association Studies

Polygenic inheritance is a defining feature of most complex traits and one of the major reasons why linkage analysis in family pedigrees is not amenable to identifying genes responsible for complex traits. The problem is further compounded by the fact that many complex traits are influence by non-genetic (environmental) factors as well. To solve this problem, researchers suggested that genetic association studies rather than linkage analysis would be more effective in identifying the responsible genetic loci under the assumption of

polygenic inheritance. Instead of examining chromosome markers that co-segregate with disease status in family pedigrees, genetic association studies examine the frequency of the allele in a large population cohort to determine if the allele frequency is correlated with the trait or disease status. Indeed, researchers have shown that for studies with the same sample sizes, genetic association studies significantly outperforms linkage analysis under the assumption of polygenic inheritance. The process of performing genetic association studies have now evolved into a process known as Genome wide association studies (GWAS), where markers on the entire genome are systematically tested at the appropriate threshold of significance such that the significant results are robust and reproducible [15]. To date, there are many successful GWAS that are published highlighting the overall success of GWAS as a methodology for identifying genetic loci associated with complex traits or diseases.

Missing heritability

Although genome wide association studies (GWAS) have been largely successful, the variants identified typically do not explain most of the trait's heritability. This result is known as the missing heritability problem and there are suggested hypotheses to explain the missing heritability [16]. One such hypothesis is that a substantial fraction of the heritability of the disease or trait is due to rare genetic variants [17]. As these variants are rare in the population, they are not well assayed by many of the genotyping arrays available nor are they amenable to imputation [18]. Another hypothesis is that there are more common variants with even smaller effect sizes and these studies are not well powered to detect these variants. A solution to answer this question would be to perform whole-genome sequencing instead of using genotyping arrays on even more number of samples although performing such an experiment can be costly as whole-genome sequencing is still significantly more expensive than genotyping arrays.

Therefore, perhaps it would be useful to determine if the exercise of performing sequencing and/or studying more samples to answer this question would be fruitful from the results from existing GWAS. In this dissertation, I present various methods to infer from GWAS results the genetic landscape that could explain the remaining trait heritability. Apart from performing GWAS, I described two different and independent approaches for making this inference without the need for performing additional whole-genome or whole-exome sequencing.

Approaches to examine genetic architecture

The first approach is one that explores the possibility of rare genetic variants contributing to the trait by examining the effect of the variants identified through GWAS on individuals at the tails of the distribution (Chapter 2). Using human height as our model phenotype, we showed that common variants identified through GWAS at the short end of the distribution are less predictive than expected. This result can be explained by the presence of rare genetic variants contributing to short stature. The second approach is one that explores the summary statistics obtained from GWAS (Chapter 3). By examining the direction of effect (odds-ratios or effect sizes), an excess of risk-increasing variants compared to risk-decreasing one can be indicative of polygenic inheritance from low-frequency or rare genetic variants, especially for dichotomous traits or diseases. In the subsequent chapter (Chapter 4), I will describe our study to determine genetic variants that can explain the heritable complex trait of body proportion using sitting-height ratio (SHR) as the phenotype. SHR is thought to be heritable and the SHR of European Americans is known to be significantly larger than African Americans. I will provide evidence that this difference in SHR is largely genetically driven as well as polygenic. Finally, I will conclude with a summary of the findings presented and discuss the potential implications and possible future research stemming from the discoveries described in this dissertation.

THE HERITABILITY OF COMPLEX TRAITS

A description of heritability

Complex traits or diseases are broadly defined as phenotypes that do not follow a Mendelian pattern of inheritance. Such traits are usually relatively common, i.e. at least 1% of the population have the trait or disease in contrast to Mendelian disorders, which are usually much rarer [19]. A main question in human biology is whether the expression of a trait of interest is due to genetic factors, environmental factors or just a product of stochasticity. To measure the contribution of genetic factors to the trait, one can measure the heritability. Heritability is a measurement of how much genetics play a role of explaining the difference of the trait between individuals of a population [20]. It can be loosely described as how much of the trait that you have is due to you inheriting it from your parents. It is also a technical term, defined as the ratio of variances, specifically the proportion of total variance in a population for a trait that is attributable to genetic variation [20]. This distinction of its varied use in literature is sometimes not made which can be a source of confusion [21]. Heritability can also be divided into 2 categories, the first being broad-sense heritability and the second being narrow-sense heritability. Broad-sense heritability (H^2) describes the attribution of total genetic variation to the trait's variability while narrow-sense heritability (h^2) describes the attribution of only additive genetic variation to the trait's variability.

Methods for estimating heritability

As heritability is not a physical trait that can be directly measured, one can only use various methods to provide an estimate. One of the first methods would be to determine if average phenotypic value of the parents (mid-parental phenotype) is correlated with the offspring's phenotypic value. This method was first used by Francis Galton over a 100 years ago

to show that human height is heritable [22]. The correlation can be measured by linear regression and studies have put the estimate of height as high as 80% ($h^2 \sim 0.8$) [23]. This method can also be adapted to use correlation estimates of full-siblings instead of parent-offspring although some other adjustments are required. Another popular way of measuring heritability would be the use of Falconer's formula in twin studies [24]. Given that dizygotic (DZ) twins on average are only 50% identical by descent (IBD) while monozygotic (MZ) twins are 100% IBD, MZ twins are therefore expected to be two times more similar than DZ twins. As DZ twins are approximately 50% IBD, heritability can be estimated by taking twice the difference of the phenotypic correlation between MZ twins and DZ twins. More recently, with the introduction of whole-genome genotyping arrays, heritability can be now be estimated by taking the correlation of phenotypic values with IBD estimates from full siblings [25] as well as using the correlation of all common SNPs in predicting the phenotype [26]. Heritability is not necessary constant over time. Heritability can decrease with increased environmental variability. It has been suggested that heritability for morphological traits will decrease in poorer environmental conditions [27], e.g. nutrient poor environment. This fits the theory that in a poor environment, competition for resources will cause increased environmental variability that will influence the outcome of the trait. Nonetheless, heritability estimates provide us with a way to determine which traits are mainly genetically influenced and which traits are mainly environmentally influenced.

Heritability and genetic architecture

It is known that it is not a single gene but a multitude of genes that are responsible for complex traits or diseases. We also find that most of these complex traits or diseases, their occurrences are not as rare as most of the Mendelian diseases with prevalence rate very much greater than 1 in 1000 individuals. For example, a study of the incidence of Schizophrenia

reported an average lifetime morbid risk for schizophrenia to be 7.2 per 1000 persons [28]. A study of prevalence of type 2 diabetes in adolescents put the prevalence as high as 110 per 1000 persons (11%) [29]. Given that the disease can be common in the population, we can ask if the genetic variants that are responsible for the disease are common or rare in the population. Asking this question would illustrate 2 concepts. The first is what is known as the "common disease, common variant hypothesis". In this scenario, it is thought that the genetic variants that give rise to risk of disease is relatively common but that each variant's contribution to disease risk is small. This means that the effect size per allele is small, that is the effect size usually less than 0.1 standard deviations or has an odds-ratio less than 1.1. In such a mode of inheritance, also known as polygenic inheritance, the genetic cause of the disease per individual or family is due to all the risk variants collectively. The next concept is what is known as the "common disease, rare variant hypothesis". In this case, the genetic variants that give rise to the risk of disease are very rare and each variant's contribution to disease risk is large. The effect size per allele can be large, perhaps more than 0.5 standard deviations or an odds-ratio greater than 1.6. For this mode of inheritance, also called locus heterogeneity, the genetic cause of the disease per individual or family is due largely to only 1 gene and other individuals or family with the disease have other genes responsible for their disease. Although these "hypotheses" are seemingly different, they do not have to be mutually exclusive. Effectively, these "hypotheses" can be unified by addressing the effect sizes and variant frequencies for the spectrum of genetic variants that give rise to the disease. For such traits, the variant cannot be common and have a large effect. If that is true, the trait or disease would be monogenic and would be classified as a Mendelian disorder. As such, it is not inconceivable that a disease could have both rare large effect alleles as well as common small effect alleles. For example, even when GWAS show that most variants that are associated

with height have small effects [30], there are very rare alleles that can give rise to short stature, e.g. Achondroplasia [31] as well as rare alleles that give rise to tall stature, e.g. Marfan syndrome [32]. The same can be said for many other complex traits or disease and it is important to be aware of the genetic architecture giving rise to the trait or disease.

Heritability and polygenic inheritance

To explain the heritability of non-Mendelian complex traits and diseases, the pattern of inheritance is usually assumed to be polygenic, i.e. many variants across multiple genes each contribute a small fraction of the heritability. Examples of complex traits include asthma, schizophrenia, type 2 diabetes, inflammatory bowel disease and coronary heart disease. These traits are highly heritable [33–37] even though they do not follow a Mendelian pattern of inheritance. In type 2 diabetes, the first notable gene with variation conferring risk to the disease was TCF7L2 [38]. While not completely penetrant, individuals having a single copy of the risk allele are 1.45 times more likely to get type 2 diabetes than individuals without the risk allele. Since then, studies with much larger sample sizes have yielded about 30 distinct loci that are associated with the risk of getting type 2 diabetes [39]. A similar situation exists for schizophrenia, where prior to having sufficiently large sample sizes, no single locus or gene was determined to be significantly associated with schizophrenia [40]. However, in one of the earlier studies of schizophrenia with just over 3,000 cases and 3,000 controls, the authors reported a significant signal of polygenic inheritance from common variants [41]. In that study, the authors used the common variants that were marginally associated in their samples to model a "polygenic score", a score that represents the overall cumulative predictability of these common variants to schizophrenia risk. They found that the polygenic score is significantly predictive of schizophrenia in an independent cohort of individuals. This suggest that there are many, perhaps

thousands of variants that modulates the risk of acquiring schizophrenia, each of which have only a very small effect on the overall risk and they are not discovered to be significant because the study is simply underpowered and further studies with many more samples would be necessary. Indeed, when larger sample sizes were available for association, we begin to see significant loci emerge [42]. Other complex traits have similar stories where multiple loci have been discovered, each of which confers only a fraction of the total risk.

Quantitative traits and polygenic inheritance

Quantitative traits that are approximately normally distributed in a population are usually complex traits as well. If such a trait is heritable, then it is unlikely for variation only within a single gene or locus to influence the trait. Traits like height, body mass index (BMI), lipid levels, fasting glucose levels, blood pressure are just some notable examples. There are now well over a hundred loci that are associated with human height [30], each locus only has a very small effect on the overall height. For example, a variant in the HMGA2 locus (rs1351394), one of the first loci discovered to be associated with height, has an allele frequency of 49% and an effect size of 0.054 standard deviations or approximately 0.3 centimeters. That means every height increasing allele of this variant predicts on average an increase of only a 0.3 centimeter increase in overall height, which is just a small effect. For other quantitative traits, similar results were reported from association studies like BMI [43], LDL cholesterol [44] and blood pressure [45], etc, where many common variants have been found to be associated with these traits with each of these variants explaining only a small fraction of the overall trait. Because of the highly polygenic nature of such complex traits, methods like linkage analysis that were successful in identify loci for Mendelian disorders can be suboptimal when applied on complex traits. Therefore, new methodologies and paradigms were developed to map the variants in genes that influence

complex traits. We will discuss this in more detail in the next section.

METHODS FOR STUDYING GENETICS OF COMPLEX TRAITS

Single nucleotide polymorphisms (SNPs)

The principle for determining if variants in a gene cause or modulate risk for a disease or a trait is to be able to determine if they are associated with the disease or trait in a non-random way. As whole genome deep coverage sequencing in a large number of individuals is not feasible at this point in time (too costly), we rely on genetic markers for mapping a trait or disease to a genetic locus. The genetic marker that is currently very widely used for such a purpose is single-nucleotide polymorphisms (SNPs). SNPs are single base pair differences within the genome that is polymorphic in a population. As we are diploid for most of our chromosomes (males are largely hemizygous for the X-chromosome), some individuals in the population might have a different pair of alleles than other individuals for any particular SNP. These usually bi-allelic markers are found in abundance throughout the genome, much more frequently than STRs [46]. For each SNP, because of their bi-allelic nature as well as being diploid, each individual would largely be of only 3 genotypic states. For example, if the alleles for the SNP are “A” and “C”, then the 3 possible genotypic states would be homozygous “AA”, heterozygous “AC” or homozygous “CC”. SNPs were discovered and made publically available in a major way from the efforts of the International Hapmap Project [47]. In the phase 2 release of the project, they reported more than 3 million SNPs from 4 geographically diverse populations [9]. To find and characterize even more SNPs, the 1000 genomes project, a project that aims to characterize genomic variation from whole genome sequencing, reported their findings of about 15 million SNPs [48].

SNP genotyping strategies

While it might not be difficult or tedious to determine the genotypic state of any SNP in an individual, genotyping many SNPs for many individuals can be challenging, both from a technical as well as a cost perspective. Methods such as Sanger sequencing [49] and PCR-RFLP [50] were possible methods for performing SNP genotyping but is too tedious and expensive to perform them in high-throughput (many SNPs) over many individuals. As such, there was a need for a relatively cheap and fast technology that could genotype thousand of SNPs efficiently in many individuals. With success from efforts to characterize SNPs within human populations, that knowledge made it possible for the design of high-density SNP genotyping arrays. SNP genotyping arrays work in principle by probing for sequence variation of many targets in parallel by immobilizing the probe sequences on a surface and determine the genotype by reading out the strength to which these probe sequences are bound to their targets. These arrays can easily genotype many SNPs across the genome in a cost efficient manner [51]. There are now many companies that sell these high-density SNP genotyping arrays that can perform genotyping for over a million SNPs per sample. However, high-density SNP genotyping arrays might become less and less utilized with the growth and availability of whole-genome sequencing. Whole-genome sequencing cost have gone down significantly and it may come to a point in the near future that whole-genome sequencing will be the major strategy used by researchers to perform genotyping of genetic variants on a large scale.

Genotype imputation

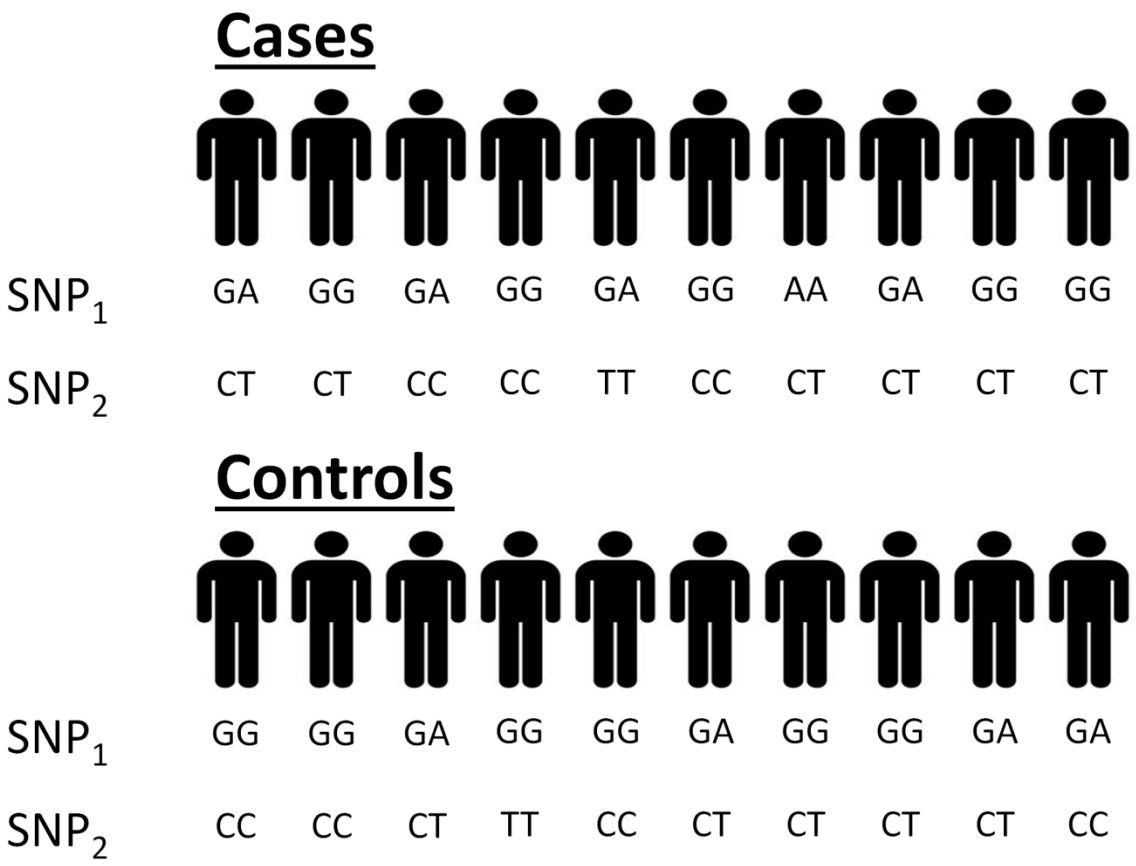
While it is possible to genotype many SNPs in parallel, it is still not possible to genotype all or most of the known SNPs in the human genome from SNP genotyping arrays. This is because there are just too many SNPs and it is impossible to fit all or most of them onto a single

genotyping array. As such, these SNP genotyping arrays have only a subset of the total possible SNPs from the human genome. Another potential problem would be that different companies that design and sell these arrays do not use the same subset of SNPs. This problem can be solved by performing genotype imputation. Genotype imputation is the process of determining the genotypes of unknown markers with some level of certainty using the genotype information of neighboring markers. This is possible because linkage disequilibrium, that variants within the genome are not independent [52]. This is because the human population is relatively new and variants that were introduced into the population tend to travel together. With enough time, recombination events between the variants will break the variants' correlation and bring about linkage equilibrium which will make imputation impossible. Genotype imputation can be performed computationally with the use of a reference panel. The reference panel is typically a more complete catalog of SNP genotypes obtained from a large cohort of individuals. Some examples of these panels would be those provided by the International Hapmap Project [47] as well as the 1000 genomes project [48] although it is not uncommon to use panels from other sources as well. With these panels together with the genotypes of one's samples, one can computationally impute the variants that are present in the panels but not genotyped in the samples. Some of the more utilized software for this purpose include BEAGLE [53], MACH [54] and IMPUTE2 [55] just to name a few. With imputation, SNPs that were directly genotyped in one set of samples that were not directly genotyped in other sets of samples can now be use for association studies.

Performing genome wide association

Linkage analysis has been shown to be less successful at identifying loci associated with complex traits than with Mendelian traits [13]. An arguably more effective approach would be to

perform a Genome wide association study (GWAS). Instead of tracking genetic markers in affected familial pedigrees, one can instead design a study and determine if the frequency of the genetic markers are significantly different between case individuals and control individuals. In such study designs, case individuals (cases) are usually randomly selected unrelated individuals that are affected and control individuals (controls) are randomly selected unrelated individuals that are unaffected. Assuming a scenario where 2 SNPs are genotyped in 1000 cases and 1000 controls (Figure 1.1), one can measure the frequency of the alleles in both SNPs to determine if the allele frequencies are significantly different by performing a chi-squared test. In this example, SNP_1 is significantly associated ($P = 2.82 \times 10^{-13}$) at a genome wide significance. The genome wide significance threshold is taken to be $P < 5 \times 10^{-8}$ although it has been suggested that it could be relaxed just a little [56]. The genome wide significance threshold has to be stringent to correct for multiple hypothesis testing given that GWAS test multiple markers at the same time [57]. SNP_2 on the other hand is only marginally associated ($P = 0.001$) and does not reach genome wide significance. This process can be systematically pursued for all the SNPs that were genotyped via the high-density SNP arrays and subsequently imputed from a reference panel. The first successful GWAS was performed on a disease called Age-related macular degeneration in 2005 [58]. In that study of 96 cases and 50 controls, they reported 2 strongly associated SNPs ($P < 10^{-7}$) in the complement factor H gene (CFH). Since then, there are many more GWAS performed with more than 10,000 SNPs identified as genome wide significant for various different traits and diseases in more than 1000 publications [59]. The large growth of GWAS can be attributed to the affordability of high density SNP arrays as well as freely available bioinformatics tools like PLINK [60] for data analysis. Besides performing



SNP	Allele	Case frequency	Control frequency	P-value
SNP1	A	30%	20%	2.82×10^{-13}
	G	70%	80%	
SNP2	T	40%	35%	0.001
	C	60%	75%	

Figure 1.1: An example of GWAS on cases versus controls. SNP_1 and SNP_2 are genotyped in 1000 cases and 1000 controls (1 stickman = 100 individuals). SNP_1 is significantly associated with disease status while SNP_2 is only marginally associated and does not reach genome wide significance. Genome wide significance is assumed to be $P < 5 \times 10^{-8}$.

case-control analyses, GWAS can also be performed on quantitative traits like height, BMI, blood pressure and lipid levels. Since there are no cases or controls, GWAS on quantitative traits seek to determine if the allele dosages for each SNP is significantly trending, either increasing or decreasing with the trait. This is usually done by linear regression of the allele dosages against the quantitative trait via a simple linear model [61]. For example, we simulated a scenario where a SNP with minor allele frequency of 30% have a 0.5 standard deviation effect (β) on the phenotype. After performing a linear regression of the allele dosages against the phenotypic score, we find a strong correlation between the SNP and the phenotype (Figure 1.2A) resulting in an estimated β of 0.47 and a very strong association signal ($P=2.97 \times 10^{-22}$). On the other hand, when we simulated a scenario where the SNP has no effect, then there is no strong correlation (Figure 1.2B). This example shows that GWAS can be use not only for dichotomous traits, but also for quantitative traits.

GWAS ineffective if causal variants not linked to SNPs

Linkage disequilibrium (LD) is a major factor for the success of GWAS. This is because, the vast majority of the time, SNP markers tested for association with diseases are not the actual genetic variant that has an effect but rather simply a marker that is in linkage disequilibrium with the disease variant. The disease variants could be SNPs, copy number polymorphisms (CNVs), short tandem repeats, insertion or deletion polymorphisms (indels) and perhaps even inversion polymorphisms. In most cases, there should be a SNP that is in LD (tagging) the causal variant. For example, many SNPs have been shown to be strong tagging the common inversion polymorphism on the human chromosome 17 [62]. It has also been shown that some SNPs from GWAS hits are strongly tagging CNVs and that these CNVs are suggested to be the causal variants [63]. However, we cannot discount the possibility that the causal variant is not

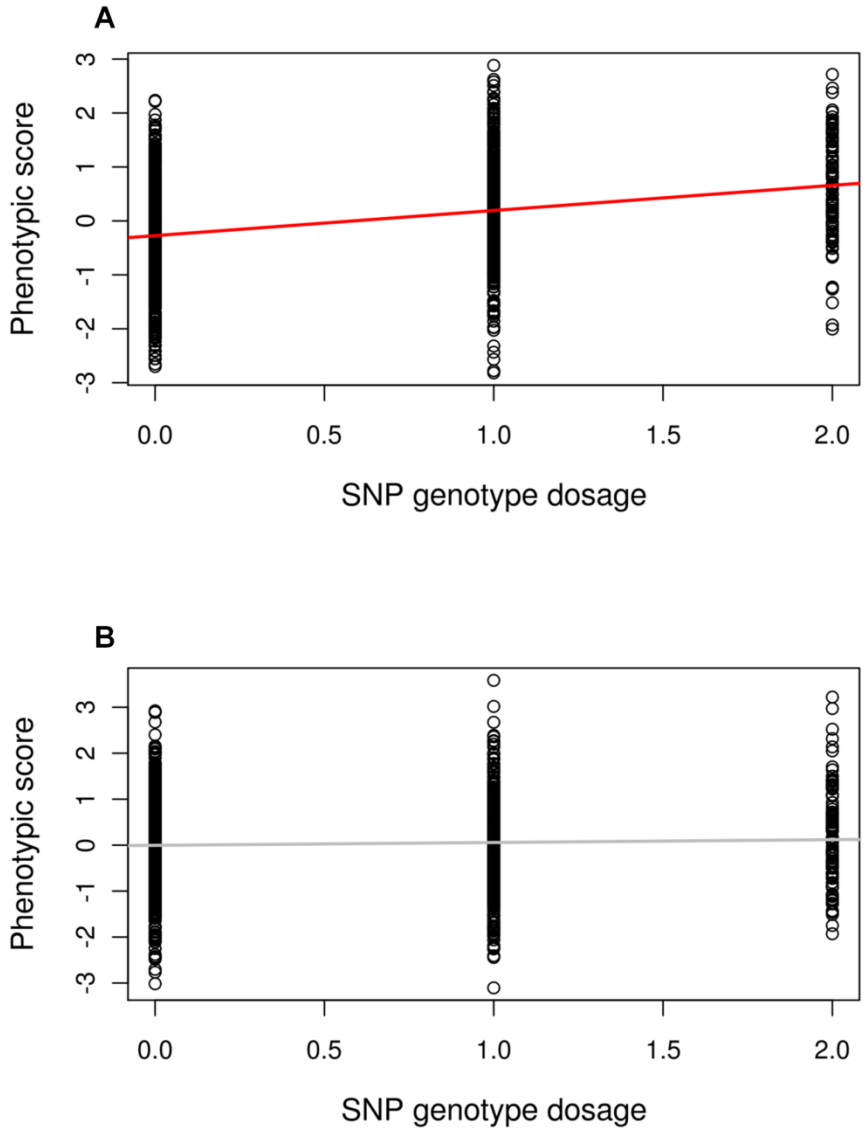


Figure 1.2: An example of GWAS on quantitative trait. Phenotypic score represents the quantitative trait. SNP genotype dosage is the number of effect alleles (0, 1 or 2) that each individual has. The association of between genotype and phenotype is shown by the least-squared regression line. (**A**) The least squared regression line (red) shows a positive correlation of genotype dosage with phenotype ($\beta=0.47$, $P=2.97 \times 10^{-22}$). (**B**) The least squared regression line (grey) shows no correlation of genotype dosage with phenotype ($\beta=0.06$, $P=0.21$).

well-tagged by SNPs. For example, a recent study showed that CNVs in two amylase genes (AMY1 and AMY2) are associated with obesity and that these regions are hard to be mapped by SNPs [64]. Only by genotyping the copy number did the authors observed the association. Therefore, besides performing GWAS using only SNPs as markers, it may be in some cases, useful to also genotype other potential markers, especially in genomic regions not well covered by SNPs.

Not straightforward to implicate causal gene from GWAS locus

While linkage disequilibrium allows one to find loci associated with disease, it is not clear which gene within the identified locus is the gene that is causal. Because of linkage disequilibrium, the region implicated in GWAS can span many genes and in that respect, linkage disequilibrium is more of a problem than a solution. To overcome this, solutions such as systems approaches that examines all the loci associated with the disease to determine its molecular architecture may be the way forward [65]. Using methods to determine if certain genes within various loci identified through GWAS are more biologically connected, those genes are more likely to be the causal gene within each of their locus. For example, in one study, the authors used a variety of biological functional databases to determine the degree of connectivity between genes [66]. In another study, the authors described an approach to form relationships between genes by analyzing PubMed abstracts [67]. These approaches have been successfully applied to results from GWAS and can prioritize the genes within each locus as to which of them are more likely to be the causal gene.

Population stratification

Genome wide association studies (GWAS) may also be confounded by population stratification. Unlike linkage analysis where studies are perform on familial pedigrees; GWAS on

the other hand compares genetic markers between unrelated cases against controls. As such, markers that reflect differences in the underlying structure of the population between cases and controls may have significant associations when performing GWAS. For example, a SNP in the LCT gene locus had significant association with height but the association is largely driven by population stratification [68]. Many methods have been developed to try and correct for population stratification. One of the more popular methods would be to include principal components as covariates when performing the statistical association with linear or logistic regression [69]. A study performed to determine the efficacy of the available methods showed that most of the methods work comparatively well to address the problem of population stratification [70]. Therefore population stratification is now not a major problem and can be adequately corrected for.

Admixture mapping

Another possible method besides GWAS would be admixture mapping. Admixture mapping, also known as "mapping by admixture linkage disequilibrium" (MALD) is a method that uses genetically mixed populations to determine if the local ancestry of different ancestral populations is correlated with a trait or disease [71]. For example, African Americans have genetic ancestry of largely African descent with a proportion being of European origin [72]. If one could determine the genomic regions of European ancestry, one could test if having European ancestry in these regions is associated with trait differences between individuals. Following this idea, methods were developed to accurately determine which regions in an individual's genome are of any particular ancestry. One of the first approach that is used extensively for this purpose is to perform the prediction using a hidden markov model (HMM) [73]. By systematically walking through each marker consecutively, HMM can be use to predict

the most likely ancestral state of the genetic marker given its frequency in each ancestral population. The more divergent the frequencies are in different population, the more likely the prediction will be accurate. The accuracy can be further improved by incorporating both linked markers as well as the use of an explicit population genetic model [74]. Admixture mapping has been performed on a multitude of phenotypes, including prostate cancer [75], body mass index [76,77], blood lipids [78], just to name a few. One of the reasons why admixture mapping might perform better than GWAS is because of admixture linkage disequilibrium. One of the initial reasons why GWAS on African populations might yield fewer results than GWAS performed on European or non-African populations is because the average linkage disequilibrium (LD) block in Africans is much smaller as they are a relatively older population [79,80]. As such, when there are relatively few SNPs genotyped for performing GWAS, it might be sufficient for studies in non-African populations but inadequate in populations of African ancestry. However, since admixture LD, LD of genomic regions due to admixture from a different population, is much stronger, this allows association signals to be discovered even with relatively lower marker density. However, this also means that if an admixture signal were to be discovered, it would be much harder to pinpoint the gene responsible for the association. GWAS on the other hand would be more sensitive and better powered if there is high density coverage of the genome, either from using high density SNP arrays or whole genome sequencing strategies. Given that high density SNP arrays are now widely used, GWAS might now be a better strategy to uncover genetic loci associated with disease.

COMPLEX PHENOTYPES

Human height is a classical complex trait

Human height is probably the best example of a heritable trait that has a polygenic architecture [81]. It is the example that Fisher used to reconcile how quantitative traits could also adhere to Mendelian inheritance [3]. Instead of having a single gene influencing the outcome of one's height, having many genes do so can explain the distribution of height in the population, which is in most cases, normally distributed [82]. However, we do know about diseases that are caused by rare mutations that have large effects on one's stature. These diseases, in most cases, have other obvious phenotypes besides the change in stature. For example, Achondroplasia, the most common cause of dwarfism is caused by a rare mutation in the FGFR3 gene. Individual carrying the mutant allele have on average about a 6 standard deviation decrease in height. The prevalence of Achondroplasia is extremely rare, affecting only about 1 in 25,000 individuals [83]. Another example would be Marfan syndrome, a genetic disorder caused by mutations in the FBN1 gene. Individuals with this Marfan syndrome are unusually tall, on average about 2 standard deviations taller. The prevalence of Marfan syndrome is rare, affecting only about 1 in 9802 individuals [84]. In both of these examples, individuals with Achondroplasia or Marfan syndrome have other consequential phenotypes as well besides their short or tall stature. Achondroplasia individuals usually present with other phenotypes like short fingers and toes [85]. Individuals with Marfan syndrome normally present with cardiovascular or vision problems too [86]. Nonetheless, rare Mendelian diseases like these do not explain for most of the variation of height in the population.

The alleles of height

Most of the variation of height is probably due to common variants that have small effect sizes. Indeed, the first such gene implicated in height is HMGA2 [87]. Identified from an initial GWAS of just under 5000 individuals, it harbors a common variant that has only an estimated

effect size of 0.4 cm per allele. Since then, many more common variants with small effects robustly associated with height have been discovered [30]. Among these variants, there are some that are associated with human syndromes characterized by abnormal skeletal growth. For example, the gene ACAN, of which there is a signal of common variant association with height, have been shown to be responsible for syndromes like Osteochondritis dissecans [88] and Spondyloepimetaphyseal dysplasia [89]. This suggests that while the common variants might be altering the gene activity in a minor way resulting in a small change in overall height, deleterious variants in these genes can cause severe reduction in stature. Thus the question remains as to what the genetic architecture is for non-syndromic individuals with short or tall stature. Is there a contribution of such large effect variants that can explain a person's tall or short stature in the general population? Or is a person's tall or short stature driven mainly by small effect common variants? In chapter 2, we shall discuss a method to infer the genetic architecture of individuals at the tails of the height distribution by examining the recently discovered common variants associated with height.

Body proportion is more constrained than height

While height is a commonly measured anthropometric that varies within a population, our heights are not as constrained and individuals can be relatively short or tall without any adverse effect on our health. Most of the problems associated with extreme tall or short stature are usually because of other adverse phenotypes associated with the tall or short stature. For example, individuals with Turner syndrome, a disease cause by monosomy X have short stature but commonly have other problems like Lymphedema or cardiovascular related problems. Also, given that women are about 2 standard deviations shorter than men shows that short stature itself is does not necessary have any health consequences and can vary within the population. On the

other hand, our body proportions are more well-defined. Humans have expected ratios of limb lengths that are vastly different from other species. For example, unlike humans, chimpanzees have arms longer than their legs [90].

Sitting height ratio as a measurement of body proportion

There certain measurements other than our full body height that can be use to judge our body proportions. Iliac length, subischial leg length, thigh length, knee height, sitting height are just some such measurements [91]. Another such measurement is arm span, which is a good proxy for overall height [92]. These measurements can be measured in a clinic but require either precise instruments or trained practitioners that they are usually not measured of patients when they pay a visit to their doctors even though they may be as informative as knowing our overall height and weight. However, one of the measurements that exist in some publically available data-sets is measurements of sitting height. Sitting height is the total stature that is comprised by the head and trunk. It is usually measured by first having the person sit on a table, then taking the measurement of the distance from the surface of the table to the top of the person's head. If one were to divide the sitting height with a person's height, one can calculate the sitting height ratio (SHR) which can then be a measure of body proportion. While short and tall stature is the characteristic of many skeletal dysplasia and overgrowth syndromes respectively, many of these syndromes can also cause severe deviations of SHR. For example, adult individuals with Achondroplasia have average SHR values of 0.66, very much higher than the population average, which is around 0.53 [93]. Another type of dysplasia, Spondyloepiphyseal dysplasia, is a syndrome characterized by severe short spines and neck. These patient's hands and feet are of normal length suggesting that their SHR values will be lower than average [94]. Next, individuals with Marfan syndrome have above average heights and may have lower than average

SHR values [95]. However, some individuals with mutations causing severe short stature might have SHR within the normal range. For example, a patient with premature pubarche and severe short stature has normal SHR [96]. SHR has also been used as a rudimentary predictor of phenotypes like body mass index, Age of Menarche and risk of diabetes [97]. Sitting height ratio (SHR) is a measurement that changes with age. More of our stature is due to our head and trunk as children than as adults, evidenced from the gradual decreasing of SHR till we reach adulthood [95].

Sitting height ratio and ancestry

SHR also differs significantly from individuals with different ancestries. Accordingly, individuals of Asian ancestry have higher SHRs than individuals of European ancestry and individuals of European ancestry have higher SHRs than individuals of African ancestry [91]. The question remains as to whether genetics is the primary driving force for the difference between SHR in different populations and whether these SHR differences between populations is a polygenic phenomenal or driven by only a single or a few genes. In chapter 4, I shall present some recent findings that will reveal more about the genetic architecture of SHR.

ACCUMULATING EVIDENCE FROM MULTIPLE STUDIES

Being underpowered

While Genome wide association studies (GWAS) have been very successful at elucidating loci that are associated with complex traits and diseases [98], this has not always been the case. Studies performed with limited samples are just underpowered for any genome wide significant associations to be discovered. The power to detect any SNP to be associated

with the trait is directly correlated to the variance of the phenotype explained by the SNP. This means that the larger the effect size or the more frequent the SNP is, the more power there is for the SNP to be detected as genome wide significant. However, given that for complex traits, the effect sizes for any given variant is very small, larger numbers of samples are required for any loci to be discovered.

Combining results by meta-analysis

While most studies may be underpowered due to small sample sizes, different studies performed on different samples with similar phenotypes could be combined or pooled together in an effort to increase the power of the study. Ideally, the genotypes and phenotypes could be shared among different research groups such that every group would have access to other group's data to perform the joint study. However, this is usually not feasible due to data sharing constraints such as the lack of storage space, privacy issues as well as the unwillingness of research groups to share their data prior to publication of their results. As such, for a typical GWAS, the association is performed on individual cohorts. Each of these cohorts has whole genome SNP data, usually produced by genotyping arrays as well as their corresponding phenotypes. The phenotype can be either a quantitative one, e.g. height, body mass index, blood pressure, etc, where there is a numerical value attached to each individual or a dichotomous one, e.g. type 2 diabetes, schizophrenia, etc, where each individual is either affected with the disease (cases) or are unaffected (controls). A dichotomous phenotype can be modeled as a phenotype with an underlying quantitative trait distribution, of which individuals whose trait value exceeded a threshold are affected and individuals who do not are unaffected [99]. For example, in the case of obesity, the underlying phenotype can be body mass index (bmi) and individuals whose bmi exceeds 30 can be classified as obese while those whose bmi are below 30 are not

[100]. However, in most dichotomous traits or diseases, this underlying trait is usually unobservable or unknown. Testing for genetic factors associated with the trait or disease is usually done by performing linear regression (quantitative trait), logistic regression (dichotomous traits) or some other test of correlation of the SNP dosages with the phenotype. Performing the test will produce resulting statistics for each SNP and by combining the statistics produced across cohorts through a process known as meta-analysis [101], one would be able to obtain the resulting summary statistics for the GWAS.

GWAS summary statistics

The resulting summary statistics contains the necessary information to determine which SNPs are significantly associated with the trait or disease in question. Typically, the summary statistics is reported in the following manner. Each row represents the result of the test for a SNP and each column reports a specific result for that SNP. There would a SNP identifier, usually the dbSNP rs-number [102], the allele frequency, the odds-ratio or effect size as well as the significance of the result, reported as the 2-tailed P-value. For a dichotomous trait, the odds-ratio (OR) would tell us the direction of effect of the allele, whether they are associated with increased or decreased risk for being affected by the trait or disease. An OR > 1 would indicate increased risk while an OR < 1 would indicate decreased risk. For a quantitative trait, the effect size (β) would be the equivalent, with a positive β indicating that the allele is associated with increased trait values and vice-versa. In either case, the P-value gives us the strength of the association and a P-value < 5×10^{-8} is suggested to be the genome-wide significant threshold [57]. SNPs that have P-values that are less than this threshold are said to have reached genome-wide significance and they are usually reported to be significantly associated with the trait or disease in question. Genes in the vicinity of such SNPs are then suspected to be involved with the trait or disease

etiology and are usually reported as well.

Independent loci

Even though by performing the GWAS, the SNPs that achieved genome-wide significance are not necessary independent and one must consider the effect of linkage disequilibrium (LD). As discussed earlier, there are many variants, SNPs included, in the genome that are correlated with one another due to LD. The SNP with the lowest P-value is called the "lead SNP" with SNPs in strong LD (usually taken as $r^2 > 0.5$) with it labeled as tagging SNPs. Together these SNPs represent only a single locus of association as the signal may well be only from a single source of variation within this region in the genome. To determine the total number of independent loci that is significantly associated with the trait, one can perform a process known as LD-pruning. This process orders the SNPs from most significant to least and systematically takes away significant SNPs that are in LD with any of the SNPs prior. On top of LD-pruning, one could also perform conditional analysis where SNPs in LD with the lead SNP can be tested again with the dosage of the lead SNP as a covariate. If the significant association is solely due to LD, the resulting P-value would not be significant. However, if the resulting P-value is still significant, then that SNP's significance cannot be explained just by LD and therefore could be counted as a separate locus associated with the trait or disease. This process could also be done in a high throughput manner taking existing summary statistics and LD information [103].

Not quite genome wide significant

While the genome-wide significant signals are said to be robust associations discovered with the trait or disease, SNPs that do not reach genome-wide significance could still be informative. While these marginally associated SNPs cannot be individually considered as robust

associations, they may inform us about the genetic architecture of the trait or disease as a whole. For example, in the GWAS of human height, the QQ-plots show a significant deviation of the marginally associated SNPs from the null model. This is indicative of the fact that there are more marginally associated SNPs than that expected under the null model and thus informs us that there are more associations to be discovered. In other cases, the QQ-plot might now show much of a deviation. Nonetheless, this suggests that the marginally associated SNPs can be informative even if most GWAS publications chose only to report the genome-wide significant ones. In chapter 3, I will discuss in depth a new method that can exploit the marginally associated SNPs to determine if there is evidence of polygenic inheritance.

SUMMARY

It is well known that many human traits and diseases do not follow a Mendelian pattern of inheritance; that most of these traits and diseases are influenced by variation in many genes, each of which contribute a small effect to the total heritability of the trait. These traits and diseases are highly heritable and therefore, mapping these traits and diseases to their genetic locus can be useful for understanding the disease etiology thereby informative for the development of potential therapeutics. Performing genetic association studies (GWAS), where one performs genetic genotyping on many genetic markers to determine if they are associated with the trait or disease has become a common technique for identifying such genetic loci. However, these loci discovered in most GWAS do not account for most of the heritability and thus our genetic understanding of these diseases is far from complete. This dissertation aims to leverage on results from GWAS to infer the genetic architecture of various complex human traits and diseases which could lead to increasing our understanding of disease etiology.

REFERENCES

1. Wills C (2007) Principles of Population Genetics, 4th edition. J Hered 98: 382–382. doi:10.1093/jhered/esm035.

2. Miko I (2008) Gregor Mendel and the principles of inheritance. Nat Educ 1: 134.

3. Fisher RA (1918) The correlation between relatives on the supposition of Mendelian inheritance. Trans R Soc Edinb 52: 399–433.

4. Kaminsky ZA, Tang T, Wang S-C, Ptak C, Oh GHT, et al. (2009) DNA methylation profiles in monozygotic and dizygotic twins. Nat Genet 41: 240–245. doi:10.1038/ng.286.

5. Hearne CM, Ghosh S, Todd JA (1992) Microsatellites for linkage analysis of genetic traits. Trends Genet TIG 8: 288–294.

6. Bornman DM, Hester ME, Schuetter JM, Kasoji MD, Minard-Smith A, et al. (2012) Short-read, high-throughput sequencing technology for STR genotyping. BioTechniques 0: 1–6. doi:10.2144/000113857.

7. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860–921. doi:10.1038/35057062.

8. The International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437: 1299–1320. doi:10.1038/nature04226.

9. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851–861. doi:10.1038/nature06258.

10. Elston RC (1998) Methods of linkage analysis--and the assumptions underlying them [see comment]. Am J Hum Genet 63: 931–934.

11. MORTON NE (1955) Sequential tests for the detection of linkage. Am J Hum Genet 7: 277–318.

12. Jorde LB (2000) Linkage Disequilibrium and the Search for Complex Disease Genes. Genome Res 10: 1435–1444. doi:10.1101/gr.144500.

13. Altmüller J, Palmer LJ, Fischer G, Scherb H, Wjst M (2001) Genomewide scans of complex human diseases: true linkage is hard to find. Am J Hum Genet 69: 936–950. doi:10.1086/324069.

14. D'Andrea AD (2010) Susceptibility Pathways in Fanconi's Anemia and Breast Cancer. N Engl J Med 362: 1909–1919. doi:10.1056/NEJMra0809889.

15. Cantor RM, Lange K, Sinsheimer JS (2010) Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application. Am J Hum Genet 86: 6–

22. doi:10.1016/j.ajhg.2009.11.017.

16. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. (2009) Finding the missing heritability of complex diseases. Nature 461: 747–753. doi:10.1038/nature08494.

17. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, et al. (2010) Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet 11: 446–450. doi:10.1038/nrg2809.

18. Li Y, Willer C, Sanna S, Abecasis G (2009) Genotype Imputation. Annu Rev Genomics Hum Genet 10: 387–406. doi:10.1146/annurev.genom.9.081307.164242.

19. McKusick VA (2007) Mendelian Inheritance in Man and Its Online Version, OMIM. Am J Hum Genet 80: 588–604. doi:10.1086/514346.

20. Visscher PM, Hill WG, Wray NR (2008) Heritability in the genomics era -- concepts and misconceptions. Nat Rev Genet 9: 255–266. doi:10.1038/nrg2322.

21. Jacquard A (1983) Heritability: one word, three concepts. Biometrics 39: 465–477.

22. Aulchenko YS, Struchalin MV, Belonogova NM, Axenovich TI, Weedon MN, et al. (2009) Predicting human height by Victorian and genomic methods. Eur J Hum Genet 17: 1070–1075. doi:10.1038/ejhg.2009.5.

23. Silventoinen K (2003) Determinants of variation in adult body height. J Biosoc Sci 35: 263–285.

24. Teikari JM, Kaprio J, Koskenvuo MK, Vannas A (1988) Heritability estimate for refractive errors--a population-based sample of adult twins. Genet Epidemiol 5: 171–181. doi:10.1002/gepi.1370050304.

25. Visscher PM, Medland SE, Ferreira MAR, Morley KI, Zhu G, et al. (2006) Assumption-Free Estimation of Heritability from Genome-Wide Identity-by-Descent Sharing between Full Siblings. PLoS Genet 2: e41. doi:10.1371/journal.pgen.0020041.

26. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, et al. (2010) Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42: 565–569. doi:10.1038/ng.608.

27. Charmantier A, Garant D (2005) Environmental quality and evolutionary potential: lessons from wild populations. Proc Biol Sci 272: 1415–1425. doi:10.1098/rspb.2005.3117.

28. McGrath J, Saha S, Chant D, Welham J (2008) Schizophrenia: A Concise Overview of Incidence, Prevalence, and Mortality. Epidemiol Rev 30: 67–76. doi:10.1093/epirev/mxn001.

29. Hotu S, Carter B, Watson P, Cutfield W, Cundy T (2004) Increasing prevalence of type 2 diabetes in adolescents. J Paediatr Child Health 40: 201–204. doi:10.1111/j.1440-

1754.2004.00337.x.

30. Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, et al. (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature. Available: http://www.ncbi.nlm.nih.gov.ezp-prod1.hul.harvard.edu/pubmed/20881960. Accessed 4 October 2010.

31. Richette P, Bardin T, Stheneur C (2008) Achondroplasia: from genotype to phenotype. Jt Bone Spine Rev Rhum 75: 125–130. doi:10.1016/j.jbspin.2007.06.007.

32. Gillis E, Kempers M, Salemink S, Timmermans J, Cheriex EC, et al. (2014) An FBN1 Deep Intronic Mutation in a Familial Case of Marfan Syndrome: An Explanation for Genetically Unsolved Cases? Hum Mutat. doi:10.1002/humu.22540.

33. Fagnani C, Annesi-Maesano I, Brescianini S, D'Ippolito C, Medda E, et al. (2008) Heritability and shared genetic effects of asthma and hay fever: an Italian study of young twins. Twin Res Hum Genet Off J Int Soc Twin Stud 11: 121–131. doi:10.1375/twin.11.2.121.

34. O'Donovan MC, Williams NM, Owen MJ (2003) Recent advances in the genetics of schizophrenia. Hum Mol Genet 12: R125–R133. doi:10.1093/hmg/ddg302.

35. Almgren P, Lehtovirta M, Isomaa B, Sarelin L, Taskinen MR, et al. (2011) Heritability and familiality of type 2 diabetes and related quantitative traits in the Botnia Study. Diabetologia 54: 2811–2819. doi:10.1007/s00125-011-2267-5.

36. Brant SR (2011) Update on the heritability of inflammatory bowel disease: The importance of twin studies. Inflamm Bowel Dis 17: 1–5. doi:10.1002/ibd.21385.

37. Mayer B, Erdmann J, Schunkert H (2007) Genetics and heritability of coronary artery disease and myocardial infarction. Clin Res Cardiol Off J Ger Card Soc 96: 1–7. doi:10.1007/s00392-006-0447-y.

38. Grant SFA, Thorleifsson G, Reynisdottir I, Benediktsson R, Manolescu A, et al. (2006) Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. Nat Genet 38: 320–323. doi:10.1038/ng1732.

39. Morris AP, Voight BF, Teslovich TM, Ferreira T, Segrè AV, et al. (2012) Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat Genet 44: 981–990. doi:10.1038/ng.2383.

40. Bergen SE, Petryshen TL (2012) Genome-wide association studies (GWAS) of schizophrenia: does bigger lead to better results? Curr Opin Psychiatry 25: 76–82. doi:10.1097/YCO.0b013e32835035dd.

41. Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, et al. (2009) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460: 748–752. doi:10.1038/nature08185.

42. Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium (2011) Genome-wide association study identifies five new schizophrenia loci. Nat Genet 43: 969–976. doi:10.1038/ng.940.

43. Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, et al. (2010) Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat Genet 42: 937–948. doi:10.1038/ng.686.

44. Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, et al. (2010) Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466: 707–713. doi:10.1038/nature09270.

45. Newton-Cheh C, Johnson T, Gateva V, Tobin MD, Bochud M, et al. (2009) Eight blood pressure loci identified by genome-wide association study of 34,433 people of European ancestry. Nat Genet 41: 666–676. doi:10.1038/ng.361.

46. Evans DM, Cardon LR (2004) Guidelines for Genotyping in Genomewide Linkage Studies: Single-Nucleotide–Polymorphism Maps Versus Microsatellite Maps. Am J Hum Genet 75: 687–692. doi:10.1086/424696.

47. International HapMap Consortium (2003) The International HapMap Project. Nature 426: 789–796. doi:10.1038/nature02168.

48. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56–65. doi:10.1038/nature11632.

49. Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci 74: 5463–5467.

50. Novel detection assay by PCR–RFLP and frequency of the CYP3A... : Pharmacogenetics and Genomics (n.d.). Available: http://journals.lww.com/jpharmacogenetics/Fulltext/2002/06000/Novel_detection_assay_by_PCR_RFLP_and_frequency_of.9.aspx. Accessed 2 April 2014.

51. Mei R, Galipeau PC, Prass C, Berno A, Ghandour G, et al. (2000) Genome-wide Detection of Allelic Imbalance Using Human SNPs and High-density DNA Arrays. Genome Res 10: 1126–1137. doi:10.1101/gr.10.8.1126.

52. Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, et al. (2001) Linkage disequilibrium in the human genome. Nature 411: 199–204. doi:10.1038/35075590.

53. Browning BL, Browning SR (2009) A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. Am J Hum Genet 84: 210–223. doi:10.1016/j.ajhg.2009.01.005.

54. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol 34: 816–834.

doi:10.1002/gepi.20533.

55. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR (2012) Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet 44: 955–959. doi:10.1038/ng.2354.

56. Panagiotou OA, Ioannidis JPA, Genome-Wide Significance Project (2012) What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. Int J Epidemiol 41: 273–286. doi:10.1093/ije/dyr178.

57. Johnson RC, Nelson GW, Troyer JL, Lautenberger JA, Kessing BD, et al. (2010) Accounting for multiple comparisons in a genome-wide association study (GWAS). BMC Genomics 11: 724. doi:10.1186/1471-2164-11-724.

58. Klein RJ, Zeiss C, Chew EY, Tsai J-Y, Sackler RS, et al. (2005) Complement Factor H Polymorphism in Age-Related Macular Degeneration. Science 308: 385–389. doi:10.1126/science.1109557.

59. Welter D, MacArthur J, Morales J, Burdett T, Hall P, et al. (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res 42: D1001–1006. doi:10.1093/nar/gkt1229.

60. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559–575. doi:10.1086/519795.

61. Anderson CA, McRae AF, Visscher PM (2006) A Simple Linear Regression Method for Quantitative Trait Loci Linkage Analysis With Censored Observations. Genetics 173: 1735–1745. doi:10.1534/genetics.106.055921.

62. Stefansson H, Helgason A, Thorleifsson G, Steinthorsdottir V, Masson G, et al. (2005) A common inversion under selection in Europeans. Nat Genet 37: 129–137. doi:10.1038/ng1508.

63. Gamazon ER, Nicolae DL, Cox NJ (2011) A Study of CNVs As Trait-Associated Polymorphisms and As Expression Quantitative Trait Loci. PLoS Genet 7. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3033384/. Accessed 3 April 2014.

64. Falchi M, El-Sayed Moustafa JS, Takousis P, Pesce F, Bonnefond A, et al. (2014) Low copy number of the salivary amylase gene predisposes to obesity. Nat Genet advance online publication. Available: http://www.nature.com/ng/journal/vaop/ncurrent/full/ng.2939.html. Accessed 3 April 2014.

65. Civelek M, Lusis AJ (2014) Systems genetics approaches to understand complex traits. Nat Rev Genet 15: 34–48. doi:10.1038/nrg3575.

66. Franke L, Bakel H van, Fokkens L, de Jong ED, Egmont-Petersen M, et al. (2006) Reconstruction of a Functional Human Gene Network, with an Application for Prioritizing

Positional Candidate Genes. Am J Hum Genet 78: 1011–1025. doi:10.1086/504300.

67. Raychaudhuri S, Plenge RM, Rossin EJ, Ng ACY, Purcell SM, et al. (2009) Identifying Relationships among Genomic Disease Regions: Predicting Genes at Pathogenic SNP Associations and Rare Deletions. PLoS Genet 5: e1000534. doi:10.1371/journal.pgen.1000534.

68. Campbell CD, Ogburn EL, Lunetta KL, Lyon HN, Freedman ML, et al. (2005) Demonstrating stratification in a European American population. Nat Genet 37: 868–872. doi:10.1038/ng1607.

69. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38: 904–909. doi:10.1038/ng1847.

70. Wu C, DeWan A, Hoh J, Wang Z (2011) A comparison of association methods correcting for population stratification in case-control studies. Ann Hum Genet 75: 418–427. doi:10.1111/j.1469-1809.2010.00639.x.

71. McKeigue PM (2005) Prospects for admixture mapping of complex traits. Am J Hum Genet 76: 1–7. doi:10.1086/426949.

72. Tang H, Jorgenson E, Gadde M, Kardia SLR, Rao DC, et al. (2006) Racial admixture and its impact on BMI and blood pressure in African and Mexican Americans. Hum Genet 119: 624–633. doi:10.1007/s00439-006-0175-4.

73. Patterson N, Hattangadi N, Lane B, Lohmueller KE, Hafler DA, et al. (2004) Methods for high-density admixture mapping of disease genes. Am J Hum Genet 74: 979–1000. doi:10.1086/420871.

74. Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, et al. (2009) Sensitive Detection of Chromosomal Segments of Distinct Ancestry in Admixed Populations. PLoS Genet 5: e1000519. doi:10.1371/journal.pgen.1000519.

75. Freedman ML, Haiman CA, Patterson N, McDonald GJ, Tandon A, et al. (2006) Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. Proc Natl Acad Sci U S A 103: 14068–14073. doi:10.1073/pnas.0605832103.

76. Cheng C-Y, Kao WHL, Patterson N, Tandon A, Haiman CA, et al. (2009) Admixture mapping of 15,280 African Americans identifies obesity susceptibility loci on chromosomes 5 and X. PLoS Genet 5: e1000490. doi:10.1371/journal.pgen.1000490.

77. Basu A, Tang H, Arnett D, Gu CC, Mosley T, et al. (2009) Admixture mapping of quantitative trait loci for BMI in African Americans: evidence for loci on chromosomes 3q, 5q, and 15q. Obes Silver Spring Md 17: 1226–1231. doi:10.1038/oby.2009.24.

78. Basu A, Tang H, Lewis CE, North K, Curb JD, et al. (2009) Admixture mapping of quantitative trait loci for blood lipids in African-Americans. Hum Mol Genet 18: 2091–

2098. doi:10.1093/hmg/ddp122.

79. Sawyer SL, Mukherjee N, Pakstis AJ, Feuk L, Kidd JR, et al. (2005) Linkage disequilibrium patterns vary substantially among populations. Eur J Hum Genet EJHG 13: 677–686. doi:10.1038/sj.ejhg.5201368.

80. Kang SJ, Chiang CWK, Palmer CD, Tayo BO, Lettre G, et al. (2010) Genome-wide association of anthropometric traits in African- and African-derived populations. Hum Mol Genet 19: 2725–2738. doi:10.1093/hmg/ddq154.

81. Visscher PM (2008) Sizing up human height variation. Nat Genet 40: 489–490. doi:10.1038/ng0508-489.

82. Schilling MF, Watkins AE, Watkins W (2002) Is Human Height Bimodal? Am Stat 56: 223–229. doi:10.1198/00031300265.

83. Wynn J, King TM, Gambello MJ, Waller DK, Hecht JT (2007) Mortality in achondroplasia study: a 42-year follow-up. Am J Med Genet A 143A: 2502–2511. doi:10.1002/ajmg.a.31919.

84. Gray JR, Bridges AB, Faed MJ, Pringle T, Baines P, et al. (1994) Ascertainment and severity of Marfan syndrome in a Scottish population. J Med Genet 31: 51–54. doi:10.1136/jmg.31.1.51.

85. Scott CI Jr (1976) Achondroplastic and hypochondroplastic dwarfism. Clin Orthop: 18–30.

86. Roberts WC, Honig HS (1982) The spectrum of cardiovascular disease in the Marfan syndrome: A clinico-morphologic study of 18 necropsy patients and comparison to 151 previously reported necropsy patients. Am Heart J 104: 115–135. doi:10.1016/0002-8703(82)90650-0.

87. Weedon MN, Lettre G, Freathy RM, Lindgren CM, Voight BF, et al. (2007) A common variant of HMGA2 is associated with adult and childhood height in the general population. Nat Genet 39: 1245–1250. doi:10.1038/ng2121.

88. Stattin E-L, Wiklund F, Lindblom K, Onnerfjord P, Jonsson B-A, et al. (2010) A missense mutation in the aggrecan C-type lectin domain disrupts extracellular matrix interactions and causes dominant familial osteochondritis dissecans. Am J Hum Genet 86: 126–137. doi:10.1016/j.ajhg.2009.12.018.

89. Tompson SW, Merriman B, Funari VA, Fresquet M, Lachman RS, et al. (2009) A recessive skeletal dysplasia, SEMD aggrecan type, results from a missense mutation affecting the C-type lectin domain of aggrecan. Am J Hum Genet 84: 72–79. doi:10.1016/j.ajhg.2008.12.001.

90. Zihlman AL, Stahl D, Boesch C (2008) Morphological variation in adult chimpanzees (Pan troglodytes verus) of the Taï National Park, Côte d'Ivoire. Am J Phys Anthropol 135: 34–41. doi:10.1002/ajpa.20702.

91. Bogin B, Varela-Silva MI (2010) Leg length, body proportion, and health: a review with a note on beauty. Int J Environ Res Public Health 7: 1047–1075. doi:10.3390/ijerph7031047.

92. Chhabra SK (2008) Using arm span to derive height: Impact of three estimates of height on interpretation of spirometry. Ann Thorac Med 3: 94–99. doi:10.4103/1817-1737.39574.

93. Stokes DC, Pyeritz RE, Wise RA, Fairclough D, Murphy EA (1988) Spirometry and chest wall dimensions in achondroplasia. Chest 93: 364–369.

94. Spranger JW, Densler J (1970) Spondyloepiphyseal Dysplasia Congenita. Radiology 94: 313–322. doi:10.1148/94.2.313.

95. Fredriks A, van Buuren S, van Heel WJM, Dijkman-Neerincx R, Verloove-Vanhoric... S, et al. (2005) Nationwide age references for sitting height, leg length, and sitting height/height ratio, and their diagnostic value for disproportionate growth disorders. Arch Dis Child 90: 807–812. doi:10.1136/adc.2004.050799.

96. Noordam C, Dhir V, McNelis JC, Schlereth F, Hanley NA, et al. (2009) Inactivating PAPSS2 Mutations in a Patient with Premature Pubarche. N Engl J Med 360: 2310–2318. doi:10.1056/NEJMoa0810489.

97. Conway BN, Shu X-O, Zhang X, Xiang Y-B, Cai H, et al. (2012) Age at Menarche, the Leg Length to Sitting Height Ratio, and Risk of Diabetes in Middle-Aged and Elderly Chinese Men and Women. PLoS ONE 7. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3309033/. Accessed 16 March 2014.

98. Clarke AJ, Cooper DN (2010) GWAS: heritability missing in action? Eur J Hum Genet 18: 859–861. doi:10.1038/ejhg.2010.35.

99. Dempster ER, Lerner IM (1950) Heritability of Threshold Characters. Genetics 35: 212–236.

100. Sharma AM, Kushner RF (2009) A proposed clinical staging system for obesity. Int J Obes 33: 289–295. doi:10.1038/ijo.2009.2.

101. Willer CJ, Li Y, Abecasis GR (2010) METAL: fast and efficient meta-analysis of genomewide association scans. Bioinforma Oxf Engl 26: 2190–2191. doi:10.1093/bioinformatics/btq340.

102. Sherry ST, Ward M-H, Kholodov M, Baker J, Phan L, et al. (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29: 308–311. doi:10.1093/nar/29.1.308.

103. Yang J, Ferreira T, Morris AP, Medland SE, Consortium GI of AnT (GIANT), et al. (2012) Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat Genet 44: 369–375. doi:10.1038/ng.2213.

Chapter 2

Common variants show predicted polygenic effects on height in the tails of the distribution, except in extremely short individuals

Yingleong Chan[1,2,3*], Oddgeir L Holmen[4,5*], Andrew Dauber[2,3*], Lars Vatten[6], Aki S Havulinna[7], Frank Skorpen[8], Kirsti Kvaløy[4], Kaisa Silander[7,9], Thutrang T Nguyen[3], Cristen Willer[10], Michael Boehnke[10], Markus Perola[7,9,11], Aarno Palotie[2,9,12,13], Veikko Salomaa[7], Kristian Hveem[4], Timothy M Frayling[14*], Joel N Hirschhorn[1,2,3*], Michael N Weedon[14*]

[1] Harvard Medical School, Department of Genetics, Boston, Massachusetts, USA.

[2] Broad Institute, Cambridge, Massachusetts, USA.

[3] Children's Hospital Boston, Boston, Massachusetts, USA.

[4] HUNT Research Centre, Department of Public Health and General Practice, Norwegian University of Science and Technology, Levanger, Norway.

[5] St. Olav Hospital, Trondheim University Hospital, Trondheim, Norway.

[6] Department of Public Health and General Practice, Norwegian University of Science and Technology, Trondheim, Norway.

[7] National Institute for Health and Welfare, Helsinki, Finland.

[8] Department of Laboratory Medicine, Children's and Women's Health, Norwegian University of Science and Technology, Trondheim, Norway.

[9] Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland.

[10] Department of Internal Medicine, Division of Cardiovascular Medicine, University of Michigan, Ann Arbor, Michigan, USA.

[11] Estonian Genome Project, University of Tartu, Tartu, Estonia.

[12] Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK.

[13] Department of Medical Genetics, University of Helsinki and University Central Hospital, Helsinki, Finland.

[14] Genetics of Complex Traits, Peninsula Medical School, University of Exeter, Exeter, UK.

[*] These authors contributed equally to this work

Originally published as:

Chan Y, Holmen OL, Dauber A, et. al., PLOS Genetics (2011). DOI: 10.1371/e1002439

ABSTRACT

Common genetic variants have been shown to explain a fraction of the inherited variation for many common diseases and quantitative traits, including height, a classic polygenic trait. The extent to which common variation determines the phenotype of highly heritable traits such as height is uncertain, as is the extent to which common variation is relevant to individuals with more extreme phenotypes. To address these questions, we studied 1,214 individuals from the top and bottom extremes of the height distribution (tallest and shortest ~1.5%), drawn from ~78,000 individuals from the HUNT and FINRISK cohorts. We found that common variants still influence height at the extremes of the distribution: common variants (49/141) were nominally associated with height in the expected direction more often than is expected by chance ($p < 5 \times 10^{-28}$) and the odds ratios in the extreme samples were consistent with the effects estimated previously in population-based data. To examine more closely whether the common variants have the expected effects, we calculated a weighted allele score (*WAS*), which is a weighted prediction of height for each individual based on the previously estimated effect sizes of the common variants in the overall population. The average *WAS* is consistent with expectation in the tall individuals, but was not as extreme as expected in the shortest individuals ($p<0.006$), indicating that some of the short stature is explained by factors other than common genetic variation. The discrepancy was more pronounced ($p<10^{-6}$) in the most extreme individuals (height <0.25 percentile). The results at the extreme short tails are consistent with a large number of models incorporating either rare genetic, non-additive or rare non-genetic factors that decrease height. We conclude that common genetic variants are associated with height at the extremes as well as across the population, but that additional factors become more prominent at the shorter extreme.

AUTHOR SUMMARY

Although there are many loci in the human genome that have been discovered to be significantly associated with height, it is unclear if these loci have similar effects in extremely tall and short individuals. Here, we examine hundreds of extremely tall and short individuals in 2 population-based cohorts to see if these known height determining loci are as predictive as expected in these individuals. We found that these loci are generally as predictive of height as expected in these individuals but that they begin to be less predictive in the most extremely short individuals. We showed that this result is consistent with models that not only include the common variants but also multiple low frequency genetic variants that substantially decrease height. However, this result is also consistent with non-additive genetic effects or rare non-genetic factors that substantially decrease height. This finding suggests the possibility of a major role of low frequency variants, particularly in individuals with extreme phenotypes and has implications on whole-genome or whole-exome sequencing efforts to discover rare genetic variation associated with complex traits.

INTRODUCTION

Height is a highly heritable trait, with estimates of heritability as high as 90% [1]. Recent genome-wide association studies of height have discovered over 180 common variants associated with height [2]. These variants have small effect sizes and collectively explain approximately 10% of the heritability. While these 180 common variants are robustly associated with height when studied as a quantitative trait in the general population, it is not known whether these variants have similar associations with stature in individuals at the extreme tails of the height distribution. If these common variants do not show the expected association with stature

at the extremes (based on their continuous distribution effect sizes), then other factors beyond common variants must contribute to extreme stature. Although there are multiple possible scenarios, one possible explanation is the existence of rare or low frequency variants with larger effect sizes, which have been proposed to explain a portion of the heritability not accounted for by the known common variants [3–5] and which may provide novel biological insights into mechanisms that affects height. Understanding the role of common variants in the tails of the height distribution will also provide methodological insight into the utility of extreme tails analysis for future genetic studies of quantitative traits.

In this chapter, we describe our approach to determine whether common alleles known to be associated with height in the general population have the expected distribution in individuals from the extremes of the height distribution. We used DNA samples from individuals with extreme heights from two population-based cohorts of Finnish (FINRISK) and Norwegian (HUNT) ancestry and genotyped them for common variants known to be associated with height. Under a polygenic model in which there are many variants and each variant additively contributes a small effect to the phenotype, we found that for individuals within ~2.81 standard deviations of the mean, the common variants have the predicted associations with height, consistent with their effect sizes estimated from the previous population study [2]. However, in individuals with more extreme short stature (the shortest 0.25% of the distribution), common variants play a less prominent role in explaining phenotype, and the data are consistent with various models in which rare variants, non-additive effects or rare non-genetic factors contribute to short stature in these individuals.

RESULTS

Individual common variants are associated with height in the extremes

We attempted to genotype SNPs at the 180 loci previously associated with height in individuals from the short and tall extremes of the FINRISK and HUNT cohorts and then performed association analyses for each SNP with height using the Cochran-Mantel-Haenszel test and logistic regression respectively. In FINRISK, SNPs at 158 of the height loci were successfully genotyped in 181 short and 192 tall individuals from the 1% tails of the height distribution. In the HUNT study, SNPs at 160 of the height loci were successfully genotyped in 385 short and 456 tall individuals from the ~1.5% tails of height. Here we focus on the 279 short and 309 tall individuals from the 1% tails of the HUNT study, so as to provide consistency with the FINRISK study. In both cohorts, the majority of SNPs had effect directions consistent with the published results [2] (HUNT 137/160, $p<0.0001$; FINRISK 122/155, $p<0.0001$) and there was a significant enrichment in SNPs reaching nominal significance for association with height (Table 2.1; Table 2.2). We then combined the data from both cohorts in a meta-analysis of 141 overlapping loci (Table 2.3). Ninety-one percent of SNPs (128/141, $p<0.0001$) had directions of effect consistent with previously published results [2] and 49 SNPs had p-values <0.05, as opposed to 7 expected by chance ($p<5x10^{-28}$). This result confirms that, as a group, SNPs found to be associated with height in the general population are also associated with height at the extremes of the phenotypic spectrum.

The effect sizes of individual common variants on height are similar in the extremes and the general population

We next tested whether the observed odds ratios (OR) are consistent with the expected odds ratios, based on the previously estimated effect sizes from the GIANT study [2] and study

Table 2.1: Individual SNP analysis for HUNT cohort

Rsid	Chr	Pos	Closest gene	Effect allele	Effect size	Freq	Observed OR	Expected OR
rs425277	1	2059032	PRKCZ	t	0.0240	0.28	1.19	1.14
rs6657613	1	17200787	MFAP2	t	0.0328	0.52	1.19	1.19
rs2903545	1	23413695	HTR1D	t	0.0215	0.59	1.16	1.12
rs4601530	1	24916698	CLIC4	c	0.0238	0.74	1.02	1.14
rs7532866	1	26614131	LIN28	a	0.0222	0.68	1.16	1.13
rs11209376	1	41270935	SCMH1	g	0.0319	0.22	1.33	1.19
rs17391694	1	78396214	GIPC2	t	0.0399	0.13	1.25	1.24
rs6699417	1	88896031	PKN2	t	0.0217	0.63	1.12	1.12
rs10874746	1	93096559	RPL5	c	0.0217	0.64	1.12	1.12
rs12731372	1	118654498	SPAG17	c	0.0379	0.75	1.25	1.22
rs11205277	1	148159496	SF3B4	g	0.0452	0.44	1.46	1.27
rs17346473	1	170349716	DNM3	g	0.0365	0.30	0.95	1.21
rs1014719	1	175069389	PAPPA2	t	0.0253	0.55	1.21	1.14
rs1046934	1	182290152	TSEN15	c	0.0459	0.35	1.21	1.28
rs10863936	1	210304421	DTL	g	0.0220	0.47	1.13	1.12
rs6684205	1	216676325	TGFB2	g	0.0328	0.26	1.15	1.19
rs11118346	1	217810342	LYPLAL1	c	0.0264	0.58	1.12	1.15
rs1172294	2	25022704	DNAJC27	a	0.0334	0.50	1.42	1.19
rs1545552	2	33213842	LTBP1	g	0.0246	0.72	1.27	1.14
rs2341459	2	44621706	C2orf34	t	0.0276	0.28	1.18	1.16
rs3791675	2	55964813	EFEMP1	c	0.0496	0.75	1.65	1.30
rs1913671	2	88680998	EIF2AK3	c	0.0268	0.37	1.31	1.15
rs7567288	2	134151294	NCKAP5	c	0.0309	0.22	1.11	1.18
rs3770047	2	178393780	PDE11A	g	0.0402	0.06	0.99	1.24
rs12470505	2	219616613	CCDC108/IHH	t	0.0483	0.91	0.85	1.29
rs6756793	2	224737163	SERPINE2	t	0.0248	0.55	1.30	1.14
rs12694997	2	241911659	SEPT2	g	0.0274	0.77	1.07	1.16
rs2597513	3	13530836	HDAC11	c	0.0392	0.10	1.73	1.23
rs13088462	3	51046753	DOCK3	c	0.0543	0.07	1.49	1.34
rs2336725	3	53093779	RTF1	c	0.0263	0.46	0.96	1.15
rs9833926	3	56625218	C3orf63	a	0.0216	0.50	1.24	1.12
rs17806888	3	67499012	SUCLG2	t	0.0399	0.90	1.40	1.24
rs11128265	3	72538487	RYBP	a	0.0304	0.80	1.27	1.18
rs6765930	3	130503468	C3orf47	g	0.0352	0.79	1.24	1.21
rs9844666	3	137456906	PCCB	g	0.0284	0.76	1.36	1.16
rs724016	3	142588260	ZBTB38	g	0.0670	0.46	1.52	1.43

Table 2.1 (Continued)

Rsid	Chr	Pos	Closest gene	Effect allele	Effect size	Freq	Observed OR	Expected OR
rs572169	3	173648421	GHSR	t	0.0355	0.33	1.32	1.21
rs720390	3	187031377	IGF2BP2	a	0.0305	0.36	1.24	1.18
rs2247341	4	1671115	SLBP/FGFR3	a	0.0251	0.38	1.29	1.14
rs6449353	4	17642586	LCORL	t	0.0714	0.86	1.52	1.46
rs17081935	4	57518233	POLR2B	t	0.0306	0.19	1.01	1.18
rs7697556	4	73734177	ADAMTS3	t	0.0219	0.48	1.21	1.12
rs1975474	4	82397961	PRKG2/BMP3	g	0.0376	0.30	1.30	1.22
rs10010325	4	106325802	TET2	a	0.0214	0.47	1.21	1.12
rs7689420	4	145787802	HHIP	c	0.0687	0.84	1.28	1.44
rs955748	4	184452669	WWC2	g	0.0243	0.78	1.23	1.14
rs13154066	5	32867427	NPR3	t	0.0350	0.39	1.12	1.20
rs6897117	5	55022532	SLC38A9	t	0.0278	0.27	1.28	1.16
rs6894139	5	88363538	MEF2C	t	0.0266	0.56	1.18	1.15
rs13177718	5	108141243	FER	c	0.0412	0.90	1.48	1.25
rs274546	5	131727766	SLC22A5	g	0.0278	0.59	1.27	1.16
rs526896	5	134384604	PITX1	t	0.0315	0.72	1.14	1.18
rs4282339	5	168188818	SLIT3	g	0.0352	0.81	1.08	1.21
rs12153391	5	171136043	FBXW11	c	0.0329	0.76	1.21	1.19
rs889014	5	172916720	BOD1	c	0.0290	0.67	1.15	1.17
rs422421	5	176449932	FGFR4/NSD1	c	0.0332	0.79	1.13	1.19
rs6879260	5	179663620	GFPT2	c	0.0281	0.60	1.24	1.16
rs12198986	6	7665058	BMP6	a	0.0359	0.46	1.34	1.21
rs806794	6	26308656	Histone cluster	a	0.0528	0.71	0.98	1.32
rs3129109	6	29192211	OR2J3	c	0.0257	0.64	0.90	1.15
rs2596530	6	31495352	MICA	g	0.0341	0.53	1.38	1.20
rs6457617	6	32771829	HLA locus	c	0.0238	0.52	1.38	1.14
rs2780226	6	34307070	HMGA1	c	0.0790	0.08	1.39	1.52
rs6457821	6	35510783	PPARD/FANCE	c	0.1210	0.98	0.98	1.90
rs12530016	6	44974300	SUPT3H/RUNX2	g	0.0305	0.80	1.55	1.18
rs310405	6	81857081	FAM46A	a	0.0300	0.52	1.08	1.17
rs7759938	6	105485647	LIN28B	c	0.0420	0.35	1.00	1.25
rs3757235	6	109818534	ZBTB24	c	0.0216	0.58	0.91	1.12
rs6915129	6	117629512	VGLL2	c	0.0216	0.60	1.13	1.12
rs1490384	6	126892853	C6orf173	t	0.0370	0.54	1.17	1.22
rs6569648	6	130390812	L3MBTL3	c	0.0358	0.24	1.26	1.21
rs7763064	6	142838982	GPR126	g	0.0445	0.72	1.31	1.27
rs543650	6	152152636	ESR1	g	0.0318	0.59	1.25	1.18

Table 2.1 (Continued)

Rsid	Chr	Pos	Closest gene	Effect allele	Effect size	Freq	Observed OR	Expected OR
rs12206717	6	158830686	TULP4	g	0.0487	0.95	0.91	1.30
rs798489	7	2768329	GNA12	c	0.0515	0.75	1.37	1.32
rs1708299	7	28156471	JAZF1	a	0.0417	0.34	1.43	1.25
rs6959212	7	38094851	STARD3NL	c	0.0229	0.68	1.24	1.13
rs42235	7	92086012	CDK6	t	0.0548	0.30	1.25	1.34
rs822552	7	148281567	PDIA4	g	0.0302	0.27	1.20	1.17
rs17088190	8	24167275	ADAM28	c	0.0278	0.75	0.91	1.16
rs6473015	8	78341040	PEX2	c	0.0320	0.32	1.25	1.19
rs6470764	8	130794847	GSDMC	c	0.0469	0.83	0.92	1.28
rs894345	8	135682763	ZFAT	c	0.0297	0.59	1.17	1.17
rs7864648	9	16358732	BNC2	t	0.0246	0.35	0.96	1.14
rs11144688	9	77732106	PCSK5	g	0.0548	0.87	1.14	1.34
rs296886	9	85781846	C9orf64	g	0.0250	0.19	1.04	1.14
rs181338	9	88297981	ZCCHC6	t	0.0234	0.53	1.20	1.13
rs2814828	9	90001002	SPIN1	t	0.0268	0.24	1.31	1.15
rs9969804	9	94468941	IPPK	a	0.0281	0.45	0.94	1.16
rs1257763	9	95933766	PTPDC1	a	0.0685	0.06	1.31	1.44
rs473902	9	97296056	PTCH1/FANCC	t	0.0741	0.93	1.01	1.48
rs7027110	9	108638867	ZNF462	a	0.0337	0.22	0.81	1.20
rs1468758	9	112846903	LPAR1	c	0.0258	0.76	1.03	1.15
rs751543	9	118162163	PAPPA	t	0.0287	0.69	1.13	1.17
rs7466269	9	132453905	FUBP3	a	0.0359	0.66	1.39	1.21
rs12338076	9	138261561	QSOX2	c	0.0304	0.29	0.95	1.18
rs7909670	10	12958770	CCDC3	c	0.0219	0.55	1.11	1.12
rs7332	10	80784066	PPIF	g	0.0252	0.50	1.30	1.14
rs11599750	10	101795432	CPN1	c	0.0230	0.66	1.12	1.13
rs2237886	11	2767307	KCNQ1	t	0.0429	0.11	1.16	1.26
rs7937898	11	12660137	TEAD1	g	0.0239	0.48	0.90	1.14
rs1330	11	17272605	NUCB2	t	0.0241	0.38	1.16	1.14
rs2904315	11	48066524	PTPRJ/SLC39A13	a	0.0311	0.30	1.04	1.18
rs1814175	11	49515748	FOLH1	t	0.0230	0.44	1.15	1.13
rs3782089	11	65093395	SSSCA1	c	0.0583	0.93	1.10	1.36
rs7112925	11	66582736	RHOD	c	0.0229	0.64	1.20	1.13
rs606452	11	74953826	SERPINH1	a	0.0397	0.16	1.12	1.24
rs494459	11	118079885	TREH	t	0.0207	0.43	0.99	1.12
rs654723	11	128091365	FLI1	a	0.0237	0.61	1.19	1.13
rs2954980	12	11750815	ETV6	t	0.0295	0.36	1.11	1.17

Table 2.1 (Continued)

Rsid	Chr	Pos	Closest gene	Effect allele	Effect size	Freq	Observed OR	Expected OR
rs10770705	12	20748734	SLCO1C1	a	0.0314	0.34	1.14	1.18
rs2638953	12	28425682	CCDC91	c	0.0356	0.69	1.53	1.21
rs2066807	12	55026949	STAT2	g	0.0520	0.08	1.68	1.32
rs1351394	12	64638093	HMGA2	t	0.0535	0.55	1.31	1.33
rs10748128	12	68113925	FRS2	t	0.0347	0.37	1.36	1.20
rs11107116	12	92502635	SOCS2	t	0.0524	0.21	1.68	1.32
rs12298826	12	122394981	SBNO1	g	0.0350	0.21	0.85	1.20
rs7332115	13	32045548	PDS5B/BRCA2	g	0.0250	0.39	1.10	1.14
rs3118906	13	50004789	DLEU7	g	0.0518	0.75	1.24	1.32
rs4773624	13	90817730	GPC5	g	0.0286	0.40	1.11	1.16
rs1950500	14	23900690	NFATC4	t	0.0323	0.24	1.55	1.19
rs10483727	14	60142628	SIX6	t	0.0322	0.38	1.25	1.19
rs6573834	14	67878151	RAD51L1	c	0.0253	0.80	1.01	1.14
rs862031	14	74061608	LTBP2	g	0.0224	0.64	1.42	1.13
rs10150088	14	91573329	TRIP11	t	0.0270	0.60	1.05	1.15
rs16964211	15	49317787	CYP19A1	g	0.0511	0.95	1.32	1.31
rs7178424	15	60167551	C2CD4A	c	0.0235	0.51	1.12	1.13
rs10152591	15	67835211	TLE3	a	0.0447	0.88	1.15	1.27
rs3759901	15	70298469	MYO9A	a	0.0555	0.02	0.94	1.34
rs5742915	15	72123686	PML	c	0.0308	0.47	0.91	1.18
rs11259936	15	82371586	ADAMTSL3	c	0.0419	0.48	1.30	1.25
rs16942341	15	87189909	ACAN	c	0.1335	0.98	1.26	2.04
rs4965598	15	98577137	ADAMTS17	c	0.0353	0.30	1.20	1.21
rs1659127	16	14295806	MKL2	a	0.0240	0.29	1.05	1.14
rs4640244	17	21224816	KCNJ12	a	0.0279	0.56	1.01	1.16
rs3110496	17	24941897	ANKRD13B	g	0.0229	0.67	0.81	1.13
rs3764419	17	26188149	ATAD5/RNF135	c	0.0374	0.62	1.12	1.22
rs17780080	17	27367259	LRRC37B	a	0.0344	0.17	1.01	1.20
rs1043515	17	34175722	PIP4K2B	g	0.0219	0.54	1.38	1.12
rs4986172	17	40571807	ACBD4	c	0.0283	0.68	1.02	1.16
rs11652146	17	44777362	ZNF652	g	0.0255	0.30	0.96	1.15
rs227723	17	52133903	NOG	t	0.0272	0.28	1.13	1.16
rs2079795	17	56851431	TBX2	t	0.0395	0.33	1.23	1.23
rs12325866	17	59109706	CSH1/GH1	a	0.0343	0.28	1.36	1.20
rs11867479	17	65601802	KCNJ16/KCNJ2	t	0.0240	0.36	1.24	1.14
rs4800452	18	18981609	CABLES1	t	0.0475	0.80	1.33	1.29
rs2078286	18	45132860	DYM	a	0.0372	0.41	1.20	1.22

Table 2.1 (Continued)

Rsid	Chr	Pos	Closest gene	Effect allele	Effect size	Freq	Observed OR	Expected OR
rs6567160	18	55980115	MC4R	c	0.0245	0.27	1.15	1.14
rs12980348	19	2132607	DOT1L	g	0.0323	0.36	1.30	1.19
rs891088	19	7135762	INSR	g	0.0251	0.23	0.91	1.14
rs4542783	19	8548160	ADAMTS10	t	0.0313	0.55	1.07	1.18
rs2279008	19	17144303	MYO9B	t	0.0308	0.75	1.04	1.18
rs17318596	19	46628935	ATP5SL	a	0.0290	0.38	1.31	1.17
rs1741344	20	4049800	SMOX	c	0.0263	0.39	1.14	1.15
rs2145272	20	6574218	BMP2	g	0.0386	0.34	1.25	1.23
rs7274811	20	31796842	ZNF341	g	0.0402	0.76	1.47	1.24
rs143384	20	33489170	GDF5	g	0.0639	0.41	1.42	1.41
rs1567865	20	47315374	ZNFX1	t	0.0337	0.22	1.15	1.20
rs2834440	21	34612369	KCNE2	a	0.0247	0.62	1.23	1.14
rs4821083	22	31386341	SYN3	t	0.0332	0.84	1.15	1.19

The table shows the results for the SNPs used in the individual association analysis in the HUNT cohort.

Table 2.2: Individual SNP analysis for FINRISK cohort

Rsid	Chr	Pos	Closest gene	Effect allele	Effect size	Freq	Observed OR	Expected OR
rs425277	1	2069172	PTCH1/FANCC	T	0.024	0.30	1.18	1.14
rs2284746	1	17306675	FAM46A	A	0.0354	0.52	1.52	1.21
rs1738475	1	23536891	NPR3	C	0.0216	0.61	1.14	1.12
rs4601530	1	24916698	FUBP3	A	0.0238	0.27	1.25	1.14
rs2154319	1	41745770	OR2J3	T	0.0335	0.74	1.18	1.20
rs17391694	1	78623626	SLC38A9	T	0.0399	0.12	2.31	1.24
rs6699417	1	89123443	SBNO1	T	0.0217	0.66	0.75	1.12
rs10874746	1	93323971	DNM3	T	0.0217	0.36	0.98	1.12
rs9428104	1	118855587	ADAMTS10	C	0.0375	0.22	1.09	1.22
rs11205277	1	149892872	TGFB2	G	0.0452	0.62	1.76	1.28
rs17346452	1	172053287	WWC2	A	0.038	0.77	1.37	1.23
rs1325598	1	175058872	RTF1	T	0.0256	0.45	1.12	1.15
rs1046934	1	184023529	SCMH1	C	0.0459	0.63	1.47	1.28
rs10863936	1	212237798	SF3B4	G	0.022	0.53	1.10	1.13
rs6684205	1	218609702	CCDC53/GNPTAB	A	0.0328	0.67	1.16	1.19
rs11118346	1	219743719	TSEN15	C	0.0264	0.50	1.30	1.15
rs10799445	1	227911883	SPAG17	A	0.0306	0.72	1.73	1.18
rs4665736	2	25187599	PPIF	T	0.0335	0.59	1.22	1.20
rs6714546	2	33361425	LTBP1	C	0.0254	0.28	1.53	1.15
rs17511102	2	37960613	CEP120	A	0.0601	0.90	2.05	1.38
rs2341459	2	44768202	ZBTB24	T	0.0276	0.31	1.59	1.16
rs3791675	2	56111309	BNC2	G	0.0496	0.25	1.25	1.31
rs11684404	2	88924622	DNAJC27	C	0.027	0.65	0.89	1.16
rs7567288	2	134151294	IGF1R	G	0.0309	0.75	1.05	1.18
rs1351164	2	217980143	RYBP	T	0.0279	0.75	1.62	1.16
rs12470505	2	219908369	NCKAP5	T	0.0483	0.88	1.37	1.30
rs2629046	2	225047744	GPR126	T	0.0247	0.55	1.11	1.14
rs2580816	2	232797966	C6orf173	C	0.0412	0.20	1.65	1.25
rs12694997	2	241911659	NPPC	C	0.0274	0.24	0.88	1.16
rs2597513	3	13555836	L3MBTL3	C	0.0392	0.88	0.77	1.24
rs13088462	3	51071713	DOCK3	C	0.0543	0.92	1.02	1.34
rs2336725	3	53093779	LIN28B	C	0.0263	0.55	1.48	1.15
rs9835332	3	56642722	GDF5	A	0.0217	0.48	1.03	1.12
rs9863706	3	72437413	KCNE2	A	0.0304	0.21	0.98	1.18
rs9844666	3	135974216	ZNFX1	G	0.0284	0.23	0.99	1.17
rs724016	3	142588510	C2CD4A	C	0.067	0.58	1.58	1.44
rs572169	3	172165727	SERPINH1	T	0.0355	0.33	1.06	1.21
rs720390	3	185548683	CYP19A1	A	0.0305	0.43	1.17	1.18
rs2247341	4	1671115	HMGA1	A	0.0251	0.40	0.96	1.14
rs6449353	4	18033488	ETV6	T	0.0714	0.87	1.40	1.47
rs17081935	4	57823476	TET2	T	0.0306	0.22	1.00	1.18
rs7697556	4	73515313	CTU2/GALNS	T	0.0219	0.49	1.49	1.13
rs10010325	4	106106353	PRKCZ	A	0.0214	0.45	1.12	1.12
rs7689420	4	145568352	MKL2	G	0.0687	0.19	1.81	1.45
rs955748	4	184215675	BMP2	A	0.0243	0.24	0.87	1.14

Table 2.2 (Continued)

Rsid	Chr	Pos	Closest gene	Effect allele	Effect size	Freq	Observed OR	Expected OR
rs1173727	5	32830521	LTBP1	T	0.0356	0.38	1.47	1.21
rs11958779	5	55001899	EFEMP1	T	0.0282	0.72	0.89	1.16
rs10037512	5	88354675	MFAP2	T	0.0267	0.58	1.45	1.15
rs1582931	5	122657199	SLBP/FGFR3	G	0.0254	0.49	0.93	1.15
rs274546	5	131727766	TULP4	A	0.0278	0.45	1.58	1.16
rs526896	5	134384604	SSSCA1	T	0.0315	0.72	1.27	1.19
rs4282339	5	168256240	ZNF462	G	0.0352	0.22	0.90	1.21
rs12153391	5	171203438	EIF2AK3	T	0.0329	0.28	1.31	1.19
rs889014	5	172984114	MC4R	T	0.029	0.40	1.17	1.17
rs422421	5	176517326	PTPDC1	G	0.0332	0.22	1.19	1.20
rs6879260	5	179731014	PDS5B/BRCA2	T	0.0281	0.38	1.38	1.16
rs3812163	6	7670759	PCSK5	A	0.0366	0.57	1.08	1.22
rs1047014	6	19949472	PKN2	C	0.0291	0.75	1.19	1.17
rs806794	6	26200677	KCNJ16/KCNJ2	A	0.0528	0.64	1.21	1.33
rs3129109	6	29084232	PEX2	A	0.0257	0.37	0.88	1.15
rs2256183	6	31380529	PPARD/FANCE	A	0.0345	0.37	1.29	1.20
rs2780226	6	34199092	TWISTNB	T	0.079	0.92	1.19	1.53
rs9472414	6	44946506	SMOX	T	0.0306	0.22	0.92	1.18
rs9360921	6	76265642	INSR	G	0.0479	0.85	1.83	1.29
rs310405	6	81800362	CDK6	A	0.03	0.48	1.40	1.18
rs7759938	6	105378954	KCNJ12	G	0.042	0.71	1.48	1.25
rs1046943	6	109783941	GIPC2	A	0.0223	0.53	0.96	1.13
rs961764	6	117522156	ZNF341	G	0.0228	0.41	1.47	1.13
rs1490384	6	126851160	GHSR	T	0.037	0.49	1.11	1.22
rs6569648	6	130349119	SOCS2	T	0.0358	0.76	1.26	1.21
rs7763064	6	142797289	ANKRD13B	G	0.0445	0.29	1.92	1.27
rs543650	6	152110943	RHOD	C	0.0318	0.47	1.34	1.19
rs9456307	6	158929442	MYO9B	T	0.0499	0.07	1.48	1.31
rs798489	7	2801803	NOG	A	0.0515	0.33	1.19	1.32
rs4470914	7	19616522	PAPPA	T	0.0328	0.17	1.22	1.19
rs12534093	7	23502974	TNS1	T	0.0298	0.20	1.24	1.17
rs1708299	7	28189946	HHIP	A	0.0417	0.32	0.95	1.25
rs6959212	7	38128326	DLEU7	G	0.0229	0.30	0.93	1.13
rs42235	7	92248076	SPIN1	T	0.0548	0.32	1.33	1.34
rs822552	7	148650634	DYM	G	0.0302	0.74	1.22	1.18
rs7460090	8	57194163	ADAMTSL3	T	0.0546	0.86	1.23	1.34
rs6473015	8	78178485	STAT2	G	0.032	0.74	1.41	1.19
rs6470764	8	130725665	CCDC91	C	0.0469	0.19	1.21	1.29
rs12680655	8	135637337	SERPINE2	C	0.0298	0.55	1.21	1.17
rs7864648	9	16358732	PIP4K2B	T	0.0246	0.35	1.18	1.14
rs11144688	9	78542286	DTL	A	0.0548	0.12	0.91	1.34
rs7853377	9	86552205	LRRC37B	A	0.0256	0.76	1.88	1.15
rs2778031	9	90835726	GNA12	T	0.0273	0.23	1.25	1.16
rs1257763	9	96893945	CCDC108/IHH	A	0.0685	0.06	1.17	1.45

Table 2.2 (Continued)

Rsid	Chr	Pos	Closest gene	Effect allele	Effect size	Freq	Observed OR	Expected OR
rs473902	9	98256235	CPN1	T	0.0741	0.90	1.08	1.49
rs7027110	9	108638867	GPC5	A	0.0337	0.23	0.98	1.20
rs751543	9	119122342	ADAMTS17	T	0.0287	0.73	1.23	1.17
rs7466269	9	133464084	ACAN	A	0.0359	0.60	1.03	1.21
rs7849585	9	138251691	ATAD5/RNF135	T	0.0324	0.32	1.06	1.19
rs7909670	10	12918764	ACBD4	C	0.0219	0.46	1.02	1.13
rs2145998	10	81121696	Histone	G	0.0252	0.56	0.98	1.15
rs11599750	10	101805442	JMJD4	A	0.023	0.34	1.17	1.13
rs2237886	11	2810731	MICA	T	0.0429	0.11	1.58	1.26
rs7926971	11	12698040	NME2	C	0.0244	0.58	1.16	1.14
rs1330	11	17316029	C3orf63	T	0.0241	0.33	1.07	1.14
rs1814175	11	49559172	FBXW11	T	0.023	0.30	1.31	1.13
rs3782089	11	65336819	ZFAT	C	0.0583	0.07	0.84	1.37
rs7112925	11	66826160	NFATC4	C	0.0229	0.32	1.05	1.13
rs634552	11	75282052	FLI1	T	0.0412	0.17	1.43	1.25
rs494459	11	118574675	TEAD1	T	0.0207	0.41	1.13	1.12
rs654723	11	128586155	HMGA2	A	0.0237	0.64	1.13	1.14
rs2856321	12	11855773	JAZF1	A	0.0298	0.65	1.22	1.17
rs10770705	12	20857467	RPL5	A	0.0314	0.30	1.29	1.18
rs2638953	12	28534415	ESR1	C	0.0356	0.70	0.96	1.21
rs2066807	12	56740682	FGFR4/NSD1	T	0.052	0.93	1.19	1.32
rs1351394	12	66351826	PCCB	T	0.0535	0.49	1.45	1.33
rs11107116	12	93978504	PAPPA2	T	0.0524	0.24	1.04	1.33
rs7971536	12	102373788	ZNF652	G	0.0247	0.41	0.78	1.14
rs11830103	12	122389499	CDC42EP3	T	0.0351	0.78	1.12	1.21
rs1809889	12	124801226	SLIT3	T	0.0315	0.31	1.33	1.19
rs7332115	13	33147548	PML	C	0.025	0.58	0.92	1.14
rs3118905	13	51105334	SDR16C5	T	0.052	0.33	1.31	1.32
rs7319045	13	92024574	MYO9A	A	0.029	0.45	1.49	1.17
rs1950500	14	24830850	BOD1	T	0.0323	0.30	1.06	1.19
rs1570106	14	68813115	IGF2BP2	G	0.0256	0.23	1.38	1.15
rs7155279	14	91555634	RAD51L1	C	0.0285	0.35	0.95	1.17
rs16964211	15	49317787	ADAMTS3	C	0.0511	0.09	1.39	1.32
rs7178424	15	62380259	TRIP11	T	0.0235	0.54	1.34	1.14
rs12902421	15	72161403	SEPT2	A	0.0691	0.95	0.96	1.45
rs5742915	15	74336633	TREH	T	0.0308	0.56	0.98	1.18
rs11259936	15	84580582	LYPLAL1	T	0.0419	0.47	1.82	1.25
rs16942341	15	89388905	POLR2B	C	0.1335	0.05	1.90	2.06
rs4965598	15	98577137	NUCB2	T	0.0353	0.68	1.17	1.21
rs2871865	15	99194896	STARD3NL	C	0.0535	0.88	1.45	1.33
rs1659127	16	14388305	LCORL	A	0.024	0.35	1.09	1.14
rs8052560	16	87304743	TBX2	A	0.0392	0.79	1.07	1.24
rs4640244	17	21284223	CCDC3	A	0.0279	0.59	1.24	1.16
rs3110496	17	24941897	PDIA4	C	0.0229	0.36	1.01	1.13
rs3764419	17	29164023	GSDMC	C	0.0374	0.39	1.18	1.22

Table 2.2 (Continued)

Rsid	Chr	Pos	Closest gene	Effect allele	Effect size	Freq	Observed OR	Expected OR
rs17780086	17	30343282	SLC22A5	A	0.0346	0.16	1.16	1.21
rs1043515	17	36922196	CLIC4	C	0.0219	0.46	1.14	1.13
rs4986172	17	43216281	FOLH1	C	0.0283	0.39	0.97	1.16
rs4605213	17	46599746	QSOX2	C	0.0234	0.32	1.21	1.13
rs2072153	17	47390014	GFPT2	C	0.0264	0.30	0.99	1.15
rs227724	17	52133816	SUPT3H/RUNX2	A	0.0272	0.61	1.07	1.16
rs2079795	17	59496649	BMP6	T	0.0395	0.30	1.58	1.24
rs11867479	17	68090207	C2orf34	T	0.024	0.34	0.97	1.14
rs9967417	18	46959500	SYN3	C	0.0381	0.62	0.98	1.23
rs17782313	18	57851097	PITX1	G	0.0249	0.80	1.15	1.14
rs12982744	19	2177193	HDAC11	C	0.0325	0.61	1.36	1.19
rs891088	19	7184762	DOT1L	G	0.0251	0.70	0.96	1.14
rs4072910	19	8644031	C9orf64	G	0.0289	0.48	1.18	1.17
rs2279008	19	17283303	SENP6	T	0.0308	0.68	1.45	1.18
rs1741344	20	4101800	MEF2C	C	0.0263	0.64	0.94	1.15
rs2145272	20	6626218	ID4	C	0.0386	0.69	1.42	1.23
rs7274811	20	32333181	TLE3	A	0.0402	0.24	1.29	1.24
rs143384	20	33489170	ZBTB38	G	0.0639	0.55	1.68	1.41
rs237743	20	47903019	VGLL2	A	0.0338	0.21	1.30	1.20
rs2834442	21	35690786	IGF2BP3	A	0.0269	0.67	1.16	1.16
rs4821083	22	31386341	KCNQ1	T	0.0332	0.84	1.09	1.20

The table shows the results for the SNPs used in the individual association analysis in the FINRISK cohort.

Table 2.3: Meta-analysis of individual SNPs for HUNT and FINRISK cohort

Rsid	Chr	Pos	Closest gene	Effect allele	Freq	Effect Size	Observed OR	P-value	Expected OR
rs425277	1	2069172	PRKCZ	t	0.28	0.02	1.19	0.0932	1.14
rs2284746	1	17306675	MFAP2	t	0.50	0.03	1.30	0.0045	1.19
rs1738475	1	23536891	HTR1D	t	0.59	0.02	1.15	0.1364	1.12
rs4601530	1	24916698	CLIC4	c	0.74	0.02	1.10	0.3495	1.14
rs2154319	1	41745770	SCMH1	g	0.23	0.03	1.27	0.0355	1.19
rs17391694	1	78623626	GIPC2	t	0.11	0.04	1.56	0.0008	1.24
rs6699417	1	89123443	PKN2	t	0.62	0.02	0.97	0.7864	1.12
rs10874746	1	93323971	RPL5	c	0.63	0.02	1.06	0.5108	1.12
rs9428104	1	118855587	SPAG17	c	0.76	0.04	1.19	0.1106	1.22
rs11205277	1	149892872	SF3B4	g	0.42	0.05	1.56	0.0000	1.27
rs17346452	1	172053287	DNM3	g	0.27	0.04	1.07	0.5121	1.21
rs1325598	1	175058872	PAPPA2	a	0.57	0.03	1.17	0.0992	1.14
rs1046934	1	184023529	TSEN15	c	0.36	0.05	1.32	0.0050	1.28
rs10863936	1	212237798	DTL	g	0.46	0.02	1.12	0.2176	1.12
rs6684205	1	218609702	TGFB2	g	0.29	0.03	1.15	0.1579	1.19
rs11118346	1	219743719	LYPLAL1	c	0.54	0.03	1.19	0.0702	1.15
rs4665736	2	25187599	DNAJC27	a	0.53	0.03	1.35	0.0017	1.19
rs6714546	2	33361425	LTBP1	g	0.72	0.02	1.36	0.0033	1.14
rs2341459	2	44768202	C2orf34	t	0.27	0.03	1.34	0.0040	1.16
rs3791675	2	56111309	EFEMP1	c	0.77	0.05	1.49	0.0002	1.30
rs11684404	2	88924622	EIF2AK3	c	0.33	0.03	1.11	0.2966	1.15
rs7567288	2	134151294	NCKAP5	c	0.20	0.03	1.08	0.4521	1.18
rs12470505	2	219908369	CCDC108/IHH	t	0.90	0.05	1.05	0.7572	1.29
rs2629046	2	225047744	SERPINE2	t	0.55	0.02	1.21	0.0364	1.14
rs12694997	2	241911659	SEPT2	g	0.76	0.03	1.00	0.9930	1.16
rs2597513	3	13555836	HDAC11	c	0.11	0.04	1.28	0.1184	1.23
rs13088462	3	51071713	DOCK3	c	0.06	0.05	1.32	0.1658	1.34
rs2336725	3	53093779	RTF1	c	0.46	0.03	1.14	0.1695	1.15
rs9835332	3	56642722	C3orf63	a	0.54	0.02	1.16	0.1039	1.12
rs9863706	3	72437413	RYBP	a	0.79	0.03	1.15	0.2250	1.18
rs9844666	3	135974216	PCCB	g	0.74	0.03	1.21	0.0929	1.16
rs724016	3	142588510	ZBTB38	g	0.43	0.07	1.55	0.0000	1.43
rs572169	3	172165727	GHSR	t	0.31	0.04	1.21	0.0543	1.21
rs720390	3	185548683	IGF2BP2	a	0.39	0.03	1.21	0.0528	1.18
rs2247341	4	1671115	SLBP/FGFR3	a	0.36	0.03	1.13	0.2082	1.14
rs6449353	4	18033488	LCORL	t	0.85	0.07	1.48	0.0111	1.46
rs17081935	4	57823476	POLR2B	t	0.19	0.03	1.01	0.9645	1.18
rs7697556	4	73515313	ADAMTS3	t	0.48	0.02	1.31	0.0036	1.12
rs10010325	4	106106353	TET2	a	0.49	0.02	1.17	0.0885	1.12
rs7689420	4	145568352	HHIP	c	0.84	0.07	1.45	0.0043	1.44
rs955748	4	184215675	WWC2	g	0.75	0.02	1.08	0.4925	1.14
rs1173727	5	32830521	NPR3	t	0.39	0.04	1.26	0.0176	1.21
rs11958779	5	55001899	SLC38A9	t	0.30	0.03	1.10	0.3480	1.16
rs10037512	5	88354675	MEF2C	t	0.56	0.03	1.29	0.0086	1.15

Table 2.3 (Continued)

Rsid	Chr	Pos	Closest gene	Effect allele	Freq	Effect Size	Observed OR	P-value	Expected OR
rs274546	5	131727766	SLC22A5	g	0.61	0.03	1.39	0.0005	1.16
rs526896	5	134384604	PITX1	t	0.73	0.03	1.18	0.1189	1.18
rs4282339	5	168256240	SLIT3	g	0.80	0.04	1.01	0.9540	1.21
rs12153391	5	171203438	FBXW11	c	0.75	0.03	1.25	0.0350	1.19
rs889014	5	172984114	BOD1	c	0.64	0.03	1.16	0.1339	1.17
rs422421	5	176517326	FGFR4/NSD1	c	0.78	0.03	1.15	0.2751	1.19
rs6879260	5	179731014	GFPT2	c	0.61	0.03	1.29	0.0067	1.16
rs3812163	6	7670759	BMP6	a	0.47	0.04	1.22	0.0320	1.21
rs806794	6	26200677	Histone	a	0.71	0.05	1.08	0.4548	1.33
rs3129109	6	29084232	OR2J3	c	0.60	0.03	0.89	0.2318	1.15
rs2256183	6	31380529	MICA	g	0.45	0.03	1.35	0.0016	1.20
rs2780226	6	34199092	HMGA1	c	0.08	0.08	1.32	0.1486	1.52
rs9472414	6	44946506	SUPT3H/RUNX2	g	0.79	0.03	1.26	0.0438	1.18
rs310405	6	81800362	FAM46A	a	0.53	0.03	1.20	0.0566	1.17
rs7759938	6	105378954	LIN28B	c	0.32	0.04	1.17	0.1087	1.25
rs1046943	6	109783941	ZBTB24	c	0.58	0.02	0.93	0.4634	1.12
rs961764	6	117522156	VGLL2	c	0.59	0.02	1.26	0.0156	1.12
rs1490384	6	126851160	C6orf173	t	0.50	0.04	1.15	0.1290	1.22
rs6569648	6	130349119	L3MBTL3	c	0.24	0.04	1.26	0.0383	1.21
rs7763064	6	142797289	GPR126	g	0.71	0.04	1.50	0.0002	1.27
rs543650	6	152110943	ESR1	g	0.60	0.03	1.28	0.0075	1.18
rs9456307	6	158929442	TULP4	g	0.94	0.05	1.01	0.9579	1.30
rs798489	7	2801803	GNA12	c	0.71	0.05	1.29	0.0106	1.32
rs1708299	7	28189946	JAZF1	a	0.31	0.04	1.22	0.0498	1.25
rs6959212	7	38128326	STARD3NL	c	0.68	0.02	1.12	0.2541	1.13
rs42235	7	92248076	CDK6	t	0.31	0.05	1.28	0.0188	1.34
rs822552	7	148650634	PDIA4	g	0.25	0.03	1.21	0.0714	1.17
rs6473015	8	78178485	PEX2	c	0.29	0.03	1.29	0.0147	1.19
rs6470764	8	130725665	GSDMC	c	0.79	0.05	1.01	0.9297	1.28
rs12680655	8	135637337	ZFAT	a	0.60	0.03	1.18	0.0854	1.17
rs7864648	9	16358732	BNC2	t	0.32	0.02	1.04	0.6868	1.14
rs11144688	9	78542286	PCSK5	g	0.89	0.05	1.05	0.7566	1.34
rs7853377	9	86552205	C9orf64	g	0.23	0.03	1.30	0.0289	1.14
rs2778031	9	90835726	SPIN1	t	0.24	0.03	1.29	0.0167	1.15
rs1257763	9	96893945	PTPDC1	a	0.04	0.07	1.27	0.2484	1.44
rs473902	9	98256235	PTCH1/FANCC	t	0.92	0.07	1.04	0.8135	1.48
rs7027110	9	108638867	ZNF462	a	0.23	0.03	0.87	0.2405	1.20
rs751543	9	119122342	PAPPA	t	0.71	0.03	1.17	0.1351	1.17
rs7466269	9	133464084	FUBP3	a	0.64	0.04	1.24	0.0246	1.21
rs7849585	9	138251691	QSOX2	c	0.33	0.03	0.99	0.9361	1.18
rs7909670	10	12918764	CCDC3	c	0.57	0.02	1.07	0.4500	1.12
rs2145998	10	81121696	PPIF	g	0.52	0.03	1.16	0.1120	1.14
rs11599750	10	101805442	CPN1	c	0.61	0.02	1.14	0.1950	1.13
rs2237886	11	2810731	KCNQ1	t	0.11	0.04	1.28	0.1129	1.26
rs7926971	11	12698040	TEAD1	g	0.46	0.02	0.99	0.9516	1.14

Table 2.3 (Continued)

Rsid	Chr	Pos	Closest gene	Effect allele	Freq	Effect Size	Observed OR	P-value	Expected OR
rs1330	11	17316029	NUCB2	t	0.35	0.02	1.13	0.2302	1.14
rs1814175	11	49559172	FOLH1	t	0.34	0.02	1.21	0.0438	1.13
rs3782089	11	65336819	SSSCA1	c	0.94	0.06	1.01	0.9532	1.36
rs7112925	11	66826160	RHOD	c	0.64	0.02	1.14	0.1905	1.13
rs634552	11	75282052	SERPINH1	a	0.14	0.04	1.26	0.0734	1.24
rs494459	11	118574675	TREH	t	0.40	0.02	1.04	0.6646	1.12
rs654723	11	128586155	FLI1	a	0.61	0.02	1.17	0.1041	1.13
rs2856321	12	11855773	ETV6	t	0.36	0.03	1.15	0.1499	1.17
rs10770705	12	20857467	SLCO1C1	a	0.33	0.03	1.19	0.0721	1.18
rs2638953	12	28534415	CCDC91	c	0.68	0.04	1.28	0.0133	1.21
rs2066807	12	56740682	STAT2	g	0.08	0.05	1.49	0.0334	1.32
rs1351394	12	66351826	HMGA2	t	0.49	0.05	1.36	0.0009	1.33
rs11107116	12	93978504	SOCS2	t	0.22	0.05	1.34	0.0106	1.32
rs11830103	12	122389499	SBNO1	g	0.22	0.04	0.95	0.6298	1.21
rs1809889	12	124801226	FAM101A	t	0.29	0.03	1.28	0.0150	1.18
rs7332115	13	33147548	PDS5B/BRCA2	g	0.38	0.03	1.03	0.7892	1.14
rs3118905	13	51105334	DLEU7	g	0.71	0.05	1.27	0.0230	1.32
rs7319045	13	92024574	GPC5	g	0.39	0.03	1.24	0.0188	1.16
rs1950500	14	24830850	NFATC4	t	0.30	0.03	1.33	0.0075	1.19
rs1570106	14	68813115	RAD51L1	c	0.79	0.03	1.15	0.2214	1.14
rs7155279	14	91555634	TRIP11	t	0.62	0.03	1.01	0.9025	1.15
rs16964211	15	49317787	CYP19A1	g	0.95	0.05	1.35	0.1532	1.31
rs7178424	15	62380259	C2CD4A	c	0.54	0.02	1.19	0.0699	1.13
rs12902421	15	72161403	MYO9A	a	0.03	0.06	0.94	0.8741	1.34
rs5742915	15	74336633	PML	c	0.47	0.03	0.94	0.4832	1.18
rs11259936	15	84580582	ADAMTSL3	c	0.52	0.04	1.49	0.0000	1.25
rs16942341	15	89388905	ACAN	c	0.97	0.13	1.43	0.2533	2.04
rs4965598	15	98577137	ADAMTS17	c	0.32	0.04	1.19	0.0952	1.21
rs1659127	16	14388305	MKL2	a	0.34	0.02	1.07	0.5156	1.14
rs4640244	17	21284223	KCNJ12	a	0.61	0.03	1.09	0.3451	1.16
rs3110496	17	24941897	ANKRD13B	g	0.67	0.02	0.89	0.2488	1.13
rs3764419	17	29164023	ATAD5/RNF135	c	0.61	0.04	1.14	0.1612	1.22
rs17780086	17	30343282	LRRC37B	a	0.15	0.03	1.06	0.6451	1.20
rs1043515	17	36922196	PIP4K2B	g	0.54	0.02	1.28	0.0067	1.12
rs4986172	17	43216281	ACBD4	c	0.65	0.03	1.00	0.9611	1.16
rs2072153	17	47390014	ZNF652	g	0.31	0.03	0.98	0.8110	1.15
rs227724	17	52133816	NOG	t	0.32	0.03	1.10	0.2980	1.16
rs2079795	17	59496649	TBX2	t	0.33	0.04	1.34	0.0032	1.23
rs11867479	17	68090207	KCNJ16/KCNJ2	t	0.35	0.02	1.14	0.1713	1.14
rs9967417	18	46959500	DYM	a	0.42	0.04	1.12	0.2481	1.22
rs17782313	18	57851097	MC4R	c	0.24	0.02	1.15	0.2121	1.14
rs12982744	19	2177193	DOT1L	g	0.41	0.03	1.32	0.0043	1.19
rs891088	19	7184762	INSR	g	0.26	0.03	0.93	0.5104	1.14
rs4072910	19	8644031	ADAMTS10	t	0.56	0.03	1.12	0.2580	1.18
rs2279008	19	17283303	MYO9B	t	0.75	0.03	1.20	0.0657	1.18

Table 2.3 (Continued)

Rsid	Chr	Pos	Closest gene	Effect allele	Freq	Effect Size	Observed OR	P-value	Expected OR
rs1741344	20	4101800	SMOX	c	0.37	0.03	1.07	0.4895	1.15
rs2145272	20	6626218	BMP2	g	0.35	0.04	1.31	0.0099	1.23
rs7274811	20	32333181	ZNF341	g	0.77	0.04	1.40	0.0018	1.24
rs143384	20	33489170	GDF5	g	0.42	0.06	1.52	0.0000	1.41
rs237743	20	47903019	ZNFX1	t	0.21	0.03	1.20	0.0900	1.20
rs2834442	21	35690786	KCNE2	a	0.62	0.02	1.20	0.0613	1.14
rs4821083	22	31386341	SYN3	t	0.83	0.03	1.13	0.3619	1.19

The table shows the results for the SNPs used in the meta-analysis of the HUNT and FINRISK cohorts.

specific allele frequencies (see Materials and Methods). Overall, the number of SNPs with observed odds ratio greater than expected odds ratios was no different than expectation under the model of equal effect sizes in extremes and the general population (HUNT 79/160 SNPs, p=0.94; FINRISK 75/155 SNPs, p=0.48 and combined 75/141, p=0.45); (Table 2.1; Table 2.2 and Table 2.3). Next, for each SNP we tested for a difference between the expected and observed odds ratio in the individual studies and in the meta-analysis. Overall there were no more or fewer significant associations than would be expected under the equal effect size model (Figure 2.1). This result demonstrates that the individual SNPs have similar effects at the extremes as in the general population.

Weighted Allele Score (*WAS*) analysis: The additive effect of the common variants differs significantly from expected in the short extremes

After determining that the individual SNPs have similar effects at the extremes of the height distribution as in the general population, we then performed additional analyses on the combined set of height-associated variants. We asked whether extremely short and extremely tall individuals show overall enrichment of height-decreasing and height-increasing alleles, respectively, to the extent expected under a purely polygenic additive model. If the enrichment is less than expected, this result would suggest that the common variants are not explaining as much of the phenotypic variation in the extremes as in the general population. To test this possibility, we first calculated the weighted allele score (*WAS*) for each individual using the height-associated SNPs previously described. The *WAS* is the cumulative effect of all of the SNPs on height weighted by each SNP's estimated effect size (β). In Figure 2.2, we show a plot of each individual's *WAS* based on the 143 loci genotyped in both cohorts versus the individual

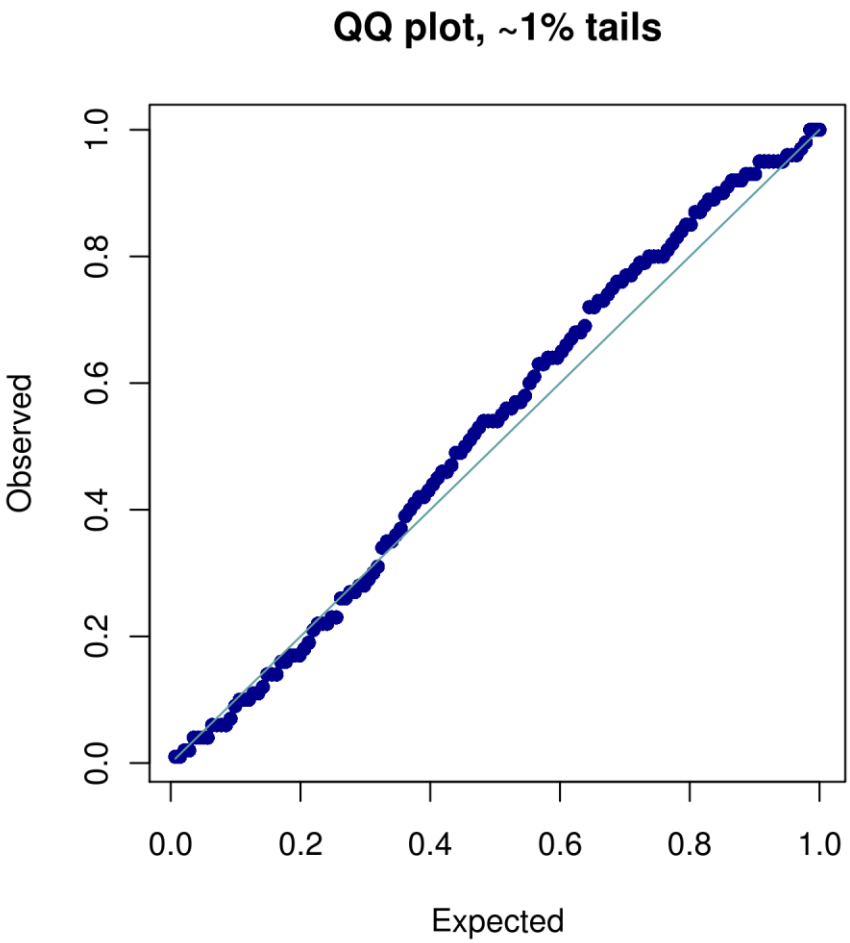


Figure 2.1: QQ Plot of p-values for individual SNPs based on the meta-analysis of HUNT and FINRISK. The figure shows a Q-Q plot of the p-values of the difference between the observed odd-ratios and the expected odd-ratios.

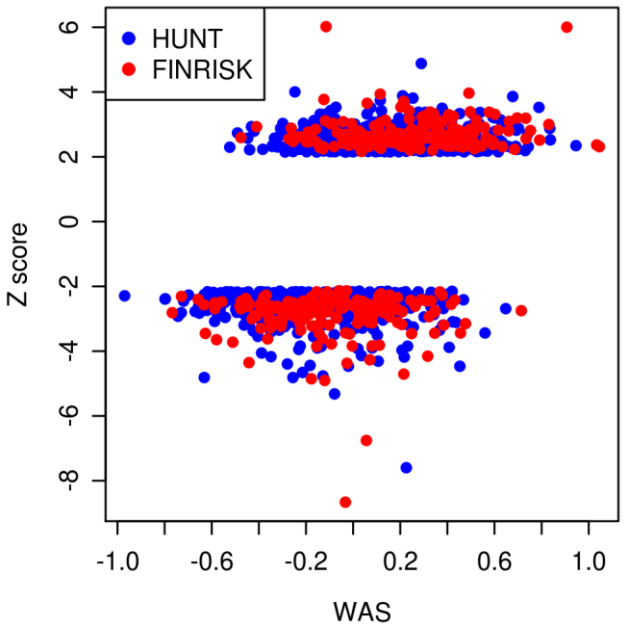


Figure 2.2: Plot of weighted allele scores (*WAS*) against Height Z-scores for HUNT and FINRISK Cohorts. The plot shows the WAS, a measure of the genetic prediction of height by known common variants, against the height Z-scores. The tall individuals (Z-score > 2.14) have generally larger WAS than the short individuals (Z-score < -2.14). Individuals from the HUNT study are labeled blue and individuals from the FINRISK study are labeled red.

height Z-scores. As expected, the *WAS* are significantly different between the tall extremes and the short extremes ($p<3x10^{-86}$), with individuals in the tall extreme having higher *WAS* on average than individuals in the short extremes.

We then tested whether the *WAS* in the short and tall groups are within expectations based on the population specific allele frequencies and previously estimated effect sizes of these SNPs, assuming a purely polygenic model. To generate the distribution of *WAS* under these expectations, we simulated populations that mimicked our ascertainment of extreme samples from the HUNT and FINRISK populations (see Materials and Methods). For each cohort, we compared the observed mean *WAS* with the distribution of mean *WAS* under the simulated model (Figure 2.3 and Figure 2.4). For the HUNT study the sample of 1224 individuals from the middle of the distribution suggest our modeling is behaving as expected (Figure 2.3). Finally, we analyzed the data by combining both studies using the 143 SNPs present in both data-sets (Figure 2.5). In each study separately and in the combined analysis, the mean observed *WAS* for the tall individuals was within expectation, but we observed a significant upward deviation of the mean observed *WAS* in the short extremes (p=0.006 for the combined-analysis). These results suggest that the collective effect of the common variants in the short extremes do not account for as much of the phenotypic variation in height as predicted from the effects seen in the general population.

The reduced effect of common variants is limited to the most extreme short individuals

Having established that the common variants do not explain as much phenotypic variation in the short extremes, we then sought to determine if this finding was accentuated in

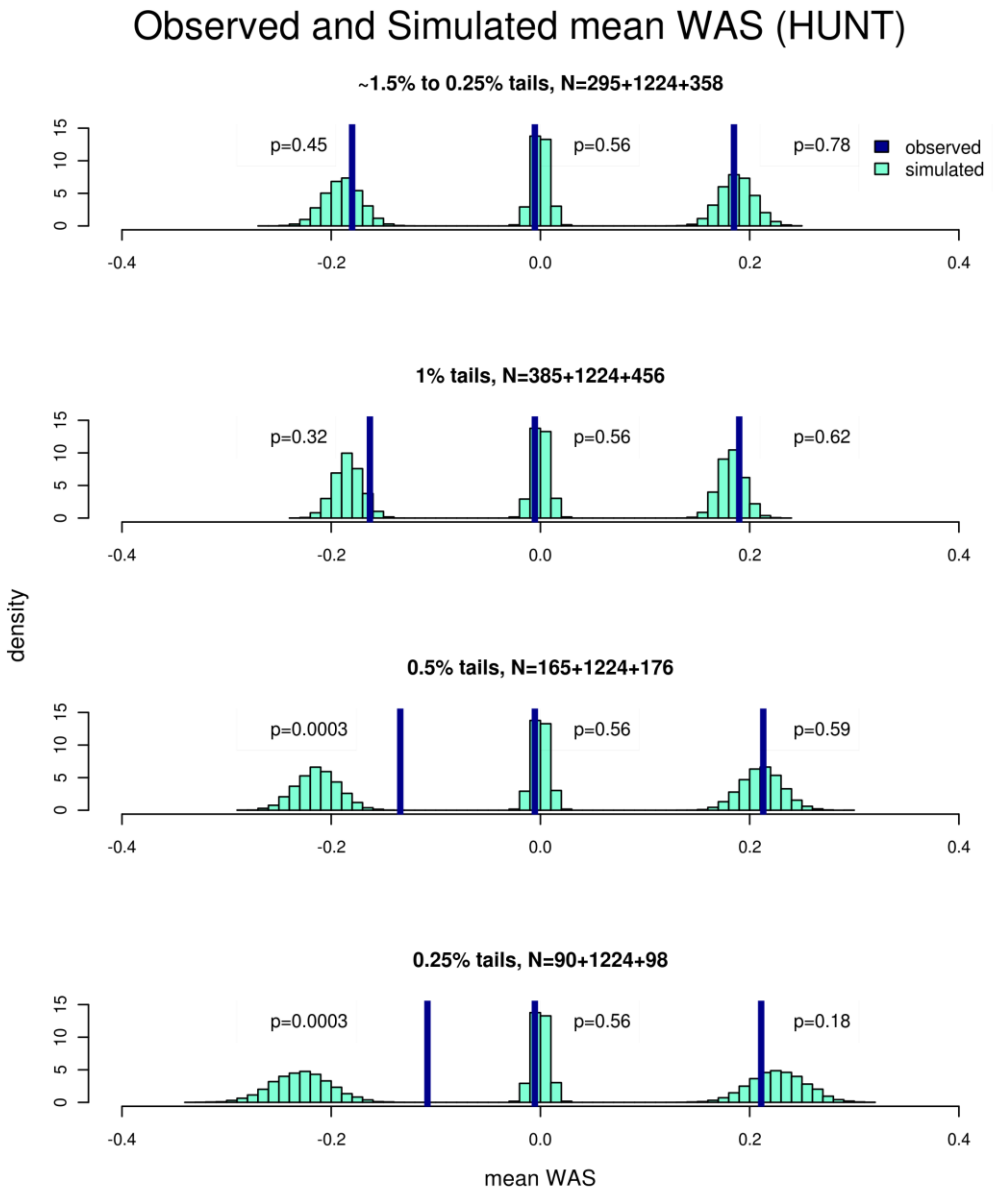


Figure 2.3: Comparison of the observed versus simulated mean weighted allele score (*WAS*) in the HUNT study. The plot shows the result of comparing the mean *WAS* of the short and tall individuals observed in the HUNT cohort against that obtained from simulation. Each row represents a different stratification of the extremes identical to those defined in Figure 2.5. The plot also show the mean *WAS* of 1224 non-extreme individuals taken from the middle of the height distribution. There is no difference between the mean *WAS* of the non-extreme individuals from that obtained from simulation (p=0.56).

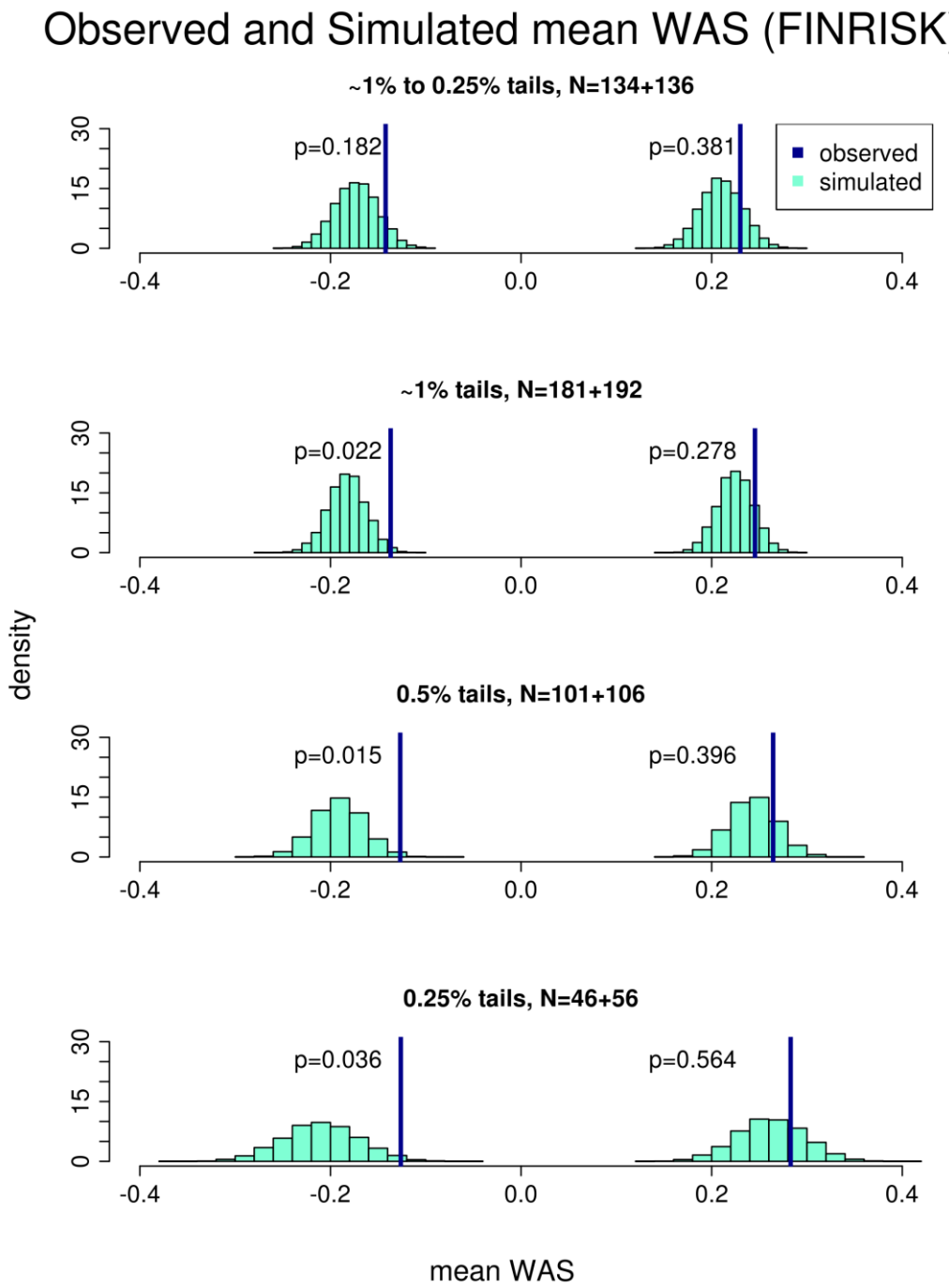


Figure 2.4: Comparison of the observed versus simulated mean weighted allele score (*WAS*) in the FINRISK study. The plot shows the result of comparing the mean *WAS* of the short and tall individuals observed in the FINRISK cohort against that obtained from simulation. Each row represents a different stratification of the extremes identical to those defined in Figure 2.5.

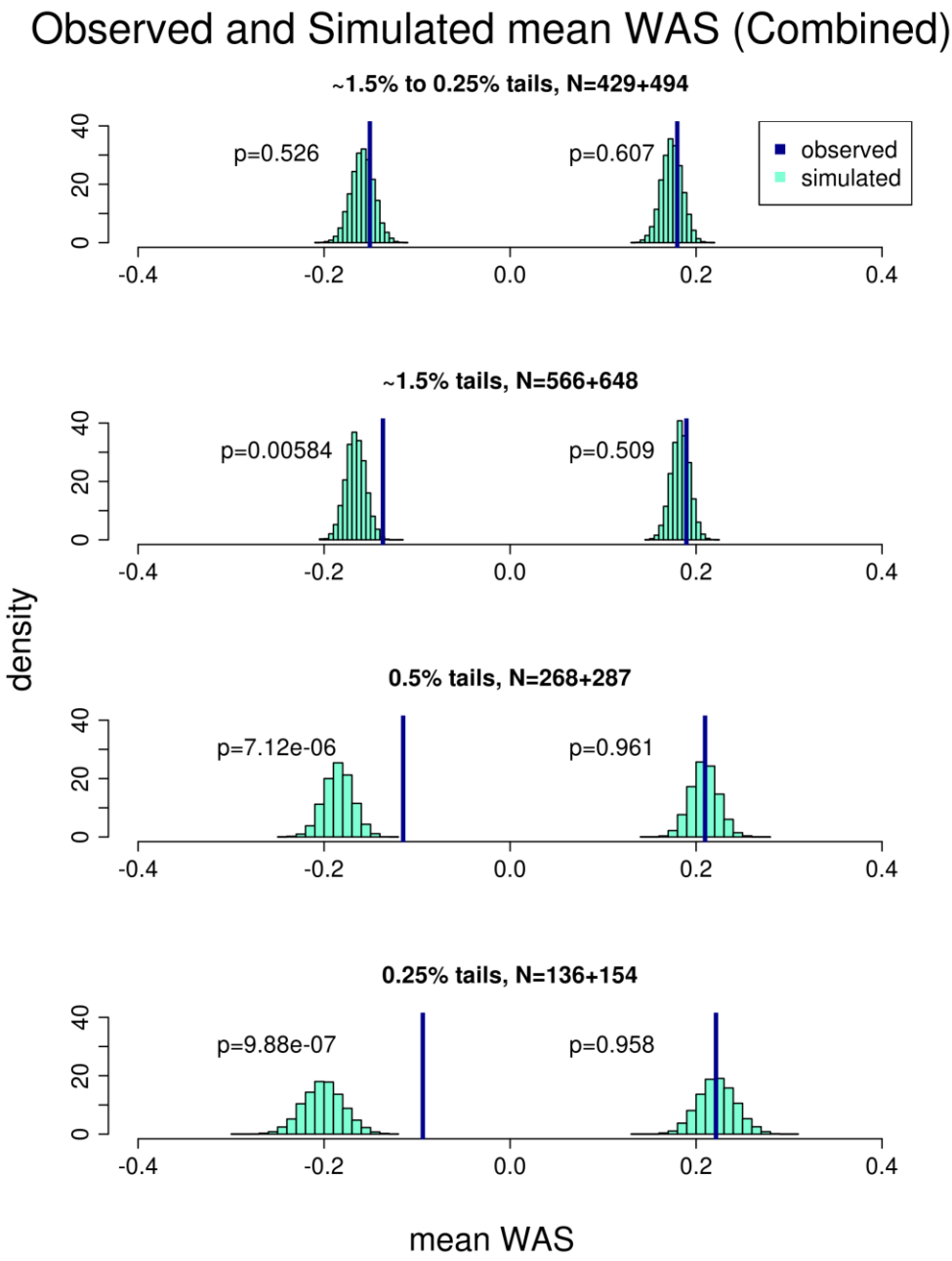


Figure 2.5: Comparison of the observed versus simulated mean weighted allele score (*WAS*) in the combined cohort. The plot shows the result of comparing the mean *WAS* of the short and tall individuals observed from both the HUNT and FINRISK cohorts against that obtained from simulation. Each row represents a different stratification of the extremes. The percentiles and numbers of individuals in the short and tall extreme respectively are listed for each stratum. The p-values represent the comparison between the observed and simulated mean *WAS*. The observed mean *WAS* for the tall individuals were not different from the simulation in any of the strata. The observed mean *WAS* for the short individuals was not different from the simulation in the first stratum. As a progressively more extreme sample is used, the short individuals' mean *WAS* becomes progressively more significantly different than the simulation.

individuals with the most extreme short stature. We stratified our analysis in several ways (Figure 2.5; Figure 2.3; Figure 2.4). First, we removed the most extreme individuals: those below the 0.25 percentile and above the 99.75 percentile. In the combined cohorts, the mean observed *WAS* in the short extremes was no longer significantly different than expected (p=0.526), indicating that the shift in *WAS* is driven by the most extremely short individuals. To further explore this hypothesis, we then selected more extreme individuals at two thresholds, including only the top and bottom 0.5% or 0.25% of the population (see Materials and Methods). For both strata, there was a more pronounced deviation of the mean observed *WAS* in the short extremes (p=7.12 x 10^{-6} and p=9.88 x 10^{-7} for the 0.5% and 0.25% extremes respectively), but again no deviation in the tall extremes. Similar observations occurred when we analyzed the cohorts separately using the same stratification procedure (Figure 2.3; Figure 2.4). We repeated the analysis using Z-scores based on inverse normal transformation, and with the three -6 SD outliers removed, and the results were essentially unchanged. The difference observed in the *WAS* analysis is also supported by the individual SNP analysis: when we performed the combined analysis described above for the 0.25% extremes rather than the entire cohort, 60% (84/139) of the SNPS have an observed effect size smaller than expected (p=0.02) (data not shown). This analyses clearly suggest that the initial marginally significant shift of the mean observed *WAS* in the short extremes is primarily driven by the most extreme short individuals. Therefore, in general, as one selects individuals with more extreme short stature, in particular those with heights below the 0.25 percentile, the common variants play a much smaller role in explaining stature, indicating that there must be other factors contributing to the phenotypic variation in these extremely short individuals.

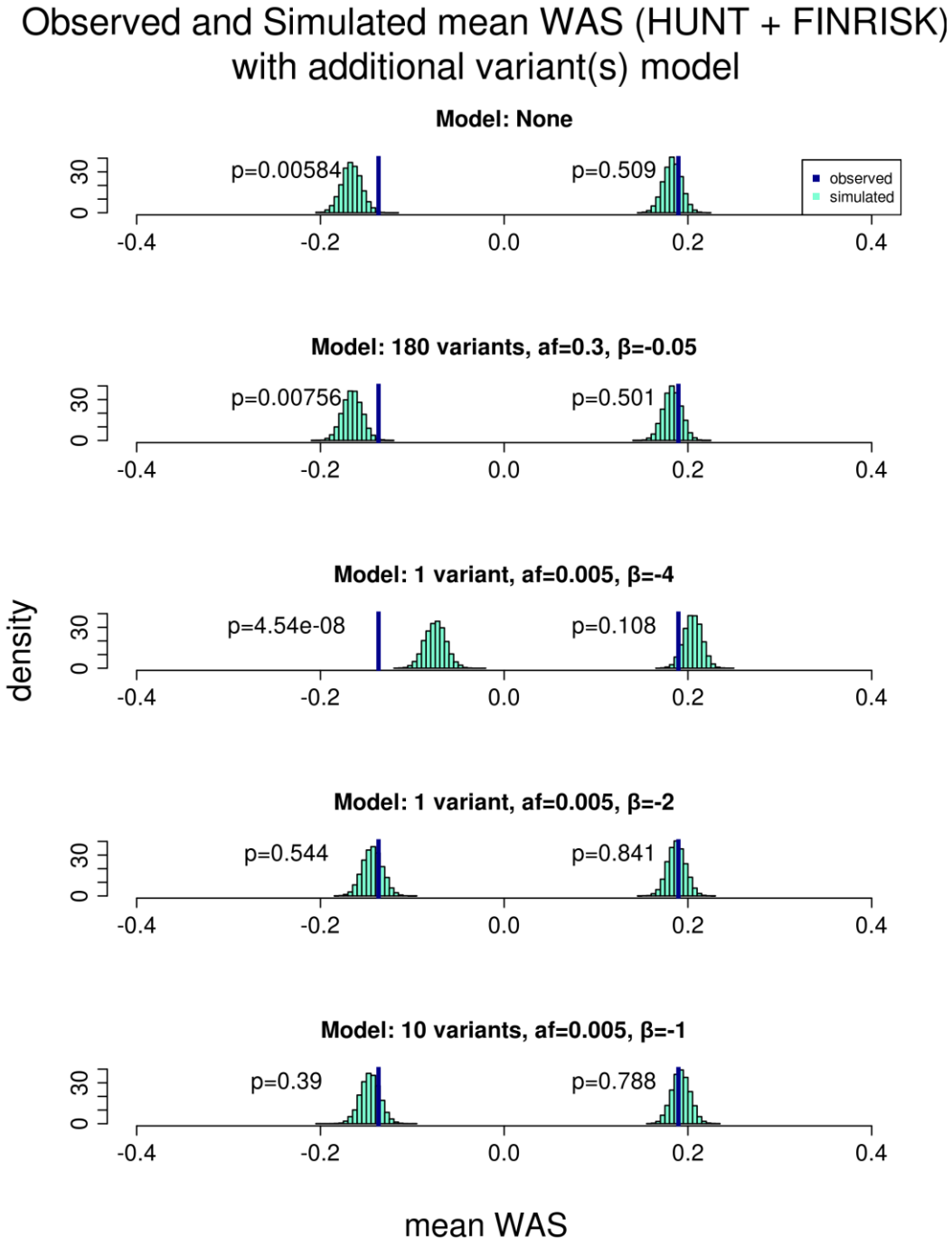
Low frequency or rare variants with larger effect sizes could explain the phenotypic variation in the short extremes

We hypothesized that lower frequency and rare genetic variants with larger effect sizes than the common variants may explain the phenotypic variation in the short extremes. To test this hypothesis, we performed population simulations with rare-variants of various allele frequencies and effect sizes, and asked if our observed data were consistent with these simulated scenarios (Figure 2.6). As a negative control, we first modeled an additional 180 SNPs, each with allele frequency of 0.3 and average effect sizes of -0.05 SD, which is similar to the allele frequency and effect size for previously discovered common variants associated with height. In this simulation, the mean *WAS* distribution did not change, indicating that adding additional common variants of similar effect sizes cannot explain the phenotypic variation in the short extremes. We then modeled a single rare variant of very large effect: frequency 0.005 and effect size of -4 SD. In this model, the mean *WAS* distribution in the extremely short individuals shifts more than we observed in our population. This simulation essentially excludes the possibility of a 0.5% variant of very large effect within our cohort. Such a variant would also be likely to be discovered in linkage studies of several thousand sib-pairs [6].

However, there are several rare variant models that would likely not have been detected in previous linkage analyses of height and generate a shift in the mean *WAS* consistent with our observed data (Figure 2.6). One such possibility is a single low frequency variant (allele frequency = 0.005) with an effect size of -2 SD; another model consistent with our data includes 10 variants each with an allele frequency of 0.005 and a moderate effect size of -1 SD. These simulations suggest that individuals with very short stature may harbor small numbers of low frequency variants of moderately large effect or a greater number of low

Figure 2.6: Comparison of the observed versus simulated mean *WAS* with models incorporating additional variants. The plot shows the result of comparing the mean *WAS* of the short and tall individuals observed from both the HUNT and FINRISK cohorts against that obtained from simulation with different scenarios of additional variants. All rows use the approximate 1.5% tails of the height distribution as extremes, resulting in 566 short and 648 tall individuals. The 1st row shows the result where the model has no additional variants affecting height and thus is identical to that from the 2nd row of Figure 2.5. The 2nd row shows a model where there are 180 additional common variants that slightly decreases height (allele frequency = 0.3 and effect size (β) = -0.05). This model does not result in any significant change to the simulated *WAS* of the short individuals and the observed *WAS* is still significantly different (p=0.00756). The 3rd row shows a model where there is 1 additional low frequency variant with a large height decreasing effect (allele frequency = 0.005 and effect size (β) = -4). This model results in a large shift in the simulated *WAS* of the short individuals to the right. The observed *WAS* is still significantly different (p=4.54 x 10^{-8}) than the simulation but in the opposite direction and thus is not consistent with our data. The 4th row shows a model where there is 1 additional low frequency variant that decreases height significantly (allele frequency = 0.005 and effect size (β) = -2). This model results in a shift in the simulated *WAS* of the short individuals to the right such that the observed *WAS* is no longer different from the simulation (p=0.544). The 5th row shows a model where there are 10 additional low frequency variants that moderately decreases height (allele frequency = 0.005 and effect size (β) = -1). This model also results in a shift in the simulated *WAS* of the short individuals to the right such that the observed WAS is no longer different from the simulation (p=0.39). The final two models are consistent with our observed data.

Figure 2.6 (Continued)



frequency variants of moderate effects contributing to their short stature. This result stands in contrast to the remainder of the height distribution in which a polygenic effect of common and rare variants with small effects could explain the majority of the heritability of height, even though only a small percentage of height-associated common variants have been identified.

Sibling analysis provides support for a different genetic architecture in extreme short individuals

To provide further support for a different genetic architecture in individuals in the extreme short tails we performed an analysis in siblings from the HUNT study. We queried the entire HUNT database (N=106,455) and identified 21,365 siblings pairs. The correlation of age and gender adjusted height between siblings was high (r = 0.466). We then identified 98 individuals (aged between 20-70yrs) with a Z-score < -2.81 (~0.25% tails) and 80 with a Z-score > 2.81 who also had at least one sibling in the database (the results are similar if we use inverse normal transformation). The average height Z-score for the siblings of the extreme short group was -0.97 (95% CI: -0.80, -1.15); the average Z-score for the full siblings of the extreme tall group was 1.29 (95% CI: 1.14, 1.45) which are significantly different (t-test, p=0.007 after reversing signs for the short group). We then performed this same analysis for the 0.25% to 1.5% tails individuals and there was no significant difference in z-scores of siblings between the short (-1.05 95% CI: -1.13, -0.97) and tall (1.11 95% CI: 1.03, 1.18) groups (t-test, p=0.28). So the differential regression to the mean appears to be limited to the shortest ~0.25% of individuals with this group regressing more quickly than the tall extreme group. This is consistent with the results we observe with the weighted allele score (*WAS*) approach. We do not have the twin data that would allow us to separate out the environmental and genetic effects in this group and our

data is consistent with both. If the effect were due to genetics, then a model with *de novo* mutations and/or multiple recessive rare variants could cause an increased regression to the mean in extremely short individuals, although there are other plausible explanations.

DISCUSSION

We have assessed whether common variants robustly associated with height in the general population also associate with height at the extreme tails of the height distribution. We further tested whether this association is to the extent expected under a purely polygenic model. By genotyping ~160 height SNPs identified from the GIANT study [2] (that explain ~10% of the population variation in height) in individuals from the ~1% tails of height from two large population based cohorts, we have shown that the polygenic model can explain the associations in the ~1% tails of height. However, our data indicate that the polygenic model starts to break down in extreme short individuals near the 0.25 percentile cut off. This conclusion is supported by our sibling analysis, which demonstrated that siblings in the 0.25% short tail regress to the mean more than those in the 0.25% tall group. Interestingly, the overall height distribution also shows a slight asymmetric deviation from normality, with an excess of individuals with extremely short stature but not for extremely tall stature.

While in general the individuals in the ~1% tails carry as many height increasing alleles as would be predicted based on their height, there was a clear deviation for individuals in the shortest 0.25% tail. On average, these individuals carry significantly more “tall” alleles at the 160 SNPs than would be expected if common alleles were explaining their short stature. This suggests that the heights of these individuals are explained by factors other than common variants. Our simulations suggest that rare variants could explain this difference in the 0.25%

shortest tail. For example, 10 rare variants with modest effects on height (1SD) are consistent with our observed data, as is a single variant with a 2SD effect. The sibling analysis also suggests a role for *de novo* or multiple recessive variants in the extreme short individuals. While rare height-decreasing variants of large effect are a plausible explanation, there are many other genetic models consistent with our data, including a mixture of height-decreasing with a smaller number of height-increasing rare variants, or variants having non-additive effects. While non-additive genetic effects could explain the data, no evidence was found for dominance or gene-gene interaction effects for the SNPs used in this study in the original GIANT publication [2]. It is also possible that these individuals are short for non-genetic reasons. One could suggest that these individuals are short because of differences in ancestry, but we have taken steps to remove any possible ethnic outliers from our extremes (see Materials and Methods). Measurement or recording error is another possibility, although the fact that the tall group does not show this effect (which presumably is equally likely to contain measurement error as the short group) suggests this is an unlikely explanation. Non-genetic factors could also be a possibility, for example, poor early-life nutrition, severe infection, or other chronic childhood diseases could have prevented these individuals from reaching their genetic height potential.

This result also suggests that these families would be good candidates to investigate in sequencing studies, as they may be enriched for rare or de novo, higher penetrance alleles. More generally, the weighted allele score (*WAS*) method developed here could be used to select individuals to sequence in the search for these types of rarer variants, not only for height but also for other polygenic traits and diseases. Specifically, individuals in the extreme tails of a trait distribution who have an unexpectedly high or low weighted allele score may be particularly useful to sequence, especially if multiple relatives with these characteristics were present in the

extreme tails.

Our study also demonstrates empirically that selecting individuals from the extreme tails of a complex trait distribution is an efficient approach for genetic studies, as was proposed both for linkage studies [7,8] and association studies [9,10]. Despite a quite modest sample size (N<1000), we replicated a large fraction of the individual SNPs identified in the GIANT study in our extreme height analysis. Ninety-one percent of the SNPs had odds ratios that were directionally consistent with the direction in the published GIANT study (p<0.0001), and 35% (49/141) of SNPS had p<0.05 in the consistent direction. Our analyses also demonstrate that, outside of the 0.25% tails, this level of association is entirely consistent with that expected given the extreme tail ascertainment of our samples and the individual SNP continuous distribution effect sizes. Given this result, the ascertainment of our 923 samples from the ~1% to 0.25% tails provides equivalent power to approximately 6000 samples randomly selected from the general population for a variant explaining approximately 0.1% of the variation in height. Indeed, the ability to detect associations in samples ascertained for extreme phenotypes has been recently demonstrated in studies of bone mineral density [11], body mass index [12], triglyceride levels [13], and type 2 diabetes (using a liability threshold model [14]). Also, our results suggest that the statistical power of detecting these small effect variants would be reduced if we were to include the most extreme tails of the phenotypic distribution (in our case, the shortest 0.25% of individuals), consistent with predictions made based on simulation studies of mixtures of common and rare variants [15]. Nonetheless, our findings suggest that the use of individuals with the most extreme phenotypes could be particularly valuable to detect rarer variants with larger effect sizes more efficiently.

In conclusion, we have shown that common genetic variants associated with height in the

general population are also associated with height at the ~1% tails of the height distribution. Our data suggest that common variants play less of a role, and the effect of rarer larger-effect alleles and/or strong environmental factors start to predominate around the 0.25% extreme. This finding may also have broader implications for studies of disease, in that the polygenic model may apply well to those diseases that represent the tails of an underlying normal distribution, but perhaps less well to diseases that correspond to more extreme phenotypes.

MATERIALS AND METHODS

Ethics statement

Both studies were conducted according to the principles expressed in the Declaration of Helsinki. Attendance was voluntary, and each participant signed a written informed consent including information on genetic analyses. Local institutional review boards approved study protocols.

Subjects

The HUNT study

The Nord-Trøndelag Health Study (HUNT) is a comprehensive population based health study (www.ntnu.edu/hunt) with personal and family medical histories on approximately 120,000 people from Nord-Trøndelag County, Norway, collected during three intensive studies (HUNT 1, 2, and 3). Inviting all citizens aged 20 and over, information was collected from self-reported questionnaires consisting of >200 health-related questions, standardized clinical

examinations, urine and non-fasting venous blood sample. The population in Nord-Trøndelag County is ethnically homogeneous, <3% of non-Caucasian ethnicity, making it especially suitable for epidemiological genetic research. Height was measured by trained personnel to the nearest 1.0 cm with the participants wearing light clothes without shoes according to standardized methods [16].

For this study we sourced data from HUNT 2 (1995-97) in which 65,258 individuals participated (71.2% of invited). We generated age and gender standardized height for the whole population, and selected the shortest 1000 individuals and the tallest 1000 individuals from the 54,909 participants aged between 18 and 70yrs. We removed known 1st degree relatives based on information from the Medical Birth Registry of Norway, those reporting to be living outside of Norway their first year of life, and those with low DNA concentrations. We then genotyped the remaining shortest 471 individuals (<-2.14 SDs) and the tallest 479 individuals (>2.14 SDs) from the cohort. We also genotyped 1,458 individuals of all ages with a Z-score between +/- 2 SDs as our middle group.

The FINRISK Study

FINRISK is a Finnish national survey on risk factors of chronic and non-communicable diseases. It is carried out every five years since 1972 using independent, random and representative population samples from different parts of Finland [17]. For this study, we selected individuals from 4 different sub-populations divided by geography (East vs. West Finland) and gender (Table 2.4). Individuals aged 25 to 74 years were included. We then took approximately the tallest and shortest 50 individuals (Table 2.4) from each tail of the distribution

Table 2.4: The FINRISK cohort divided into 4 sub-populations

Cohort	No. of Individuals	No. of short extremes	No. of tall extremes
men/west	4271	53	51
men/east	6582	52	52
women/west	5025	52	52
women/east	7610	52	52
Total	23488	209	207
Total successfully genotyped		186	192
Total with genotypes used		181	192

The table shows the number of individuals used for each of the FINRISK sub-populations. The FINRISK cohort is sub-divided between male and female as well as individuals from east and west Finland.

from each sub-population (extremes) and performed genotyping.

Genotyping and Quality Control

HUNT study

Blood sampling was done whenever subjects attended HUNT 2. DNA was extracted from peripheral blood leukocytes from whole blood or blood clots stored in the HUNT Biobank, using the Puregene kit (Gentra Systems, Minneapolis, MN) manually or with an Autopure LS (Gentra Systems). Laboratory technicians were blinded to the results of the height measurements. Details on the DNA extraction and the HUNT Biobank are described elsewhere [16].

Genotyping of short and tall individuals were done at the Norwegian University of Science and Techonology, Norway using the iSelect Metabochip (Illumina, San Diego, CA) and the Infinium HD ultra protocol. Each 96-well plate included both tall and short individuals and one sample of identical reference DNA. Genotype calling was done using GenTrain version 2.0 in GenomeStudio V2010.3 (Illumina, San Diego, CA). Genotyping of the middle group was done on the Metabochip at the Center for Inherited Disease Research (CIDR, MD) and called with BeadStudio 3.3.7 with Gentrain version 1.0 (Illumina, San Diego, CA).

Samples that did not meet a 99% completion threshold were excluded from further analysis (N=19; 0.7%). Additional post-genotyping exclusions based on gender discrepancy (N=11) and first-degree relatedness (pi-hat >0.2; N=152, 6.3%) were done using PLINK [18]. Ethnic outliers (N=174, 7.2%) were excluded using the EIGENSTRAT software package [19]. After quality assessment 2,063 individuals (85.7%) remained for further analysis, 385 (81.2%) short, 456

(95.2%) tall and 1,224 (83.9%) individuals in the middle group.

106 SNPs of the 180 GIANT height hits were directly typed on the Metabochip. In addition, we used the SNP Annotation and Proxy Search to map 54 of the remaining 74 SNPs with a HapMap r2 > 0.8 linkage disequilibrium proxy result [20]. These 160 SNPs (i.e. 106 directly typed and 54 proxies) were used in subsequent analyses. All SNPs showed a genotyping success rate >98% and were in Hardy Weinberg equilibrium.

FINRISK study

We directly genotyped the samples for the 180 previously identified height SNPs. The genotyping was done at Children's Hospital Boston using Sequenom iPLEX genotyping (Sequenom, Inc, San Diego, CA, USA). In total, 186 short individuals and 192 tall individuals were successfully genotyped for 158 SNPs. All 158 SNPs had a genotyping success rate ≥ 90% and the overall genotyping rate was 97.85%. One of these SNPs (rs1809889) is not part of the 180 GIANT SNPs, but data were available for this SNP from the GIANT meta-analysis so it was included in our analysis.

We genotyped an additional 49 ancestry informative markers (AIMs) to identify ethnic outliers [21]. We inputted genotype data from our subjects as well as the reference HAPMAP samples (CEU, YRI, CHB+JPT) for the 49 AIMs together with 130 height SNPs into Structure 2.3.3 [22]. We detected 5 ethnic outliers with >10% Asian ancestry who were excluded from further analysis leaving a total of 181 short and 192 tall individuals as our FINRISK study group.

Statistical Analysis

Individual SNP analysis

For FINRISK, we calculated the observed odds ratio for each of our 158 SNPs using the Cochran-Manzel-Hansel test, which is a stratified chi-square test. We stratified the individuals into 4 sub-cohorts based on geography and gender (Table 2.4) and performed the test using PLINK [18]. The observed odds-ratio for each SNP was recorded, along with the 95% confidence interval. For HUNT the observed odds-ratio and 95% confidence intervals and the single association analysis was performed using logistic regression in PLINK.

For both cohorts, we calculated the expected odds ratio for each SNP by estimating the odds of the height-increasing versus the height-decreasing allele in both the tall extremes (cases) and the short extremes (controls) assuming a standard normal distribution for standardized height, i.e. height ~ *Normal(0,1)*. For a given SNP, we defined the height-increasing effect size as β and the height-increasing allele frequency as p. The mean height for the height-increasing allele would be $M_i = \beta\ (1 - p)$ and the mean height for the height-decreasing allele would be $M_d = -\beta\ p$. The variance of height for the both alleles would be $V = 1 - \beta^2\ p\ (1-p)$. We then calculated the odds of observing the height-increasing allele versus the height-decreasing allele for both the tall extremes (cases) and the short extremes (controls) by taking the ratio of the probabilities of each allele being seen in the cases and the controls respectively. These are calculated as:

$$Odds_{cases} = \frac{\int_{2.326}^{\infty} N(x \mid M_i, V)\,dx}{\int_{2.326}^{\infty} N(x \mid M_d, V)\,dx}$$

$$Odds_{controls} = \frac{\int_{-\infty}^{-2.326} N(x \mid M_i, V)\,dx}{\int_{-\infty}^{-2.326} N(x \mid M_d, V)\,dx}$$

where $N(x|M,V)$ denotes the density function at x of a Normal distribution with mean M and variance V. We use a cut-off of +/-2.326 to denote the approximate 1% tails. We then calculated the expected odds-ratio by taking the ratio between $Odds_{cases}$ over $Odds_{controls}$, i.e.

$$Expected\ Odds\ Ratio = \frac{Odds_{cases}}{Odds_{controls}}$$

To assess whether individual SNPs had odds ratios significantly different from expectation, we generated upper and lower 95% confidence limits for the expected distribution based on the GIANT beta and standard errors estimates as above, and used the natural log of these confidence limits to estimate an approximate standard error for the expected odds ratio, i.e.

$$SE_{Expected\_OR} = \frac{\ln(Upper) - \ln(Lower)}{2 \times 1.96}$$

We then assessed significance by a Z-test of the difference between observed odds ratio and expected odds ratio to obtain the Zscore, i.e.

$$Zscore = \frac{\ln(OR_{observed}) - \ln(OR_{expected})}{\sqrt{SE_{Observed\_OR}{}^2 + SE_{Expected\_OR}{}^2}}$$

Meta-analysis

The HUNT and FINRISK studies genotyped different sets of SNPS, with only 98 of the SNPs matching exactly across the studies. We therefore used forty-three of the HUNT SNPs that had $r^2 > 0.8$ HapMap proxies with a genotyped FINRISK SNPs. We used the inverse variance method to meta-analyze the odds ratios for these 141 SNPs from the two studies. As opposed to the individual studies, where study specific allele frequencies were used, we used the GIANT allele frequency information to generate the expected odds ratios for the meta-analysis. This did

not appreciably affect the results for individual SNP analysis within the individual studies, and the meta-analyzed results were consistent to those in the two individual studies.

Modeling the Weighted Allele Score (WAS)

To calculate the Weighted Allele Score (*WAS*) for each individual, we took the sum of the effective allele dosages of the height SNPs multiplied by their respective estimated effect sizes (βs) using the Stage 1 betas from the GIANT study, as shown in the formula below.

$$WAS = \sum_{i=1}^{N} \beta_i \times SNP_i - \alpha$$

β and *SNP* are the effect size and effective allele dosage (0, 1 or 2) of the height SNPs and *WAS* is the weighted allele score. *N* is the total number of SNPs available to calculate the weighted allele score. α is the mean of the sum such that the expected *WAS* is 0 as shown by the formula below.

$$\alpha = \sum_{i=1}^{N} 2 \times \beta_i \times Frequency_i$$

Frequency is the allele frequency of the effect allele obtained from the Finnish or HUNT estimates.

We calculated the statistical difference between the *WAS* of the short versus the tall individuals by performing a 2-tailed 2-sample t-test to obtain the respective p-value. All the calculations were done using the R statistical software package.

Obtaining Finnish allele frequency estimates

The allele frequency estimate for each SNP was obtained by taking only the Finnish individuals from the GIANT height study and calculating the expected allele frequency. The cohorts used were the FUSION NIDDM Case control study from Finland, the GenMets Case control study from Finland and the FINRISK component of the MIGen cohort. The total number of individuals used for obtaining the estimates is 3618.

Simulating the distribution of WAS under the null model

The null model assumes that the only factors determining height (Z-score) are the cumulative additive effects of the GIANT height SNPs and noise. We modeled the Z-score with the formula below.

$$Zscore = WAS + N(0, {\sigma^2}_{remaining})$$

Zscore is the height Z-score, $N(0, \sigma^2_{\text{remaining}})$ is a normally distributed random variable with mean 0 and variance $\sigma^2_{\text{remaining}}$. $\sigma^2_{\text{remaining}}$ is calculated such that the variance of *Zscore* is 1, i.e. $\sigma^2_{\text{remaining}}$ is 1 – var(*WAS*). The variance of *WAS* can be calculated with the formula below,

$$var(WAS) = \sum_{i=1}^{N} 2 \times {\beta_i}^2 \times Frequency_i \times (1 - Frequency_i)$$

On the other hand, a simulated individual's effective allele dosage is obtained by sampling from a set of binomial distributions with *N*=2 and *p* being the allele frequencies of each SNP. The simulated effective allele dosages can then be used to calculate each individual's *WAS*. The

simulation approach for each cohort was modeled to mirror the methods of subject selection.

Simulating FINRISK

For the FINRISK study, the simulations were performed using the following steps. We first generated the effective allele dosages for each SNP for 200,000 individuals by random sampling. We then randomly sampled 4271, 6582, 5025 and 7610 individuals to represent the 4 sub-populations and obtained their Z-scores using the previously described modeling. For each subgroup, we picked the appropriate number of the most extreme individuals to mimic the actual sample selection. We then pooled the short and tall extremes together and randomly dropped individuals to obtain exactly 181 short extremes and 192 tall extremes. We then randomly drop SNPs from the simulated individuals to mimic the missing genotype rate in FINRISK and then calculate the Weighted Allele Score (*WAS*) for each simulated individual. This simulation process was repeated 10,000 times. For the stratified analyses of various height cut-offs, we adjusted the numbers of selected individuals in each strata by taking the floor of the expected number of individuals in that strata. In our cohort, the top 0.5% extremes included 21, 32, 25 and 38 individuals from each tail of the 4 sub-populations respectively, and for the top 0.25% extremes included 10, 16, 12 and 19 individuals from each tail of the 4 sub-populations. For the top ~1% to 0.25% extremes, we included all our extremes but excluded the top 10, 16, 12 and 19 individuals from each tail of the 4 sub-populations.

Simulating HUNT

The simulations for HUNT were performed as follows. We generated the effective allele

dosages for each SNP for 400,000 individuals by random sampling. We then randomly selected 50,000 individuals and obtained their Z-scores.

We then selected all short and tall extremes with a Z-score cut-off of -2.14 and +2.14 respectively. Next, we randomly selected 385 short extremes and 456 tall extremes and calculated the *WAS*. This process was repeated 10,000 times. As in the FINRISK simulation, the number of individuals varies for each stratified analysis. Because we performed stratified analyses for varying levels of height cut-offs, our definition for the top 0.5% extremes is a Z-score cut-off below -2.57 and above +2.57 and for the top 0.25% extremes is a Z-score cut-off below -2.81 and above +2.81. For the top ~1.5% to 0.25% extremes, we used only extremes that had Z-scores between -2.14 and -2.81 for the short extremes and between 2.14 and 2.81 for the tall extremes.

Determining if the mean observed *WAS* is significantly different from the simulated expectation

We evaluated the significance of the mean observed *WAS* by determining the p-value of the mean observed *WAS* from the null distribution of the mean *WAS* obtained from the simulations. The two-tailed p-value is calculated by evaluating the mean observed *WAS* from Normal($\mu_{simulation}$, $\sigma^2_{simulation}$) where $\mu_{simulation}$ is the mean of the mean *WAS* and $\sigma^2_{simulation}$ is the variance of the mean *WAS* from the simulations.

Modeling Rare-variants with moderate to large effect sizes

Modeling the rare-variant effect into the simulation is accomplished by adding an additional

rare-variant term into the calculation of the height Z-score without changing the definition of *WAS* as shown in the equation below.

$$Zscore = WAS + \left(\sum_{i=1}^{n} B_i \times V_i\right) - \alpha_{rv} + N(0, \sigma^2{}_{remaining})$$

where *n* is the number of independent rare-variants, *B* represents the effect size of the rare-variants, and V is the allele dosage of the rare-variant. α_{rv} is the mean of the rare-variants score such that the rare-variants do not change the expected Z-score, i.e. the expected Z-score is still 0. Similarly, α_{rv} can be calculated by the following formula,

$$\alpha_{rv} = \sum_{i=1}^{n} 2 \times B_i \times F_i$$

$\sigma^2_{remaining}$ in this case will have to be adjusted for the rare-variants such that the variance of the Z-score remains at 1, i.e. $\sigma^2_{remaining}$ is 1 – var(*WAS*) – var(Σ *B V*). *F* is the allele frequency of the rare-variants. Simulations done with modeling rare-variants are identical to the prior simulations of FINRISK or HUNT except that the new terms are used for calculating the Z-score.

ACKNOWLEDGEMENTS

We would like to thank Sailaja Vedantam for the calculation of the Finnish allele frequencies, Minttu Sauramo and Elina Mäkinen for aliquotting the FINRISK DNA samples.

REFERENCES

1. Visscher PM, Macgregor S, Benyamin B, Zhu G, Gordon S, et al. (2007) Genome Partitioning of Genetic Variation for Height from 11,214 Sibling Pairs. Am J Hum Genet

81: 1104–1110. doi:10.1086/522934.

2. Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, et al. (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature. Available: http://www.ncbi.nlm.nih.gov.ezp-prod1.hul.harvard.edu/pubmed/20881960. Accessed 4 October 2010.

3. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, et al. (2010) Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet 11: 446–450. doi:10.1038/nrg2809.

4. Cirulli ET, Goldstein DB (2010) Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet 11: 415–425. doi:10.1038/nrg2779.

5. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. (2009) Finding the missing heritability of complex diseases. Nature 461: 747–753. doi:10.1038/nature08494.

6. Sham PC, Cherny SS, Purcell S, Hewitt JK (2000) Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data. Am J Hum Genet 66: 1616–1630. doi:10.1086/302891.

7. Lander ES, Botstein D (1989) Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121: 185–199.

8. Risch N, Zhang H (1995) Extreme discordant sib pairs for mapping quantitative trait loci in humans. Science 268: 1584–1589. doi:10.1126/science.7777857.

9. Van Gestel S, Houwing-Duistermaat JJ, Adolfsson R, van Duijn CM, Van Broeckhoven C (2000) Power of selective genotyping in genetic association analyses of quantitative traits. Behav Genet 30: 141–146.

10. Abecasis GR, Cookson WOC, Cardon LR (2001) The Power to Detect Linkage Disequilibrium with Quantitative Traits in Selected Samples. Am J Hum Genet 68: 1463–1474. doi:10.1086/320590.

11. Duncan EL, Danoy P, Kemp JP, Leo PJ, McCloskey E, et al. (2011) Genome-Wide Association Study Using Extreme Truncate Selection Identifies Novel Genes Affecting Bone Mineral Density and Fracture Risk. PLoS Genet 7: e1001372. doi:10.1371/journal.pgen.1001372.

12. Cotsapas C, Speliotes EK, Hatoum IJ, Greenawalt DM, Dobrin R, et al. (2009) Common body mass index-associated variants confer risk of extreme obesity. Hum Mol Genet 18: 3502–3507. doi:10.1093/hmg/ddp292.

13. Hegele RA, Ban MR, Hsueh N, Kennedy BA, Cao H, et al. (2009) A polygenic basis for four classical Fredrickson hyperlipoproteinemia phenotypes that are characterized by hypertriglyceridemia. Hum Mol Genet 18: 4189–4194. doi:10.1093/hmg/ddp361.

14. Guey LT, Kravic J, Melander O, Burtt NP, Laramie JM, et al. (2011) Power in the phenotypic extremes: a simulation study of power in discovery and replication of rare variants. Genet Epidemiol 35: 236–246. doi:10.1002/gepi.20572.

15. Allison DB, Heo M, Schork NJ, Wong S-L, Elston RC (1998) Extreme Selection Strategies in Gene Mapping Studies of Oligogenic Quantitative Traits Do Not Always Increase Power. Hum Hered 48: 97–107. doi:10.1159/000022788.

16. Holmen J, Midthjell K, Krüger Ø, Langhammer A, Holmen TL, et al. (2003) The Nord-Trøndelag Health Study 1995-97 (HUNT 2): Objectives, contents, methods and participation. Nor Epidemiol 13: 19–32.

17. Vartiainen E, Laatikainen T, Peltonen M, Juolevi A, Männistö S, et al. (2010) Thirty-five-year trends in cardiovascular risk factors in Finland. Int J Epidemiol 39: 504–518. doi:10.1093/ije/dyp330.

18. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559–575. doi:10.1086/519795.

19. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38: 904–909. doi:10.1038/ng1847.

20. Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, et al. (2008) SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics 24: 2938–2939. doi:10.1093/bioinformatics/btn564.

21. Egyud MRL, Gajdos ZKZ, Butler JL, Tischfield S, Marchand L, et al. (2009) Use of weighted reference panels based on empirical estimates of ancestry for capturing untyped variation. Hum Genet 125: 295–303. doi:10.1007/s00439-009-0627-8.

22. Pritchard JK, Stephens M, Donnelly P (2000) Inference of Population Structure Using Multilocus Genotype Data. Genetics 155: 945–959.

Chapter 3

An excess of risk-increasing low frequency variants can be a signal of polygenic inheritance in complex diseases

Yingleong Chan[1,2,3], Elaine T Lim[1,2,4], Niina Sandholm[5,6,7], Sophie R Wang[1,2,3], Amy Jayne McKnight[8], Stephan Ripke[2,4], DIAGRAM Consortium, GENIE Consortium, GIANT Consortium, IIBDGC Consortium, PGC Consortium, Mark J Daly[1,2,4], Benjamin M Neale[2,4], Rany M Salem[1,2,3], Joel N Hirschhorn[1,2,3]

[1] Harvard Medical School, Department of Genetics, Boston, Massachusetts, USA.
[2] Program in Medical and Population Genetics, Broad Institute, Cambridge, Massachusetts, USA.
[3] Department of Endocrinology, Boston Children's Hospital, Boston, Massachusetts, USA.
[4] Analytic and Translational Genetics Unit, Massachusetts General Hospital, Massachusetts , USA.
[5] Folkhälsan Institute of Genetics, Folkhälsan Research Center, Biomedicum Helsinki, Helsinki, Finland
[6] Division of Nephrology, Department of Medicine, Helsinki University Central Hospital, Helsinki, Finland
[7] Department of Biomedical Engineering and Computational Science, Aalto University School of Science, Helsinki, Finland
[8] Nephrology Research, Centre for Public Health, Queen's University of Belfast, Belfast, Northern Ireland, UK

Originally published as:

Y Chan, et. al., American Journal of Human Genetics (2014), Volume 94, Issue 3, Pages 437–452

ABSTRACT

In most complex diseases, much of the heritability remains unaccounted for by common variants. It has been postulated that lower frequency variants contribute to the remaining heritability. Here, we describe a method to test for polygenic inheritance from lower frequency variants using GWAS summary association statistics. We explored scenarios with many causal low frequency variants and showed that there is more power to detect risk variants than protective variants, resulting in an increase in the ratio of detected risk to protective variants (R/P ratio). Such an excess can also occur if risk variants are present and kept at lower frequencies because of negative selection. The R/P ratio can be falsely elevated because of reasons unrelated to polygenic inheritance, such as uneven sample sizes or asymmetric population stratification, so precautions to correct for these confounders are essential. We tested our method on published GWAS results and observed a strong signal in some diseases (schizophrenia and type 2 diabetes) but not others. We also explored the shared genetic component in overlapping phenotypes related to inflammatory bowel disease (Crohn's disease [CD] and ulcerative colitis [UC]) and diabetic nephropathy (macroalbuminuria and end stage renal disease [ESRD]). While the signal was still present when both CD and UC were jointly analyzed, the signal was lost when macroalbuminuria and ESRD were jointly analyzed, suggesting that these phenotypes should best be studied separately. Thus, our method may also help guide the design of future genetic studies of various traits and diseases.

INTRODUCTION

Most common diseases involve a mix of both genetic and environmental factors and do

not follow simple patterns of Mendelian inheritance. In such diseases, the genetic component is usually polygenic: genetic variation in many genes individually contribute a small or a moderate component of disease risk [1]. Genome-wide association studies (GWAS) have identified numerous genomic loci in which common variants (≥5% frequency) are associated with complex diseases [2]. Even in some of the largest and most successful GWAS to date, much of the genetic contribution to phenotype remains unexplained (sometimes called "missing heritability") [3,4], suggesting that lower frequency variants, not well surveyed by GWAS, may also contribute to the missing heritability. Indeed, in some diseases such as autism spectrum disorders (ASD [MIM 209850]), inherited rare (<1% frequency) and low frequency (<5% frequency) variants have been recently shown to play an important role in the genetic architecture of the disorder [5,6], suggesting that more loci with low frequency variants could be identified if appropriate additional studies were performed. In other diseases, there is as yet little evidence of a substantial role for low frequency variation, leaving open the question of whether studies of low frequency variation will be fruitful for those diseases.

The relative success of different approaches in identifying more contributing loci will depend on what type of variation accounts for the missing heritability. Low frequency variants may remain undetected because they may not be well-represented or well-tagged by markers on genotyping arrays and therefore would not be well-imputed [7]. Along these lines, the statistical power to detect low frequency variants in GWAS is much lower than common variants if their underlying effect sizes are similar[8]. Knowing whether low frequency variants contribute to the missing heritability of a disease is important because approaches better-suited to identify additional common variants differ from those aimed at identifying rarer variants (genotyping arrays with common variants compared to arrays with lower frequency variants or sequencing).

Methods for detecting a contribution from common variants to the missing heritability have been described previously. In a GWAS of schizophrenia (SCZ [MIM 181500]) [9], Purcell and colleagues developed the concept of a polygenic score by combining the effects of multiple common variants that are modestly associated with schizophrenia. They showed that the score is predictive of schizophrenia in an independent cohort, thus indicating that there is a polygenic signal from many yet-to-be-detected common variants in schizophrenia. Yang and colleagues adopted a different approach by assessing the narrow-sense heritability of human height with a linear-model analysis using hundreds of thousands of common variants [10]. They found that at least 45% of the variance of height can be accounted for by common variants, indicating that there are many common variants associated with height that have yet to be discovered. Although both methods can be used to detect a signal of polygenic inheritance from common variants in complex diseases, these tests were not designed to specifically test for low frequency variants, and also require individual-level genotype data.

In this chapter, we describe an approach that can be applied directly to GWAS summary statistics to ascertain the presence of polygenic inheritance from low frequency variants. We observed that, if low frequency variants contribute to disease susceptibility, there can be an excess of associated risk variants compared to protective variants at a given significance level. Here, risk variants are defined as variants for which the minor allele is associated with increased risk of disease and protective variants are defined as variants for which the minor allele is associated with decreased risk of disease. Under the null model, there should be no excess of associated risk variants compared to protective variants. We calculated the risk to protective ratio (R/P ratio): the ratio of the number of detected risk variants over the number of detected protective variants, to test for such an excess of risk variants. We explored various scenarios that

could give rise to an increased in the R/P ratio. First, we showed empirically and analytically that when low allele frequency variants contribute to polygenic inheritance of a disease with low prevalence, there is an elevated R/P ratio because of greater power to detect risk variants than protective variants. Next, we showed through simulations that under a scenario of polygenic inheritance that includes negative selection, risk variants can have lower average frequencies than protective variants leading to an elevated R/P ratio within the lower frequency range. However, we also showed that such an elevated R/P ratio can occur because of reasons unrelated to polygenic inheritance. First, we showed that an uneven sample size of having substantially more controls than cases can produce an apparent increase in the R/P ratio and therefore, where the sample size is not balanced between cases and controls, one should compare the observed R/P ratio against that obtained through simulations with the same number of cases and controls. Next, we showed that particular scenarios of asymmetric population stratification can produce a similar excess of low frequency risk variants and recommend that precautions for detecting and correcting for such stratification should be performed before one can confidently interpret an excess of risk variants as being a signal of polygenic inheritance.

We then applied our method to results from published GWAS for several diseases, including schizophrenia [11], bipolar disorder (BIP [MIM 125480]) [12], major depressive disorder (MDD [MIM 608516]) [13], type 2 diabetes (T2D [MIM 125853]) [14] and various classes of obesity (OB [MIM 601665]) [15]. We observed strong signals of increased risk variants in several of the diseases but little or no signal in others, suggesting that efforts to discover low frequency and rare variants will be more fruitful for the diseases with such a signal. We further used our method to test whether apparently related phenotypes share low frequency or rare genetic contributors and hence should be analyzed together or separately. By applying the

method to phenotypes related to diabetic nephropathy (DN [MIM 603933] [16] and inflammatory bowel disease (IBD [MIM 266600]) [17], we found that the polygenic signal was eliminated when individuals with macroalbuminuria and individuals with end stage renal disease were analyzed together, whereas we still observed a significant signal when individuals with Crohn's disease and ulcerative colitis were analyzed together. Thus, our method has the potential to guide the strategy in searching for additional genetic loci as well as in prioritizing the choice of phenotype for future studies of rare genetic variation in polygenic traits and diseases.

MATERIALS AND METHODS

Testing for an excess of risk variants from GWAS summary statistics

Calculating the R/P ratio statistic from observed GWAS summary statistics

The four input fields we used for R/P ratio calculations for each SNP are: an identifier (rsID), the minor allele frequency, the association P-value, and a field to determine the direction of effect, i.e. either an odds-ratio (OR) or an effect size (β). The ORs or βs were adjusted to reflect the effect of the minor allele by inverting the ORs or changing the sign of the βs if they were reported for the major allele. Each variant was assigned as risk if the $OR > 1$ or $\beta > 0$ and protective if the $OR < 1$ or $\beta < 0$. Neutral variants, i.e. $OR = 1$ or $\beta = 0$ were discarded from the analysis. We removed SNPs not present in the Hapmap CEU population (phase 2 release 28) [18,19],not in the 1,000 Genomes EUR population [20] as well as SNPs with minor allele frequency less than 1%. We sorted the remaining variants in order from most significant to least and performed LD-pruning by systematically going through the variants and removing variants that have an $r^2 > 0.1$ with any of the more significantly associated variants. We used PLINK [21]

to calculate r^2 correlations of variant-pairs within a 1 mega-base window from 379 EUR individuals of the 1,000 Genomes. To measure the excess of risk variants in the lower frequency range, we separated the low frequency variants into 3 distinct bins, i.e. 1%-5%, 5%-10% and 10%-15%. We also included the 30%-50% bin as a negative-control, where we should not observe any excess of risk variants. For each bin, we counted the number of detected risk variants and the number of detected protective variants that meet significance cutoffs of $P < 0.001$ and $P < 0.01$. We calculated the R/P ratio as,

$$R/P\ ratio = \frac{No.\ of\ detected\ risk\ variants}{No.\ of\ detected\ protective\ variants}$$

Assessing the significance of the observed Risk/Protective (R/P) ratio

To assess the significance of an elevation in R/P ratio, we simulated individuals using HAPGEN [22] by using parameters from the Hapmap CEU population (phase 3, r2) to obtain the null distribution of the $\log_2$ R/P ratio statistic. We first simulated 100,000 individuals to form a pool of individuals that we can subsequently sample from. Next, we randomly sampled the same number of individuals in cases and controls as were used in the actual GWAS, performed the association test using PLINK, with LD-pruning and R/P ratio calculations identical to the procedure described above. We repeated this process 1000 times to obtain accurate estimates of the sample mean (μ) and standard deviation (σ) of the $\log_2$ R/P ratio under the null for each of our frequency bins and P-value cutoffs. We calculated the significance of the observed $\log_2$ R/P ratio by performing a one-tailed Z-test to obtain the Zscore and P-value (P), i.e.

$$Zscore = \frac{observed\ log_2 R/P\ ratio - \mu}{\sigma}$$

$$P = \int_{Zscore}^{\infty} N(x, 0,1)dx$$

We defined $P < 0.01$ as our significance threshold for calling a significant excess of risk variants.

Calculating non-centrality parameter (NCP) for comparing power between risk and protective variants

Power calculation

The power of a variant is expressed by calculating the expected non-centrality parameter (NCP) of the χ^2 distribution for the alternative distribution. The greater the NCP, the more power there is to detect the effective variant. The algorithm for calculating NCP is identical to the genetic power calculator[8] for case-control threshold-selected quantitative traits, assuming an additive model of the QTL effect, i.e. the dominance to additive QTL effect parameter is set to 0. The variance explained for a SNP with allele frequency as p and effect size as β is $\beta^2 2p(1-p)$. For risk variants, we calculated the NCP (NCP_{risk}) for multiple values of effect sizes (β), ranging from 0 to 0.5 with intervals of 0.01. Similarly, for protective variants, we calculated the NCP ($NCP_{protective}$) for multiple values of β, ranging from 0 to -0.5 with intervals of 0.01. The relative difference in power between risk and protective variants is measured by the NCP ratio. The NCP ratio is calculated as,

$$NCP\ ratio = \frac{NCP_{risk}}{NCP_{protective}}$$

Base Model

We define the base model as a set of parameters used for calculating NCP. 10,000 cases, 10,000 controls, effective and marker variant frequency set to 1%. The prevalence is set as 1%, i.e. the trait threshold's lower and upper limit is 2.33 and 9 respectively for cases and -9 and 2.33 for controls. We have used 9 and -9 as surrogates for infinity ($+\infty$ and $-\infty$ respectively) but any sufficiently large number will not change the conclusions of the downstream analyses. Complete linkage disequilibrium (LD) between the causal variant and marker variant is assumed, i.e. D' = 1.

Simulating R/P ratios for negative selection

Obtaining frequencies and effect sizes

If the variants that have an effect on the phenotype are under negative selection, it can lead to scenarios where there are more risk variants than protective variants to begin with, especially for low frequency variants. To illustrate this, we simulated neutral variants and causal variants under negative selection using previously published models and parameters that result in an allele spectrum similar to that observed in European population [23,24]. We used the forward simulation package ForSim [25] to simulate coding sequence variation in the European population in 1000 genes. The average gene coding length was set as 1500bp. We used a mutation rate per site of $2x10^{-8}$ and a uniform locus-wide recombination rate of 2Mb/cM. We modeled the distribution of selection coefficients (s) for *de novo* missense mutations by a gamma distribution [26]. We used the conventional 4-parameter model of the history of the European population with long-term constant size (N=8100 for 45,000 generations) followed by a bottleneck (N=2000) and then by exponential growth (1.5% increase per generation for 370

generations) to achieve a final population size of approximately 500,000 individuals [23,24]. We obtained 823 non-neutral variants that have minor allele frequencies ≥ 1% and assigned them as effective variants and assuming that the allele under negative selection confers risk, i.e. positive effect (Figure 3.1). By considering only additive genetic effects, we assigned effect sizes as: $\beta=s^{\tau}(1+\varepsilon)$ as suggested in Eyre-Walker [27]. Here, β is the variant's additive effect on the quantitative trait; s is the absolute value of the variant's selection coefficient and ε is a normally distributed random noise parameter which was set to having mean 0 and standard deviation 0.05. τ is the degree of coupling between β and s and was set at 0.5 for our analyses. The effect sizes are scaled so that these 823 variants explain 60% of the phenotypic variance.

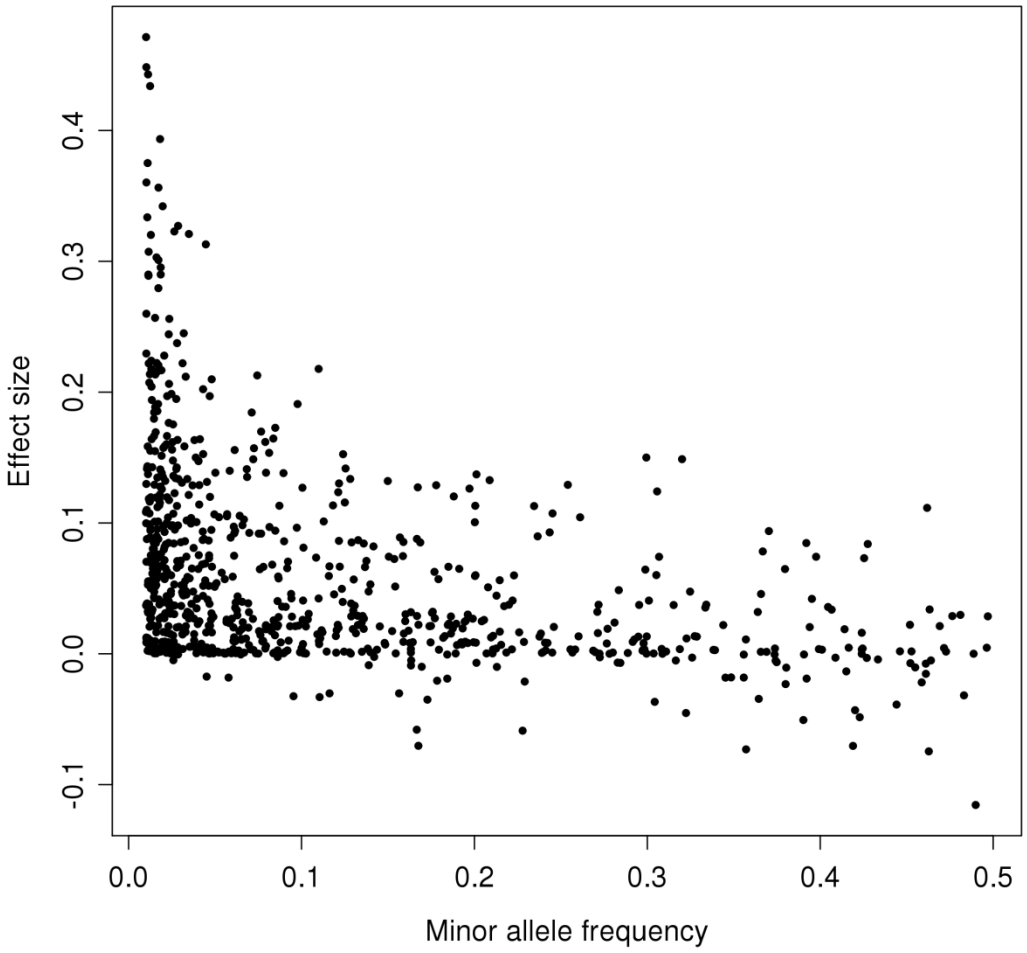


Figure 3.1: The frequency and effect sizes for the 823 SNPs under selection. The plot shows the minor allele frequency (x-axis) and effect size in standard deviation units (y-axis) for the 823 SNPs that were obtained through simulating a trait under negative selection.

Obtaining phenotypes and calculating R/P ratio for the selection model

We use the 100,000 HAPGEN simulated individuals and selected 823 matched SNPs such that the frequency matches the variants generated by ForSim. We then assigned these matched SNPs with effect sizes determined earlier. We calculated the phenotypic Zscore for each of our 100,000 individuals in the same way that we did in a previous study [28], i.e. by calculating the weighted allele score (*WAS*) and adding it to a randomly generated variable sampled from a normal distribution of mean 0 and variance 0.4 such that the total variance explained is 1. We then sampled 2,000 individuals with phenotypic Zscores > 1.645 (5% prevalence) as cases and another 2,000 individuals with phenotypic Zscores ≤ 1.645 as controls. We used PLINK to perform the association test on all the variants and calculated the R/P ratio within the same frequency bins as well as P-value cutoffs as described above. This process was repeated 1,000 times to obtain the distribution of the R/P ratio. For the control model, we randomly sampled 2,000 individuals as cases and 2,000 individuals as controls and calculated the R/P ratio as described above.

Simulating R/P ratios for population stratification

We use HAPGEN to simulate 4,000 distinct individuals from the Hapmap CEU population (phase 3, r2) as well as another 4000 distinct individuals from the Hapmap TSI population (phase 3, r2). For complete stratification, we randomly sampled 1,000 individuals from the CEU pool as controls and 1,000 individuals the TSI pool as cases. We simulated asymmetric mixtures of 1, 5 and 10 percent by randomly sampling 1000 individuals from the CEU pool as controls and sampled 10, 50 and 100 individuals from the TSI pool as cases,

respectively, and made up the remainder of the cases from the CEU pool. We used PLINK to perform the association test on all the variants and calculated the R/P ratio within the same frequency bins as well as P-value cutoffs as described above. Each process was repeated 1,000 times to obtain the distribution of the R/P ratio. All PCA analysis was performed using smartpca from the EIGENSOFT 3.0 package [29]. All meta-analysis of GWAS summary statistics were performed using METAL[30]. Inflation of the GWAS test statistic due to population stratification was assessed by genomic control inflation factor (λ_{GC}) [31].

Calculating R/P ratio from published GWAS summary statistics

Schizophrenia, major depressive disorder and bipolar disorder

GWAS summary statistics were provided from published results of schizophrenia [11], bipolar disorder [12] and major depressive disorder [13]. SNPs that failed imputation (INFO < 0.6) were discarded. The number of cases and controls used for simulating the null distribution are as follows: Schizophrenia (SCZ), 9,394 cases and 12,462 controls; major depressive disorder (MDD), 9,240 cases and 9,519 controls; bipolar disorder (BIP), 7,481 cases and 9,250 controls.

Type 2 diabetes

GWAS summary statistics were provided from published results of type 2 diabetes [14]. SNPs that passed imputation for less than 15,000 individuals (N_{cases} < 15,000) were discarded. The number of cases and controls used for simulating the null distribution are 15,000 cases and 50,337 controls.

Obesity

GWAS summary statistics were provided from published results of various classes of obesity[15]. SNPs that passed imputation for less than 50,000 individuals ($N_{cases} < 50,000$), 10,000 individuals ($N_{cases} < 10,000$), 2,000 individuals ($N_{cases} < 2,000$) and 1,000 individuals ($N_{cases} < 1,000$) were discarded for the overweight (BMI > 25), class1 (BMI > 30), class2 (BMI > 35) and class3 (BMI > 40) datasets respectively. The number of cases and controls used for simulating the null distribution are as follows: overweight, 50,000 cases and 35,715 controls; class1, 10,000 cases and 20325 controls; class2, 2,000 cases and 12,466 controls; Class3, 1,000 cases and 18,346 controls.

Inflammatory bowel disease

GWAS summary statistics were provided from published results of Crohn's disease (CD) [32], ulcerative colitis (UC) [33] and the combined case cohort of both Crohn's disease and ulcerative colitis (CD+UC) [17]. SNPs that failed imputation (INFO < 0.6) were discarded. The number of cases and controls used for simulating the null distribution are as follows: CD, 5,956 cases and 14,927 controls; UC, 6,968 cases and 20,464 controls; CD+UC, 12,882 cases and 21,770 controls.

Diabetic nephropathy

GWAS summary statistics were provided from published results of phenotypes related to

diabetic nephropathy [16] which are Macroalbuminuria (MACRO) and End stage renal disease (ESRD). SNPs that failed imputation in at least 1 cohort were discarded. The number of cases and controls used for simulating the null distribution are as follows: macroalbuminuria versus control ($MACRO_{ctrl}$), 1,478 cases and 3,315 controls; end stage renal disease versus control ($ESRD_{ctrl}$), 1,399 cases and 3,315 controls; ESRD versus controls that include MACRO ($ESRD_{ctrl+macro}$), 1,399 cases and 5,253 controls; combined MACRO and ESRD versus control ($[MACRO + ESRD]_{ctrl}$), 2,916 cases and 3,315 controls.

RESULTS

We developed a method to detect and assess the significance of an excess of risk variants, measured by the ratio of risk variants to protective variants (R/P ratio) within a series of frequency bins and P-value cutoffs (see Materials and Methods). We proceeded to show that under an assumption of polygenic inheritance from low frequency variants, there is more statistical power to detect risk variants than protective variants, which can result in an increased R/P ratio. We also showed that such an excess can also occur if risk variants are kept at lower frequencies because of negative selection. However, such an excess can also occur because of reasons unrelated to a contribution of rare variants to disease risk: uneven sample sizes or asymmetric population stratification. Therefore, steps have to be taken to account for these latter possibilities before one can confidently interpret the excess of risk variants as a true signal of polygenic inheritance. Finally, we applied the method to GWAS summary statistics from several published studies.

Significantly higher power to detect low frequency risk variants of moderate to large effect

The liability threshold model for disease [34] has been shown to be consistent with results from GWAS for multiple diseases [35]. This model assumes that there is an underlying unmeasured trait related to disease risk, and that individuals are affected with disease only when the value of the trait exceeds a particular threshold. Under such a model, we discovered that the statistical power to detect risk variants is higher than the power to detect protective variants, even when they have the same effect size with respect to the underlying unmeasured trait. For example, we calculated power using a pre-defined set of parameters defined as the 'base model' (see Materials and Methods). From our calculations, we observed that, as effect size increases, there is significantly more power to detect risk than protective variants as indicated by the increase in the NCP ratio (Figure 3.2). This result shows that for this scenario, where the number of risk and protective variants are equal and have similar absolute effect sizes, the difference in power can create an excess of detected risk variants over protective variants which can result in an increased R/P ratio.

The difference in power is larger under certain scenarios

We explored how the difference in power to detect risk and protective variants would be affected when we varied the parameters in the model under which we calculated power. First, we calculated power using the base model but varied the minor allele frequency from 1% to 15%. The difference in power for risk and protective variants decreases as the variant frequency increases (Figure 3.3A). Second, we varied the disease prevalence from 1% (trait Z-score > 2.33) to 15% (trait Z-score > 1.03). Here, the difference in power decreases with increasing disease

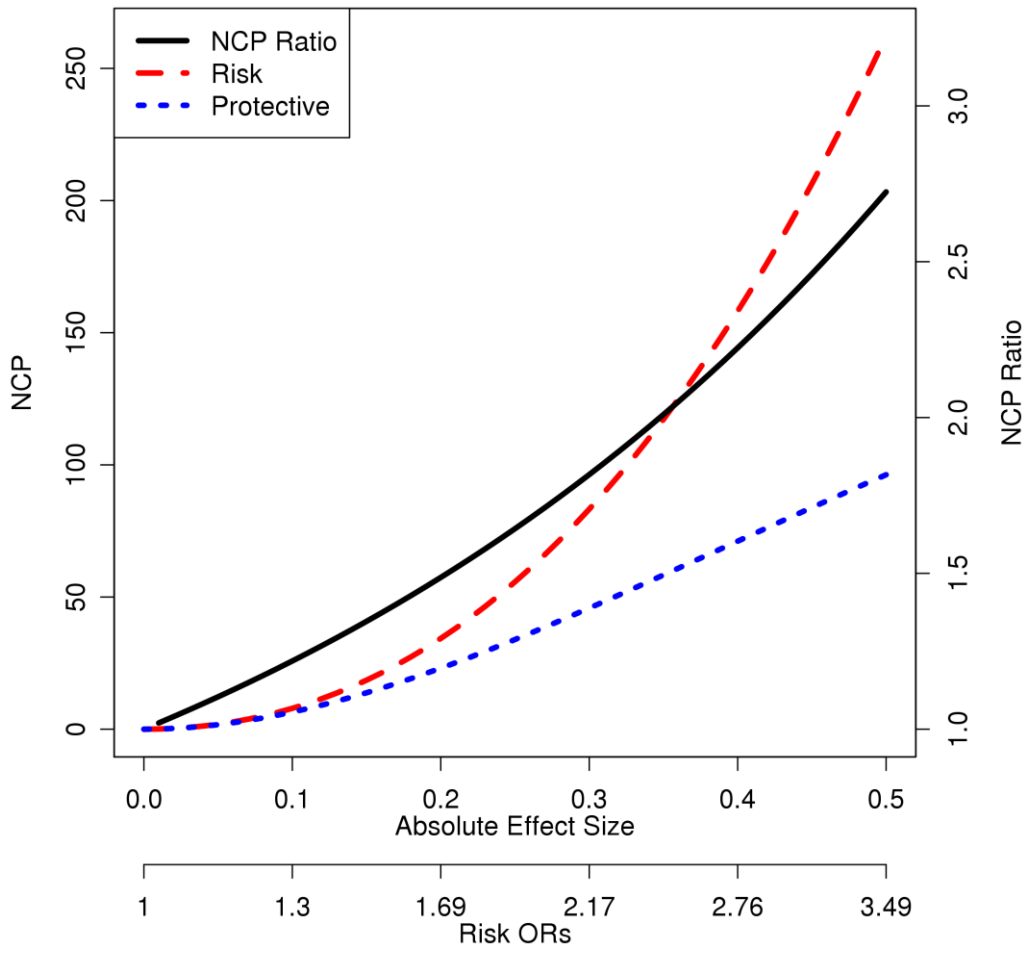


Figure 3.2: Comparing the power to detect risk and protective variants with the same underlying effect size. The plot shows the power as the non-centrality parameter (NCP) for detecting minor alleles that confer risk (risk variants) and minor alleles that confer protection (protective variants) with varying absolute effect sizes ($0 < \beta < 0.5$ in standard deviation units) using parameters from the base model (see Materials and Methods). It also shows the NCP Ratio, which is the NCP of risk variants divided by the NCP of protective variants with the same absolute effect size (right vertical axis). The equivalent odds-ratio (OR) for the risk variants is also shown on the horizontal axis.

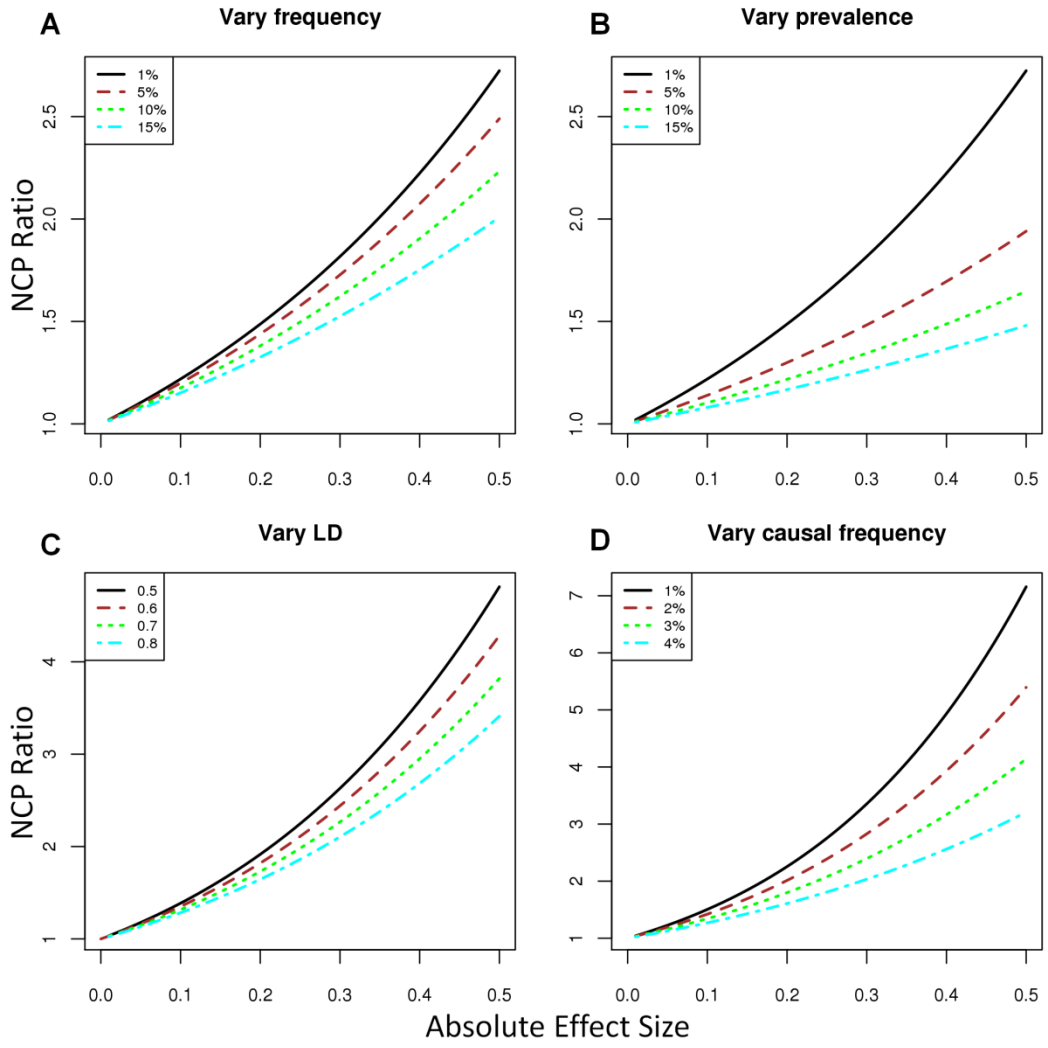


Figure 3.3: Effects of varying various parameters on the NCP Ratio. The plots show the difference in power for detecting risk versus protective variants through the NCP Ratio under varying parameters. Unless otherwise specified, the parameters used for calculating NCP are from the base model (see Materials and Methods). (**A**) Minor allele frequency of the associated variant varying from 1% to 15%. (**B**) Disease prevalence (threshold of liability) varying from 1% to 15%. (**C**) Linkage disequilibrium (LD) between the causal variant and the marker variant as a function of D' (varying from 0.5 to 0.8). (**D**) The marker variant frequency is set at 5% with the causal variant frequency ranging from 1% to 4%.

prevalence (Figure 3.3B), and there is no difference in power at any effect size when the disease prevalence is exactly 50%. Third, we varied the linkage disequilibrium (LD) between the associated variant and the causal variant from moderate LD (D' = 0.5) to strong LD (D' = 0.8). While there is a general loss of power with decreasing LD, the difference in power between risk and protective variants increases with decreasing LD (Figure 3.3C). Along similar lines, when we assumed that low frequency causal variants are being tagged by variants of higher frequencies (fixing the frequency of the tagged variant at 5% and varying the frequency of the causal variant from 4% to 1%), we also observed a greater difference in power as the causal variant frequency decreased (Figure 3.3D). These results show that the difference in power between risk and protective variants should be more obvious when testing variants within the low frequency range (< 5% frequency), in polygenic diseases with lower prevalence, and when the markers being tested are proxies for lower frequency causal variants. The driving force behind this result is that cases are ascertained from individuals with an extreme distribution of liability scores whereas controls have a much broader distribution of liability scores. Consequently, given an equal number of cases and controls, the increase in minor allele count of a risk variant in the cases is greater than the increase in minor allele count of an equally strong protective variant in the controls, leading to higher power for detecting the risk variant (see Appendix for derived formulae that confirm the increase in power). Thus, if rare or low frequency variants play a substantial role in certain diseases with polygenic architecture, these results predict that we could observe an increased R/P ratio for low frequency variants in the GWAS summary statistics for these diseases.

Excess of risk variants can be caused by negative selection

Beyond the differences in power, an excess of risk compared to protective variants can also occur if there is negative selection against the disease, leading risk variants to be kept at lower frequencies than protective variants. To illustrate this scenario, we simulated negative selection by coupling effects on evolutionary fitness and on a quantitative trait for a set of variants (frequency ≥ 1%), and then assigning case-control status based on the trait values (see Materials and Methods). We observed an increase in the R/P ratio for the frequency bins within 1% to 15% but not for the 30-50% frequency bin (Figure 3.4). These results show that under a model where rare variants contribute to disease and are under negative selection, we could also observe an increase in the R/P ratio for low frequency variants in the GWAS summary statistics for these diseases.

Excess of risk variants arise from having more controls than cases

The previous results show that polygenic inheritance from lower frequency variants can lead to an increase in the R/P ratio, but such an increase can also occur in other settings. Under the null hypothesis, one would expect that on average, the number of detected risk variants to be equal to the number of detected protective variants resulting in an expected R/P ratio of 1. However, in our simulations, we observed that the expected R/P ratio can deviate from 1 because of an imbalance between the number of cases and controls. Specifically, if there are substantially more controls than cases, a feature present in some GWAS of dichotomous traits, it would result in the increase of the expected R/P ratio (R/P ratio > 1). To illustrate this, we randomly simulated 1,000 cases and 3,000 controls (1k/3k) and measured the distribution of the R/P ratio under a null

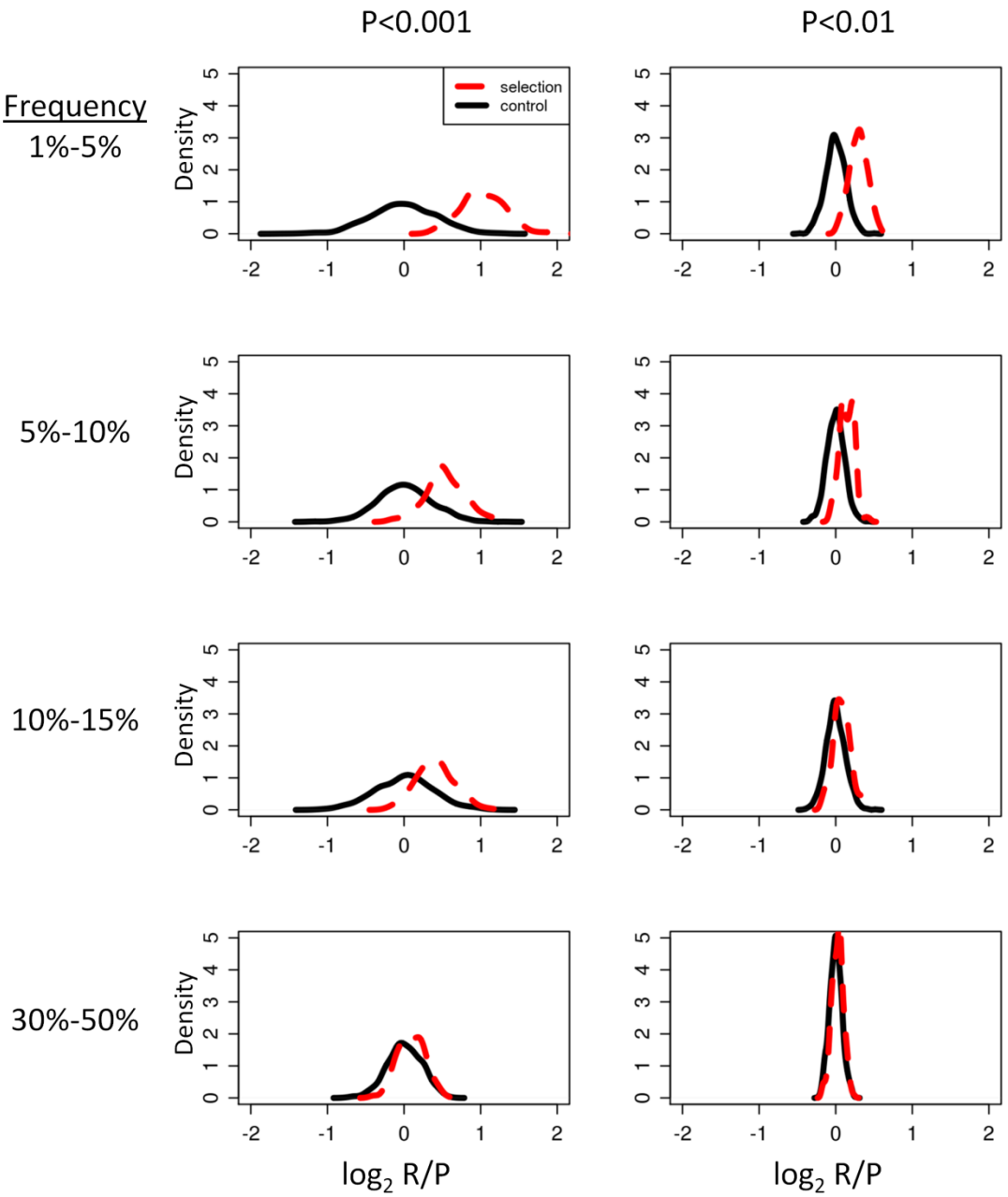


Figure 3.4: The distribution of the R/P ratio from simulating variants under negative selection. The figure shows the distribution of the $\log_2$ R/P ratio for various frequency bins and P-value cutoffs from simulating variants under negative selection. The selection model (red) uses the 823 effective variants while the control (black) model assumes no variants affect the phenotype.

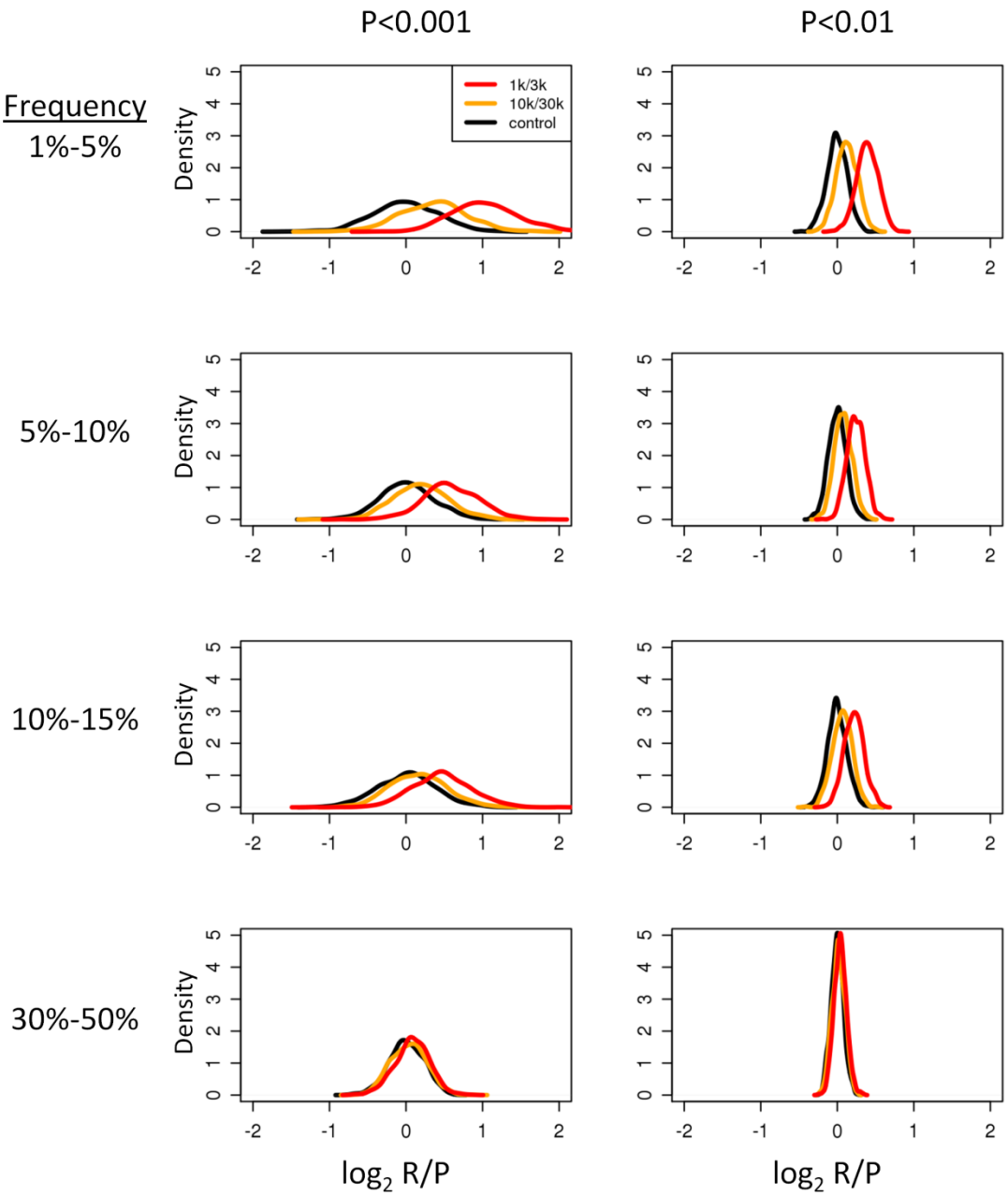


Figure 3.5: The null distribution of the R/P ratio with larger number of controls than cases. The figure shows the distribution of the $\log_2$ R/P ratio for various frequency bins and P-value cutoffs from simulating larger number of controls than cases. The 1k/3k (red) model simulates the null distribution of the $\log_2$ R/P ratio for 1,000 cases and 3,000 controls. The 10k/30k (orange) model simulates the null distribution of the $\log_2$ R/P ratio for 10,000 cases and 30,000 controls. The control (black) model simulates the null distribution of the $\log_2$ R/P ratio for 1,000 cases and 1,000 controls.

model of no association (see Materials and Methods). We observed that there is an increase in the R/P ratio distribution for 1k/3k for the low frequency bins (Figure 3.5). This increase is not seen with common variants (30-50% frequency bin), nor if the number of cases and controls are equal (Figure 3.5). Of note, with larger sample sizes (10,000 cases and 30,000 controls; 10k/30k, we observed that the increase in R/P ratio is substantially attenuated (Figure 3.5). These results show that an excess of controls can increase the expected R/P ratio, and should be accounted for by comparing the observed R/P ratio against those obtained through simulations under a null model. These results also show that with sufficiently large number of cases (e.g. > 10,000 cases), the increase in the expected R/P ratio due to this imbalance will be minimal.

Excess of risk variants can be due to asymmetric population stratification

We also considered whether an excess of risk variants could be seen in GWAS that are confounded by population stratification. As a first test, we randomly simulated 1,000 individuals of either northern European ancestry (CEU, based on allele frequencies in the CEU HapMap sample) or southern European ancestry (TSI, based on allele frequencies in the TSI HapMap sample). In one experiment, we simulated 1,000 CEU individuals as controls and 1,000 TSI individuals as cases (see Materials and Methods), and as a stratification-free experiment, we simulated 1,000 CEU controls and 1,000 CEU cases. We found that while there was a large excess of apparent associations for both risk and protective variants, leading to enormous inflation of the genomic control test statistic ($\lambda_{GC} \sim 22.9$), the resulting R/P ratio did not deviate substantially from expectations under the null (Figure 3.6). Therefore, even extreme scenarios with the usual forms of population stratification should not cause substantial deviations of the

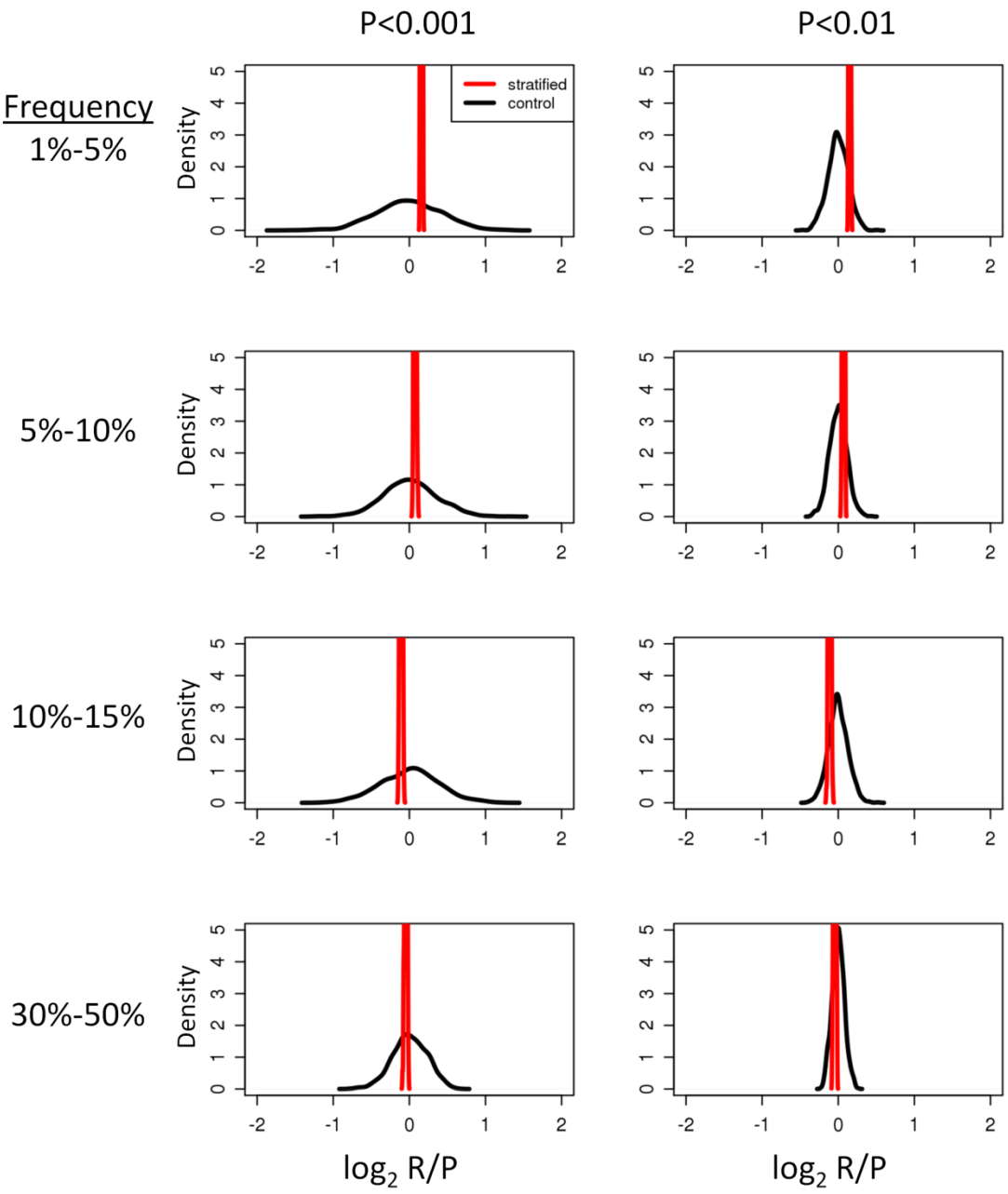


Figure 3.6: The distribution of the R/P ratio from simulating population stratification. The figure shows the distribution of the log2 R/P ratio for various frequency bins and P-value cutoffs from simulating population stratification. The stratification model (red) simulates the association perform with cases only from the TSI population and controls only from the CEU population. The control model (black) simulates both cases and controls from the CEU population.

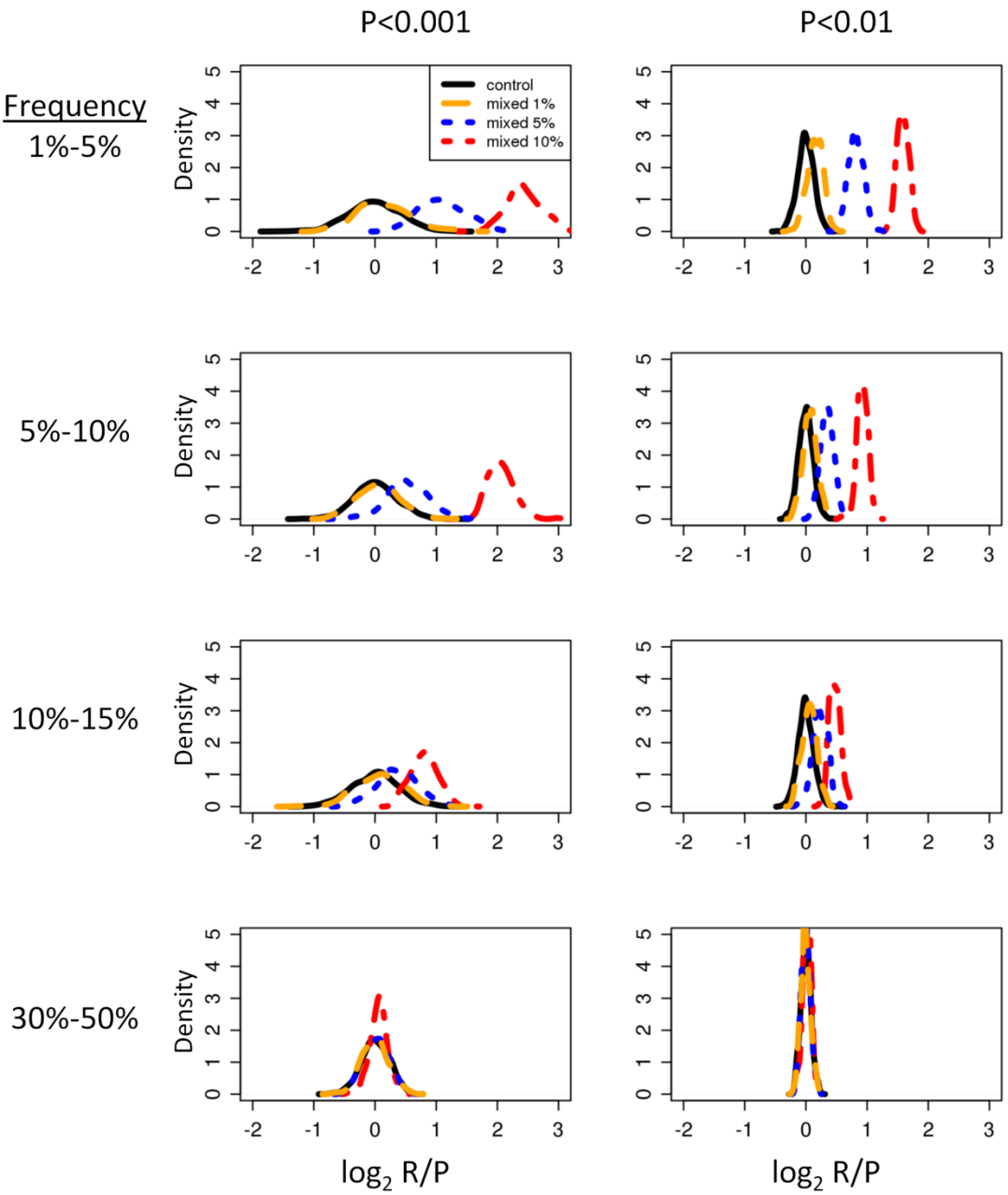


Figure 3.7: The distribution of the R/P ratio from simulating asymmetric population stratification. The figure shows the distribution of the log2 R/P ratio for various frequency bins and P-value cutoffs from simulating asymmetric population stratification. The models for asymmetric population stratification are as follows. Mixed 10%, 5% and 1% indicates 10%, 5% and 1% of the cases are TSI individuals respective while the rest of the individuals used are of CEU ancestry. The control model comprises of only CEUs without any population stratification.

R/P ratio. However, we reasoned that a special case of asymmetric population stratification could potentially cause the R/P ratio to depart from expectations under the null. Specifically, if there were a mixture of different populations only in cases but not in controls, or vice-versa, it could lead to an increase or decrease of the R/P ratio. To test this, we randomly simulated a series of models where controls are homogenous (CEU), while cases are a mixture of CEU and TSI (see Materials and Methods). At a 1% mixture in cases ($\lambda_{GC} \sim 1.01$), we did not observe any significant excess of risk variants, but at 5% mixture ($\lambda_{GC} \sim 1.06$), we observed an excess of risk variants within the low frequency ranges (Figure 3.7). This excess is even larger with a 10% mixture ($\lambda_{GC} \sim 1.24$) (Figure 3.7). Variants within the common frequency range do not show an excess of risk variants (Figure 3.7). These results show that such asymmetric population stratification can increase the R/P ratio, with only moderate increases in the genomic control statistics. As a corollary, if the mixture were to exist in controls but not cases, we would expect the R/P ratio to decrease.

Finally, we meta-analyzed the results from the asymmetrically stratified GWAS with results from non-stratified GWAS (see Materials and Methods) to determine the effect on the R/P ratio if only a subset of the studies had asymmetric population stratification. We found that the increase in the R/P ratio is attenuated after meta-analysis (Figure 3.8). These results indicate that while asymmetric population stratification can give rise to an excess of risk variants, combining such results with non-stratified results can reduce the magnitude of the signal. Because this particular type of stratification is unlikely to be present in most of the cohorts prior to meta-analysis, it may be useful to examine the summary statistics of each study individually to determine if the increased R/P ratio is derived from a subset of studies in the GWAS

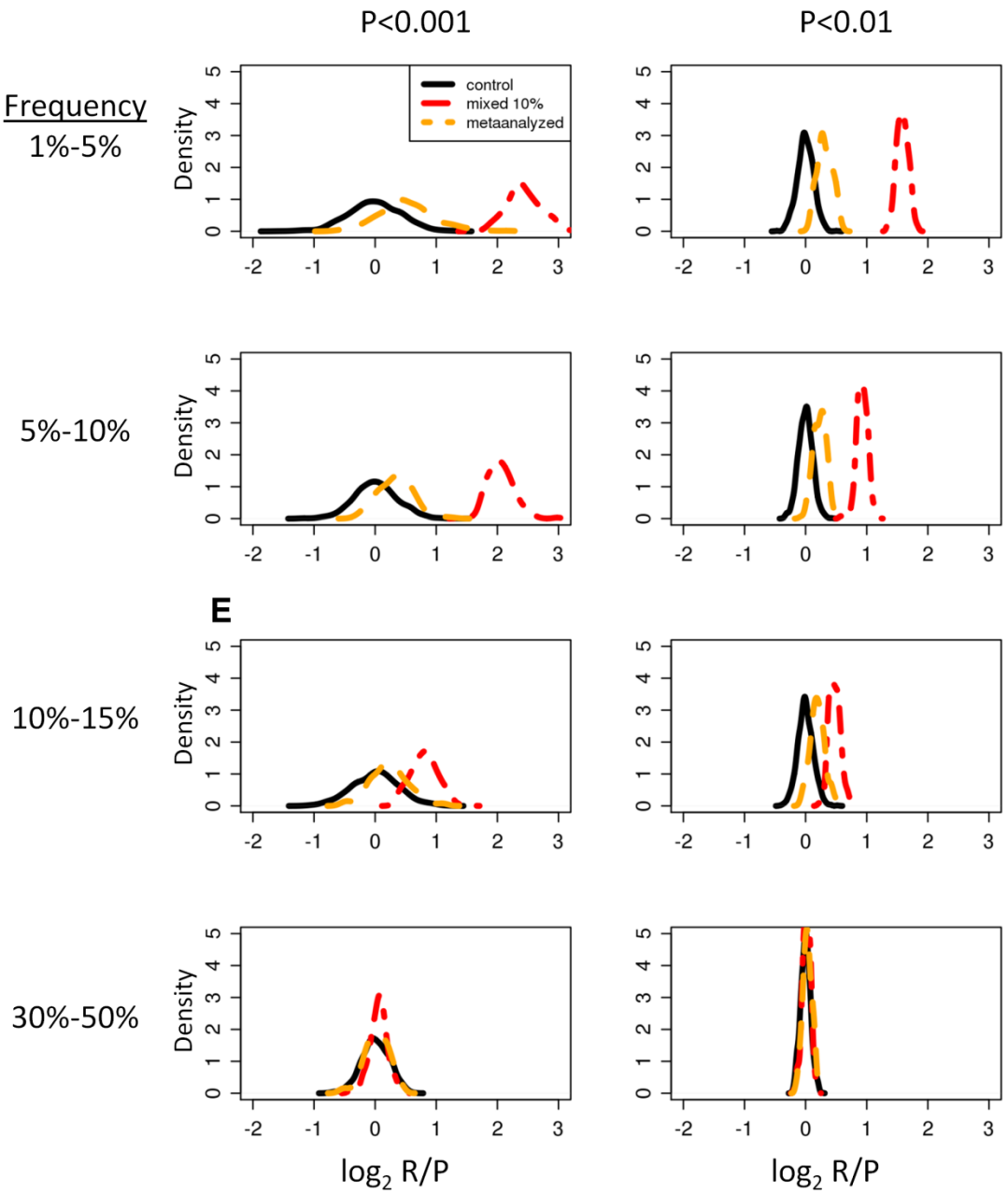


Figure 3.8: The distribution of the R/P ratio from simulating asymmetric population stratification after meta-analysis. The figure shows the distribution of the log2 R/P ratio for various frequency bins and P-value cutoffs from simulating asymmetric population stratification after meta-analysis with non-stratified data. The model "mixed 10%" and "metaanalyzed" refers to asymmetric population stratification of 10% mixture of TSI individuals of the cases before and after being meta-analyzed with 4 other datasets without such stratification respectively. The control model indicates no asymmetric population stratification.
meta-analysis.

Ideally, if an increased R/P ratio is observed, principal component analysis or other methods should also be applied to the primary data to search for outliers present exclusively in cases, to further rule out asymmetric population stratification as a cause of an increased R/P ratio.

Using the R/P ratio in actual GWAS results to search for signals of low frequency variants contributing to disease risk

Schizophrenia, major depressive disorder and bipolar disorder

We applied our method to data from several psychiatric disorders: schizophrenia [11], bipolar disorder [12] and major depressive disorder [13]. We observed a significant increase in the R/P ratio only for schizophrenia in the 1-5% frequency bin, at a cutoff of $P < 0.01$ ($P = 2.42 \times 10^{-7}$) (Table 3.1). We did not observe any significant differences in the other frequency bins nor for any of the other psychiatric disorders (Table 3.1). These results are indicative of polygenic inheritance from low frequency variants in schizophrenia but do not provide similar support for a role of low frequency variants in major depressive disorder or bipolar disorder.

Type 2 diabetes

Next, we applied our method to GWAS results of type 2 diabetes [14]. The R/P ratio for type 2 diabetes was significantly increased in the low frequency bins (Table 3.2). The most significant difference was observed in the 1-5% bin with cutoff of $P < 0.01$ ($P = 3.08 \times 10^{-15}$).

Table 3.1: Schizophrenia, Major depressive disorder and Bipolar disorder

		SCZ			MDD			BIP		
Freq (%)	Pvalue cutoff	*O(R/P)*	*E(R/P)*	*P*	*O(R/P)*	*E(R/P)*	*P*	*O(R/P)*	*E(R/P)*	*P*
1-5	0.001	1.864	1.127	0.0298	1.210	1.058	0.269	0.884	1.110	0.748
	0.01	1.623	1.032	**2.42e-7**	1.169	1.006	0.048	0.953	1.028	0.778
5-10	0.001	1.348	1.057	0.1279	0.933	1.039	0.623	1.038	1.077	0.509
	0.01	1.230	1.019	0.0111	0.914	1.005	0.865	0.973	1.013	0.678
10-15	0.001	1.050	1.082	0.4926	1.348	1.035	0.126	1.038	1.055	0.473
	0.01	1.054	1.019	0.3335	1.193	1.005	0.027	1.046	1.015	0.349
30-50	0.001	1.063	1.022	0.3736	1.098	1.003	0.264	1.122	1.039	0.291
	0.01	1.001	1.003	0.5010	0.944	1.001	0.836	1.070	1.009	0.165

The observed, expected R/P ratios and P-values obtained from analyzing GWAS summary statistics of psychiatric disorders: Schizophrenia (SCZ), major depressive disorder (MDD) and bipolar disorder (BIP). *O(R/P)* refers to the observed R/P ratio while *E(R/P)* refers to the expected R/P ratio obtained through simulations. *P* refers to the p-value obtained from a 1-tailed Z-test (In bold: $P < 0.01$).

Table 3.2: Type 2 diabetes

		T2D		
Freq (%)	Pvalue cutoff	*O(R/P)*	*E(R/P)*	*P*
1-5	0.001	3.833	1.205	**5.89e-6**
	0.01	2.009	1.069	**3.08e-15**
5-10	0.001	1.636	1.131	0.043
	0.01	1.439	1.051	**2.28e-5**
10-15	0.001	1.660	1.081	0.031
	0.01	1.400	1.033	**8.36e-4**
30-50	0.001	1.041	1.038	0.459
	0.01	1.035	1.008	0.308

The observed, expected R/P ratios and P-values obtained from analyzing GWAS summary statistics of type 2 diabetes (T2D). *O(R/P)* refers to the observed R/P ratio while *E(R/P)* refers to the expected R/P ratio obtained through simulations. *P* refers to the p-value obtained from a 1-tailed Z-test (In bold: $P < 0.01$).

We also observed a significant excess of risk variants in the 10-15% bin ($P < 0.01$, $P = 2.28 \times 10^{-5}$). As the difference in power between risk and protective variants becomes minimal as the variant frequency increases, this observed excess of risk variants is more likely due to negative selection on diabetes risk alleles, tagging of low frequency variants by the more common SNPs in this frequency range, and/or possibly asymmetric population stratification. Nonetheless, these results are indicative of polygenic inheritance from low frequency variants in type 2 diabetes.

Obesity

We also applied our method to GWAS results for various classes of obesity [15]: overweight (BMI > 25), class 1 (BMI > 30), class 2 (BMI > 35) and class 3 (BMI > 40). The controls used for each class of obesity were individuals with BMI < 25. We observed a significant increase in the 1-5% frequency bin with a cutoff of $P < 0.01$ for only the class 1 dataset ($P = 8.8 \times 10^{-6}$) (Table 3.3). Also, while we generally observed a gradual increase in the R/P ratio with increasing BMI definitions of obesity, which could be consistent with a role of lower frequency variants, the increase in R/P ratio could also be explained by having more controls than cases. We did not observe any significant excess of risk variants for the low frequency bins in the class 2 or class 3 datasets, likely because of the severely reduced sample sizes for the more extreme BMI definitions of obesity.

Testing whether related phenotypes are likely to share low frequency causal variants

To increase the power of GWAS, some studies have pooled apparently related phenotypes

Table 3.3: Obesity

		Overweight			**Class1**		
Freq (%)	Pvalue cutoff	*O(R/P)*	*E(R/P)*	*P*	*O(R/P)*	*E(R/P)*	*P*
1-5	0.001	1.188	0.997	0.228	0.917	1.164	0.758
	0.01	1.120	0.986	0.078	1.536	1.050	**8.8e-6**
5-10	0.001	1.026	0.998	0.408	1.139	1.098	0.393
	0.01	1.023	0.991	0.328	0.937	1.023	0.838
10-15	0.001	0.784	0.999	0.826	0.971	1.087	0.610
	0.01	1.109	1.003	0.113	1.013	1.028	0.544
30-50	0.001	1.121	0.991	0.194	1.059	1.020	0.380
	0.01	1.022	0.999	0.340	1.045	1.004	0.225
		Class2			**Class3**		
Freq (%)	Pvalue cutoff	*O(R/P)*	*E(R/P)*	*P*	*E(R/P)*	*O(R/P)*	*P*
1-5	0.001	2.462	2.410	0.410	3.700	3.454	0.354
	0.01	1.533	1.376	0.114	1.814	1.617	0.111
5-10	0.001	0.697	1.640	0.999	1.857	2.067	0.607
	0.01	1.108	1.222	0.871	1.227	1.346	0.845
10-15	0.001	1.276	1.567	0.713	1.385	1.766	0.779
	0.01	1.066	1.208	0.883	1.269	1.267	0.479
30-50	0.001	0.949	1.094	0.763	1.019	1.112	0.696
	0.01	0.985	1.035	0.816	0.955	1.044	0.946

The observed, expected R/P ratios and P-values obtained from analyzing GWAS summary statistics of clinical classes of obesity: Overweight (BMI > 25), Class1 (BMI > 30), Class2 (BMI > 35) and Class3 (BMI > 40). *O(R/P)* refers to the observed R/P ratio while *E(R/P)* refers to the expected R/P ratio obtained through simulations. *P* refers to the p-value obtained from a 1-tailed Z-test (In bold: $P < 0.01$).

into a single case group [16,17]. We applied our method to measure the R/P ratio on published GWAS results of these related phenotypes. We reasoned that our method could also be used to test if pooling related phenotypes would increase power to detect low frequency variants, using only the GWAS summary statistics. We applied our method to GWAS results from two different pairs of related phenotypes, one pair for inflammatory bowel disease and one pair for diabetic nephropathy.

Inflammatory bowel disease

The two major types of inflammatory bowel disease are Crohn's disease (CD) and ulcerative colitis (UC)[36]. We examined the R/P ratio in GWAS results for Crohn's disease [32], ulcerative colitis[33] and the combined case cohort of both Crohn's disease and ulcerative colitis [17]. We observed significant increases in the R/P ratio for both Crohn's disease and ulcerative colitis within the low frequency bins (Table 3.4). The most significant increases were found in the 1-5% bin with cutoff of $P < 0.01$ (CD: $P = 1.55 \times 10^{-10}$, UC: $P = 2.25 \times 10^{-9}$), consistent with a polygenic role of low frequency variants in both diseases. However, when Crohn's disease and ulcerative colitis were combined as a single case group (CD + UC), the increase in R/P ratio is less significant than in the individual GWAS results (Table 3.4). These results suggest that there are some low frequency genetic contributors to Crohn's disease and ulcerative colitis that are not shared by both diseases. However, because the signal is still present (albeit attenuated) when both diseases were studied together, it also suggests that the two diseases do share some overlapping low frequency genetic contributors, although the attenuated signal could reflect

Table 3.4: Inflammatory bowel disease: Crohn's disease and Ulcerative Colitis

		CD			UC			CD+UC		
Freq (%)	Pvalue cutoff	*O(R/P)*	*E(R/P)*	*P*	*O(R/P)*	*E(R/P)*	*P*	*O(R/P)*	*E(R/P)*	*P*
1-5	0.001	2.545	1.347	0.017	1.958	1.358	0.075	1.385	1.159	0.222
	0.01	1.994	1.111	**1.55e-10**	1.866	1.106	**2.25e-9**	1.457	1.048	**1.6e-4**
5-10	0.001	1.148	1.162	0.477	1.490	1.192	0.153	1.099	1.107	0.463
	0.01	1.314	1.069	**1.4e-3**	1.460	1.066	**8.59e-5**	1.239	1.027	0.012
10-15	0.001	1.200	1.181	0.424	1.279	1.186	0.337	1.583	1.076	0.059
	0.01	1.043	1.059	0.551	1.213	1.066	0.075	1.104	1.026	0.205
30-50	0.001	0.925	1.035	0.743	1.163	1.037	0.217	1.036	1.026	0.445
	0.01	1.052	1.018	0.266	1.004	1.009	0.524	1.043	1.005	0.251

The observed, expected R/P ratios and P-values obtained from analyzing GWAS summary statistics of inflammatory bowel diseases: Crohn's disease (CD), Ulcerative colitis (UC) and the combined CD and UC as a single case group (CD+UC). *O(R/P)* refers to the observed R/P ratio while *E(R/P)* refers to the expected R/P ratio obtained through simulations. *P* refers to the p-value obtained from a 1-tailed Z-test (In bold: $P < 0.01$).

Table 3.5: Diabetic Nephropathy: Macroalbuminuria and End stage renal disease

		$MACRO_{ctrl}$			**$ESRD_{ctrl}$**		
Freq (%)	Pvalue cutoff	*O(R/P)*	*E(R/P)*	*P*	*O(R/P)*	*E(R/P)*	*P*
1-5	0.001	2.000	1.655	0.205	1.944	1.706	0.283
	0.01	1.560	1.198	**1.4e-3**	1.705	1.207	**6.4e-5**
5-10	0.001	1.563	1.359	0.253	1.278	1.404	0.585
	0.01	1.200	1.116	0.175	1.240	1.143	0.147
10-15	0.001	0.893	1.275	0.892	1.343	1.304	0.403
	0.01	1.208	1.104	0.150	1.190	1.128	0.258
30-50	0.001	1.122	1.066	0.343	1.198	1.051	0.197
	0.01	0.990	1.023	0.690	1.152	1.014	0.017
		$ESRD_{ctrl+macro}$			**$[MACRO + ESRD]_{ctrl}$**		
Freq (%)	Pvalue cutoff	*O(R/P)*	*E(R/P)*	*P*	*E(R/P)*	*O(R/P)*	*P*
1-5	0.001	2.667	2.008	0.146	1.087	1.133	0.504
	0.01	2.270	1.285	**9e-11**	1.026	1.042	0.550
5-10	0.001	1.533	1.584	0.496	0.875	1.071	0.754
	0.01	1.552	1.187	**2.9e-4**	1.045	1.017	0.352
10-15	0.001	1.462	1.397	0.380	0.912	1.038	0.640
	0.01	1.310	1.160	0.078	1.053	1.009	0.290
30-50	0.001	0.968	1.076	0.719	1.037	1.001	0.382
	0.01	1.038	1.032	0.449	0.981	1.003	0.652

The observed, expected R/P ratios and P-values obtained from analyzing GWAS summary statistics of diabetic nephropathy: macroalbuminuria ($MACRO_{ctrls}$), end stage renal disease ($ESRD_{ctrls}$), ESRD versus controls that include MACRO ($ESRD_{ctrls+macro}$) and the combined MACRO and ESRD as a single case group ($[MACRO + ESRD]_{ctrls}$). *O(R/P)* refers to the observed R/P ratio while *E(R/P)* refers to the expected R/P ratio obtained through simulations. *P* refers to the p-value obtained from a 1-tailed Z-test (In bold: $P < 0.01$).

persistence of two separate individual signals that are diluted after combination of the two sets of cases.

Diabetic nephropathy

We performed a similar analysis on two phenotypes used to characterize diabetic nephropathy[16]: macroalbuminuria (MACRO) and end stage renal disease (ESRD). Unlike inflammatory bowel disease, MACRO and ESRD are not necessarily distinct; MACRO is a milder form of diabetic nephropathy and some of those individuals progress to develop ESRD. The controls used for that study were diabetic individuals that did not develop nephropathy. We analyzed the GWAS results performed for individuals with macroalbuminuria versus controls ($MACRO_{ctrl}$), individuals with end stage renal disease versus controls ($ESRD_{ctrl}$), individuals with end stage renal disease versus controls that also include individuals with macroalbuminuria ($ESRD_{ctrl+macro}$) and a combined case cohort that includes both individuals with macroalbuminuria and end stage renal disease versus controls ($[MACRO + ESRD]_{ctrl}$). For the analyses of $MACRO_{ctrl}$ and of $ESRD_{ctrl}$, we observed significant increases to the R/P ratio in the 1-5% bin with cutoff of $P < 0.01$ ($MACRO_{ctrl}$: $P = 0.001$, $ESRD_{ctrl}$: $P = 6.4 \times 10^{-5}$) (Table 3.5). For the $ESRD_{ctrl+macro}$ analysis, where individuals with macroalbuminuria are included within the controls, there is an even larger increase of the R/P ratio ($ESRD_{ctrl+macro}$: $P = 9 \times 10^{-11}$) (Table 3.5). However, when $MACRO_{ctrl}$ and $ESRD_{ctrl}$ were combined into a single case group ($[MACRO + ESRD]_{ctrl}$), none of the frequency bins showed significant increases in the R/P ratio (Table 3.5). These results suggest that while there are low frequency contributors to both macroalbuminuria and end stage renal disease, these contributors do not substantially overlap.

There is no detectable increase in the R/P ratio when both phenotypes are combined, unlike our observations for inflammatory bowel disease. Thus, these results indicate that studies of low frequency variation for diabetic nephropathy would be more fruitful if MACRO and ESRD are tested separately.

DISCUSSION

We have shown that our method for measuring the R/P ratio can be used as a test for the presence of multiple low frequency or rare genetic contributors to disease risk. This method can be applied to GWAS summary statistics, even if there are few or no genome-wide significant associations. We analyzed results from multiple published GWAS studies, and found significant signals in some but not all diseases. These results support the hypotheses that the diseases where the R/P ratio is increased have a polygenic contribution from as-yet undetected low frequency or rare variants.

Some existing methods for detecting polygenic inheritance [9,10,37] use variants that achieve nominal significance in GWAS to determine if they are informative as predictors of phenotype. Because our method assesses the direction of effect of these variants against the null model, our method represents a rather different, independent approach for assessing polygenic inheritance of low frequency variants. Furthermore, our method does not require having identified associated loci or the availability of individual level data. For example, in schizophrenia, it has been shown that a substantial proportion of schizophrenia disease risk is the result of variants with frequency > 1% [38]. Our finding suggests that some of disease risk is accounted for by variants within the low frequency range (frequency < 5%). In a recent exome

sequencing study of 2,536 schizophrenia cases and 2,543 controls[39], Purcell and colleagues showed a polygenic burden of rare disruptive mutations, which is consistent with our observation. Similarly, for type 2 diabetes, our results suggest the presence of low frequency or rare variants contributing to disease risk, even though most of the variants known to be associated with disease risk are common (frequency ≥ 5%) [14].

We also showed that negative selection under polygenic inheritance can increase the R/P ratio for low frequency variants, because risk variants would be kept at lower frequencies while the protective variants could drift to higher frequencies. Indeed, in a previous study [40], Park and colleagues showed that across most qualitative traits, minor alleles conferred risk more often than protection which they concluded to be evidence for purifying selection. While this can be the case for some diseases, we also showed that this increase in the R/P ratio can also arise because there is more power to detect risk variants than protective variants. Furthermore, we have established that if there are substantially more controls than cases, a feature present in many GWAS, this imbalance can distort the null distribution such that there would appear to be more risk than protective variants. However, this imbalance can be accounted for through simulations, as we have demonstrated.

Our method also provides a simple and early way of assessing the utility of different phenotype definitions for genetic studies of low frequency variation simply from GWAS summary statistics. Our results for inflammatory bowel disease are consistent with the idea that Crohn's disease and ulcerative colitis have some overlapping genetic contributors. Indeed, a previous study exploring the effect of common Crohn's disease variants on ulcerative colitis identified significant overlaps between the two diseases, but also loci specific to Crohn's disease [41]. For diabetic nephropathy, where there are few established loci from which to draw

conclusions from, we observed signals for both macroalbuminuria and particularly for end stage renal disease when analyzed separately, but no significant signal when both diseases were combined as a single case group. This suggests that macroalbuminuria and end-stage renal disease are distinct in their genetic architecture and would be more productive if they were to be studied separately. Interestingly, the same GWAS on diabetic nephropathy discovered a single genome-wide significant locus only when end stage renal disease was treated separately from macroalbuminuria [16], consistent with our observation.

Finally, asymmetric population stratification between cases and controls can lead to both false positive associations (as evidenced by an increased genomic control inflation factor) [42], and also an increase in the R/P ratio. Thus, while our observations of higher than expected R/P ratios in some of the published GWAS datasets are suggestive of a role of low frequency variants, we cannot completely rule out that some of these signals could be in part explained by asymmetric population stratification. Of note, none of the R/P ratios showed a deficit of risk variants (which would be expected under some models of asymmetric population stratification), suggesting that asymmetric population stratification is not widespread. Furthermore, these GWAS have used methods to detect and correct for population stratification.

In conclusion, our method can be used to screen for polygenic inheritance from low frequency or rare variants in diseases where GWAS have been performed. Our method can also be extended to other summary statistics, e.g. studies from sequencing or exome-chip genotyping, to assess low frequency variants that were directly genotyped rather than imputed. This method can serve as a simple approach to guide researchers in prioritizing strategies in searching for as yet unexplained heritability for specific diseases. For example, in a study of epilepsy [43], Heinzen and colleagues failed to identify any rare variants of large effect through exome

sequencing; analysis of GWAS data for epilepsy can in theory help guide decisions about embarking on additional studies of low frequency or rare variants with larger sample sizes. Although a lack of a signal from our method does not rule out a role for low frequency variants, and may reflect a combination of small sample sizes, and a set of effect sizes and frequencies that do not significantly alter the R/P ratio, a positive signal can provide greater confidence about the likelihood that low frequency or rare variants contribute to disease risk.

APPENDIX

Calculating NCP from various given parameters

We define the following parameters required to calculate the non-centrality parameter (NCP) as a function of effect size of minor allele (β), minor allele frequency (p), liability threshold (t), number of case individuals (N_d) and number of control individuals (N_c). We denote the minor allele (effect allele) as a_1 and the major allele (non-effect allele) as a_2. As such, the liability distribution of a_1 is $N(x, \mu_1, \sigma^2)$ and the liability distribution of a_2 is $N(x, \mu_2, \sigma^2)$ such that $N(x, \mu, \sigma^2)$ is the probability density function of a normal distribution with mean μ and variance σ^2.

The mean liabilities for a_1 and a_2 are as follows:

$$\text{Mean liability for } a_1 = \mu_1 = \beta - \beta p = \beta q$$

$$\text{Mean liability for } a_2 = \mu_2 = -\beta p$$

where q is the major allele frequency such that $p + q = 1$. The variance remaining σ^2 is:

$$\text{Variance remaining} = \sigma^2 = 1 - \beta^2 p q$$

Next, we calculate a series of conditional probabilities as follows:

$$P(case|a_1) = \int_t^{\infty} N(x, \mu_1, \sigma^2)\ \mathrm{dx}$$

$$P(case|a_2) = \int_t^{\infty} N(x, \mu_2, \sigma^2)\ \mathrm{dx}$$

$$P(control\ |a_1) = \int_{-\infty}^{t} N(x, \mu_1, \sigma^2)\ \mathrm{dx}$$

$$P(control\ |a_2) = \int_{-\infty}^{t} N(x, \mu_2, \sigma^2)\ \mathrm{dx}$$

With these conditional probabilities, we proceed to calculate the expected allele frequencies of both the minor allele and major allele in both cases and controls using Bayes' theorem. These are calculated as:

$$P_{d1} = P(a_1|case) = \frac{P(case|a_1)\ p}{\int_t^{\infty} N(x, 0,1)dx}$$

$$P_{d2} = P(a_2|case) = \ 1 - P_{d1}$$

$$P_{c1} = \ P(a_1|control) = \ \frac{P(control|a_1)\ p}{\int_{-\infty}^{t} N(x, 0,1)dx}$$

$$P_{c2} = P(a_2|control) = \ 1 - P_{c1}$$

We then calculate the NCP by the χ^2 statistic from a 2 by 2 contingency table for the expectation of the observed number of a_1 and a_2 in both cases and controls.

	Case	Control	Total
a_1	$2\,N_d\,P_{d1}$	$2\,N_c\,P_{c1}$	$2\,A$
a_2	$2\,N_d\,(1-P_{d1})$	$2\,N_c\,(1-P_{c1})$	$2\,B$
Total	$2\,N_d$	$2\,N_c$	$2\,T$

where,

$$A = N_d\,P_{d1} + N_c\,P_{c1}$$

$$B = N_d\,(1-P_{d1}) + N_c\,(1-P_{c1})$$

$$T = A + B = N_d + N_c$$

The expected number for each cell is the row total times the column total divided by the grand total.

Thus, the NCP is calculated as:

$$NCP = \sum_{Each\ cell} \frac{(Observed - Expected)^2}{Expected}$$

$$NCP = \frac{\left(2\,N_d\,P_{d1} - \frac{4AN_d}{2T}\right)^2}{\frac{4AN_d}{2T}} + \frac{\left(2\,N_c\,P_{c1} - \frac{4AN_c}{2T}\right)^2}{\frac{4AN_c}{2T}} + \frac{\left(2\,N_d\,P_{d2} - \frac{4BN_d}{2T}\right)^2}{\frac{4BN_d}{2T}} + \frac{\left(2\,N_c\,P_{c2} - \frac{4BN_c}{2T}\right)^2}{\frac{4BN_c}{2T}}$$

$$NCP = \frac{2TN_d{P_{d1}}^2}{A} + \frac{2AN_d}{T} - 4N_dP_{d1} + \frac{2TN_c{P_{c1}}^2}{A} + \frac{2AN_c}{T} - 4N_cP_{c1} + \frac{2TN_d(1-P_{d1})^2}{B} + \frac{2BN_d}{T}$$

$$- 4N_d(1-P_{d1}) + \frac{2TN_c(1-P_{c1})^2}{B} + \frac{2BN_c}{T} - 4N_c(1-P_{c1})$$

$$NCP = \frac{2TN_d{P_{d1}}^2}{A} + \frac{2TN_c{P_{c1}}^2}{A} + \frac{2TN_d(1-P_{d1})^2}{B} + \frac{2TN_c(1-P_{c1})^2}{B} - 2T$$

After some algebra and simplification,

$$NCP = \frac{2T}{AB} N_d N_c (P_{d1} - P_{c1})^2$$

Therefore,

$$NCP = 2N_d N_c (P_{d1} - P_{c1})^2 \left(\frac{N_d + N_c}{(N_d P_{d1} + N_c P_{c1})(N_d P_{d2} + N_c P_{c2})}\right)$$

We verified that these formulae were correct by comparing to simulated results.

Determining NCP ratio between risk and protective variants with the same magnitude of effect

We formulated the various probabilities between risk and protective variants. Assuming β to be positive, the risk variant would have the following probabilities,

$$P_{d1} = \frac{p\int_t^{\infty} N(x, \beta q, \sigma^2)\mathrm{dx}}{\int_t^{\infty} N(x, 0, 1)\mathrm{dx}}$$

$$P_{c1} = \frac{p\int_{-\infty}^{t} N(x, \beta q, \sigma^2)\mathrm{dx}}{\int_{-\infty}^{t} N(x, 0, 1)\mathrm{dx}}$$

and the protective variant with the same magnitude of effect would have the following probabilities,

$$P_{d1} = \frac{p\int_t^{\infty} N(x, -\beta q, \sigma^2)\mathrm{dx}}{\int_t^{\infty} N(x, 0, 1)\mathrm{dx}}$$

$$P_{c1} = \frac{p\int_{-\infty}^{t} N(x,-\beta q,\sigma^2)\mathrm{dx}}{\int_{-\infty}^{t} N(x,0,1)\mathrm{dx}}$$

Assuming that there are equal number of cases and controls ($N_1 = N_2$), then

$$NCP\ \alpha\ \left(\frac{\int_{t}^{\infty} N(x,\beta q,\sigma^2)\mathrm{dx}}{\int_{t}^{\infty} N(x,0,1)\mathrm{dx}} - \frac{\int_{-\infty}^{t} N(x,\beta q,\sigma^2)\mathrm{dx}}{\int_{-\infty}^{t} N(x,0,1)\mathrm{dx}}\right)^2$$

The ratio between risk and protective variants with the similar magnitude of β is therefore

$$NCP\ ratio = \frac{\left(\frac{\int_{t}^{\infty} N(x,\beta q,\sigma^2)\mathrm{dx}}{\int_{t}^{\infty} N(x,0,1)\mathrm{dx}} - \frac{\int_{-\infty}^{t} N(x,\beta q,\sigma^2)\mathrm{dx}}{\int_{-\infty}^{t} N(x,0,1)\mathrm{dx}}\right)^2}{\left(\frac{\int_{t}^{\infty} N(x,-\beta q,\sigma^2)\mathrm{dx}}{\int_{t}^{\infty} N(x,0,1)\mathrm{dx}} - \frac{\int_{-\infty}^{t} N(x,-\beta q,\sigma^2)\mathrm{dx}}{\int_{-\infty}^{t} N(x,0,1)\mathrm{dx}}\right)^2}$$

We can transform the distributions such that,

$$NCP\ ratio = \frac{\sigma\left(\frac{\int_{\frac{t-\beta q}{\sigma}}^{\infty} N(z,0,1)\mathrm{dz}}{\int_{t}^{\infty} N(x,0,1)\mathrm{dx}} - \frac{\int_{-\infty}^{\frac{t-\beta q}{\sigma}} N(z,0,1)\mathrm{dz}}{\int_{-\infty}^{t} N(x,0,1)\mathrm{dx}}\right)^2}{\sigma\left(\frac{\int_{\frac{t+\beta q}{\sigma}}^{\infty} N(y,0,1)\mathrm{dy}}{\int_{t}^{\infty} N(x,0,1)\mathrm{dx}} - \frac{\int_{-\infty}^{\frac{t+\beta q}{\sigma}} N(y,0,1)\mathrm{dy}}{\int_{-\infty}^{t} N(x,0,1)\mathrm{dx}}\right)^2}$$

where $z = \frac{x-\beta q}{\sigma}$, $y = \frac{x+\beta q}{\sigma}$ and $dx = \sigma dz = \sigma dy$

Then,

$$NCP\ ratio = \frac{\left(\frac{\int_{\frac{t-\beta q}{\sigma}}^{t} N(z,0,1)\mathrm{dz} + \int_{t}^{\infty} N(z,0,1)\mathrm{dz}}{\int_{t}^{\infty} N(x,0,1)\mathrm{dx}} - \frac{\int_{-\infty}^{t} N(z,0,1)\mathrm{dz} - \int_{\frac{t-\beta q}{\sigma}}^{t} N(z,0,1)\mathrm{dz}}{\int_{-\infty}^{t} N(x,0,1)\mathrm{dx}}\right)^2}{\left(\frac{\int_{t}^{\infty} N(y,0,1)\mathrm{dy} - \int_{t}^{\frac{t+\beta q}{\sigma}} N(y,0,1)\mathrm{dy}}{\int_{t}^{\infty} N(x,0,1)\mathrm{dx}} - \frac{\int_{-\infty}^{t} N(y,0,1)\mathrm{dy} + \int_{t}^{\frac{t+\beta q}{\sigma}} N(y,0,1)\mathrm{dy}}{\int_{-\infty}^{t} N(x,0,1)\mathrm{dx}}\right)^2}$$

$$NCP\ ratio = \frac{(1 + \frac{\int_{\frac{t-\beta q}{\sigma}}^{t} N(z,0,1)\mathrm{dz}}{\int_{t}^{\infty} N(x,0,1)\mathrm{dx}} - (1 - \frac{\int_{\frac{t-\beta q}{\sigma}}^{t} N(z,0,1)\mathrm{dz}}{\int_{-\infty}^{t} N(x,0,1)\mathrm{dx}}))^2}{(1 - \frac{\int_{t}^{\frac{t+\beta q}{\sigma}} N(y,0,1)\mathrm{dy}}{\int_{t}^{\infty} N(x,0,1)\mathrm{dx}} - (1 + \frac{\int_{t}^{\frac{t+\beta q}{\sigma}} N(y,0,1)\mathrm{dy}}{\int_{-\infty}^{t} N(x,0,1)\mathrm{dx}}))^2}$$

$$NCP\ ratio = \frac{(\frac{\int_{\frac{t-\beta q}{\sigma}}^{t} N(z,0,1)\mathrm{dz}}{\int_{t}^{\infty} N(x,0,1)\mathrm{dx}} + \frac{\int_{\frac{t-\beta q}{\sigma}}^{t} N(z,0,1)\mathrm{dz}}{\int_{-\infty}^{t} N(x,0,1)\mathrm{dx}})^2}{(-1^2)(\frac{\int_{t}^{\frac{t+\beta q}{\sigma}} N(y,0,1)\mathrm{dy}}{\int_{t}^{\infty} N(x,0,1)\mathrm{dx}} + \frac{\int_{t}^{\frac{t+\beta q}{\sigma}} N(y,0,1)\mathrm{dy}}{\int_{-\infty}^{t} N(x,0,1)\mathrm{dx}})^2}$$

$$NCP\ ratio = \frac{(\int_{\frac{t-\beta q}{\sigma}}^{t} N(z,0,1)\mathrm{dz})^2}{(\int_{t}^{\frac{t+\beta q}{\sigma}} N(y,0,1)\mathrm{dy})^2}$$

When prevalence is 50% (*t=0*),

$$\int_{\frac{-\beta q}{\sigma}}^{0} N(z,0,1)\mathrm{dz} = \int_{0}^{\frac{+\beta q}{\sigma}} N(y,0,1)\mathrm{dy}$$

and therefore

$$NCP\ ratio = 1$$

This shows that when prevalence is 50% (*t*=0) and there are equal sample numbers in cases and controls ($N_1 = N_2$), the NCP between risk and protective variants with identical magnitudes of effect (β) would be the same regardless of any other parameters.

For the case where t > 0, if

$$\int_{\frac{t-\beta q}{\sigma}}^{t} N(z,0,1)\mathrm{dz} - \int_{t}^{\frac{t+\beta q}{\sigma}} N(y,0,1)\mathrm{dy} > 0$$

then the NCP for risk variants will be greater than the NCP for protective variants and the NCP ratio will be greater than 1. When $t > \beta q$, this will be true because the normal distribution is monotonic decreasing above z=0 (y=0).

To extend this to the more general case of t>0, we first examine the individual components,

$$\int_{\frac{t-\beta q}{\sigma}}^{t} N(z,0,1)\mathrm{dz} = \int_{-\infty}^{t} N(z,0,1)\mathrm{dz} - \int_{-\infty}^{\frac{t-\beta q}{\sigma}} N(z,0,1)\mathrm{dz}$$

$$= \frac{1}{2}[1 + \mathrm{eft}\left(\frac{\mathrm{t}}{\sqrt{2}}\right)] - \frac{1}{2}[1 + \mathrm{eft}\left(\frac{\mathrm{t}-\beta q}{\sigma\sqrt{2}}\right)]$$

$$= \frac{1}{2}[\mathrm{eft}\left(\frac{\mathrm{t}}{\sqrt{2}}\right) - \mathrm{eft}\left(\frac{\mathrm{t}-\beta q}{\sigma\sqrt{2}}\right)]$$

where eft is the error function. Similarly,

$$\int_{t}^{\frac{t+\beta q}{\sigma}} N(y,0,1)\mathrm{dy} = \int_{-\infty}^{\frac{t+\beta q}{\sigma}} N(y,0,1)\mathrm{dy} - \int_{-\infty}^{t} N(y,0,1)\mathrm{dy}$$

$$= \frac{1}{2}[1 + \mathrm{eft}\left(\frac{\mathrm{t}+\beta q}{\sigma\sqrt{2}}\right)] - \frac{1}{2}[1 + \mathrm{eft}\left(\frac{\mathrm{t}}{\sqrt{2}}\right)]$$

$$= \frac{1}{2}[\mathrm{eft}\left(\frac{\mathrm{t}+\beta q}{\sigma\sqrt{2}}\right) - \mathrm{eft}\left(\frac{\mathrm{t}}{\sqrt{2}}\right)]$$

Therefore,

$$\int_{\frac{t-\beta q}{\sigma}}^{t} N(z,0,1)\mathrm{dz} - \int_{t}^{\frac{t+\beta q}{\sigma}} N(y,0,1)\mathrm{dy}$$

$$= \frac{1}{2}\left[\mathrm{eft}\left(\frac{\mathrm{t}}{\sqrt{2}}\right) - \mathrm{eft}\left(\frac{\mathrm{t}-\beta q}{\sigma\sqrt{2}}\right)\right] - \frac{1}{2}[\mathrm{eft}\left(\frac{\mathrm{t}+\beta q}{\sigma\sqrt{2}}\right) - \mathrm{eft}\left(\frac{\mathrm{t}}{\sqrt{2}}\right)]$$

$$= \text{eft}\left(\frac{\text{t}}{\sqrt{2}}\right) - \frac{1}{2}\,\text{eft}\left(\frac{\text{t}-\beta q}{\sigma\sqrt{2}}\right) - \frac{1}{2}\,\text{eft}\left(\frac{\text{t}+\beta q}{\sigma\sqrt{2}}\right)$$

Taking the first 2 terms of the Taylor-series expansion of the error function and approximating σ to 1 (σ ≈ 1),

$$\text{eft}\left(\frac{\text{t}}{\sqrt{2}}\right) - \frac{1}{2}\,\text{eft}\left(\frac{\text{t}-\beta q}{\sigma\sqrt{2}}\right) - \frac{1}{2}\,\text{eft}\left(\frac{\text{t}+\beta q}{\sigma\sqrt{2}}\right)$$

$$\cong \frac{2}{\sqrt{\pi}}\left(\frac{\text{t}}{\sqrt{2}} - \frac{\text{t}^3}{6\sqrt{2}}\right) - \frac{1}{2}\left(\frac{2}{\sqrt{\pi}}\right)\left(\frac{\text{t}-\beta q}{\sqrt{2}} - \frac{(\text{t}-\beta q)^3}{6\sqrt{2}}\right) - \frac{1}{2}\left(\frac{2}{\sqrt{\pi}}\right)\left(\frac{\text{t}+\beta q}{\sqrt{2}} - \frac{(\text{t}+\beta q)^3}{6\sqrt{2}}\right)$$

$$= \frac{1}{\sqrt{\pi}}\left(\frac{12\text{t}}{6\sqrt{2}} - \frac{2\text{t}^3}{6\sqrt{2}} - \frac{6\text{t}-6\beta q}{6\sqrt{2}} + \frac{(\text{t}-\beta q)^3}{6\sqrt{2}} - \frac{6\text{t}+6\beta q}{6\sqrt{2}} + \frac{(\text{t}+\beta q)^3}{6\sqrt{2}}\right)$$

$$= \frac{1}{\sqrt{\pi}}\left(\frac{-2\text{t}^3 + (\text{t}-\beta q)^3 + (\text{t}+\beta q)^3}{6\sqrt{2}}\right)$$

$$= \frac{1}{\sqrt{\pi}}\left(\frac{-2\text{t}^3 + \text{t}^3 - 3\text{t}^2\beta q + 3t(\beta q)^2 - (\beta q)^3 + \text{t}^3 + 3\text{t}^2\beta q + 3t(\beta q)^2 + (\beta q)^3}{6\sqrt{2}}\right)$$

$$= \frac{1}{\sqrt{\pi}}\left(\frac{t(\beta q)^2}{\sqrt{2}}\right)$$

As such, if *t >0,*

$$\frac{1}{\sqrt{\pi}}\left(\frac{t(\beta q)^2}{\sqrt{2}}\right) > 0$$

Therefore, if *t > 0*,

$$\int_{\frac{t-\beta q}{\sigma}}^{t} N(z,0,1)\text{dz} > \int_{t}^{\frac{t+\beta q}{\sigma}} N(y,0,1)\text{dy}$$

$$NCP\ ratio > 1$$

Therefore, for diseases with low prevalence (t > 0), there is more power to detect risk variants

compared with the protective variant.

REFERENCES

1. Hirschhorn JN, Gajdos ZKZ (2011) Genome-Wide Association Studies: Results from the First Few Years and Potential Implications for Clinical Medicine. Annual Review of Medicine 62: 11–24. doi:10.1146/annurev.med.091708.162036.

2. Cantor RM, Lange K, Sinsheimer JS (2010) Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application. The American Journal of Human Genetics 86: 6–22. doi:10.1016/j.ajhg.2009.11.017.

3. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. (2009) Finding the missing heritability of complex diseases. Nature 461: 747–753. doi:10.1038/nature08494.

4. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, et al. (2010) Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet 11: 446–450. doi:10.1038/nrg2809.

5. Lim ET, Raychaudhuri S, Sanders SJ, Stevens C, Sabo A, et al. (2013) Rare complete knockouts in humans: population distribution and significant role in autism spectrum disorders. Neuron 77: 235–242. doi:10.1016/j.neuron.2012.12.029.

6. Yu TW, Chahrour MH, Coulter ME, Jiralerspong S, Okamura-Ikeda K, et al. (2013) Using whole-exome sequencing to identify inherited causes of autism. Neuron 77: 259–273. doi:10.1016/j.neuron.2012.11.002.

7. Iyengar SK, Elston RC (2007) The genetic basis of complex traits: rare variants or "common gene, common disease"? Methods Mol Biol 376: 71–84. doi:10.1007/978-1-59745-389-9_6.

8. Purcell S, Cherny SS, Sham PC (2003) Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. Bioinformatics 19: 149–150.

9. Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, et al. (2009) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460: 748–752. doi:10.1038/nature08185.

10. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, et al. (2010) Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42: 565–569. doi:10.1038/ng.608.

11. Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium (2011) Genome-wide association study identifies five new schizophrenia loci. Nat Genet 43: 969–

976. doi:10.1038/ng.940.

12. Psychiatric GWAS Consortium Bipolar Disorder Working Group (2011) Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. Nat Genet 43: 977–983. doi:10.1038/ng.943.

13. Major Depressive Disorder Working Group of the Psychiatric GWAS Consortium, Ripke S, Wray NR, Lewis CM, Hamilton SP, et al. (2013) A mega-analysis of genome-wide association studies for major depressive disorder. Mol Psychiatry 18: 497–511. doi:10.1038/mp.2012.21.

14. Morris AP, Voight BF, Teslovich TM, Ferreira T, Segrè AV, et al. (2012) Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat Genet 44: 981–990. doi:10.1038/ng.2383.

15. Berndt SI, Gustafsson S, Mägi R, Ganna A, Wheeler E, et al. (2013) Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. Nat Genet. doi:10.1038/ng.2606.

16. Sandholm N, Salem RM, McKnight AJ, Brennan EP, Forsblom C, et al. (2012) New Susceptibility Loci Associated with Kidney Disease in Type 1 Diabetes. PLoS Genet 8: e1002921. doi:10.1371/journal.pgen.1002921.

17. Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, et al. (2012) Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature 491: 119–124. doi:10.1038/nature11582.

18. Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, et al. (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467: 52–58. doi:10.1038/nature09298.

19. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851–861. doi:10.1038/nature06258.

20. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56–65. doi:10.1038/nature11632.

21. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559–575. doi:10.1086/519795.

22. Su Z, Marchini J, Donnelly P (2011) HAPGEN2: simulation of multiple disease SNPs. Bioinformatics 27: 2304–2305. doi:10.1093/bioinformatics/btr341.

23. Adams AM, Hudson RR (2004) Maximum-Likelihood Estimation of Demographic Parameters Using the Frequency Spectrum of Unlinked Single-Nucleotide Polymorphisms.

Genetics 168: 1699–1712. doi:10.1534/genetics.104.030171.

24. Agarwala V, Flannick J, Sunyaev S, GoT2D Consortium, Altshuler D (2013) Evaluating empirical bounds on complex disease genetic architecture. Nat Genet. doi:10.1038/ng.2804.

25. Lambert BW, Terwilliger JD, Weiss KM (2008) ForSim: a tool for exploring the genetic architecture of complex traits with controlled truth. Bioinformatics 24: 1821–1822. doi:10.1093/bioinformatics/btn317.

26. Kryukov GV, Shpunt A, Stamatoyannopoulos JA, Sunyaev SR (2009) Power of deep, all-exon resequencing for discovery of human trait genes. Proc Natl Acad Sci USA 106: 3871–3876. doi:10.1073/pnas.0812824106.

27. Eyre-Walker A (2010) Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. Proc Natl Acad Sci U S A 107: 1752–1756. doi:10.1073/pnas.0906182107.

28. Chan Y, Holmen OL, Dauber A, Vatten L, Havulinna AS, et al. (2011) Common variants show predicted polygenic effects on height in the tails of the distribution, except in extremely short individuals. PLoS Genet 7: e1002439. doi:10.1371/journal.pgen.1002439.

29. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38: 904–909. doi:10.1038/ng1847.

30. Willer CJ, Li Y, Abecasis GR (2010) METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics 26: 2190–2191. doi:10.1093/bioinformatics/btq340.

31. Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55: 997–1004.

32. Franke A, McGovern DPB, Barrett JC, Wang K, Radford-Smith GL, et al. (2010) Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat Genet 42: 1118–1125. doi:10.1038/ng.717.

33. Anderson CA, Boucher G, Lees CW, Franke A, D'Amato M, et al. (2011) Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. Nature Genetics 43: 246–252. doi:10.1038/ng.764.

34. Falconer DS (1965) The inheritance of liability to certain diseases, estimated from the incidence among relatives. Annals of Human Genetics 29: 51–76. doi:10.1111/j.1469-1809.1965.tb00500.x.

35. Slatkin M (2008) Exchangeable models of complex inherited diseases. Genetics 179: 2253–2261. doi:10.1534/genetics.107.077719.

36. Baumgart DC, Carding SR (2007) Inflammatory bowel disease: cause and immunobiology.

Lancet 369: 1627–1640. doi:10.1016/S0140-6736(07)60750-8.

37. Yang J, Lee SH, Goddard ME, Visscher PM (2013) Genome-wide complex trait analysis (GCTA): methods, data analyses, and interpretations. Methods Mol Biol 1019: 215–236. doi:10.1007/978-1-62703-447-0_9.

38. Lee SH, DeCandia TR, Ripke S, Yang J, Sullivan PF, et al. (2012) Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. Nat Genet 44: 247–250. doi:10.1038/ng.1108.

39. Purcell SM, Moran JL, Fromer M, Ruderfer D, Solovieff N, et al. (2014) A polygenic burden of rare disruptive mutations in schizophrenia. Nature. doi:10.1038/nature12975.

40. Park J-H, Gail MH, Weinberg CR, Carroll RJ, Chung CC, et al. (2011) Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. Proc Natl Acad Sci U S A 108: 18026–18031. doi:10.1073/pnas.1114759108.

41. Anderson CA, Massey DCO, Barrett JC, Prescott NJ, Tremelling M, et al. (2009) Investigation of Crohn's disease risk loci in ulcerative colitis further defines their molecular relationship. Gastroenterology 136: 523–529.e3. doi:10.1053/j.gastro.2008.10.032.

42. Price AL, Zaitlen NA, Reich D, Patterson N (2010) New approaches to population stratification in genome-wide association studies. Nat Rev Genet 11: 459–463. doi:10.1038/nrg2813.

43. Heinzen EL, Depondt C, Cavalleri GL, Ruzzo EK, Walley NM, et al. (2012) Exome Sequencing Followed by Large-Scale Genotyping Fails to Identify Single Rare Variants of Large Effect in Idiopathic Generalized Epilepsy. Am J Hum Genet 91: 293–302. doi:10.1016/j.ajhg.2012.06.016.

Chapter 4

Genome wide association in European and African Americans discover novel loci associated with sitting height ratio

Yingleong Chan[1,2,3], Rany M Salem[1,2,3], Joel N Hirschhorn[1,2,3]

[1] Department of Genetics, Harvard Medical School, Boston, MA, USA

[2] Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, USA.

[3] Department of Endocrinology, Boston Children's Hospital, Boston, MA, USA.

ABSTRACT

Body proportion is a phenotype that is determined by the ratio of different components of the human anatomy. While there are many genetic studies that have been performed for height, little is known about the genetics underlying our body proportion and the genes regulating our proportion might play a more important role for growth and development. Here we report our findings of our analysis on sitting height ratio (SHR), the ratio of sitting height to overall height. We show that genetics contribute in a major way to explain the difference in SHR between African Americans and European Americans. After adjusting for height, age, sex, body mass index and the relevant principal components, the genome-wide association study (GWAS) in African and Europeans Americans uncover 3 loci associated with SHR. One of the loci (rs5959358) resides on the X-chromosome and was reported to be also associated with height. Comparing the known loci associated with height with the results of SHR reveal that most of the loci are associated with alterations of SHR too. While these confirm that SHR is largely genetically determined, nonetheless more samples are required to reveal the full genetic architecture in SHR determination.

INTRODUCTION

Human height is a commonly used trait to illustrate a highly heritable that is polygenic. Our height however, is in reality a summation of many different components, e.g. head length, trunk length, leg length, etc. One of the first reports on how these individual lengths should correlate with each other was given by Leonardo da Vinci in his illustration of *Vitruvian Man* circa 1490. In it, he recorded the expected proportions of these measurements in relation with each other for the human body. Further research have postulated that some of these

measurements may be predictors of diseases [1]. For example, there is evidence that leg length can be a predictor of metabolic disorders underlying type 2 diabetes [2].

One such measurement is sitting height. Sitting height is defined as total stature that is comprised by head and trunk and is usually measured by having the individual sit on a table and measuring the length from the table surface to the top of the person's head. Since sitting height is a component of an individual's total height, the sitting height ratio (SHR), defined as the sitting height divided by total height is an indicator of an individual's body proportion. The SHR of an individual changes as we grow. Unlike height which increases as we age, SHR rapidly decreases as we progress from being a baby to being a teenager an increases slightly as we become adults [3,4]. In the extreme case, individuals affected with skeletal dysplasias not only have short stature, but also have disproportionate SHR [5]. Depending on the type of skeletal dysplasia, the SHR can be severely increased. For example, individuals with Achondroplasia have average SHR of 0.66 (normal range: 0.52-0.53) [6]. On the other hand, individuals with spondyloepiphyseal and spondylometaepiphyseal dysplasias may have normal SHR values [7].

The SHR is also slightly different between people from different ancestries. Individuals of Asian ancestry have higher SHR than individuals of European ancestry and individuals of European ancestry have higher SHR than individuals of African ancestry [8]. This difference is assumed to be due to genetic factors, although it remains unclear whether the difference is due to many variants with small effect sizes or a few variants with large effect sizes.

In this chapter, we described our approach to determine if there is a strong genetic influence on the SHR difference between individuals of different ancestry. We found that SHR is highly correlated with the degree to which African Americans have admixed of European ancestry. The more European ancestry an African American has, the higher his or her SHR,

consistant with the reported observations. We performed genome wide association study of SHR with both European Americans as well as African Americans and reported 3 loci associated with SHR. We then examine several variants that were known to be associated with height and observed that many of these variants were also marginally associated with SHR. These results suggest that variants associated with height that are also associated with SHR might be in genes that regulate development of the growth plate.

RESULTS

European Americans have higher sitting height ratios (SHR) than African Americans

We used the ARIC [9] and CARDIA [10] cohorts as they include both European and African Americans with both sitting height and height measurements. After removing individuals that failed our quality control (see Materials and Methods), we have 7,257 European American individuals and 2,354 African American individuals from ARIC. For CARDIA, we have 1,047 European American individuals and 715 African American individuals. Comparing the sitting height ratio (SHR) between European and African American individuals, we find that European Americans have higher SHR values than their African Americans (Figure 4.1). In both ARIC and CARDIA, the mean SHR for European Americans is 0.53 while the mean SHR for African Americans is 0.51. After correcting SHR for covariates like height, age, sex, BMI and expressed SHR in terms of a Zscore (see Materials and Methods), we observed that there is more than a 1 standard deviation difference (ARIC = 1.16, CARDIA = 1.06) between European Americans and African Americans. This result is consistent with earlier findings that European Americans have higher SHR than African Americans [8].

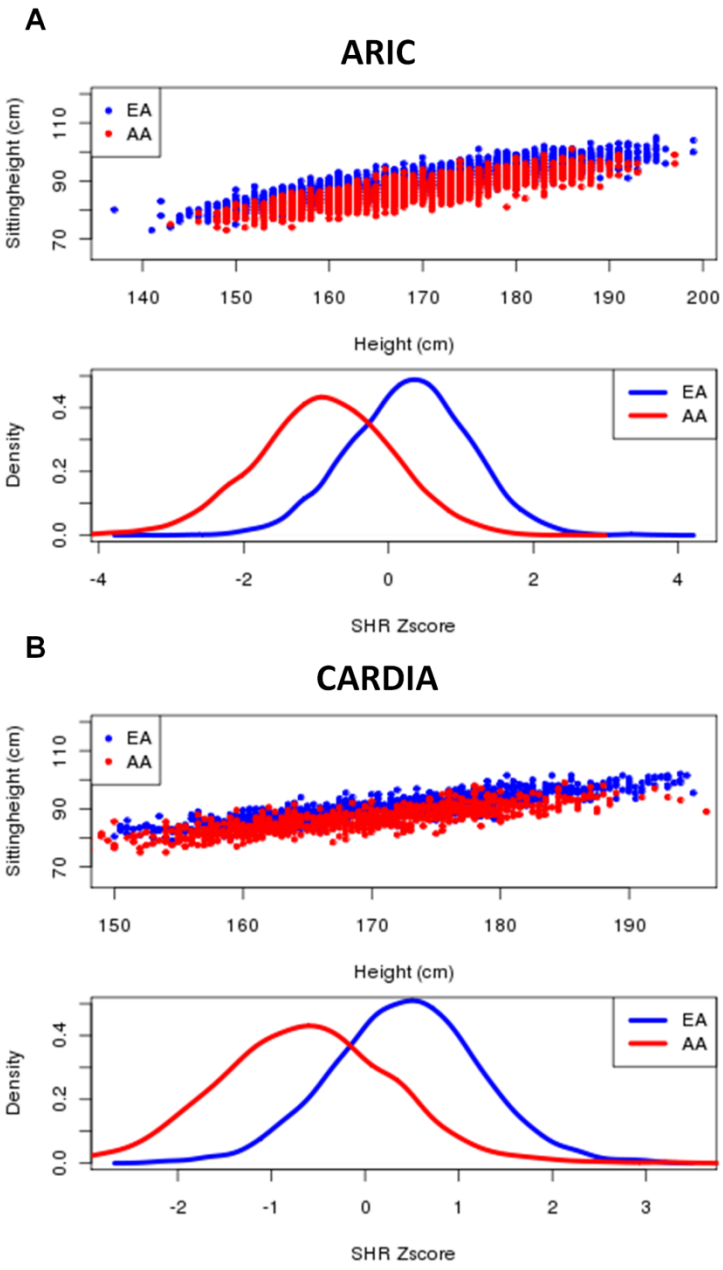


Figure 4.1: Sitting height, height and sitting height ratio (SHR) distribution. We examined the sitting height, heights and sitting height ratios (SHRs) for individuals in both the ARIC and CARDIA cohorts. European Americans (EA) are colored in blue while African Americans (AA) are colored in red. (**A**) The top panel plots the sitting heights versus total height for the individuals in the ARIC cohort (N=9,611). The bottom panel represents the histogram of SHR of European American and African Americans where there is about a 1.18 standard deviation difference between the 2 populations. (**B**) The CARDIA cohort (N=1,762). The bottom panel shows the histogram of SHR of EA and AA where there is about a 1.06 standard deviation difference between the 2 populations.

Degree of European admixture is predictive of sitting height ratio (SHR) in African Americans

To determine if the SHR difference between European and African Americans has a genetic component, we reasoned that we could test this by exploring the genetic landscape of African Americans. As it is common for African Americans to high levels (>10%) of European ancestry [11], the level of European ancestry in any given African American should be correlated with SHR, if there is a genetic component to this difference. Given that European Americans have higher SHR than African Americans, we expect this correlation to be positive. To test this, we used principal component analysis to determine the degree of European admixture for the African Americans in both the ARIC and CARDIA (see Materials and Methods). We observed that there is a gradient of percentage European admixture in the African Americans (Figure 4.2A-B) with some African Americans having as much as 60% European ancestry. There are significance positive correlations between the percentage European admixture and normalized sitting height ratios (SHR) (Figure 4.2C-D). This result shows that the SHR difference between European and African Americans has a significant genetic component.

Analysis of African American individuals identifies variant associated with sitting height ratio (SHR)

Given evidence for a genetic component, we proceeded to test for genetic markers that are associated with sitting height ratio (SHR). We performed genome wide association on the using the genotypes of the African American individuals from both the ARIC and CARDIA cohorts and performed the meta-analysis by combining the results from both cohorts (see Materials and Methods). We observed a genome-wide significant signal at the chromosome

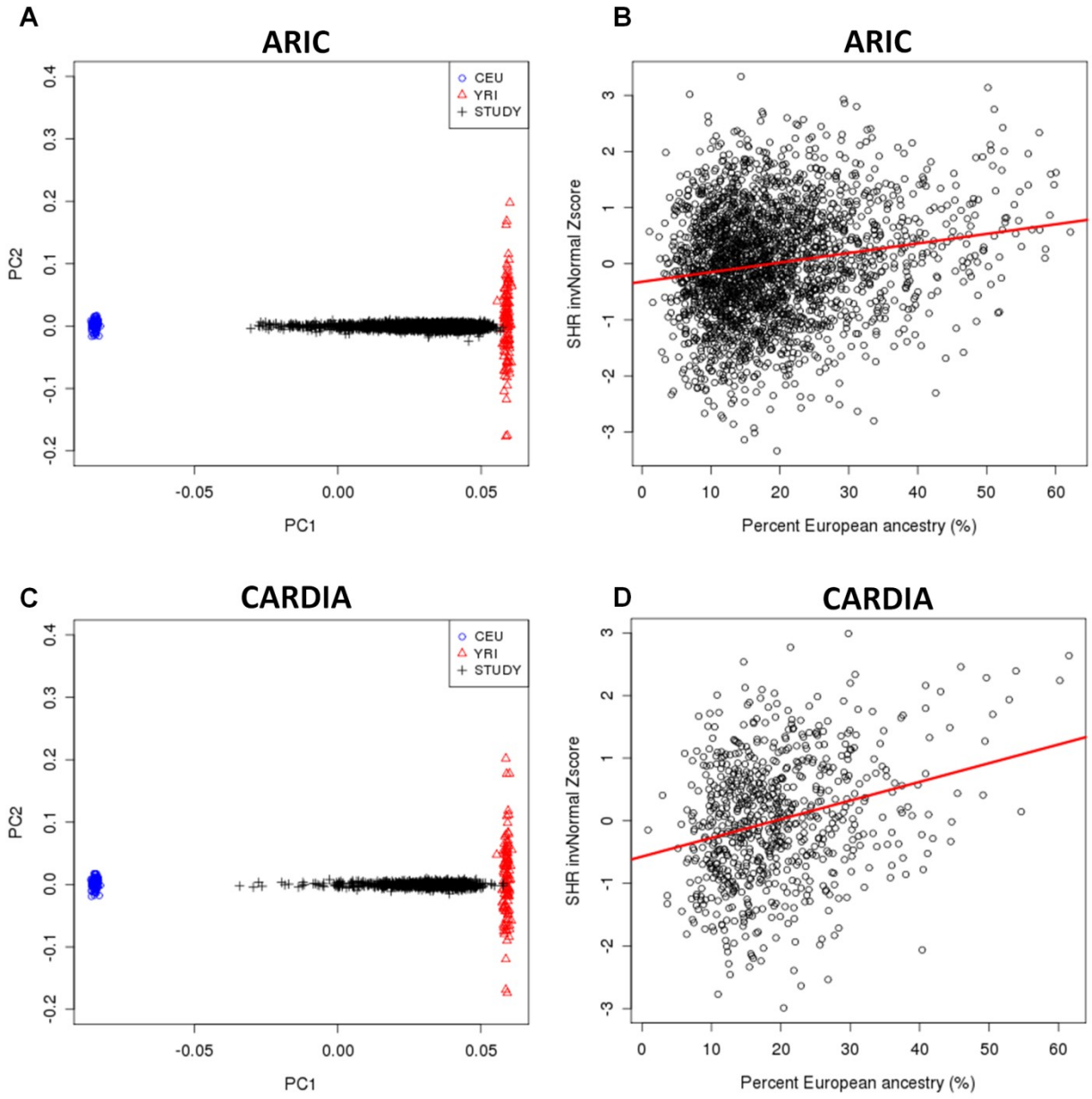


Figure 4.2: Association of global European ancestry with sitting height ratio (SHR). The plots show the degree of European admixture for each African American individual and how it correlates with SHR. A and C show the degree of European admixture in the 2 cohorts by principal component analysis. Individuals closer to CEU (blue) have more European ancestry than individuals close to YRI (red). B and D show the association of European ancestry with SHR using linear regression. (**A**) Global European ancestry for ARIC. (**B**) Correlating global European ancestry with SHR for ARIC. (**C**) Global European ancestry for CARDIA. (**D**) Correlating global European ancestry with SHR for CARDIA.

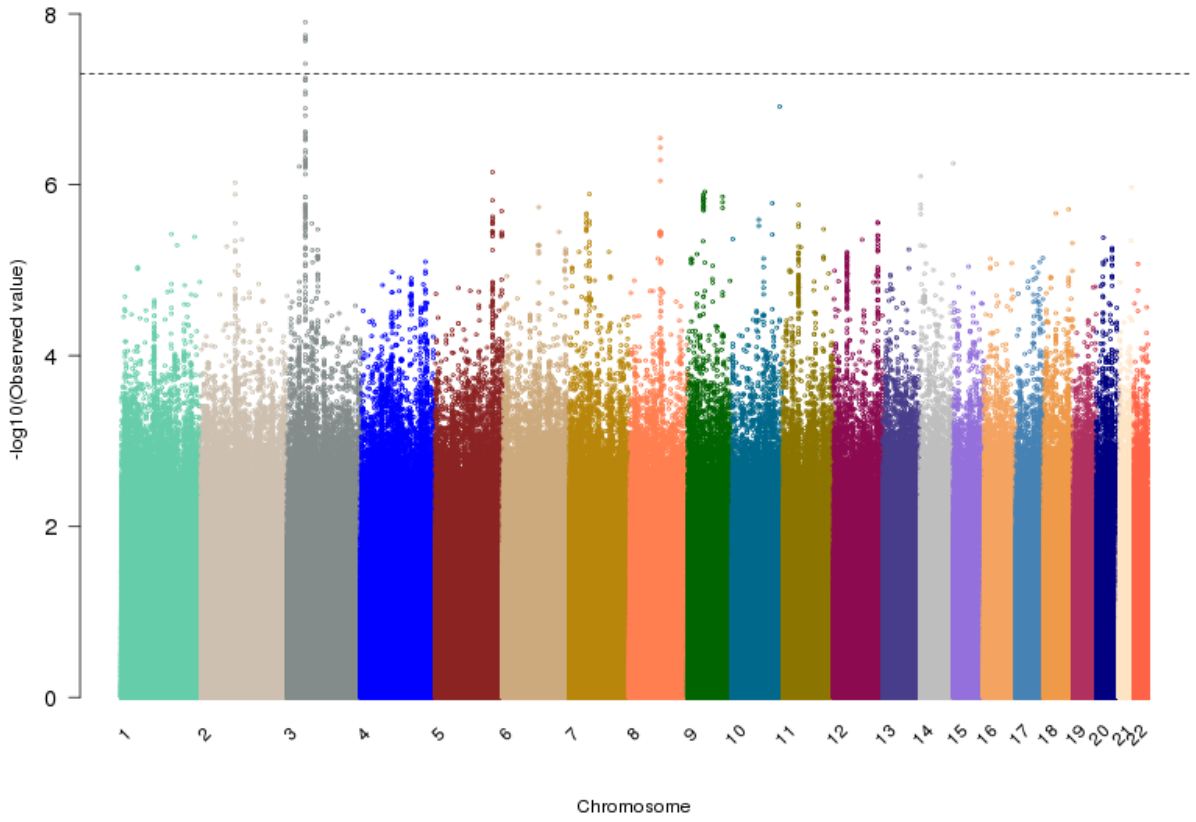


Figure 4.3: Genome wide association study (GWAS) of African American individuals. The Manhattan plot of the GWAS performed for the African American individuals from the ARIC and CARDIA cohorts. Only 1 locus (rs201786365) reached genome wide significance.

3p21.33 locus of which the lead variant (rs201786365) has an association statistic of $P=1.252 \times 10^{-8}$ with the minor allele (MAF = 0.14) associated with increased SHR ($\beta = 0.21$) (Figure 4.3). This variant is present only in African Americans and is fixed as the major allele in European Americans. As such, this variant does not explain for the SHR difference between African and European Americans. The closest gene in the locus to the lead SNP is *ABHD5* of which mutations in the gene has been associated with Chanarin-Dorfman syndrome [12].

Analysis of European American individuals identifies 2 loci associated with sitting height ratio (SHR)

We continued to explore for genetic associations for SHR by performing the test on our European American individuals. We performed the test on the European American individuals in the ARIC, CARDIA, CHS, FHS cohorts (see Materials and Methods). We observed a genome-wide significant signal at the chromosome 18p11.23 locus of which the lead variant (rs140449984) has an association statistic of $P=3.70 \times 10^{-9}$ with the minor allele (MAF = 0.07) associated with decreased SHR ($\beta = -0.149$) (Figure 4.4). This variant lies within an intron of the PTPRM gene, which the protein encoded is a member of protein tyrosine phosphatase (PTP) family.

Additionally, we observed a significant signal on the X-chromosome (rs5959358) that the minor allele (MAF = 0.37) is associated with decrease SHR ($\beta = -0.097$, $P = 9.71 \times 10^{-8}$) only in women (Figure 4.5). Interestingly, the locus, which is in the vicinity of ITM2A, has been shown to be associated with height and also reported to escape dosage compensation [13]. That reported variant (rs1751138) is also associated with decrease SHR ($\beta = -0.0945$, $P = 3.18 \times 10^{-7}$) and is in strong linkage disequilibrium (LD) with rs5959358.

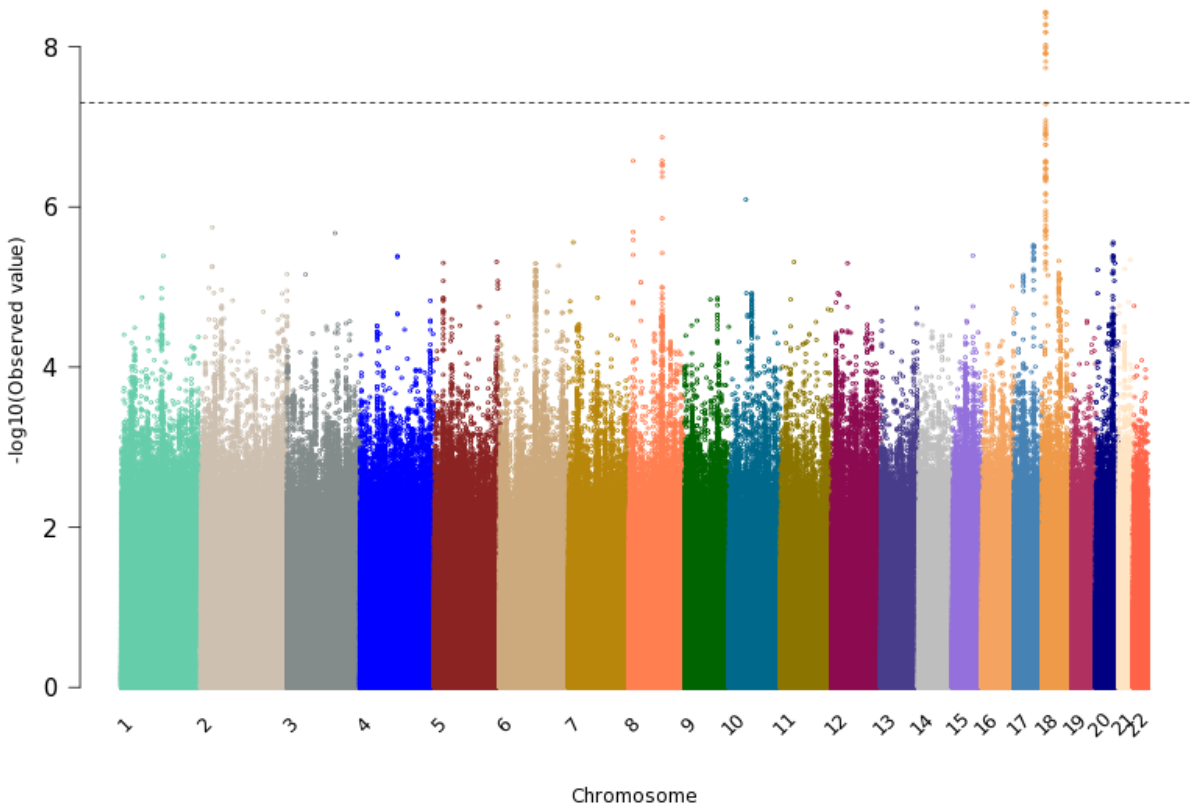


Figure 4.4: Genome wide association study (GWAS) of European American individuals. The Manhattan plot of the GWAS performed for the European American individuals from the ARIC, CARDIA, CHS and FHS cohorts. Only 1 locus reached genome wide significance. The lead variant (rs140449984) is in the *PTPRM* gene.

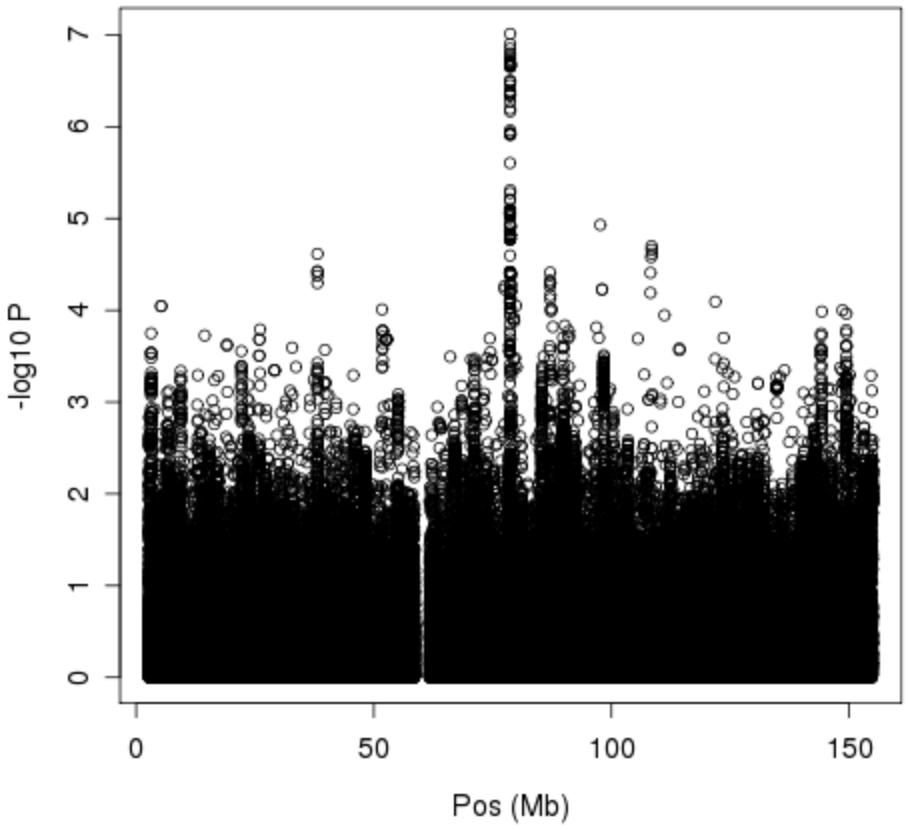


Figure 4.5: Genome wide association study (GWAS) of the X-chromosome in European American women. The plot of the X-chromosome association performed for the European American women from the ARIC, CARDIA, CHS and FHS cohorts. The strongest association signal (rs5959358) has the closest gene (ITM2A) that was previously reported to harbor variants that escape dosage compensation.

Variants associated with height are also associated with sitting height ratio (SHR)

As sitting height is one of the components of height, we reasoned that variants that alter our height may be enriched for variants that also alter our SHR. To test this, we obtained a set of 421 LD-independent variants that have been shown to be robustly associated with height (Wood et. al., unpublished) and determine if they are also associated with SHR. Although none of these 421 variants reached genome-wide significance, we observed that as a whole, the 421 height associated variants are also significantly associated with SHR (Figure 4.6). We observed 49 of the 421 variants to have SHR P-values less than 0.05, which is significant (*Expected=21.05/421; P=2x10*$^{-8}$). The strongest associated variant (rs2079795) has an association with SHR with a P-value of approximately 3 x 10^{-6}. Also, the variant associated with height in GDF5, which was previously suggested to also have some association with sitting height [14] had some marginal association with SHR (*P=0.01*) (Table 4.1). These results are indicative that SHR is polygenic and a substantial number of height associated alleles do alter the SHR as well.

DISCUSSION

We have shown that body proportion as determined by our sitting height ratio (SHR) is mainly genetically driven. SHR also appears to be more constraint than height as while a standard deviation (SD) of height is 6.08 cm (ARIC), an SD of sitting height adjusted for height is just 1.95 centimeters (ARIC). Also, in general, men and women can differ in heights as much as 12cm [15], the sitting height adjusted for height difference between men and women is just approximately 0.47cm (ARIC). This is suggestive that the genes underlying the variability of SHR might just be more relevant than height to development as there is more selective pressure to keep our SHR within an acceptable range. However, we and others have shown that there is a

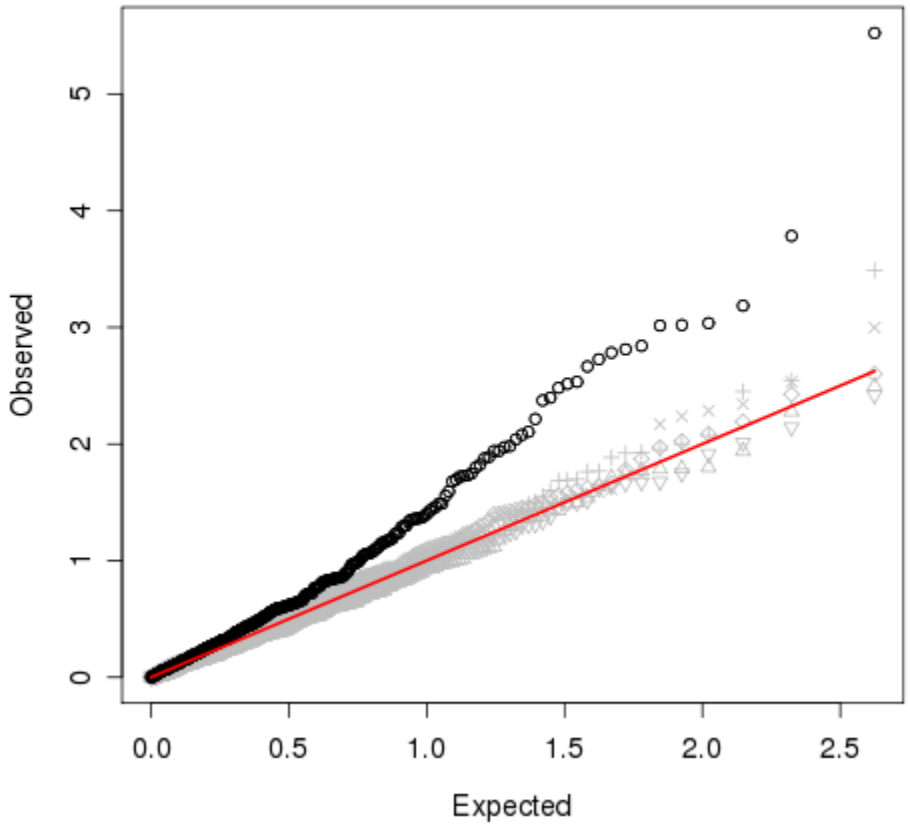


Figure 4.6: QQ-plot of the 421 LD-independent SNPs known to be associated with height. The plot shows the 421 height SNPs are as a group, also associated with sitting height ratio (SHR) even if none of them reached genome wide significance. The x-axis is the expected $-\log_{10}$ of the P-values while the y-axis is the observed $-\log_{10}$ of the P-values obtained from the association with SHR from the European American individuals. The gray points represent 5 different random samplings of 421 different variants from the GWAS of SHR from the European American individuals.

Table 4.1: The effect sizes, P-values of the 421 height associated SNPs with sitting height ratio (SHR).

						Height		Sitting height ratio		
Rsid	**Chr**	**Position**	**Ref**	**Alt**	**Reffreq**	**Effect Size**	**P-value**	**Effect Size**	**P-value**	**Closest Gene**
rs2079795	17	56851431	t	c	0.33	0.045	8.60E-48	-0.067	2.99E-06	C17orf82
rs310421	6	81848782	t	g	0.54	0.028	1.40E-21	0.0508	0.0001647	FAM46A
rs4369779	18	18989406	c	t	0.79	0.056	1.40E-54	-0.0557	0.0006526	CABLES1
rs1614303	10	123386796	t	g	0.83	0.023	1.50E-09	-0.0579	0.0009193	FGFR2
rs42039	7	92082358	t	c	0.27	0.051	4.10E-39	-0.0518	0.0009542	CDK6
rs217181	16	70671503	t	c	0.2	0.024	2.20E-10	-0.055	0.0009664	HPR
rs3790086	16	68445208	c	g	0.56	0.023	9.80E-16	0.0417	0.001443	WWP2
rs4803468	19	46614192	a	g	0.42	0.029	1.20E-20	0.0421	0.001541	BCKDHA
rs3807931	7	20348199	a	g	0.45	0.027	4.00E-21	-0.0425	0.001648	ITGB8
rs3825199	12	92501085	g	a	0.23	0.054	1.90E-53	0.0504	0.001875	SOCS2
rs3791679	2	55950396	a	g	0.77	0.084	1.30E-96	-0.0467	0.002166	EFEMP1
rs2224538	20	37985492	t	c	0.65	0.018	4.90E-09	0.0404	0.002936	MAFB
rs3760318	17	26271841	g	a	0.63	0.054	2.30E-59	-0.0399	0.00306	CENTA2
rs8006657	14	54314899	g	a	0.59	0.024	3.70E-15	0.0396	0.003301	SAMD4A
rs1966913	16	65941727	a	t	0.96	0.042	1.00E-08	-0.092	0.004018	LRRC36
rs7733195	5	172927230	g	a	0.64	0.028	3.40E-20	0.0404	0.004234	FAM44B
rs6485978	11	12634991	c	t	0.46	0.022	2.10E-14	-0.036	0.006102	TEAD1
rs12323101	13	32041406	a	g	0.37	0.021	2.40E-12	0.0374	0.007877	PDS5B
rs11642612	16	29937696	c	a	0.4	0.017	3.20E-08	-0.0351	0.008347	FLJ25404
rs17081935	4	57518233	t	c	0.2	0.03	5.00E-16	-0.0444	0.009117	C4orf14
rs9428104	1	118657110	g	a	0.75	0.044	1.10E-37	-0.0386	0.01043	SPAG17
rs16968242	15	74527274	g	c	0.07	0.034	6.20E-09	-0.0634	0.01065	SCAPER
rs143384	20	33489170	g	a	0.42	0.063	1.30E-71	0.0343	0.01146	GDF5
rs314263	6	105499438	c	t	0.32	0.043	3.10E-43	-0.0365	0.01148	LIN28B
rs2888893	12	105862761	c	t	0.51	0.017	7.30E-09	-0.0334	0.01289	C12orf23
rs11659752	18	75323850	t	g	0.7	0.025	2.10E-13	-0.0361	0.0133	NFATC1
rs212524	1	21455898	c	t	0.6	0.021	4.70E-12	0.0326	0.01502	ECE1
rs10877030	12	56542981	t	g	0.68	0.023	2.80E-13	0.0347	0.01597	CTDSP2
rs2871865	15	97012419	c	g	0.88	0.059	8.10E-32	0.0499	0.01789	IGF1R
rs3116168	2	232698075	c	t	0.73	0.022	2.40E-09	-0.0337	0.01866	DIS3L2
rs10770705	12	20748734	a	c	0.34	0.03	4.80E-22	-0.0338	0.01879	SLCO1C1
rs1797625	3	114309105	t	a	0.36	0.018	1.00E-09	-0.0318	0.01912	C3orf17
rs1884897	20	6560832	a	g	0.36	0.038	4.70E-33	-0.0313	0.02021	BMP2
rs1658351	3	57988613	c	t	0.35	0.023	3.00E-13	0.0318	0.021	FLNB
rs2597513	3	13530836	c	t	0.11	0.042	1.10E-18	-0.0476	0.02557	HDAC11
rs953199	9	99522797	c	a	0.76	0.02	6.10E-09	0.0331	0.0278	XPA

Table 4.1 (Continued)

						Height		Sitting height ratio		
Rsid	**Chr**	**Position**	**Ref**	**Alt**	**Reffreq**	**Effect Size**	**P-value**	**Effect Size**	**P-value**	**Closest Gene**
rs7517682	1	103292177	g	a	0.44	0.022	9.20E-14	0.0282	0.03245	COL11A1
rs7544462	1	37735343	a	c	0.91	0.032	1.10E-09	-0.0517	0.03251	C1orf149
rs1326023	20	54275785	a	g	0.3	0.024	9.80E-14	0.0299	0.03458	MC3R
rs17349981	15	80018975	a	t	0.85	0.028	2.40E-11	0.0399	0.03573	MEX3B
rs9825951	3	100752611	t	a	0.35	0.02	4.20E-10	-0.0287	0.03733	COL8A1
rs7701414	5	131613857	g	a	0.44	0.041	4.90E-42	-0.0276	0.0396	PDLIM4
rs12519505	5	77541632	c	t	0.78	0.023	3.20E-10	0.0327	0.04161	AP3B1
rs3814333	1	182273742	t	c	0.32	0.049	1.90E-53	-0.0288	0.04288	GLT25D2
rs2763273	6	168577472	c	t	0.76	0.022	1.80E-10	-0.0325	0.04334	SMOC2
rs1171615	10	61139096	c	t	0.22	0.022	4.50E-09	-0.0371	0.04364	SLC16A9
rs11640018	16	73885809	c	t	0.37	0.019	2.10E-09	-0.0275	0.04409	CFDP1
rs12779328	10	12983979	c	t	0.72	0.028	1.50E-17	-0.0298	0.0446	CCDC3
rs4953951	2	135903815	c	t	0.9	0.035	4.30E-11	-0.0433	0.04661	ZRANB3
rs12120956	1	113004094	g	a	0.77	0.025	9.90E-13	-0.0313	0.05032	CAPZA1
rs7033487	9	118169078	t	c	0.79	0.041	3.50E-29	0.0316	0.05056	PAPPA
rs10948222	6	45352393	c	t	0.58	0.032	8.70E-22	0.0266	0.05165	SUPT3H
rs1055144	7	25837634	t	c	0.19	0.022	1.80E-09	-0.0327	0.05777	NFE2L3
rs2272566	11	234552	a	g	0.48	0.016	2.40E-08	-0.0246	0.05786	PSMD13
rs606452	11	74953826	a	c	0.14	0.043	6.40E-23	-0.0352	0.05948	SERPINH1
rs936339	3	144018195	t	c	0.19	0.022	2.00E-08	-0.031	0.06448	PCOLCE2
rs4802134	19	43038525	a	g	0.21	0.027	2.90E-11	0.0279	0.06557	SIPA1L3
rs1546391	3	116180147	g	c	0.07	0.042	2.50E-12	-0.0446	0.06734	ZBTB20
rs7534365	1	148142748	c	t	0.19	0.045	3.50E-20	-0.0336	0.06785	SV2A
rs7985356	13	114045564	t	a	0.77	0.023	2.50E-11	0.0285	0.07016	CDC16
rs26868	16	2189377	a	t	0.47	0.025	2.70E-13	0.0264	0.07022	CASKIN1
rs12186664	5	95655981	t	a	0.32	0.021	6.00E-12	-0.0255	0.07265	PCSK1
rs9650315	8	57318152	g	t	0.87	0.057	2.50E-34	0.0352	0.07642	CHCHD7
rs798497	7	2762483	a	g	0.7	0.057	2.70E-71	0.0262	0.07826	GNA12
rs1405212	6	117597357	c	t	0.59	0.023	4.60E-14	0.0242	0.08067	VGLL2
rs2166898	2	121329129	g	a	0.84	0.027	8.70E-11	-0.0305	0.08378	GLI2
rs6446315	4	5086488	g	a	0.17	0.025	3.60E-09	-0.0309	0.08545	CYTL1
rs2034172	3	55386803	g	a	0.68	0.018	2.50E-08	0.0247	0.08616	WNT5A
rs4686904	3	188921216	c	t	0.35	0.022	1.00E-12	-0.0232	0.08732	BCL6
rs8103992	19	19526643	a	c	0.2	0.024	3.60E-10	0.0274	0.08779	PBX4
rs2338115	17	34183104	t	c	0.54	0.024	1.10E-16	0.0222	0.08954	PIP4K2B
rs1461503	11	122350285	c	a	0.57	0.018	3.70E-10	-0.0222	0.09167	BSX
rs2093210	14	60027032	c	t	0.42	0.039	7.50E-36	0.022	0.09649	C14orf39

Table 4.1 (Continued)

						Height		Sitting height ratio		
Rsid	**Chr**	**Position**	**Ref**	**Alt**	**Reffreq**	**Effect Size**	**P-value**	**Effect Size**	**P-value**	**Closest Gene**
rs9967417	18	45213498	g	c	0.43	0.037	1.20E-32	0.0218	0.101	DYM
rs2275325	1	202067358	c	g	0.28	0.019	5.00E-09	-0.0241	0.1031	ZC3H11A
rs3885668	2	10095930	c	t	0.43	0.022	6.90E-14	0.022	0.1049	KLF11
rs6955948	7	150139653	t	c	0.28	0.031	8.80E-20	0.0242	0.1052	TMEM176A
rs4868126	5	171216074	g	t	0.6	0.025	2.70E-11	-0.025	0.108	FBXW11
rs2633761	3	4703104	a	g	0.5	0.017	3.70E-09	0.021	0.1086	ITPR1
rs820848	5	74000416	g	a	0.29	0.021	3.60E-09	0.0239	0.1169	HEXB
rs1681630	11	47925728	t	c	0.34	0.031	1.10E-23	-0.0214	0.1218	PTPRJ
rs422421	5	176449932	c	t	0.78	0.034	1.70E-20	-0.0243	0.1277	FGFR4
rs17574650	5	42472673	c	a	0.11	0.036	1.50E-11	-0.0353	0.1278	GHR
rs3802758	11	45892611	a	g	0.94	0.041	5.10E-10	-0.0349	0.1346	PEX16
rs1562975	4	109628057	a	g	0.3	0.025	4.00E-15	-0.0217	0.1376	RPL34
rs7659107	4	114961698	g	a	0.23	0.024	6.60E-12	-0.024	0.1385	CAMK2D
rs10997979	10	69607198	g	a	0.5	0.018	3.50E-10	0.0196	0.1405	MYPN
rs862034	14	74060499	g	a	0.64	0.03	2.60E-23	-0.0199	0.1408	LTBP2
rs6441170	3	159289654	c	t	0.38	0.022	8.60E-14	-0.0198	0.1409	SHOX2
rs6694089	1	170350504	a	g	0.28	0.027	2.00E-13	0.0211	0.141	DNM3
rs6061231	20	60390312	c	a	0.72	0.02	1.70E-10	-0.0212	0.1436	RPS21
rs11144688	9	77732106	g	a	0.89	0.064	5.90E-24	-0.03	0.1437	PCSK5
rs10767838	11	30304503	a	g	0.72	0.025	1.80E-14	0.0212	0.1457	C11orf46
rs17792664	14	20960523	g	c	0.14	0.033	2.70E-14	-0.0275	0.1459	CHD8
rs12209223	6	76221309	a	c	0.12	0.046	1.90E-20	-0.0321	0.1466	FILIP1
rs2326458	16	83545180	c	a	0.25	0.022	4.50E-10	-0.0218	0.1478	ZDHHC7
rs6584575	10	105567399	a	g	0.1	0.032	1.20E-09	-0.0323	0.1492	SH3PXD2A
rs291979	10	121119787	a	g	0.23	0.03	5.50E-18	-0.023	0.1492	GRK5
rs17410035	5	31576899	t	g	0.33	0.017	1.70E-08	-0.0205	0.1524	C5orf22
rs1935157	1	219383881	g	c	0.3	0.024	3.10E-14	-0.0206	0.1525	HLX
rs9816693	3	38022958	c	g	0.17	0.031	3.60E-15	0.0248	0.1548	VILL
rs8052560	16	87304743	a	c	0.79	0.037	8.40E-17	-0.0235	0.1634	C16orf84
rs17511102	2	37814117	t	a	0.09	0.049	2.80E-17	-0.0334	0.1665	CDC42EP3
rs17264185	15	64784141	g	a	0.27	0.021	1.30E-10	-0.0203	0.1679	SMAD6
rs7971536	12	100897919	t	a	0.54	0.028	5.00E-18	0.0185	0.1685	CCDC53
rs181338	9	88297981	t	c	0.51	0.029	5.70E-24	-0.0178	0.1697	ZCCHC6
rs4735677	8	78310746	t	a	0.28	0.036	1.20E-29	-0.0203	0.1706	PXMP3
rs552707	7	28171828	t	c	0.31	0.047	7.40E-49	-0.02	0.1752	JAZF1
rs2956605	8	76045609	a	c	0.38	0.027	1.70E-17	-0.0186	0.1891	CRISPLD1
rs13177718	5	108141243	c	t	0.92	0.054	5.40E-19	-0.0337	0.1902	FER

Table 4.1 (Continued)

						Height		Sitting height ratio		
Rsid	**Chr**	**Position**	**Ref**	**Alt**	**Reffreq**	**Effect Size**	**P-value**	**Effect Size**	**P-value**	**Closest Gene**
rs2834442	21	34612656	a	t	0.64	0.024	5.70E-15	-0.0179	0.1907	KCNE2
rs3809790	17	24979666	c	t	0.53	0.022	1.50E-13	0.0173	0.1914	SSH2
rs6894139	5	88363538	t	g	0.56	0.031	4.50E-25	-0.0175	0.1918	MEF2C
rs11612228	12	447245	t	c	0.38	0.02	3.70E-10	-0.0192	0.1925	B4GALNT3
rs2164747	12	102868966	g	a	0.1	0.028	9.10E-09	-0.0278	0.1969	HSP90B1
rs301901	5	37082383	a	g	0.57	0.026	4.50E-19	0.0172	0.1997	NIPBL
rs12228415	12	14411968	g	a	0.45	0.017	2.70E-08	-0.0173	0.2032	ATF7IP
rs7727731	5	64710202	t	c	0.11	0.033	2.10E-11	-0.0259	0.2171	ADAMTS6
rs9858528	3	184838099	a	g	0.74	0.021	8.50E-11	-0.0179	0.2196	KLHL24
rs9766	17	38106367	a	g	0.54	0.021	2.40E-13	-0.0161	0.2203	EZH1
rs12214804	6	34296844	c	t	0.08	0.087	1.60E-52	0.0302	0.2221	HMGA1
rs8756	12	64646019	c	a	0.49	0.054	1.30E-71	-0.0163	0.2235	HMGA2
rs17450430	20	47205671	t	a	0.24	0.034	6.20E-24	-0.0188	0.2251	STAU1
rs8180991	8	126569532	c	g	0.77	0.029	2.80E-16	-0.0193	0.2274	TRIB1
rs891088	19	7135762	g	a	0.26	0.027	1.30E-15	-0.0177	0.229	INSR
rs273945	7	137262106	c	a	0.58	0.018	2.90E-09	0.0164	0.2322	CREB3L2
rs17806888	3	67499012	t	c	0.88	0.033	4.30E-12	0.0236	0.2343	SUCLG2
rs2211866	21	38609977	a	g	0.41	0.022	1.90E-13	-0.0157	0.2349	KCNJ15
rs9977276	21	46260755	g	t	0.78	0.023	1.30E-10	0.0185	0.2354	COL6A1
rs12855	1	51212681	t	c	0.09	0.036	1.00E-12	0.0266	0.2363	CDKN2C
rs584828	17	35852756	c	t	0.6	0.028	3.30E-20	0.0159	0.2396	IGFBP4
rs7043114	9	94427804	c	t	0.44	0.028	1.30E-22	-0.0154	0.2401	IPPK
rs1812175	4	145794294	g	a	0.84	0.052	8.40E-30	-0.0211	0.2404	HHIP
rs11867479	17	65601802	t	c	0.35	0.025	2.00E-15	-0.0164	0.243	KCNJ16
rs11616380	13	79603316	t	g	0.28	0.02	1.20E-09	-0.0176	0.2431	SPRY2
rs955748	4	184452669	g	a	0.76	0.028	4.80E-16	0.0184	0.2436	WWC2
rs2117563	17	70880580	g	a	0.83	0.025	2.10E-10	-0.0196	0.2467	GRB2
rs4548838	15	98578713	t	c	0.46	0.034	9.10E-30	-0.0152	0.2487	ADAMTS17
rs12190423	6	72259432	g	c	0.62	0.016	4.30E-08	0.016	0.2497	OGFRL1
rs11152213	18	56003928	c	a	0.25	0.025	9.20E-13	-0.0176	0.2518	MC4R
rs9880211	3	137590239	g	a	0.75	0.032	1.30E-20	-0.0173	0.2531	STAG1
rs4974480	3	135661252	t	a	0.68	0.037	5.70E-23	-0.0158	0.2552	ANAPC13
rs12470505	2	219616613	t	g	0.9	0.046	4.00E-20	-0.0246	0.256	CCDC108
rs4875421	8	4814740	t	a	0.46	0.019	1.10E-10	0.0153	0.2567	CSMD1
rs4725061	7	8053164	g	a	0.44	0.02	1.50E-10	0.016	0.2569	GLCCI1
rs7181724	15	92352611	g	a	0.45	0.02	2.40E-10	0.0155	0.2573	MCTP2
rs7259684	19	12047611	g	a	0.07	0.039	1.70E-09	0.0271	0.2587	LOC729747

Table 4.1 (Continued)

						Height		Sitting height ratio		
Rsid	**Chr**	**Position**	**Ref**	**Alt**	**Reffreq**	**Effect Size**	**P-value**	**Effect Size**	**P-value**	**Closest Gene**
rs7567851	2	178392966	c	g	0.08	0.039	2.20E-12	0.0268	0.2604	PDE11A
rs10083886	17	67434950	t	c	0.26	0.02	3.40E-09	-0.0163	0.2621	SOX9
rs11047239	12	24099047	g	c	0.3	0.022	4.10E-12	0.0166	0.2622	SOX5
rs7319045	13	90822575	a	g	0.39	0.024	3.70E-15	0.0156	0.2668	GPC5
rs1113765	7	55856828	g	a	0.82	0.025	1.00E-10	0.019	0.2709	SEPT14
rs6949739	7	46383928	t	a	0.91	0.037	4.80E-12	0.0285	0.2725	IGFBP3
rs10131337	14	36214267	t	c	0.24	0.027	5.60E-13	0.0173	0.2743	PAX9
rs16834765	1	32144029	t	c	0.06	0.045	1.40E-12	-0.0305	0.2747	PTP4A2
rs11750568	5	178468319	a	g	0.33	0.019	4.20E-10	-0.0153	0.2805	ADAMTS2
rs12513181	4	124055106	c	a	0.26	0.019	2.10E-08	0.0161	0.285	NUDT6
rs6879260	5	179663620	c	t	0.61	0.027	1.10E-17	-0.0146	0.286	GFPT2
rs13088462	3	51046753	c	t	0.06	0.053	1.10E-14	0.0322	0.2916	DOCK3
rs4239020	17	77769930	c	t	0.33	0.021	1.50E-11	0.0146	0.2968	CCDC57
rs17113369	1	95559811	t	c	0.97	0.07	2.40E-08	0.0374	0.2995	RWDD3
rs4656220	1	168915901	t	c	0.39	0.022	7.50E-12	0.0146	0.2998	PRRX1
rs6794009	3	61488535	g	a	0.44	0.016	2.80E-08	-0.0134	0.3011	PTPRG
rs2306694	12	54966903	g	a	0.07	0.047	1.20E-16	0.0277	0.3015	CS
rs6920372	6	109830632	g	a	0.59	0.026	5.70E-19	0.0138	0.3081	PPIL6
rs2662027	5	56290242	g	t	0.9	0.032	1.40E-11	0.022	0.3082	MIER3
rs10880969	12	45113290	c	t	0.7	0.023	1.10E-12	0.0146	0.3117	SLC38A2
rs14062	18	17704301	g	a	0.67	0.018	1.60E-08	0.014	0.3175	MIB1
rs7834383	8	13317848	t	g	0.36	0.021	1.90E-11	-0.0144	0.3186	DLC1
rs2581830	3	53109138	t	c	0.4	0.025	7.60E-16	0.013	0.328	RFT1
rs2748483	6	146377253	a	t	0.55	0.018	2.40E-09	-0.0132	0.3281	GRM1
rs3782089	11	65093395	c	t	0.94	0.053	1.00E-15	-0.0258	0.3288	SSSCA1
rs199515	17	42211804	c	g	0.8	0.023	1.60E-09	0.0163	0.3297	WNT3
rs6696239	1	225816691	g	a	0.81	0.038	2.80E-25	-0.0164	0.3299	ZNF678
rs6420435	16	80741702	a	c	0.21	0.025	1.80E-11	-0.0151	0.3308	MPHOSPH6
rs2306596	4	39020335	a	c	0.52	0.02	1.80E-11	0.0132	0.3354	RFC1
rs316618	15	39583790	t	a	0.78	0.026	9.80E-13	-0.0155	0.3393	LTK
rs724016	3	142588260	g	a	0.44	0.078	1.10E-156	0.0123	0.3484	ZBTB38
rs12621643	2	223626227	g	t	0.7	0.019	1.70E-08	-0.0133	0.3489	KCNE4
rs7692995	4	17545732	t	c	0.85	0.101	5.20E-100	-0.0169	0.3516	LCORL
rs1265097	6	31214438	c	a	0.89	0.04	6.50E-15	0.02	0.3538	PSORS1C1
rs7033940	9	6430419	g	c	0.87	0.024	3.80E-08	0.0182	0.354	UHRF2
rs692964	18	13084132	g	a	0.4	0.019	2.30E-10	-0.0123	0.3556	CEP192
rs1036821	8	135719665	g	a	0.7	0.047	2.80E-38	-0.0137	0.3563	ZFAT

Table 4.1 (Continued)

						Height		Sitting height ratio		
Rsid	**Chr**	**Position**	**Ref**	**Alt**	**Reffreq**	**Effect Size**	**P-value**	**Effect Size**	**P-value**	**Closest Gene**
rs780094	2	27594741	c	t	0.61	0.021	7.50E-12	0.0122	0.3564	GCKR
rs7466269	9	132453905	a	g	0.64	0.033	1.70E-26	-0.0126	0.3606	FUBP3
rs989393	9	100783157	t	c	0.71	0.023	5.80E-13	-0.013	0.3611	COL15A1
rs26024	5	127723921	c	a	0.35	0.023	3.90E-13	-0.0128	0.3654	FBN2
rs4640244	17	21224816	a	g	0.61	0.025	6.60E-14	-0.0136	0.3668	KCNJ12
rs11221442	11	128082834	g	c	0.75	0.027	3.00E-14	-0.0137	0.3672	FLI1
rs11683207	2	97699722	t	c	0.8	0.024	5.70E-09	-0.0183	0.3723	ZAP70
rs12987566	2	171860892	t	c	0.27	0.023	1.30E-12	0.013	0.377	METTL8
rs1832871	6	158642022	a	g	0.34	0.021	9.20E-12	-0.0124	0.3786	TULP4
rs3739707	9	112832527	c	a	0.75	0.024	1.60E-11	-0.0133	0.3793	LPAR1
rs870183	17	546561	g	a	0.53	0.017	3.80E-09	0.0114	0.381	VPS53
rs4812586	20	34978087	a	g	0.84	0.033	1.70E-16	-0.0163	0.3847	SAMHD1
rs2961830	5	50490489	a	t	0.35	0.019	1.10E-09	0.0121	0.3886	ISL1
rs1036477	15	46702218	a	g	0.9	0.032	2.80E-11	-0.0179	0.39	FBN1
rs354196	2	54819911	g	a	0.53	0.021	1.90E-12	-0.0112	0.3937	SPTBN1
rs17038954	2	1624680	t	c	0.06	0.04	1.10E-10	-0.0241	0.3959	PXDN
rs2510396	11	68174228	c	g	0.86	0.029	2.60E-12	0.0158	0.3969	GAL
rs5742915	15	72123686	c	t	0.47	0.038	1.20E-34	0.0115	0.4033	PML
rs34651	5	72179761	c	t	0.08	0.042	4.20E-13	-0.0207	0.4079	TNPO1
rs13416119	2	42316434	a	g	0.9	0.029	4.90E-09	0.0197	0.409	EML4
rs7273787	20	4046567	g	a	0.35	0.022	3.00E-12	-0.0113	0.411	SMOX
rs6974574	7	38076598	t	a	0.69	0.031	2.30E-19	-0.0116	0.4115	STARD3NL
rs10779751	1	11206923	a	g	0.28	0.02	5.80E-10	0.0118	0.4139	FRAP1
rs6952113	7	120564855	g	a	0.62	0.018	1.10E-09	-0.0112	0.4145	C7orf58
rs738288	22	38237607	g	a	0.47	0.019	1.50E-10	0.0106	0.4151	SMCR7L
rs1047014	6	19949472	c	t	0.25	0.033	7.50E-20	0.0135	0.4209	ID4
rs17807185	7	77146231	g	a	0.38	0.022	3.30E-13	0.0107	0.4358	RSBN1L
rs12904334	15	70629759	a	g	0.02	0.094	1.50E-13	0.0417	0.4384	ARIH1
rs3132297	9	136441687	g	a	0.83	0.024	6.40E-09	0.0136	0.4409	RXRA
rs2815379	1	67283062	g	a	0.71	0.018	2.50E-08	-0.011	0.4445	SLC35D1
rs1923367	10	80802835	g	c	0.52	0.029	3.20E-22	0.0104	0.4457	ZCCHC24
rs7177711	15	60167263	a	g	0.54	0.021	1.60E-13	0.0101	0.4471	FAM148A
rs6988484	8	49576333	c	t	0.25	0.023	4.20E-12	-0.0116	0.4482	EFCAB1
rs2057291	20	56905438	a	g	0.34	0.02	4.80E-10	0.0104	0.4508	GNAS
rs12137162	1	19635983	a	c	0.28	0.019	4.90E-09	0.0109	0.4535	CAPZB
rs1550162	8	117632713	g	a	0.29	0.024	2.90E-14	0.011	0.4641	EIF3H
rs6813055	4	88849055	a	t	0.49	0.017	5.50E-09	0.0098	0.4663	DMP1

Table 4.1 (Continued)

						Height		Sitting height ratio		
Rsid	**Chr**	**Position**	**Ref**	**Alt**	**Reffreq**	**Effect Size**	**P-value**	**Effect Size**	**P-value**	**Closest Gene**
rs4072910	19	8550031	g	c	0.56	0.031	9.90E-18	-0.0101	0.4685	ADAMTS10
rs7551732	1	88911629	a	t	0.61	0.027	2.50E-19	-0.0096	0.471	PKN2
rs1401795	17	52194651	a	g	0.51	0.03	5.00E-25	-0.0093	0.4711	C17orf67
rs4425077	2	216118761	g	c	0.41	0.02	1.40E-11	-0.0095	0.4726	FN1
rs9841435	3	192593854	g	a	0.32	0.019	6.10E-10	0.0099	0.4745	CCDC50
rs4350272	10	25096124	a	g	0.28	0.02	2.00E-09	0.0108	0.4747	ARHGAP21
rs7980687	12	122388664	a	g	0.2	0.036	1.30E-21	0.0118	0.4763	SBNO1
rs888403	18	2756938	g	a	0.36	0.019	1.00E-08	-0.0104	0.4772	SMCHD1
rs564914	1	47687820	t	a	0.39	0.025	4.10E-17	0.0096	0.4773	FOXD2
rs17122659	12	58243190	g	a	0.12	0.032	6.80E-11	-0.0154	0.4782	SLC16A7
rs749234	2	144947819	a	g	0.32	0.018	1.30E-08	0.0099	0.4809	ZEB2
rs4605213	17	46599746	c	g	0.34	0.019	2.00E-09	-0.0098	0.482	NME1-NME2
rs567401	1	85760746	t	c	0.17	0.028	3.10E-11	-0.0132	0.4906	DDAH1
rs7743622	6	132772065	g	c	0.58	0.018	4.60E-08	-0.0094	0.4917	MOXD1
rs9292468	5	32854830	t	c	0.4	0.053	4.80E-46	-0.0092	0.4987	C5orf23
rs9434723	1	9214869	a	g	0.16	0.028	8.60E-12	-0.0122	0.5023	H6PD
rs3812040	5	39461777	t	c	0.72	0.024	3.50E-13	0.0102	0.5048	DAB2
rs757081	11	17308259	g	c	0.34	0.024	3.20E-14	-0.0092	0.5057	NUCB2
rs9309101	2	43483116	g	a	0.33	0.02	3.20E-10	0.0092	0.5068	THADA
rs11687941	2	241840083	c	g	0.75	0.025	4.00E-13	-0.01	0.5077	HDLBP
rs7112925	11	66582736	c	t	0.64	0.023	2.60E-14	-0.0089	0.5116	RHOD
rs11835818	12	120979192	c	t	0.49	0.017	4.30E-09	0.0088	0.5121	BCL7A
rs11090631	22	44225035	t	c	0.2	0.022	1.50E-08	-0.011	0.5143	RIBC2
rs17783015	12	88755517	c	t	0.84	0.025	5.20E-10	0.0122	0.5169	ATP2B1
rs318095	17	44329733	t	c	0.46	0.023	3.30E-15	0.0083	0.5228	ATP5G1
rs7849585	9	138251691	t	g	0.33	0.036	9.80E-29	0.0091	0.5248	QSOX2
rs16964211	15	49317787	g	a	0.95	0.044	4.80E-09	0.0188	0.5248	CYP19A1
rs817300	9	97420043	g	a	0.93	0.07	2.20E-23	-0.0177	0.5274	PTCH1
rs7567288	2	134151294	c	t	0.2	0.028	3.80E-13	-0.0107	0.529	NAP5
rs486359	6	160694431	c	g	0.49	0.017	1.60E-08	0.0085	0.5299	SLC22A3
rs17250196	7	99655132	t	g	0.07	0.044	8.70E-10	0.0205	0.5335	GATS/
rs2023693	16	20787541	g	a	0.6	0.017	1.40E-08	0.0081	0.5388	DCUN1D3
rs2682587	19	48774269	a	c	0.2	0.024	4.30E-10	-0.0103	0.5442	XRCC1
rs429433	8	8785304	a	g	0.05	0.046	6.70E-11	-0.0228	0.5449	MFHAS1
rs1659127	16	14295806	a	g	0.34	0.03	1.20E-19	-0.0087	0.546	MKL2
rs2145357	6	116558135	g	a	0.27	0.022	1.70E-11	0.0091	0.5461	NT5DC1
rs763318	4	12572672	g	a	0.53	0.025	4.40E-17	-0.008	0.5476	RAB28

Table 4.1 (Continued)

						Height		Sitting height ratio		
Rsid	**Chr**	**Position**	**Ref**	**Alt**	**Reffreq**	**Effect Size**	**P-value**	**Effect Size**	**P-value**	**Closest Gene**
rs32855	5	79871948	a	g	0.78	0.024	4.50E-11	0.0098	0.5518	FAM151B
rs2631676	10	93027389	g	a	0.19	0.028	1.50E-12	0.0103	0.5539	PCGF5
rs2974438	5	168183481	g	a	0.8	0.038	3.90E-26	-0.0098	0.5541	SLIT3
rs4883972	13	73956482	c	g	0.55	0.019	1.70E-10	0.008	0.5565	KLF12
rs2806561	1	23377382	a	g	0.57	0.027	2.70E-21	0.0076	0.5588	LUZP1
rs2856321	12	11747040	g	a	0.36	0.031	1.00E-25	-0.0081	0.5603	ETV6
rs12871822	13	48099041	g	t	0.34	0.018	2.70E-09	-0.0082	0.5645	CYSLTR2
rs7162825	15	61226239	t	c	0.5	0.016	2.80E-08	0.0073	0.5716	LACTB
rs6435143	2	202902501	a	c	0.44	0.019	2.40E-10	0.0074	0.5737	NOP5
rs7027110	9	108638867	a	g	0.23	0.032	2.30E-20	-0.0087	0.5738	ZNF462
rs4896582	6	142745570	g	a	0.7	0.051	6.90E-58	-0.0082	0.5769	GPR126
rs497273	12	119689065	c	g	0.38	0.019	2.80E-10	-0.0079	0.5774	SPPL3
rs12882130	14	102948527	c	g	0.63	0.024	2.90E-14	-0.0077	0.5819	MARK3
rs6540834	1	212694042	c	t	0.66	0.028	1.70E-17	0.0073	0.5822	PTPN14
rs6561319	13	46010121	a	c	0.64	0.021	1.50E-11	-0.0076	0.5827	LRCH1
rs11855014	15	83529838	g	a	0.71	0.022	1.40E-10	-0.0078	0.5926	PDE8A
rs2013265	8	24148445	c	t	0.75	0.027	9.20E-17	0.008	0.6017	ADAM28
rs165189	5	139125931	g	a	0.15	0.031	2.70E-11	0.0104	0.6018	PSD2
rs6962887	7	134696326	t	g	0.68	0.022	9.60E-11	0.0078	0.6019	CNOT4
rs10748128	12	68113925	t	g	0.35	0.038	4.60E-29	0.0078	0.6033	FRS2
rs39623	5	129082520	a	t	0.08	0.045	7.20E-17	0.0129	0.6075	ADAMTS19
rs2715094	7	50697946	g	a	0.25	0.021	1.20E-09	-0.008	0.6112	GRB10
rs2289195	2	25316987	a	g	0.43	0.042	3.00E-34	-0.0068	0.613	DNMT3A
rs4986172	17	40571807	c	t	0.65	0.038	1.60E-31	0.0069	0.6196	ACBD4
rs7652177	3	173451771	g	c	0.51	0.037	1.00E-36	0.0065	0.6207	FNDC3B
rs540652	2	169415674	t	c	0.46	0.021	6.20E-13	-0.0065	0.6215	NOSTRIN
rs6688100	1	158666210	t	c	0.48	0.016	2.20E-08	-0.0064	0.6225	VANGL2
rs3818416	13	77372469	c	a	0.78	0.021	4.60E-09	0.0078	0.6246	EDNRB
rs7899004	10	104331425	t	c	0.56	0.024	4.30E-17	-0.0066	0.6268	SUFU
rs9993613	4	73694878	t	g	0.47	0.03	7.80E-25	-0.0065	0.6312	ADAMTS3
rs6761041	2	224738373	t	c	0.55	0.024	1.70E-16	-0.0062	0.6318	SERPINE2
rs16895130	6	42032909	g	a	0.28	0.025	2.00E-14	-0.0072	0.6319	CCND3
rs12474201	2	46774789	a	g	0.36	0.029	1.70E-20	-0.0065	0.6333	SOCS5
rs6971575	7	95877584	c	g	0.29	0.022	3.60E-10	-0.007	0.6408	SLC25A13
rs1996422	4	48382108	g	a	0.28	0.022	1.30E-11	-0.0073	0.6428	FRYL
rs10790381	11	119762705	a	g	0.82	0.027	1.20E-12	-0.0081	0.643	ARHGEF12
rs2123731	19	4880473	a	g	0.73	0.025	6.80E-13	-0.0067	0.6462	UHRF1

Table 4.1 (Continued)

						Height		Sitting height ratio		
Rsid	**Chr**	**Position**	**Ref**	**Alt**	**Reffreq**	**Effect Size**	**P-value**	**Effect Size**	**P-value**	**Closest Gene**
rs6902771	6	152199574	t	c	0.46	0.029	1.50E-21	0.0062	0.6485	ESR1
rs6658763	1	145158997	c	t	0.92	0.034	5.10E-10	0.011	0.6504	FMO5
rs6439168	3	130533633	g	a	0.79	0.038	5.20E-26	-0.0071	0.6545	H1FX
rs2120335	2	68348506	g	a	0.59	0.018	8.40E-10	0.0059	0.6546	PPP3R1
rs6080830	20	17719113	a	g	0.56	0.016	1.50E-08	-0.0058	0.6559	BANF2
rs2149163	9	16445833	c	g	0.4	0.02	2.90E-11	0.0059	0.656	BNC2
rs4843367	16	84975391	c	t	0.66	0.02	7.10E-10	-0.006	0.6634	FOXF1
rs10152739	15	36271158	t	a	0.25	0.023	3.00E-11	-0.0064	0.6699	SPRED1
rs13113518	4	56094405	c	t	0.36	0.017	1.60E-08	0.0059	0.6711	CLOCK
rs17556750	4	82374592	a	c	0.31	0.044	2.80E-43	-0.0061	0.6781	PRKG2
rs10972628	9	35927611	g	a	0.74	0.02	1.20E-08	-0.006	0.6846	OR2S2
rs11156098	6	156629523	t	c	0.12	0.029	5.40E-10	-0.0088	0.6867	ARID1B
rs7007200	8	109854114	g	c	0.69	0.017	4.70E-08	0.0059	0.6896	TMEM74
rs3812423	8	25354627	g	c	0.64	0.021	1.00E-12	0.0056	0.6905	KCTD9
rs11648796	16	732191	g	a	0.25	0.034	7.70E-19	-0.0068	0.6928	NARFL
rs11799609	1	241684940	t	g	0.16	0.026	7.00E-10	0.0073	0.6986	SDCCAG8
rs10794175	10	126348063	t	g	0.43	0.021	4.00E-12	0.0052	0.6991	FAM53B
rs11880992	19	2127403	a	g	0.4	0.032	1.10E-26	-0.0051	0.7003	DOT1L
rs932445	6	2112224	t	c	0.59	0.021	1.10E-11	0.0053	0.7006	GMDS
rs4601530	1	24916698	c	t	0.74	0.026	5.90E-15	0.0056	0.704	CLIC4
rs2072268	17	63814947	g	a	0.52	0.021	1.70E-11	-0.0053	0.7058	ARSG
rs11616067	12	114877557	a	g	0.76	0.02	1.00E-08	0.0059	0.709	MED13L
rs7716219	5	54990828	t	c	0.31	0.029	2.50E-21	0.0055	0.7092	SLC38A9
rs4624820	5	141661972	a	g	0.52	0.018	1.80E-10	-0.005	0.7095	SPRY4
rs12693589	2	191540907	c	t	0.25	0.022	5.50E-11	0.0055	0.7114	STAT1
rs2302580	4	8659534	c	t	0.58	0.029	1.20E-15	-0.0052	0.7155	CPZ
rs992157	2	218863025	a	g	0.57	0.018	1.60E-09	-0.0048	0.7166	PNKD
rs761391	6	85504822	c	t	0.46	0.021	6.10E-10	-0.0049	0.7183	TBX18
rs2811594	1	93115870	g	a	0.63	0.023	3.10E-13	0.0048	0.7187	FAM69A
rs2058092	14	73002719	t	c	0.56	0.017	8.40E-09	0.0047	0.7243	NUMB
rs6714546	2	33214929	g	a	0.72	0.035	2.40E-24	0.0051	0.7327	LTBP1
rs17330192	6	17697354	c	t	0.28	0.019	1.20E-08	-0.0052	0.7348	FAM8A1
rs10883563	10	102674370	a	c	0.55	0.023	3.20E-15	0.0045	0.7387	FAM178A
rs2175513	3	68705056	g	a	0.43	0.017	2.00E-08	0.0044	0.7396	FAM19A1
rs7853235	9	85850602	t	c	0.2	0.029	8.80E-15	0.0053	0.7503	RMI1
rs3923086	17	60979950	c	a	0.6	0.025	2.80E-14	-0.0046	0.7511	AXIN2
rs1980850	14	67716941	g	a	0.83	0.029	2.40E-13	-0.0053	0.7554	RAD51L1

Table 4.1 (Continued)

						Height		Sitting height ratio		
Rsid	**Chr**	**Position**	**Ref**	**Alt**	**Reffreq**	**Effect Size**	**P-value**	**Effect Size**	**P-value**	**Closest Gene**
rs2298265	1	149525667	c	t	0.88	0.03	6.90E-11	-0.0061	0.7574	ZNF687
rs1155939	6	126907826	a	c	0.5	0.042	1.30E-47	0.0041	0.7606	C6orf173
rs7253628	19	35739109	g	a	0.16	0.024	6.20E-10	0.0054	0.7612	ZNF536
rs1325596	1	175060689	a	g	0.57	0.025	2.10E-18	0.0039	0.7621	PAPPA2
rs6746356	2	174524144	a	c	0.75	0.019	1.80E-08	0.0045	0.7647	SP3
rs11618507	13	29070751	t	g	0.25	0.023	1.80E-10	0.0049	0.7678	SLC7A1
rs975210	15	68151406	a	g	0.18	0.034	7.90E-17	0.0053	0.7698	TLE3
rs929637	7	12243047	g	t	0.78	0.022	1.90E-10	0.0046	0.7731	TMEM106B
rs1571892	9	93298657	c	a	0.29	0.017	5.00E-08	0.0041	0.7733	NFIL3
rs12669267	7	72942572	c	t	0.87	0.029	2.60E-08	-0.0065	0.7742	WBSCR28
rs6838153	4	122940449	g	a	0.34	0.021	7.90E-12	-0.0039	0.7831	EXOSC9
rs9835332	3	56642722	g	c	0.54	0.028	3.00E-22	-0.0035	0.7845	C3orf63
rs8058684	16	52072619	a	g	0.3	0.021	6.40E-11	0.0038	0.787	RBL2
rs7261425	20	20016635	c	g	0.71	0.021	5.10E-10	0.0039	0.7874	C20orf26
rs999599	9	116051416	t	c	0.37	0.017	9.70E-09	-0.0036	0.7884	COL27A1
rs8097893	18	73112043	a	g	0.95	0.044	1.30E-10	0.0088	0.7893	GALR1
rs12639764	4	106435654	t	c	0.62	0.027	5.00E-20	0.0036	0.7913	TET2
rs1074683	20	31768314	c	g	0.76	0.047	2.40E-42	0.004	0.7939	PXMP4
rs1420023	12	12767378	c	g	0.88	0.028	1.60E-08	0.0056	0.7965	CDKN1B
rs2284746	1	17179262	g	c	0.52	0.04	1.20E-40	0.0033	0.8027	MFAP2
rs7154721	14	91497101	t	c	0.57	0.027	1.30E-20	0.0032	0.8072	TRIP11
rs2345835	2	18438433	c	t	0.54	0.019	3.40E-10	-0.0033	0.809	RDH14
rs12435366	14	34908140	c	t	0.73	0.023	3.60E-11	0.0038	0.8099	NFKBIA
rs1552173	17	74230437	c	t	0.46	0.018	2.00E-10	-0.0031	0.8137	PSCD1
rs6600365	1	41328840	c	t	0.43	0.027	9.90E-21	-0.0031	0.8151	SCMH1
rs833152	2	182927346	c	a	0.42	0.016	4.00E-08	0.003	0.8195	PDE1A
rs12538407	7	23487841	a	g	0.6	0.043	1.00E-35	-0.0031	0.8212	IGF2BP3
rs11624136	14	58758573	a	g	0.5	0.017	3.30E-09	-0.0029	0.8215	DAAM1
rs2829941	21	26130806	t	g	0.61	0.017	3.20E-08	-0.003	0.8222	APP
rs632124	11	118118445	a	t	0.42	0.022	2.20E-14	0.0029	0.8267	DDX6
rs13006748	2	20015300	c	g	0.3	0.023	1.00E-11	0.003	0.8387	WDR35
rs568610	8	27583914	t	c	0.24	0.023	1.20E-11	-0.0031	0.8416	SCARA3
rs2280470	15	87196630	a	g	0.33	0.031	5.50E-21	0.0028	0.8419	ACAN
rs9217	17	7303812	c	t	0.37	0.03	4.40E-23	-0.0026	0.846	ZBTB4
rs1950500	14	23900690	t	c	0.3	0.031	2.70E-22	0.0028	0.8481	NFATC4
rs4785393	16	48816984	g	a	0.16	0.023	1.80E-08	0.0032	0.8497	PAPD5
rs806794	6	26308656	a	g	0.71	0.055	7.80E-59	-0.0028	0.8512	HIST1H2BF

Table 4.1 (Continued)

						Height		Sitting height ratio		
Rsid	**Chr**	**Position**	**Ref**	**Alt**	**Reffreq**	**Effect Size**	**P-value**	**Effect Size**	**P-value**	**Closest Gene**
rs2059877	19	52880621	t	g	0.26	0.019	2.30E-08	-0.0027	0.8556	GLTSCR1
rs3958122	4	1663729	t	c	0.35	0.027	1.20E-17	-0.0025	0.8564	SLBP
rs7284476	22	36459278	a	g	0.43	0.016	4.60E-08	-0.0024	0.8576	TRIOBP
rs9291926	5	67635412	t	g	0.49	0.019	9.30E-10	0.0024	0.8582	PIK3R1
rs4733724	8	130792910	a	g	0.8	0.05	2.60E-42	-0.003	0.8588	MLZE
rs7740107	6	130416154	t	a	0.26	0.034	5.10E-22	0.0027	0.8606	L3MBTL3
rs11049611	12	28491511	c	t	0.7	0.037	6.30E-30	-0.0025	0.8613	CCDC91
rs991946	6	166249852	c	t	0.52	0.022	6.80E-14	-0.0023	0.8628	T
rs1582931	5	122685098	g	a	0.52	0.028	2.70E-20	0.0023	0.8635	CCDC100
rs8017130	14	22828996	g	a	0.69	0.023	6.50E-12	-0.0025	0.8655	HOMEZ
rs13388725	2	108413622	g	a	0.41	0.018	2.40E-09	0.0022	0.8672	GCC2
rs10863936	1	210304421	g	a	0.47	0.021	9.00E-13	-0.0022	0.8677	DTL
rs6911389	6	144121322	t	g	0.35	0.018	1.40E-08	-0.0024	0.8686	PHACTR2
rs1599473	8	120544539	g	t	0.75	0.026	4.10E-14	-0.0024	0.8772	NOV
rs9395264	6	47582981	g	t	0.68	0.02	1.10E-10	-0.0022	0.8805	CD2AP
rs10995319	10	52432893	t	c	0.76	0.019	2.20E-08	0.0022	0.8871	PRKG1
rs4332428	10	4955434	a	g	0.88	0.036	1.10E-15	-0.0028	0.8916	AKR1C1
rs3915129	3	41218746	g	t	0.47	0.016	3.80E-08	0.0017	0.8946	CTNNB1
rs11783655	8	145109561	t	a	0.61	0.019	5.40E-10	0.0018	0.8956	PLEC1
rs11684404	2	88705737	c	t	0.34	0.032	2.30E-25	-0.0018	0.8967	EIF2AK3
rs1544196	1	222699405	g	a	0.77	0.019	2.80E-08	-0.002	0.8991	WDR26
rs13150868	4	152400121	t	g	0.44	0.018	1.20E-09	-0.0016	0.9087	ESSPL
rs9392918	6	7653630	c	t	0.47	0.041	2.40E-43	0.0015	0.9126	BMP6
rs10780910	9	90039075	t	a	0.43	0.028	6.20E-21	0.0014	0.9148	SPIN1
rs3118905	13	50003335	g	a	0.72	0.044	1.60E-33	0.0016	0.9167	DLEU7
rs8103068	19	17383869	t	c	0.86	0.032	5.00E-12	0.0021	0.9176	BST2
rs2247870	5	90187345	a	g	0.55	0.017	1.60E-08	-0.0013	0.9211	GPR98
rs915506	10	97795064	g	a	0.65	0.019	1.50E-10	0.0013	0.9275	CCNJ
rs7069985	10	27930837	g	a	0.25	0.023	1.30E-11	-0.0014	0.9301	RAB18
rs1945237	11	55986645	c	t	0.09	0.03	8.70E-09	-0.0021	0.9312	OR5M9
rs1576900	9	18619792	g	a	0.7	0.019	6.50E-09	-0.0012	0.9342	ADAMTSL1
rs2781373	14	64637968	g	a	0.62	0.021	2.90E-12	0.0011	0.9376	MAX
rs425277	1	2059032	t	c	0.28	0.028	4.80E-17	-0.0011	0.9407	PRKCZ
rs17391694	1	78396214	t	c	0.12	0.04	4.00E-14	-0.0015	0.9437	GIPC2
rs8102380	19	10662185	g	a	0.31	0.021	5.90E-12	0.0008	0.9538	ILF3
rs11779459	8	124049732	t	c	0.35	0.018	3.40E-08	-0.0008	0.9573	ZHX2
rs822531	7	148260692	t	c	0.78	0.035	1.10E-18	0.001	0.9574	EZH2

Table 4.1 (Continued)

						Height		Sitting height ratio		
Rsid	**Chr**	**Position**	**Ref**	**Alt**	**Reffreq**	**Effect Size**	**P-value**	**Effect Size**	**P-value**	**Closest Gene**
rs720390	3	187031377	a	g	0.38	0.068	8.70E-58	-0.0007	0.9626	IGF2BP2
rs12330322	3	72538045	c	t	0.78	0.034	3.50E-22	0.0007	0.9668	RYBP
rs991967	1	216682074	c	a	0.28	0.038	4.10E-32	0.0006	0.9687	TGFB2
rs3763631	9	35798334	c	g	0.69	0.021	3.40E-11	0.0005	0.9711	NPR2
rs897080	2	44627706	c	t	0.26	0.033	2.60E-21	-0.0004	0.9778	C2orf34
rs7162542	15	82305294	g	c	0.55	0.03	7.70E-16	-0.0003	0.9806	ADAMTSL3
rs7568069	2	71437993	g	a	0.42	0.021	1.40E-13	0.0003	0.9811	ZNF638
rs6462432	7	32902049	a	g	0.39	0.017	1.80E-08	-0.0003	0.9842	KBTBD2
rs6691924	1	54726833	t	c	0.9	0.031	4.70E-10	-0.0004	0.9853	ACOT11
rs526896	5	134384604	t	g	0.73	0.037	2.60E-27	-0.0003	0.9855	PITX1
rs2854207	17	59300839	g	c	0.27	0.04	4.20E-28	0.0001	0.9921	CSH2
rs2074977	19	3385028	c	a	0.36	0.028	4.60E-20	-0.0001	0.9927	NFIC
rs1199734	13	20468246	g	t	0.81	0.021	4.00E-08	0.0001	0.9942	LATS2
rs2237886	11	2767307	t	c	0.11	0.042	1.60E-17	-0.0001	0.9969	KCNQ1
rs8069300	17	11924957	g	c	0.47	0.016	1.70E-08	0	0.9985	MAP2K4

The 421 height associated SNPs and their effect sizes and P-values with sitting height ratio (SHR). The reference allele has been aligned such that the effect size for height is always positive. The variants are ordered with decreasing significance to SHR.

large difference of SHR between people of different ancestral background and there is more than 1 standard deviation difference between the SHR of European American and African Americans. While uncovering the underlying genetic reason for such a difference could improve our understanding of developmental biology during growth, differences of other phenotypes between European and African Americans could also be studied to determine if genetics is the primary cause for the difference. For example, studies of cancer rates have shown that African Americans have significantly higher incident rates of cancer [16] and subsequent genetic studies have uncovered common variants associated with prostate cancer that could explain for the greater incidence in African Americans [17,18].

Interestingly, we managed to observe some loci that reach genome wide significance even with our relatively small sample size. The lead variant (rs201786365) discovered in our African American samples is not in any genes. The closest gene (120kb upstream) is *ABHD5*, where mutations in *ABHD5* (also known as CGI-58) has been associated with Chanarin-Dorfman syndrome, a syndrome characterized by the individual's inability to process triglycerides which can lead to having short stature [19]. The variants discovered from our studies in European Americans, rs140449984 (PTPRM) and rs5959358 (ITM2A) are also interesting. PTPRM, while not known to be associated with height, is associated with a syndrome called deletion 18p syndrome which can lead to mental and growth retardation, and craniofacial dysmorphism [20]. While the variant (rs5959358) does not lie in any gene, the closest gene, ITM2A (70kb upstream), a gene found on the X-chromosome is associated with SHR in women but not in men. The locus have also been reported to be strongly associated with height [13]. This result suggests that the variant responsible for altering SHR plays a role in escaping dosage compensation in women that results in altered SHR [13,21,22]. Finally, we also show that most of the SNPs

associated with height do show effects that alter SHR. While variants that increase height have slightly higher probability to be associated with decreased SHR (e.g. FGFR2, CDK6), some variants that increase height are also associated with increase SHR (e.g. FAM46A, WWP2). Other variants while having a strong effect on overall height, they do not seem to be associated with SHR (e.g. HMGA2, ZBTB38). These 3 classes of genes might be clustered distinct biological pathways that have very different mechanism on how they alter overall height.

In conclusion, this study is a large scale whole genome experiment to discover the underlying genetic basis for differences in body proportion using the sitting height ratio (SHR) as a read out. We uncovered a few loci that are significantly associated with SHR and that there are a significant number of loci associated with height that also alters the SHR. These results suggest that SHR is also polygenic and further studies of larger sample sizes is required to explain the full genetic spectrum of SHR.

MATERIALS AND METHODS

Quality control (QC)

The data were downloaded from dbGAP and passed through our quality control pipeline. The QC is largely done using PLINK [23] software. Samples that have ambiguous or incorrect gender were filtered out (using --check-sex option in PLINK). SNPs that have > 5% missing rate were filtered out. Samples that have > 2% missing SNPs were removed. SNPs that have minor allele frequencies < 1% were dropped. We then examine samples that have extreme heterozygosity and removed samples that were +/- 4 standard deviations (using –het option in PLINK). The SNP annotations for chromosome and base-pair positions were set to the coordinates of hg19

(GRCh37) using liftover. We then calculated pairwise IBD/IBS (using –genome option in PLINK) and remove individuals that have excessive matching with other individuals (PI_HAT > 0.05). The samples were then superimposed on the HAPMAP [24] version 3 by comparing principal components using SMARTPCA [25]. Samples that do not belong to the right PCA cluster were removed. SNPs that have excessive plate-effects ($P < 1 \times 10^{-7}$) were dropped. For samples that are of European ancestry, SNPs that have excessive deviation from Hardy-Weinberg equilibrium ($P < 1 \times 10^{-7}$) were dropped.

Determining global European ancestry in African American individuals

For the African American individuals (ARIC and CARDIA cohorts), the global European ancestry was calculated by SMARTPCA from the CEU and YRI samples of HAPMAP version 3. The CEU individuals are proxies of European ancestry while the YRI individuals are proxies of African ancestry. The principal components were calculated using only the CEU and YRI individuals while projecting them onto the ARIC and CARDIA African Americans. The first principal component is taken to be the axis that represents the degree of global European admixture for each of our individuals.

Genotype Imputation

The genotypes were phased using SHAPEIT2 [26] and imputed using IMPUTE2 [27]. The imputation panel used were from the 1000 genomes [28] containing 379 Europeans, 246 Africans and African-Americans, 286 Asians and 181 Latin Americans. The imputation panel consists of approximately 22 million variants (SNPs and indels). For the X-chromosome, only the non-

pseudo autosomal region was imputed. The phasing and imputation were done separately for males and females.

Genome wide association

The associations were performed using sitting height ratio (SHR) adjusted for sex, height, age and body mass index (BMI). SHR was calculated by taking sitting height divided by total height. Individuals that were missing for SHR or any of the above covariate were discarded. Only unrelated individuals were used, i.e. no pair of individuals has PI_HAT > 0.05. The SHR were inverse-normalized per cohort. The top 10 principal components (PCs) were calculated using SMARTPCA and any PCs that had an association with SHR ($P < 0.05$) were used as a covariate as well. For African American individuals, the global percentage European admixture was included as an additional covariate. The association for the imputed variants with SHR was performed by a linear regression (--linear command with PLINK). The resulting association results for each cohort were then meta-analyzed together using METAL [29] with GC correction turned on. Variants on the X-chromosome were analyzed separately between males and females.

Atherosclerosis Risk In Communities (ARIC) cohort

We obtained genotypic and phenotypic data from dbGAP. There were initially 13,113 samples (European + African Americans) and after performing the quality control (QC) procedure, there were 7,257 (3,551 males and 3,706 females) European Americans and 2,354 African Americans (894 males and 1,460 females). The genotypes were typed using the Affymetrix Genome-Wide Human SNP Array 6.0 platform.

Coronary Artery Risk Development in Young Adults (CARDIA) cohort

We obtained genotypic and phenotypic data from dbGAP. There were initially 1,675 European American samples and after performing the quality control (QC) procedure, there were 1,047 (494 males and 553 females) samples remaining. For African Americans, there were initially 1,393 samples and after performing the QC procedure, there were 715 (275 males and 440 females) samples remaining. The genotypes were typed using the Affymetrix Genome-Wide Human SNP Array 6.0 platform.

Cardiovascular Health Study (CHS) cohort

We obtained genotypic and phenotypic data from dbGAP. There were initially 3,980 European American samples and after performing the quality control (QC) procedure, there were 2,926 (1,163 males and 1,763 females) samples remaining. The genotypes were typed using the Illumina HumanCNV370v1-Duo platform.

Framingham Heart Study (FHS) cohort

We obtained genotypic and phenotypic data from dbGAP. The FHS cohort is largely data with family pedigrees. Sitting height measurements were only observed for the original cohort. After removing samples that do not have sitting height measurements and that are unrelated, there were 713 (269 males and 444 females) samples remaining. The genotypes were typed using the Affymetrix 500K platform.

REFERENCES

1. Wadsworth MEJ, Hardy RJ, Paul AA, Marshall SF, Cole TJ (2002) Leg and trunk length at 43 years in relation to childhood health, diet and family circumstances; evidence from the 1946 national birth cohort. Int J Epidemiol 31: 383–390. doi:10.1093/ije/31.2.383.

2. Johnston LW, Harris SB, Retnakaran R, Gerstein HC, Zinman B, et al. (2013) Short leg length, a marker of early childhood deprivation, is associated with metabolic disorders underlying type 2 diabetes: the PROMISE cohort study. Diabetes Care 36: 3599–3606. doi:10.2337/dc13-0254.

3. De Arriba Muñoz A, Domínguez Cajal M, Rueda Caballero C, Labarta Aizpún JI, Mayayo Dehesa E, et al. (2013) Sitting height/standing height ratio in a spanish population from birth to adulthood. Arch Argent Pediatría 111: 309–314. doi:10.1590/S0325-00752013000400009.

4. Fredriks A, van Buuren S, van Heel WJM, Dijkman-Neerincx R, Verloove-Vanhoric... S, et al. (2005) Nationwide age references for sitting height, leg length, and sitting height/height ratio, and their diagnostic value for disproportionate growth disorders. Arch Dis Child 90: 807–812. doi:10.1136/adc.2004.050799.

5. Krakow D, Rimoin DL (2010) The skeletal dysplasias. Genet Med 12: 327–341. doi:10.1097/GIM.0b013e3181daae9b.

6. Stokes DC, Pyeritz RE, Wise RA, Fairclough D, Murphy EA (1988) Spirometry and chest wall dimensions in achondroplasia. Chest 93: 364–369.

7. Hertel NT, Müller J (1994) Anthropometry in skeletal dysplasia. J Pediatr Endocrinol 7: 155–161.

8. Bogin B, Varela-Silva MI (2010) Leg length, body proportion, and health: a review with a note on beauty. Int J Environ Res Public Health 7: 1047–1075. doi:10.3390/ijerph7031047.

9. The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. The ARIC investigators (1989). Am J Epidemiol 129: 687–702.

10. Gardin JM, Wagenknecht LE, Anton-Culver H, Flack J, Gidding S, et al. (1995) Relationship of Cardiovascular Risk Factors to Echocardiographic Left Ventricular Mass in Healthy Young Black and White Adult Men and Women The CARDIA Study. Circulation 92: 380–387. doi:10.1161/01.CIR.92.3.380.

11. Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, et al. (2009) The genetic structure and history of Africans and African Americans. Science 324: 1035–1044. doi:10.1126/science.1172257.

12. Redaelli C, Coleman RA, Moro L, Dacou-Voutetakis C, Elsayed SM, et al. (2010) Clinical and genetic characterization of Chanarin-Dorfman syndrome patients: first report of large

deletions in the ABHD5 gene. Orphanet J Rare Dis 5: 33. doi:10.1186/1750-1172-5-33.

13. Tukiainen T, Pirinen M, Sarin A-P, Ladenvall C, Kettunen J, et al. (2014) Chromosome X-Wide Association Study Identifies Loci for Fasting Insulin and Height and Evidence for Incomplete Dosage Compensation. PLoS Genet 10: e1004127. doi:10.1371/journal.pgen.1004127.

14. Sanna S, Jackson AU, Nagaraja R, Willer CJ, Chen W-M, et al. (2008) Common variants in the GDF5-BFZB region are associated with variation in human height. Nat Genet 40: 198–203. doi:10.1038/ng.74.

15. KUCZMARSKI MF, KUCZMARSKI RJ, NAJJAR M (2001) Effects of Age on Validity of Self-Reported Height, Weight, and Body Mass Index: Findings from the Third National Health and Nutrition Examination Survey, 1988–1994. J Am Diet Assoc 101: 28–34. doi:10.1016/S0002-8223(01)00008-6.

16. Landis SH, Murray T, Bolden S, Wingo PA (1999) Cancer statistics, 1999. CA Cancer J Clin 49: 8–31. doi:10.3322/canjclin.49.1.8.

17. Amundadottir LT, Sulem P, Gudmundsson J, Helgason A, Baker A, et al. (2006) A common variant associated with prostate cancer in European and African populations. Nat Genet 38: 652–658. doi:10.1038/ng1808.

18. Haiman CA, Patterson N, Freedman ML, Myers SR, Pike MC, et al. (2007) Multiple regions within 8q24 independently affect risk for prostate cancer. Nat Genet 39: 638–644. doi:10.1038/ng2015.

19. Bruno C, Bertini E, Di Rocco M, Cassandrini D, Ruffa G, et al. (2008) Clinical and genetic characterization of Chanarin–Dorfman syndrome. Biochem Biophys Res Commun 369: 1125–1128. doi:10.1016/j.bbrc.2008.03.010.

20. Portno?? M-F, Gruchy N, Marlin S, Finkel L, Denoyelle F, et al. (2007) Midline defects in deletion 18p syndrome: clinical and molecular characterization of three patients: Clin Dysmorphol 16: 247–252. doi:10.1097/MCD.0b013e328235a572.

21. Bondy CA, Cheng C (2009) Monosomy for the X chromosome. Chromosome Res 17: 649–658. doi:10.1007/s10577-009-9052-z.

22. Castagné R, Zeller T, Rotival M, Szymczak S, Truong V, et al. (2011) Influence of sex and genetic variability on expression of X-linked genes in human monocytes. Genomics 98: 320–326. doi:10.1016/j.ygeno.2011.06.009.

23. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559–575. doi:10.1086/519795.

24. International HapMap Consortium (2003) The International HapMap Project. Nature 426: 789–796. doi:10.1038/nature02168.

25. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38: 904–909. doi:10.1038/ng1847.

26. Delaneau O, Zagury J-F, Marchini J (2013) Improved whole-chromosome phasing for disease and population genetic studies. Nat Methods 10: 5–6. doi:10.1038/nmeth.2307.

27. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR (2012) Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet 44: 955–959. doi:10.1038/ng.2354.

28. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56–65. doi:10.1038/nature11632.

29. Willer CJ, Li Y, Abecasis GR (2010) METAL: fast and efficient meta-analysis of genomewide association scans. Bioinforma Oxf Engl 26: 2190–2191. doi:10.1093/bioinformatics/btq340.

Chapter 5

Concluding Remarks

OVERVIEW

The development of whole genome genotyping technologies (DNA microarrays, whole genome sequencing, etc) coupled with computational capabilities for performing genotype-phenotype associations have allowed genome wide association studies (GWAS) to be successful at identifying genetic variants associated with many complex traits and diseases [1]. This is because GWASs are best suited for identifying variants for traits with polygenic architecture where many loci have only a small effect on the resulting phenotype [2]. There is now compelling evidence that many of these variants result in changes of RNA expression levels which could be the reason behind their association with the phenotype [3,4]. While GWASs have been largely successful, much of the heritability has not been explained by the currently discovered variants although as the sample sizes increase, the better powered GWASs will be in detecting variants with smaller effect sizes [5]. Nonetheless, even if GWASs yield no new variants associated with phenotypes, the landscape of the genetic association statistics from GWASs might still be informative in teaching us about the genetic architecture of the phenotype. In this dissertation, we demonstrated how one can leverage the results from GWASs to infer the role of rare and common variants to polygenic architecture.

MAJOR FINDINGS AND IMPLICATIONS

In chapter 2, we discussed experiments that analyze the common variant's effects on height at the tails of the height distribution. The findings are:

- Single SNP analysis shows that common variants have expected effects at the tails.
- The short individuals have less than expected number of common short alleles (alleles

that are shown to reduce stature).

- This effect is driven by the shortest individuals.
- This result is consistent with rare variants having moderate effects on short stature.

Given that the short individuals have less than expected number of common short alleles, there is a fair chance that there are many such rare variants that have moderate effects on short stature in the population. Studies have been performed to determine what some of these variants might be [6], but might still prove difficult given the lack of power as the allele frequencies of such variants are very low. There is also evidence that rare copy number variants (CNVs) in genomic regions can explain the short stature in some patients [7,8]. Therefore, one of the implications of our results is that if one wants to have a strategy for identifying rare variants that cause short stature in the population, the recruitment of individuals with short stature is critical. It would be better to first genotype individuals with short stature for their height-associated common variants and determine if these individuals have a deficit of height decreasing alleles. As the short stature individuals could be short because of rare variants and/or common variants, enriching for individuals with a deficit of common height decreasing alleles would enrich for individuals harboring rare variants. Our results also implicate the use of 'extreme' individuals for genetic studies, that such studies can be used to compliment our knowledge about the genetic architecture of the trait in question.

In chapter 3, we discussed a method to determine polygenic inheritance from low frequency variants by examining if there is an excess of risk conferring variants from summary statistics of association studies. The findings are,

- An excess of low frequency risk-increasing variants can be a signal of polygenic inheritance as measured by an increase in the risk to protective (R/P) ratio.

- This excess can be due to risk-increasing variants being more statistically powered than risk-decreasing variants with the same magnitude of effect.
- This excess can also be due to having more risk variants to begin with because of negative selection keeping risk variants at low frequencies.
- There is a higher probability for false positive associations to be risk variants if there are substantially more controls than cases.
- This excess can also be due to asymmetric population stratification because of badly designed GWAS.
- An analysis of some published GWAS summary statistics reveal significantly increased R/P ratios for schizophrenia, type 2 diabetes and obesity.
- Significant increased R/P ratios were observed for macroalbuminuria and end stage renal disease but not if these subtypes of diabetic nephropathy were combined into a single case group.

These findings suggest that one could simply test for an excess of risk conferring variants to determine if the low frequency variants contribute as a whole to disease risk. Methods to detect for a contribution of low frequency or rare genetic variants to disease risk are crucial as they can inform researchers whether pursuing the hypothesis would be a fruitful endeavor. While methods like GCTA [9] and polygene score [10] can be adapted to perform such analyses, examining for the excess of risk conferring variants provide an independent support for low frequency polygenic contributors to disease risk and requires only summary statistics without the need for primary genotype data. Besides having such a method, the findings suggest that most GWAS are designed to better discover low frequency variants that confer risk to disease. While this is useful for explaining disease etiology, it may be suboptimal for discovering genes that might be useful

as drug targets for treatment. This is because genes that have low frequency variants that confer protection to disease are best suited as drug targets assuming that the variants confer some loss of function effect on the gene. For example, low frequency loss of function variants in PCSK9 have been found to have a protective effect against coronary heart disease [11] and now has become a drug target for lowering LDL cholesterol [12]. Our results suggest that if GWAS were designed such that cases are individuals strongly protected against disease and controls are everyone else, that design will be better optimized to discover rare protective variants.

In chapter 4, we examine the extent of genetic contribution to sitting height ratio (SHR) by performing genome wide association studies on African and European Americans. The findings are,

- Degree of European admixed ancestry in African Americans strongly associated with sitting height ratio (SHR) suggests strong genetic contribution.
- GWAS in African Americans discover a locus associated with SHR.
- GWAS in European Americans discover 2 loci associated with SHR.
- More than expected height-associated variants show association with SHR as well.
- Some of these height-increasing allele decreases SHR while other increases SHR.

These results show that the difference of sitting height ratios (SHRs) between European and African Americans is genetic and that GWAS performed can reveal variants that are associated with SHR. However, the few variants discovered through GWAS do not explain the difference between European and African Americans suggesting that this difference is polygenic. As such, to fully uncover the full extent of such a difference, many more samples are required. The excess of known height-associated variants associated with SHR is also interesting. While sitting height is a component of total height, the sitting height ratio is not. Given that we corrected for total

height when performing the linear regression, the association statistic represents the change between the upper-body to lower-body ratio. As such, these height-associated SNPs can be grouped into 3 categories, i.e. the height-increasing allele does not alter SHR, the height-increasing allele decreases SHR and the height-increasing allele increases SHR. While we perhaps do not have enough height loci to investigate this, there is a strong hypothesis that the height-increasing alleles that increases SHR are probably in genes that function to increase spine length or that the alternate allele decreases femur or tibia length. On the other hand the height-increasing alleles that decrease SHR may perhaps be working to increase the length of the femur or tibia. The variants that have no effect on SHR may perhaps be regulating hormonal output. Perhaps examining these 3-classes of variants will shed more light on the biology of growth and the relevant developmental pathways involved.

Genome wide association studies (GWAS) can inform us about the genetic architecture of traits and diseases. We argue that one should not merely look at only the genome wide significant results from GWASs and ignore variants that are insignificant. By performing computational modeling on the full range of results, one would be able to infer the genetic architecture of the trait or disease and perhaps shed light on the biological mechanism responsible for producing the change in the phenotype.

FUTURE DIRECTIONS

In this section, we focus on the results from this dissertation to the understanding of disease etiology, the broader implications and potential future research directions and goals towards the broader aim of improving our understanding of genetic diseases as well as towards the discovery therapeutic strategies. It has been suggested that while there is a plethora of effort

for performing disease mapping through the use of GWAS, little has been discovered about the mechanisms of how these variants influence the disease pathology and even less so in terms of therapeutics. This trend might change in the future but there are several issues that might hinder the effort for understanding the mechanism of common variants to disease. First, as the effect sizes of these common variants are small, studying the variant's effect either in in-vitro systems or animal models may not be feasible as the magnitude of effect may be too small to be observed from the readout. Next, given the many number of such variants, it may be impractical to simultaneously study most of them. As such, rare-variants with large effects might be better suited for such follow up studies.

We have observed from the results from studying individuals from the extreme ends of the height distribution, individuals with short stature could potentially be short because of rare variants of moderate effects. These effect sizes could be large enough to register a read out from studying animal models. In fact, it has been shown that human alleles could be introduced into zebrafish causing these zebrafish to have similar phenotypes [13]. Therefore, given that rare variants with moderate effects are not likely to be discovered from GWAS as the SNP markers from GWAS are mainly common, new approaches for rare variant discovery are needed. Some have suggested and performed either whole-genome, whole-exome or exome-chip experiments as an effort to discover rare variants associated with diseases. Results from our work suggest that analyzing the GWAS results may be informative as to how likely such efforts would be fruitful. From our studies of individuals with short and tall stature, we found that there is a less than expected number of short alleles for the short individuals suggesting that they may have rare variants that moderately cause a decrease in height. If one were to sample from short individuals where their common variant profile predicts tall stature or above-average stature, these

individuals would more likely to harbor such rare-variants. Such rare-variants could lie in genes there are known in pathways that regulate growth or could be from genes without much known biological function or mechanism.

There could be several approaches to studying these genes to elucidate their unknown biological function or mechanism. One approach would be to introduce these variants via genome-engineering methods into some model organism. Since, if these organisms have homologous genes such that the human version of these genes is still functional, it is possible to replace the organism's endogenous gene with a human version harboring these variants. If the human allele of the gene causes a similar phenotype in the organism, in this case, short stature, it would be evidence that the allele is the causal variant responsible for the human phenotype and subsequent studies into the mechanism of action can be studied via the model organism. This strategy could be extended to phenotypes of other quantitative traits like body-mass-index (BMI), lipid levels and blood pressure. Although not widely done, modeling disease outcome using human alleles has been demonstrated to be successful in zebrafish [14]. The key would be to identify the rare-variants with large effects and we have shown that studying the phenotypic extremes can be more optimal for doing so.

While identifying rare variants with relatively larger effect sizes may be useful for understanding disease etiology, it may not be as useful for the development of therapeutics, in particular, the genes underlying these rare variants do not make good candidate drug targets. This is because these variants are usually deleterious variants and therefore targeting these genes is predictive of increasing risk to disease. Also, even if the variants are gain of function variants, targeting these genes would only work for individuals that have the risk allele, which would still be rare in the population. The truth is that most individuals are affected by complex diseases not

because of rare variants but by the cumulative effect of common variants in many genes together with environmental stimuli. Even when certain traits, like sitting height ratio, are highly differentiated between populations, the reason behind that differentiation is usually polygenic. Unless it becomes feasible to have drugs that target many genes concurrently where each target is only modestly affected, one would perhaps need a better solution for treating a polygenic disease.

One possibility would be to target genes where there are rare deleterious variants that have moderate protective effects. However, we have shown that for case-control association studies, there is more power to detect risk than protective variants. Therefore, in order to optimize power to detect protective variants, the "case" individuals used in a case-control association should be individuals that are protected against the disease. Finding such individuals however, is a challenge on its own as individuals who are protected against disease do not show up at a clinic. One possibility would be use a quantitative trait measurement that is a proxy for the disease. For example, one criterion for having type 2 diabetes is having fasting glucose levels above 125 mg/dL. If one were to be able to recruit individuals that have lower than normal fasting glucose levels as cases and controls to be anyone else, then that case-control study design would be more optimized for detecting such protective variants. Another approach would be to use unaffected individuals that have strong environmental exposure to getting the disease. For example, the use of healthy middle-aged adults that are obese but do not have type 2 diabetes could be used as cases. Since there is a high probability of getting type 2 diabetes if one is obese, non-diabetic obese individuals might harbor protective variants against type 2 diabetes. Perhaps, such a new paradigm for performing GWAS might be the way forward for optimizing the power to detect rare protective variants.

A POSTSCRIPT

We are at a point in time when research in human genetics for understanding complex diseases is in its critical moment. It was not too long ago where we do not have even a single gene or locus associated with any complex disease but now we have many, perhaps too many to even comprehend how it is possible to move forward. As genomic techniques improve and sequencing cost gets reduced, perhaps having a whole genome sequence for any single individual would be easily achieved. In the near future, having a genomic profile for any patient would be like measuring blood pressure today. It would be quick, easy and inexpensive. Therefore, the challenge of the future would be to determine how one could harness the genome's sequence of every patient to improve our understanding of disease mechanisms as well as to aid in the development of new therapeutics. It is incumbent on us scientist to make that a reality and I strongly believe that we will succeed. It is only a matter of time.

REFERENCES

1. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, et al. (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 9: 356–369. doi:10.1038/nrg2344.

2. Stranger BE, Stahl EA, Raj T (2011) Progress and Promise of Genome-Wide Association Studies for Human Complex Trait. Genetics 187: 367–383. doi:10.1534/genetics.110.120907.

3. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, et al. (2010) Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. PLoS Genet 6: e1000888. doi:10.1371/journal.pgen.1000888.

4. Nica AC, Montgomery SB, Dimas AS, Stranger BE, Beazley C, et al. (2010) Candidate Causal Regulatory Effects by Integration of Expression QTLs with Complex Trait Genetic Associations. PLoS Genet 6: e1000895. doi:10.1371/journal.pgen.1000895.

5. Visscher PM, Brown MA, McCarthy MI, Yang J (2012) Five Years of GWAS Discovery. Am

J Hum Genet 90: 7–24. doi:10.1016/j.ajhg.2011.11.029.

6. Wang SR, Carmichael H, Andrew SF, Miller TC, Moon JE, et al. (2013) Large-Scale Pooled Next-Generation Sequencing of 1077 Genes to Identify Genetic Causes of Short Stature. J Clin Endocrinol Metab 98: E1428–E1437. doi:10.1210/jc.2013-1534.

7. Dauber A, Yu Y, Turchin MC, Chiang CW, Meng YA, et al. (2011) Genome-wide Association of Copy-Number Variation Reveals an Association between Short Stature and the Presence of Low-Frequency Genomic Deletions. Am J Hum Genet 89: 751–759. doi:10.1016/j.ajhg.2011.10.014.

8. Zahnleiter D, Uebe S, Ekici AB, Hoyer J, Wiesener A, et al. (2013) Rare Copy Number Variants Are a Common Cause of Short Stature. PLoS Genet 9: e1003365. doi:10.1371/journal.pgen.1003365.

9. Yang J, Lee SH, Goddard ME, Visscher PM (2013) Genome-wide complex trait analysis (GCTA): methods, data analyses, and interpretations. Methods Mol Biol Clifton NJ 1019: 215–236. doi:10.1007/978-1-62703-447-0_9.

10. Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, et al. (2009) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460: 748–752. doi:10.1038/nature08185.

11. Cohen JC, Boerwinkle E, Mosley TH, Hobbs HH (2006) Sequence Variations in PCSK9, Low LDL, and Protection against Coronary Heart Disease. N Engl J Med 354: 1264–1272. doi:10.1056/NEJMoa054013.

12. Stein EA, Mellis S, Yancopoulos GD, Stahl N, Logan D, et al. (2012) Effect of a Monoclonal Antibody to PCSK9 on LDL Cholesterol. N Engl J Med 366: 1108–1118. doi:10.1056/NEJMoa1105803.

13. Khanna H, Davis EE, Murga-Zamalloa CA, Estrada-Cuzcano A, Lopez I, et al. (2009) A common allele in RPGRIP1L is a modifier of retinal degeneration in ciliopathies. Nat Genet 41: 739–745. doi:10.1038/ng.366.

14. Davis EE, Zhang Q, Liu Q, Diplas BH, Davey LM, et al. (2011) TTC21B contributes both causal and modifying alleles across the ciliopathy spectrum. Nat Genet 43: 189–196. doi:10.1038/ng.756.