Finding the Missing Heritability:

Gene Mapping Strategies for Complex Pedigrees

by

Kaanan Pradeep Shah

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Human Genetics)
in the University of Michigan
2013

Doctoral Committee:

Associate Professor Julie A. Douglas, Chair
Assistant Professor Anthony Antonellis
Professor David T. Burke
Professor Patricia A. Peyser

Dedication

To Marilyn Aunty, your fight was an inspiration

To my family for always listening to my stories

Acknowledgements

This work represents my research over the last 6 years of graduate school. None of this would have been possible without the endless support of my family, friends, mentors and colleagues.

First, and foremost I have to thank Julie Douglas. As my mentor, she has always allowed me the freedom to pursue my own scientific ideas, and freely shared her own with me. I cannot count the number of hours we spent discussing science in her office. I cannot thank her enough for her dedication to my research career and professional development. She has always been willing to have a "quick meeting" at a moment's notice—even if that turned into hours.

My family has always been supportive of my education, and encouraged me to go to graduate school. They were always willing to listen to me telling tales of my adventures in graduate school from thousands of miles away. Even though I was far away from my own parents, my Michigan and Ohio family have always treated me like a daughter. I could not have done this without the support and encouragement of my aunts and uncles.

I also have to thank my friends, for their encouragement and advice over the years. My friends, Valerie Schaibley and Michele Gornick, taught me that no problem was too big to be solved over a long lunch. Without the help of Kristen Purrington and Michele Gornick I would not have made it through all of my statistics classes. And to my Thursday girls, Valerie Schaibley, Ilea Swinehart, Heather McLauglin, Stephanie Coomes, Cheryl Jacobs-Smith, and Kadee Luderman-- thank you for giving me a life outside of science.

The research presented in this dissertation could not have been done without the help, guidance, and support of my colleagues at the University of Michigan. Thank you to all the current and former Douglas Lab members: Chris Plotts, Cris Van Hout, James MacDonald, and Al Levin. In addition to the small group in the Douglas Lab, I have had the pleasure of working with friendliest lab-neighbors: The Sing Lab. This research would not have been possible without the generous sharing of computing resources by Charlie Sing, Paul Kopek, Ken Weiss, and the entire, always smiling, Sing Lab. I also have to thank my dissertation committee, David Burke, Patrica Peyser, and Anthony Antonellis, for their guidance with my dissertation projects. Finally, being part of the Genome Sciences Training Program provided me with 3 years of funding and an intellectual community at Michigan that I will forever be grateful for.

Lastly, I have to thank the Amish community; none of this work would have been possible without their willingness to participate in genetics research.

Table of Contents

List of Tables

List of Figures

List of Appendices

Abstract

Geneticists have been working for decades to identify genetic factors that underlie variation in complex traits. Yet much of the variance attributed to additive genetic factors remains unaccounted for, the so-called “missing heritability problem.” Factors that may account for some of the missing heritability include the following: rare variants, structural variants, gene-gene interactions, and gene-environment interactions. In this dissertation, I evaluate the contribution of rare variants and gene-gene interactions to the missing heritability problem. Specifically, I develop and evaluate research strategies that take advantage of complex pedigree information. I apply these strategies to quantitative traits in the Old Order Amish, a population isolate in which most individuals are related through a single, complex pedigree.

In Chapter 2, I describe a new statistical test to identify quantitative traits that are likely influenced by rare variants of large effect. I found evidence for the presence of rare variants influencing a few traits, including (remarkably) one for which a null mutation was previously identified. In Chapter 3, I evaluate the performance of Markov-chain Monte Carlo (MCMC) algorithms for linkage analysis of quantitative traits with complex pedigrees and dense genetic maps. I discovered that current algorithms fail to converge, resulting in highly variable LOD (logarithm of the odds) scores between MCMC runs. Despite this variability, I found consistent evidence of linkage for one trait for which a locus of large effect was previously mapped. Together, results from chapters

2 and 3 imply that rare variants of large effect are unlikely to explain much of the missing heritability of these traits.

In Chapter 4, I consider that heritability might be overestimated rather than missing. To explore this possibility, I evaluate a new regression-based method to estimate heritability that is not inflated by gene-gene interactions. As suggested by Zuk *et al.* (2012), this method is ideal for use in population isolates but has not been investigated in realistic data settings. Unexpectedly, I discovered that the method produces biased estimates of the narrow-sense heritability, even for purely polygenic traits. Thus, caution should be exercised before using this method and attributing the missing heritability to gene-gene interactions.

Chapter 1

Introduction

1.1 Gene Mapping and the Missing Heritability Problem

The major goal of human geneticists is to understand how genetic variation contributes to phenotypic variation in the population. To this end, geneticists have been mapping genes for traits and diseases in humans and model organisms since well before the sequencing of the first human genome. Ultimately, these studies can lend insights into human health and disease by learning about the genetic basis for traits and diseases. For example, gene mapping studies have uncovered a substantial overlap in the genetic factors that are associated with risk of many autoimmune and inflammatory diseases [1]. The links between many of these diseases were not previously appreciated.

Advances in microarray technologies since the sequencing of the first human genome have facilitated a dramatic increase in the rate of discovery of genes for both Mendelian diseases and complex traits [2-4]. Simultaneously genotyping millions of single nucleotide polymorphisms (SNPs) cheaply and efficiently using microarrays has allowed researchers to assess genetic variation genome-wide to identify variants associated with quantitative traits and complex diseases. These genome-wide association studies (GWAS's) are now a common approach to identifying common variants (minor allele frequency, MAF > 5%) associated with complex traits without prior biological candidates. The underlying hypothesis for these GWAS's is that risk for common diseases will be associated with variants that are common the population [5, 6]. Thus, by testing common variants for association with common diseases and quantitative traits, GWAS's have identified >9,000 loci associated with >700 traits [7].

Despite this success, perhaps one of the biggest lessons from GWAS's over the last 8 years has been that the genetic architecture of complex diseases is far more complex than we anticipated. In other words, there are likely many more variants spanning the allele frequency spectrum that influence trait variation or disease risk in more complicated ways than previously anticipated. Even with sample sizes on the order of 100,000 individuals, much of the genetic component, or heritability of many complex traits remains unaccounted for. For example, the nearly 180 common loci associated with height collectively explain only 12% of the heritability of height [4]. In fact, some studies suggest that there may be 1,000s of loci for height distributed uniformly across the genome [8]. Thus, like most complex traits, height is extremely polygenic, influenced by variation at numerous loci, most with individually weak effects.

The inability to account for much of the heritability of complex traits by common variants has been deemed the "missing heritability problem" [9]. Many have hypothesized about where the missing heritability might be found: rare variants, structural variants, gene-gene interactions, and gene-environment interactions are all possible sources of genetic variation that could further explain the heritability of complex traits [10, 11]. It is likely that a combination of these factors will play a role in complex traits; however, the relative importance of each class of variation will likely be trait specific. Different study designs and strategies will be necessary to identify and associate variants in these different classes with complex traits. Therefore prior insights into the contribution of different types of genetic variation to specific complex traits can inform the design of studies and enable future gene mapping efforts.

A current challenge in the field is to evaluate the potential contribution of different types of genetic variation to quantitative traits. I evaluate the potential contribution of rare variants of large effect to the missing heritability in Chapters 2 and 3 of this dissertation. In Chapter 4, I set out to evaluate the contribution of gene-gene interactions to the missing heritability.

1.1.1 Rare Variants and the Missing Heritability Problem

Nearly a decade of GWAS's has clearly identified a role for common variants in complex traits. However, these variants alone have not fully accounted for the heritability of complex traits. Therefore, scientists have put forth alternative hypotheses to the common disease, common variant hypothesis that was largely the basis for GWAS's [5, 6]. One such hypothesis, the common disease, rare variant hypothesis suggests that rare variants may account for much of the missing heritability [12]. Rare variants, defined here as genetic variants with a MAF <5%, tend to have a larger impact when they are associated with phenotypic variation [10]. In the context of quantitative traits, this means that carriers of a rare, trait-associated variant may have a trait value substantially displaced from the population mean. By definition, carriers are rare in the general population. However, studying individuals with extreme phenotypes and/or large families can enrich for many more copies of the rare variant, and therefore making them easier to discover in gene mapping studies [13, 14].

While the potential for rare variants to account for the heritability of complex traits has been appreciated for over a decade, it was not until recent advances in DNA sequencing technologies that their role of could be assessed. While the cost of next-generation sequencing is decreasing [15, 16], it is not yet practical to sequence all individuals in a large study to identify rare variants. Therefore, the goal of the methods in Chapter 2 and 3 is to prioritize traits, individuals, and loci for sequencing studies to identify rare variants. The approaches developed in these two chapters may facilitate the discovery of rare trait-associated variants. Therefore, by applying these methods to numerous quantitative traits, I evaluate the potential for rare variants to account for more of the missing heritability.

For quantitative traits, one approach to discovering rare, trait-associated variants is to sequence individuals with extreme phenotypes. For this approach to be cost effective, it is important to carefully select the trait and individuals to sequence. In Chapter 2 of this dissertation, I describe a novel statistical test to prioritize traits and individuals in family-based studies for sequencing by identifying traits likely to be influenced by rare variants of large effect using only phenotype and pedigree information. Another, classical approach to identifying loci harboring rare, trait-associated variants is genome-wide linkage analysis. In Chapter 3, I evaluate a Markov-chain Monte Carlo (MCMC) method for linkage analysis with complex pedigrees and dense genetic marker maps. This MCMC approach has not previously been evaluated with pedigrees as large as our Amish pedigrees (bitsize up to 100) or a dense genetic marker map (1 SNP/cM).

1.1.2 Gene-Gene Interactions and the Missing Heritability Problem

The non-additive effects at alleles at multiple loci, or gene-gene interactions, may also contribute to the missing heritability. In fact, a number of models have been put forth that demonstrate how gene-gene interactions may explain some of the missing heritability [17-21]. While these theoretical models are plausible, detecting interacting loci in humans has been extremely challenging. Genome-wide interaction studies are incredibly underpowered due to the multiple testing burden of testing millions combinations of variants across the genome [22, 23]. For example, to conduct all pair-wise tests between ~2.5million common SNPs genome-wide requires ~10^{12} tests, therefore the multiple testing is substantially worse than a single marker genome-wide association study ($2.5\text{x}10^{6}$ tests). Despite the challenges and results of interaction studies in humans, model organism studies have shown that the effects of gene-gene interactions can be quite significant, accounting for up to half of trait heritabilities in some instances [24-26].

As an alternative to detecting specific interacting loci, I set out to assess the potential contribution of gene-gene interactions to the missing heritability. I do this by comparing estimates of trait heritabilities with and without the influence of interactions. I used a newly proposed regression-based estimator of heritability that is not confounded by the effects of interactions [21]. In Chapter 4 of this dissertation I apply this regression-based estimator and evaluate the statistical properties of the estimator via simulations.

1.2 The Old Order Amish

The motivation and data for the studies in this dissertation come from our on-going efforts to map genes related to mammographic density and cardiovascular disease risk factors in our studies in the Old Order Amish population. The unique population, traits, and studies are described below.

1.2.1 Population History

The Old Order Amish originated as followers of Jakob Ammann, a Mennonite bishop who split from the church in 1693. After splitting from the Mennonites, the Amish migrated to Switzerland and Eastern France. In the 1700's, the Amish began migrating to the United States to escape religious persecution. The first wave of ~500 Amish immigrants, known today as the Old Order Amish, founded the settlement in Lancaster County, PA [27-29]. Throughout this dissertation, 'Amish' refers specifically to the Old Order Amish population.

Today, the Amish remain a closed population. There is little, if any, influx of outsiders into the population because the Amish do not believe in spreading their faith [30]. Furthermore, 10% of Amish children leave the community each generation [30]. Despite this, the high birth rate among the Amish has resulted in constant population growth [27, 31]. An Amish woman has, on average, ~7 children [32]. Thus, from the original ~500 founders, the Amish have grown to 30,000-50,000 in and around Lancaster County, PA [28, 33].

1.2.2 Genetics Research within the Amish Population

It has long been recognized that the Amish population is particularly well suited for genetic studies. The Amish are a genetically closed population with extensive genealogical records [34, 35], a high standard of living, and a willingness to participate in medical genetics research [29]. The first genetic studies in this population began in 1962 when Victor McKusick and John Hostetler became interested in recessive metabolic conditions in the population [36]. Since then, the Amish have contributed to advances in our understanding of numerous Mendelian disorders (reviewed in [33, 37]) and complex traits [for examples, [38-52]).

The closed population and demographic history of the Amish have resulted in elevated inbreeding within the population. On average, any two Amish individuals are more closely related than 2nd cousins but less than 1st cousins, once removed. The bottleneck at the time of population founding suggests that the spectrum of rare genetic variation will be compressed. Specifically, we expect fewer rare variants segregating and that these variants may be in higher copy number in the Amish than more cosmopolitan European populations. Based on gene dropping simulations conditional on the pedigree structure from our study of mammographic density (described in Section 1.2.2.1), we expect ~17% of variants with a minor allele frequency (MAF) ≤ 5% in the population founders to have a MAF > 5% in the current day population. Therefore, the Amish may be ideally suited for gene mapping studies to identify rare, trait-associated variants. In fact, to-date there are a few examples of rare variants associated with quantitative traits that have increased in frequency in the Amish [39, 45, 48].

The work in this dissertation is motivated by our gene mapping efforts in the Amish population. Specifically, I will use data from our study of mammographic density and the Heredity and Phenotype Intervention (HAPI) Heart study to apply the methods discussed in the subsequent chapters. Both of these are studies of complex quantitative traits related to disease risk. Details about each study are given below.

1.2.2.1 Family-Based Study of Mammographic Density

The primary goal of our study was to map genes related to mammographic density. Mammographic density is a well-established, heritable risk factor for breast cancer. This area appears light on a mammogram and likely represents the ductal, epithelia, and connective tissue of the breast. Women with greater than 75% density have a 4-6 fold increased risk of breast cancer compared to women with less than 25% density [53]. Starting in 2005, we began recruiting healthy Amish women to participate in our study of mammographic density with the goal of identifying genetic factors that are associated with density, and potentially to breast cancer risk more generally.

In total, we recruited 1,521 women over 5 years to participate in our study. These women represent nearly half of the current census population of Amish women aged 40-80 years old in Lancaster County, PA. Initially, women ≥ 40 years old were only eligible for our study if they had a living sister ≥ 40 years old. We later dropped this criterion to recruit all women ≥ 40 years old, regardless of whether they had an eligible sister. Recruitment was done primarily through word-of-moth and door-to-door interviews. Therefore, while we sampled nearly have of the eligible population, the sample is likely not a random sample from the population. Specific inclusion and exclusion criteria are described elsewhere [54].

The final sample included complete phenotype information on 1,481 women. In addition to mammographic density, we collected and analyzed a number of other quantitative traits related to breast cancer risk including measures of body size, reproductive and menstrual traits, and several serum hormones and growth factors. Details about the traits are in Appendix 1.The women in our study were also genotyped at ~2.5 million SNPs genome-wide. After quality control of our genome-wide SNP data, we had 1,472 women with phenotype and genotype information.

All of the women in our study were connected into a single 13-generation pedigree using the extensive genealogical records from the Anabaptist Genealogical Database (AGDB) [34, 35]. Our study sample includes many closely related women, specifically 274 parent-offspring pairs and 1,254 sibling pairs. The women in our study are divided between pre- and post-menopausal women (728 and 753, respectively). As is typical of the Amish, these women were highly parous (91%) and rarely used exogenous hormones (10%).

1.2.2.2 Heredity and Phenotype Intervention Heart Study

The goal of the Heredity and Phenotype Intervention (HAPI) Heart study was to identify genetic and environmental factors that are related to cardiovascular disease risk. This study, started in 2002, was designed and conducted in collaboration with colleagues at the University of Michigan and the University of Maryland. In addition to measuring baseline cardiovascular risk, the study participants also participated in a series of intervention studies. The interventions were designed as short term challenges that mimic long term explosions that affect cardiovascular disease risk.

Amish men and women over the age of 20 and generally healthy enough to participate in the interventions were eligible for the study. Participants were initially recruited through participation in other studies, word-of-moth, community mailings, and local physicians. Eligible family members of participants were also requested to participate. Thus the sample was enriched for close relatives. Specific eligibility criteria for the study and each intervention are detailed elsewhere [55]. In total, 1,003 individuals identified to participate in the study, but 123 individuals refused to participate or did not meet all of the general eligibility criteria. Therefore, in total, the final sample included 868 individuals.

The men and women in the HAPI Heart study were phenotyped for a large number of baseline cardiovascular risk factors and participated in a number of interventions, including a high-fat diet challenge, a cold pressor stress test (CPT), a dietary salt intervention, and a low-dose aspirin therapy treatment. Multiple measures of cardiovascular response and health were measured during and after the intervention studies. Details about the baseline phenotyping methods and intervention studies are described in detail elsewhere [55]. For the purposes of this dissertation, I focused only on baseline quantitative traits related to cardiovascular disease risk, detailed in Appendix 3. The study participants were also genotyped at ~500,000 SNPs genome-wide.

All of the individuals in the study can be connected into a single extended pedigree. Consistent with the recruiting strategy, there were any close relative pairs including 314 parent-offspring pairs and 592 sibling pairs. The sample included 460 men and 408 women. Also consistent with the eligibility criteria, the study participants were generally healthy, with < 1% having diabetes and ~1% on lipid lowering medications prior to the study [55].

Chapter 2

A Method to Prioritize Quantitative Traits and Individuals for Sequencing in Family-Based Studies[1]

Owing to recent advances in DNA sequencing, it is now technically feasible to evaluate the contribution of rare variation to complex traits and diseases. However, it is still cost prohibitive to sequence the whole genome (or exome) of all individuals in each study. For quantitative traits, one strategy to reduce cost is to sequence individuals in the tails of the trait distribution. However, the next challenge becomes how to prioritize traits and individuals for sequencing since individuals are often characterized for dozens of medically relevant traits. In this article, we describe a new method, the Rare Variant Kinship Test (RVKT), which leverages relationship information in family-based studies to identify quantitative traits that are likely influenced by rare variants.

Conditional on nuclear families and extended pedigrees, we evaluate the power of the RVKT via simulation. Not unexpectedly, the power of our method depends strongly on effect size, and to a lesser extent, on the frequency of the rare variant and the number and type of relationships in the sample. As an illustration, we also apply our method to data from two genetic studies in the Old Order Amish, a founder population with extensive genealogical records. Remarkably, we implicate the presence of a rare

[1] Shah KP, Douglas JA (2013) A Method to Prioritize Quantitative Traits and Individuals for Sequencing in Family-Based Studies. PLoS ONE 8(4): e62545. doi:10.1371/journal.pone.0062545

variant that lowers fasting triglyceride levels in the Heredity and Phenotype Intervention (HAPI) Heart study (p=0.044), consistent with the presence of a previously identified null mutation in the *APOC3* gene that lowers fasting triglyceride levels in HAPI Heart study participants.

2.1 Introduction

The genetic architecture of most complex traits and diseases is poorly understood. Indeed, genome-wide association studies (GWAS's) have identified hundreds of loci with relatively weak effects on complex traits and diseases, leaving much of their heritability unaccounted for [10]. This is expected (in part) since the genotyping technology used in these studies captures primarily common sequence variation, namely, single nucleotide polymorphisms (SNPs) with minor allele frequencies (MAFs) of at least 5%. Rare variants (MAF<5%), which are poorly captured by standard GWA arrays [56], may have larger effect sizes than common variants and may make an important contribution to complex traits and diseases. In fact, results from large-scale sequencing studies (n>10,000) suggest a much higher load of rare variants than was previously appreciated and may bear on the heritability unexplained by GWAS [57, 58].

Recent advances in DNA sequencing technology have dramatically increased the capacity to discover rare variants. However, it is still cost prohibitive to sequence whole genomes (or even whole exomes) on the scale of a GWAS, e.g., by sequencing all study participants. For studies of quantitative traits, one strategy to reduce cost is to sequence individuals with extreme phenotypes. Simulation studies [59, 60] and empirical studies of candidate genes suggest that this is a powerful approach for identifying rare trait-associated alleles. For example, this approach has been successfully used to identify rare variants in candidate genes associated with body mass index (BMI) [61], high-density lipoprotein (HDL) [62], low-density lipoprotein (LDL) [63, 64], and sterol absorption [63].

The power of extreme-trait sequencing or selective genotyping, originally introduced by Lander and Botstein [65], derives from the fact that rare trait-influencing alleles with modest to large effects will be enriched in frequency in the upper or lower tail of the trait distribution. The success of this strategy, however, depends (in part) on the careful selection of traits and individuals to sequence. In theory, the most powerful approach is to select and sequence the most extreme individuals from each tail of the trait distribution. In practice, however, power may be lost by sequencing too few or too many individuals or by choosing a suboptimal trait. To optimize the selection of traits and individuals for an extreme-trait sequencing study, we develop a new statistical test, the Rare Variant Kinship Test (RVKT). Our test is designed for use in family-based studies in which individuals have already been phenotyped – but not necessarily genotyped – for dozens of quantitative traits relevant to human health and disease.

Briefly, the RVKT leverages the relatedness of individuals in family-based studies to identify quantitative traits that are most likely to be influenced by rare variants. The premise of our test is that rare variants with at least modest effects will be enriched in the tails of the trait distribution and preferentially carried by closely related individuals. Unlike complex segregation analysis, which attempts to identify a particular mode of inheritance, our approach makes few assumptions about the trait architecture. We assess the power of our test via simulation and apply it to dozens of quantitative traits from two of our studies in the Old Order Amish population.

2.2 Methods

2.2.1 Ethics Statement

All human subject research was previously reviewed and approved by the Institutional Review Boards at the University of Michigan and the University of Maryland. Written informed consent was obtained from all study participants.

2.2.2 Overview

Here we describe the RVKT, simulations to assess power, and applications to two family-based studies. The RVKT requires a sample of families with pedigree and phenotype data and assumes that each of the quantitative traits to be tested has a narrow-sense heritability that is significantly different from zero. The null hypothesis of the RVKT is that a given trait is purely polygenic, meaning influenced by multiple additive, independent loci of small effect. Under the null hypothesis, individuals in the tail of the trait distribution carry trait-influencing alleles at many loci. The alternative hypothesis of the RVKT is that at least one locus of modest to large effect influences the trait, and accordingly, that the trait-associated allele(s) is necessarily rare (the rare variant). Under the alternative hypothesis, individuals in the tail of the trait distribution should preferentially carry the rare variant and thus may be more closely related when measured against the null hypothesis.

2.2.3 The Rare Variant Kinship Test

For each trait, we define and calculate the RVKT statistic as the mean of the pair-wise kinship coefficients between individuals in the tail of the quantitative trait distribution. Tail membership is determined by ordering individual trait values. Conditional on the pedigrees in the sample, the kinship coefficient between two individuals is the probability that a randomly chosen allele from one individual and a randomly chosen allele from the other individual at an autosomal locus are inherited identical by descent from a recent, common ancestor. We calculate pair-wise kinship coefficients using the matrix method described by Lange [66] and implemented in MENDEL version 10.0.0. Since the kinship coefficient depends only on the structure of the pedigree connecting a pair of individuals, the RVKT requires pedigree data but no genetic data. Thus, it can be applied before carrying out expensive genotyping or sequencing experiments.

To assess statistical significance, we compare the observed RVKT statistic for each trait to its expected distribution under a purely polygenic model (the null hypothesis) (described below). Under the alternative hypothesis, the observed RVKT statistic may exceed its expected value, meaning individuals in the tail of the trait distribution may be more closely related than expected under the null hypothesis. Thus, we use a one-sided test. Because the genetic architecture of each trait is unknown, we conduct the RVKT for both tails of the trait distribution (upper and lower) and multiple tail sizes. Tail size is the proportion of individuals in the tail of the trait distribution. We then select the RVKT statistic with the minimum p-value in each tail (p_{min}).

The expected distribution of the RVKT statistic depends on the actual pedigrees and the narrow-sense heritability of the trait. Thus, we use simulation to generate an empirical null distribution for each trait. Specifically, using MORGAN version 3.0 [67] we simulate 10,000 replicates of a purely polygenic trait with heritability equal to the narrow-sense heritability estimated from the observed data. Simulations are done conditional on the observed pedigrees. We calculate the RVKT statistic for each replicate using the same tail sizes tested in the observed data. The resulting RVKT statistics form an empirical null distribution for each trait and tail size. From this distribution, we determine a rejection region based on the prescribed size of the test (false-positive rate).

2.2.4 Assessment of the Test by Computer Simulation

To evaluate the power of our test, we conducted gene dropping simulations conditional on our Amish pedigrees (described below), and for comparison, four-person nuclear families (two parents and two offspring) with sample sizes corresponding to our Amish studies. Specifically, we simulated a single additive, bi-allelic locus with a trait-influencing allele frequency of 0.5, 1, 2, 3, or 4% (the rare variant) that accounted for 2, 5, 10, 20, or 30% of the total trait variance. In each simulation, we assumed that multiple additive, independent genetic factors, including the rare variant, accounted for 40, 60, or 80% of the trait variance (the narrow-sense heritability). For each set of parameters, we simulated 1,000 replicates using MORGAN version 3.0 and tested tail sizes of 1, 2, 4, 6, and 8%. For each tail size, power was calculated as the proportion of replicates for which the RVKT statistic equaled or exceeded the 95th percentile of the empirical null distribution, i.e., using a significance level of 0.05. We generated a single null distribution (as described above) for narrow-sense heritabilities of 40, 60, and 80% and repeatedly compared each replicate under the alternative hypothesis to this distribution.

A subset of the simulations above were conducted on pedigree structures connecting 1,481 women from our genetic study of mammographic density [54] and 868 men and women from the Heritability and Phenotype Intervention (HAPI) Heart study, a genetic and environmental study of cardiovascular risk factors [55]. Individuals in both studies were from the Old Order Amish population of Lancaster County, Pennsylvania. Using the extensive genealogical information available from the Anabaptist Genealogical Database [34, 35], we were able to connect subjects within each study into a single, 13-generation pedigree. Table 2.1 gives the number and types of pair-wise relationships after merging in only two generations from the complete pedigree, i.e., by merging in the parents and grandparents of all study subjects, and trimming the resulting pedigrees using PedCut [68] with a maximum bit size of 100. To assess the impact of pedigree complexity on power, we repeated simulations using the complete 13-generation pedigree.

2.2.5 Application of the Test to Empirical Data

We applied the RVKT to dozens of quantitative traits from the two genetic studies described above, with the goal of prioritizing traits and individuals for extreme-trait sequencing. Specifically, we applied the RVKT to 35 quantitative traits from our study of mammographic density (n=1,481), including absolute measures of the dense and non-dense area of the breast, percent mammographic density, total breast size, measures of body size, reproductive and menstrual traits, and several serum hormones and growth factors (Appendix 1). These traits, all of which are heritable, are of interest because of their associations with breast cancer risk. Prior to testing, we transformed each trait to approximate univariate normality, when necessary, and adjusted for age and menopausal status. For the hormones and growth factors, we carried out menopausal-specific analyses using batch-specific z-scores adjusted for age.

We also applied the RVKT to 37 quantitative traits from the HAPI Heart study (n=868), including measures of body size, fasting lipid levels, and measures of vascular health and arterial stiffness (Appendix 3). These traits, which are also heritable, are of interest because of their associations with cardiovascular disease. Prior to testing, we transformed each trait to approximate univariate normality, when necessary, and adjusted for age and sex.

As in our simulations, we tested tail sizes of 1, 2, 4, 6, and 8%, corresponding to 15, 30, 59, 89, and 118 subjects from our study of mammographic density and 9, 17, 34, 51, and 68 subjects from the HAPI Heart study. We then selected the RVKT statistic with the minimum p-value (p_{min}) in the upper and lower tail of each trait distribution. To control for multiple testing of traits, some of which may be correlated, we calculated the effective number of tests using the method described by Li and Ji [69] and applied a Bonferroni correction to p_{min}, denoted $p_{min,\ corrected}$.

2.3 Results

2.3.1 Size of the Rare Variant Kinship Test

To assess power, we used simulation to generate the null distribution of the RVKT statistic and determine the size (false-positive rate) of the test. As expected, the cumulative distribution function (CDF) was discrete. However, it became increasingly discrete as the number and types of relative pairs in the tail or sample decreased. For example, Figure 2.1 shows the top quintile of the CDF for a purely polygenic trait with a narrow-sense heritability of 40% and two sample structures: four-person nuclear families (n=1,484) and trimmed Amish pedigrees from our study of mammographic density (n=1,481). In the top quintile, the RVKT statistic assumed 36 values for the trimmed Amish pedigrees (Figure 2.1C) but only 6 values for nuclear families (Figure 2.1A), assuming a tail size of 1%. None of these values, however, coincided with the 95th percentile of the CDF. Thus, in our power calculations below, we selected a rejection region having size as close as possible to 0.05, without exceeding 0.05, in order to maintain a significance level of 0.05.

2.3.2 Power of the Rare Variant Kinship Test

Under the alternative hypothesis, power was generally maximized when the tail size matched the expected carrier frequency of the rare variant (data not shown). In other words, if q denotes the frequency of the rare variant, power was greatest for a tail size of $2(1-q)q+q^2$. Thus, we report results below for tail sizes that maximized power.

As expected, power of the RVKT increased as the effect size increased, meaning as the rare variant accounted for an increasing proportion of the trait variance. For example, using the trimmed pedigrees from our study of mammographic density (Table 2.1) and assuming a narrow-sense heritability of 40% and a rare variant frequency of 2%, power ranged from approximately 6 to 87% for effect sizes of 2 to 30%, respectively (Figure 2.2A). Similarly, based on the trimmed pedigrees from the HAPI Heart study and the same set of parameters, power ranged from approximately 7 to 61% (Figure 2.2B). As expected, power also increased as the sample size increased (Figure 2.2) and/or the rare variant frequency decreased (Figure 2.3). Power did not change much as the narrow-sense heritability of the trait varied from 40 to 80% (Figure 2.4).

Power degraded substantially as pedigree complexity increased, meaning as the number and types of distantly related pairs in a sample increased. For example, consider a sample of 1,481 individuals, a narrow-sense heritability of 40%, and a rare variant with frequency 2% and effect size 20%. Under these parameters, power decreased from 64% for the trimmed Amish pedigrees to 26% for the complete 13-generation pedigree (grey versus black bars in Figure 2.2A). In fact, power was actually higher with four-person nuclear families (n=1,484 individuals; white bars in Figure 2.2A) than with our trimmed Amish pedigrees (75% versus 64%). Pedigree complexity also reduced power for smaller effect sizes (Figure 2.2) and for pedigree structures in the HAPI Heart study (Figure 2.2B).

2.3.3 Application of the Rare Variant Kinship Test

After evaluating the power of the RVKT via simulation, we applied our test to dozens of quantitative traits from our two Amish studies. Figure 2.5 and Figure 2.6 summarize RVKT p-values (p_{min}) from our study of mammographic density and the HAPI

Heart study, respectively. The RVKT statistic was nominally significant for 8 of the 35 traits in the density study ($p_{min} \leq 0.05$). After correcting for multiple testing (26 effective tests), the RVKT remained significant for 3 of the 8 traits, including free estradiol and prolactin in pre-menopausal women and estradiol in post-menopausal women ($p_{min, corrected} \leq 0.05$). Similarly, in the HAPI Heart study, the RVKT was nominally significant for 14 of the 37 traits, one of which, namely, fasting triglyceride levels, remained significant after correcting for 22 effective tests ($p_{min, corrected}=0.044$).

In total, after multiple test correction, the RVKT statistic was significant for 4 of 72 quantitative traits across our two genetic studies. Table 2.2 gives results for each of these 4 traits for the tail size corresponding to the smallest empirical p-value (p_{min}). For example, in pre-menopausal women from our study of mammographic density, p_{min}, which corresponded to a tail size of 2% (14 of 728 women), was 0.0004 for prolactin. These 14 women had the lowest batch-standardized and age-adjusted levels of prolactin and a mean pair-wise kinship coefficient of 0.080 compared to an expected value of 0.068 under a purely polygenic model (approximate 95% confidence interval of 0.067 to 0.070). For each of the other 3 traits, the RVKT statistic was also significant when testing the lower but not upper tail of the trait distribution.

2.4 Discussion

The advantage of using the RVKT to prioritize traits and individuals for sequencing in family-based studies is best illustrated by results from the HAPI Heart study. In testing 37 quantitative traits, many of which are established risk factors for cardiovascular disease, we found significant evidence of excess relatedness between individuals in the lower tail of the distribution for fasting triglycerides. For tail sizes of 1 to 8%, the mean pair-wise kinship coefficient ranged from 0.114 to 0.020, respectively, and was significantly different from the kinship coefficient expected under a purely polygenic model of trait architecture ($p \leq 0.05$). Although differences between significance levels were not pronounced for different tail sizes, the significance of the RVKT was minimized for the 17 individuals with the lowest age- and sex-adjusted triglyceride levels, or equivalently, for a tail size 0of approximately 2%.

Remarkably, Pollin *et al.* [45] previously identified a null mutation in the *APOC3* gene (R19X, rs76353203) with a frequency of 0.024 that lowers fasting triglyceride levels in HAPI Heart study participants. This mutation was discovered because it was tagged by another SNP (rs10892151, MAF = 0.028) in the context of a GWAS ($p=4.1x10^{-13}$, $r^2=0.85$ between rs76353203 and rs10892151). Had we sequenced the 17 individuals in the lower tail of the age- and sex-adjusted triglyceride distribution, we would have discovered *APOC3* R19X since 7 of these individuals were mutation carriers, an 8-fold enrichment compared to the ~5% of individuals who were carriers in the overall sample. Notably, none of the 17 individuals in the upper tail of the distribution carried the mutation.

As expected, the power of the RVKT was low for small to modest effect sizes. In fact, the power of our test to implicate the presence and influence of *APOC3* R19X on fasting triglycerides in the HAPI Heart study was less than 25%. As such, it cannot be used to exclude the presence of rare trait-associated alleles, unless these alleles account for a large proportion of the phenotypic variance. However, when multiple medically relevant quantitative traits are available, the RVKT may be a valuable starting point for prioritizing traits and individuals for sequencing. For example, even though *APOC3* R19X carriers in the HAPI Heart study had cardio-protective profiles for several lipids, including higher HDL and lower LDL cholesterol and lower triglyceride levels, Pollin *et al.* [45] discovered R19X because its tag SNP had an exclusive genome-wide significant association with fasting triglyceride levels. Consistent with their findings, we singled out fasting triglycerides – out of 37 traits – as the basis for an extreme-trait sequencing study by applying the RVKT.

The power of the RVKT is heavily influenced by the number and types of relationships in a sample. Specifically, the power of the RVKT increases as the number of closely related pairs increases. In contrast, power is lost as the number of distantly related pairs multiplies. For instance, in our simulations (Figures Figure 2.2, Figure 2.3, and Figure 2.4), power was actually greater with the trimmed Amish pedigrees than with the complete 13-generation pedigree, with differences as great as 20-30% for large effect sizes. To understand why, it's helpful to consider the impact of trimming on the mean kinship coefficient under the null and alternative hypotheses. Under both hypotheses, trimming decreases the mean since individuals who are distantly related, say third cousins, appear to be unrelated. However, it does so to a lesser extent under the alternative hypothesis. This is because the mean under the alternative is dominated by closely related pairs, which are maintained regardless of trimming. As a result, the difference between the mean kinship coefficient under the null and alternative hypotheses is larger – and in turn, power is greater – with trimming than without.

Although trimming increases power, the RVKT may actually be conservative when pedigrees are too simple. In fact, it may be impossible to choose a rejection region from the empirical null distribution of the RVKT statistic such that the size of the test does not exceed the significance level. For example, consider a single pair of siblings from each of 741 families (n=1,482). To obtain the null distribution, we simulated 1,000 replicates of a purely polygenic trait with a narrow-sense heritability of 40%. However, when we tested a 1% tail size, we obtained only three values of the RVKT statistic (data not shown). The largest value occurred 17 times; therefore, the smallest possible test size was 0.017. In other words, it would have been impossible to conduct a 0.01 level test. This problem was especially pronounced for modest sample sizes and small tail sizes due to discontinuities in the empirical null distribution of the mean pair-wise kinship coefficient (data not shown).

An implicit assumption of the RVKT is that – within each family – a specific allele at the same locus has an effect on the trait of interest. In other words, the power of the test depends on the extent of allelic and locus homogeneity within each family but does not require homogeneity between families. For example, if multiple rare variants influence a trait, then phenotypically extreme individuals from the same family are more likely to share the same trait-influencing alleles than phenotypically extreme individuals from different families. Thus, the RVKT statistic may still exceed its expected value since individuals in the tail of the trait distribution may be more closely related than expected under the null hypothesis. From this perspective, isolates like the Amish are an ideal population in which to apply the RVKT and carry out extreme-trait sequencing since many copies of the same rare trait-associated allele are likely to be segregating within a family due to a combination of founder effect and genetic drift.

In our simulations, we considered rare variant frequencies ranging from 0.5 to 4%. We did so for two reasons. First, the Old Order Amish population of Lancaster County, PA derives from a small number of European ancestors (~500) who immigrated nearly 250 years ago and has since increased in size to approximately 45,000 individuals (census size) [28]. Thus, many alleles that were initially rare or private in the ancestral population, e.g., MAF<0.5% in HapMap or 1KG projects, have either been eliminated from the Amish or increased in frequency due to founder effect and/or genetic drift. Second, even in the presence of the allelic heterogeneity typical of non-founder populations, the aggregate trait-associated allele frequency at a single locus may still be greater than 0.5% and thus potentially amenable to detection by the RVKT.

Prioritizing traits and individuals for sequencing using the RVKT requires only pedigree and phenotype data and thus can be done before carrying out costly sequencing experiments. This process, however, requires accurate pedigree and phenotype data. Likewise, it is important to consider the impact of adjusting for covariates or stratifying the analysis by subgroups before identifying individuals with extreme trait values. For example, in our study of mammographic density, we found significant evidence for the presence of rare variants influencing the dense area of the breast in post- but not pre-menopausal women. Specifically, after adjustment for age, p_{min} for the RVKT was 0.018 and 0.031 for the lower and upper tails, respectively

(Appendix 4). These results suggest the presence of at least one variant that lowers density and another variant that increases density in post-menopausal women.

We developed the RVKT to inform the selection of traits and individuals for sequencing and rare variant discovery. Predictably, the power of our test depended – above all – on the effect size of the rare variant. Indeed, it was underpowered to detect rare variants unless those variants had large effects. However, our analysis of over 70 quantitative traits from our Amish studies suggests that the results may still be informative to prioritize sequencing efforts.

Table 2.1: Pair-wise relationships between individuals from our study of mammographic density (n=1,481) and the HAPI Heart study (n=868) after pedigree trimming

	Number of Pairs	
Relationship Pair	**Mammographic density study**	**HAPI Heart study**
Parent-offspring	276	314
Siblings	1,254	592
Grandparent-grandchild	0	21
Avuncular	1,125	732
1st cousins	4,676	1,379
1st cousins, once removed	2,993	1,508
2nd cousins	1,345	905
Other	871	807

Note – Pedigree trimming yielded 177 families with 1-44 study participants per family (average of 8) in our study of mammographic density and 138 families with 1-46 study participants per family (average of 6) in the HAPI Heart study

Table 2.2: Rare variant kinship test (RVKT) results from two genetic studies in the Amish

Trait	Tail size (n)	Observed mean pair-wise kinship coefficient		Expected mean pair-wise kinship coefficient (approximate 95% CI)[d]	P-value[d]	
		Lower tail	Upper tail		Lower tail	Upper tail
Prolactin[a]	2% (14)	0.080	0.067	0.068 (0.067-0.070)	0.0004	1.0000
Free estradiol[a]	8% (57)	0.021	0.019	0.018 (0.017-0.019)	0.0008	0.1650
Estradiol[b]	6% (44)	0.026	0.025	0.024 (0.022-0.025)	0.0015	0.0737
Fasting triglycerides[c]	2% (17)	0.074	0.056	0.058 (0.056-0.063)	0.0020	0.7999

[a] Based on 728 pre-menopausal women from our study of mammographic density, and after standardizing by batch and adjusting for age, an estimated narrow-sense heritability of approximately 24% (for prolactin) and 34% (for free estradiol)

[b] Based on 753 post-menopausal women from our study of mammographic density, and after standardizing by batch and adjusting for age, an estimated narrow-sense heritability of approximately 35%

[c] Based on 868 men and women from the HAPI Heart study, and after adjusting for age and sex, an estimated narrow-sense heritability of approximately 49%

[d] Based on 10,000 simulations under the null hypothesis of a purely polygenic trait architecture

Figure 2.1: Top quintile of the cumulative distribution function of the RVKT statistic
Distribution is based on 1,000 replicates of a purely polygenic trait with a narrow-sense heritability of 40% and (panels A and B) four-person nuclear families (n=1,484) or (panels C and D) trimmed pedigrees from our study of mammographic density (n=1,481). Panels A and C are based on a tail size of 1% (15 individuals), and panels B and D are based on a tail size of 8% (118 individuals). Dashed line denotes the 95th percentile.

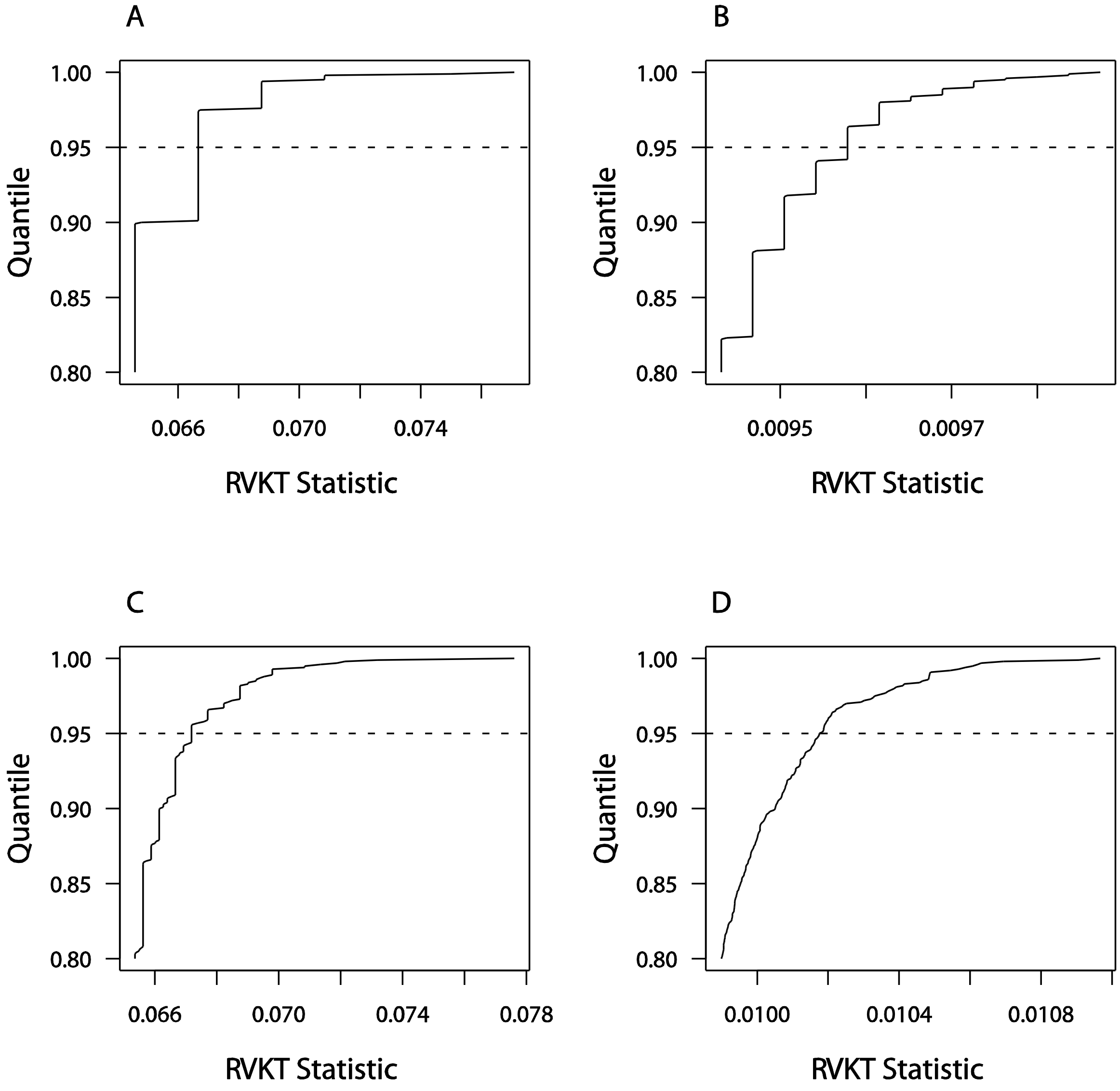


Figure 2.2: Power of the RVKT as a function of effect size

Effect size is the proportion of the trait variance explained by the rare variant. Results are based on 1,000 simulations of a quantitative trait and assume a rare variant allele frequency of 2%, a narrow-sense heritability of 40%, and pedigrees from (panel A) our study of mammographic density (n=1,481) or (panel B) the HAPI Heart study (n=868). Power is shown for trimmed Amish pedigrees (gray bars) and the complete 13-generation Amish pedigree (black bars). For comparison, power is also shown for four-person nuclear families (two parents and two offspring), with sample sizes equivalent to the sizes of our Amish studies (white bars). The significance level was set at 0.05.

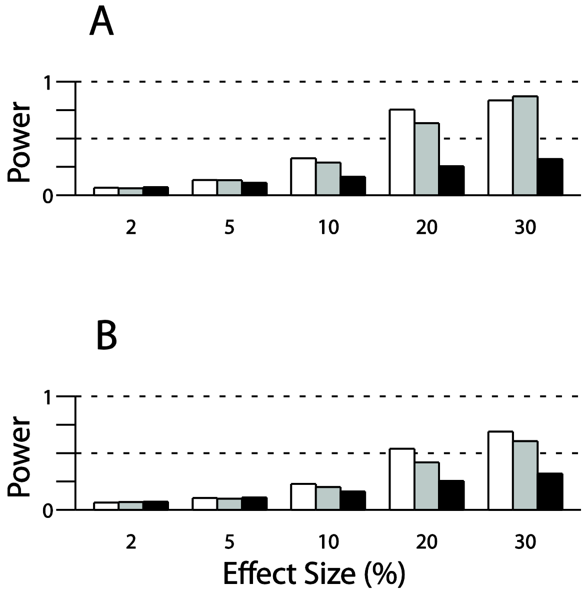


Figure 2.3: Power of the RVKT as a function of the rare variant allele frequency (RAF) Results are based on 1,000 simulations of a quantitative trait and assume a rare variant that accounts for 5% of the trait variance, a narrow-sense heritability of 40%, and pedigrees from (panel A) our study of mammographic density (n=1,481) or (panel B) the HAPI Heart study (n=868). Power is shown for trimmed Amish pedigrees (gray bars) and the complete 13-generation Amish pedigree (black bars). For comparison, power is also shown for four-person nuclear families (two parents and two offspring), with sample sizes equivalent to the sizes of our Amish studies (white bars). The significance level was set at 0.05.

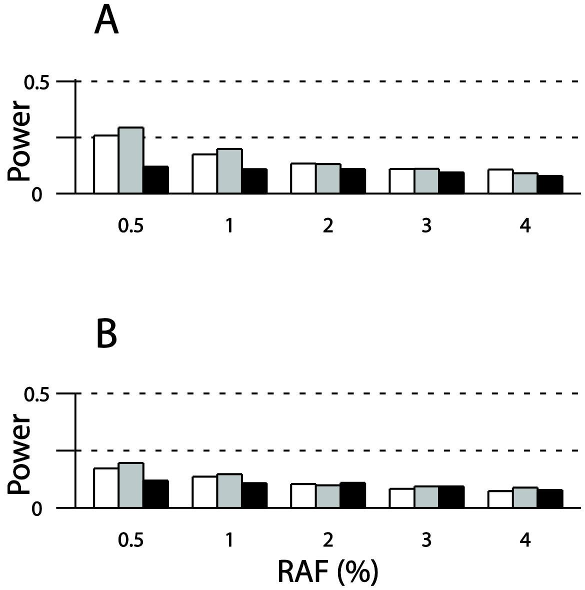


Figure 2.4: Power of the RVKT as a function of the narrow-sense heritability

Results are based on 1,000 simulations of a quantitative trait and assume a rare variant with an allele frequency of 2% that accounts for 1/8th of the genetic variance and pedigrees from (panel A) our study of mammographic density (n=1,481) or (panel B) the HAPI Heart study (n=868). Power is shown for trimmed Amish pedigrees (gray bars) and the complete 13-generation Amish pedigree (black bars). For comparison, power is also shown for four-person nuclear families (two parents and two offspring), with sample sizes equivalent to the sizes of our Amish studies (white bars). The significance level was set at 0.05.

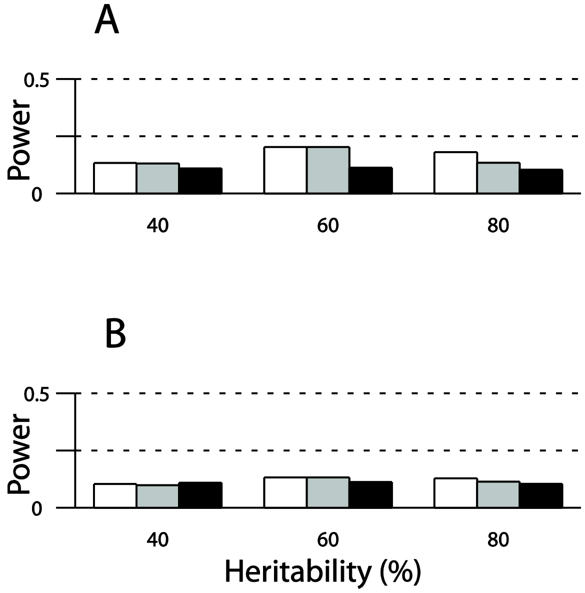


Figure 2.5: RVKT p-values (p_{min}) for 35 quantitative traits from our study of mammographic density

Each bar represents the result for a single trait. Black bars, significant ($p_{min} \leq 0.05$); gray bars, not significant. Dashed line denotes p-value threshold corrected for multiple testing. Before applying the RVKT, traits were transformed to approximate normality, when necessary, and adjusted for age and menopausal status, except for the hormones and growth factors, which were standardized by batch, adjusted for age, and analyzed separately for pre- and post-menopausal women.

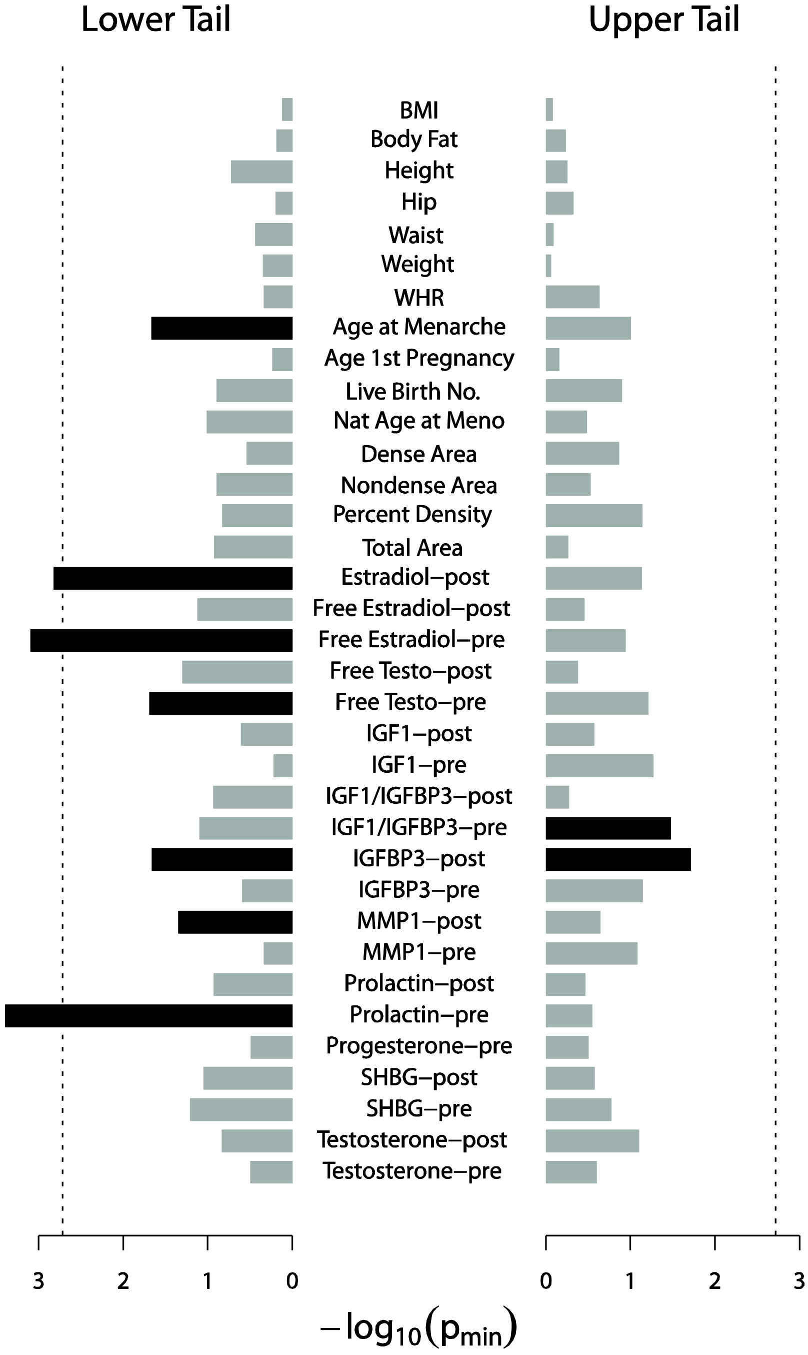
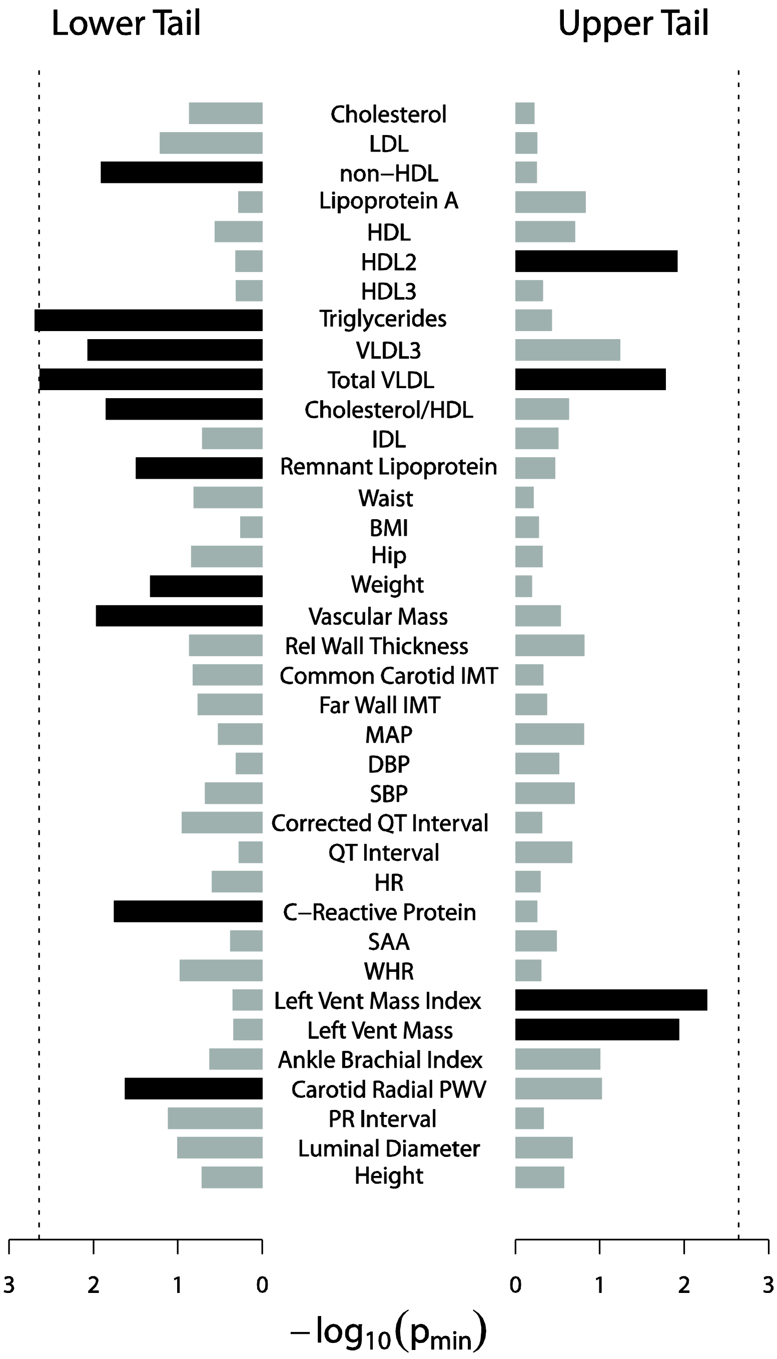


Figure 2.6: RVKT p-values (pmin) for 37 quantitative traits from the HAPI Heart study
Each bar represents the result for a single trait. Black bars, significant ($p_{min} \leq 0.05$); gray bars, not significant. Dashed line denotes p-value threshold corrected for multiple testing. Traits were transformed to approximate normality, when necessary, and adjusted for age and sex. Traits are ordered such that highly correlated traits are closer together.



Chapter 3

Multipoint Linkage Analysis Using Markov-Chain Monte Carlo Methods Fails to Converge for Complex Pedigrees and Dense Genetic Maps[2]

Dense arrays of SNPs are routinely genotyped and used to conduct association analyses of complex traits. In the context of family-based studies, these arrays also afford the opportunity to carry out genome-wide linkage analyses. However, exact calculation of identity by decent (IBD) sharing probabilities for linkage analysis becomes computationally intractable when dense genetic maps are used with complex pedigrees. Here we apply a Markov-chain Monte Carlo (MCMC) sampling method to estimate IBD sharing probabilities in complex pedigrees with considerable missing data and a dense genetic map. To our knowledge, the performance of MCMC-based methods has not been tested in this setting.

Our goal is to analyze quantitative traits from our genetic study of mammographic density in the Old Order Amish. Although the women from this study can be connected into a single 13-generation pedigree, we analyze them as a set of 177 trimmed pedigrees. We found substantial variation in LOD scores between MCMC runs due to a lack of convergence of the MCMC algorithm. For example, based on 12

[2] Shah KP and Douglas JA (2013) Multipoint Linkage Analysis Using Markov-Chain Monte Carlo Methods Fail to Converge for Complex Pedigrees and Dense Genetic Maps. *In preparation*

independent MCMC runs, the maximum LOD score for mammographic density ranged from 1.3 to 3 (mean of 1.9).

To improve the precision of our LOD score estimates while remaining computationally feasible, we adopt a combined strategy of windowing and averaging across the genome. Even with non-converged IBD sharing probability estimates, we were able to detect a previously validated linkage peak for serum matrix metalloproteinase (MMP1) levels (LOD score ranged from 9 to 11). While we were able to detect large linkage signals, improved MCMC methods are necessary to detect weaker signals that may still be informative for prioritizing association and sequencing studies.

3.1 Introduction

Dense arrays of SNPs are routinely used to conduct association analyses of complex traits. In the context of family-based studies, these arrays also afford the opportunity to carry out genome-wide linkage analyses. Although linkage analysis of complex traits is underpowered to detect common variants with modest effects, it may be useful in conjunction with sequencing and association analyses to identify rare variants with larger effects. For example, linkage information can be used to filter the variants discovered from whole exome or whole genome sequencing [14, 70, 71]. This approach has been successful both in the context of Mendelian disease [72] and complex quantitative traits [73, 74]. Similarly, Roeder *et al.* showed that the power of a genome-wide association study may be improved when linkage information is used to weight association p-values [75].

Multipoint linkage analysis can be broken down into two steps: 1) estimate allele sharing probabilities conditional on pedigree and genotype data and 2) test for linkage by modeling the trait as a function of these allele sharing probabilities. For each pair of study participants, we need to estimate the probability that they inherited 0, 1, or 2 alleles identical by decent (IBD) from a recent common ancestor. The choice of method to estimate IBD sharing probabilities depends on the pedigree complexity and marker density. Pedigree complexity is commonly measured as the bitsize, or twice the number of non-founders (*n*) in the pedigree minus the number of founders (*f*) or *2n-f*. A founder is defined as an individual with no parents in the pedigree. In contrast, a non-founder has both parents present in the pedigree. For large, complex pedigrees (bitsize ≥ 32) with only a few genetic markers (5-10 microsatellites), exact IBD sharing probabilities can be calculated efficiently via the Elston-Stewart algorithm [76]. Similarly, with small pedigrees (bitsize < 32) and many markers (>100 SNPs along a full chromosome), exact IBD sharing probabilities can be calculated via the Lander-Green algorithm [77]. For large, complex pedigrees and many markers, as is the case in our study, Markov-chain Monte Carlo (MCMC) sampling methods are necessary to estimate IBD sharing probabilities.

Previous efforts to evaluate the performance of MCMC methods for linkage analysis have focused on evaluating accuracy by comparison to exact methods [78-81]. Therefore, they were limited to evaluating only small pedigrees or only a few genetic markers. In 2006, Wijsman *et al.* evaluated the accuracy of MCMC-based IBD sharing probability estimates using their combined locus and meiosis sampler, implemented in the program lm_markers in the MORGAN software package [81]. They found small differences in LOD score estimates by comparing MCMC-based results to exact methods for a small pedigree and dense marker map and for a large pedigree and sparse marker map. For example, with a large pedigree (bitsize = 124) and sparse microsatellite marker map (1 marker/10 cM), the MCMC-based LOD score was on average 6% different than the exact LOD score [81]. MCMC-based methods, however, have yet to be evaluated using both large, complex pedigrees and a dense SNP marker map.

We set out to conduct genome-wide linkage analysis of mammographic density and known or suspected breast cancer risk factors from our study in the Amish population of Lancaster County, PA. For this study, we use a set of trimmed Amish pedigrees (up to a bitsize of 100) and a dense SNP marker map (~1 SNP/cM). Therefore, we evaluate the mixing performance of current MCMC methods for linkage analysis. Specifically, conditional on real and simulated data from our mammographic density study, we evaluate both LOD score variability and concordance of IBD sharing probability estimates.

3.2 Methods

3.2.1 Family-Based Study of Mammographic Density

We recruited 1,521 healthy Amish women between 2005 and 2010 to participate in our study of mammographic density. These women represent nearly half of the current census population of Amish women aged 40-80 years old in Lancaster County, PA. Our study design focused on recruitment of sibships. Initially, women ≥ 40 years old were only eligible for our study if they had a living sister ≥ 40 years old. We later dropped this criterion to recruit all women ≥ 40 years old, regardless of whether they had an eligible sister. Additional study design details, including eligibility criteria, are described elsewhere [54].

Our primary goal was to conduct linkage analysis to identify genetic factors related to mammographic density and selected quantitative traits related to breast cancer risk (listed in Appendix 5). Prior to analysis, we transformed the traits to approximate normality, when necessary, and adjusted for age and menopausal status. Our final analysis ready data set for linkage included phenotype and genotype information on 1,472 women. These women can be connected into a single 13-generation pedigree using the extensive genealogical records from the Anabaptist Genealogical Database (AGDB) [34, 35].

3.2.2 Linkage Methods

Here we describe the genetic marker map, pedigree structures, and MCMC method used for linkage analysis.

3.2.2.1 Genetic Map

The women in our study were genotyped on the Illumina HumanOmni2.5-4v1_B Array (2,443,179 SNPs) by the Center for Inherited Disease Research (CIDR). After extensive quality control (see Appendix 2 for details), we retained 1,452,421 SNPs with a minor allele frequency (MAF) ≥ 1% in our Amish sample. In order to conduct multipoint linkage analysis, we selected a set of SNPs in low linkage disequilibrium (LD). To select these SNPs, we first estimated pair-wise LD between SNP using a set of 400 minimally related women. These women were optimally selected using PedMine [82]. We then used the --*indep-pairwise* option in Plink V1.07 [83] to prune SNPs based on linkage disequilibrium. In Plink, we used a 5,000 SNP window, with 1 SNP step size and retained SNPs with a pair-wise $R^2 < 7\%$. After pruning, we retained 3,523 SNPs for linkage analysis. These SNPs were densely spaced across the genome (mean spacing = 1.1 cM), highly polymorphic (mean MAF = 0.42), and nearly uncorrelated (mean pair-wise $R^2 = 1\%$, maximum pair-wise $R^2 = 13\%$).

To determine a unique position for each SNP in our genetic map, we used the Rutgers V2 sex-averaged genetic map [84] and linear interpolation. We converted the Kosambi map positions in the original Rutgers map to Haldane units in SOLAR V4.3.1 [85]. Linkage analysis using the variance components model in SOLAR requires using a Haldane map function. The Haldane map function assumes no crossover interference, i.e. it assumes recombination events are independent. Thus we can use a Markov-chain to model the IBD process along the chromosome (see Section 3.2.2.3).

3.2.2.2 Pedigree Trimming

The 1,472 women in our study can be connected into a single 13-generation 5,044-member pedigree, including 201 founders and 4,843 non-founders. Because this pedigree is not computationally tractable to estimate IBD sharing probabilities, we use a set of trimmed pedigrees to estimate IBD sharing probabilities and conduct linkage analysis. We created these pedigrees by merging in the parents and grandparents of all study participants and then trimming the pedigrees with PedCut [68] to a maximum bitsize of 100. PedCut was developed to trim pedigrees to a constrained bitsize while maximizing the number of phenotyped individuals per pedigree [68]. Using a bitsize of

100 maintained most first- and second- degree relationships between study participants. Extensive trimming (bitsize ≤ 32) would have allowed exact calculation of IBD sharing probabilities; however, many first-degree relative pairs would have been lost in the trimming process. Ignoring these close relationships can falsely increase [86] or decrease [87] linkage evidence. Therefore, our choice of bitsize was a trade-off between computational practicality and the accuracy of linkage results.

3.2.2.3 Estimation of IBD Sharing Probabilities

Estimating the probability that a pair of individuals inherited 0, 1, or 2 alleles IBD at a specific locus from a recent common ancestor is done by enumerating all the possible inheritance patterns at that locus that are consistent with the known pedigree and marker data. The IBD process along a chromosome can be modeled as a Markov-chain, where the hidden states correspond to the true IBD at each position along the chromosome and the transition probabilities between states are related to genetic distance between adjacent loci, i.e., the probability of a recombination event. Meiosis indicators, S_{ij}, are used to track the IBD state through the pedigree for $i = 1 .. n$ individuals at each $j = 1 .. m$ markers along the chromosome. S_{ij} is 0 when the gamete inherits the parent's paternal allele and 1 when the gamete inherits the parent's maternal allele. A meiosis vector, S_M, represents the complete IBD state for all individuals, i, and markers, j. In other words, S_M is the vector of all S_{ij}'s for $i = 1 .. n$ individuals at each $j = 1 .. m$ markers. The number of different possible meiosis vectors that are consistent with the data grows has the number of individuals, n, and markers, m, increases.

As pedigree complexity (bitsize) and marker density increases, it is computationally impossible to enumerate all possible meiosis patterns that are consistent with the observed data. Instead, MCMC methods can be used to sample meiosis vectors conditional on the observed marker data and pedigree configurations to estimate IBD sharing probabilities. We used the combined locus and meiosis sampler, lm-sampler, implemented in Loki V2.4.7.4 [88] to estimate IBD sharing conditional on our Amish pedigrees and genetic marker map. This implementation of the MCMC algorithm is equivalent to the lm_markers program in the MORGAN software package [67, 81]. The lm-sampler uses a block-Gibbs MCMC sampler that updates a subset of parameters during each iteration. At each iteration, either $S_{.j}$, all meioses at a single locus (l-step) [88], or $S_{i.}$, all loci at a single meiosis (m-step) [89] are updated. The l-sampler has poor mixing performance with tightly linked markers, while the m-sampler has poor mixing performance with large pedigrees with a lot of missing data in the parental generations. In order to deal with both of these challenges in our Amish data, we used a 1:1 ratio of l- to m- steps. We ran the MCMC algorithm using 500,000 burn-in iterations and 1 million or 30 million total iterations. We estimated genome-wide IBD sharing probabilities for all pairs of women along a 1cM.

3.2.2.4 Comparison of IBD Sharing Probabilities between MCMC Runs

To evaluate convergence of the MCMC algorithm, we ran the algorithm with multiple random start seeds and compared results. We calculated the difference in local kinship estimates, φ_j, along chromosome 10 between the two MCMC runs. The kinship coefficient, φ_j, is defined as the probability that at locus j, an allele sampled from one individual is IBD with an allele sampled from another individual. We estimated these local kinships at a 1cM grid along chromosome 10 from *j = 0 … 179* cM for all pairs of individuals in our study. Chromosome 10 was used as a representative chromosome because it is nearly average length in terms of number of SNP markers. Also, this chromosome was of particular interest because our top linkage result for mammographic density was on chromosome 10.

3.2.2.5 Calculation of LOD Scores

To calculate LOD scores, we used standard multipoint variance components methods implemented in SOLAR V4.3.1 [85]. These methods model the quantitative trait variance as function of covariates, a major gene at the locus being tested, and the residual additive genetic effects. We evaluated the evidence for linkage by comparing the likelihood of a model with and without linkage using a standard likelihood ratio test. The LOD score is defined as the logarithm, base 10, of the likelihood ratio, or the logarithm of the odds of linkage. We calculated LOD scores at a 1cM grid across the genome, corresponding to the points at which we estimated IBD sharing probabilities using the MCMC algorithm described above.

3.2.2.6 Averaging LOD Scores

We ran the MCMC algorithm 12 times, each with a different random start seed, to estimate IBD sharing probabilities. We then used each set of IBD sharing probabilities to estimate LOD scores. Finally, we combined these 12 LOD scores by averaging them at each corresponding grid point across the genome. We used the mean LOD score over separate MCMC runs as our measure of evidence for linkage. Using the mean helps to reduce the noise in linkage results that comes from variability in estimating the IBD sharing probabilities.

3.2.2.7 Evaluating the False-Positive Rate

To evaluate the false-positive rate of our averaging approach to linkage analysis, we used gene dropping simulations in MORGAN V3.0.3 [67] conditional on the full 13-generation Amish pedigree and the genetic map from chromosome 10. Specifically, we assigned genotypes for each marker on chromosome 10 to the pedigree founders conditional on the observed marker allele frequencies. We then simulated meioses through the pedigree to create marker genotypes for all non-founders in the pedigree. For each set of simulated genotypes, we also simulated a polygenic quantitative trait with total heritability equal to 40%. The quantitative trait was simulated under a null model of no linkage to the simulated chromosome. We created 50 replicate data sets with genotype and quantitative trait values for all phenotyped study participants. During the simulations, we kept track of actual IBD sharing and used these values to calculate 'actual LOD scores' across the chromosome. We then compared the LOD scores from a single MCMC run and the average over 12 MCMC runs to the actual LOD scores.

Our gene dropping simulations allowed us to evaluate chromosome-wide significance only for linkage peaks on chromosome 10. Ideally, we would have conducted gene dropping simulations genome-wide to determine appropriate empirical significance thresholds, however, this was computationally impossible. As an alternative to extensive gene dropping simulations, we used the thresholds set forth by Lander and Kruglyak (1995) to determine genome-wide significance. A LOD score greater than 1.9

indicates suggestive evidence for linkage and a LOD score greater than 3.3 indicates significant evidence for linkage. Under the null hypothesis of no linkage, we expect to observe a LOD score ≥ 1.9 once across the genome and a LOD score ≥ 3.3 only 1 in 20 times.

3.2.3 Computing Resources

3.2.3.1 Cluster Configuration

All analyses were run on our shared Linux compute cluster. The full cluster is comprised of 65 compute nodes with a total of 588 cores. Forty-one of the compute nodes have an Intel Xeon Processor (X5680), dual 6-cores (12 cores), a processor speed of 3.3GHz, and 48 GBs of RAM. The remaining 24 compute nodes have an Intel Xeon Processor (5160), dual 2-cores (8 cores), a processor speed of 3.00GHz, and 16GB of RAM. Each of the 588 cores works as a single central processing unit, or CPU.

3.2.3.2 Windowing

MCMC processes are computationally intensive, especially as pedigree complexity and marker density increase. Computing time is therefore an important and practical consideration when using MCMC methods to estimate IBD sharing probabilities. In order to run the MCMC algorithm multiple times genome-wide in a reasonable amount of time, we need to use our full compute cluster. However, we had limited access to all 588 cores because our compute cluster is a shared lab resource. Consequently, we adopted the windowing strategy to maximize use of our available computing resources. The benefits of windowing will depend on individual computing resources; however, windowing did not change the overall results of our linkage analysis. Specifically, the distribution of the differences in local kinship estimates along chromosome 10 was similar with and without windowing (data not shown).

To estimate IBD sharing probabilities we divided the genome into 20-SNP windows with a 10-SNP overlap. To cover the genome once required 361 windows. Each window was run in parallel on a single CPU to estimate IBD sharing across the genome. Prior simulation work showed a minimal loss in information content using a 20-SNP window compared to a full chromosome of SNP data (personal communication with Albert Levin, data not shown). Without windowing, each chromosome could be run in parallel, however, this would not fully utilize our computing resources and each individual process would take longer to run.

For each trait in our linkage analysis, we calculated LOD scores at a 1cM grid within each window. For the overlapping regions of each window, we averaged LOD scores across windows. This resulted in a single LOD score at each point along the genome. After repeating this process 12 times, we averaged the replicate runs as described in section 3.2.2.6

3.3 Results

3.3.1 Trimming Amish Pedigrees

We trimmed our Amish pedigrees in order to create a set of computationally tractable pedigrees while maintaining the close relative pairs that are informative for linkage analysis. An example pedigree after trimming (bitsize = 96) is shown in Figure 3.1. This 84-member pedigree had 42 phenotyped individuals in the bottom two generations. Large sibships without genotype information for individuals in the preceding generations are especially challenging for estimating IBD sharing probabilities via MCMC methods. This pattern of missing data is characteristic of our example pedigree (Figure 3.1) and our Amish pedigrees more generally. Consistent with our recruiting strategy, there were a large number of sibling pairs (1,254) and many fewer parent-offspring pairs (274) in our trimmed pedigrees (Table 3.1). In total, the 1,472 women in our study were grouped into 177 trimmed pedigrees with bitsize ≤ 100 (Figure 3.2a). There were with 1-42 phenotyped individuals (Figure 3.2b) and 1-35 missing individuals (Figure 3.2c) per pedigree.

3.3.2 MCMC Convergence

We found that LOD score estimates were highly variable across different MCMC runs. For example, Figure 3.3 shows the LOD scores across chromosome 10 for mammographic density, our primary trait of interest, based on IBD sharing probabilities from 12 separate MCMC runs (1 million iterations each). At the linkage peak with the maximum mean LOD score, we observed LOD scores ranging from 1.3 to 3.1 (Figure 3.3, gray lines).

The observed variation in LOD scores was due to differences in estimates of IBD sharing probabilities between MCMC runs. For example, Figure 3.4 shows the cumulative distribution function (CDF) of differences in local kinship estimates, φ_j, for all pairs of women in our study at a 1cM grid along chromosome 10 for 2 MCMC runs. After 1 million iterations, only 13% of φ_j estimates converged to the same value (Figure 3.4, black curve). Surprisingly, the majority of the differences in φ_j estimates were small. In fact, only 3.3% of φ_j had a difference ≥ 0.125 (Figure 3.4, black curve).

The differences in kinship estimates between MCMC runs persisted even after a 30-fold increase in the number of MCMC iterations. For example, after 30 million iterations, 2.3% of local kinship coefficients still had a difference ≥ 0.125 (Figure 3.4, red curve). This suggests that the MCMC algorithm is stuck at a local maximum and that increasing run time will not substantially improve IBD sharing estimates.

3.3.3 Linkage Analysis of Mammographic Density and Selected Quantitative Traits

To evaluate evidence for linkage despite the variability in estimates of IBD sharing probabilities between MCMC runs, we averaged LOD scores over 12 separate MCMC runs. Our top linkage peak was on chromosome 10, with a mean LOD score of 1.88 (details in Table 3.2). Genome-wide linkage results for mammographic density are shown in Figure 3.5. The same variability in LOD scores that were observed on chromosome 10 (Figure 3.3) were observed genome-wide (Figure 3.5).

The high variability in LOD scores based on different MCMC runs was seen across all of the quantitative traits we analyzed. The top linkage peaks (mean LOD score ≥ 1.5) for all traits are shown in Table 3.2. Still, when a known locus of large effect was present, we were able to detect it. For example, the significant linkage peak on chromosome 11 for serum MMP1 levels (mean LOD = 10) (Table 3.2) was previously reported in another Amish study [90]. This linkage peak can be accounted for by 3 non-coding SNPs (rs495366, rs12289128, and rs11226373) near the MMP gene cluster on chromosome 11q; collectively, they account for 31% of the phenotypic variance in serum MMP1 levels [90].

3.3.4 False-Positive Rate of the Mean LOD Score

Based on simulations under the null hypothesis of no linkage, conditional on our Amish pedigrees, we found that using the mean LOD score had lower LOD scores overall compared to the a single LOD score estimate and the actual LOD scores. Table 3.3 shows the comparison of actual LOD scores to mean LOD scores over the 180 grid points on chromosome 10 for the 50 replicates (9,000 total LOD score estimates). The mean LOD score never exceeded 1.5, whereas the actual LOD score was greater than 1.5 under a model of no linkage 27 times. Large LOD scores in the absence of an actual trait locus are likely a result of using the trimmed pedigrees for analysis. The mean LOD score appears to average out some of the biases introduced by pedigree trimming.

In our simulations, the maximum mean LOD score on chromosome 10 was 1.2. Using this threshold, our maximum mean LOD score for mammographic density (1.88) has a chromosome-wide p-value ≤ 0.02. At this threshold (LOD > 1.2), we may have additional linkage peaks on chromosomes 8p, 2p, and 22p for mammographic density. Details about these linkage peaks are in Table 3.2. It is important to note that the linkage peaks for mammographic density would not have met our criterion for genome-wide suggestive evidence of linkage, i.e. mean LOD scores were all less than 1.9.

3.3.5 Computing Time

MCMC processes are computationally intensive. The decision to window was a practical one, specific to our compute resources and CPU availability. The run time of the MCMC algorithm depends on the pedigree structures, sample size, map density, chromosome length, and number of iterations. Using chromosome 10 as a representative chromosome, we extrapolated run times of the MCMC algorithm with and without windowing for our study (Table 3.4). The MCMC algorithm ran for 4 continuous days to estimate IBD sharing probabilities across the full chromosome 10. In contrast, each 20-SNP window took 0.5 days to run the MCMC algorithm. Chromosome 10 breaks down into 16 overlapping windows. If we were only using a single CPU, it would be more efficient to run the MCMC algorithm for the full chromosome compared to windowing (4 versus 8 days, Table 3.4). However, with multiple available CPUs, we improved efficiency by running multiple processes in parallel. Using our full compute cluster (588 CPUs), estimating IBD sharing probabilities 12 times genome-wide was expected to take 4 days with or without windowing (Table 3.4). In the case of windowing, however, the computer run time could be divided over 0.5 day blocks of time. In contrast, running full chromosomes would require continuous computer run time. Because our compute cluster is a shared lab resource, we chose to estimate IBD sharing probabilities genome-wide using the windowing approach. This allowed us estimate IBD sharing probabilities genome-wide 12 times over 2 weekends without disrupting others using the compute cluster during the week.

For the same pedigrees and genetic marker map, run time scales linearly with the number of MCMC iterations. To evaluate convergence, we ran the MCMC algorithm for 30 million iterations for the 16 windows covering chromosome 10. Each window took 15 days to run. Therefore, while we were able to evaluate a single chromosome, this approach would not be a feasible genome-wide.

3.4 Discussion

In family-based studies, genome-wide SNP markers allow researchers to conduct both linkage and association analyses. However, linkage analysis with SNPs requires a denser marker map compared to previous microsatellite marker maps in order to be informative for estimating IBD sharing probabilities [91]. In combination with large pedigrees, dense marker maps require the use of MCMC methods to conduct linkage analysis. Therefore, evaluating MCMC methods in this context is an important research problem.

The lm-sampler MCMC algorithm failed to converge on stable estimates of IBD sharing probabilities for our complex pedigree structures and dense SNP marker map. For our study of mammographic density, regardless of how long we ran the MCMC algorithm, only 12-13% of local kinship estimates, φ_j, were the same between MCMC runs. Overall, the majority of estimates showed only small differences in probability estimates between MCMC runs. While only a small proportion of pairs had large discrepancies in φ_j estimates, these pairs contribute disproportionately to LOD scores because they have higher local kinship estimates. Thus, we suspect that the aggregation of many small differences in φ_j estimates and a few very large differences resulted in large differences in LOD score estimates between MCMC runs when conducting linkage analysis. In the future, it will be important to identify which types of relative pairs have the most variable φ_j estimates to better understand the observed variability in LOD scores.

Prior to this study, an evaluation of MCMC methods had not been done for both complex pedigrees (bitsize ≤ 100) and a dense genetic marker map (1 SNP/cM). Our results are, however, consistent with previous observations about the performance of MCMC methods for linkage analysis, even though these studies used smaller pedigrees and/or fewer markers. Wijsman *et al.* observed differences between MCMC based LOD score estimates and exact LOD scores for both a small pedigree with a dense SNP marker map and a large pedigree with a sparse set of STR markers [81]. Sieh *et al.* also noted that MCMC-based linkage results were "noisier" and likely contained more false-positives when using a SNP marker map compared to a microsatellite map [78].

In 2008, Tong and Thompson also recognized the need to improve MCMC methods for IBD estimation [79]. Their approach extends the m-step of the lm-sampler by sampling a subset of meioses at each iteration instead of just a single meiosis. While the software was not available at the time our study was conducted, the sampler has now been implemented as lm_auto in the MORGAN V3.0.3 software package [67]. Tong and Thompson showed that with a large pedigree (bitsize of 70) and sparse genetic map (0.17 SNP/cM), their multiple-meiosis sampler combined with the locus-sampler gave more consistent LOD score estimates compared to the original lm-sampler (used in this study) [79]. For example, in the center of the marker map the mean difference between the estimated and exact LOD score was 0.3 using the lm-sampler and 0.01 using the multiple-meiosis sampler after 1 million MCMC iterations. However, our recent, limited evaluation of the lm_auto program using default parameters with our trimmed pedigrees and marker data from chromosome 10 showed larger differences in IBD sharing probability estimates between MCMC runs than the lm-sampler, e.g. 6% of φ_j had a difference ≥ 0.125 versus 3% with the lm-sampler. In the future, a more extensive evaluation of the lm_auto program, including optimizing run parameters and starting configurations, will be necessary to determine how this sampler performs with complex pedigrees and a dense genetic map in terms of consistency and computing time.

As an interim approach to conducting linkage analysis using current MCMC methods, we averaged LOD scores from multiple MCMC runs. Our averaging approach likely provided more reliable evidence for linkage compared to the any single linkage scan. Using a single MCMC run to estimate LOD scores could have resulted in very different interpretations depending on the run, and in this way could be more susceptible to false interpretation of linkage signals. For example, in one MCMC run, the maximum LOD score on chromosome 10 for mammographic density was 3, and in another run it was 1.3. Thus, in one situation, we would have a nearly genome-wide significant result, whereas in the other we would not even have suggestive evidence for linkage.

There is clear room for improvement of MCMC methods for estimating IBD sharing probabilities with complex pedigrees and dense genetic maps. A limitation of the current study is that it is restricted to pedigrees from our Amish study. In the future, an evaluation of MCMC methods with more general, but large, pedigrees may be useful to others conducting linkage analysis. Specifically, determining what types of pedigree configurations, e.g. sibship sizes and missing data rates result in non-converged estimates of IBD sharing probabilities for current MCMC methods. This could provide guidance both in the design of future studies, and making decisions on how to optimally trim pedigrees when using MCMC methods to estimate IBD sharing probabilities.

In our study, we chose to trim pedigrees to a bitsize ≤ 100 while maximizing the number of phenotyped individuals per pedigree. Trimming our pedigrees to be small enough for exact IBD probability calculations, i.e. bitsize < 32, will likely remove too many o close relative pairs between and therefore dramatically reduce the power of our linkage analysis. Nonetheless, evaluating the impact of different amounts of trimming, e.g. bitsize limits of 75 or 50, may be useful in determining the trade-off between accuracy, computing time, and the validity of linkage results.

Linkage analysis, and more generally IBD sharing methods to prioritize regions of the genome harboring rare variants for complex traits, will be increasingly useful as next-generation sequencing costs go down. The vast number of variants discovered in a sequencing study will require methods to filter the list of likely candidates prior to genotyping and testing in large samples of individuals. In family-based studies, linkage analysis has the potential to be a useful filter, provided there are accurate methods to conduct linkage analyses with dense SNP maps [73, 74]. Therefore, it is important to continue to develop and evaluate methods for conducting linkage analysis with complex pedigrees and dense genetic maps.

Table 3.1: Selected pair-wise relationships among the 1,472 women in our mammographic density study

Pair-wise relationships were determined using SOLAR V4.3.1 [85] after pedigree trimming.

Relationship Type	Number of Pairs
Parent-offspring	274
Siblings	1,254
Avuncular	1,119
Half siblings	5
1st cousins	4,638
1st cousins, once removed	2,948
2nd cousins	1,341
Other	866

Table 3.2: Linkage peaks with LOD ≥ 1.5 for selected traits

The mean LOD score is the average LOD score based on 12 separate MCMC runs to estimate IBD sharing probabilities. All traits were transformed to approximate normality, when necessary, and adjusted for age and menopausal status.

Trait	Chr	Position (cM)	Mean LOD (range)
dense area	2	1	1.79 (1.03 - 2.90)
dense area	8	19	1.78 (1.23 - 2.39)
dense area	10	89	1.88 (1.34 - 3.01)
non-dense area	17	52	2.02 (1.46 - 3.04)
non-dense area	17	108	2.36 (1.73 - 3.21)
percent density	8	26	1.91 (1.30 - 2.83)
percent density	11	90	1.60 (1.19 - 1.96)
total area	17	108	1.59 (1.06 - 2.27)
age at menarche	1	223	1.79 (1.25 - 2.15)
age at menarche	12	139	1.65 (1.24 - 2.33)
age at menarche	16	24	1.82 (1.49 - 2.40)
age at menarche	19	10	1.64 (1.03 - 2.26)
live birth no.	3	201	1.78 (1.12 - 2.52)
live birth no.	16	57	3.38 (2.84 - 4.43)
live birth no.	21	2	2.98 (2.24 - 3.80)
live birth no.	23	110	1.61 (1.51 - 1.82)
live birth no.	23	210	3.49 (3.14 - 3.87)
BMI	3	126	1.96 (1.23 - 2.67)
BMI	6	122	2.13 (1.59 - 2.96)
BMI	9	141	1.72 (1.38 - 2.08)
hip	6	124	2.12 (1.65 - 2.71)
hip	9	137	2.02 (1.21 - 5.52)
waist	5	3	1.66 (0.70 - 3.07)
waist	6	122	2.84 (2.32 - 3.51)
waist	9	138	1.64 (1.24 - 2.07)
waist	10	146	2.45 (1.85 - 3.22)
waist	17	108	1.57 (1.17 - 2.49)

Trait	Chr	Position (cM)	Mean LOD (range)
weight	3	126	1.71 (1.16 - 2.10)
weight	4	155	1.51 (1.11 - 1.98)
weight	6	122	1.79 (1.28 - 2.53)
MMP1	2	0	2.37 (1.20 - 3.66)
MMP1	2	99	1.85 (1.23 - 2.82)
MMP1	3	132	1.78 (1.06 - 2.35)
MMP1	5	180	1.84 (1.09 - 3.04)
MMP1	6	84	2.34 (1.76 - 3.01)
MMP1	8	61	3.77 (3.08 - 4.78)
MMP1	11	108	10.16 (8.97 - 11.2)
MMP1	17	84	2.08 (1.20 - 2.96)

Table 3.3: Agreement between actual LOD scores and the mean LOD score under the null hypothesis of no linkage

We conducted 50 replicate gene dropping simulations for chromosome 10 based on the pedigrees, marker map, and allele frequencies observed in our Amish study under the null hypothesis of no linkage. For each replicate, we calculated LOD scores at 180 grid points along chromosome 10 (1 point/cM), resulting in 9,000 LOD score estimates.

		Actual LOD					
		0-0.5	0.5-1	1-1.5	1.5-2	2-2.5	TOTAL
Mean LOD	0-0.5	**8,219**	344	54	8	0	8,625
	0.5-1	181	**115**	34	16	1	347
	1-1.5	14	7	**5**	2	0	28
	1.5-2	0	0	0	**0**	0	0
	2-2.5	0	0	0	0	**0**	0
	TOTAL	8,414	466	93	26	1	9,000

Table 3.4. Number of days to compute IBD sharing probabilities

Compute run times are extrapolated from the time to estimate IBD sharing probabilities for chromosome 10 with and without windowing. Chromosome 10 had a total of 161 SNPs and was covered by 16 20-SNP overlapping windows. IBD sharing probability estimates are based on 1 million iterations of the MCMC algorithm conditional on our study sample (N = 1,472) and genetic marker map (3,523 SNPs total).

	Without windowing				**With windowing**			
		Number of CPUs				**Number of CPUs**		
	# chrs	**1**	**32**	**588**	**# windows**	**1**	**32**	**588**
Chromosome 10 (1 time)	1	4	4	4	16	8	0.5	0.5
Whole Genome (1 time)	23	92	4	4	361	180.5	6	0.5
Whole Genome (12 times)	276	1,104	36	4	4332	2,166	68	4

Figure 3.1: Example Amish pedigree with a bitsize of 96

This pedigree contains 84 individuals. The squares are males and circles are females. Marriage nodes are shown as small dark circles. Filled in solid circles in the bottom 2 generations are the 42 women in our study of mammographic density with genotype and phenotype information in this family.

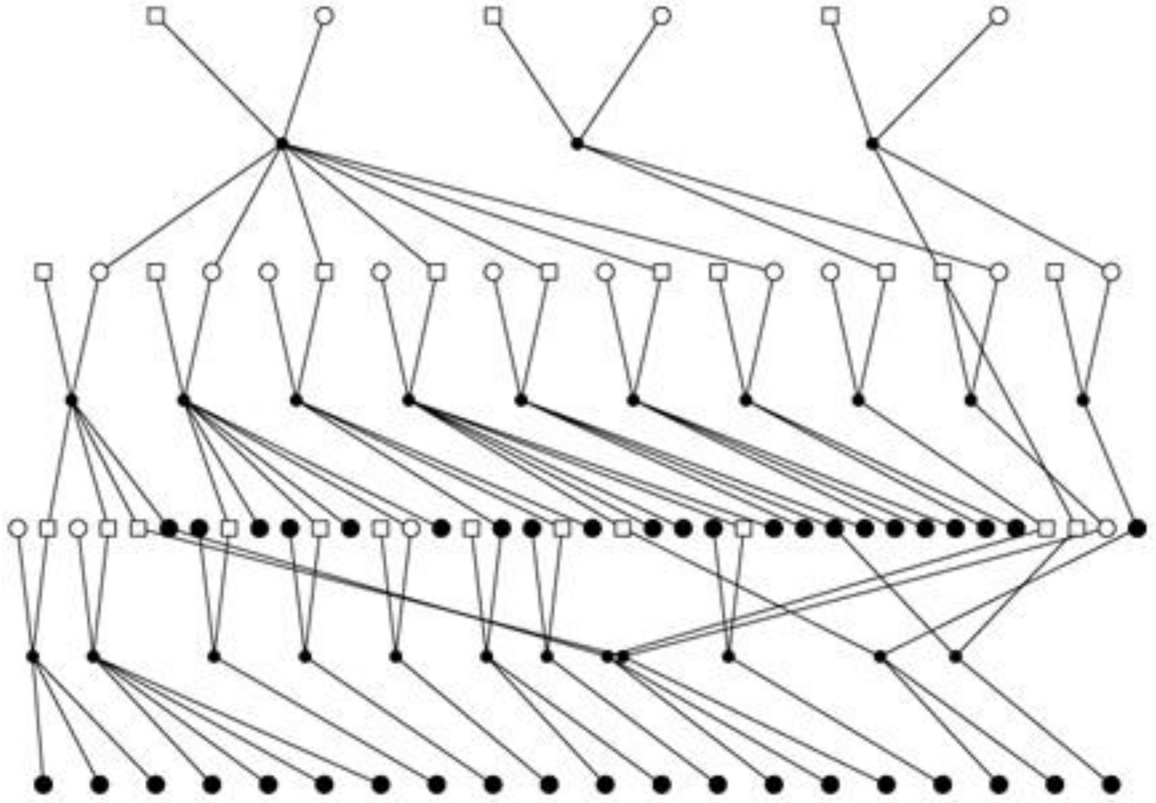


Figure 3.2: Trimmed Amish pedigree size characteristics
Estimating IBD sharing probabilities via MCMC methods is more difficult as pedigree complexity (bitsize = *2n-f*) and missing data increase. These histograms show the distribution of pedigree bitsize (A), the number of phenotyped individuals per pedigree (B), and the number of individuals in each pedigree with missing data (C) in our 177 pedigrees after trimming.

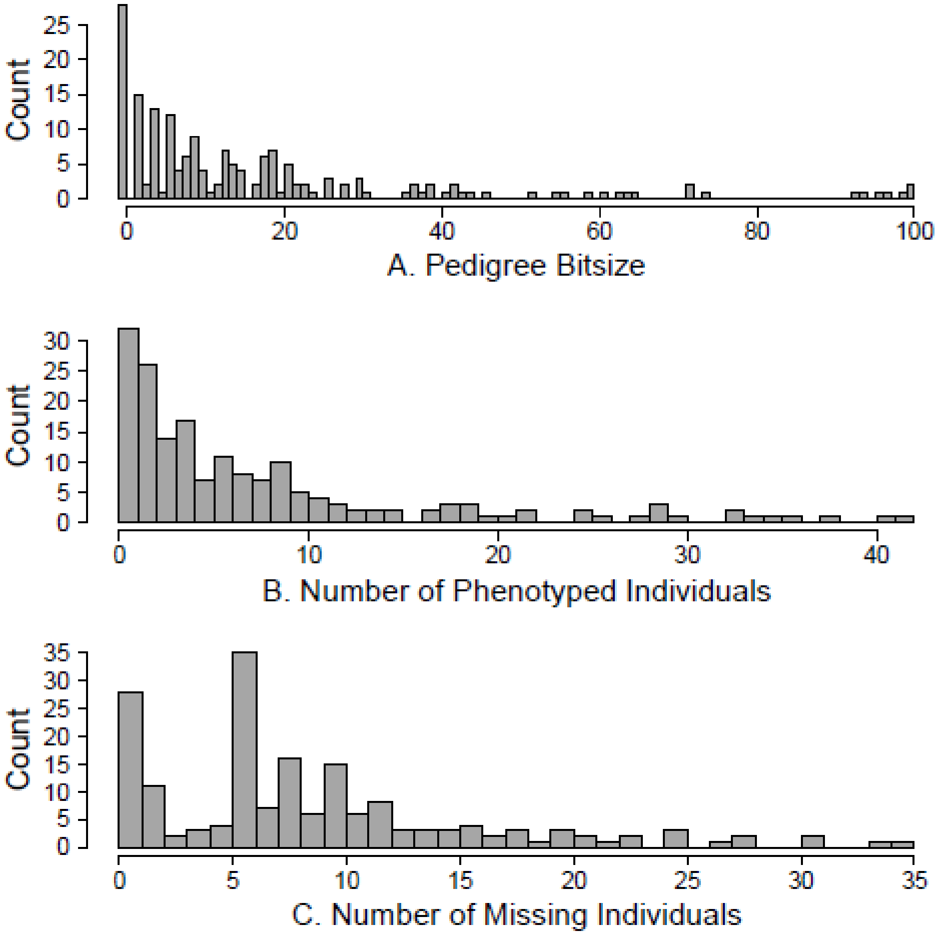


Figure 3.3: LOD scores for mammographic density on chromosome 10

Detailed plot of LOD scores across chromosome 10 for log-transformed absolute dense area adjusted for age and menopausal status. The 12 gray lines show results from the individual MCMC runs, and the red line is the mean LOD score. IBD sharing estimates were based on MCMC runs with the lm-sampler in Loki after 1 million iterations. Chromosome 10q had the largest mean LOD score at 89cM with a 1-LOD support interval of 78-102.

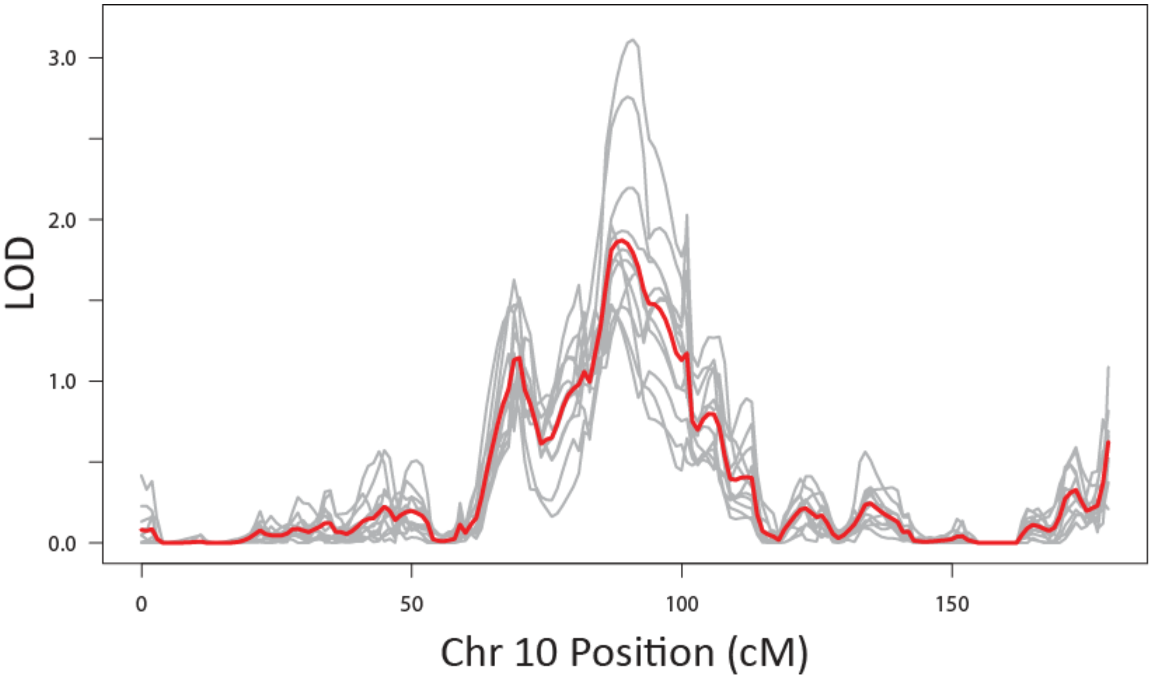


Figure 3.4: The cumulative distribution function of differences in pair-wise IBD sharing probability estimates between MCMC runs

The cumulative distribution function of the differences in IBD sharing probabilities, Δ(φ_j), estimated along a 1cM grid on chromosome 10 (*j = 0 .. 179*) for all pairs of women in our study from 2 independent MCMC runs. We evaluated convergence of the lm-sampler in Loki with 1 million (black solid line) or 30 million (red dashed line) iterations [88].

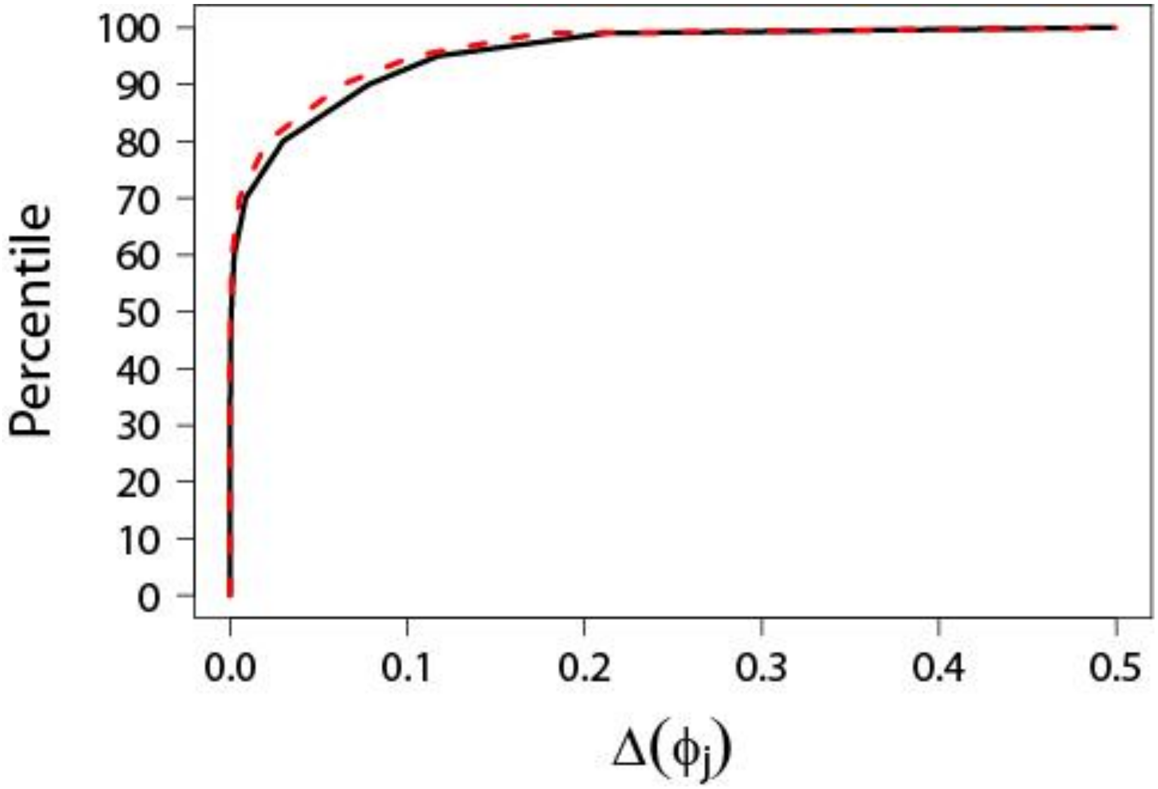
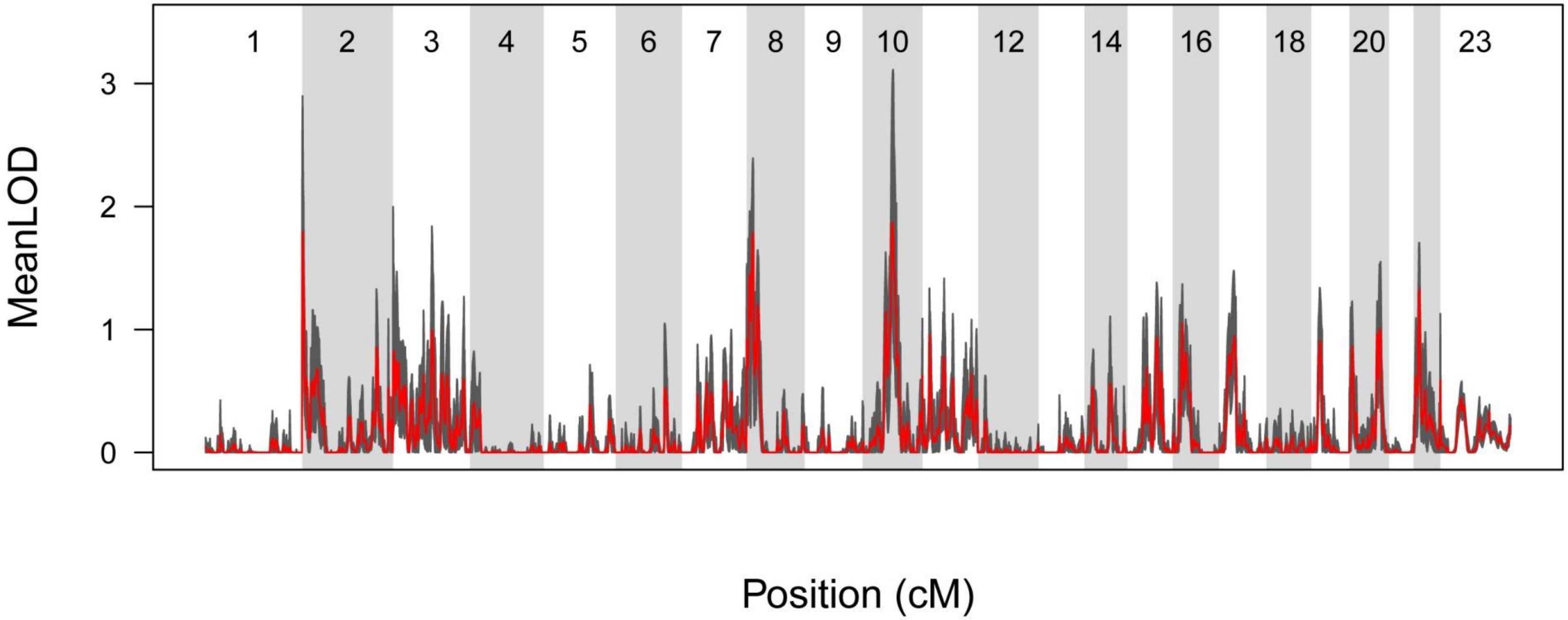


Figure 3.5: Genome-wide LOD scores for mammographic density
The mean LOD scores (red) are plotted across the genome for log-transformed absolute dense area, adjusted for age and menopausal status. Individual chromosomes are alternately shaded. The high variability in LOD scores between MCMC runs was seen across the genome. The range of LOD score estimates from the 12 separate MCMC runs is shown as the dark gray shadow.



Chapter 4

Determining Sources of Bias in Estimates of the Narrow-Sense Heritability[3]

Since the sequencing of the human genome, gene mapping efforts have identified thousands of loci associated with hundreds of medically relevant complex traits. Still, a substantial fraction of the estimated heritability of these traits remains to be explained. One explanation for this so-called "missing heritability problem" is that estimates of the heritability of many traits has been over estimated. It has long been appreciated that variance components methods can overestimate the additive genetic variance, or narrow-sense heritability of a trait, in the presence of gene-gene interactions, gene-environment interactions, and/or gene-environment covariance. Furthermore, studies in model organisms suggest that these interactions may underlie up to half of the observed phenotypic variation [24].

Recently, Zuk *et al.* proposed a new, regression-based method to estimate the narrow-sense heritability that is, in theory, robust to the underlying genetic architecture [21]. We applied their regression-based method selected traits from our study of mammographic density in the Amish and estimated the heritabilites of all traits to be ~40% lower than variance components based estimates. Before interpreting these results as evidence for pervasive interaction effects, we sought to evaluate the

[3] Shah KP and Douglas JA (2013) Determining Sources of Bias in Estimates of the Narrow-Sense Heritability. *In preparation*

statistical properties of the proposed estimator using simulations. We evaluate the method conditional on our Amish pedigree and evaluate the sensitivity of the method to sources of population structure. This was the first study to use a realistic forward genetic simulation model to assess the properties of the proposed regression-based estimator of heritability. We found that the estimator showed a downward bias, and that this bias increased as the expected trait heritability increased. Our results suggest caution when interpreting Zuk *et al.*'s regression-based estimates of the narrow sense heritability.

4.1 Introduction

While geneticists have been working for decades to identify genetic variants underlying trait heritabilities, much of the variance attributed to additive genetic factors has yet to be accounted for, the so-called "missing heritability problem" [9]. To date, genome-wide association studies (GWAS) have identified >9,000 loci associated with >700 quantitative traits and complex diseases [7]. Yet, these loci only explain a small fraction of the heritability for these traits.

Height is a classic example of the missing heritability problem. The heritability of height is estimated to be >80% in most populations, meaning that genetic factors can account for 80% of the phenotypic variation of height [92]. Despite this high heritability, the ~180 loci associated with height, in total, only explain ~12% of the heritability of the trait because each locus has a small effect size [4]. Yang *et al.* suggests that ~45% of the variance of height can be attributed to common variants present on SNP genotyping arrays, although the effects of these variants are individually too small to detect via

GWAS [8]. Even after incorporating aggregate effects of common variants, half of the heritability of height remains unaccounted for. This apparent missing heritability has led scientists to question the validity of the heritability estimates themselves.

One explanation for the missing heritability is that the heritability may have been over-estimated. In other words, identified genetic factors cannot account for the estimated heritability of most traits, resulting in the appearance of missing heritability. The variance of a trait (V_T) can be partitioned into the variance explained by additive and independent genetic factors (V_A), dominant genetic factors (V_D), and environmental factors (V_E). Heritability, in the broad-sense, is the total trait variance explained by all genetic factors, $H^2 = (V_A+V_D)/V_T$. In contrast, the narrow-sense heritability is defined as the proportion of phenotypic variance attributable to only additive genetic effects, $h^2 = V_A/V_T$. Throughout this chapter, ‘heritability’ refers to the narrow-sense heritability, unless otherwise noted. Heritability can be estimated by partitioning the observed phenotypic correlations between relative pairs in a variance components model. However, the variance components model assumes that there is no gene-gene interaction, gene-environment interaction, or gene-environment covariance [93]. In reality, ignoring the presence of interactions or gene-environment covariance can lead to overestimates of the heritability [17-20, 93].

Zuk *et al*. present a new method to estimate h^2 that is not confounded by the effects of gene-gene interactions, gene-environment interactions, and/or gene-environment covariance [21]. Their method involves using only distantly related individuals to estimate h^2. As opposed to close relatives, distant relative pairs share a very small proportion of their genome identical by decent (IBD), and therefore are less likely to have inherited many interacting loci IBD. Thus, using these pairs to estimate h^2 captures the additive genetic effects, free from the confounding of gene-gene interactions. Zuk *et al*. proposes estimating h^2 by regressing the phenotype similarity for pairs of individuals on their genetic similarity [21]. They state that their method is best applied in isolated populations where the mean relationship coefficient is non-zero, and there are many distantly related pairs of individuals from which to estimate h^2. In addition to helping explain the apparent missing heritability, large differences between the regression-based and variance components based h^2 estimates may provide evidence that interactions will account for some fraction of trait variance.

To evaluate the evidence for possible interactions, we set out to apply the regression-based estimate of h^2 to selected traits from our study of mammographic density in the Amish. Our Amish study provides the unique opportunity to compare heritability estimates from the variance components and regression methods using the same dataset and population. Interestingly, we found that the regression-based heritability estimates were, on average, 40% lower than the variance components estimates regardless of the trait. Before interpreting these large differences in h^2 estimates as evidence for pervasive gene-gene interactions, it was necessary to evaluate the statistical properties of the regression-based h^2 estimator using a realistic simulation model. We evaluate the regression-based estimator using simulations conditional on the pedigree structures from our Amish study. We also assess the effects of non-random mating in the population and non-random sampling of study participants on the estimates of h^2. This study is the first to evaluate the properties of the regression-based h^2 estimator proposed by Zuk *et al.* using a realistic model of IBD sharing between study participants.

4.2 Methods

4.2.1 Family-Based Study of Mammographic Density

We applied and evaluated the regression-based estimator of h^2 using data from our family-based study of mammographic density. All of the women in our study come from the Old Order Amish population in Lancaster County, PA. This is an isolated founder population that migrated from Europe in the 1700's. The Amish originated from ~500 founder individuals, although approximately 95% of the current population can be traced back to just 128 founders [28]. Approximately 10-20% of Amish children leave the community each generation [30]. However, the high birth rate among the Amish has resulted in overall population growth [27, 31]. The current Amish population is approximately 30,000-50,000 individuals [28].

Our study includes complete phenotype information on 1,481 healthy Amish women ≥ 40 years old. These women represent nearly half of the current census population of Amish women aged 40-80 years old in Lancaster County, PA. Our study design focused on recruitment of sibships. Initially, women ≥ 40 years old were only eligible for our study if they had a living sister who was also eligible. We later dropped this criterion to recruit all women ≥ 40 years old, regardless of whether they had an eligible sister. Recruitment was conducted primarily through word-of-mouth and door-to-door interviews. Therefore, while we sampled nearly half of the eligible population, our sample is likely not a random sample of the population and is enriched for close relatives, e.g., sisters and first cousins. Further details about our recruitment strategy, including specific inclusion and exclusion criteria, are detailed elsewhere [54].

All women in our study can be connected into a single 13-generation pedigree using the extensive genealogical records from the Anabaptist Genealogical Database (AGDB) [34, 35]. We estimated the heritability of selected quantitative traits (11) from our study of mammographic density, including absolute measures of the dense and non-dense area of the breast, percent density, measures of body size, reproductive, and menstrual traits. Traits were transformed to approximate normality, when necessary. We used a linear regression model to adjust each trait for age and menopausal status and obtained studentized residuals from the regression. This was done with the lm() function in R, version 2.15.2 [94]. The studentized residuals were used to estimate heritabilites.

4.2.2 Estimating the Narrow-Sense Heritability, h^2

Our initial goal was to apply and compare regression-based and variance components-based estimates of h^2 for traits from our Amish study to assess the potential contribution of gene-gene interactions to the missing heritability. We also apply the h^2 estimators to quantitative traits simulated under a purely polygenic model in order to evaluate the potential for bias in the estimators. Both estimators are expected be unbiased when applied to a purely polygenic trait [21, 95]. Applying the variance components estimator in this context allows us to make sure our trait simulations are not biased. We can then assess the statistical properties of the regression-based estimator using these trait simulations.

4.2.2.1 Regression-Based h² Estimate

Zuk *et al.* showed a theoretical proof that the slope of the regression line, β, from the regression of phenotypic similarity on genetic similarity using distantly related pairs of individuals was a consistent estimator of the narrow-sense heritability, regardless of the traits genetic architecture. Because distantly related pairs of individuals share very little of their genomes identical by decent (IBD), they are unlikely to share alleles at interacting loci. Therefore, using only these pairs to estimate h^2, should result in an estimate of h^2 that is not influenced by the effects of gene-gene interactions. The regression only includes pairs of individuals with a relationship coefficient in some small neighborhood, ϵ, of the mean relationship coefficient, $\bar{r}$. Zuk *et al*. proved that $\widehat{h^2} = (1-\bar{r})\hat{\beta}$ is a consistent estimator of the heritability as $\epsilon \rightarrow 0$.

Following the suggestion of Zuk *et al.* we defined distantly related pairs as those with a relationship coefficient, $\bar{r} \pm \epsilon$, where $\epsilon = \bar{r}$. Therefore, the range of relationship coefficients we used was 0 to $2\bar{r}$. The relationship coefficient is defined as the proportion of the genome shared identical by decent (IBD) between a pair of individuals, or twice the pair-wise kinship coefficient. We calculated pedigree-based pair-wise kinship coefficients (and in turn relationship coefficients) using the matrix method described by Lange [66] and implemented in MENDEL version 13.0.0.

We used ordinary least squares regression to model the relationship between the products of the trait values between pairs (phenotypic similarity) on their relationship coefficients (genetic similarity) for distantly related pairs of individuals. Regression was done using the lm() function in R, version 2.15.2 [94].

4.2.2.2 Variance Components Estimate of h^2

We used the variance components model as an alternative method to estimate h^2. The variance components model we used partitions the total trait variance into the variance explained by additive genetic factors and residual unshared environmental factors. In order to accurately account for marriage loops in our pedigree, we estimated the expected probabilities of Jacquard's condensed identity coefficients (described in [66]) using IdCoefs V2.1 [96]. From the 9 identity coefficients, Δ1 – Δ9, we calculate the pair-wise kinship coefficients between all pairs of study participants. In order to estimate the heritability using the variance components model in SOLAR V4.3.1, we need the estimates of the kinship coefficient and Δ7, or the probability a pair of individuals shares 2 alleles IBD. We imported these values into the variance components model implemented in the polygenic function in SOLAR V4.3.1 [85]. For computational reasons, we were only able to estimate h^2 using variance components for the traits simulated based on our Amish pedigree. The pedigrees from our simulated isolated population (described below) were too large and computationally intensive to obtain a variance components-based h^2 estimate. Specifically, the 48GB of RAM on our shared compute cluster nodes was not sufficient to calculate identity coefficients in IdCoefs V2.1 on the full simulated pedigrees. There was also insufficient memory to run SOLAR V4.3.1 on the full simulated pedigrees.

4.2.3 Simulations to Evaluate the Regression-Based Estimator of h^2

4.2.3.1 Simulated Population Isolate

We used forward genetic simulations to create a set of related individuals from which to draw a study sample to evaluate the regression-based estimator of h^2. To simulate an isolated founder population, we started with a set of 100 unrelated founders, and then simulated mating forward in time for 14 non-overlapping generations. At each generation, we allowed only 50% of the population to successfully reproduce to generate offspring into the next generation. Each pair of individuals that successfully reproduces has N offspring, were N ~Poisson (6), meaning the average number of offspring is 6. We choose these parameters to create a final population size that was similar to the current-day Amish population [30]. Each simulation results in a 14-generation pedigree connecting all individuals in the population to the original 100 founders. All of our simulations were done in R version 2.15.2 [94].

4.2.3.2 Quantitative Trait Simulations with a Polygenic Model

In order to evaluate the bias of the regression-based h^2 estimator, we simulated a purely polygenic quantitative trait using the genedrop function in MORGAN [67]. Specifically, we simulated a polygenic trait by sampling from a multivariate normal distribution. We repeatedly simulated a quantitative trait with h^2 equal to 20, 40, 60, or 80% of the total trait variance. We chose these simulation parameters to test the full range of estimated heritablities from our real data (Table 4.2). To evaluate the performance of the regression-based estimator in the context of our mammographic density study, we simulated 1,000 replicate traits for each heritability value tested, conditional on our Amish pedigree. To evaluate the effects of non-random mating or non-random sampling, we simulated 10 study samples per sampling scheme (described below) and then simulated 100 replicate traits per study sample to result in 1,000 simulations per heritability.

We standardized each trait to have a mean of 0 and variance of 1 before estimating h^2. We then estimated the h^2 of the trait using the regression-based method described by Zuk *et al*. [21] and in section 4.2.2.1 above. We truncated our h^2 estimates at 0 and 1 if they were estimated to be < 0 or > 1, respectively. We determined the bias by subtracting the mean h^2 estimate from the expected (simulated) h^2. Since we simulated an additive trait using the polygenic model, we expect the estimated h^2 to be equal to the simulated value. The bias was significantly different from 0 if the mean h^2 estimate ± 1.96 times the standard error of the mean, i.e. the 95% confidence interval around the mean, did not include the expected value. In other words, we used a p-value threshold 5% to determine significance.

4.2.3.3 Random Mating, Random Sampling

Using the scheme described above, we simulated an isolated population with random mating. To select a study sample, we then randomly chose 1,500 individuals from the final 2 generations. This was meant to mimic the current-day population. In our results, we refer to this as the ‘Random’ sample. We simulated 10 replicate study samples. For each study sample, we trimmed the pedigree in MENDEL version 13.0.0 [97] to include only the individuals necessary to relate the 1,500 study participants. We calculated kinship coefficients based on the pedigree using MENDEL version 13.0.0 [66].

4.2.3.4 Random Mating, Non-Random Sampling

From our randomly mating simulated populations, we also selected a non-random study sample to mimic our Amish study recruitment strategy. We selected a study sample (N = 1,500) from the final 2 generations of the simulate population based on sibships. We sampled sets of siblings in proportion to our observed distribution of sibship sizes for our Amish study (Table 4.1). In our results, we refer to this as the 'Siblings' sample.

To create a more extreme non-random sampling scheme, we also selected a study sample around probands. The goal of our proband sampling scheme was to be an extreme case of the word-of-mouth recruitment we used in our Amish sample. The Amish live in small church districts of ~12 closely related families [98]. Thus, the word-of-mouth recruitment in our density study likely led to recruiting sets of close relatives. To create our study sample, we selected an individual, the proband, at random from the final 2 generations of the simulated population, and then sampled all of the proband's relatives out to 2nd cousins. We repeated this process until we had a study sample of 1,500 individuals. If the sample was over 1,500 individuals, we randomly removed the extra individuals from the sample. In our results, we refer to this as the 'Probands' sample.

We created 10 replicate study samples each for the 'Siblings' and 'Probands' simulations. For each of the resulting study samples, we trimmed the pedigree and calculated kinship coefficients in MENDEL version 13.0.0 [66, 97].

4.2.3.5 Non-Random Mating, Random Sampling

Finally, we used forward genetic simulations, with the parameters described above, to simulate a structured population with non-random mating. We did this by simulating 1 generation of random mating between the 100 founders, followed by 13 generations of non-random mating. Specifically, after the first generation, we split the population into 2 subpopulations and only allowed random mating within each subpopulation for the subsequent 13 generations. We then selected a study sample (N = 1,500) at random from the final 2 generations of the simulated population. Similar to the previous simulations, we trimmed the pedigree and calculated kinship coefficients in MENDEL version 13.0.0 [66, 97] for 10 replicate study samples. In our results, we refer to this as the 'Structured' sample.

4.3 Results

4.3.1 Pedigree Characteristics

The distribution of relationship coefficients, *r*, between all pairs and distantly related pairs of individuals from our Amish study is shown in Figure 4.1. Our Amish sample had a mean relationship coefficient of 0.0719. Therefore, on average, any two Amish individuals were more closely related than 2nd cousins but less than 1st cousins, once removed. We classified 98.4% of pairs as distantly related and included them in the regression-based estimates of h^2 (Figure 4.1B, Table 4.3). The distribution of sibship sizes is shown in Table 4.1. On average, the sibships in our Amish study contained 2 individuals, and 76% of women had at least one sister in the study. This is consistent with our emphasis on recruiting sisters early in the study.

For each of our simulated study scenarios, the distribution of relationship coefficients for 1 of the 10 replicates is shown, as a representative example, in Figure 4.2 - Figure 4.5. Summary statistics of the relationship coefficient distribution for the representative simulated study samples are shown in Table 4.3. Details for all replicate study samples are shown in Appendix 6.

The mean relationship coefficient, $\bar{r}$, for all pairs of women was between 0.0557 – 0.0668 for our simulated study samples. For the randomly mating simulations, as expected, the 'Random' sample had the lowest mean relationship coefficient because this sample was not selected to enrich for close relative pairs (Table 4.3). As with our Amish study, greater than 90% of pairs from our simulated pedigrees had a relationship coefficient between 0 and $\bar{r}$, and were used in the regression-based estimates of h^2 (Table 4.3).

The sibship size distribution from the 'Siblings' sample is similar to our Amish sibship size distribution, as intended (Table 4.1). The mean sibship size was 2 for both the 'Siblings' sample and the Amish sample. Not surprisingly, the 'Proband' sample had the largest mean sibship size (~5). This is consistent with the sampling of all close relatives of the proband in this sampling scheme. For our 10 'Proband' samples, we selected, on average, 5 probands, and ~319 close relatives (out to 2nd cousins) around each proband to generate a sample of 1,500 individuals.

The non-randomly mating, randomly sampled (Structured) simulated sample had a mean relationship coefficient similar to that of the randomly mating 'Random' sample, 0.0557 and 0.0608 respectively (Table 4.3). However, the overall distribution of the pair-wise relationship coefficients from these two simulated samples is very different (Figure 4.5 and Figure 4.2, respectively). The distribution of *r* from the 'Structured' sample has 3 peaks (Figure 4.5). The 2 smaller peaks, centered at 0.15 and 0.07, correspond to the within group mean relationship coefficients in the 2 subpopulations (Figure 4.5). By chance, one subpopulation was smaller than the other, and therefore had a larger within group mean relationship coefficient. The large peak in the relationship coefficient distribution at 0.02 corresponds to the pairs of individuals between subpopulations (Figure 4.5). Because of the 1st generation of random mating, followed by separation for 13 generations, these pairs are more distantly related than the within group pairs, and therefore are expected to have lower relationship coefficients.

4.3.2 Heritability Estimates: Amish Study Sample

We estimated the narrow-sense heritability using the variance components model and the regression-based method for 11 quantitative traits from our Amish study of mammographic density. The variance components h^2 estimates ranged from 0.45-0.83. In contrast, the regression-based estimates ranged from 0.03-0.47. All of the traits had h^2 estimates significantly greater than 0 using the variance components model ($p < 0.05$, Table 4.2). The regression-based h^2 estimates were significantly different from 0 for all traits, except BMI and age at natural menopause ($p < 0.05$, Table 4.2). On average, the regression-based h^2 estimates were 40% lower than the variance components estimates (Table 4.2).

Because we observed consistently lower regression-based h^2 estimates for the traits from our Amish study, we used simulations to evaluate the potential bias of the regression-based h^2 estimator conditional on the Amish pedigrees. Our simulations showed that regardless of the simulated h^2 of the quantitative trait (20-80%), the regression-based estimate was ~60% lower than expected. For example, when we simulated a polygenic trait with h^2 = 40% conditional on our Amish pedigree, the mean regression-based h^2 estimate was 16% (standard deviation = 11%) (Figure 4.6). In contrast, when we estimated h^2 for our simulated traits using the variance components model, the mean estimate (from 1,000 replicate simulations) was not significantly different from expectation. Specifically, for each mean h^2 estimate using the variance components method, the 95% confidence interval around the mean included the expected simulated value (Figure 4.6). As expected, the variance components model was unbiased under additivity. The downward bias of the regression-based h^2 estimator, conditional on the pedigree from our Amish study, is consistent with the decreased h^2 estimates seen in our real data (Table 4.2).

4.3.3 Heritability Estimates: Simulated Study Samples

The regression-based h^2 estimator was biased downward for the simulated polygenic quantitative traits conditional on our randomly mating, randomly sampled (Random) study sample when h^2 > 20% (Figure 4.7). When h^2 = 20%, there was no significant bias. However, as the simulated h^2 increased, the bias increased. For example, when the simulated h^2 was 80%, the estimated h^2 was 70.5% (standard deviation = 25%), or about 12% lower than expected. This slight downward bias, even with the randomly mating, randomly sampled simulated study sample, was consistent regardless of sample size. For example, when we doubled the study sample to 3,000 individuals, the regression-based h^2 estimates were still, on average, 10% lower than expected when the trait had a simulated h^2 of 80% (data not shown).

The downward bias of the regression-based h^2 estimate was larger when we simulated a randomly mating, non-randomly sampled study sample. Again, the bias was only seen when the simulated h^2 exceeded 20%. For example, with a simulated trait with h^2 = 80%, the mean regression-based estimate of h^2 for the 'Siblings' and 'Probands' samples were 66% (standard deviation = 31%) and 56% (standard deviation = 38%), respectively (Figure 4.7). These estimates are 18% and 30% lower than expected.

Not surprisingly, the h^2 estimates from our non-randomly mating, randomly sampled simulations (Structured) had the largest downward bias of our simulated study samples. As with our other simulations, the bias increased with h^2. For example, the mean regression-based h^2 estimate was 8% lower than expected when the trait was simulated with h^2 = 20%, and 42% lower than expected when h^2 was set to 80% (Figure 4.7). For a trait with simulated h^2 = 20%, the bias was, however, not significantly different from 0 (p-value > 0.05).

4.4 Discussion

The goal of this study was to apply a new regression-based estimator of h^2 to quantitative traits from our Amish study of mammographic density. We set out to compare h^2 estimates from this new regression-based method, not confounded by gene-gene interactions, with the variance components estimates of h^2, in order to evaluate the potential role of gene-gene interactions in accounting for the missing heritability. Motivated by our finding that the regression-based h^2 estimates were consistently lower than the variance components-based estimate by ~40%, we used simulations to assess the potential bias in the h^2 estimators under a purely polygenic model.

We comprehensively evaluated the bias of the estimator, conditional on the Amish pedigree from our mammographic density study and simulated pedigrees with varying degrees of population structure. Surprisingly, the regression-based h^2 estimator consistently underestimated the narrow-sense heritability when applied to a quantitative trait simulated under a polygenic model. The bias became more pronounced as the simulated trait heritability increased. The downward bias was highest in our trait simulations conditional on our Amish pedigree; estimates were on average 60% lower than expectation. Our simulations confirmed that the lower heritability estimates seen for our real data are likely an artifact of the estimation method, and therefore should be interpreted with caution.

Unexpectedly, we found a small but significant bias in the regression-based h^2 estimates with our randomly mating, randomly sampled simulations. These simulations were designed to represent an ideal random sample from a randomly mating population. However, even in this best-case scenario, there was a downward bias in the h^2 estimates. In their original description of the regression-based estimator, Zuk *et al*. performed a limited evaluation of the consistency of the estimator using simulations [21]. To create a study sample of related individuals, they generated a set of 1,000 individuals with chromosomes generated as a mosaic of 28 founders. Their approach did not fully model mating over many generations to create an isolated population as our, more realistic, simulations did. They simulated a quantitative trait using their limited pathway model of interactions with narrow-sense heritability of 25%. Their results showed the estimator to be unbiased in this scenario. These results are similar to our observation that the estimator is unbiased at a heritability of 20%. However, they did not evaluate and report a broader range of heritabilities so we could not make further comparisons to the bias we observed as the simulated h^2 increased.

Compared to those of Zuk *et al.*, the simulation models we used are very different in terms of how we simulated a study sample and quantitative trait. While both methods resulted in a set of distantly related individuals with quantitative trait values to estimate the narrow-sense heritability, the distribution of relationship coefficients was quite different. The mean relationship coefficient in the simulations conducted by Zuk *et al.* was 3.6%, whereas in our randomly mating, randomly sampled (Random) simulation the mean relationship coefficient was 6.1%. Therefore, when defining distantly related pairs as those with a relationship coefficient between 0 and $2\bar{r}$, our range was much larger. Specifically, the range of relationship coefficients used by Zuk *et al.* was from 0 to 7.2%, whereas the range for our 'Random' sample was 0 to 12.2%. However, Zuk *et al*. showed that the regression-based estimator is consistent as the neighborhood of relationship coefficients around the mean goes to 0. Interestingly, we observed a similar downward bias of the heritability estimates for the 'Random' sample even when we used a smaller range of relationship coefficients in the regression. For example, when we simulated a trait with h^2 equal to 80%, the mean regression-based estimate was 71% when the range of relationship coefficients was from 0 to 12.2% and 68% when the range was decreased to 3..1% to 9.2% (Appendix 7).

The downward bias we observed in our non-randomly mating simulated study sample (Structured) was consistent with our expectations. The regression-based method assumes that the study sample is selected from a randomly mating population. Our structured population was, by design, not randomly mating. These results are consistent with previous observations of downward bias when applying the method to a study sample from Northern Finland, an isolated population with well documented population structure [99]. As observed with our 'Structured' population, the higher relationship coefficients within subpopulations and lower relationship coefficients between subpopulations drive the regression-based heritability estimates toward the null hypothesis of 0, and therefore create an overall downward bias. Using our simulated data, we also showed that when we split the study sample into homogenous subgroups using only pairs of individuals from the same subpopulation, the bias in the h^2 estimator is reduced and similar to that seen with the randomly mating, randomly sampled study sample (data not shown).

We also observed a downward bias, however to a lesser extent, in our randomly mating, but non-randomly sampled study samples (Siblings and Probands). We suspect that this bias comes from the small amount of population structure induced by our sampling scheme. By over-sampling close relative pairs, we created groups within the study sample that are more closely related to each other than individuals from other groups, similar to the subpopulations seen with our structured study sample. The

‘Probands’ sample was more biased than the ‘Siblings’ sample, likely because of the inclusion of many more close relatives in the sample. Therefore, regardless of detectable non-random mating in the source population, or non-random sampling of study participants, the regression-based h^2 estimator increasingly underestimates h^2 as the expected value of h^2 increases. In the case of our Amish study sample, the observed bias could be caused by both non-random mating and non-random sampling.

The regression-based method to estimate h^2 was originally proposed for use with observed IBD sharing estimated from genome-wide marker data [21]. For this study, we use the expected, instead of observed, IBD sharing determined by the pedigree relationship between pairs of study participants. In our forward genetic simulations, the use of expected relationship coefficients dramatically reduces the computational complexity of the simulation model. In our Amish study, we found that pedigree-based and genome-based estimates of the relationship coefficient were largely consistent (data not shown), and therefore we expect h^2 estimates to agree. If IBD sharing can be estimated accurately, using the observed sharing between individuals to calculate h^2 should result in more precise h^2 estimates. Thus, while the use of expected relationship coefficients increases the overall error of the h^2 estimate, it should not induce the downward bias we observed in our simulations.

Genome-wide association studies have identified 1000’s of variants associated with 100’s of traits, however, for most complex traits, much of the heritability remains unexplained [4, 10]. A possible explanation for this so-called “missing heritability” problem is that the narrow-sense heritability of many traits has been overestimated because of the presence of gene-gene interactions, gene-environment interactions, and/or gene-environment covariance. While statistical methods are typically underpowered to identify interacting loci in human studies of complex traits [22, 23], observations in model organisms suggest that interactions may play a large role in the heritability of most traits [24-26]. In order to evaluate the extent to which the heritability of traits in our own study of mammographic density have been overestimated, we sought to apply a newly proposed regression-based method to estimate h^2. In theory, the regression-based method should yield an unbiased estimate of h^2, regardless of the underlying genetic architecture. However, we found that when applied to data simulated under a polygenic model, the estimator is biased downward. While these results may be specific to our Amish pedigree and/or the parameters of our simulation model, they do warrant caution when interpreting the results based on this method. In the future, it will be important to use simulations to evaluate the potential for bias in specific study samples before applying the regression-based method to estimate trait heritabilities.

Table 4.1: Distribution of sibship sizes for our Amish sample and simulated samples
The distribution of sibship sizes for our density study (Amish, N = 1,481) and our simulated study samples (N = 1,500): random mating, random sampling (Random), random mating, non-random sampling (Siblings and Probands), and non-random mating, random sampling (Structured).

	Frequency (%)				
Sibship Size	**Amish**	**Siblings**[1]	**Random**[1]	**Probands**[1]	**Structured**[1]
1	356 (48)	349 (47)	1159 (88)	6 (2)	1099 (85)
2	197 (26)	207 (28)	151 (11)	19 (7)	173 (13)
3	100 (13)	104 (14)	12 (1)	36 (13)	17 (1)
4	52 (7)	50 (7)	1 (<1)	46 (16)	1 (<1)
5	25 (3)	25 (3)	-	49 (18)	0.2 (<1)
6	10 (1)	11 (2)	-	44 (16)	-
7	3 (<1)	3 (<1)	-	35 (12)	-
8	1 (<1)	1 (<1)	-	19 (7)	-
9	1 (<1)	1 (<1)	-	13 (5)	-
10	-	-	-	5 (2)	-
11	-	-	-	4 (2)	-
12	-	-	-	2 (1)	-
13	-	-	-	1 (<1)	-
14	-	-	-	0.2 (<1)	-
15	-	-	-	0.2 (<1)	-

[1] Distribution is the mean over 10 replicate simulations

Table 4.2: Comparison of heritability estimates for selected traits from our study of mammographic density

We transformed traits to approximate normality, and adjusted for age and menopausal status. We estimated h^2 using the studentized residuals after covariant adjustment. Both methods incorporate the full 13-generation Amish pedigree to determine pair-wise relationships. Except where noted, all traits had an estimated narrow-sense heritability that was significantly greater than 0 (p-value < 0.05). The sample size for all traits was 1,481 except where noted.

	h^2 Estimate (S.E.)		Ratio of h^2 Estimates:
Trait	Variance Components	Regression[c]	Regression/Variance Components
Dense Area	0.49 (0.07)	0.20 (0.05)	0.41
Percent Density	0.45 (0.06)	0.20 (0.05)	0.44
Nondense Area	0.72 (0.06)	0.31 (0.05)	0.44
Age at Menarche	0.56 (0.07)	0.21 (0.05)	0.37
Nat Age at Meno[a,b]	0.47 (0.11)	0.03 (0.10)	0.06
Height	0.83 (0.06)	0.40 (0.05)	0.48
Weight	0.60 (0.07)	0.17 (0.05)	0.28
BMI[b]	0.50 (0.07)	0.09 (0.05)	0.18
Waist	0.51 (0.06)	0.29 (0.05)	0.57
Hip	0.56 (0.07)	0.12 (0.05)	0.22
WHR	0.52 (0.06)	0.47 (0.05)	0.89

[a] Sample size for Nat Age at Meno = 671

[b] The regression-based estimates of h^2 for Nat Age at Meno (p = 0.79) and BMI (p = 0.053) were not significantly different from 0.

[c] The regression-based estimates of h^2 included pairs of individuals with a relationship coefficient between 0 and $2\bar{r}$, where $\bar{r}$ is the mean relationship coefficient in the full sample

Table 4.3: Summary statistics of the relationship coefficient (r) distribution for our Amish sample and simulated samples
Summary statistics for all pairs (All) and distantly related pairs (Distant) for each study sample. Summary for the Amish (N = 1,481) and simulated study samples (N = 1,500): random mating, random sampling (Random), random mating, non-random sampling (Siblings and Probands), and non-random mating, random sampling (Structured). A pair of individuals is defined as distantly related if their relationship coefficient, $r = [0,2\bar{r}]$, where $\bar{r}$ is the overall mean relationship coefficient. Only distantly related pairs were included in the regression-based h^2 estimates.

	Number of Pairs		Mean r (SD)		Range of r	
Sample	**All**	**Distant**	**All**	**Distant**	**All**	**Distant**
Amish	1,095,940	1,078,528 (98.4%)	0.0719 (0.031)	0.0693 (0.019)	0.0009 - 0.5990	0.0009 - 0.1437
Random[a]	1,124,250	1,121,585 (99.8%)	0.0608 (0.011)	0.0604 (0.005)	0.0523 - 0.7870	0.0523 - 0.1216
Siblings[a]	1,124,250	1,120,584 (99.7%)	0.0613 (0.019)	0.0604 (0.005)	0.0530 - 0.6130	0.0530 - 0.1225
Probands[a]	1,124,250	1,103,551 (98.2%)	0.0668 (0.037)	0.0628 (0.010)	0.0536 - 0.6031	0.0536 - 0.1329
Structured[a]	1,124,250	1,033,846 (92.0%)	0.0557 (0.038)	0.0475 (0.025)	0.0167 - 0.6264	0.0167 - 0.1112

[a]Values for our simulated study samples are based on a single representative sample of the 10 replicates (corresponding to Figure 4.2 - Figure **4.5**). Details for each replicate are given in Appendix 6.

Figure 4.1: Distribution of pair-wise relationship coefficients from our study of mammographic density

The distributions of relationship coefficients, *r*, for all pairs (panel A) or distantly related pairs (panel B) from our Amish study of mammographic density. The dashed line shows the mean pair-wise relationship coefficient, $\bar{r}$, based on all pairs of women.

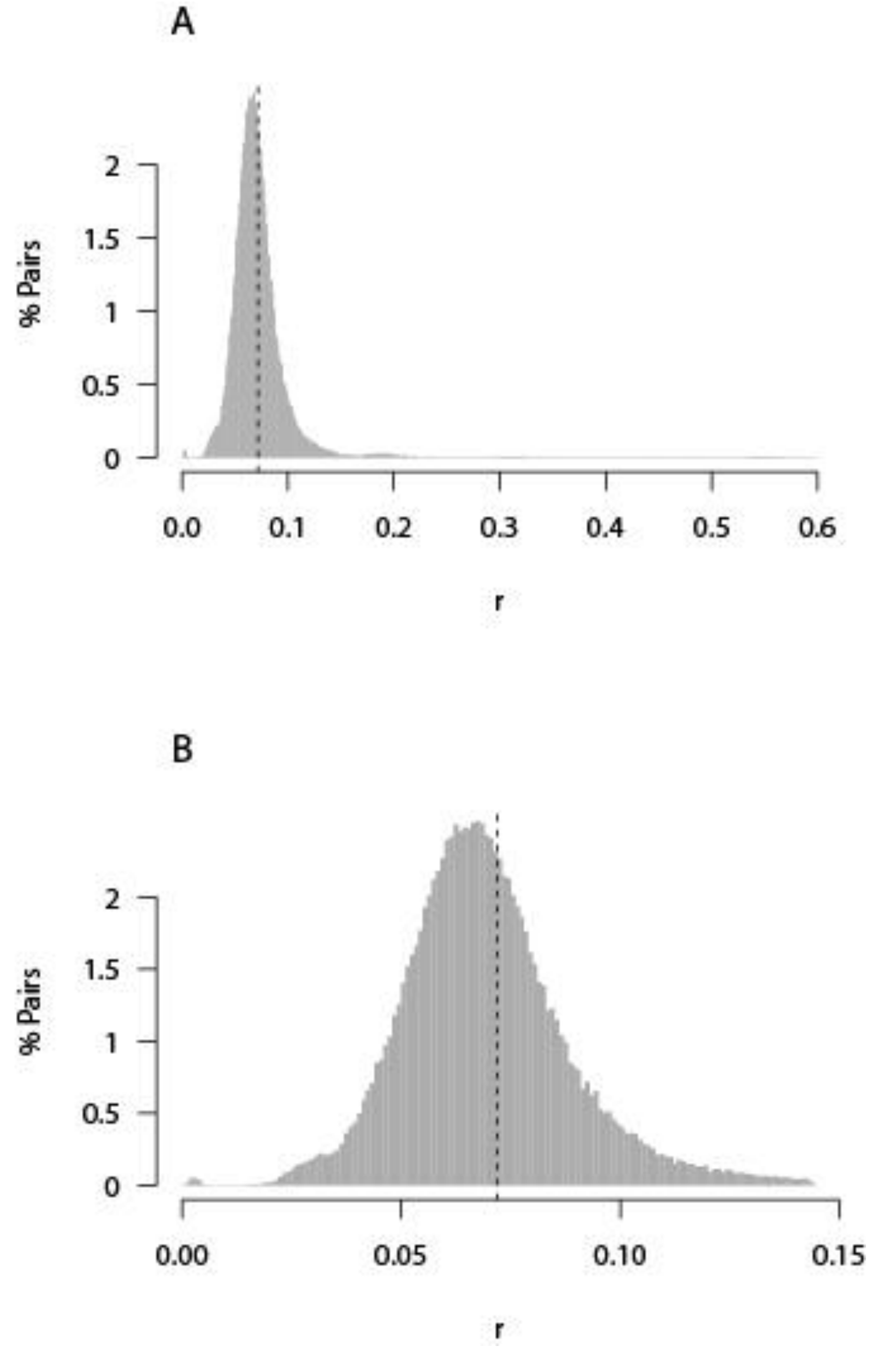


Figure 4.2: Distribution of pair-wise relationship coefficients for the simulated 'Random' sample

Representative distributions of relationship coefficients, *r*, for all pairs (panel A) or distantly related individuals (panel B) from our randomly mating, randomly sampled study sample (Random). The dashed line shows the mean pair-wise relationship coefficient, $\bar{r}$, based on all pairs of individuals.

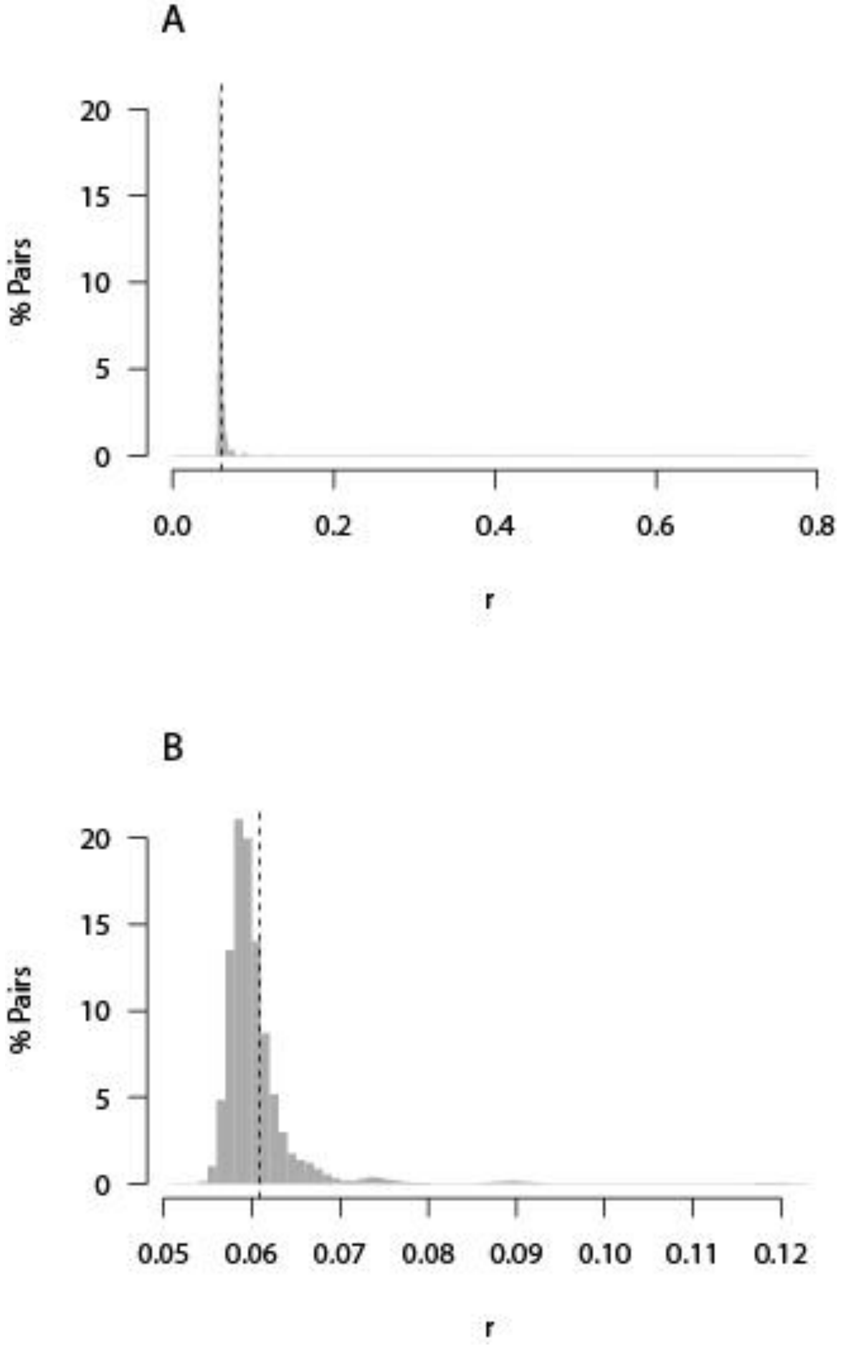


Figure 4.3: Distribution of pair-wise relationship coefficients for the simulated 'Siblings' sample

Representative distributions of relationship coefficients, *r*, for all pairs (panel A) or distantly related individuals (panel B) from our randomly mating study sample, sampled around sibships (Siblings). The dashed line shows the mean pair-wise relationship coefficient, $\bar{r}$, based on all pairs of individuals.

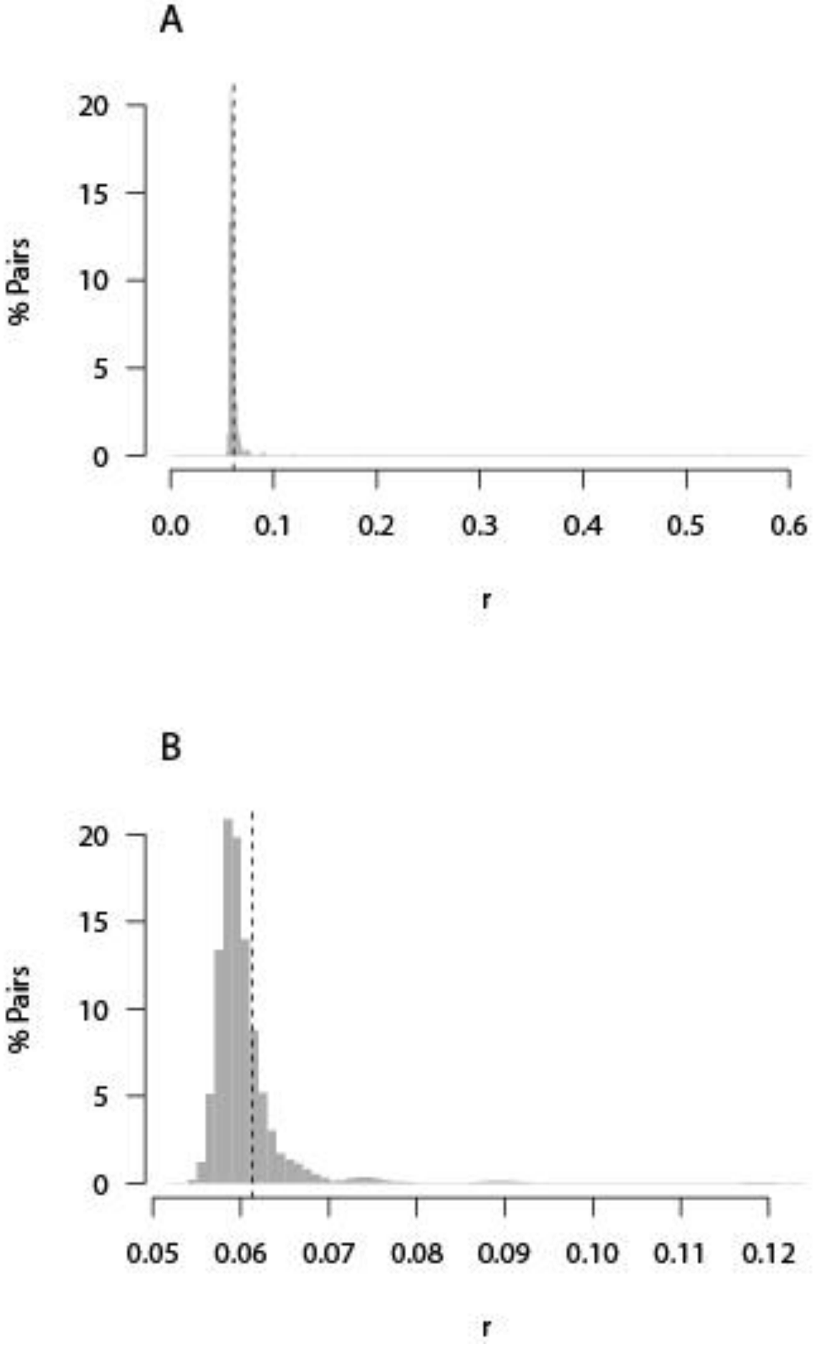


Figure 4.4: Distribution of pair-wise relationship coefficients for the simulated 'Probands' sample

Representative distributions of relationship coefficients, *r*, for all pairs (panel A) or distantly related individuals (panel B) from our randomly mating study sample, sampled around probands (Probands). The dashed line shows the mean pair-wise relationship coefficient, $\bar{r}$, based on all pairs of individuals.

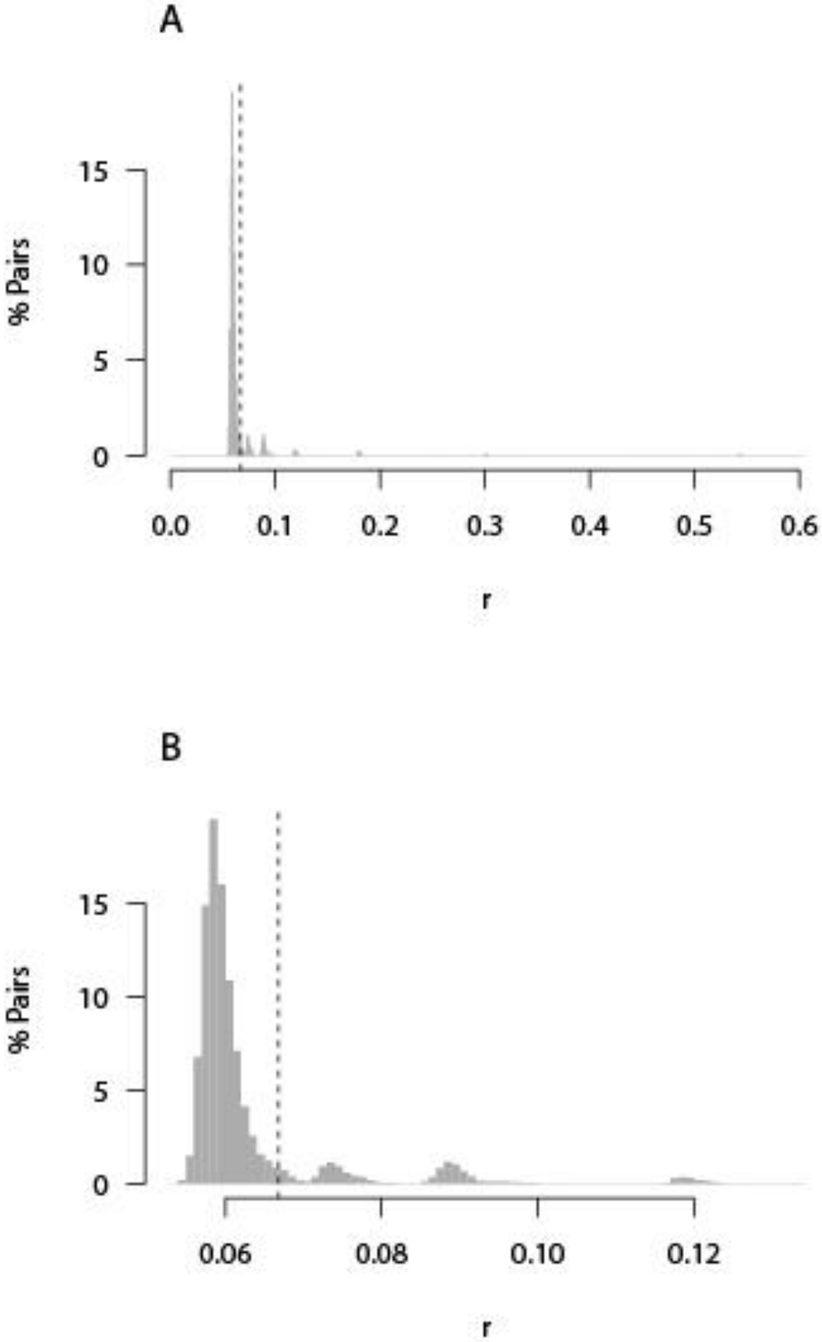


Figure 4.5: Distribution of pair-wise relationship coefficients for the simulated 'Structured' sample

Representative distributions of relationship coefficients, *r*, for all pairs (panel A) or distantly related individuals (panel B) from our non-randomly mating, randomly sampled study sample (Structured). The dashed line shows the mean pair-wise relationship coefficient, $\bar{r}$, based on all pairs of individuals.

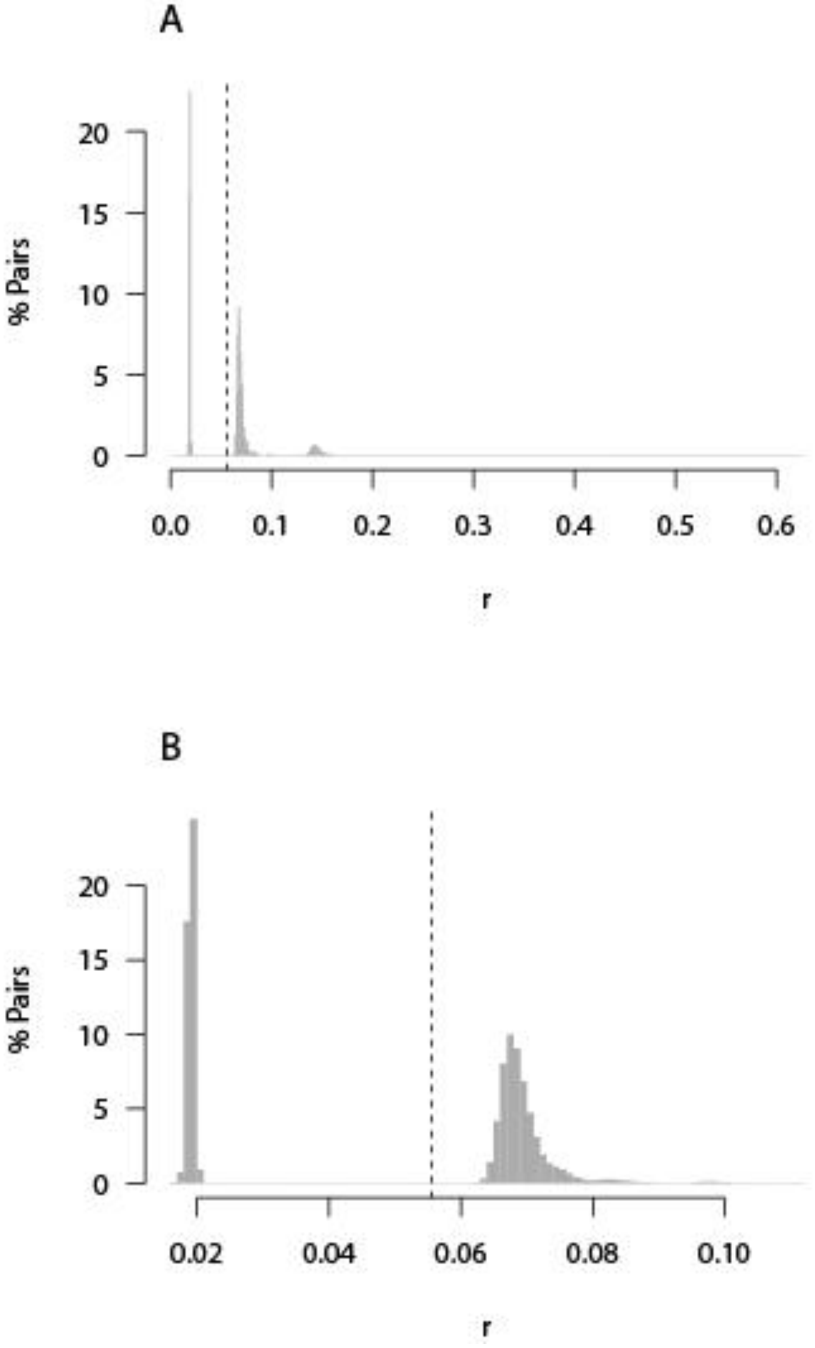


Figure 4.6: Variance components- and regression-based h^2 estimates conditional on the pedigree from our Amish study of mammographic density

A comparison of the mean heritability estimates over the 1,000 replicate simulations of a purely polygenic quantitative trait for each simulated trait heritability using the variance components model and regression method. The error bars represent the 95% confidence interval around the mean.

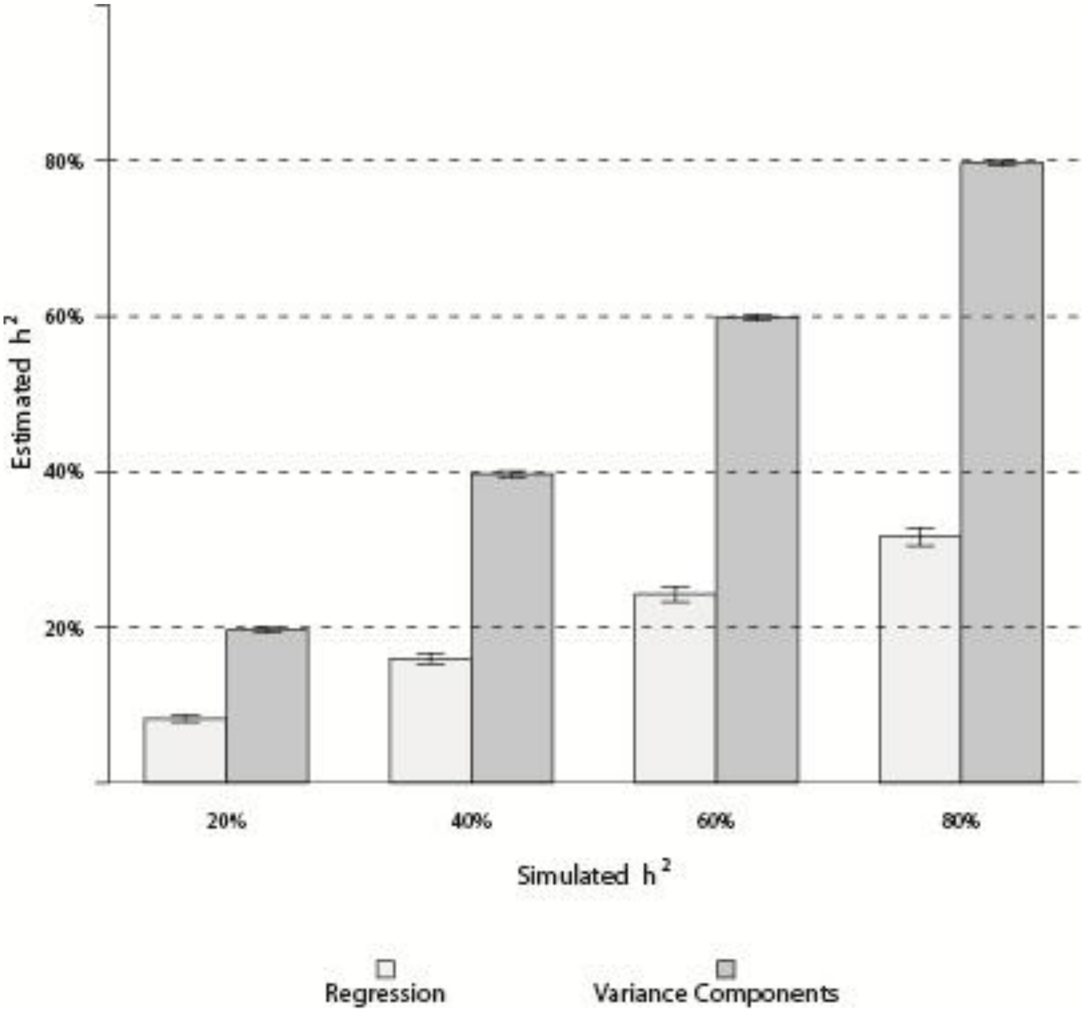
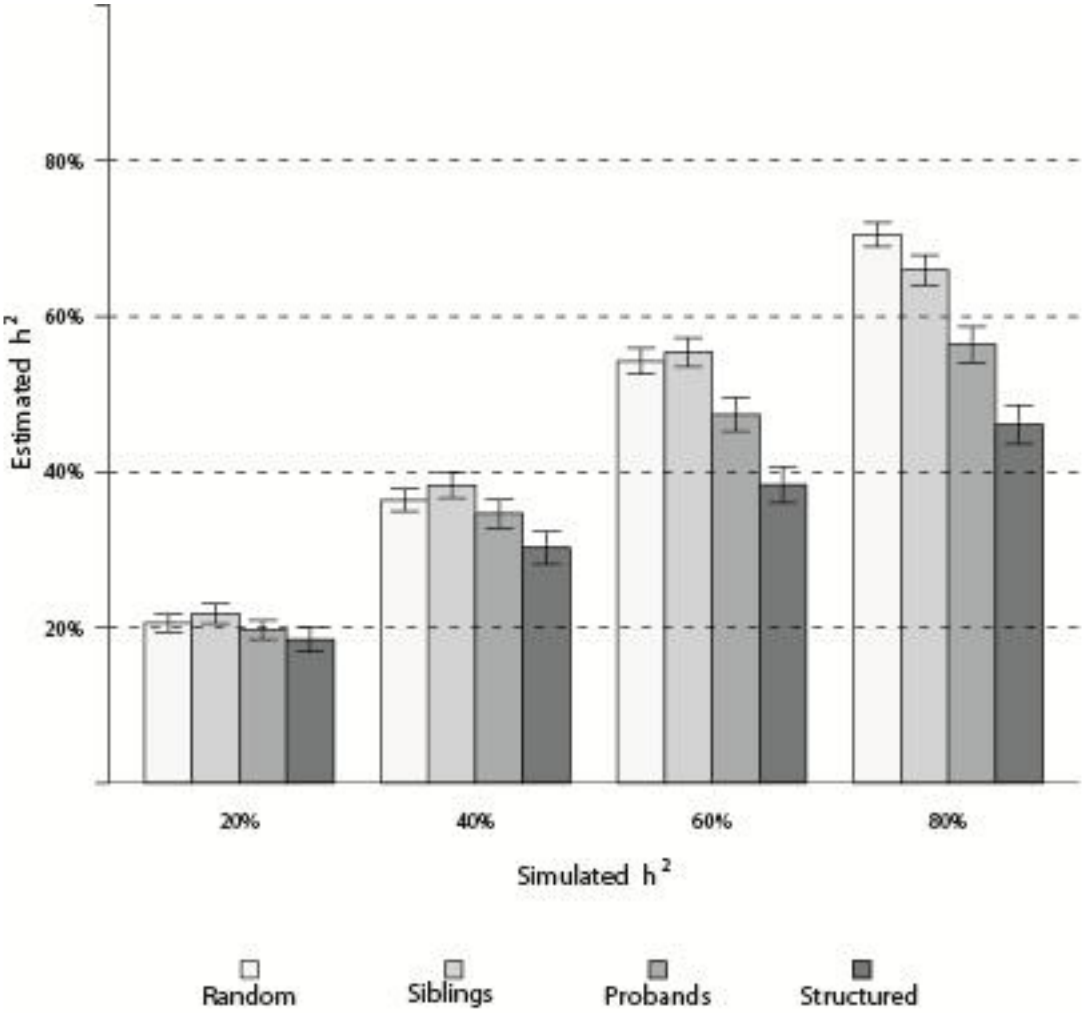


Figure 4.7: Regression-based h^2 estimates for simulated study samples
The mean heritability estimate using the regression-based method over the 1,000 replicate simulations for each sampling scheme (Random, Siblings, Probands, or Structured). The error bars represent the 95% confidence interval around the mean.



Chapter 5

Conclusions

After nearly a decade of genome-wide association studies, with thousands of identified loci for hundreds of traits, it is clear that the genetic architecture of complex traits is more complicated than previously anticipated. While most traits are heritable, finding the individual genetic factors that underlie this heritability has proven to be a challenge. The missing heritability will likely be accounted for by a combination of common variants with smaller effects, rare variants, structural variants, gene-gene interactions, and gene-environment interactions. The studies in this dissertation aimed to develop and evaluate strategies to assess possible sources of the missing heritability by taking advantage of the relatedness of individuals in family-based studies. In Chapters 2 and 3, I evaluated the potential contribution of rare variants to the missing heritability of quantitative traits. In Chapter 4, I evaluated the statistical properties of a new regression-based heritability estimator with the intention of assessing the contribution of gene-gene interactions to the missing heritability of quantitative traits.

5.1 The Contribution of Rare Variants to the Missing Heritability

Large scale DNA sequencing projects of thousands of individuals worldwide have discovered many more rare variants than previously anticipated [100-102]. Each one of us may carry ~10,000 rare non-synonymous variants that alter protein sequence [100]. However, the phenotypic consequence of these variants is largely unknown. Gene mapping efforts will allow us to better understand how specific variants relate to variation in phenotypes between individuals. The goal of Chapter 2 and 3 was to assess the possible contribution of rare variants to the missing heritability of quantitative traits.

In Chapter 2, I developed and evaluated a novel statistical test, the rare variant kinship test (RVKT), to prioritize traits likely to be influenced by a rare variant, and are thus likely to be successful in an extreme trait sequencing study. The test identifies traits with an excess of closely related individuals in the tail of the quantitative trait distribution relative to the expected relatedness under a polygenic model. While this test was underpowered, I found evidence for a rare variant influencing 4 of the quantitative traits from two of our Amish studies. One of the traits, triglyceride levels in the HAPI Heart study, served as a proof of principle. A mutation in the *APOC3* gene associated with lower triglycerides was already identified in this study [45]. The remaining 3 traits identified using the RVKT: prolactin, free estradiol, and estradiol, are now prime candidates for extreme trait sequencing studies. In the future, sequencing the individuals in the tails of the trait distribution for these traits may identify rare, trait-influencing variants. Therefore, there is the potential to further dissect the heritability of these traits. While I had 4 significant results, for the remaining 72 traits I tested across both Amish studies had little evidence for the presence of a rare variant of large effect. Therefore, for these traits, in this population, rare variants of large effect are unlikely to account for the remaining missing heritability.

In Chapter 3, I conducted linkage analysis for a number of traits related to breast cancer risk and identified some of the major challenges of conducting linkage with complex pedigrees and dense SNP marker maps. In general, I found that current Markov-chain Monte Carlo (MCMC) methods do not converge on stable estimates of IBD sharing probabilities in our study. Furthermore, this led to a high degree of variability in LOD score estimates from linkage analyses. As a starting point to conduct linkage analysis despite the highly variable LOD score estimates, I used the average LOD score over many MCMC runs as our measure of evidence for linkage. Using this approach I identified candidate regions for a number of traits, however, very few traits had genome-wide significant evidence for linkage (LOD > 3.3). Regardless, the identified linkage peaks may be a useful starting place to filter variants identified by DNA sequencing in the future.

As dense SNP markers maps become commonplace for linkage analysis, it will be important to continue work in this area to improve methods for linkage analysis. A key next step to improving MCMC methods will be to better understand why the linkage results were so variable. To do this I will need to identify the individuals and families that had largest differences in local kinship estimates between MCMC runs. By knowing specifically the types of pedigree configurations and relative pairs that are challenging for MCMC methods, I can start to explore potential solutions to the problem. One potential way to improve MCMC convergence is by optimizing the multiple-meiosis sampler [79]. Unfortunately, at the time that our linkage analysis was conducted, the software implementation of the multiple-meiosis sampler was not available. However, in the future, I can optimize the multiple-meiosis sampler to sub-sample the types of relative pairs that are most challenging for MCMC methods. For example, the multiple-meiosis sampler currently has the option to sample a full sibship or 3-generation pedigree simultaneously [79].

While the results of Chapters 2 and 3 are not surprising, they may be useful for guiding future gene mapping efforts. It is an exceptional case to identify a rare variant that has a large effect on a complex trait. While there are few known examples in the Amish [39, 45, 48], the work in these two chapters shows that these are likely the exception, not the rule. It is therefore important to gain evidence of the presence of a rare, trait-associated variant prior to conducting expensive DNA sequencing studies. The RVKT described in Chapter 2 allows users to assess the evidence for a rare variant of large effect without conducting expensive genotyping or sequencing studies. With the addition of genome-wide marker data, linkage analysis too provides evidence for rare variants of large effect in a relatively cost effective manner. Therefore, the methods developed and evaluated in Chapters 2 and 3 of this dissertation provide valuable insights into the genetic architecture of quantitative traits. For most of the traits from our Amish studies, it is clear that there are unlikely new rare variants of large effect to be discovered.

5.2 The Contribution of Gene-Gene Interactions to the Missing Heritability

Our current knowledge about the operation of biological systems in pathways suggests that gene-gene interactions may play a role in explaining the missing heritability. While genome-wide interaction studies are difficult to undertake in humans, in model organisms it is clear that up to half of the heritability of many gene expression quantitative traits can be explained by gene-gene interactions [24]. Therefore, in Chapter 4 of this dissertation I set out to evaluate the overall contribution of gene-gene interactions to selected quantitative traits from our study of mammographic density. To do so, I planned on comparing heritability estimates from a variance components model, known to be inflated in the presence of the interactions, and a regression-based approach that is not confounded by the presence of interactions. In the process, I discovered that the regression-based estimator of heritability had a downward bias for even a purely polygenic trait, meaning no gene-gene interactions. Therefore, I could not interpret differences between heritability estimates as evidence for the influence of gene-gene interactions.

While the results of Chapter 4 were unable to provide insights into the contribution of gene-gene interactions to the missing heritability, the work raises many other questions about the utility of the newly proposed regression-based heritability estimator. It is clear that the estimator is highly sensitive to population structure; however, in the future it will be important to identify a set of population parameters, if any, under which the estimator can be unbiased. Only then can we decide if these parameters are realistic for any human population. Using the forward genetic simulation framework developed in Chapter 4, I can vary the population founder size, reproductive success rate, mean number of offspring, or study sample size in order to find a set of parameters that result in an unbiased estimator. The small founder size (100 individuals) and small sample size (1,500 individuals) used in the simulation study in Chapter 4 may have resulted in some population structure that could be eliminated by increasing either of these two parameter values.

In the case of our Amish study of mammographic density, the estimator was, on average, 60% biased downward when applied to a polygenic quantitative trait simulated conditional on our Amish pedigree. This is likely due to a combination of undetected population structure within the Amish and non-random sampling of our study participants. A potential way forward, to apply the regression-based estimator to this population, is to identify a subset of individuals from our study for which the estimator is unbiased. I could cluster individuals based on their relationship coefficients to get subgroups that are similarly related to one another. While this approach would allow us to get an estimate of the heritability without the confounding of gene-gene interactions, the smaller sample size would reduce power and result in extremely large standard errors around the heritability estimates. Therefore, I may be unable to distinguish these estimates from 0 or 1, making them potentially uninformative.

Overall, the work in Chapter 4 to evaluate the statistical properties of the newly proposed regression-based estimator of the heritability has opened a door of new questions about consistency and utility of the estimator in practice. Further exploration using simulations and real data will help us understand under what specific circumstances, if any, the estimator can be reliably applied. In the case of our Amish study, the estimator cannot yet be applied without substantial bias.

Appendices

Appendix 1: Trait descriptions from our study of mammographic density

Trait descriptions for selected quantitative traits from our study of mammographic density in the Amish. These traits were used for analysis in Chapter 2.

Trait	Description	N	h^2
BMI	Body Mass Index	1,481	0.48
Body Fat	Percent body fat	1,109	0.52
Height	Height	1,481	0.94
Hip	Hip circumference	1,481	0.21
Waist	Waist circumference	1,481	0.52
Weight	Weight	1,481	0.57
WHR	waist circumference/ hip circumference	1,481	0.54
Age at Menarche	Age at first menstrual period	1,479	0.54
Age 1st Pregnancy	Age at first child	1,349	0.25
Live Birth No.	Number of live births	1,355	0.40
Nat Age at Meno	Age at natural menopause	671	0.42
Dense Area	Dense breast area	1,481	0.49
Nondense Area	Non-dense breast area	1,481	0.70
Percent Density	Percent density = dense area/total area * 100	1,481	0.45
Total Area	Total breast area	1,481	0.75
Estradiol-post	Estradiol in postmenopausal women	752	0.35
Free Estradiol-post	Estradiol indexed for SHBG in postmenopausal women	752	0.28
Free Estradiol-pre	Estradiol indexed for SHBG in premenopausal women	728	0.34
Free Testo-post	Testosterone indexed for SHBG in postmenopausal women	752	0.50
Free Testo-pre	Testosterone indexed for SHBG in premenopausal women	728	0.72
IGF1-post	Insulin-like growth factor 1 in postmenopausal women	546	0.77
IGF1-pre	Insulin-like growth factor 1 in premenopausal women	477	0.60

Trait	Description	N	h^2
IGF1/IGFBP3-post	IGF1/IGFBP3 in postmenopausal women	546	0.74
IGF1/IGFBP3-pre	IGF1/IGFBP3 in premenopausal women	477	0.31
IGFBP3-post	Insulin-like growth factor binding protein 3 in postmenopausal women	547	0.60
IGFBP3-pre	Insulin-like growth factor binding protein 3 in premenopausal women	477	0.69
MMP1-post	MMP-1 measured in serum in postmenopausal women	693	0.55
MMP1-pre	MMP-1 measured in serum in premenopausal women	668	0.87
Prolactin-post	Prolactin in postmenopausal women	752	0.26
Prolactin-pre	Prolactin in premenopausal women	728	0.24
Progesterone-pre	Progesterone in premenopausal women	728	0.35
SHBG-post	Steroid hormone binding globulin in postmenopausal women	752	0.33
SHBG-pre	Steroid hormone binding globulin in premenopausal women	728	0.70
Testosterone-post	Testosterone in postmenopausal women	752	0.50
Testosterone-pre	Testosterone in premenopausal women	728	0.55

h^2 is the narrow sense heritability of each trait after adjusting for age and menopausal status, and for the hormones and growth factors, after standardizing by batch, adjusting for age, and stratifying analyses by menopausal status; h^2 was significantly different from 0 ($p \leq 0.05$) for all traits except estradiol in premenopausal women and progesterone in postmenopausal women

Appendix 2: Number of SNPs from our study of mammographic density

Autosomal and X-chromosome SNPs were genotyped on 1,472 women from our study of mammographic density on the Illumina HumanOmni2.5-4v1_B Array at The Center for Inherited Disease Research (CIDR). We retained all high quality SNPs with a MAF ≥ 1% in our sample for analysis.

Quality Control Filter	Total
Total genotyped	2,443,179
>2 duplicate inconsistency[1]	671
>5% missing data[2]	47,579
Mendelian inconsistencies[2,3]	3,449
$P < 10^{-6}$ for HWE test[4]	2,032
Mitochondrial	93
Passed QC filter[5]	2,391,559
Duplicate SNPs	9,562
Non-uniquely mapped SNPs	2
Passed QC and unique	2,381,995
Passed QC	
Monomorphic[2]	715,458
Non-monomorphic[2]	1,666,537
MAF ≥ 0.01	**1,452,421**

[1] Based on 34 Amish and 32 CEU individuals who were genotyped in duplicate; SNPs with more than two duplicate genotype discrepancies were excluded.

[2] Based on 1,472 Amish individuals.

[3] SNPs with > 5 Mendelian inconsistencies.

[4] Based on 400 minimally related individuals.

[5] SNPs may fail QC in more than one way, so rows do not sum to the subtotal passing QC.

Appendix 3: Trait descriptions from the HAPI Heart study

Trait descriptions for selected quantitative traits from the Heredity and Phenotype Intervention (HAPI) Heart study in the Amish. These traits were used for analysis in Chapter 2.

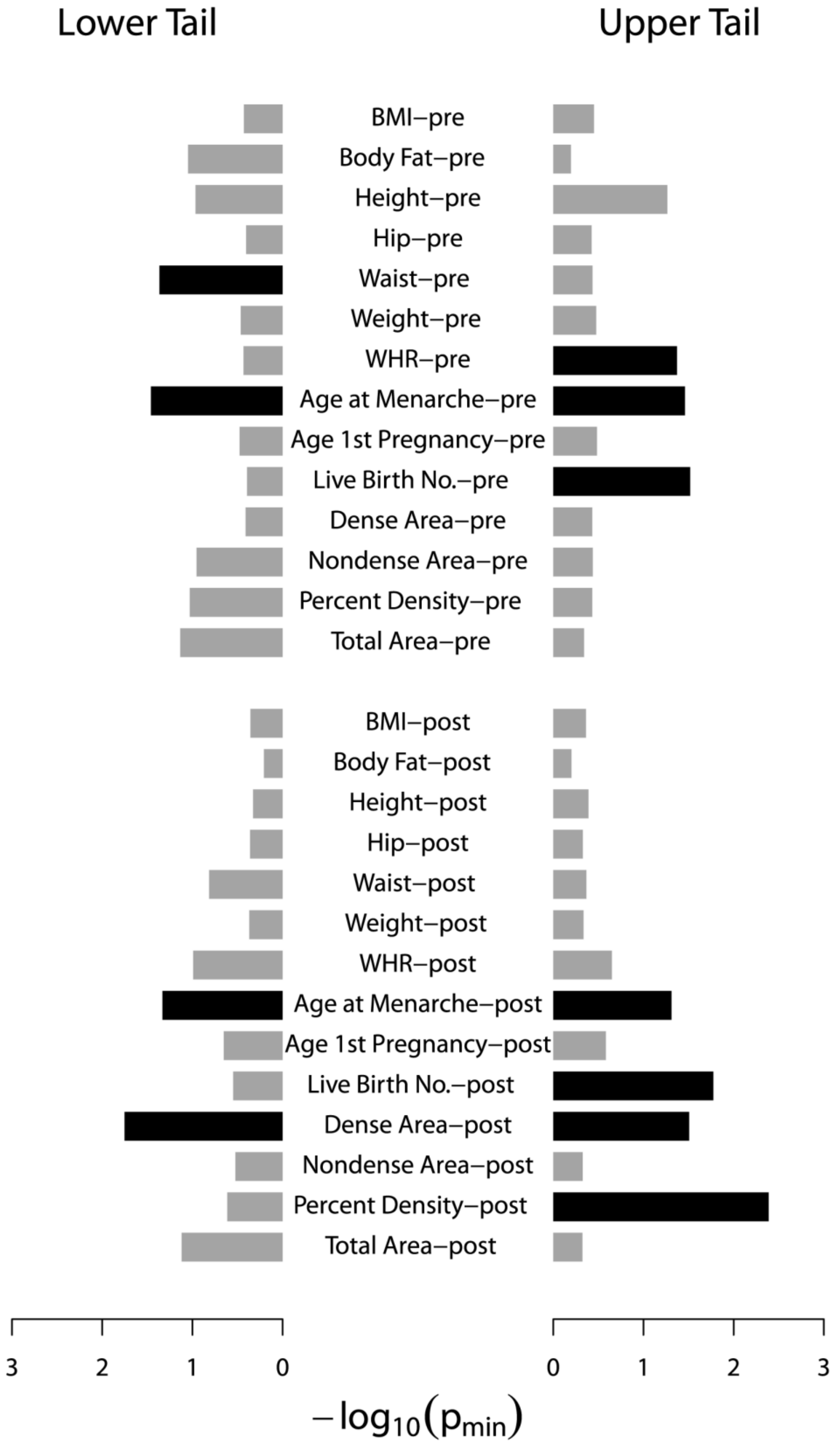
Trait	Description	N	h^2
BMI	Body Mass Index	868	0.49
Height	Height	868	0.71
Hip	Hip circumference	868	0.42
Waist	Waist circumference	868	0.51
Weight	Weight	868	0.59
WHR	Waist circumference/ hip circumference	868	0.38
Cholesterol	Fasting Total Cholesterol	858	0.73
Cholesterol/HDL	Total cholesterol/HDL cholesterol	858	0.59
C-Reactive Protein	C-Reactive Protein levels	857	0.33
HDL	Fasting HDL Cholesterol	858	0.58
HDL2	Fasting HDL sub fraction 2	850	0.51
HDL3	Fasting HDL sub fraction 3	850	0.50
IDL	Fasting intermediate density lipoprotein	850	0.44
LDL	Fasting LDL Cholesterol	857	0.73
Lipoprotein A	Fasting lipoprotein A	850	0.62
non-HDL	Fasting non-HDL cholesterol	850	0.68
Remnant Lipoprotein	Fasting remnant lipoprotein cholesterol	849	0.46
SAA	Serum Amyloid A	510	0.34
Total VLDL	Fasting very low density lipoprotein cholesterol	850	0.42
Triglycerides	Fasting triglycerides	858	0.50
VLDL3	Fasting VLDL sub fraction 3	850	0.44
Corrected QT Interval	QT Interval from the EKG corrected for heart rate	866	0.52
DBP	Diastolic blood pressure	868	0.14
HR	Heart rate	866	0.19
MAP	Mean arterial pressure = 2/3 DBP + 1/3 SBP	868	0.21
PR Interval	PR Interval from an EKG	799	0.38
QT Interval	QT Interval from the EKG	866	0.26
SBP	Systolic blood pressure	868	0.32
Carotid Radial PWV	Pulse wave velocity in the radial carotid	664	0.18
Common Carotid IMT	Common carotid artery Intimal Medial Thickness	819	0.33
Far Wall IMT	Common carotid artery Intimal Far Wall Max Thickness, mean of 4 measures	809	0.37

Trait	Description	N	h^2
Left Vent Mass	Left ventricle mass measured at echocardiogram	835	0.28
Left Vent Mass Index	Left Ventricular Mass / Body Surface Area	835	0.24
Luminal Diameter	Diameter of the common carotid artery at the end diastole	809	0.53
Ankle Brachial Index	Average of right and left ankle brachial index (mmHg)	861	0.23
Rel Wall Thickness	CommonCarotidIMT / LuminalDiameter	809	0.24
Vascular Mass	1.06*pi*((LuminaDiameter/2 + CommonCarotidIMT)^2 - (LuminalDiameter/2)^2)	809	0.29

Note – h^2 is the narrow sense heritability of each trait after adjusting for age and sex; h^2 was significantly different from 0 ($p \leq 0.05$) for all traits

Appendix 4: RVKT p-values (p_{min}) for quantitative traits from our study of mammographic density stratified by menopausal status

Each bar represents the result for a single trait. Black bars, significant ($p_{min} \leq 0.05$); gray bars, not significant. Before applying the RVKT, traits were transformed to approximate normality, when necessary, and adjusted for age.



Appendix 5: Trait descriptions from our study of mammographic density used for linkage analysis

Trait	Description	N	h^2 (S.E.)
BMI	Body Mass Index	1,481	0.48 (0.07)
Body Fat	Percent body fat	1,109	0.52 (0.08)
Height	Height	1,481	0.94 (0.06)
Hip	Hip circumference	1,481	0.21 (0.07)
Waist	Waist circumference	1,481	0.52 (0.07)
Weight	Weight	1,481	0.57 (0.07)
WHR	waist circumference/ hip circumference	1,481	0.54 (0.06)
Age at Menarche	Age at first menstrual period	1,479	0.54 (0.06)
Age 1st Pregnancy	Age at first child	1,349	0.25 (0.08)
Live Birth No.	Number of live births	1,355	0.40 (0.07)
Nat Age at Meno	Age at natural menopause	671	0.42 (0.11)
Dense Area	Dense breast area	1,481	0.49 (0.07)
Nondense Area	Non-dense breast area	1,481	0.70 (0.06)
Percent Density	Percent density = dense area/total area * 100	1,481	0.45 (0.07)
Total Area	Total breast area	1,481	0.75 (0.06)
MMP1	Matrix metalloproteinase 1	1,361	0.55 (0.06)

Note – h^2 is the narrow sense heritability of each trait after adjusting for age and menopausal status; h^2 was significantly different from 0 ($p \leq 0.05$) for all traits.

Appendix 6: Relationship coefficient summary statistics for simulated study samples

Summary statistics for all pairs (All) and distantly related pairs (Distant) for each simulated study sample (N = 1,500): random mating, random sampling (Random), random mating, non-random sampling (Siblings and Probands), and non-random mating, random sampling (Structured). A pair of individuals is defined as distantly related if their relationship coefficient is between 0 and $2\bar{r}$, where $\bar{r}$ is the overall mean relationship coefficient. Only distantly related pairs were included in the regression-based h^2 estimates.

		# Pairs		Mean r (SD)		Range of r	
Pedigree	Rep	All	Distant	All	Distant	All	Distant
Random	1	1,124,250	1120,773 (99.7%)	0.0586 (0.010)	0.0582 (0.004)	0.0515 - 0.6008	0.0515 - 0.1172
	2	1,124,250	1,122,028 (99.8%)	0.0697 (0.011)	0.0693 (0.005)	0.0605 - 0.6221	0.0605 - 0.1392
	3	1,124,250	1,120,417 (99.7%)	0.0545 (0.010)	0.0541 (0.003)	0.0478 - 0.5996	0.0478 - 0.1004
	4	1,124,250	1,122,134 (99.8%)	0.0678 (0.011)	0.0674 (0.005)	0.0589 - 0.6098	0.0589 - 0.1354
	5	1,124,250	1,121,585 (99.8%)	0.0608 (0.011)	0.0604 (0.005)	0.0523 - 0.7870	0.0523 - 0.1216
	6	1,124,250	1,120,403 (99.7%)	0.0503 (0.010)	0.0499 (0.004)	0.0431 - 0.5556	0.0431 - 0.0959
	7	1,124,250	1,121,660 (99.8%)	0.0678 (0.012)	0.0674 (0.005)	0.0580 - 0.6104	0.0580 - 0.1356
	8	1,124,250	1,120,751 (99.7%)	0.0594 (0.012)	0.0590 (0.005)	0.0511 - 0.5610	0.0511 - 0.1189
	9	1,124,250	1,120,439 (99.7%)	0.0592 (0.012)	0.0587 (0.005)	0.0506 - 0.6037	0.0506 - 0.1184
	10	1,124,250	1,120,471 (99.7%)	0.0537 (0.010)	0.0533 (0.003)	0.0470 - 0.5998	0.0470 - 0.0988
Siblings	1	1,124,250	1,120,117 (99.6%)	0.0591 (0.019)	0.0582 (0.004)	0.0511 - 0.6058	0.0511 - 0.1182
	2	1,124,250	1,121,015 (99.7%)	0.0703 (0.019)	0.0694 (0.005)	0.0611 - 0.7916	0.0611 - 0.1399
	3	1,124,250	1,119,429 (99.6%)	0.0550 (0.019)	0.0541 (0.003)	0.0478 - 0.6023	0.0478 - 0.1057
	4	1,124,250	1,120,939 (99.7%)	0.0683 (0.019)	0.0674 (0.005)	0.0589 - 0.6108	0.0589 - 0.1364
	5	1,124,250	1,120,584 (99.7%)	0.0613 (0.019)	0.0604 (0.005)	0.0530 - 0.6130	0.0530 - 0.1225
	6	1,124,250	1,119,301 (99.6%)	0.0508 (0.019)	0.0499 (0.004)	0.0430 - 0.5986	0.0430 - 0.1000
	7	1,124,250	1,120,555 (99.7%)	0.0683 (0.020)	0.0674 (0.005)	0.0584 - 0.6101	0.0584 - 0.1360
	8	1,124,250	1,120,253 (99.6%)	0.0599 (0.019)	0.0590 (0.005)	0.0507 - 0.5989	0.0507 - 0.1197
	9	1,124,250	1,119,991 (99.6%)	0.0597 (0.019)	0.0588 (0.005)	0.0508 - 0.7888	0.0508 - 0.1195
	10	1,124,250	1,119,333 (99.6%)	0.0542 (0.019)	0.0533 (0.003)	0.0473 - 0.5550	0.0473 - 0.0977

		# Pairs		Mean r (SD)		Range of r	
Pedigree	**Rep**	**All**	**Distant**	**All**	**Distant**	**All**	**Distant**
Probands	1	1,124,250	1,100,900 (97.9%)	0.0646 (0.038)	0.0603 (0.011)	0.0519 - 0.6008	0.0519 - 0.1262
	2	1,124,250	1,098,178 (97.7%)	0.0770 (0.041)	0.0719 (0.011)	0.0608 - 0.7913	0.0608 - 0.1494
	3	1,124,250	1,097,778 (97.6%)	0.0619 (0.041)	0.0568 (0.011)	0.0488 - 0.5585	0.0488 - 0.1235
	4	1,124,250	1,095,459 (97.4%)	0.0776 (0.042)	0.0723 (0.014)	0.0596 - 0.6146	0.0596 - 0.1481
	5	1,124,250	1,103,551 (98.2%)	0.0668 (0.037)	0.0628 (0.010)	0.0536 - 0.6031	0.0536 - 0.1329
	6	1,124,250	1,099,417 (97.8%)	0.0583 (0.040)	0.0535 (0.012)	0.0438 - 0.5955	0.0438 - 0.1166
	7	1,124,250	1,102,476 (98.1%)	0.0743 (0.037)	0.0701 (0.011)	0.0594 - 0.6082	0.0594 - 0.1479
	8	1,124,250	1,100,237 (97.9%)	0.0654 (0.039)	0.0607 (0.010)	0.0518 - 0.7894	0.0518 - 0.1308
	9	1,124,250	1,103,132 (98.1%)	0.0652 (0.037)	0.0612 (0.011)	0.0514 - 0.7850	0.0514 - 0.1304
	10	1,124,250	1,094,334 (97.3%)	0.0627 (0.042)	0.0572 (0.013)	0.0476 - 0.5590	0.0476 - 0.1214
Structured	1	1,124,250	1,106,923 (98.5%)	0.0738 (0.054)	0.0720 (0.052)	0.0188 - 0.6560	0.0188 - 0.1475
	2	1,124,250	1,025,489 (91.2%)	0.0508 (0.035)	0.0450 (0.030)	0.0145 - 0.5895	0.0145 - 0.1016
	3	1,124,250	951,929 (84.7%)	0.0702 (0.050)	0.0545 (0.034)	0.0228 - 0.8358	0.0228 - 0.1405
	4	1,124,250	1,016,957 (90.5%)	0.0644 (0.050)	0.0514 (0.030)	0.0170 - 0.7053	0.0170 - 0.1275
	5	1,124,250	1,033,846 (92.0%)	0.0557 (0.038)	0.0475 (0.025)	0.0167 - 0.6264	0.0167 - 0.1112
	6	1,124,250	1,004,933 (89.4%)	0.0567 (0.045)	0.0448 (0.029)	0.0130 - 0.6654	0.0130 - 0.1133
	7	1,124,250	1,099,756 (97.8%)	0.0626 (0.040)	0.0607 (0.037)	0.0214 - 0.6325	0.0214 - 0.1253
	8	1,124,250	1,080,522 (96.1%)	0.0629 (0.043)	0.0596 (0.039)	0.0206 - 0.8170	0.0206 - 0.1258
	9	1,124,250	1,040,032 (92.5%)	0.0555 (0.039)	0.0481 (0.029)	0.0130 - 0.6572	0.0130 - 0.1111
	10	1,124,250	1,060,782 (94.4%)	0.0652 (0.044)	0.0606 (0.039)	0.0213 - 0.8263	0.0213 - 0.1305

Appendix 7: Regression-based h^2 estimates for the randomly mating, randomly sampled simulated sample (Random) with different relationship coefficient ranges

The mean heritability estimate using the regression-based method over the 1,000 replicate simulations conditional on the randomly mating, randomly sampled study sample (Random). The estimates are based on various ranges of the relationship coefficient, r, around the mean relationship coefficient, $\bar{r}$, used to include pairs in the regression.

		Mean Estimated h^2 (SD) (%)		
		Relationship Coefficient (*r*) Range		
		$[0.5\bar{r}, 1.5\bar{r}]$	$[0,2\bar{r}]$	$[0,2.5\bar{r}]$
Simulated h^2 (%)	20	22 (21)	21 (19)	20 (18)
	40	36 (27)	36 (24)	36 (22)
	60	52 (30)	54 (27)	55 (25)
	80	68 (29)	71 (25)	72 (23)

References

1. Voight, B.F. and C. Cotsapas, Human genetics offers an emerging picture of common pathways and mechanisms in autoimmunity. Curr Opin Immunol, 2012. **24**(5): p. 552-7.

2. Glazier, A.M., J.H. Nadeau, and T.J. Aitman, Finding genes that underlie complex traits. Science, 2002. **298**(5602): p. 2345-9.

3. Green, E.D. and M.S. Guyer, Charting a course for genomic medicine from base pairs to bedside. Nature, 2011. **470**(7333): p. 204-13.

4. Lander, E.S., Initial impact of the sequencing of the human genome. Nature, 2011. **470**(7333): p. 187-97.

5. Chakravarti, A., Population genetics--making sense out of sequence. Nat.Genet., 1999. **21**(1 Suppl): p. 56-60.

6. Lander, E.S., The new genomics: global views of biology. Science, 1996. **274**(5287): p. 536-539.

7. Hindorff, L., et al. A Catalog of Published Genome-Wide Association Studies. 12/14/12; Available from: www.genome.gov/gwastudies.

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

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

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

11. McCarthy, M.I. and J.N. Hirschhorn, Genome-wide association studies: potential next steps on a genetic journey. Hum Mol Genet, 2008. **17**(R2): p. R156-65.

12. Pritchard, J.K., Are rare variants responsible for susceptibility to complex diseases? Am.J.Hum.Genet., 2001. **69**(1): p. 124-137.

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

14. Wijsman, E.M., The role of large pedigrees in an era of high-throughput sequencing. Hum Genet, 2012. **131**(10): p. 1555-63.

15. Mardis, E.R., Next-Generation Sequencing Platforms. Annu Rev Anal Chem (Palo Alto Calif), 2013.

16. Quail, M.A., et al., A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics, 2012. **13**: p. 341.

17. Cheverud, J.M. and E.J. Routman, Epistasis and its contribution to genetic variance components. Genetics, 1995. **139**(3): p. 1455-1461.

18. Culverhouse, R., et al., A perspective on epistasis: limits of models displaying no main effect. Am.J.Hum.Genet., 2002. **70**(2): p. 461-471.

19. Falconer, D.S. and T.F. Mackay, Introduction to Quantitative Genetics. 1996.

20. Song, Y.S., F. Wang, and M. Slatkin, General epistatic models of the risk of complex diseases. Genetics, 2010. **186**(4): p. 1467-73.

21. Zuk, O., et al., The mystery of missing heritability: Genetic interactions create phantom heritability. Proc Natl Acad Sci U S A, 2012. **109**(4): p. 1193-8.

22. Cordell, H.J., Detecting gene-gene interactions that underlie human diseases. Nat Rev Genet, 2009. **10**(6): p. 392-404.

23. Culverhouse, R.C., A comparison of methods sensitive to interactions with small main effects. Genet Epidemiol, 2012. **36**(4): p. 303-11.

24. Brem, R.B. and L. Kruglyak, The landscape of genetic complexity across 5,700 gene expression traits in yeast. Proc Natl Acad Sci U S A, 2005. **102**(5): p. 1572-7.

25. Carlborg, O. and C.S. Haley, Epistasis: too often neglected in complex trait studies? Nat.Rev.Genet., 2004. **5**(8): p. 618-625.

26. Huang, W., et al., Epistasis dominates the genetic architecture of Drosophila quantitative traits. Proc Natl Acad Sci U S A, 2012. **109**(39): p. 15553-9.

27. Cross, H.E., Population studies and the Old Order Amish. Nature, 1976. **262**(5563): p. 17-20.

28. Lee, W.J., et al., PedHunter 2.0 and its usage to characterize the founder structure of the Old Order Amish of Lancaster County. BMC Med Genet, 2010. **11**: p. 68.

29. McKusick, V.A., J.A. Hostetler, and J.A. Egeland, Genetic Studies of the Amish, Background and Potentialities. Bull Johns Hopkins Hosp, 1964. **115**: p. 203-22.

30. Kraybill, D.B., The riddle of Amish culture. 2003: Johns Hopkins University Press.

31. McKusick, V.A., The Amish. Endeavour, 1980. **4**(2): p. 52-7.

32. Ericksen, J.A., et al., Fertility patterns and trends among the Old Order Amish. Popul Stud (Camb), 1979. **33**(2): p. 255-76.

33. Strauss, K.A. and E.G. Puffenberger, Genetics, medicine, and the Plain people. Annu Rev Genomics Hum Genet, 2009. **10**: p. 513-36.

34. Agarwala, R., et al., Software for constructing and verifying pedigrees within large genealogies and an application to the Old Order Amish of Lancaster County. Genome Res, 1998. **8**(3): p. 211-21.

35. Agarwala, R., A.A. Schaffer, and J.F. Tomlin, Towards a complete North American Anabaptist Genealogy II: analysis of inbreeding. Hum Biol, 2001. **73**(4): p. 533-45.

36. Francomano, C.A., V.A. McKusick, and L.G. Biesecker, Medical genetic studies in the Amish: historical perspective. Am J Med Genet C Semin Med Genet, 2003. **121C**(1): p. 1-4.

37. Morton, D.H., et al., Pediatric medicine and the genetic disorders of the Amish and Mennonite people of Pennsylvania. Am J Med Genet C Semin Med Genet, 2003. **121C**(1): p. 5-17.

38. Cheng, Y.C., et al., Genetic effects on postprandial variations of inflammatory markers in healthy individuals. Obesity (Silver Spring), 2010. **18**(7): p. 1417-22.

39. Daley, E., et al., Variable bone fragility associated with an Amish COL1A2 variant and a knock-in mouse model. J Bone Miner Res, 2010. **25**(2): p. 247-61.

40. Ma, L., et al., Evaluation of A2BP1 as an obesity gene. Diabetes, 2010. **59**(11): p. 2837-45.

41. McArdle, P.F., et al., Association of a common nonsynonymous variant in GLUT9 with serum uric acid levels in old order amish. Arthritis Rheum, 2008. **58**(9): p. 2874-81.

42. Musunuru, K., et al., Association of single nucleotide polymorphisms on chromosome 9p21.3 with platelet reactivity: a potential mechanism for increased vascular disease. Circ Cardiovasc Genet, 2010. **3**(5): p. 445-53.

43. Njajou, O.T., et al., A common variant in the telomerase RNA component is associated with short telomere length. PLoS ONE, 2010. **5**(9): p. e13048.

44. Parsa, A., et al., Hypertrophy-associated polymorphisms ascertained in a founder cohort applied to heart failure risk and mortality. Clin Transl Sci, 2011. **4**(1): p. 17-23.

45. Pollin, T.I., et al., A null mutation in human APOC3 confers a favorable plasma lipid profile and apparent cardioprotection. Science, 2008. **322**(5908): p. 1702-5.

46. Roghmann, M.C., et al., Persistent Staphylococcus aureus colonization is not a strongly heritable trait in Amish families. PLoS ONE, 2011. **6**(2): p. e17368.

47. Shen, H., et al., Association of the vitamin D metabolism gene CYP24A1 with coronary artery calcification. Arterioscler Thromb Vasc Biol, 2010. **30**(12): p. 2648-54.

48. Shen, H., et al., Familial defective apolipoprotein B-100 and increased low-density lipoprotein cholesterol and coronary artery calcification in the old order amish. Arch Intern Med, 2010. **170**(20): p. 1850-5.

49. Shen, H., et al., Glucokinase regulatory protein gene polymorphism affects postprandial lipemic response in a dietary intervention study. Hum Genet, 2009. **126**(4): p. 567-74.

50. Shuldiner, A.R., et al., Association of cytochrome P450 2C19 genotype with the antiplatelet effect and clinical efficacy of clopidogrel therapy. JAMA, 2009. **302**(8): p. 849-57.

51. Tarasov, K.V., et al., COL4A1 is associated with arterial stiffness by genome-wide association scan. Circ Cardiovasc Genet, 2009. **2**(2): p. 151-8.

52. Wang, Y., et al., From the Cover: Whole-genome association study identifies STK39 as a hypertension susceptibility gene. Proc Natl Acad Sci U S A, 2009. **106**(1): p. 226-31.

53. Boyd, N.F., et al., Mammographic density and the risk and detection of breast cancer. N.Engl.J.Med., 2007. **356**(3): p. 227-236.

54. Douglas, J.A., et al., Mammographic breast density--evidence for genetic correlations with established breast cancer risk factors. Cancer Epidemiol Biomarkers Prev, 2008. **17**(12): p. 3509-16.

55. Mitchell, B.D., et al., The genetic response to short-term interventions affecting cardiovascular function: rationale and design of the Heredity and Phenotype Intervention (HAPI) Heart Study. Am Heart J, 2008. **155**(5): p. 823-8.

56. Zeggini, E., et al., An evaluation of HapMap sample size and tagging SNP performance in large-scale empirical and simulated data sets. Nat Genet, 2005. **37**(12): p. 1320-2.

57. Coventry, A., et al., Deep resequencing reveals excess rare recent variants consistent with explosive population growth. Nat Commun, 2010. **1**: p. 131.

58. Nelson, M.R., et al., An Abundance of Rare Functional Variants in 202 Drug Target Genes Sequenced in 14,002 People. Science, 2012.

59. Guey, L.T., et al., Power in the phenotypic extremes: a simulation study of power in discovery and replication of rare variants. Genet Epidemiol, 2011. **35**(4): p. 236-246.

60. Li, B. and S.M. Leal, Discovery of rare variants via sequencing: implications for the design of complex trait association studies. PLoS Genet, 2009. **5**(5): p. e1000481.

61. Ahituv, N., et al., Medical sequencing at the extremes of human body mass. Am J Hum Genet, 2007. **80**(4): p. 779-91.

62. Cohen, J.C., et al., Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science, 2004. **305**(5685): p. 869-72.

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

64. Kotowski, I.K., et al., A spectrum of PCSK9 alleles contributes to plasma levels of low-density lipoprotein cholesterol. Am J Hum Genet, 2006. **78**(3): p. 410-22.

65. Lander, E.S. and D. Botstein, Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics, 1989. **121**(1): p. 185-99.

66. Lange, K., Mathematical and Statistical Methods for Genetic Analysis. 2 ed. Statistics for Biology and Health, ed. K. Dietz, et al. 2002, New York: Springer-Verlag.

67. MORGAN, 3.0.3, Available from: http://www.stat.washington.edu/thompson/Genepi/MORGAN/Morgan.shtml.

68. Liu, F., et al., An approach for cutting large and complex pedigrees for linkage analysis. Eur J Hum Genet, 2008. **16**(7): p. 854-60.

69. Li, J. and L. Ji, Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. Heredity, 2005. **95**(3): p. 221-7.

70. Bailey-Wilson, J.E. and A.F. Wilson, Linkage analysis in the next-generation sequencing era. Hum Hered, 2011. **72**(4): p. 228-36.

71. Do, R., S. Kathiresan, and G.R. Abecasis, Exome sequencing and complex disease: practical aspects of rare variant association studies. Hum Mol Genet, 2012. **21**(R1): p. R1-9.

72. Smith, K.R., et al., Reducing the exome search space for mendelian diseases using genetic linkage analysis of exome genotypes. Genome Biol, 2011. **12**(9): p. R85.

73. Rosenthal, E.A., et al., Linkage and association of phospholipid transfer protein activity to LASS4. J Lipid Res, 2011. **52**(10): p. 1837-46.

74. Bowden, D.W., et al., Molecular basis of a linkage peak: exome sequencing and family-based analysis identify a rare genetic variant in the ADIPOQ gene in the IRAS Family Study. Hum Mol Genet, 2010. **19**(20): p. 4112-20.

75. Roeder, K., et al., Using linkage genome scans to improve power of association in genome scans. Am.J.Hum.Genet., 2006. **78**(2): p. 243-252.

76. Elston, R.C. and J. Stewart, A general model for the genetic analysis of pedigree data. Hum.Hered., 1971. **21**(6): p. 523-542.

77. Lander, E.S. and P. Green, Construction of multilocus genetic linkage maps in humans. Proc.Natl.Acad.Sci.U.S.A, 1987. **84**(8): p. 2363-2367.

78. Sieh, W., et al., Comparison of marker types and map assumptions using Markov chain Monte Carlo-based linkage analysis of COGA data. BMC Genet, 2005. **6 Suppl 1**: p. S11.

79. Tong, L. and E. Thompson, Multilocus lod scores in large pedigrees: combination of exact and approximate calculations. Hum Hered, 2008. **65**(3): p. 142-53.

80. Wijsman, E.M., Summary of Group 8: Development and extension of linkage methods. Genet Epidemiol, 2003. **25 Suppl 1**: p. S64-71.

81. Wijsman, E.M., J.H. Rothstein, and E.A. Thompson, Multipoint linkage analysis with many multiallelic or dense diallelic markers: Markov chain-Monte Carlo

provides practical approaches for genome scans on general pedigrees. Am J Hum Genet, 2006. **79**(5): p. 846-58.

82. Douglas, J.A. and C.I. Sandefur, PedMine--a simulated annealing algorithm to identify maximally unrelated individuals in population isolates. Bioinformatics, 2008. **24**(8): p. 1106-8.

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

84. Matise, T.C., et al., A second-generation combined linkage physical map of the human genome. Genome Res, 2007. **17**(12): p. 1783-6.

85. Almasy, L. and J. Blangero, Multipoint quantitative-trait linkage analysis in general pedigrees. Am.J Hum.Genet., 1998. **62**(5): p. 1198-1211.

86. Liu, F., et al., Ignoring distant genealogic loops leads to false-positives in homozygosity mapping. Ann Hum Genet, 2006. **70**(Pt 6): p. 965-70.

87. Dyer, T.D., et al., The effect of pedigree complexity on quantitative trait linkage analysis. Genet.Epidemiol., 2001. **21 Suppl 1**: p. S236-S243.

88. Heath, S.C., Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. Am.J Hum.Genet., 1997. **61**(3): p. 748-760.

89. Thompson, E.A. and S.C. Heath, Estimation of conditional multilocus gene identity among relatives. Lecture Notes-Monograph Series, 1999: p. 95-113.

90. Cheng, Y.C., et al., Genome-wide association scan identifies variants near Matrix Metalloproteinase (MMP) genes on chromosome 11q21-22 strongly associated with serum MMP-1 levels. Circ Cardiovasc Genet, 2009. **2**(4): p. 329-37.

91. Goddard, K.A. and E.M. Wijsman, Characteristics of genetic markers and maps for cost-effective genome screens using diallelic markers. Genet.Epidemiol., 2002. **22**(3): p. 205-220.

92. Visscher, P.M., et al., Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. PLoS Genet, 2006. **2**(3): p. e41.

93. Jacquard, A., Heritability: one word, three concepts. Biometrics, 1983. **39**(2): p. 465-77.

94. Ihaka, R. and R. Gentleman, R: A language for data analysis and graphics. Journal of computational and graphical statistics, 1996. **5**(3): p. 299-314.

95. Lange, K., An approximate model of polygenic inheritance. Genetics, 1997. **147**(3): p. 1423-1430.

96. Abney, M., M.S. McPeek, and C. Ober, Estimation of variance components of quantitative traits in inbred populations. Am J Hum Genet, 2000. **66**(2): p. 629-50.

97. Lange, K. and J.S. Sinsheimer, The pedigree trimming problem. Hum Hered, 2004. **58**(2): p. 108-11.

98. Hou, L., et al., Amish revisited: next-generation sequencing studies of psychiatric disorders among the Plain people. Trends Genet, 2013.

99. Browning, S.R. and B.L. Browning, Identity-by-descent-based heritability analysis in the Northern Finland Birth Cohort. Hum Genet, 2013. **132**(2): p. 129-38.

100. Abecasis, G.R., et al., A map of human genome variation from population-scale sequencing. Nature, 2010. **467**(7319): p. 1061-73.

101. Nelson, M.R., et al., An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science, 2012. **337**(6090): p. 100-4.

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