Methods and Applications for Collection, Contamination Estimation, and Linkage Analysis of Large-scale Human Genotype Data

by

Gregory J.M. Zajac

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in the University of Michigan
2020

Doctoral Committee:

Professor Gonçalo Abecasis, Chair
Professor Michael Boehnke
Associate Professor Hyun Min Kang
Professor Patricia Peyser
Professor Sebastian Zöllner

Gregory J.M. Zajac

gzajac@umich.edu

ORCID iD: 0000-0001-6411-9666

Dedication

To my wife Cynthia, and my parents Oleh and Renee

Acknowledgements

I extend my humblest and most sincere gratitude to my advisor, Goncalo Abecasis, for seven years of research mentorship and for giving me a stable job during graduate school. It was truly an honor to study under him and with the incredible team he has assembled at Michigan. I also thank Mike Boehnke for his mentorship as part of the Genome Science Training Program and for his service on my doctoral committee, along with Sebastian Zollner, Pat Peyser, and Hyun Min Kang. Their insights and constructive criticism of my work have been invaluable in shaping what was written in this dissertation and making it better.

Aside from my committee, I consider Lars Fritsche to be a great mentor who helped me get my first exposure to working with genetic data when I was a new MS student. I also thank Cristen Willer for the opportunity she gave me to work in her lab with Sarah Graham and learn an entire new set of skills. I extend my gratitude to Xiang Zhou, Shuang Feng, and Sarah Gagliano Taliun, for all their guidance on my research and letting me ask them countless questions. I thank Bob Henson for the opportunity to teach over the summer at ICPSR. I also thank fellow Wolverine Suzie Weekes for my first research experience in 2012.

I am grateful for the serendipity of having Chaolong Wang sit in the cubicle next to mine for a semester so I could ask him questions while developing the ancestry results for Genes for Good. To Anita Pandit, Kate Brieger, Kevin Li, Scott Vrieze, Johanna Foerster, Chris Clark, Aubrey Annis, Ellen Schmidt, Melissa and Stephanie Bachoura, Laura Baker, Irene Feliceti, and Chrissy Dobski: Genes for Good was one of the most exciting things I have ever had the opportunity to be a part of, and I will always remember working with each of you.

Being able to work with friends like Alan Kwong, Peter Van de Haar, and Santy Das helped to make this time in grad school one of many laughs and memories. I also thank the great folks I met in my cohort, like Chris Lee and Vincent Tan. I thank Gerhardt Hellman for telling me about the field of statistics and encouraging me to go to graduate school. I thank an anonymous grad student who encouraged me to learn how to program while I was a new undergraduate student in statistics. It was truly great advice!

I am grateful to Jonathan LeFaive, Joshua Weinstock, Andy Boughton, and Mary Kate Wing for their guidance and advice on my many programming problems. I also thank Sean Caron, Chris Scheller, Harsha Dushetty, and Chris Scheller for keeping the cluster running though many analyses. I certainly did my best to make good use of that resource while at Michigan, and probably caused you all plenty of headaches.

I am eternally grateful to my wife Cynthia for her love and support, hard work, and patience while I worked on this degree. I also thank her parents Raul and Rosario for their help in getting our life together established, and for all their help caring for us, our home, and our dog, Prosperous. I owe an enormous debt to my parents, brother, sister, and maternal grandparents for providing me the most wonderful intellectual environment for a boy to grow in and develop a love of learning and scholarly pursuits. I also thank my paternal grandparents, Wolodymyr and Anna, for risking their lives to flee their homeland during wartime and come to America so we could live lives full of opportunity and freedom. I thank my uncle Myron for encouraging me to leave California and start my adventure in higher education at the University of Missouri.

Finally, I thank all my teachers, professors, supervisors, colleagues, peers, friends, relatives, and church brethren, who are too many to list here, for believing in me and encouraging me to pursue great things with my life. I hope to make you all proud one day. Thank you.

Table of Contents

List of Tables

List of Figures

Abstract

In recent decades statistical genetics has contributed substantially to our knowledge of human health and biology. This research has many facets: from collecting data, to cleaning, to analyzing. As the scope of the scientific questions considered and the scale of the data continue to increase, these bring additional challenges to every step of the process. In this dissertation, I describe novel approaches for each of these three steps, focused on the specific problems of participant recruitment and engagement, DNA contamination estimation, and linkage analysis with large data sets. In Chapter 1, we introduce the subject of this dissertation and how it fits with other developments in the generation, analysis and interpretation of human genetic data.

In Chapter 2, we describe Genes for Good, a new platform for engaging a large, diverse participant pool in genetics research through social media. We developed a Facebook application where participants can sign up, take surveys related to their health, and easily invite interested friends to join. After completing a required number of these surveys, we send participants a spit kit to collect their DNA. In a statistical analysis of 27,000 individuals from all over the United States genotyped in our study, we replicated health trends and genetic associations, showing the utility of our approach and accuracy of self-reported phenotypes we collected.

In Chapter 3, we introduce VICES (Verify Intensity Contamination from Estimated Sources), a statistical method for joint estimation of DNA contamination and its sources in genotyping arrays. Genotyping array data are typically highly accurate but sensitive to mixing of DNA samples from multiple individuals before or during genotyping. VICES jointly estimates the

total proportion of contaminating DNA and identify which samples it came from by regressing deviations in probe intensity for a sample being tested on the genotypes of another sample. Through analysis of array intensity and genotype data from HapMap samples and the Michigan Genomics Initiative, we show that our method reliably estimates contamination more accurately than existing methods and implicates problematic steps to guide process improvements.

In Chapter 4, we propose Population Linkage, a novel approach to perform linkage analysis on genome-wide genotype data from tens of thousands of arbitrarily related individuals. Our method estimates kinship and identical-by-descent segments (IBD) between all pairs of individuals, fits them as variance components using Haseman-Elston regression, and tests for linkage. This chapter addresses how to iteratively assess evidence of linkage in large numbers of individuals across the genome, reduce repeated calculations, model relationships without pedigrees, and determine segregation of genomic segments between relatives using single-nucleotide polymorphism (SNP) genotypes. After applying our method to 6,602 individuals from the National Institute on Aging (NIA) SardiNIA study and 69,716 individuals from the Trøndelag Health Study (HUNT), we show that most of our signals overlapped with known GWAS loci and many of these could explain a greater proportion of the trait variance than the top GWAS SNP.

In Chapter 5, we discuss the impact and future directions for the work presented in this dissertation. We have proposed novel approaches for gathering useful research data, checking its quality, and detecting associations in the investigation of human genetics. Also, this work serves as an example for thinking about the process of human genetic discovery from beginning to end as a whole and understanding the role of each part.

Chapter 1 Introduction

Statistics has been applied to the field of genetics since its early days, both as a modelling tool and for analysis. Today, the field of statistical genetics is primarily concerned with how to study the effects of the naturally occurring genetic variation in humans as opposed to inducing genetic variation in cells or model organisms in the laboratory. Similar to other application areas of statistics, genetics has been revolutionized by advances in data collection, computing technology, and methods for large data sets. There are several major themes in statistical genetics that we will explore in this chapter and connect to other topics covered in later chapters of this dissertation.

The predominant theme in statistical genetics research has been in the improvement of genotyping technology. Sanger sequencing was introduced in 1977 but was prohibitively expensive for sequencing more than a few genomic segments in more than a handful of individuals (Sanger, Nicklen, & Coulson, 1977). This was soon followed by RFLP (Botstein, White, Skolnick, & Davis, 1980) and PCR genotyping (1984) which allowed faster genotyping of a single known variant to up to a few thousand short tandem repeats (Broman, Murray, Sheffield, White, & Weber, 1998). Genotyping arrays, invented in 1998, first allowed genotyping 1,500 variants in parallel and have continued to become more dense, with modern arrays sporting up to 4.3 million markers (Illumina, 2016; LaFramboise, 2009; Wang et al., 1998). Short-read sequencing, introduced in 2005 (Mukhopadhyay, 2009), led to similar improvements and the number of whole human genomes sequenced has increased to over one hundred thousand in one study alone (Kowalski et

al., 2019). This improvement has coincided with a drop in the cost of sequencing a human genome from $9 million in 2007 to $1,000 in 2019 (Wetterstrand, 2019). These technologies also switched the focus from highly polymorphic indels to single-nucleotide polymorphisms (SNPs), which account for the majority of genetic variation in humans (Auton et al., 2015). While still in development, long-read sequencing technologies are seeing increasing application in human genetics for detecting structural variants (Merker et al., 2018) and in rapidly generating more accurate reference genomes (Miga et al., 2020). In sum, these improvements have led to gains of several orders of magnitude in the amount and types of variation that can be captured, and in the number of individuals assayed. These technologies form an ever-widening foundation for statistical genetics research in humans.

These improvements in genotyping technology have also helped lead to many changes in how human genetic data is analyzed, in particular for mapping traits to genomic regions. Linkage analysis started for mapping traits to genomic regions with limited genotype data in families around (Morton, 1955). The first types of association tests were used in candidate gene studies but the majority of these had small sample sizes and had poor replicability (Hirschhorn, Lohmueller, Byrne, & Hirschhorn, 2002). With the introduction of affordable, dense genotyping arrays, these became ubiquitous as genome-wide association studies (Buniello et al., 2019). Genotype imputation (Li, Willer, Ding, Scheet, & Abecasis, 2010), improved reference panels (Auton et al., 2015), and GWAS based on whole-genome sequencing allowed testing more variants and combining information across multiple studies to achieve larger sample sizes in meta-analysis (Willer, Li, & Abecasis, 2010; Willer et al., 2013). Despite the success of these methods in finding associated loci, significant GWAS variants have failed to explain more than a tiny fraction of the heritability of complex traits (Manolio et al., 2009). This paradox has motivated scientists to

innovate with new types of analysis that have a greater functional interpretation, like gene-based tests to focus on coding sequence changes or eQTL and TWAS analysis to focus on the role of changes in expression on disease (Gusev et al., 2016; Wu et al., 2011). Others have turned to using PGRSs to capture the infinitesimal contribution of many variants beyond GWAS hits to model and study the genetics of traits (Wray, Goddard, & Visscher, 2007). Still others have developed methods for testing for gene-environment interactions to explain the missing heritability (Hahn, Ritchie, & Moore, 2003; Manning et al., 2011). In summary, the analysis of human genetic data has diversified considerably and now allows scientists to answer more questions than what genomic locations appear to influence a particular trait or disease.

The next step after mapping a trait to a particular gene or genomic region is often an experiment or analysis of additional -omics data to determine the function of a gene and how it influences the associated trait at the molecular level. While in vitro studies in human cell lines and in vivo studies in model organisms are considered the gold standard for functional characterization, these are expensive, time consuming, and findings may translate poorly to humans (Forstag & Anestidou, 2018). To complement these experimental methods, a number of bioinformatic approaches have been developed. At the most fundamental level, software has been developed to predict protein structure and function from its amino acid sequence (Yang et al., 2015), and changes in structure from a coding variant (Adzhubei et al., 2010). A variety of such algorithms for variant effect prediction were aggregated into CADD scores to predict the deleterious effect of any possible variant, even in noncoding or intergenic regions (Kircher et al., 2014; Rentzsch, Witten, Cooper, Shendure, & Kircher, 2019). Some methods use GWAS summary statistics to attempt to narrow down a causal variant through fine-mapping (Mägi et al., 2017) or constructing credible sets for causal variants (Maller et al., 2012). Many methods incorporate other types of -

omics data and summary statistics, like co-localization studies between GWAS variants and eQTLs (Plagnol, Smyth, Todd, & Clayton, 2009), peaks from ChIP-seq (Anand, Kalesinskas, Smail, & Tanigawa, 2019), or genetic interactions captured by Hi-C (Martin et al., 2015). In particular, methods that can characterize large numbers of genes or variants have become more important as gene-mapping analyses have transitioned from finding few variants of very large effect driving Mendelian disorders to associating many loci of uncertain function in complex traits.

As the aforementioned improvements in genotyping and sequencing technology have allowed researchers to assay an ever-increasing number and type of genetic variants, the number of individuals assayed has also rapidly increased and how cohorts are recruited, with new attention being turned to increasing the diversity of participants in genetic studies. Early linkage studies typically recruited a few dozen individuals from families that were enriched for the disease of interest (Fisher, Vargha-Khadem, Watkins, Monaco, & Pembrey, 1998; Tsui et al., 1985). The first GWAS studies typically recruited 100s to 1,000s of case and control individuals carefully matched on their demographics, and all of a single genetic ancestry (Consortium, 2007; Klein et al., 2005). More recent studies that have produced many findings include comprehensive genotyping of an entire community (Krokstad et al., 2013; Pilia et al., 2006), hospital system (Fritsche et al., 2018; Gaziano et al., 2016; Roden et al., 2008) or a national-level biobank (Hakonarson, Gulcher, & Stefansson, 2003; Metspalu, Köhler, Laschinski, Ganten, & Roots, 2004; Nagai et al., 2017; Sudlow et al., 2015). Direct-to-consumer genetics companies have also been able to engage millions in research with the services they provide (Stoekle, Mamzer-Bruneel, Vogt, & Herve, 2016). Even though these study designs have all done a great deal to increase the inclusiveness and diversity of genetics research, many types of diversity are still poorly captured in genetic studies to date (Popejoy & Fullerton, 2016). This problem is well recognized and there is a

concerted push to address it (Hindorff et al., 2018). Several projects currently underway or in planning phases are making great strides in this respect and we can expect this to be a major theme in genetics studies of humans in the near future (Mapes et al., 2020; Nhlbi, 2018).

Parallel to this trend in the number and types of individuals recruited for human genetic studies, there has been a diversification in how phenotypes are collected. While the largest meta-analyses may still focus on a single or small number of traits (M. Liu et al., 2019), several of the most prolific individual studies include information on hundreds or even thousands of traits (Bycroft et al., 2018; Gagliano Taliun et al., 2020). In particular, studies that collect electronic health records are able to extract up to thousands of phenotypes on their participants by analyzing insurance billing codes (Carroll, Bastarache, & Denny, 2014), extracting values from laboratory tests generated during patient care (Goldstein et al., 2020), or text mining of physician notes from office visits (Denny, 2012). Many genetic studies are relying solely on self-reported phenotype data, particularly those recruited from direct-to-consumer genomics companies (Tsoi et al., 2017). In addition, many studies are collecting traits that are primarily behavioral or psychological rather than health-related, or that may not have an obvious genetic mechanism but which still yield genetic associations (Barban et al., 2016).

While there are many constants in the area of data cleaning and quality checking (QC), this has also grown to meet the challenges brought by innovations in how data is collected and analyzed for genetic studies. As genetics studies have become larger and more diverse, classical methods for Hardy-Weinberg Equilibrium (HWE) testing were adapted for cohorts with multiple ancestries (Kwong et al., 2020). As studies transitioned from linkage to GWAS and sample sizes grew, faster methods were developed to estimate genetic relatedness to identify duplicates, sample swaps, and exclude close relatives in a robust fashion (Manichaikul et al., 2010). As meta-analyses came to

include millions of samples across hundreds of cohorts, it became more difficult to verify that a given sample was not a participant in multiple cohorts that contributed to the meta-analysis, particularly since researchers typically do not have permission to share individual-level data. A variety of approaches have been developed to address this, including calculating kinship on encrypted genotypes (Zhao, 2019) and estimating the overlap between samples using only the shared summary statistics (Sengupta, 2018). Self-reported and EHR-derived phenotypes are especially prone to over reporting of cases or mislabeling similar phenotypes as one another (like type 1 and type 2 diabetes), so approaches were developed to detect and potentially correct for these issues (Duffy et al., 2004; Ekstrøm & Feenstra, 2012). Some existing methods also found new applications, for example methods for contamination detection in sequence data were extended to identify droplets with multiple cells in single-cell RNA sequencing (Kang et al., 2018), or maternal DNA in fetal samples (Nabieva et al., 2020), or host and tumor DNA in cancer samples (Bergmann, Chen, Arora, Vacic, & Zody, 2016).

The process of discovery in human genetic is multi-faceted and includes many steps in the collection, preparation, and generation of genetic and phenotypic data beyond gene mapping and functional characterization. In this dissertation, I present three chapters that represent advances in 5 of these 6 themes in statistical genetics: recruitment, phenotyping, genotyping, quality control, and analysis. Specifically, I describe a novel strategy for building a genetics cohort over social media, a statistical method for joint estimation of DNA contamination and its sources, and a scalable framework for performing linkage analysis in population cohorts with tens of thousands of individuals. Figure 1-1 shows how these three projects intersect and connect to these larger themes in human genetics. The following sections in this chapter provide more details on what the

shortcomings and limitations of existing methods in these areas and how those shortcomings motivated the work presented in this dissertation.

Participant recruitment and engagement

As mentioned previously in this chapter, a great variety of study designs have been successfully applied in human genetics to study a wide array of human traits and conditions (Buniello et al., 2019). Despite this success, there continue to be many shortcomings in the recruitment of samples and collection of phenotype data that limit their diversity and translatability of findings to traditionally underserved communities. Academic efforts, while generally free for individuals who participate, often recruit participants from a particular medical center or health system and typically exclude people outside their geographic reach or who lack access to medical care (Shavers-Hornaday, Lynch, Burmeister, & Torner, 1997). Even biobank efforts that attempt to be more inclusive and include willing participants from across an entire country—like the UK Biobank—still require a lengthy, in-person assessment at one of their recruitment centers that represents a significant barrier to participation for many individuals (Bycroft et al., 2018). Direct-to-consumer genomics companies have made great progress on some of these limitations and have recruited an impressive number of participants from around the world (Ehm et al., 2017), but since individuals usually must purchase the companies' genotyping service to be included, these cohorts are limited in terms of the socioeconomic and racial diversity of their participants.

To address some of these shortcomings, in Chapter 2, I introduce Genes for Good, a novel platform for participant contact, recruitment, and engagement over social media. Typically, prospective participants find the study through one of their social media contacts. After consenting, they have the option to take several surveys about their health, behavior, and psychology. Once a participant completes a required number of these, we would mail a spit kit to them to collect their

DNA. Upon returning this, their sample is genotyped and we return their decoded genotypes and an estimate of their genetic ancestry to them. We then use the data our participants shared to run analyses and contribute to large consortia conducting meta-analyses. Because anyone in the US over 18 and with a Facebook account can join and participation is completely free, Genes for Good represents an effort to explore potential solutions to the aforementioned issues with both academic and direct-to-consumer genotyping efforts.

Chapter 2 fits primarily into the themes of direct recruitment of participants and collection of self-reported phenotypes as it aims to address shortcomings in these areas. Figure 1-1 indicates that it also intersects with the theme of genotyping since it is an application of genotyping array technology in building this cohort. In addition, we involved data cleaning and quality checks into Genes for Good, particularly to test the validity and utility of self-reported phenotype data from volunteers. Finally, Chapter 2 is an application of GWAS analysis methods to validate the accuracy of the data collected in the study.

Contamination estimation

Contamination, defined as the mixture of genetic material from individuals of the same species before or during genotyping, is a prolific problem known to affect array genotype calls and downstream analyses (Flickinger, Jun, Abecasis, Boehnke, & Kang, 2015). Existing methods for genotyping arrays include a hypothesis test for the presence of an individual in a specific DNA sample (Homer et al., 2008) and a variety of estimation algorithms based on allele frequencies (Jun et al., 2012). Researchers conducting a large-scale genotyping study are potentially interested in both estimating the proportion of contaminating DNA in a sample to determine its exclusion from further analysis and identifying the source of contamination to guide improvements in sample preparation and potential re-extraction and genotyping of clean DNA. In addition, contamination

methods based on allele frequencies become biased when the allele frequencies are calculated in a different population than the contaminating DNA is from, as can occur in a diverse study with samples from a variety of genetic ancestries. Currently, there is no unified framework to estimate the total proportion of contaminating DNA in a sample in a fashion that is robust to genetic ancestry and the individuals that contributed to it.

In Chapter 3 I introduce Verify Intensity Contamination from Estimated Sources (VICES), a statistical method that aims to directly address these goals by jointly estimating contamination and its sources in genotyped samples. The intuition behind VICES rests on the fact that contamination causes the distribution of array probe intensities to deviate from their expectation. VICES regresses these deviations in the array probe intensity of one sample on the genotypes of another sample. After using allele frequencies to control for the effect of greater dissimilarity between samples at common variants, the regression coefficient of the sample genotypes provides an estimate of the contribution of that sample to the overall mixture. These estimates and the estimate of total contamination can then be refined by jointly regressing the array probe intensities on the genotypes of all identified contaminating samples. VICES runs all these steps automatically, seamlessly, and efficiently to make it easy to use and scalable for large genotyping projects. Because VICES replaces allele frequencies with contamination sample genotypes, VICES aims to address the issues of biased estimates resulting from miss-specified allele frequencies in addition to providing researchers with more information about how contamination occurred.

The predominant theme in Chapter 3 is that of sample-level quality checks. It moves this sub-field forward by giving researchers better information to make decisions about excluding contaminated samples (particularly in a diverse setting) or choosing to regenotype them based on how contamination occurred. Chapter 3 is also highly integrated with the theme of genotyping

array technology, since the method is tailored to contamination estimation in arrays and depends on specific aspects of array genotyping. The chapter also includes some discussion of how contamination affects downstream analyses.

Linkage analysis

Genome-wide association studies (GWAS) continue to report novel associations with ever-increasing sample size, the collection of more phenotypes, denser genotyping arrays, and the increasing availability of short-read sequence data. However, new results, while impressive in number, often represent marginal gains in the proportion of trait variance explained and actual biological insights into the traits studied (Fritsche et al., 2013). One explanation that some have proposed for this "missing heritability" in GWAS is that single-variant tests do a poor job of capturing the contributions of ungenotyped variation, allelic heterogeneity and epistatic interaction in many traits (Manolio et al., 2009). Linkage analysis, a class of methods for testing for the co-segregation of a trait with genomic segments within families, has been proposed as a solution to several of these problems (Hodge, Hager, & Greenberg, 2016; Manichaikul et al., 2010). These features, and the presence of large numbers of related individuals in many cohorts recruited for GWAS (particularly those from a population-based study) might make linkage an attractive choice for a secondary analysis in many studies. However, existing methods for linkage analysis have many drawbacks that make them impossible or impractical to run in a population cohort collected for a GWAS analysis. The first issue is that the classical methods for linkage analysis based on the Elston-Stewart or Lander-Green algorithms were developed when genotype information was relatively expensive to collect and scale poorly to data sets with hundreds of thousands of genetic markers across tens of thousands of individuals (Ott, Wang, & Leal, 2015). This problem is often solved by splitting large pedigrees into sib pairs or nuclear families (F. Liu, Kirichenko,

Axenovich, van Duijn, & Aulchenko, 2008), but the relatedness structure in a study collected for GWAS might contain more scattered pairs of loose relatives than complete pedigrees, if any pedigree information was collected from participants at all. Finally, linkage methods are often designed to use highly polymorphic microsatellite markers with many alleles for maximum informativeness for detecting recombination events between relatives, while GWAS studies typically collect single-nucleotide polymorphism (SNP) genotypes that are more ambiguous indicators of allelic segregation (Evans & Cardon, 2004).

All these problems have helped to motivate Population Linkage, a new method for scaling up linkage analysis to population-level data which I introduce in Chapter 4. Population Linkage works by first obtaining estimates of kinship and identical-by-descent (IBD) regions using genome-wide genotype data. It then fits these estimates in a fast approximation of a variance-components model for a quantitative trait known as Haseman-Elston regression to obtain a point estimate and standard error of the trait variance attributed to IBD sharing in a region. The method then takes this point estimate and its standard error to test for evidence of linkage at that locus and repeats this process across the genome. Population Linkage addresses the problem of the scalability of linkage analysis by taking advantage of efficient methods for estimating IBD segments and fitting variance components models rather than performing full-pedigree likelihood calculations. Despite the minimal information in individual SNPs for inferring recombination events in linkage analysis, the initial step of calculating IBD estimates for Population Linkage effectively combines this information across multiple SNPs. Population Linkage can work with any arbitrary pedigree structure and takes all pairwise relationships into account simultaneously. It does not depend on reported pedigree information at all, instead using only the relationship information inferred from the genotype data. Population Linkage is a method intended to

complement a GWAS to provide additional insights for the proportion of trait variance explained by a region, and to capture the effects of ungenotyped variation, allelic heterogeneity, and epistatic effects that might be missed in a GWAS.

Chapter 4 is primarily concerned with the analysis of genetic data for mapping traits to genomic regions. It is specifically focused on applying linkage analysis in population cohorts that have many relative pairs and families. The phenotypes considered in Chapter 4 are also specifically focused on those that are chemically determined in the lab from biological samples, like blood lipid measurements. Finally, similar to Chapters 2 and 3, Chapter 4 is focused on conducting these analyses with genotyping array data. Chapter 4 also includes a brief discussion on how applying Population Linkage in a sequencing study may be different.

Summary of objectives

In this dissertation, we propose to demonstrate solutions for existing limitations through the following aims:

1. Show success in recruiting a nation-wide genetics cohort based on self-reported phenotype data
2. Design a more useful estimator of DNA contamination that reveals the source and probable cause of DNA contamination
3. Develop a method for computationally tractable linkage analysis on population-scale genotype data with power to reveal novel insights about the traits being studied.

These three objectives are addressed in Chapter 2 (Objective 1), Chapter 3 (Objective 2), and Chapter 4 (Objective 3). More detailed background, motivation, and results can be found in each of these chapters.

Figures

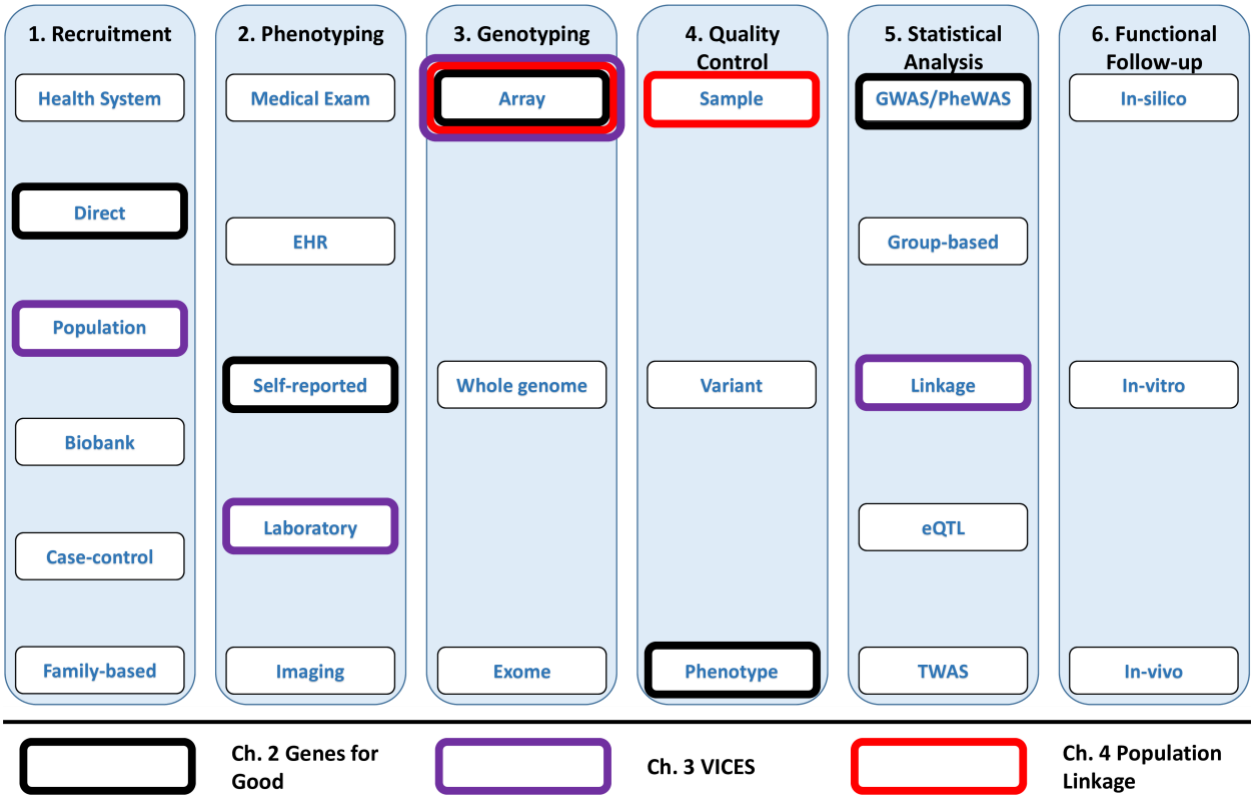


Figure 1-1 Sequence of discovery in human genetics

This figure details some of the steps in human genetic discovery, with particular emphasis on the approaches and technologies commonly used in statistical genetics. Within each step, several methods used in that step are listed. The methods for each step that intersect with the three projects in this dissertation are boxed.

References

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., . . . Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nat Methods, 7*(4), 248-249. doi: 10.1038/nmeth0410-248

Anand, S., Kalesinskas, L., Smail, C., & Tanigawa, Y. (2019). SNPs2ChIP: Latent Factors of ChIP-seq to infer functions of non-coding SNPs. *Pac Symp Biocomput, 24*, 184-195.

Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., . . . Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature, 526*(7571), 68-74. doi: 10.1038/nature15393 [doi]

Barban, N., Jansen, R., de Vlaming, R., Vaez, A., Mandemakers, J. J., Tropf, F. C., . . . Study, L. C. (2016). Genome-wide analysis identifies 12 loci influencing human reproductive behavior. *Nat Genet, 48*(12), 1462-1472. doi: 10.1038/ng.3698

Bergmann, E. A., Chen, B. J., Arora, K., Vacic, V., & Zody, M. C. (2016). Conpair: concordance and contamination estimator for matched tumor-normal pairs. *Bioinformatics, 32*(20), 3196-3198. doi: 10.1093/bioinformatics/btw389

Botstein, D., White, R. L., Skolnick, M., & Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet, 32*(3), 314-331.

Broman, K. W., Murray, J. C., Sheffield, V. C., White, R. L., & Weber, J. L. (1998). Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am J Hum Genet, 63*(3), 861-869. doi: 10.1086/302011

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., . . . Parkinson, H. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res, 47*(D1), D1005-D1012. doi: 10.1093/nar/gky1120

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., . . . Marchini, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature, 562*(7726), 203-209. doi: 10.1038/s41586-018-0579-z

Carroll, R. J., Bastarache, L., & Denny, J. C. (2014). R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics, 30*(16), 2375-2376. doi: 10.1093/bioinformatics/btu197

Consortium, W. T. C. C. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature, 447*(7145), 661-678. doi: 10.1038/nature05911

Denny, J. C. (2012). Chapter 13: Mining electronic health records in the genomics era. *PLoS Comput Biol, 8*(12), e1002823. doi: 10.1371/journal.pcbi.1002823

Duffy, S. W., Warwick, J., Williams, A. R. W., Keshavarz, H., Kaffashian, F., Rohan, T. E., . . . Sadeghi-Hassanabadi, A. (2004). A simple model for potential use with a misclassified binary outcome in epidemiology. *Journal of Epidemiology and Community Health, 58*(8), 712-717. doi: 10.1136/jech.2003.010546

Ehm, M. G., Aponte, J. L., Chiano, M. N., Yerges-Armstrong, L. M., Johnson, T., Barker, J. N., . . . Waterworth, D. M. (2017). Phenome-wide association study using research participants' self-reported data provides insight into the Th17 and IL-17 pathway. *PLoS One, 12*(11), e0186405. doi: 10.1371/journal.pone.0186405

Ekstrøm, C. T., & Feenstra, B. (2012). Detecting sample misidentifications in genetic association studies. *Stat Appl Genet Mol Biol, 11*(3), Article 13. doi: 10.1515/1544-6115.1772

Evans, D. M., & Cardon, L. R. (2004). Guidelines for genotyping in genomewide linkage studies: single-nucleotide-polymorphism maps versus microsatellite maps. *Am J Hum Genet, 75*(4), 687-692. doi: 10.1086/424696

Fisher, S. E., Vargha-Khadem, F., Watkins, K. E., Monaco, A. P., & Pembrey, M. E. (1998). Localisation of a gene implicated in a severe speech and language disorder. *Nat Genet, 18*(2), 168-170. doi: 10.1038/ng0298-168

Flickinger, M., Jun, G., Abecasis, G. R., Boehnke, M., & Kang, H. M. (2015). Correcting for Sample Contamination in Genotype Calling of DNA Sequence Data. *Am J Hum Genet, 97*(2), 284-290. doi: 10.1016/j.ajhg.2015.07.002

Forstag, E. H., & Anestidou, L. (2018). Advancing disease modeling in animal-based research in support of precision medicine

proceedings of a workshop Retrieved from http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=e000xna&AN=1841450

Fritsche, L. G., Chen, W., Schu, M., Yaspan, B. L., Yu, Y., Thorleifsson, G., . . . Consortium, A. G. (2013). Seven new loci associated with age-related macular degeneration. *Nat Genet, 45*(4), 433-439, 439e431-432. doi: 10.1038/ng.2578

Fritsche, L. G., Gruber, S. B., Wu, Z., Schmidt, E. M., Zawistowski, M., Moser, S. E., . . . Mukherjee, B. (2018). Association of Polygenic Risk Scores for Multiple Cancers in a Phenome-wide Study: Results from The Michigan Genomics Initiative. *Am J Hum Genet, 102*(6), 1048-1061. doi: 10.1016/j.ajhg.2018.04.001

Gagliano Taliun, S. A., VandeHaar, P., Boughton, A. P., Welch, R. P., Taliun, D., Schmidt, E. M., . . . Abecasis, G. R. (2020). Exploring and visualizing large-scale genetic associations by using PheWeb. *Nat Genet, 52*(6), 550-552. doi: 10.1038/s41588-020-0622-5

Gaziano, J. M., Concato, J., Brophy, M., Fiore, L., Pyarajan, S., Breeling, J., . . . O'Leary, T. J. (2016). Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J Clin Epidemiol, 70*, 214-223. doi: 10.1016/j.jclinepi.2015.09.016

Goldstein, J. A., Weinstock, J. S., Bastarache, L. A., Larach, D. B., Fritsche, L. G., Schmidt, E. M., . . . Zawistowski, M. (2020). LabWAS: novel findings and study design recommendations from a meta-analysis of clinical labs in two independent biobanks. *medRxiv*, 2020.2004.2008.19011478. doi: 10.1101/2020.04.08.19011478

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W., . . . Pasaniuc, B. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet, 48*(3), 245-252. doi: 10.1038/ng.3506

Hahn, L. W., Ritchie, M. D., & Moore, J. H. (2003). Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *Bioinformatics, 19*(3), 376-382. doi: 10.1093/bioinformatics/btf869

Hakonarson, H., Gulcher, J. R., & Stefansson, K. (2003). deCODE genetics, Inc. *Pharmacogenomics, 4*(2), 209-215. doi: 10.1517/phgs.4.2.209.22627

Hindorff, L. A., Bonham, V. L., Brody, L. C., Ginoza, M. E. C., Hutter, C. M., Manolio, T. A., & Green, E. D. (2018). Prioritizing diversity in human genomics research. *Nat Rev Genet, 19*(3), 175-185. doi: 10.1038/nrg.2017.89

Hirschhorn, J. N., Lohmueller, K., Byrne, E., & Hirschhorn, K. (2002). A comprehensive review of genetic association studies. *Genet Med, 4*(2), 45-61. doi: 10.1097/00125817-200203000-00002

Hodge, S. E., Hager, V. R., & Greenberg, D. A. (2016). Using Linkage Analysis to Detect Gene-Gene Interactions. 2. Improved Reliability and Extension to More-Complex Models. *PLoS One, 11*(1), e0146240. doi: 10.1371/journal.pone.0146240

Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., . . . Craig, D. W. (2008). Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet, 4*(8), e1000167. doi: 10.1371/journal.pgen.1000167

Illumina. (2016). Infinium ® Omni5-4 v1.2 BeadChip. San Diego, CA.

Jun, G., Flickinger, M., Hetrick, K. N., Romm, J. M., Doheny, K. F., Abecasis, G. R., . . . Kang, H. M. (2012). Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am J Hum Genet, 91*(5), 839-848. doi: 10.1016/j.ajhg.2012.09.004

Kang, H. M., Subramaniam, M., Targ, S., Nguyen, M., Maliskova, L., McCarthy, E., . . . Ye, C. J. (2018). Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat Biotechnol, 36*(1), 89-94. doi: 10.1038/nbt.4042

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet, 46*(3), 310-315. doi: 10.1038/ng.2892

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J. Y., Sackler, R. S., Haynes, C., . . . Hoh, J. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science, 308*(5720), 385-389. doi: 10.1126/science.1109557

Kowalski, M. H., Qian, H., Hou, Z., Rosen, J. D., Tapia, A. L., Shan, Y., . . . Group, T. H. H. W. (2019). Use of >100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. *PLoS Genet, 15*(12), e1008500. doi: 10.1371/journal.pgen.1008500

Krokstad, S., Langhammer, A., Hveem, K., Holmen, T. L., Midthjell, K., Stene, T. R., . . . Holmen, J. (2013). Cohort Profile: the HUNT Study, Norway. *International journal of epidemiology, 42*(4), 968-977. doi: 10.1093/ije/dys095

Kwong, A. M., Blackwell, T. W., LeFaive, J., de Andrade, M., Barnard, J., Barnes, K. C., . . . Kang, H. M. (2020). Robust, flexible, and scalable tests for Hardy-Weinberg Equilibrium across diverse ancestries. *bioRxiv*, 2020.2006.2023.167759. doi: 10.1101/2020.06.23.167759

LaFramboise, T. (2009). Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Res, 37*(13), 4181-4193. doi: 10.1093/nar/gkp552

Li, Y., Willer, C. J., Ding, J., Scheet, P., & Abecasis, G. R. (2010). MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol, 34*(8), 816-834. doi: 10.1002/gepi.20533

Liu, F., Kirichenko, A., Axenovich, T. I., van Duijn, C. M., & Aulchenko, Y. S. (2008). An approach for cutting large and complex pedigrees for linkage analysis. *Eur J Hum Genet, 16*(7), 854-860. doi: 10.1038/ejhg.2008.24

Liu, M., Jiang, Y., Wedow, R., Li, Y., Brazel, D. M., Chen, F., . . . Psychiatry, H. A.-I. (2019). Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nat Genet, 51*(2), 237-244. doi: 10.1038/s41588-018-0307-5

Maller, J. B., McVean, G., Byrnes, J., Vukcevic, D., Palin, K., Su, Z., . . . Consortium, W. T. C. C. (2012). Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat Genet, 44*(12), 1294-1301. doi: 10.1038/ng.2435

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W. M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics, 26*(22), 2867-2873. doi: 10.1093/bioinformatics/btq559

Manning, A. K., LaValley, M., Liu, C. T., Rice, K., An, P., Liu, Y., . . . Dupuis, J. (2011). Meta-analysis of gene-environment interaction: joint estimation of SNP and SNP × environment regression coefficients. *Genet Epidemiol, 35*(1), 11-18. doi: 10.1002/gepi.20546

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., . . . Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature, 461*(7265), 747-753. doi: 10.1038/nature08494

Mapes, B. M., Foster, C. S., Kusnoor, S. V., Epelbaum, M. I., AuYoung, M., Jenkins, G., . . . Program, A. o. U. R. (2020). Diversity and inclusion for the All of Us research program: A scoping review. *PLoS One, 15*(7), e0234962. doi: 10.1371/journal.pone.0234962

Martin, P., McGovern, A., Orozco, G., Duffus, K., Yarwood, A., Schoenfelder, S., . . . Eyre, S. (2015). Capture Hi-C reveals novel candidate genes and complex long-range interactions with related autoimmune risk loci. *Nat Commun, 6*, 10069. doi: 10.1038/ncomms10069

Merker, J. D., Wenger, A. M., Sneddon, T., Grove, M., Zappala, Z., Fresard, L., . . . Ashley, E. A. (2018). Long-read genome sequencing identifies causal structural variation in a Mendelian disease. *Genet Med, 20*(1), 159-163. doi: 10.1038/gim.2017.86

Metspalu, A., Köhler, F., Laschinski, G., Ganten, D., & Roots, I. (2004). [The Estonian Genome Project in the context of European genome research]. *Dtsch Med Wochenschr, 129 Suppl 1*, S25-28. doi: 10.1055/s-2004-824840

Miga, K. H., Koren, S., Rhie, A., Vollger, M. R., Gershman, A., Bzikadze, A., . . . Phillippy, A. M. (2020). Telomere-to-telomere assembly of a complete human X chromosome. *Nature*. doi: 10.1038/s41586-020-2547-7

Morton, N. E. (1955). Sequential tests for the detection of linkage. *Am J Hum Genet, 7*(3), 277-318.

Mukhopadhyay, R. (2009). DNA sequencers: the next generation. *Anal Chem, 81*(5), 1736-1740. doi: 10.1021/ac802712u

Mägi, R., Horikoshi, M., Sofer, T., Mahajan, A., Kitajima, H., Franceschini, N., . . . COGENT-Kidney Consortium, T. D.-G. C. (2017). Trans-ethnic meta-regression of genome-wide association studies accounting for ancestry increases power for discovery and improves fine-mapping resolution. *Hum Mol Genet, 26*(18), 3639-3650. doi: 10.1093/hmg/ddx280

Nabieva, E., Sharma, S. M., Kapushev, Y., Garushyants, S. K., Fedotova, A. V., Moskalenko, V. N., . . . Yarotsky, D. (2020). Accurate fetal variant calling in the presence of maternal cell contamination. *Eur J Hum Genet*. doi: 10.1038/s41431-020-0697-6

Nagai, A., Hirata, M., Kamatani, Y., Muto, K., Matsuda, K., Kiyohara, Y., . . . Group, B. J. C. H. (2017). Overview of the BioBank Japan Project: Study design and profile. *J Epidemiol, 27*(3S), S2-S8. doi: 10.1016/j.je.2016.12.005

Nhlbi, U. (2018). The NHLBI Trans-Omics for Precision Medicine (TOPMed) Whole Genome Sequencing Program. *BRAVO variant browser*.

Ott, J., Wang, J., & Leal, S. M. (2015). Genetic linkage analysis in the age of whole-genome sequencing. *Nat Rev Genet, 16*(5), 275-284. doi: 10.1038/nrg3908

Pilia, G., Chen, W. M., Scuteri, A., Orrú, M., Albai, G., Dei, M., . . . Schlessinger, D. (2006). Heritability of cardiovascular and personality traits in 6,148 Sardinians. *PLoS Genet, 2*(8), e132. doi: 10.1371/journal.pgen.0020132

Plagnol, V., Smyth, D. J., Todd, J. A., & Clayton, D. G. (2009). Statistical independence of the colocalized association signals for type 1 diabetes and RPS26 gene expression on chromosome 12q13. *Biostatistics, 10*(2), 327-334. doi: 10.1093/biostatistics/kxn039

Popejoy, A. B., & Fullerton, S. M. (2016). Genomics is failing on diversity. *Nature, 538*(7624), 161-164. doi: 10.1038/538161a

Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., & Kircher, M. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res, 47*(D1), D886-D894. doi: 10.1093/nar/gky1016

Roden, D. M., Pulley, J. M., Basford, M. A., Bernard, G. R., Clayton, E. W., Balser, J. R., & Masys, D. R. (2008). Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin Pharmacol Ther, 84*(3), 362-369. doi: 10.1038/clpt.2008.89

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A, 74*(12), 5463-5467. doi: 10.1073/pnas.74.12.5463

Sengupta, S. (2018). Improved Analysis of Large Genetic Association Studies Using Summary Statistics Retrieved from http://hdl.handle.net/2027.42/143992

Shavers-Hornaday, V. L., Lynch, C. F., Burmeister, L. F., & Torner, J. C. (1997). Why are African Americans under-represented in medical research studies? Impediments to participation. *Ethn Health, 2*(1-2), 31-45. doi: 10.1080/13557858.1997.9961813

Stoekle, H. C., Mamzer-Bruneel, M. F., Vogt, G., & Herve, C. (2016). 23andMe: a new two-sided data-banking market model. *BMC medical ethics, 17*, 9. doi: 10.1186/s12910-016-0101-9 [doi]

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., . . . Collins, R. (2015). UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med, 12*(3), e1001779. doi: 10.1371/journal.pmed.1001779

Tsoi, L. C., Stuart, P. E., Tian, C., Gudjonsson, J. E., Das, S., Zawistowski, M., . . . Elder, J. T. (2017). Large scale meta-analysis characterizes genetic architecture for common psoriasis associated variants. *Nat Commun, 8*, 15382. doi: 10.1038/ncomms15382

Tsui, L. C., Buchwald, M., Barker, D., Braman, J. C., Knowlton, R., Schumm, J. W., . . . Plavsic, N. (1985). Cystic fibrosis locus defined by a genetically linked polymorphic DNA marker. *Science, 230*(4729), 1054-1057. doi: 10.1126/science.2997931

Wang, D. G., Fan, J. B., Siao, C. J., Berno, A., Young, P., Sapolsky, R., . . . Lander, E. S. (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms

in the human genome. *Science, 280*(5366), 1077-1082. doi: 10.1126/science.280.5366.1077

Wetterstrand, K. A. (2019). DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Retrieved August 11, 2020, from https://www.genome.gov/sequencingcostsdata

Willer, C. J., Li, Y., & Abecasis, G. R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics, 26*(17), 2190-2191. doi: 10.1093/bioinformatics/btq340

Willer, C. J., Schmidt, E. M., Sengupta, S., Peloso, G. M., Gustafsson, S., Kanoni, S., . . . Consortium, G. L. G. (2013). Discovery and refinement of loci associated with lipid levels. *Nat Genet, 45*(11), 1274-1283. doi: 10.1038/ng.2797

Wray, N. R., Goddard, M. E., & Visscher, P. M. (2007). Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res, 17*(10), 1520-1528. doi: 10.1101/gr.6665407

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., & Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet, 89*(1), 82-93. doi: 10.1016/j.ajhg.2011.05.029

Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., & Zhang, Y. (2015). The I-TASSER Suite: protein structure and function prediction. *Nat Methods, 12*(1), 7-8. doi: 10.1038/nmeth.3213

Zhao, X. (2019). Statistical Methods and Privacy Preserving Protocols for Combining Genetic Data with Electronic Health Records.

Chapter 2 Genes for Good: Engaging the Public in Genetics Research Using Social Media[1]

Introduction

More than 10,000 genetic loci have been successfully linked to common and complex diseases (Welter et al., 2014). In previous decades, the major challenge for human genetic studies was the cost and complexity of the genotyping itself; however, researchers now face the bigger hurdle of obtaining large enough samples that also include useful, linked medical and health data. The study designs typically used to collect such data are expensive and often exclude individuals based on location or demographics. We reasoned that using social media platforms would not only allow us to recruit a large population cohort, but also help us to reach populations that might not typically participate in genetic studies due to the time commitment or distance to a research center. Potential advantages of social media-based study designs include the ability to reach diverse populations and the ability to engage participants in research over time. Potential concerns include representativeness and the ability of this approach to reproduce findings obtained using more traditional designs.

We present a new study design to take advantage of recent developments in health survey methods using social media and widespread interest in direct-to-consumer genetic testing (Royal

[1] This work was published in the American Journal of Human Genetics as, "Genes for good: engaging the public in genetics research via social media," *105*(1), 65-77. I was one of three joint first authors and contributed in processing incoming genotype data, generating participant results and educational materials, running replication GWASs and accompanying tables and figures, writing sections about methods and results, and summarizing demographic and disease information in Genes for Good and comparison studies, along with creating the relevant tables and figures.

et al., 2010; Stoekle, Mamzer-Bruneel, Vogt, & Herve, 2016). Genes for Good is an ongoing, large-scale study of health, genetic, and behavioral information. We aim to engage tens of thousands of individuals in research through a Facebook application, reducing the expense of traditional epidemiologic designs and the exclusivity and high socioeconomic status associated with current direct-to-consumer efforts (Agurs-Collins et al., 2015).

Our model of using social media for genetic research invites participants to complete online health assessments at their convenience, as has been successfully applied in numerous studies of health, behavior (Pedersen & Kurz, 2016), and psychology (Kosinski, Matz, Gosling, Popov, & Stillwell, 2015), including studies of rare genetic diseases (Abiad, Robbins, Morris, & Sobreira, 2018), childbirth preferences (Arcia, 2014), and prediction of personality traits (Kosinski et al., 2015). When a consenting participant has completed a minimum number of health history and health tracking surveys, they are mailed a spit kit to collect DNA for analysis. After genotyping, we test genetic variants for association with health, disease, and environmental information collected through online assessments.

In this paper, we demonstrate that the Genes for Good study model is a viable complement to more traditional research study designs. The phenotypic and genotypic data we have collected thus far appear valid and reliable. Further, the incentive structure of Genes for Good – namely, altruism combined with the return of survey response summaries and genetic data to participants – is effective, as demonstrated by exponential recruitment from all fifty U. S. states. Importantly, the recruitment happened organically, with participants publicizing the study through their own networks, without relying on paid advertising. We briefly explored the use of study recruitment websites (such as ResearchMatch (Harris et al., 2012)), but only several hundred participants were recruited this way. We also saw large influxes of participants after online articles appeared in

Reddit (Free DNA Test from the University of Michigan, 2017) and Buzzfeed (Hughes, 2015). While resources still go toward answering questions about the study and resolving technical issues, efficient participant recruitment and engagement allowed us to dedicate a larger fraction of resources to sample collection, processing and downstream analyses. The long-term goals of the study fall broadly into five main categories: (a) to identify novel genetic loci associated with a variety of phenotypes, (b) to longitudinally track an array of health and behavioral measures, (c) to enable genotype-first study designs (such as detailed phenotypic assessments of participants with naturally-occurring knockout variants), (d) to educate participants and make the data available to them, and (e) to encourage data sharing among researchers. Here, we present our study design and methods, as well as initial findings about our sample demographics and important health indicators.

One particular advantage of hosting our study on social media is that we can reach participants in an environment that many already visit regularly as part of their daily routines. Social media use in the U.S. has dramatically increased in the last decade – rising from 7% in 2005 to over 65% in 2015 (Perrin, 2015) – and so we have the potential to reach a majority of the U.S. population through our application. In the last few years, several research groups have recognized the major advantages social media offers: flexible timing, the possibility of incentives and reminders, and the ability to reach non-urban communities. There has already been substantial success in recruiting for studies via Facebook (Fenner et al., 2012) as well as in using it to prevent loss-to-follow-up (Mychasiuk & Benzies, 2012). Further, the flexible framework of Genes for Good allows us and other research groups to continue adding new surveys and activities to address future research questions. Our study takes advantage of the opportunity for repeated contact that

social media offers and represents the first large genetic study of tens of thousands of individuals conducted via Facebook.

Considering their ubiquity and ease of use, social media and mobile devices as research tools are important avenues to explore further (Steinhubl, Muse, & Topol, 2015). However, we recognize some of the potential disadvantages we are likely to face: (a) inaccurate data, (b) low response rate (Kapp, Peters, & Oliver, 2013), (c) high attrition, and (d) a sample limited to those who have a Facebook account. In the first year of the study, we prioritized testing and combatting several of these expected limitations. With the aforementioned challenges in mind, we implemented various methods to assess the quality of our data. First, we looked at common diseases and phenotypes to validate our results – and thus our approach to data collection – by comparing them to prior findings from traditional research and meta-analysis designs. When expected phenotypic relationships hold true, such as that between BMI and Type 2 diabetes, we gain confidence in the quality of the survey responses we are collecting. Additionally, we assessed the quality of the genetic data by replicating findings from genome wide association studies (GWAS) for a variety of traits that are known to have genetic components, such as diabetes, asthma, BMI, hair color, and eye color, confirming that our data yields the expected signals. We also examined rates of chronic health conditions, such as hypertension and diabetes, to explore how our study participants compare to the overall U.S. population.

Material and Methods

We have implemented a large, IRB-approved genetic study using social media. Participants must be at least 18 years old, live in the U.S., and have a Facebook account. They are recruited via snowball sampling, i.e. by finding our Genes for Good Facebook application through friends, family, and social media connections. Once a person has consented, they are invited to complete

online health history assessments at their convenience. The surveys consist of health history questionnaires, daily tracking surveys, and an optional health conditions module in which participants can list other conditions that they have. Once they have completed a minimum number of required questionnaires, they are mailed a spit kit to collect DNA for analysis. The cost of each participant is about $80, which includes postage, DNA extraction, and genotyping; there is essentially no cost associated with recruitment or data collection. Throughout the course of the study, we have typically employed 2-3 full-time staff (study coordinator, developers), several graduate and undergraduate students, and a part-time administrative assistant to assist with sending and receiving spit kits.

Genetic analysis

DNA is genotyped at ~600,000 SNPs using either the Illumina Infinium CoreExome-24 v1.0 or v1.1 arrays for nonsynonymous exonic variants and a panel of common genome-wide markers (Illumina, 2017). The standard set of markers on the array is augmented with missense, loss of function, and potential lipids and myocardial infarction variants identified in the HUNT whole genome sequencing and whole exome sequencing projects (Krokstad et al., 2013); height-associated variants from GIANT (Wood et al., 2014); potential stop-gain variants in 96 genes at loci potentially implicated in type 2 diabetes, blood lipid levels, Alzheimer's disease, nicotine/alcohol metabolism, and several others with mutations implicated in serious but treatable health conditions; complex trait associated variants in the EBI/NHGRI GWAS catalog (Welter et al., 2014); a random subset of Neanderthal SNPs from the 1000 Genomes Project (Sankararaman et al., 2014); ancestry informative markers identified by Paschou et al. that were highly correlated with the principal components of Human Genome Diversity Project samples (Paschou, Lewis, Javed, & Drineas, 2010); and pain related variants proposed by Dr. Chad Brummett of the

University of Michigan Division of Pain Research. Genotypes at an additional >30 million variants in the 1000 Genomes Phase 3 panel (Auton et al., 2015) are imputed using Minimac3 (Das et al., 2016). After quality control, local genetic ancestry is estimated using RFMix (Maples, Gravel, Kenny, & Bustamante, 2013), global ancestry with ADMIXTURE (Alexander, Novembre, & Lange, 2009), and principal components analysis performed with TRACE (Wang, Zhan, Liang, Abecasis, & Lin, 2015), using the Human Genome Diversity Project samples as a reference panel (Li et al., 2008) for all three analyses. We provide each Genes for Good participant with a section in the app to view these estimates of genetic ancestry on the sample they provided.

For the GWAS of Genes for Good participants' BMI, the BMI measurements were calculated from the Height and Weight survey in the app, which was derived from height and weight questionnaires available from PhenX Toolkit (Hamilton et al., 2011). Weight measurements for the first several thousand genotyped participants were bottom-coded at 80 lbs and top-coded at 251 lbs; then, the top-coded value was changed to 381 lbs partway through the study to capture a greater range of variation. For participants that were pregnant at the time of answering the survey, we used their pre-pregnancy weight obtained from the same survey. The BMI values were then regressed on sex, age, array chip version, and the first five principal components; the residuals were inverse-normal transformed in order to compare effect size estimates to the largest published meta-analysis of BMI (Locke et al., 2015) and to reduce the impact of extreme observations. We used the SAIGE software (Zhou et al., 2018) to run a mixed model GWAS, accounting for sample relatedness and population structure. Polygenic risk scores were calculated using PLINK (Chang et al., 2015).

Participant engagement

We provide participants with several ways to interact with both their own data and the research study as a whole. After each health history survey is completed, we provide charts summarizing the information, in some cases comparing each participant's answers to the Genes for Good study population (example in Figure 2-7 Example Health History resultFigure 2-7). Similarly, for daily tracking surveys, we generate summaries of each participant's health behavior over time as well as summary statistics for the entire study (example in Figure 2-8). In addition to providing this ongoing feedback and summary of the survey responses, we also offer participants who submit a sample a breakdown of their genetic ancestry; the current version includes 7 continental human populations (Europe, Africa, East Asia, Central/South Asia, West Asia/North Africa, Americas, and Oceania), and results are served in the form of a global ancestry estimate, local ancestry inference, and principal components analysis using the methods described previously (RFMIX, ADMIXTURE, TRACE). Before seeing their estimates of genetic ancestry, they are required to watch a short video on how to interpret their results. Participants can also download their array and imputed genotypes.

Privacy and data security

All Genes for Good data is divided into two classes: (a) personally identifiable information, such as email addresses, Facebook user IDs, and physical mailing addresses; and (b) research information, such as survey responses and genetic data. Each class of data is stored in a distinct relational database and served from a distinct server. Extracts for outside researchers include only research-specific data. We plan to ask participants to allow use of their mailing address to link to information such as geocode pollution, built environment (for instance, the number of fast food outlets or public parks within a certain radius of one's home), and census tract data. In these cases,

the participants' physical address would still be withheld from external collaborators, but variables generated using addresses could be shared upon request.

The privacy of Genes for Good data is monitored by the University of Michigan Institutional Review Board. All genetic and survey results are stored in a secure server on campus that is not directly connected to the public internet, and DNA samples are stored in physically secure spaces with restricted access. In addition, all archived data is de-identified to protect subject privacy including participants' demographic summary and genetic information. Even though Genes for Good uses Facebook to authenticate login, Facebook does not access information we collect through the App and we do not use participant's social media postings and connections in our research. We make efforts to communicate with participants about the extensive measures we take in ensuring the privacy of their data and to ease their worries about using social media as a platform for genetic research.

All communication to and from the application is encrypted. Participants are authenticated using a Facebook account and Facebook's OAuth implementation, ensuring that participants only have access to their own data once inside the application. Communication with Facebook servers is limited to authentication only; although Facebook receives and retains information about which Facebook accounts have accessed the Genes for Good app, all other information provided by participants is provided directly to Genes for Good servers. Facebook cannot see any of the data entered by participants.

Once participants have their genetic data analyzed, they are notified that they may access results inside the app with a Results Access Code, a randomly generated alphanumeric code that must be requested by the participant and will be delivered to the email address on the participant's Genes for Good profile. Participant genotype data is processed internally on University of

Michigan servers and is distributed to participants upon request via Box, a secure third-party file-sharing platform. Participants may request their raw genotypes as often as they like from within the genetic results section of the app. Each request compresses and uploads raw genotype data and supplementary information to a private, password-protected Box account directory. For security purposes, all requested genotypes automatically expire from Box servers three days after being uploaded.

Results

Since the launch of Genes for Good on January 19th, 2015 (Martin Luther King Jr. Day), we have seen steadily increasing participant recruitment and consistent use of the Facebook application. Genes for Good now has enough participants to begin conducting meaningful analyses with the data. As of March 2019, 117,652 participants had tried the app, with 81,110 who signed the electronic consent form. Consenting users have completed over 2.9 million surveys, answering >22 million questions. Genes for Good has mailed 33,427 spit kits to eligible participants, of which 27,470 have been returned (as of March 2019). The genetic data freeze used for this paper contains data from 20,232 participants whose genotypes passed quality control checks as of mid-2018.

Sample characteristics and phenotypes

Participants were recruited successfully from all fifty states, with areas of peak participant density roughly overlapping with major U.S. population centers (Figure 2-1). About 90% of users have residential addresses outside of Michigan. Compared to the U.S. population, our sample is younger (Genes for Good median age of 33, U.S. adult median age of 44) and enriched for females (74% of participants are women, compared to 51% for US adults, Table 2-1). Our sample also closely resembles the U.S. population on household income, although it is enriched for individuals

from middle-income households with an annual income of $35,000 - $100,000; Table 2-3). In contrast, the majority of the participants in the research cohort from 23andMe are from households with an annual income over $100,000 (Tung et al., 2011). To confirm the quality of the data collected from our sample, we also compared disease rates to those in the general U.S. population (Table 2-2). In looking at important risk factors for cardiovascular disease, we observed relatively similar rates of high cholesterol, hypertension, and smoking. However, our sample had lower rates of disease outcomes such as stroke and myocardial infarction. Our genotype data freeze contained 20,232 individuals, of which 76.3% were non-Hispanic white, 3.8% Asian, 2.7% African American, 8.8% multi-racial/other, and 8.3% Hispanic/Latino as determined by self-report through our Demographics survey.

In addition to the phenotype information collected from survey responses, 12,216 participants have reported 64,401 cases of 3,067 health conditions in an optional section of the app that allows participants to search for and report disorders using the Systematized Nomenclature of Medicine (SNOMED) dictionary (Lee, Cornet, Lau, & de Keizer, 2013). These participant-entered data show that Genes for Good has attracted an unusually high proportion of individuals with certain rare diseases, like Ehlers-Danlos Syndrome (565 cases or 0.93% of GfG participants compared to ~0.02% prevalence worldwide) (Levy, 2018). The 5 most commonly reported disorders were generalized anxiety disorder (1,803 cases), asthma (1,389), hypothyroidism (941), depressive disorder (920), and migraine (918). Higher BMI was associated with increased risk for all 5 conditions in logistic regression of each of the five traits on BMI, sex, and age (odds ratios of 1.02, 1.03, 1.04, 1.01, 1.03 per unit higher BMI, p-values of 7.6×10^{-9}, 2.1×10^{-20}, 3.9×10^{-24}, 1.5×10^{-4}, 6.3×10^{-14}).

To evaluate the quality of our data, we used our survey data to verify known phenotypic relationships. Taking diabetes as an example, we analyzed the association of the disease with BMI. Given the rapidly increasing prevalence of diabetes in the U.S., this is a particularly important outcome to examine. Over the past three decades, the number of diagnosed Americans has more than tripled, from 5.6 million in 1980 to 21 million in 2012 (CDC, 2014). And because about one-third of diabetics are undiagnosed, national survey statistics consistently underestimate the true prevalence of diabetes (CDC, 2014). We compared rates of diabetes in our sample, within each BMI bracket, to those reported from nationally representative samples (Bays, Chapman, Grandy, & Group, 2007) and found a similar trend of increasing diabetes prevalence as BMI increased (Figure 2-2). We further explored this relationship by calculating the estimated effect of BMI on diabetes status, adjusting for age, sex, and race, using NHANES and Genes for Good data separately. We found that the relationship between BMI and diabetes was comparable between studies (95% CI for odds ratio per 1-unit increase in BMI, NHANES: 1.07-1.10; 95% CI, GFG: 1.08-1.10). When comparing simple correlation coefficients between BMI and diabetes status across studies, we found no notable difference between Genes for Good and NHANES (r_{GFG}=0.18, r_{NHANES}=0.19, p = 0.83). Though our sample is quite different from NHANES in terms of wealth, age distribution, and ethnic diversity, we observe similar trends in both cohorts when comparing diabetes cases and controls: diabetics typically have higher rates of obesity, higher age, lower income, and lower education (Supplementary Table 2-1).

Genetic associations

To validate the quality of our self-reported phenotypes, we analyzed a data freeze of 20,232 genotypes to see if we could replicate known genetic associations. We first analyzed traits related to pigmentation and BMI, because these traits are known to have strong genetic factors. For

example, most variation in eye color is determined by 6 SNPs in *HERC2* and *OCA2* (F. Liu et al., 2009). Figure 2-3 shows the number of participants with each combination of eye color and genotype at one of the SNPs with the strongest association signal, rs12913832. We observed strong evidence of association between eye color and genotype (X^2 = 15,599, df = 8, $p = 10^{-3376}$, N=19,974), and the direction of effects is consistent with what was previously reported. Other pigmentation traits like hair color, skin sun response, and hair texture are also consistent with prior studies. Supplementary Table 2-2 shows detailed GWAS results, and Supplementary Table 2-3 compares our results to several larger studies. We show that Genes for Good replicates the top pigmentation associations in prior studies at least nominally ($p < 0.05$), and frequently does so at genome-wide significance ($p < 5x10^{-8}$).

We next compared results for a mixed model GWAS of BMI, using measurements obtained from the Height and Weight health history survey, to results from the GIANT consortium (Locke et al., 2015). We obtained effect sizes consistent with those published for the top ten GIANT loci. We also obtained nominally significant ($p < 0.05$) association results at all 10 loci. Figure 2-4 summarizes the comparison of our results with published GIANT results, showing consistency of direction of effect, magnitude, and relative significance (Figure 2-5 shows regional association in our top signal, at FTO). Given the relatively small sample size of our data, our effect estimates necessarily have wider confidence limits compared to the meta-analysis. However, the meta-analysis point estimates are contained within these limits for nearly every SNP, which provide evidence that self-reported phenotypes collected within our cohort are reliable.

We next expanded our comparison of GWAS results obtained with Genes for Good data to include the traits of type 1 diabetes, type 2 diabetes, and asthma. For all traits except asthma, our association signals are consistent with reports from published large GWAS and show some

significant hits (Supplementary Table 2-2, Supplementary Table 2-3, and Supplementary Table 2-4; Supplementary Figure 2-1). Our asthma analysis did not give any genome-wide significant results, but when we examined the eighteen SNPs associated with asthma in the study of Demenais et al. (2018) we found that all had a consistent direction of effect in Genes for Good data but with smaller effect sizes (Supplementary Table 2-4). Our asthma cases and controls were defined based on answers to "Was your asthma ever confirmed by a doctor?" with 4,378 cases and 11,715 controls reported. Given the large proportion of cases (27.2%), we believe that some individuals who answered "Yes" did not meet the standard for an asthma diagnosis used in Demenais et al. (2018) A similar observation has been made in other studies of self-reported phenotypes — for example, in a study of psoriasis including data from 23andMe customers, it was estimated that only ~36% of individuals who self-reported having psoriasis met the criteria used in clinical studies, diluting association signals and effect size estimates (Tsoi et al., 2017). We did an adjustment proposed by Duffy et al. (2004) to account for the apparent over-reporting of cases (Duffy et al., 2004). We also did a power calculation at the 0.05 significance level to determine our ability to replicate the findings in Demenais et al. and estimated that we should replicate approximately 7 of 18 SNPs (summing estimated power across eighteen variants gives expected number of 6.8 replicated signals). After the Duffy adjustment over half of our odds ratios were closer to the effect sizes reported in Demenais et al., though some odds ratios were overcorrected to have effect sizes larger than those reported in Demenais et al. As our power calculation suggested, we were able to replicate 7 of the 18 SNPs at the 0.05 significance level (Table S4) (Demenais et al., 2018; Tsoi et al., 2017). Reassuringly, we also found that, when we calculated polygenic risk scores (PRS) for type 1 and type 2 diabetes using publicly available GWAS summary statistics (Bycroft et al., 2018; Lunshof, Church, & Prainsack, 2014), PRS for type 2

diabetes was strongly associated with self-reported type 2 diabetes status (OR increase per PRS quintile=1.47; p=7.63 x 10^{-37}) and that PRS for type 1 diabetes PRS was strongly associated with self-reported type 1 diabetes status (OR increase per PRS quintile=1.66; p=5.13 x 10^{-9}) (Figure 2-6). We found similar support for an association between asthma PRS and self-reported asthma (OR increase per PRS quintile=1.16; p=3.17 x 10^{-26}) (Figure 2-6).

Somewhat unexpectedly, we observed that in our type 2 diabetes results the signal at *CDKAL1* was stronger than at *TCF7L2*, which is typically the top signal reported for type 2 diabetes GWAS. Hypothesizing that this might be due to the younger age of Genes for Good participants, we split the Genes for Good data at the median age to test for changes in diabetes risk between the below-median age and above-median age groups for the *TCF7L2* and *CKDAL1* variants (median age = 32; $cases_{Below\text{-}Median}$ = 65, $controls_{Below\text{-}Median}$ = 8,385; $cases_{Above\text{-}Median}$ = 722, $controls_{Above\text{-}Median}$ = 7,728). Although we saw a trend to a larger diabetes risk for carriers of the *TCF7L2* variant rs7903146 in the above-median group ($OR_{Below\text{-}Median}$ = 1.21, $OR_{Above\text{-}Median}$ = 1.34), we saw the same trend for carriers of the *CDKAL1* variant rs7756992 ($OR_{Below\text{-}Median}$ = 1.04, $OR_{Above\text{-}Median}$ = 1.37). Regardless, the differences between the below-median and above-median age groups for both SNPs were not significant (p > 0.05).

Discussion

We set out to recruit a large, diverse sample of engaged volunteers that might provide information about the diverse U.S. population. For each volunteer, we used surveys to collect health and behavioral data that might inform a variety of genomic research studies. With rapid and inexpensive recruitment, we have quickly developed a participant pool with which to validate the quality of the data. We are optimistic about our ability to obtain the large sample size required for valid genetic association studies of complex diseases and behaviors. With our current analysis of

20,232 individuals, we have successfully validated several known genotype-phenotype relationships and contributed to several consortium meta-analyses (Jiang et al., 2018; M. Liu et al., 2019; Sanchez-Roige et al., 2017; Zhan et al., 2017).

We have good representation with respect to geography, age, and gender, though our sample does have some noticeable differences from a sample of random U.S. adults. One characteristic that presents both an opportunity and a challenge is the younger age of Genes for Good participants compared to the U.S. adult population. While a younger demographic may be more interesting for some measures (behavioral data, activity levels), it will be less useful for others (age-associated cancers and development of other late-onset chronic disease). We do see slightly lower rates of the chronic conditions examined here compared to the general U.S. population, which we attribute to the lower average age of our participants; even if participants have the relevant risk factors, they may not have had the time to develop those long-term outcomes. For instance, we see much lower rates of heart attack in our participants despite comparable hypertension rates, and we see lower rates of type 2 diabetes despite comparable BMI (Figure 2-2). At the same time, Genes for Good's recruitment strategy may have led to an enrichment of individuals with certain rare diseases like Ehlers-Danlos Syndrome, perhaps because of network effects within these communities.

Most participants completed the minimum number of health history surveys required to receive a spit kit (15 surveys), with many going well above that number. Completion of daily tracking surveys was modest, with most genotyped participants completing only the minimum number required to obtain a spit kit. None of our surveys are mandatory and it is certainly possible that participants will avoid surveys that are more onerous or which they are not comfortable with, introducing ascertainment biases (for example, individuals who are not skilled at reasoning puzzles

might choose to skip the reasoning). The most completed surveys were generally those that appear higher in the list of available surveys within our app (Supplementary Figure 2-2; Supplementary Figure 2-3 provides additional details of survey completion rates).

Another challenge we face is that our sample is heavily skewed female. While targeted recruitment in the future may bring the gender distribution into balance, we also recognize the immediate potential to conduct a large-scale study of women's health and have implemented relevant survey measures regarding polycystic ovarian syndrome and pregnancy outcomes.

Genetic information, privacy, and ethics

There are a number of incentives for participation in Genes for Good besides the altruistic contribution and potential positive impact of genetics research on society. Firstly, we provide interactive graphs and visualizations by which users can compare their survey responses to those of other participants (examples in Figure 2-7 and Figure 2-8). Secondly, Genes for Good allows participants to view estimates of their genetic ancestry and download their raw genetic data, which some have argued should be the fundamental right of participants who contribute DNA to research (Lunshof et al., 2014). When downloading genetic data, we require participants to review a short slide show that explains the data we generate is suitable for a research study but does not meet the standards used for clinical genetic tests. We emphasize that, compared to the data used in clinical tests, research data might be more susceptible to error. Around 70% of participants with genotypes available have requested a download link for their raw genetic data, which we provide in 23andMe format, a format known to be widely accepted at third-party interpretation sites. Many participants have told us they upload their data to third-party sites to obtain more detailed ancestry estimates, find DNA relatives, and even seek health interpretation. A recent review paper (Hollands et al., 2016) investigating reactions to a clinical genetic risk assessment concluded that in general,

patients do not engage in risk-reducing behavior after receiving information about genetic predisposition. We expect that Genes for Good participants are unlikely to base major health or life decisions on the research-grade data we have returned. In addition, we will continue to develop Genes for Good web-based software applications to promote literacy of individuals about their genetic information.

Along with raw genetic data, we also return to participants their genetic ancestry information based on DNA analysis. The primary anticipated risk of the return of ancestry information is the discovery or suspicion of non-paternity and/or secret adoption by participants, i.e. discovering one's ancestry is inconsistent with what the participant knows about the ancestry of their supposedly biological parents. This has the potential to cause emotional or psychological stress on participants and their families, and we provide education about this risk during the informed consent.

Significance and future directions

The online platform implemented in Genes for Good is a viable study design for population-based genetic research. Now in the study's fourth year, we have already had great success in recruitment, health history survey analysis, and genetic analysis. We are currently exploring the more than 300 phenotypes collected so far and continue to participate in ongoing collaborations. As the sample size grows, our power to detect novel associations and our ability to contribute more meaningful data to researchers will increase.

The flexibility of the study design and our ongoing relationship with participants also makes it possible to implement new methods of data collection with relative ease. Additional data collection techniques are being developed and validated in a wide array of studies, including wireless sensors for continuous collection of data related to physical activity (Appelboom et al.,

2014; Dobkin & Dorsch, 2011), heart rate (El-Amrawy & Nounou, 2015), body temperature, sleep (Montgomery-Downs, Insana, & Bond, 2012), and GPS location logging to infer habits and environmental exposures (Glasgow et al., 2016). These measures and more are currently available through a combination of smartphone and wrist sensors (e.g. FitBit), and many more wireless sensors exist for more specialized tasks (e.g. breathalyzers, insulin levels, QT interval). These and other novel data collection methods are developing rapidly, holding great promise in the near future for the efficient collection of large quantities of precise longitudinal data with minimal participant burden. The implementation of such devices would facilitate the collection of tracking data within Genes for Good.

Having verified the quality of our data and several known associations with particular loci, we are now poised to begin exploring new genotypic-phenotypic relationships, such as those with behavioral and health tracking information. Research in other settings with Genes for Good data show that our results are consistent with those of prior studies. Liu et al. (2017) show that a PRS calculated from SSGAC's educational attainment data is effective in predicting 4% of the trait variance, which is consistent with previously reported out-of-sample predictive power for educational attainment (Branigan, McCallum, & Freese, 2013). We are also working to streamline data sharing methods to facilitate collaborations with other researchers. Finally, we are actively developing new tools to provide participants with meaningful data summaries at the personal and study level. We believe these steps will keep participants engaged and invested in the genetic research and will also help encourage longitudinal survey completions.

As we seek opportunities for long term funding of the study, we are currently not collecting spit kits from new participants. Although enrollment has decreased since we stopped offering spit kits (we currently collect only health survey responses), interest remains high, as evidenced by the

email inquiries we receive on a weekly basis. We plan to collect and genotype additional samples when future funding becomes available; when doing so, we expect to implement several changes to study protocol that will solve issues observed throughout the course of the study. For example, we noticed that survey completion correlates with the order that the survey appears on the app homepage (Supplementary Figure 2-2); something as simple as randomizing the order upon refresh may remedy this.

Supplemental Data

Supplemental data contain four tables and three figures.

Conflicts of Interest

Gonçalo R Abecasis is currently an employee of Regeneron Pharmaceuticals and the beneficiary of stock options and grants in Regeneron. Previously, he served on scientific advisory boards for 23andMe, Regeneron Pharmaceuticals and Helix.

Acknowledgements

This research has been conducted using the UK Biobank Resource under application number 24460 (specifically, calculation of PRS for type 1 diabetes and asthma was conducted using GWAS results from UK Biobank).

This study was supported by PI discretionary funds.

Administrative Support: Irene Felicetti, Stephanie Bachoura, Samantha Bachoura, Laura Baker

IT Support: Sean Caron

UM Sequencing Core: Robert Lyons, Susan Dagenais, Christopher Krebs, David Erdody

Web Resources

Genes for Good Facebook application: https://app.genesforgood.org

Genes for Good informational website: https://www.genesforgood.org

Full text of all Genes for Good survey: https://genesforgood.org/for_researchers

Information on Box compliance with HIPAA guidelines: https://www.box.com/industries/healthcare

Figures

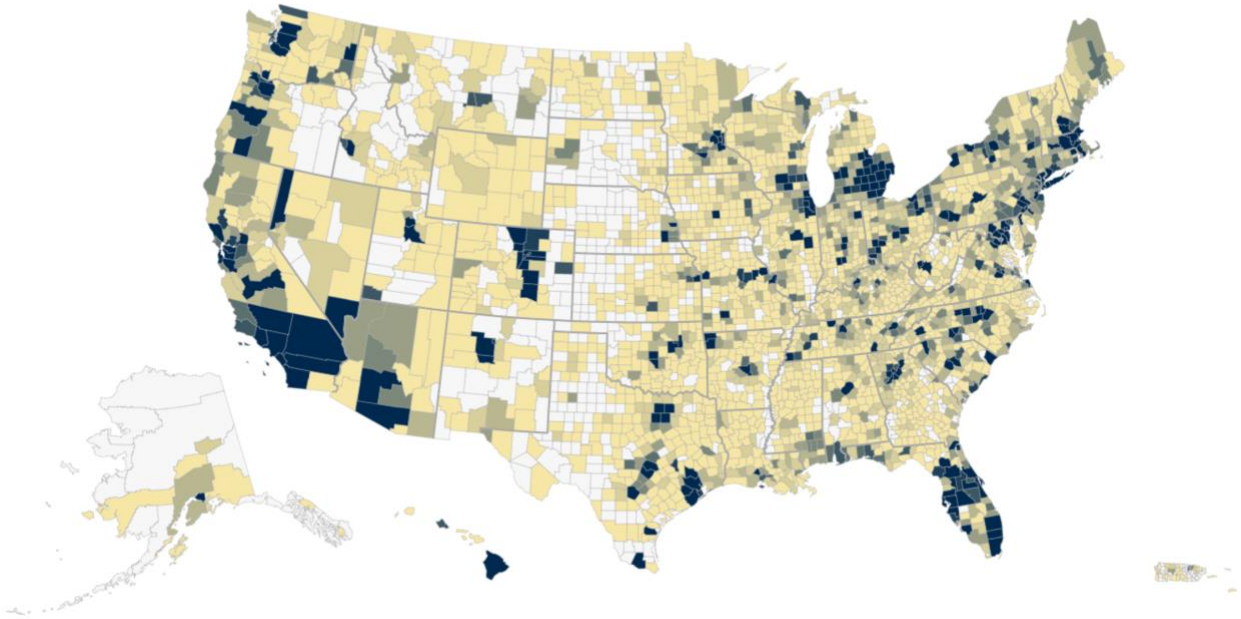


Figure 2-1 Geographic Distribution

The geographic distribution of Genes for Good participants as of October 2017. The colors indicate the number of participants who have logged into the app from that county, with darker colors representing higher density.

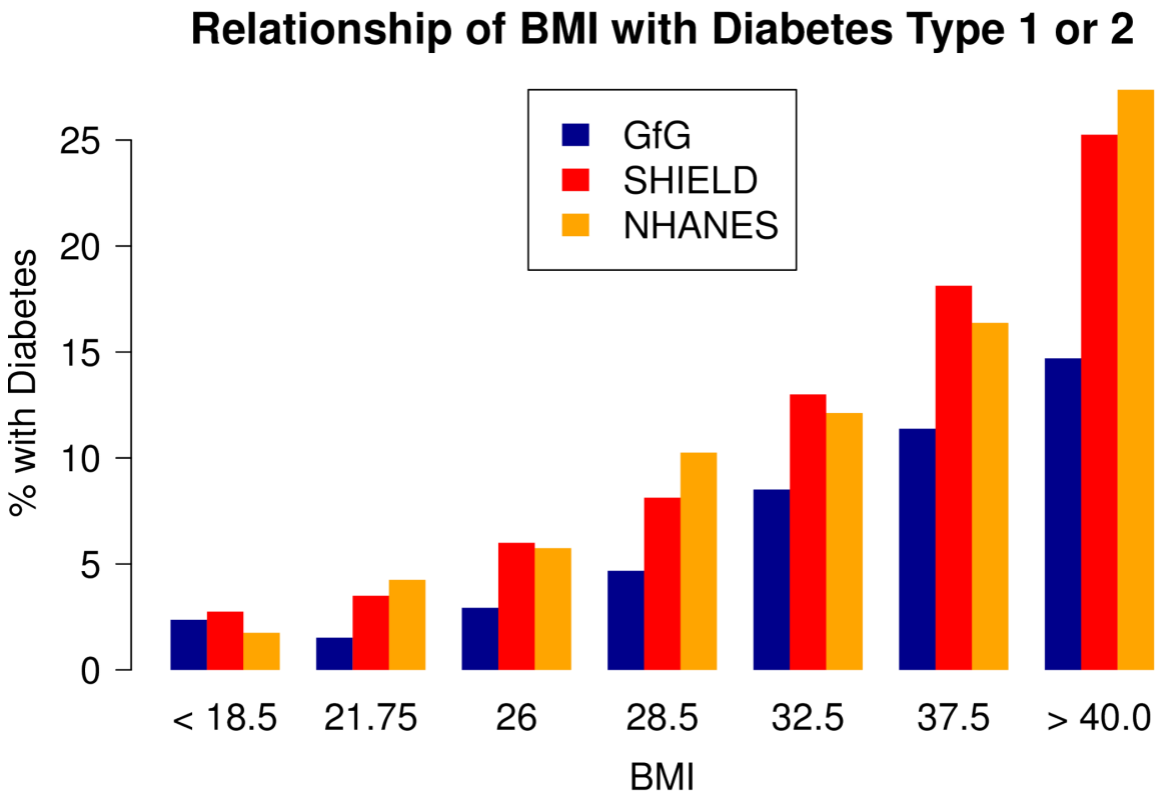


Figure 2-2 Relationship between BMI and diabetes rates

The relationship between BMI and diabetes rates in participants is consistent with that seen in the general U.S. population. Type 2 diabetes is a phenotype of particular interest because of its increasing prevalence, impact on cardiovascular health, and relatively well-characterized genetics. Here, we have compared the rates of diabetes in Genes for Good participants to the rates found in the nationally representative studies SHIELD and NHANES (Bays et al., 2007).

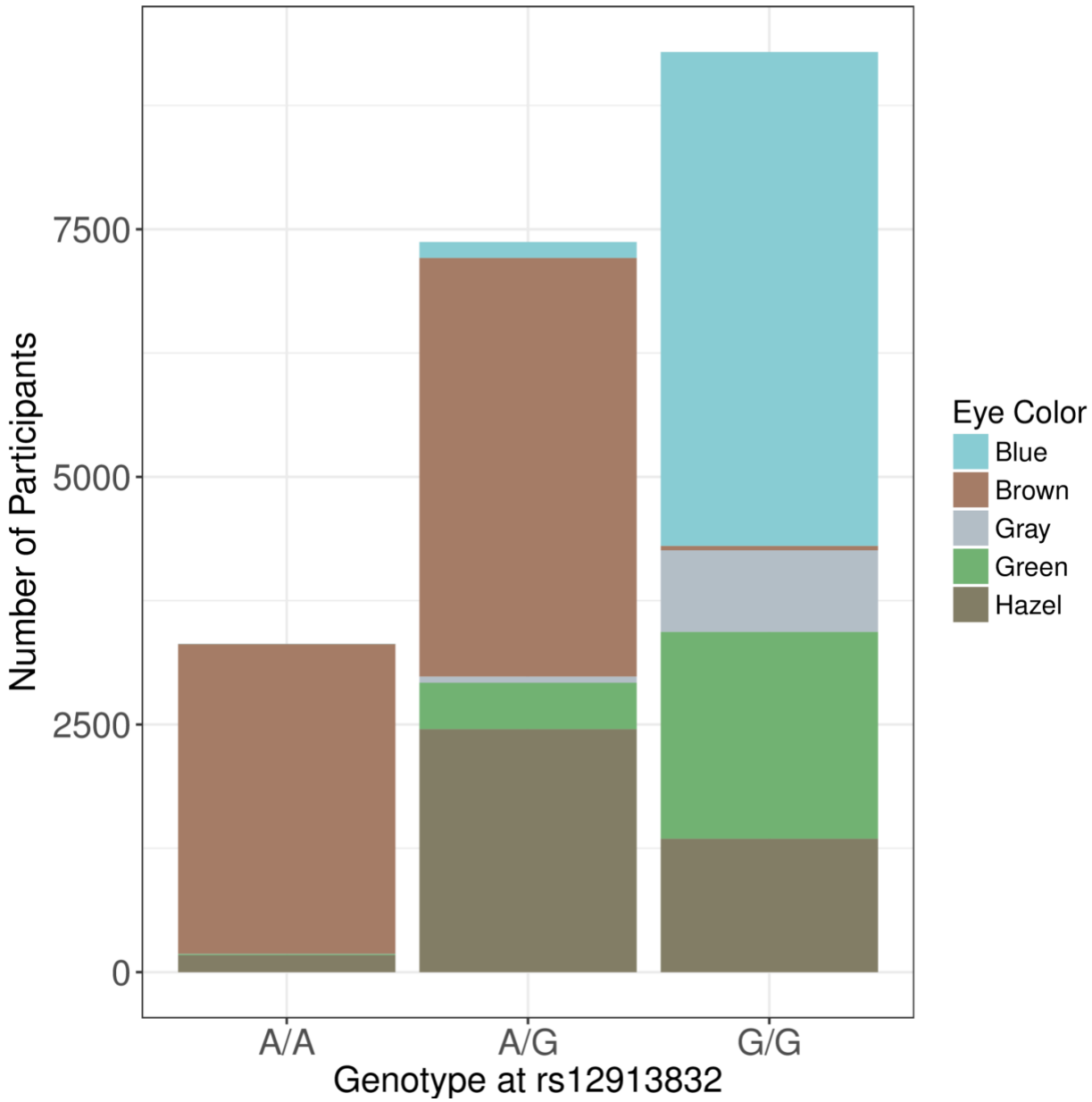


Figure 2-3 Eye Color by Genotype

Distribution of eye color among participants with different genotypes at rs12913832 (the top signal when performing GWAS using blue eye color in Genes for Good participants), a marker in the *HERC2* gene known to play a role in eye color determination.

SNP	Gene	P GfG	P Locke et al.	Comparison of Effect Estimates β and 95% CI
rs1558902	*FTO*	3×10^{-14}	1×10^{-156}	
rs6567160	*MC4R*	1×10^{-6}	7×10^{-59}	
rs13021737	*TMEM18*	6×10^{-7}	5×10^{-54}	
rs10938397	*GNPDA2*	8×10^{-7}	1×10^{-40}	
rs543874	*SEC16B*	1×10^{-9}	2×10^{-40}	
rs2207139	*TFAP2B*	2×10^{-6}	8×10^{-31}	
rs11030104*	*BDNF*	7×10^{-3}	7×10^{-30}	
rs3101336	*NEGR1*	9×10^{-3}	3×10^{-26}	
rs7138803	*BCDIN3D*	5×10^{-3}	5×10^{-26}	
rs10182181	*ADCY3*	1×10^{-5}	8×10^{-26}	

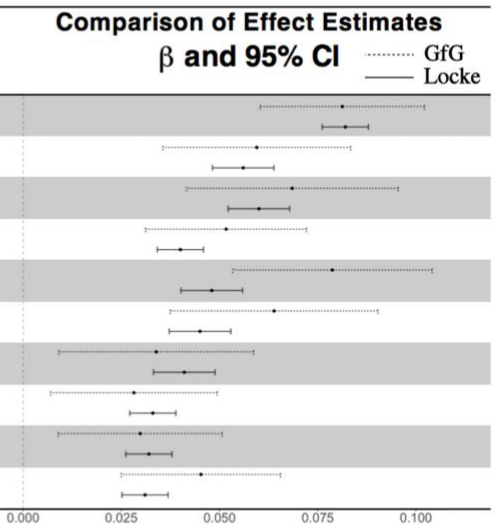


Figure 2-4 BMI GWAS Effect Sizes

Effect size estimates of a GWAS for BMI in our study sample compared to findings from a meta-analysis. We compare effect estimates from Genes for Good to published findings from the Locke et al 2015 meta-analysis of BMI GWAS (Locke et al., 2015). Specifically, we looked at the top 10 reported signals and were able to replicate all of these effects in direction and nominal significance ($p < 0.05$). The forest plot on the right compares effect size estimates across studies; the dashed lines represent the confidence intervals around the Genes for Good estimates, while the solid lines represent results from Locke et al. Given the relatively small sample size available in this data freeze, our estimates have fairly wide confidence limits. However, Locke's estimates are completely contained within our limits for 8 of 10 SNPs.
*Imputed variant

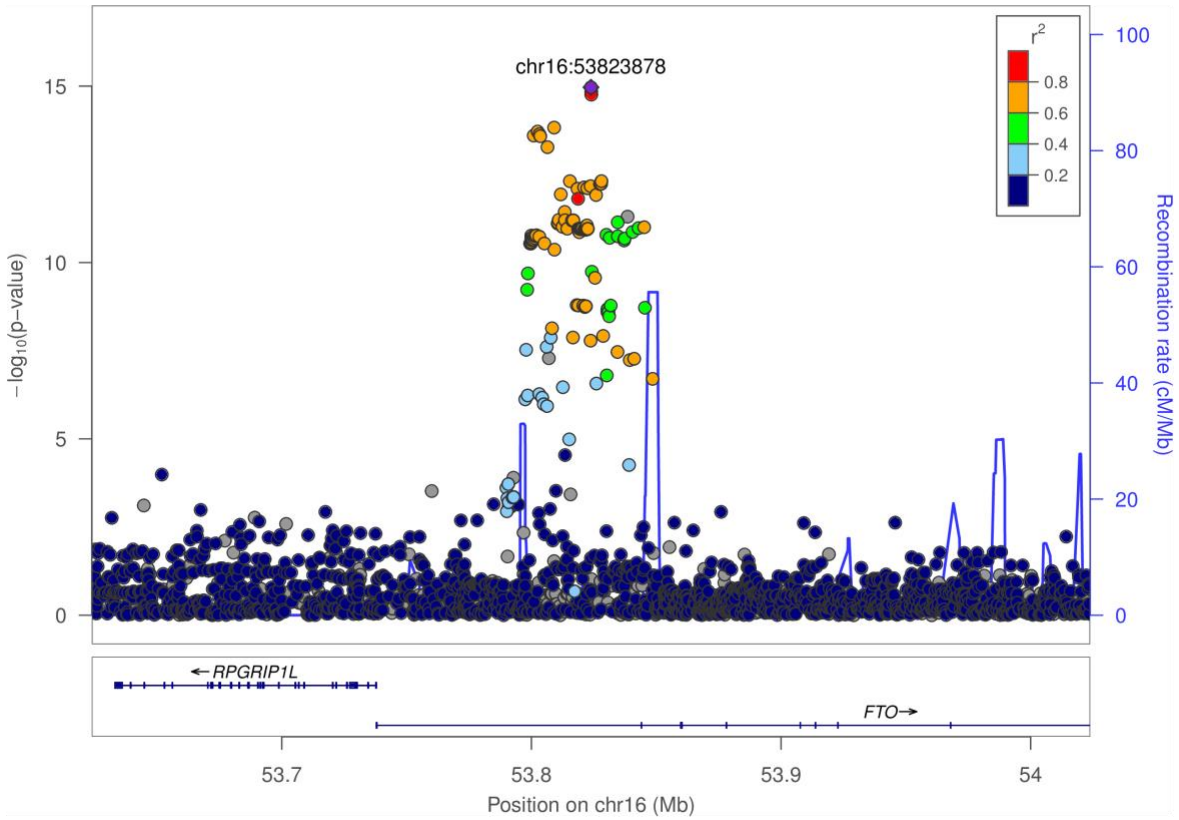


Figure 2-5 LocusZoom Plot of *FTO*

LocusZoom plot showing single-variant association results for BMI in the gene *FTO*. This result is consistent with other studies that reported their strongest evidence for association in this gene. The effect size at the nearby SNP rs1558902 (0.081) was consistent with the effect size (0.081) reported previously in Locke et al. (2015).

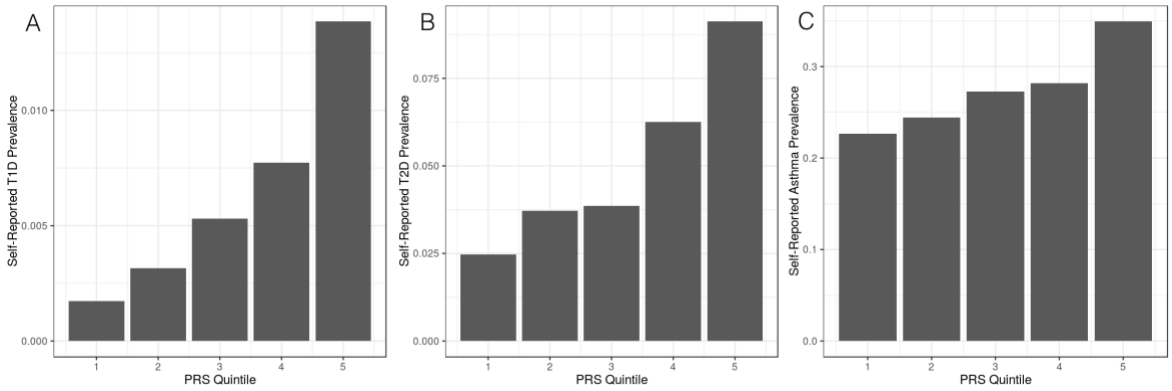


Figure 2-6 Genetic Risk of Diabetes

Prevalence for self-reported Type 1 and Type 2 diabetes across polygenic risk score quintiles (five bins of equal sample size). An increase in the genetic risk score is associated with increasing prevalence of disease. We also evaluated associations between polygenic risk score quintile and Type 1 diabetes, Type 2 diabetes, and asthma status, adjusted for age and sex. We found that all three self-reported traits were significantly associated with calculated PRS quintile (p_{T1D}=5.13 x 10^{-9}, p_{T2D}=7.63 x 10^{-37}, p_{asthma}= 3.17 x 10^{-26}).

Figure 2-7 Example Health History result

An example of how participants' results to the Personality survey are displayed within the Genes for Good app. The bars show this participant's percentile scores on the five personality attributes measured by the survey.

Figure 2-8 Example daily tracking result

An example of how participants' answers to the daily sleep tracking survey are displayed, showing (A) average hours of sleep for this participant, compared to other participants of the same age range and sex, and to all other Genes for Good participants, (B) average hours of sleep reported for different days of the week when this participant has taken the survey, (C) average hours of sleep over the past 7 days, past 30 days, and over all responses from this participant, and (D) average hours of sleep reported for different days of the week for all Genes for Good participants stratified by sex.

Tables

Table 2-1 Demographics

	Genes for Good[a]	U.S. Population[b]	Facebook-using population[c]
Age			
Median, years	33	44[d]	
18-24	17.0%	13.2%	19.5%
25-34	37.1%	17.1%	27.0%
35-44	21.6%	16.4%	19.6%
45-54	11.9%	18.3%	16.5%
55+	12.4%	35.5%	17.4%
Sex			
Male	25.9%	49.2%	49%
Female	74.1%	50.8%	51%

[a]Data source for our study data is based on all valid responses as of August 9th, 2017
[b]Data for U.S. population from the 2010 U.S. Census (Howden & Meyer, 2011)
[c]Data for Facebook population from Statistica (Distribution of Facebook users in the United States as of January 2017, by age group and gender, 2017; eMarketer & Squarespace, 2017)
[d]Median age of U.S. persons over age 18 reported in the U.S. 2010 Census

Table 2-2 Chronic health indicators in study sample compared to overall United States population

	Genes for Good[a]	U.S. Population[b]
BMI, mean, kg/m^2	29.80	29.38
Underweight (BMI < 18.5)	1.9%	1.6%
Normal weight (BMI 18.5 - 24.9)	31.6%	27.2%
Overweight (BMI 25 - 29.9)	26.0%	31.6%
Obese (BMI ≥ 30)	40.4%	39.7%
High cholesterol	26.1%	29.3%
Hypertension	24.9%	29%
Previous stroke	1.3%	2.9%
Previous MI	1.5%	4.5%
Diabetes (Type 1 or 2)	6.5%	9.3%
Current smoker	17.0%	15.1%

[a]Data source for our study data is based on all valid responses as of August 9th, 2017
[b]Data from nationally representative samples to determine U.S. rates of obesity (CDC & NCHS, 2017), high cholesterol, hypertension (Nwankwo, Yoon, Burt, & Gu, 2013), stroke (Mozaffarian et al., 2015) , MI, diabetes, and smoking (Ward, Clarke, Nugent, & Schiller, 2016)

Table 2-3 Income distribution

Income Category	Genes for Good (%)	US Population[a] (%)	23andMe[b] (%)
Less than $35,000	28.0	30.2	10.2
$35,000 to $50,000	18.9	12.9	7.2
$50,000 to $75,000	19.8	17.0	13.9
$75,000 to $100,000	14.5	12.3	14.7
More than $100,000	18.9	27.7	54.0

Distribution of household income among Genes for Good participants based on answers to the Demographics survey as of August 9, 2017 compared to the general U.S. population.
[a]Data from U.S. Census Table H-17 (Semega, Fontenot, & Kollar, 2017)
[b]Data describing 23andMe research cohort approximated from 2011 ASHG poster (Tung et al., 2011)

Supplementary Tables

Supplementary Table 2-1 Diabetes cases and controls demographics

Diabetes Cases and Controls, Genotyped Samples				
	GfG Cases (N=948)	GfG Controls (N=16,581)	NHANES Cases (N=809)	NHANES Controls (N=4,796)
BMI	35.71 (8.63)	29.11 (7.79)	32.58 (7.75)	28.80 (6.83)
Underweight	1.0%	1.9%	0.5%	1.7%
Normal weight	8.4%	34.6%	12.5%	30.0%
Overweight	17.0%	27.4%	29.0%	32.0%
Obese	73.6%	36.1%	58.0%	36.3%
Age				
<21	1.1%	6.3%	0.1%	7.0%
21-30	8.1%	40.0%	2.8%	19.1%
31-40	21.3%	28.3%	5.7%	18.3%
41-50	20.1%	11.7%	12.0%	16.4%
51-60	27.7%	8.2%	21.5%	14.9%
61-70	16.5%	4.3%	31.1%	12.3%
>70	5.2%	1.1%	26.7%	11.8%
Sex				
Female	65.8%	68.5%	45.7%	52.8%
Male	34.2%	31.5%	54.3%	47.2%
Race				
Hispanic	7.4%	8.4%	38.1%	29.8%
Asian	1.0%	3.9%	8.9%	12.5%
Black	3.1%	2.6%	23.6%	20.8%
White	79.8%	76.2%	26.1%	33.1%
Multiracial/Other	8.7%	8.9%	3.3%	3.9%
Income				
<$35K	33.7%	26.6%	50.6%	38.7%
$35K-$75K	38.0%	38.1%	30.3%	31.6%
$75K-$100K	14.4%	15.3%	6.9%	10.8%
>$100K	13.9%	20.0%	12.2%	19.0%
Education				
No HS	3.5%	2.0%	32.8%	22.7%
HS Diploma	16.3%	11.3%	21.9%	23.0%
Some college or Associate's degree	45.8%	41.3%	27.6%	29.6%
Bachelor's or higher	34.5%	45.5%	17.7%	26.2%

Comparison of Genes for Good cohort (genotyped diabetes cases and controls) to NHANES (CDC & NCHS, 2017) cohort.

Supplementary Table 2-2 Genome-wide significant hits for various pigmentation and health phenotypes

	Locus	CHR	POS (hg38)	Nearest Genes	EA	EAF	N	Beta	SE	p-value
Hair Color Depth	rs12203592	6	396,321	*IRF4*	T	0.149	19,143	0.42	0.01	2.8×10^{-289}
	rs1129038	15	28,111,713	*HERC2*	T	0.697	19,143	-0.20	0.01	1.8×10^{-97}
	rs17184180	14	92,314,043	*SLC24A4*	A	0.407	19,143	-0.13	0.01	1.7×10^{-56}
	rs12821256	12	88,934,558	*KITLG*	C	0.091	19,143	-0.20	0.01	3.8×10^{-46}
	rs16891982	5	33,951,588	*SLC45A2*	G	0.876	19,143	-0.22	0.02	3.2×10^{-39}
	rs72928978	11	69,063,896	*TPCN2*	A	0.102	19,143	-0.16	0.01	2.1×10^{-32}
	rs1805008	16	89,919,736	*MC1R*	T	0.055	19,143	-0.21	0.02	1.2×10^{-31}
	rs80293268	1	8,147,519	*ERRFI1, SLC45A1*	C	0.038	19,143	-0.18	0.02	9.3×10^{-18}
	rs1205312	20	34,261,610	*ASIP*	G	0.934	19,143	0.13	0.02	7.3×10^{-15}
	rs71443018	2	28,390,435	*FOSL2*	C	0.043	19,143	-0.14	0.02	2.9×10^{-13}
	rs17349283	2	221,225,077	*EPHA4*	G	0.427	19,143	-0.06	0.01	1.1×10^{-11}
	rs1126809	11	89,284,793	*TYR*	A	0.252	19,143	-0.06	0.01	2.0×10^{-11}
	rs9544609	13	77,818,521	*EDNRB, SLAIN1*	A	0.561	19,143	-0.05	0.01	1.9×10^{-10}
	rs112232483	17	47,879,446	*SP2, SP6*	C	0.257	19,143	-0.05	0.01	3.0×10^{-8}
	rs17248377	5	53,820,293	*ARL15, NDUFS4*	A	0.223	19,143	-0.05	0.01	4.9×10^{-8}
	Locus	**CHR**	**POS (hg38)**	**Nearest Genes**	**EA**	**EAF**	**N**	**Beta**	**SE**	**p-value**
Hair Texture	rs36010924	1	152,116,368	*TCHH*	G	0.177	19,983	-0.27	0.01	9.0×10^{-170}
	rs80293268	1	8,147,519	*ERRFI1, SLC45A1*	C	0.038	19,983	-0.21	0.02	1.9×10^{-26}
	rs121908120	2	218,890,289	*WNT10A*	A	0.021	19,983	0.26	0.03	3.3×10^{-24}
	rs56210557	20	63,533,171	*PTK6*	A	0.055	19,983	0.15	0.02	2.6×10^{-20}
	rs12951078	17	40,754,960	*KRT25, KRTAP*	A	0.529	19,983	-0.07	0.01	1.7×10^{-18}
	rs11170678	12	53,760,390	*HOXC13*	G	0.235	19,983	-0.07	0.01	6.8×10^{-16}
	rs4149433	2	108,380,806	*SULT1C4, EDAR*	T	0.056	19,983	-0.17	0.02	6.0×10^{-15}
	rs1419295	10	8,259,689	*GATA3*	G	0.140	19,983	-0.08	0.01	5.1×10^{-13}
	*rs1918719	8	116,293,163	*EIF3H*	C	0.209	19,983	-0.06	0.01	7.0×10^{-11}
	rs62405519	6	10,293,990	*OFCC1*	A	0.582	19,983	0.05	0.01	3.1×10^{-10}

	*rs7499783	16	79,768,781	*MAFTRR*	C	0.183	19,983	0.06	0.01	3.3×10^{-9}
	Locus	**CHR**	**POS (hg38)**	**Nearest Genes**	**EA**	**EAF**	**N**	**Beta**	**SE**	**p-value**
Skin Sun Response	rs12203592	6	396,321	*IRF4*	T	0.138	17,633	0.22	0.01	2.3×10^{-66}
	rs1805007	16	89,919,709	*MC1R*	T	0.060	17,633	0.26	0.02	3.0×10^{-44}
	rs62211989	20	33,950,585	*RALY*	C	0.059	17,633	0.19	0.02	9.2×10^{-26}
	rs16891982	5	33,951,588	*SLC45A2*	G	0.868	17,633	0.17	0.02	7.0×10^{-24}
	rs1126809	11	89,284,793	*TYR*	A	0.243	17,633	0.09	0.01	2.1×10^{-18}
	rs12350739	9	16,885,019	*BNC2*	A	0.508	17,633	0.08	0.01	1.0×10^{-17}
	rs116858369	7	17,211,369	*AHR*	A	0.011	17,633	0.24	0.04	9.5×10^{-9}
	rs117886461	15	27,985,232	*OCA2*	A	0.008	17,633	0.26	0.05	4.1×10^{-8}
	Locus	**CHR**	**POS (hg38)**	**Nearest Genes**	**EA**	**EAF**	**N**	**OR**	**CI**	**p-value**
Blue Eyes	rs1129038	15	28,111,713	*HERC2*	T	0.701	19,982	2.84	(2.66, 3.02)	2.7×10^{-232}
	rs1126809	11	89,284,793	*TYR*	A	0.253	19,982	1.54	(1.45, 1.62)	1.9×10^{-51}
	rs4904866	14	92,302,159	*SLC24A4*	T	0.407	19,982	1.47	(1.40, 1.55)	7.7×10^{-51}
	rs12203592	6	396,321	*IRF4*	T	0.151	19,982	1.45	(1.36, 1.55)	1.1×10^{-27}
	rs10960730	9	12,631,099	*TYRP1*	G	0.573	19,982	1.27	(1.2, 1.34)	4.7×10^{-19}
	rs16891982	5	33,951,588	*SLC45A2*	G	0.880	19,982	1.93	(1.66, 2.24)	2.2×10^{-17}
	rs9723267	22	45,969,677	*WNT7B*	T	0.315	19,982	1.17	(1.11, 1.23)	2.2×10^{-8}
	Locus	**CHR**	**POS (hg38)**	**Nearest Genes**	**EA**	**EAF**	**N**	**Beta**	**SE**	**p-value**
Height	rs62346126	4	144,639,014	*HHIP*	A	0.778	19,581	0.07	0.01	4.0×10^{-12}
	rs9892365	17	61,414,023	*TBX2*	G	0.664	19,581	-0.05	0.01	3.6×10^{-11}
	rs13077048	3	141,388,112	*ZBTB38*	T	0.417	19,581	0.05	0.01	1.5×10^{-10}
	rs1897112	2	55,888,333	*EFEMP1*	C	0.235	19,581	-0.06	0.01	7.3×10^{-10}
	rs584961	11	75,566,583	*SERPINH1*	G	0.893	19,581	-0.08	0.01	1.6×10^{-9}
	rs9634212	12	93,599,490	*SOCS2*	A	0.216	19,581	0.06	0.01	3.1×10^{-9}
	rs3760318	17	30,920,697	*CENTA2*	A	0.367	19,581	-0.05	0.01	3.1×10^{-9}
	rs1000972	20	6,641,070	*BMP2, CASC20*	A	0.653	19,581	-0.05	0.01	6.3×10^{-9}
	rs11205303	1	149,934,520	*MTMR11*	C	0.380	19,581	0.05	0.01	8.3×10^{-9}

	rs2707450	4	17,940,937	*LCORL*	T	0.733	19,581	-0.05	0.01	1.3×10^{-8}
	rs57026767	6	34,251,921	*C6orf1*	T	0.807	19,581	-0.06	0.01	1.4×10^{-8}
	rs798489	7	2,762,169	*GNA12*	T	0.242	19,581	-0.05	0.01	4.1×10^{-8}
	rs244711	5	177,082,192	*ZNF346, FGFR4*	T	0.692	19,581	0.05	0.01	4.2×10^{-8}
	Locus	**CHR**	**POS (hg38)**	**Nearest Genes**	**EA**	**EAF**	**N**	**Beta**	**SE**	**p-value**
BMI	rs28432761	16	53,789,966	*FTO*	C	0.495	19,278	0.08	0.01	1.1×10^{-15}
	rs62107261	2	422,144	*FAM150B*	C	0.039	19,278	-0.18	0.03	3.5×10^{-12}
	rs539515	1	177,919,890	*SEC16B*	C	0.192	19,278	0.08	0.01	7.2×10^{-10}
	rs55835921	3	186,113,685	*ETV5*	C	0.178	19,278	-0.08	0.01	5.7×10^{-9}
	rs118178156	8	86,283,096	*SLC7A13, WWP1*	T	0.007	19,278	0.34	0.06	1.5×10^{-8}
	rs185527056	2	12,275,975	*LPIN1, TRIB2*	G	0.005	19,278	-0.40	0.07	2.4×10^{-8}
	Locus	**CHR**	**POS (hg38)**	**Nearest Genes**	**EA**	**EAF**	**N**	**OR**	**CI**	**p-value**
Type 1 Diabetes	rs9273363	6	32,658,495	*HLA-DQB1*	A	0.263	17,529	3.77	(2.69, 5.29)	1.3×10^{-14}
	Locus	**CHR**	**POS (hg38)**	**Nearest Genes**	**EA**	**EAF**	**N**	**OR**	**CI**	**p-value**
Type 2 Diabetes	rs12660618	6	20,677,079	*CDKAL1*	T	0.169	17,529	1.52	(1.32, 1.76)	1.5×10^{-8}

*Associations not reported in previous studies.

All associations are consistent with findings in previous studies (McMahon et al., 2018) except for the hair texture hits at rs1918719 and rs7499783. CHR, chromosome; POS38, build 38 chromosome position; EA, effect allele; EAF, effect allele frequency; N, number of participants included in analysis; SE, standard error.

Supplementary Table 2-3 Comparison of Genes for Good top GWAS hits to previously reported results

				Reference				GFG			
	Published Locus	**Nearest Gene**	**EA**	**Trait**	**N**	**Beta**	**p-value**	**Trait**	**N**	**Beta**	**p-value**
Hair Color Depth	rs12913832	*HERC2*	A	1 - Blond	283,410	0.50	$<10^{-100}$	1 - Blond	19,143	0.20	1.0×10^{-96}
	rs12203592	*IRF4*	T	2 - Red	283,410	0.38	$<10^{-100}$	2 - Red	19,143	0.42	2.8×10^{-289}
	rs17184180	*SLC24A4*	A	3 - Light Brown	281,197	-0.20	$<10^{-100}$	3 - Light Brown	19,143	-0.13	1.7×10^{-56}
				4 - Dark Brown				4 - Dark Brown			
				5 - Black				5 - Black			
	Reference Study: (Hysi et al., 2018)										
				Reference				**GFG**			
	Published Locus	**Nearest Gene**	**EA**	**Trait**	**N**	**Beta**	**p-value**	**Trait**	**N**	**Beta**	**p-value**
Hair Texture	rs17646946	*TCHHL1*	A	Level 1 - Straight	28,964	−0.21	5.8×10^{-134}	1 - Straight	19,983	-0.25	1.3×10^{-153}
	rs74333950	*WNT10A*	G	Level 2 - Wavy	28,964	0.06	9.5×10^{-18}	2 - Wavy	19,983	0.09	1.1×10^{-17}
	rs11203346	*PADI3*	G	Level 3 - Curly	28,964	−0.07	9.2×10^{-17}	3 - Curly	19,983	-0.06	1.5×10^{-7}
								4 - Very Tight Curls			
	Reference Study: (F. Liu et al., 2018)										
				Reference				**GFG**			
	Published Locus	**Nearest Gene**	**EA**	**Trait**	**N**	**OR**	**p-value**	**Trait**	**N**	**Beta**	**p-value**
Skin Sun Response	rs12203592	*IRF4*	T	Level 1 - Low Tan Response	121,296	1.74	1.1×10^{-581}	1 - Never burns	17,633	0.22	2.3×10^{-66}
	rs369230	*MC1R*	G	Level 2 - High Tan Response	121,296	1.60	1.0×10^{-522}	2 - Burns rarely	17,633	0.10	4.2×10^{-24}
	rs6059655	*RALY/ASIP*	A		121,296	1.69	1.4×10^{-315}	3 - Burns moderately	17,633	0.18	4.3×10^{-24}
								4 - Often burns			
								5 - Always burns			
	Reference Study: (Visconti et al., 2018)										
				Reference				**GFG**			
	Published Locus	**Nearest Gene**	**EA**	**Trait**	**N**	**OR**	**p-value**	**Trait**	**N**	**OR**	**p-value**
Blue Eye Color	rs8039195	*HERC2*	T	Blue vs. green/brown	5,130	13.10	3.9×10^{-129}	5,148 blue vs.	19,982	2.88	3.7×10^{-132}
	rs4904868	*SLC24A4*	T		5,130	0.67	2.5×10^{-14}	14,834 not-blue	19,982	0.74	4.2×10^{-32}
	rs1408799	*TYRP1*	T		5,130	0.71	1.5×10^{-9}		19,982	0.79	8.5×10^{-17}
	Reference Study: (Sulem et al., 2008)										
				Reference				**GFG**			

	Published Locus	Nearest Gene	EA	Trait	N	Beta	p-value	Trait	N	Beta	p-value
Height	rs724016	*ZBTB38*	A	Height (m)	252,972	-0.08	3.2×10^{-158}	Height (in)	19,581	-0.05	7.1×10^{-10}
	rs143384	*GDF5*	A		247,786	-0.08	1.2×10^{-121}		19,581	-0.04	1.3×10^{-7}
	rs8756	*HMGA2*	A		253,008	-0.06	4.5×10^{-90}		19,581	-0.03	5.2×10^{-5}
	Reference Study: (Wood et al., 2014)										
				Reference				**GFG**			
	Published Locus	**Nearest Gene**	**EA**	**Trait**	**N**	**Beta**	**p-value**	**Trait**	**N**	**Beta**	**p-value**
BMI	rs1558902	*FTO*	A	BMI (kg/m²)	336,974	0.08	1.1×10^{-156}	BMI (kg/m²)	19,278	0.08	2.6×10^{-14}
	rs6567160	*MC4R*	C		339,006	0.06	6.7×10^{-59}		19,278	0.06	1.1×10^{-6}
	rs13021737	*TMEM18*	G		333,169	0.06	5.4×10^{-54}		19,278	0.07	6.4×10^{-7}
	Reference Study: (Locke et al., 2015)										
				Reference				**GFG**			
	Published Locus	**Nearest Gene**	**EA**	**Trait**	**N**	**OR**	**p-value**	**Trait**	**N**	**OR**	**p-value**
Type 1 Diabetes	rs9273364	*HLA-DQB1*	G	2,660 cases	391,416	2.26	2.8×10^{-142}	106 cases	17,529	3.77	1.3×10^{-14}
	rs2596560	*MICA*	C	288,756 controls	391,416	1.63	1.1×10^{-46}	17,424 controls	17,529	2.20	4.4×10^{-6}
	rs689	*INS-IGF2*	T		391,416	1.33	8.1×10^{-20}		17,529	1.27	0.12
	Reference Study: (Bycroft et al., 2018)										
				Reference				**GFG**			
	Published Locus	**Nearest Gene**	**EA**	**Trait**	**N**	**OR**	**p-value**	**Trait**	**N**	**OR**	**p-value**
Type 2 Diabetes	rs7903146	*TCF7L2*	T	48,286 cases	298,957	1.27	1.4×10^{-212}	848 cases	17,529	1.30	8.4×10^{-6}
	rs1558902	*FTO*	A	250,671 controls	298,957	1.12	1.6×10^{-50}	16,898 controls	17,529	1.24	8.8×10^{-5}
	rs7756992	*CDKAL1*	A		298,957	0.90	1.2×10^{-41}		17,529	0.74	5.7×10^{-7}
	Reference Study: DIAGRAM Consortium T2D GWAS meta-analysis - European Summary Statistics (Mahajan et al., 2018)										
				Reference				**GFG**			
	Published Locus	**Nearest Gene**	**EA**	**Trait**	**N**	**OR**	**p-value**	**Trait**	**N**	**OR**	**p-value**
Asthma	rs2952156	*ERBB2, PGAP3, MIEN1*	G	19,954 cases	127,669	0.86	7.6×10^{-29}	4,378 cases	16,093	0.93	0.01
	rs9272346	*HLA-DRB1,*	A	107,715 controls	127,669	1.16	4.8×10^{-28}	11,715 controls	16,093	1.09	5.6×10^{-4}

		HLA-DQA1									
	rs10455025	*SLC25A46, TSLP*	C		127,669	1.15	2.0×10^{-25}		16,093	1.06	0.04
	Reference Study: (Demenais et al., 2018)										

*Associations not reported in previous studies.

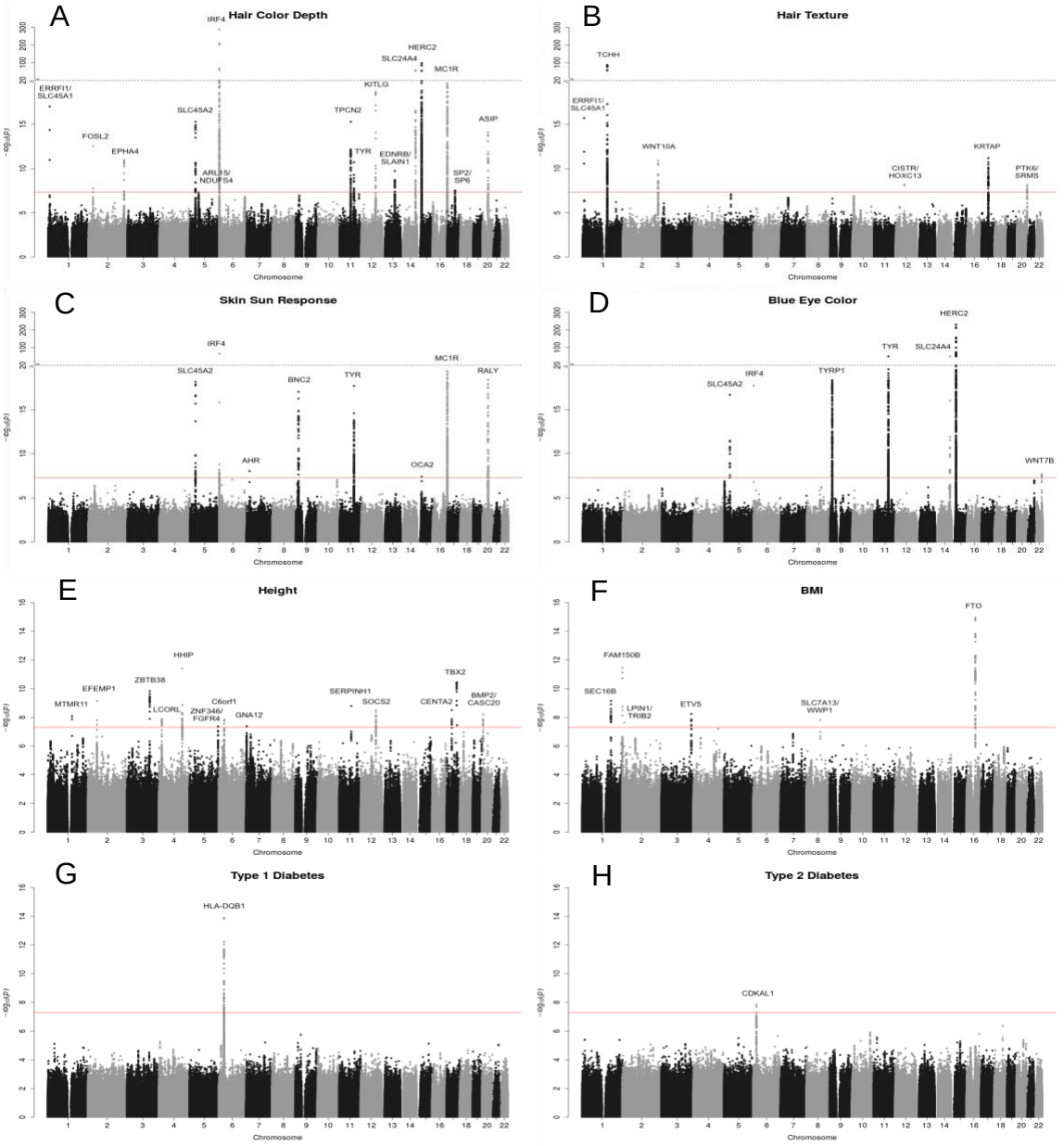
Replications of the top three hits from various studies of pigmentation and health traits (Bycroft et al., 2018; Demenais et al., 2018; Hysi et al., 2018; F. Liu et al., 2018; Locke et al., 2015; Mahajan et al., 2018; Sulem et al., 2008; Visconti et al., 2018; Wood et al., 2014). Direction of effect for all variants is consistent between the reference studies and Genes for Good, and most Genes for Good results attain at least nominal significance ($p < 0.05$). EA, effect allele; N, number of participants included in analysis; OR, odds (log-additive) ratio.

Supplementary Table 2-4 Comparison of Genes for Good asthma results to previously reported results

				Reference			GFG				
				19,954 European-ancestry cases 107,715 European-ancestry controls			Unadjusted 4,378 cases 11,715 controls		Adjusted 652 cases 15,441 controls		
	Published Locus	**Nearest Gene**	**EA**	**N**	**OR**	**p-value**	**N**	**Unadj. OR**	**Adj. OR**	**Power (α=0.05)**	**p-value**
Asthma	rs2952156	*ERBB2, PGAP3, MIEN1*	G	127,669	0.86	7.6×10^{-29}	16,093	0.93	0.80	0.65	0.005
	rs9272346	*HLA-DRB1, HLA-DQA1*	A	127,669	1.16	4.8×10^{-28}	16,093	1.09	1.84	0.67	0.001
	rs10455025	*SLC25A46, TSLP*	C	127,669	1.15	2.0×10^{-25}	16,093	1.06	1.46	0.61	0.045
	rs1420101	*IL1RL1, IL1RL2, IL18R1*	T	127,669	1.12	9.1×10^{-20}	16,093	1.02	1.11	0.46	0.565
	rs992969	*RANBP6, IL33*	G	127,669	0.85	1.1×10^{-17}	16,093	0.96	0.90	0.67	0.205
	rs20541	*IL13, RAD50, IL4*	G	127,669	0.89	1.4×10^{-14}	16,093	0.94	0.83	0.37	0.043
	rs2033784	*SMAD3, SMAD6, AAGAB*	G	127,669	1.11	2.5×10^{-14}	16,093	1.005	1.03	0.37	0.862
	rs2325291	*BACH2, GJA10, MAP3K7*	A	127,669	0.91	8.6×10^{-13}	16,093	0.98	0.94	0.31	0.454
	rs7927894	*EMSY, LRRC32*	T	127,669	1.10	3.5×10^{-11}	16,093	1.04	1.32	0.35	0.118
	rs11071558	*RORA, NARG2, VPS13C*	G	127,669	0.89	1.9×10^{-10}	16,093	0.91	0.76	0.26	0.011
	rs17806299	*CLEC16A, DEXI, SOCS1*	A	127,669	0.90	2.1×10^{-10}	16,093	0.99	0.97	0.28	0.729
	rs17637472	*ZNF652, PHB*	A	127,669	1.08	3.3×10^{-9}	16,093	1.06	1.46	0.25	0.035
	rs1233578	*GPX5, TRIM27*	G	127,669	1.11	5.3×10^{-9}	16,093	1.05	1.37	0.23	0.184
	rs2589561	*GATA3, CELF2*	G	127,669	0.90	1.4×10^{-8}	16,093	0.996	0.99	0.29	0.914
	rs2855812	*MICB, HCP5, MCCD1*	T	127,669	1.10	1.7×10^{-8}	16,093	1.02	1.12	0.28	0.589
	rs12543811	*TPD52, ZBTB10*	A	127,669	0.93	3.4×10^{-8}	16,093	0.98	0.94	0.22	0.447
	rs167769	*STAT6, NAB2, LRP1*	T	127,669	1.08	1.6×10^{-7}	16,093	1.07	1.55	0.25	0.014
	rs7705042	*NDFIP1, GNDPA1, SPRY4*	A	127,669	1.08	1.6×10^{-6}	16,093	1.01	1.09	0.24	0.637
	Reference Study: (Demenais et al., 2018)Demenais et al (2018) Nature Genetics **50**, 42-53										

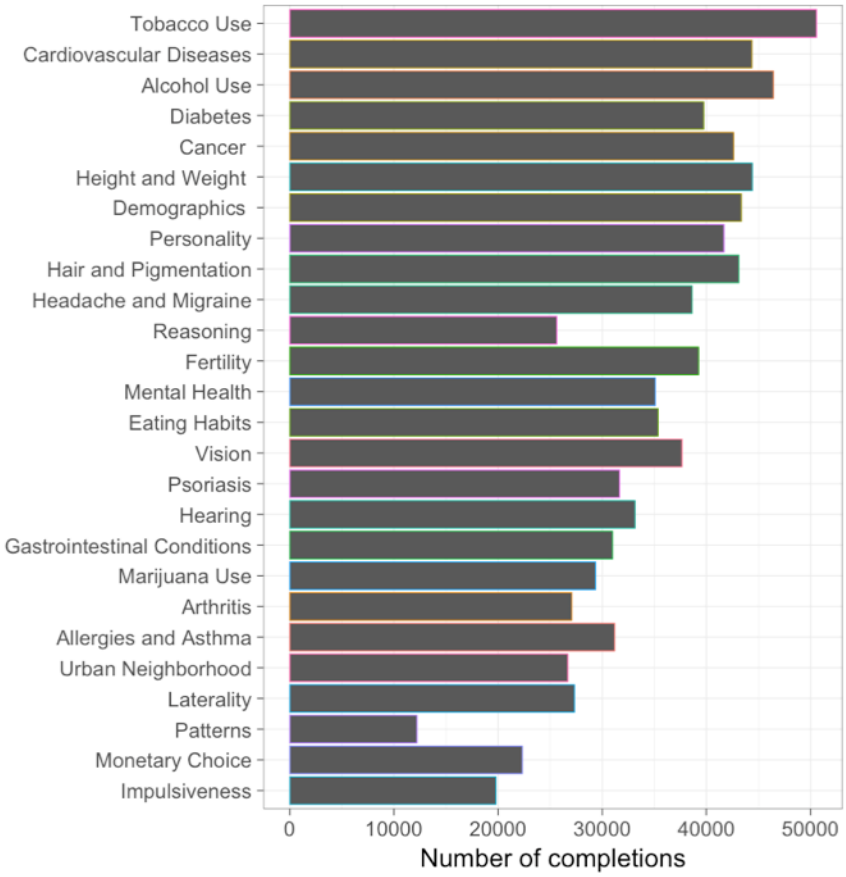
Genes for Good replications of eighteen asthma hits found in Demenais et al. (2018). Adjustments to odds ratios (OR) and sample sizes were made using the approach of Duffy et al. (2004) to correct for response misclassification. Power calculations were made at the 0.05 significance level using the Genes for Good adjusted sample size, disease frequencies and relative risk values from Demenais et al. (2018) control samples, 7.7% population prevalence, and an additive disease model. EA, effect allele; N, number of participants included in analysis. Filtered at AF > 0.005 and AC > 15

Supplementary Figures



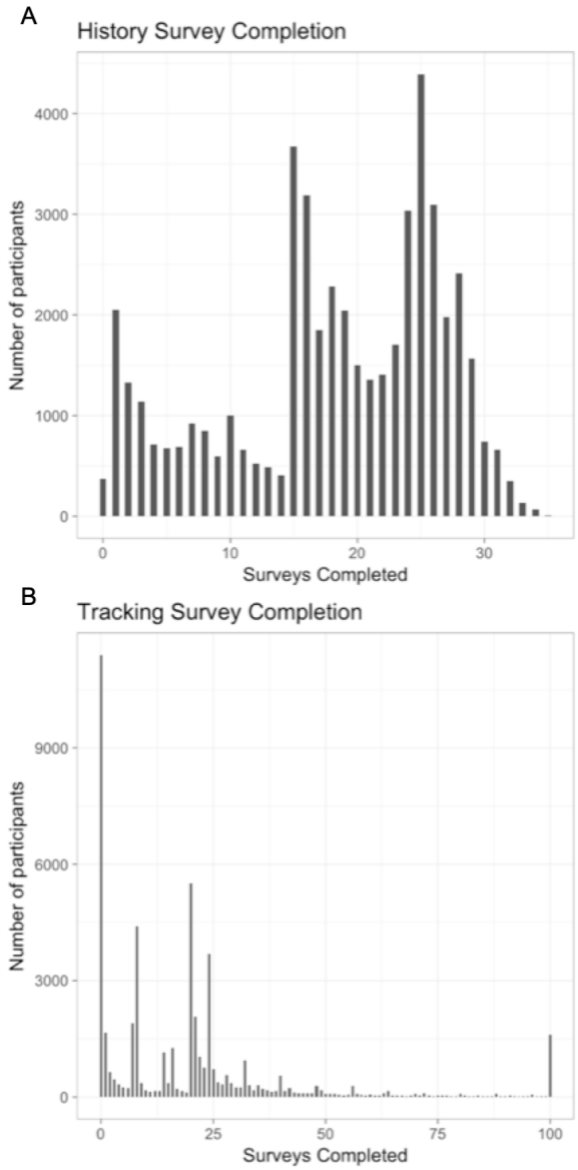
Supplementary Figure 2-1 GWAS panel of common traits in Genes for Good

Manhattan plots for GWAS analysis of various pigmentation and health traits. The x-axis indicates chromosomal location. The y-axis represents $-\log_{10}$(p-value). The red line indicates genome-wide significance ($p = 5\times10^{-8}$). Each genome-wide significant locus is labeled with the gene nearest to it.



Supplementary Figure 2-2 Survey completion count for Health History surveys available in Genes for Good

Survey completion count for Genes for Good surveys. Surveys are ordered by date implemented, with the oldest surveys at the top. The first ten surveys were all available at launch. The Reasoning and Patterns surveys are known to be on the longer side.



Supplementary Figure 2-3 Histogram of Health History and Daily Tracking survey completion

References

Abiad, J. E., Robbins, S., Morris, C., & Sobreira, M. (2018). Survey of Patients with Ollier Disease and Maffucci Syndrome Over Facebook Compared to Review of Clinical Literature (Abstract #9). *Platform talk presented at the 2018 ACMG Annual Clinical Genetics Meeting, April 10-14, 2018*, Charlotte, NC.

Agurs-Collins, T., Ferrer, R., Ottenbacher, A., Waters, E. A., O'Connell, M. E., & Hamilton, J. G. (2015). Public Awareness of Direct-to-Consumer Genetic Tests: Findings from the 2013 U.S. Health Information National Trends Survey. *Journal of cancer education : the official journal of the American Association for Cancer Education, 30*(4), 799-807. doi: 10.1007/s13187-014-0784-x [doi]

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res, 19*(9), 1655-1664. doi: 10.1101/gr.094052.109

Appelboom, G., Camacho, E., Abraham, M. E., Bruce, S. S., Dumont, E. L., Zacharia, B. E., . . . Connolly, E. S., Jr. (2014). Smart wearable body sensors for patient self-assessment and monitoring. *Arch Public Health, 72*(1), 28. doi: 10.1186/2049-3258-72-28

Arcia, A. (2014). Facebook Advertisements for Inexpensive Participant Recruitment Among Women in Early Pregnancy. *Health education & behavior : the official publication of the Society for Public Health Education, 41*(3), 237-241. doi: 10.1177/1090198113504414 [doi]

Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., . . . Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature, 526*(7571), 68-74. doi: 10.1038/nature15393 [doi]

Bays, H. E., Chapman, R. H., Grandy, S., & Group, S. I. (2007). The relationship of body mass index to diabetes mellitus, hypertension and dyslipidaemia: comparison of data from two national surveys. *Int J Clin Pract, 61*(5), 737-747. doi: 10.1111/j.1742-1241.2007.01336.x

Branigan, A. R., McCallum, K. J., & Freese, J. (2013). Variation in the Heritability of Educational Attainment: An International Meta-Analysis. *Social Forces, 92*(1), 109-140. doi: 10.1093/sf/sot076

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., . . . Marchini, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature, 562*(7726), 203-209. doi: 10.1038/s41586-018-0579-z

CDC. (2014). National Diabetes Statistics Report: Estimates of Diabetes and Its Burden in the United States, 2014. *Department of Health and Human Services*(Report), https://www.cdc.gov/diabetes/pubs/statsreport14/national-diabetes-report-web.pdf.

CDC, & NCHS. (2017). National Health and Nutrition Examination Survey Data 2015-2016. *Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease*

Control and Prevention, National Center for Health Statistics, https://wwwn.cdc.gov/Nchs/Nhanes/Search/DataPage.aspx?Component=Examination&CycleBeginYear=2015.

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience, 4*, 7. doi: 10.1186/s13742-015-0047-8

Das, S., Forer, L., Schonherr, S., Sidore, C., Locke, A. E., Kwong, A., . . . Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. *Nature genetics, 48*(10), 1284-1287. doi: 10.1038/ng.3656 [doi]

Demenais, F., Margaritte-Jeannin, P., Barnes, K. C., Cookson, W. O. C., Altmüller, J., Ang, W., . . . collaborators, A. A. G. C. A. (2018). Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. *Nat Genet, 50*(1), 42-53. doi: 10.1038/s41588-017-0014-7

Distribution of Facebook users in the United States as of January 2017, by age group and gender. (2017). *We Are Social*, https://www.statista.com/statistics/187041/us-user-age-distribution-on-facebook/.

Dobkin, B. H., & Dorsch, A. (2011). The promise of mHealth: daily activity monitoring and outcome assessments by wearable sensors. *Neurorehabil Neural Repair, 25*(9), 788-798. doi: 10.1177/1545968311425908

Duffy, S. W., Warwick, J., Williams, A. R. W., Keshavarz, H., Kaffashian, F., Rohan, T. E., . . . Sadeghi-Hassanabadi, A. (2004). A simple model for potential use with a misclassified binary outcome in epidemiology. *Journal of Epidemiology and Community Health, 58*(8), 712-717. doi: 10.1136/jech.2003.010546

El-Amrawy, F., & Nounou, M. I. (2015). Are Currently Available Wearable Devices for Activity Tracking and Heart Rate Monitoring Accurate, Precise, and Medically Beneficial? *Healthc Inform Res, 21*(4), 315-320. doi: 10.4258/hir.2015.21.4.315

eMarketer, & Squarespace. (2017). Number of Facebook users in the United States as of January 2017, by age group (in millions). *Statista*, https://www.statista.com/statistics/398136/us-facebook-user-age-groups/.

Fenner, Y., Garland, S. M., Moore, E. E., Jayasinghe, Y., Fletcher, A., Tabrizi, S. N., . . . Wark, J. D. (2012). Web-based recruiting for health research using a social networking site: an exploratory study. *J Med Internet Res, 14*(1), e20. doi: 10.2196/jmir.1978

Free DNA Test from the University of Michigan. (2017). *Reddit r/freebies*, https://www.reddit.com/r/freebies/comments/67v69c65/free_dna_test_from_the_university_of_michigan/.

Glasgow, M. L., Rudra, C. B., Yoo, E. H., Demirbas, M., Merriman, J., Nayak, P., . . . Mu, L. (2016). Using smartphones to collect time-activity data for long-term personal-level air

pollution exposure assessment. *J Expo Sci Environ Epidemiol, 26*(4), 356-364. doi: 10.1038/jes.2014.78

Hamilton, C. M., Strader, L. C., Pratt, J. G., Maiese, D., Hendershot, T., Kwok, R. K., . . . Haines, J. (2011). The PhenX Toolkit: get the most from your measures. *American Journal of Epidemiology, 174*(3), 253-260. doi: 10.1093/aje/kwr193 [doi]

Harris, P. A., Scott, K. W., Lebo, L., Hassan, N., Lightner, C., & Pulley, J. (2012). ResearchMatch: a national registry to recruit volunteers for clinical research. *Acad Med, 87*(1), 66-73. doi: 10.1097/ACM.0b013e31823ab7d2

Hollands, G. J., French, D. P., Griffin, S. J., Prevost, A. T., Sutton, S., King, S., & Marteau, T. M. (2016). The impact of communicating genetic risks of disease on risk-reducing health behaviour: systematic review with meta-analysis. *BMJ, 352*, i1102. doi: 10.1136/bmj.i1102

Howden, L. M., & Meyer, J. A. (2011). Age and Sex Composition: 2010 *Census Briefs, C2010BR-03* (pp. 1-16 https://www.census.gov/prod/cen2010/briefs/c2010br-2003.pdf). Washington, D.C.: U.S. Census Bureau.

Hughes, V. (2015). A New Facebook App Wants To Test Your DNA. *BuzzFeed News*, https://www.buzzfeed.com/virginiahughes/a-new-facebook-app-wants-to-test-your-dna.

Hysi, P. G., Valdes, A. M., Liu, F., Furlotte, N. A., Evans, D. M., Bataille, V., . . . International Visible Trait Genetics, C. (2018). Genome-wide association meta-analysis of individuals of European ancestry identifies new loci explaining a substantial fraction of hair color variation and heritability. *Nat Genet, 50*(5), 652-656. doi: 10.1038/s41588-018-0100-5

Illumina. (2017). Infinium ® CoreExome- 24 v1.2 BeadChip. https://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet_human_core_exome_beadchip.pdf.

Jiang, Y., Chen, S., McGuire, D., Chen, F., Liu, M., Iacono, W. G., . . . Liu, D. J. (2018). Proper conditional analysis in the presence of missing data: Application to large scale meta-analysis of tobacco use phenotypes. *PLoS Genet, 14*(7), e1007452. doi: 10.1371/journal.pgen.1007452

Kapp, J. M., Peters, C., & Oliver, D. P. (2013). Research recruitment using Facebook advertising: big potential, big challenges. *J Cancer Educ, 28*(1), 134-137. doi: 10.1007/s13187-012-0443-z

Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V., & Stillwell, D. (2015). Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *The American psychologist, 70*(6), 543-556. doi: 10.1037/a0039210

Krokstad, S., Langhammer, A., Hveem, K., Holmen, T. L., Midthjell, K., Stene, T. R., . . . Holmen, J. (2013). Cohort Profile: the HUNT Study, Norway. *International journal of epidemiology, 42*(4), 968-977. doi: 10.1093/ije/dys095

Lee, D., Cornet, R., Lau, F., & de Keizer, N. (2013). A survey of SNOMED CT implementations. *J Biomed Inform, 46*(1), 87-96. doi: 10.1016/j.jbi.2012.09.006

Levy, H. P. (2018). Hypermobile Ehlers-Danlos Syndrome. In M. P. Adam, H. H. Ardinger, R. A. Pagon & S. E. Wallace (Eds.), *GeneReviews®* (pp. https://www.ncbi.nlm.nih.gov/books/NBK1279/). Seattle, WA: University of Washington.

Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., . . . Myers, R. M. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. *Science, 319*(5866), 1100-1104. doi: 10.1126/science.1153717.; ID: 300 10.1126/science.1153717

Liu, F., Chen, Y., Zhu, G., Hysi, P. G., Wu, S., Adhikari, K., . . . Kayser, M. (2018). Meta-analysis of genome-wide association studies identifies 8 novel loci involved in shape variation of human head hair. *Hum Mol Genet, 27*(3), 559-575. doi: 10.1093/hmg/ddx416

Liu, F., van Duijn, K., Vingerling, J. R., Hofman, A., Uitterlinden, A. G., Janssens, A. C., & Kayser, M. (2009). Eye color and the prediction of complex phenotypes from genotypes. *Curr Biol, 19*(5), R192-193. doi: 10.1016/j.cub.2009.01.027

Liu, M., Jiang, Y., Wedow, R., Li, Y., Brazel, D. M., Chen, F., . . . Psychiatry, H. A.-I. (2019). Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nat Genet, 51*(2), 237-244. doi: 10.1038/s41588-018-0307-5

Liu, M., Rea-Sandin, G., Foerster, J., Fritsche, L., Brieger, K., Clark, C., . . . Vrieze, S. (2017). Validating Online Measures of Cognitive Ability in Genes for Good, a Genetic Study of Health and Behavior. *Assessment*, 1073191117744048. doi: 10.1177/1073191117744048

Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., Day, F. R., . . . Speliotes, E. K. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature, 518*(7538), 197-206. doi: 10.1038/nature14177

Lunshof, J. E., Church, G. M., & Prainsack, B. (2014). Information access. Raw personal data: providing access. *Science (New York, N.Y.), 343*(6169), 373.

Mahajan, A., Taliun, D., Thurner, M., Robertson, N. R., Torres, J. M., Rayner, N. W., . . . McCarthy, M. I. (2018). Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat Genet, 50*(11), 1505-1513. doi: 10.1038/s41588-018-0241-6

Maples, B. K., Gravel, S., Kenny, E. E., & Bustamante, C. D. (2013). RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am J Hum Genet, 93*(2), 278-288. doi: 10.1016/j.ajhg.2013.06.020

McMahon, A., Malangone, C., Suveges, D., Sollis, E., Cunningham, F., Riat, H. S., . . . Hall, P. (2018). The NHGRI-EBI GWAS Catalog of published genome-wide association studies,

targeted arrays and summary statistics 2019. *Nucleic Acids Research, 47*(D1), D1005-D1012. doi: 10.1093/nar/gky1120

Montgomery-Downs, H. E., Insana, S. P., & Bond, J. A. (2012). Movement toward a novel activity monitoring device. *Sleep Breath, 16*(3), 913-917. doi: 10.1007/s11325-011-0585-y

Mozaffarian, D., Benjamin, E. J., Go, A. S., Arnett, D. K., Blaha, M. J., Cushman, M., . . . Stroke Statistics, S. (2015). Heart disease and stroke statistics--2015 update: a report from the American Heart Association. *Circulation, 131*(4), e29-322. doi: 10.1161/cir.0000000000000152

Mychasiuk, R., & Benzies, K. (2012). Facebook: an effective tool for participant retention in longitudinal research. *Child Care Health Dev, 38*(5), 753-756. doi: 10.1111/j.1365-2214.2011.01326.x

Nwankwo, T., Yoon, S. S., Burt, V., & Gu, Q. (2013). Hypertension among adults in the United States: National Health and Nutrition Examination Survey, 2011-2012. *NCHS Data Brief*(133), 1-8.

Paschou, P., Lewis, J., Javed, A., & Drineas, P. (2010). Ancestry informative markers for fine-scale individual assignment to worldwide populations. *Journal of medical genetics, 47*(12), 835-847. doi: 10.1136/jmg.2010.078212 [doi]

Pedersen, E. R., & Kurz, J. (2016). Using Facebook for Health-related Research Study Recruitment and Program Delivery. *Current opinion in psychology, 9*, 38-43. doi: 10.1016/j.copsyc.2015.09.011 [doi]

Perrin, A. (2015). Social Networking Usage: 2005-2015. *Pew Research Center*, http://www.pewinternet.org/2015/2010/2008/social-networking-usage-2005-2015.

Royal, C. D., Novembre, J., Fullerton, S. M., Goldstein, D. B., Long, J. C., Bamshad, M. J., & Clark, A. G. (2010). Inferring genetic ancestry: opportunities, challenges, and implications. *American Journal of Human Genetics, 86*(5), 661-673. doi: 10.1016/j.ajhg.2010.03.011 [doi]

Sanchez-Roige, S., Fontanillas, P., Elson, S. L., the 23andMe Research, T., Pandit, A., Schmidt, E. M., . . . Palmer, A. A. (2017). Genome-wide association study of delay discounting in 23,217 adult research participants of European ancestry. *Nature Neuroscience*.

Sankararaman, S., Mallick, S., Dannemann, M., Prufer, K., Kelso, J., Paabo, S., . . . Reich, D. (2014). The genomic landscape of Neanderthal ancestry in present-day humans. *Nature, 507*(7492), 354-357. doi: 10.1038/nature12961 [doi]

Semega, J. L., Fontenot, K. R., & Kollar, M. A. (2017). Households by Total Money Income, Race, and Hispanic Origin of Householder: 1967 to 2016 *U.S. Census Bureau, Current Population Reports, P60-259, Income and Poverty in the United States: 2016* (pp. 23-29). Washington, DC: U.S. Government Printing Office.

Steinhubl, S. R., Muse, E. D., & Topol, E. J. (2015). The emerging field of mobile health. *Science translational medicine, 7*(283), 283rv283. doi: 10.1126/scitranslmed.aaa3487 [doi]

Stoekle, H. C., Mamzer-Bruneel, M. F., Vogt, G., & Herve, C. (2016). 23andMe: a new two-sided data-banking market model. *BMC medical ethics, 17*, 9. doi: 10.1186/s12910-016-0101-9 [doi]

Sulem, P., Gudbjartsson, D. F., Stacey, S. N., Helgason, A., Rafnar, T., Jakobsdottir, M., . . . Stefansson, K. (2008). Two newly identified genetic determinants of pigmentation in Europeans. *Nat Genet, 40*(7), 835-837. doi: 10.1038/ng.160

Tsoi, L. C., Stuart, P. E., Tian, C., Gudjonsson, J. E., Das, S., Zawistowski, M., . . . Elder, J. T. (2017). Large scale meta-analysis characterizes genetic architecture for common psoriasis associated variants. *Nat Commun, 8*, 15382. doi: 10.1038/ncomms15382

Tung, J. Y., Eriksson, N., Kiefer, A. K., Macpherson, J. M., Naughton, B. T., Chowdry, A. B., . . . Mountain, J. L. (2011). Characteristics of an Online Consumer Genetic Research Cohort (Abstract #914T). *Poster presented at the 61st Annual Meeting of The American Society of Human Genetics, October 11-15, 2011*, Montreal, Canada.

Visconti, A., Duffy, D. L., Liu, F., Zhu, G., Wu, W., Chen, Y., . . . Falchi, M. (2018). Genome-wide association study in 176,678 Europeans reveals genetic loci for tanning response to sun exposure. *Nat Commun, 9*(1), 1684. doi: 10.1038/s41467-018-04086-y

Wang, C., Zhan, X., Liang, L., Abecasis, G. R., & Lin, X. (2015). Improved ancestry estimation for both genotyping and sequencing data using projection procrustes analysis and genotype imputation. *Am J Hum Genet, 96*(6), 926-937. doi: 10.1016/j.ajhg.2015.04.018

Ward, B. W., Clarke, T. C., Nugent, C. N., & Schiller, J. S. (2016). Early Release of Selected Estimates Based on Data From the 2015 National Health Interview Survey. *National Center for Health Statistics, May 2016*, https://www.cdc.gov/nchs/data/nhis/earlyrelease/earlyrelease201605.pdf.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., . . . Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res, 42*(Database issue), D1001-1006. doi: 10.1093/nar/gkt1229

Wood, A. R., Esko, T., Yang, J., Vedantam, S., Pers, T. H., Gustafsson, S., . . . Frayling, T. M. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet, 46*(11), 1173-1186. doi: 10.1038/ng.3097

Zhan, X., Chen, S., Jiang, Y., Liu, M., Iacono, W. G., Hewitt, J. K., . . . Liu, D. J. (2017). Association Analysis and Meta-Analysis of Multi-allelic Variants for Large Scale Sequence Data. *bioRxiv*, 197913.

Zhou, W., Nielsen, J. B., Fritsche, L. G., Dey, R., Gabrielsen, M. E., Wolford, B. N., . . . Lee, S. (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-

scale genetic association studies. *Nat Genet, 50*(9), 1335-1341. doi: 10.1038/s41588-018-0184-y

Chapter 3 Estimation of DNA Contamination and Its Sources in Genotyped Samples[2]

Introduction

Array genotyping is the standard method to genotype large numbers of individuals for genome-wide association studies (GWAS), consumer genomics, evaluation of copy number in clinical settings, and sample quality control prior to sequencing (Diskin et al., 2008). Consortium efforts now include millions of directly genotyped samples, and array genotyping has successfully been applied to traits as diverse as height (Marouli et al., 2017), body mass index (Locke et al., 2015), blood pressure (Hoffmann et al., 2017), type 2 diabetes (Mahajan et al., 2014), schizophrenia (Goes et al., 2015), and inflammatory bowel disease (Liu et al., 2015), among many others. When coupled with imputation, genotyping arrays can achieve a similar coverage of the genome to sequencing for a fraction of the cost (Y. Li, Willer, Ding, Scheet, & Abecasis, 2010).

Typically, genotyping arrays use fluorescent-tagged nucleotides or oligonucleotides that are specific to each allele of a genetic polymorphism. Measurements of allele-specific intensities are collected in parallel at 100,000s of loci, post-processed and clustered to distinguish genotypes at different bi-allelic markers (G. Li, 2016). These steps are sensitive to DNA sample contamination and mixing so that contaminated samples will have a higher probability of missing

[2] This work appeared in Genetic Epidemiology as "Estimation of DNA contamination and its sources in genotyped samples." *43*(8), 980-995. I was first author.

or erroneous calls that can result in a loss of power (Flickinger, Jun, Abecasis, Boehnke, & Kang, 2015) or in erroneous downstream inferences.

This DNA sample contamination is a common problem in large-scale studies. For example, the 1000 Genomes project reported that 3% of the sequenced samples were excluded due to high contamination (Flickinger et al., 2015). To address this problem, there are now several methods for detecting DNA contamination in both genotyping and sequencing data. Early methods flagged contaminated samples, but did not estimate the proportion of contamination (Homer et al., 2008). Newer methods like VerifyIDintensity and BAFRegress estimate contamination proportions by examining sample-specific shifts in allele intensity clusters for each genotype (Jun et al., 2012). Similar methods exist to examine the proportion of reads in sequencing data that are from contaminating DNA, for example ContEst and VerifyBAMID (Cibulskis et al., 2011; Jun et al., 2012). Contamination estimation has even been applied to array methylation data (Heiss & Just, 2018). Although our focus here is on within-species contamination, methods also exist for estimating cross-species contamination in sequencing data (Schmieder & Edwards, 2011). However, none of these methods can simultaneously estimate both contamination and its sources in genotyping array samples.

Here we present a new method, VICES (Verification of Intensity Contamination from Estimated Sources) that estimates contamination proportions and identifies contaminating samples in genotyping array data. VICES initially uses sample allele frequencies to estimate contamination and then revises this estimate by iteratively searching for sources of contamination among other genotyped samples. When the contaminating sample can be identified, our method provides improved estimates of contamination proportions compared to existing methods VerifyIDintensity and BAFRegress. Identifying contaminating samples also helps revise laboratory protocols to

prevent future contamination. Finally, by examining data from ongoing studies, we show that VICES can help flag problematic sample processing steps where contamination occurred.

Methods

Our method has three steps: 1. identifying contaminated samples, 2. identifying likely contaminating samples for each contaminated sample, and 3. producing a final estimate of contamination, quantifying contributions from each contaminating sample (Figure 3-1).

We will first introduce some notation. We consider a set of individuals, each genotyped using an array. For each marker j, we assume two alleles, arbitrarily labelled A and B. We denote the frequency of B at this marker as AF_j. We let G_{ij} denote the estimated genotype for individual i at marker j, encoded as 0 (homozygous for A), 1 (homozygous for B), or ½ (heterozygous). Following convention, we let I_{ij} denote the relative intensity of the B-allele probe, measured on a 0 to 1 scale by interpolating allele intensity values with respect to the centers of the three genotype clusters and truncating any values that fall outside the 0 to 1 range (Illumina, 2010). Although other definitions of I_{ij} are possible, we choose this one because estimates are readily available from Illumina genotyping software.

The following model relates I_{ij} of the sample being tested to its estimated genotype and to the genotypes of each potential contaminating sample. Let α_i be the total proportion of contaminating DNA in sample i and α_{ik} the proportion of DNA mixture from sample k.

$$E(I_{ij}) = (1 - \alpha_i)G_{ij} + \sum_k \alpha_{ik} G_{kj} \qquad \text{(Equation 1)}$$

Directly fitting this model performs poorly because even in the absence of contamination, average intensity $I_{ij} \leq 1$ when $G_{ij} = 1$ and average intensity $I_{ij} \geq 0$ when $G_{ij} = 0$. Instead, we fit three genotype specific background intensity values γ_0, $\gamma_{½}$, and γ_1 which model the expected intensity for each genotype class. This results in the model:

$$E(I_{ij}) = (1 - \alpha_i)\gamma_{[G_{ij}]} + \sum_k \alpha_{ik}\gamma_{[G_{kj}]} \quad \text{(Equation 2)}$$

which requires numerical optimization to estimate the total contamination proportion, α_i, and the contamination proportions α_{ik} from each contaminating sample. Fitting the following linear regression model

$$E(I_{ij}) = \gamma_{[G_{ij}]} + \sum_k \alpha_{ik}G_{kj} \quad \text{(Equation 3)}$$

gave estimates within 0.1% of Equation 2 for the contamination proportion from each contaminating sample, α_{ik}, while using only a fraction of the computational time. The $\gamma_{[G_{ij}]}$ intercept terms allow for a different mean I_{ij} for each cluster of sample genotypes, with each α_{ik} coefficient having the convenient interpretation as the contamination proportion from sample k. Identification of the contaminated and contaminating samples in a genotyping cohort, and estimation of the contamination proportion from each contaminating sample α_{ik} proceeds as follows:

Identification of contaminated samples

We substitute the contaminating sample genotypes in Equation 3 with the allele frequencies AF_j to obtain initial estimates of the contamination proportion α_i for each sample being considered.

This enables us to exclude uncontaminated samples from the computationally intensive search for samples that contributed contaminating DNA.

We fit the following model to obtain $\hat{\alpha}_{iAF}$, an initial estimate of the contamination proportion α_i:

$$E(I_{ij}) = \gamma_{[G_{ij}]} + \alpha_{iAF} AF_j \qquad \text{(Equation 4)}$$

When fitting this model, we recommend excluding any sites with minor allele frequency less than 0.1 to reduce the influence of monomorphic and rare variants on the parameter estimation.

If this first estimate of the contamination proportion based on allele frequencies, $\hat{\alpha}_{iAF}$, is below a user-specified threshold T (we recommend T no less than 0.005), then we assume the sample is uncontaminated and estimation stops here. If it is above that threshold, then our method attempts to identify the contaminating samples among the other genotyped samples.

Find the samples that contributed contaminating DNA

After identifying the contaminated samples using allele frequencies, the next step is to estimate a set of likely samples that contributed DNA to them. To do this, we fit the following linear regression model where we regress allelic intensity on the contaminated sample genotypes, allele frequency, and the genotypes of each candidate contaminating sample in turn:

$$E(I_{ij}) = \gamma_{[G_{ij}]} + \alpha_{iAF} AF_j + \alpha_{ik} G_{kj} \qquad \text{(Equation 5)}$$

This step identifies a series of candidate contaminating samples for each contaminated sample. We specifically focus on pairings of contaminated and contaminating samples where the estimate of

$\hat{\alpha}_{ik}$ is greater than our contamination threshold T. For these potential combinations of contaminated and contaminating samples, we proceed to the final step to calculate an improved contamination estimate.

Fit the final model with all contaminating samples to produce a final estimate

After identifying likely contaminating samples, this final step fits the following regression:

$$E(I_{ij}) = \gamma_{[G_{ij}]} + \alpha_{iAF} AF_j + \sum_k \alpha_{ik} G_{kj} \qquad \text{(Equation 6)}$$

with the intensities I_{ij} and estimated genotypes G_{ij} of the contaminated sample, the allele frequencies AF_j, and the genotypes G_{kj} of all the samples whose estimated contribution $\hat{\alpha}_{ik}$ to the contamination proportion was greater than the contamination threshold T.

Since contamination only affects I_{ij} at sites where $G_{ij} \neq G_{kj}$, such sites tend to be highly polymorphic. As a result, any individual k', even if it did not contribute DNA to sample i, is likely to have many $G_{ij} \neq G_{k'j}$ at those sites with large $I_{ij} - G_{ij}$, and can appear to explain some of the contamination. Therefore, the set of potential contaminating samples identified in Step 2 may include false positives. When the contributions of these "false positive" contaminating samples are estimated jointly with those of the true contaminating samples, we expect their $\hat{\alpha}_{ik}$ coefficients to drop near zero. Therefore, we expect the best estimates of contamination proportions will be obtained after estimation in step 3 (using Equation 6). If at this point, there are any $\hat{\alpha}_{ik} < T$, we exclude the sample with the smallest $\hat{\alpha}_{ik}$ and refit the regression, repeating this step until we have excluded all candidate contaminating samples whose contributions $\hat{\alpha}_{ik}$ are below T.

After inclusion of all contaminating samples, the background contamination estimate should also drop to near or below 0. We define background contamination as α_{iAF} in equation 6.

To be consistent with this interpretation, once all samples with contamination contribution $\hat{\alpha}_{ik}$ less than T are removed, this background contamination term α_{iAF} is also dropped if it is estimated $\leq$ 0 since the proportion of contaminating DNA from any source cannot be negative.

The final model and resulting estimate of contamination can be one of the following three possibilities:

1. The estimated contamination contribution from allele frequencies, $\hat{\alpha}_{iAF}$, drops to or below 0 and the model is refit with the estimated contaminating samples only. The estimate of the total contamination proportion is then the sum of the contamination contribution from each estimated source, as in Equation 3:

$$E(I_{ij}) = \gamma_{[G_{ij}]} + \sum_k \alpha_{ik} G_{kj}. \quad \text{(Equation 3)}$$

2. No contaminating samples remain in the model, leaving only the contamination contribution from allele frequencies. This results in the model in Equation 4 and the same contamination proportion estimated in Step 1:

$$E(I_{ij}) = \gamma_{[G_{ij}]} + \alpha_{iAF} AF_j. \quad \text{(Equation 4)}$$

3. Both estimated contaminating samples and allele frequencies remain in the model. Then the $\hat{\alpha}_{iAF}$ coefficient can be interpreted as the proportion of contamination that came either from outside the genotyping cohort or from contaminating samples in the cohort but at proportions that were too small to be estimated reliably. The estimate of the total contamination proportion is then

the sum of the contamination contribution from the estimated sources and the contamination contribution from allele frequencies. In this scenario, the final model is as in Equation 6:

$$E(I_{ij}) = \gamma_{[G_{ij}]} + \alpha_{iAF} AF_j + \sum_k \alpha_{ik} G_{kj}. \quad \text{(Equation 6)}$$

Implementation

We have implemented VICES in a free software package written in C++ and available for download at http://genome.sph.umich.edu/wiki/VICES.

Experimental Data

We analyzed contamination in two sets of genotyping data. These different data sets allowed us to quantify the effect of contamination in the context of different arrays and experiments. It also allowed us to compare the performance of VICES with previous contamination methods VerifyIDintensity and BAFRegress under different scenarios (Jun et al., 2012).

Intentionally contaminated HapMap samples

To evaluate the effect of contamination on genotype calling and the performance of our method, we used intensity data and genotype calls generated by Jun et al. (2012) from 34 samples that were intentional mixtures of DNA from 4 HapMap cell lines (International HapMap et al., 2010). The samples were 100:0, 0.5:99.5, 1:99, 2:98, 3:97, 5:95, and 10:90 mixtures of mixed European ancestry (CEU) samples NA07055 and NA06990, and 0:100, 0.5:99.5, 1:99, 2:98, 5:95, and 10:90 mixtures of Yoruban (YRI) samples NA19200 and NA18504 (Table 3-1) and genotyped on the Illumina MetaboChip (Voight et al., 2012) at 196,725 markers. We obtained contaminating

sample genotypes and allele frequency estimates for contamination estimation from the 1000 Genomes Phase 3 version 5 at sites that overlapped with the MetaboChip (Genomes Project et al., 2015). We estimated contamination in these 34 samples using: (1) VICES with contaminating sample genotypes (VICES-Geno), (2) VICES with allele frequencies (VICES-AF), (3) VerifyIDintensity (VID) and (4) BAFRegress (BAFR). Specifically, we compared root-mean-squared-error (RMSE), bias, and trend in absolute error as contamination increased for the four sets of contamination estimates.

For the estimates calculated using VICES-Geno, the contaminating sample was already known in each case, so we estimated the contamination proportion by fitting the model in Equation 3. For all mixtures of HapMap YRI cell lines, we used the 1000 Genomes genotypes from sample NA19200 to estimate contamination. For the uncontaminated CEU samples from NA07055, we randomly chose an unrelated CEU sample from 1000 Genomes, NA12776, to provide the contaminating sample genotypes to fit in the model. For the CEU mixture samples, we used the metabochip genotypes of NA07055 as the contaminating sample. We only used NA19200 genotypes at sites with minor allele frequency above 10% in in 661 African ancestry samples of the 1000 Genomes Project (AFR). Similarly, we only used NA12776 or NA07055 genotypes at sites with minor allele frequency above 10% in 503 European ancestry samples of the 1000 Genomes Project (EUR).

For the estimates calculated using VICES-AF, we regressed the Metabochip intensities on their respective genotypes and allele frequencies as in Equation 4. We used 1000 Genomes EUR allele frequencies to estimate contamination in the CEU samples and 1000 Genomes AFR allele frequencies to estimate contamination in the YRI samples. As in the previous, we only used allele frequencies with MAF above 10%. We used the same sets of allele frequencies to estimate

contamination with BAFRegress and VerifyIDintensity. We ran BAFRegress with default settings and VerifyIDintensity using the per-marker analysis option recommended by the authors of the software (Jun et al., 2012).

We also used the intentionally mixed HapMap samples to illustrate the effect of using allele frequencies from a mis-specified population on contamination estimation with VICES-AF, BAFRegress, and VerifyIDintensity. For this analysis, we used the 1000 Genomes EUR allele frequencies to estimate contamination in the YRI samples, and the 1000 Genomes AFR allele frequencies to estimate contamination in the CEU samples. Again, we only used allele frequencies with MAF above 10% and the per-marker analysis option for VerifyIDintensity.

Michigan Genomics Initiative

Next, we compared estimates from VICES with VerifyIDintensity and BAFRegress, in a large genotyping study where contamination may have occurred unintentionally. For this, we used data from the Michigan Genomics Initiative (Fritsche et al., 2018), an ongoing study of genetic data and health records from patient volunteers at the University of Michigan Hospital. We used 22,366 samples genotyped at 603,583 markers on a customized Illumina Infinium HumanCoreExome-24 v1.0 array (Illumina, 2017). DNAs, extracted from blood, were assayed in batches of 288 to 576 samples (3-6 plates of 96 samples each) per run according to the Illumina Infinium HTS Assay Protocol Guide (Illumina, 2013). The smallest assay runs with 288 samples were combined with larger batches for genotype calling in GenomeStudio (Illumina, 2016), so sets of genotype calls ranged in size from 384 to 864 samples. We considered contamination between samples from different set of genotype calls to be unlikely, so we ran our method on each set of genotype calls separately using VICES with the default settings. We also ran VerifyIDintensity on each set of genotype calls separately and with the per-marker analysis option. BAFRegress was

run under default settings. For both VerifyIDintensity and BAFRegress, we used variants that overlapped with the HumanCoreExome array and whose 1000 Genomes EUR MAF was above 10% at overlapping sites. VICES calculates allele frequencies for initial estimation so no external allele frequencies were used. The true contamination proportions were not known in MGI, but we were able to compare the concordance of the three methods' contamination estimates, the proportion of samples with estimated contamination greater than 0.5%, and how strongly contamination estimates were correlated with the number of missing and excess heterozygous genotype calls as calculated by Plink 1.9 (Chang et al., 2015).

Results

HapMap

Shift in probe intensities - HapMap

We examined how contamination changed overall intensity for homozygous A/A, heterozygous, and homozygous B/B genotypes. We saw that, in each case, intensity clusters were shifted towards the contaminant genotype. This result supports the validity of the assumption in Equation 2 that the intensities shift in proportion to the contamination and the genotypes of the contaminant sample. The kernel density plots in Figure 3-2 show the distributions of the intensities for an uncontaminated sample and for a sample contaminated at the 10% level, as a function of genotypes for the contaminating sample. The distribution of the intensities in the contaminated sample is shifted towards the genotypes of the contaminating sample (for example, when the contaminating sample has genotype B/B, all intensities are shifted towards the B allele). As expected, the distribution of intensities for the uncontaminated sample is independent of the genotypes of the potential contaminating sample.

Estimation - HapMap

We next examined whether we could accurately estimate contamination in the intentionally mixed HapMap samples. These samples were prepared by Jun et al. (2012) to assess the performance of their own methods to estimate contamination. A total of 179,935 markers overlapped between the Metabochip and 1000 Genomes. Of these, we used AFR allele frequencies of 90,401 markers with MAF above 10% and EUR allele frequencies of 88,747 markers with MAF in EUR above 10%. Compared to the intended contamination, VICES-Geno had a root-mean-squared-error (RMSE) of 0.0057 and bias of -0.0035 across the 34 samples (Table 3-2 and Figure 3-3). As contamination increased, the absolute error of VICES-Geno estimates increased on average by 0.0012 for each percentage increase in contamination. VICES-Geno performed better than VICES-AF, which had RMSE of 0.0068, bias of -0.0041, and an increase in absolute error of 0.0015 for each percentage increase in contamination. This shows an additional benefit in estimating contamination by using the genotypes of the contaminating sample as opposed to sample or population allele frequencies.

VICES-Geno's performance was within 0.001 of existing method BAFRegress on the three criteria and outperformed VerifyIDintensity by a much wider margin. BAFRegress had a RMSE of 0.0054, bias of -0.0024, and absolute error increased by 0.0011 for each percentage increase in contamination, while VerifyIDintensity had RMSE of 0.0310, bias of -0.0085, and absolute error increased by 0.0056 for each percentage increase in contamination (Figure 3-3). The results of this comparison are also summarized in Table 3-2.

Estimation with Misspecified Allele Frequencies - HapMap

We next evaluated the impact of ancestral population for reference allele frequencies on estimates of contamination. We expected this choice would have only a very limited impact for

VICES-Geno as long as contaminating sample genotypes were available. However, the impact would be potentially larger for BAFRegress and VerifyIDintensity since they rely on estimated allele frequencies to estimate contamination.

We used 1000 Genomes allele frequencies calculated in EUR with MAF > 10% at 88,747 markers that overlapped with the Metabochip to estimate contamination in the intentionally mixed HapMap YRI samples. Similarly, we used 1000 Genomes allele frequencies calculated in AFR with MAF > 10% at 90,401 markers that overlapped with the Metabochip to estimate contamination in the CEU samples. Compared to the intended contamination, VICES-AF using mis-specified allele frequencies had RMSE of 0.0231, bias of -0.0140, and absolute error increased by 0.0057 for each percentage increase in contamination across the 34 samples. When the correct allele frequencies were used, VICES-AF had RMSE of 0.0068, bias of -0.0041, and a 0.0015 increase in absolute error for each percentage increase in contamination.

The other two methods also showed a similar drop in performance when using the misspecified allele frequencies. BAFRegress had a RMSE of 0.0261, bias of -0.0150, and the absolute error increased by 0.0065 for each percentage increase in contamination, while VerifyIDintensity had RMSE of 0.0312, bias of -0.0086, and the absolute error increased by 0.0056 for each percentage increase in contamination. The results of this comparison between our method, BAFRegress, and VerifyIDintensity with misspecified allele frequencies are also summarized in Table 3-3.

All three methods performed worse when the population for the allele frequencies was misspecified than when they were correctly specified, as shown in Table 3-2. This result implies that when using BAFRegress or VerifyIDintensity, prior knowledge of the ancestry of contaminating DNA is necessary to find contaminated samples and exclude their genotype calls

from downstream analyses, an impractical step in a large GWAS cohort of diverse ancestry. This result highlights the benefit to estimating samples that contributed contaminating DNA so that estimation is not as sensitive to the choice of population for allele frequencies.

Shift in allele frequencies with misspecified allele frequencies - HapMap

We further explored the previous point about how using misspecified allele frequencies can lead to an underestimation of contamination levels. Figure 3-4 shows the distribution of intensities for each genotype for a contaminated and an uncontaminated sample in different 1000 Genomes EUR minor allele frequency bins instead of contaminating sample genotypes as in Figure 3-2. As expected, contamination results in a greater shift in the intensity distribution at markers with higher allele frequencies. Figure 3-5 recapitulates Figure 3-4 but uses minor allele frequencies calculated from 1000 Genomes AFR individuals. As shown, the distribution of probe intensities is similar in the uncontaminated sample regardless of MAF of the population in which the MAFs were calculated. However, the shift in the intensity distribution at higher allele frequencies is less pronounced when using 1000 Genomes AFR MAFs compared to using 1000 Genomes EUR MAFs. This result highlights the benefit of using estimated contaminating sample genotypes for improving contamination estimation in genotyping samples.

Michigan Genomics Initiative (MGI)

Estimation - MGI

Our next aim was to investigate whether our method could accurately estimate contamination in a large-scale genotyping experiment. A test of the three methods in the 22,366 MGI samples suggests that VICES strikes a balance between the low estimates provided by BAFRegress and the higher estimates provided by VerifyIDintensity, consistent with our analysis

of intentionally contaminated HapMap samples (see Figure 3-3, Table 3-2). Among the 22,366 samples, VICES found 354 with contamination greater than 0.5%, BAFRegress found 291 samples, while VerifyIDintensity found 4,498, or 20% of the samples tested.

This last result raised the question of why VerifyIDintensity estimated contamination greater than 0.5% for 4,188 samples for which both BAFRegress and VICES estimated contamination less than 0.5%. Upon investigation, it turned out that in samples where VICES estimated contamination less than 0.5%, the VerifyIDintensity estimates tended to be higher when there was a greater mean squared difference between the probe intensity and called genotype centroid (Figure 3-7). The same relationship was not seen in the BAFRegress or VICES estimates in the same set of samples. This result shows that VerifyIDintensity is prone to overestimating contamination in samples with greater variability in their probe intensities.

The true contamination proportions were not known in MGI, but we compared the estimates from the three methods to one another to determine which represented the best consensus. We found that the samples which VICES estimated as contaminated greater than 0.5% were validated more often by the other methods than the samples estimated as contaminated greater than 0.5% by BAFRegress or VerifyIDintensity. The bar plot in Figure 3-8 shows the counts for the number of samples with estimated contamination greater than 0.5% by at least two of the three methods, which also shows that VICES had the highest number of samples (316) with estimated contamination greater than 0.5% verified by at least one other method. VICES also had lower root-mean-squared-difference with estimates from BAFRegress (0.0075) and VerifyIDintensity (0.0062) than they did with each other (0.0089).

Comparing the contamination estimates to call rate and excess heterozygosity of the MGI samples provided an independent metric which further supports the accuracy of VICES. Figure

3-9 and Figure 3-10 show that all three methods exhibited the same relationship that, as estimated contamination increased, genotype call rates decreased and excess heterozygosity increased. However, the underestimation of BAFRegress was more pronounced in samples with a high level of contamination. BAFRegress did not estimate contamination greater than 13% for any sample, even for 11 samples that VICES and VerifyIDintensity both estimated as having contamination proportions greater than 20%. For this reason, the trend between estimated contamination and excess heterozygosity, and estimated contamination and call rate was weaker with the BAFRegress estimates (R^2 0.03 for both call rate and excess heterozygosity) than VICES (R^2 0.18 for call rate, R^2 0.19 for excess heterozygosity) or VerifyIDintensity (R^2 0.11 for call rate, R^2 0.12 for excess heterozygosity).

Since the plot of sample call rate against VICES estimated contamination in Figure 3-9 appeared to show two trend lines, we sought an explanation. Specifically, we observed that many contaminated samples had a lower call rate than would be predicted by their contamination as estimated by VICES (Figure 3-9, left panel). We found that $\log_2$ R ratio, a measure of the average genotyping array probe intensity for a sample (Peiffer et al., 2006), was a strong predictor of call rate (R^2 0.48, Figure 3-11). In Figure 3-12 we removed all 165 samples with $\log_2$ R ratio 2 standard deviations below the mean before plotting sample call rate against estimated contamination. In this plot, compared to Figure 3-9, the relationship between contamination and call rate was stronger and more distinct (R^2 0.71, 0.13, and 0.55 respectively for VICES, BAFRegress, and VerifyIDintensity). This result shows that this second trend line in Figure 3-9 was not due to underestimation by our method, but by heterogeneity in the array probe intensity among the samples.

Contaminating sample search - MGI

We sought to evaluate how often our method could find contaminating samples and whether the estimates implicated a clear mechanism for contamination. We used the VICES results from the 22,366 samples genotyped in the Michigan Genomics Initiative (MGI) and found that our method found contaminating samples from the same set of genotype calls for 301 or 85% of the 354 samples with estimated contamination above 0.5%. A total of 365 contaminating samples were estimated. Of these, 342 or 94% were on the same sample processing plate of 96 samples as the contaminated sample, and 328 or 90% were on the same genotyping array of 24 samples, showing that VICES estimates of contaminating samples are not random, but in fact consistently implicate a step in the sample preparation and genotyping process where contamination often occurred.

The number of contaminating samples offers further support for the accuracy of the VICES estimates relative to the other methods. Figure 3-6 shows that BAFRegress failed to detect contamination greater than 0.5% in 38 samples where VICES estimated such a level of contamination and found a contaminating sample, and VerifyIDintensity failed to detect contamination in 31 such samples. There were 26 such samples where neither BAFRegress nor VerifyIDintensity estimated contamination greater than 0.5%. These results suggest that BAFRegress and VerifyIDintensity may be prone to false negatives in contamination estimation, allowing contaminated samples through QC filters.

Based on the data in Figure 3-6, we wondered if any of the samples estimated as contaminated by VICES but not all three methods were false positives. One reason is that VICES found contaminating samples for a higher proportion (92%) of the 279 samples with estimated contamination greater than 0.5% by all three methods than in the 75 samples estimated as contaminated greater than 0.5% by VICES alone or by VICES and only one other method (57%).

One explanation is that VICES estimated much lower contamination for the samples that were estimated to be uncontaminated by either BAFRegress or VerifyIDintensity. VICES estimated 48 (64%) of the 75 samples (estimated as contaminated by VICES but not all three methods) to be contaminated below 1%, compared to 28 (10%) of the 279 samples estimated as contaminated by all three methods. Small discrepancies in the estimates between the three methods may have pushed the estimates for some samples either just above or just below the contamination threshold T for a subset of the methods. For this reason, we expect that VICES will have more difficulty estimating sources of contamination for samples with borderline detectable contamination than for samples with high contamination.

In addition to improving estimation, finding the contaminating samples enables understanding and troubleshooting the cause of contamination. In the MGI samples, Figure 3-13 shows that the contaminated samples as estimated by VICES appear adjacent to one another on both the sample processing plate and the genotyping array. Running the contaminating sample search algorithm reveals that the estimated contaminating samples for each contaminated sample were adjacent to it on the array but not the processing plate. Since it would be more difficult to explain the pattern between contaminating and contaminated samples on the processing plate, this constitutes strong evidence for contamination occurring on the genotyping array between adjacent inlet ports during sample loading or array sections during hybridization due to leaky seals.

Discussion

Contamination, or the mixture of DNA from multiple individuals prior to genotyping, decreases the quality of genotypes. Since genotyping arrays remain the predominant tool in genetic association studies, the ability to accurately diagnose contaminating DNA and its sources has the potential to improve data quality checks and data production for many genetic studies. Our results

show that our method outperforms previous methods and can reliably find the contaminating samples, even at small contamination proportions. It can also perform contamination estimation in genotyping cohorts of mixed ancestry without relying on external allele frequency information or knowledge of the population origin of the contaminating samples. This feature makes the software appropriate for a wide range of genetic association studies. We also illustrate how one can conclude that contamination occurred on a genotyping array as opposed to during other steps in sample preparation, which may lead to improved genotyping protocols.

One of our central findings is that, compared to estimating contamination and its sources separately, doing so jointly, as described here, improves both and gives users of VICES a more useful combination of results. After contamination has been detected, researchers may be faced with several follow-up questions. For example, should a contaminated sample be excluded from downstream analyses? Can a sample be re-genotyped and yield uncontaminated genotype calls? Or is a sample's DNA fit for whole-genome or whole-exome sequencing? VICES gives researchers accurate information to answer to these questions.

The above analysis illuminated several ways in which contamination and contaminating sample identification can be further improved. One remaining issue is that the deviations in array probe intensities caused by contamination can appear to be correlated to the genotypes of any individual, and not only the contaminating sample. We observed a similar effect at the population level, with the shift in allele frequencies showing the strongest correlation with frequencies in the contaminating sample population but weaker correlation when the contaminating population allele frequencies were misspecified.

This correlation between probe intensities and the genotypes of a sample that did not contribute DNA can be partially mitigated by including the sample allele frequencies in the

regression as in Equations 5 and 6. However, at particularly high levels of contamination (greater than 25%) many false positive contaminating samples may still be identified. This problem can be improved by increasing the contaminating sample threshold for highly contaminated samples instead of the default threshold of 0.5%. There are alternatives to threshold based selection of contaminating samples that may be worthy of future exploration. For example, instead of including samples in the final model based on a point estimate for contamination contribution, inclusion could have been decided by p-value or false discovery rate-adjusted q-value, or estimating inflation in contamination contribution estimates.

An alternative strategy to make the contamination estimates more robust to the genetic ancestry of the contaminating DNA could be to iteratively estimate the ancestry of the contaminating allele frequencies instead of using the fixed allele frequencies of the sample or population. Such an approach could result in more accurate contamination estimates when no contaminating sample is found or could be used to narrow the search by the ancestry of the contaminating sample, resulting in greater computational efficiency. However, we have found that using contaminating sample genotypes improves contamination estimates compared to using population allele frequencies, even when the contaminating samples' population is correctly specified (Table 3-2). Furthermore, using population allele frequencies, the user would not gain any insight as to how contamination occurred in their study.

In addition, several potential extensions or adaptations of this method exist. For example, a cross-array contamination check might be useful in studies where multiple arrays are used. In addition, the method could be adapted to impute missing and incorrect calls to salvage contaminated samples, as the CleanCall package does with contaminated sequencing data

(Flickinger et al., 2015). Our own preliminary analyses suggest this would reduce the rate of missing and incorrect genotype calls in contaminated samples.

Genotype probe intensities are approximately normally distributed around the values of 0, ½, and 1 (depending on the underlying genotype), with truncation resulting in additional point masses at 0 and 1. Contaminating DNA results in a proportional shift in these distributions, as reflected in Figure 3-2. In principle, direct modeling of this intensity distribution (see Appendix) would enable us to predict the distribution of probe intensities for samples with different degrees of contamination, to model resulting increases in missing genotype rates (when intensities are drawn from the shifted distributions they will fall more often in ambiguous regions that lie between two expected genotype clusters) and in genotyping error rates. These models would allow predictions of the impact of genotyping error rate on power (as in Sobel, Papp, and Lange (2002)) or, potentially, methods for association analysis that model the underlying intensity data directly rather than relying on discrete genotype calls (as done in Kim, Gordon, Sebat, Ye, and Finch (2008) for structural variants, for example).

In our own work, we often must decide on acceptable thresholds for sample contamination. For simple regression-based approaches that model phenotypes as a function of genotypes and covariates, it's tempting to be lenient and analyze samples that have modest amounts of contamination – after all, a contaminated sample with a few erroneous genotypes will still provide some useful information, albeit less information than an uncontaminated sample. However, many modern genetic analyses include additional analysis steps that involve sharing of information across samples – these steps might include haplotype estimation (which relies on identification of shared IBD segments between samples and is a key step in genotype imputation analyses) and also estimation of genetic kinship matrices or principal components of ancestry (which are also key

steps for modern large scale genetic analyses that include related individuals or samples of diverse ancestry). In our experience, contaminated samples can have more deleterious effects for these analyses, corrupting the information contributed by other uncontaminated samples. Empirically, we typically recommend that samples with contamination greater than ~1% to 3% should be excluded from downstream analyses.

In conclusion, we have introduced VICES, a method that performs joint estimation of contamination and its sources in genotyping array samples. This innovation results in more accurate contamination estimates which are robust in genotyping cohorts of diverse ancestry. VICES allows researchers to estimate contamination easily without importing allele frequencies and provides additional information on how their samples were contaminated, so that it can be prevented or dealt with more effectively.

Conflicts of Interest

G.R.A. is currently an employee of Regeneron Pharmaceuticals and the beneficiary of stock options and grants in Regeneron. Previously, he served on scientific advisory boards for 23andMe, Regeneron Pharmaceuticals and Helix.

Acknowledgements

G.R.A. was supported by HG007022. The authors acknowledge the Michigan Genomics Initiative, University of Michigan Precision Health, and the University of Michigan Medical School Central Biorepository for providing the collection, biospecimen storage, management, and distribution services in support of the research reported in this publication. The Michigan Genomics Initiative was supported in part through resources and services provided by Precision Health at the University of Michigan, Ann Arbor. We thank Kim Doheny, Kurt Hetrick, Matthew

Flickinger, and Goo Jun for sharing their data from the intentionally mixed HapMap samples. Support for the collection of these data was provided by NIH contract HHSN268201700006I. We also thank Jonathan LeFaive, Peter Van de Haar, and Sayantan Das for suggestions on coding and portability, and Alan Kwong for proposing the name, VICES. Dr. Kirsten Herold of the SPH Writing Lab helped with proofreading and clarifying the manuscript.

Tables

Table 3-1 HapMap sample mixture proportions

No. Samples	NA06990 (CEU)	NA07055 (CEU)	NA18504 (YRI)	NA19200 (YRI)
6	**0%**	100%	0%	0%
2	99.5%	**0.5%**	0%	0%
2	99%	**1%**	0%	0%
2	98%	**2%**	0%	0%
2	97%	**3%**	0%	0%
2	95%	**5%**	0%	0%
2	90%	**10%**	0%	0%
6	0%	0%	100%	**0%**
2	0%	0%	99.5%	**0.5%**
2	0%	0%	99%	**1%**
2	0%	0%	98%	**2%**
2	0%	0%	95%	**5%**
2	0%	0%	90%	**10%**

Composition of 34 mixtures of HapMap cell lines from NA06990, NA07055, NA18504, and NA19200. The contamination percentages are in bold.

Table 3-2 Accuracy metrics of contamination methods, correct allele frequencies

	VICES-Geno	VICES-AF	BAFRegress	VerifyIDintensity
RMSE	0.0057	0.0068	0.0054	0.031
Bias	-0.0035	-0.0041	-0.0024	-0.0085
Increase in abs. error per 1% increase in contamination	0.0012	0.0015	0.0011	0.0056

Root-mean-squared-error (RMSE), bias, and change in absolute error per 1% higher contamination of the three methods against the intended contamination of the 34 HapMap CEU samples.

Table 3-3 Accuracy metrics of contamination methods, incorrect allele frequencies

	VICES-AF	BAFRegress	VerifyIDintensity
RMSE	0.023	0.026	0.031
Bias	-0.014	-0.015	-0.0086
Increase in abs. error per 1% increase in contamination	0.0057	0.0065	0.0056

Root-mean-squared-error, bias, and change in absolute error per 1% higher contamination of the three methods against the intended contamination of the 34 intentionally mixed HapMap samples when 1000 Genomes allele frequencies from the incorrect population were used.

Figures

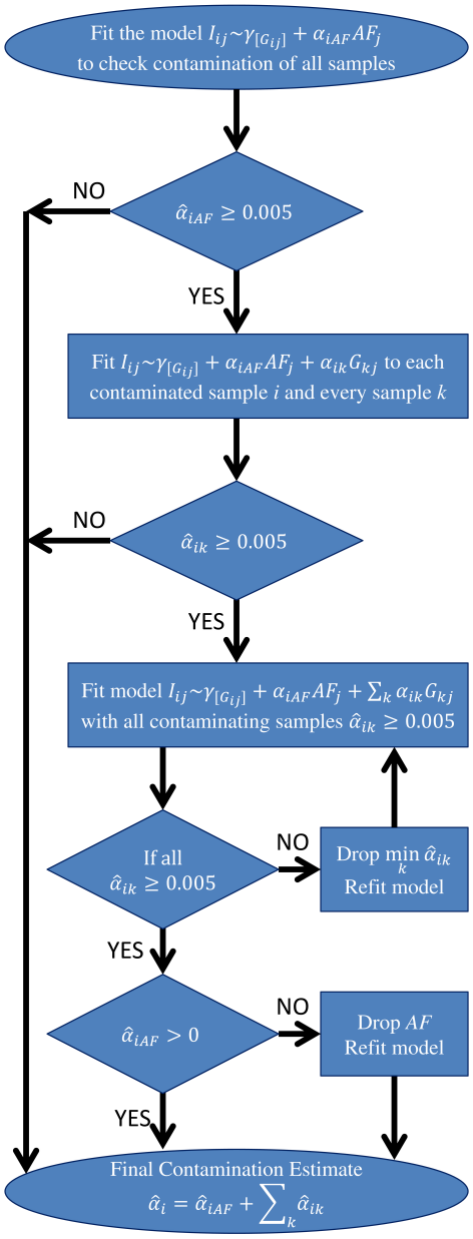


Figure 3-1 Flowchart of the contamination estimation algorithm

The flowchart shows how the algorithm progresses as contaminated samples are identified using allele frequencies, then potential contaminating samples are found for them and model selection performed to prune contaminating samples and calculate the final estimates.

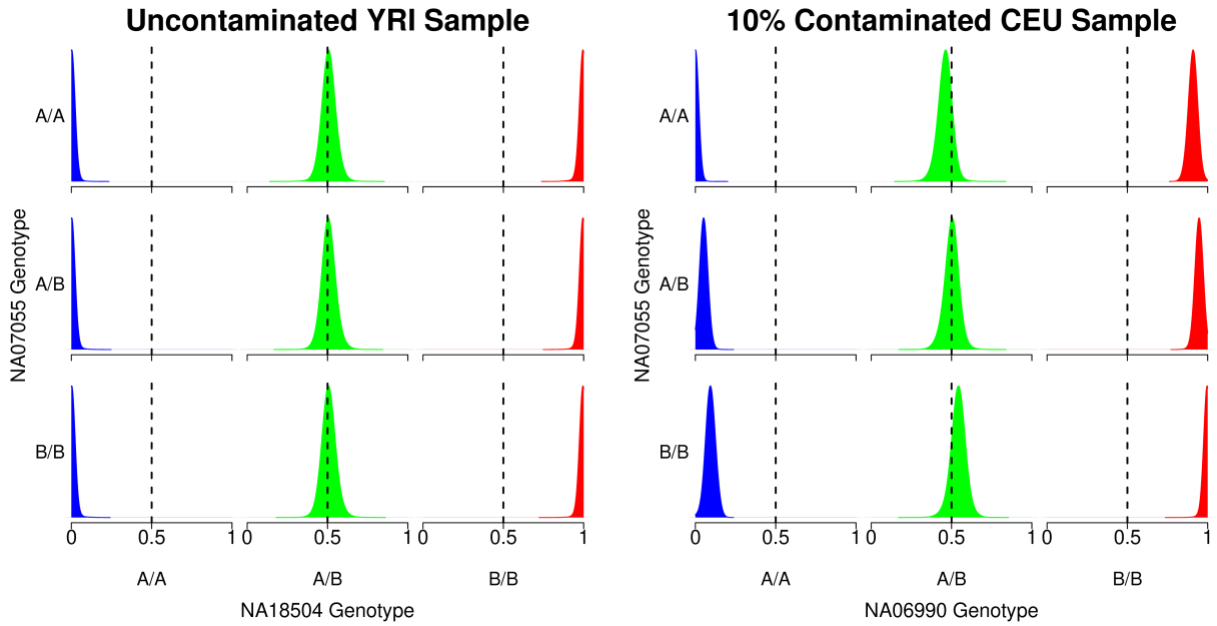


Figure 3-2 Distribution of array probe intensities by genotype

Kernel density plots showing the distribution of array probe intensities for an uncontaminated HapMap Yoruban sample (NA18504, left) and a 10% contaminated HapMap European sample (NA06990, right) as a function of the genotypes of NA07055. It is apparent that the intensities of the contaminated sample shift in the direction of NA07055 genotypes.

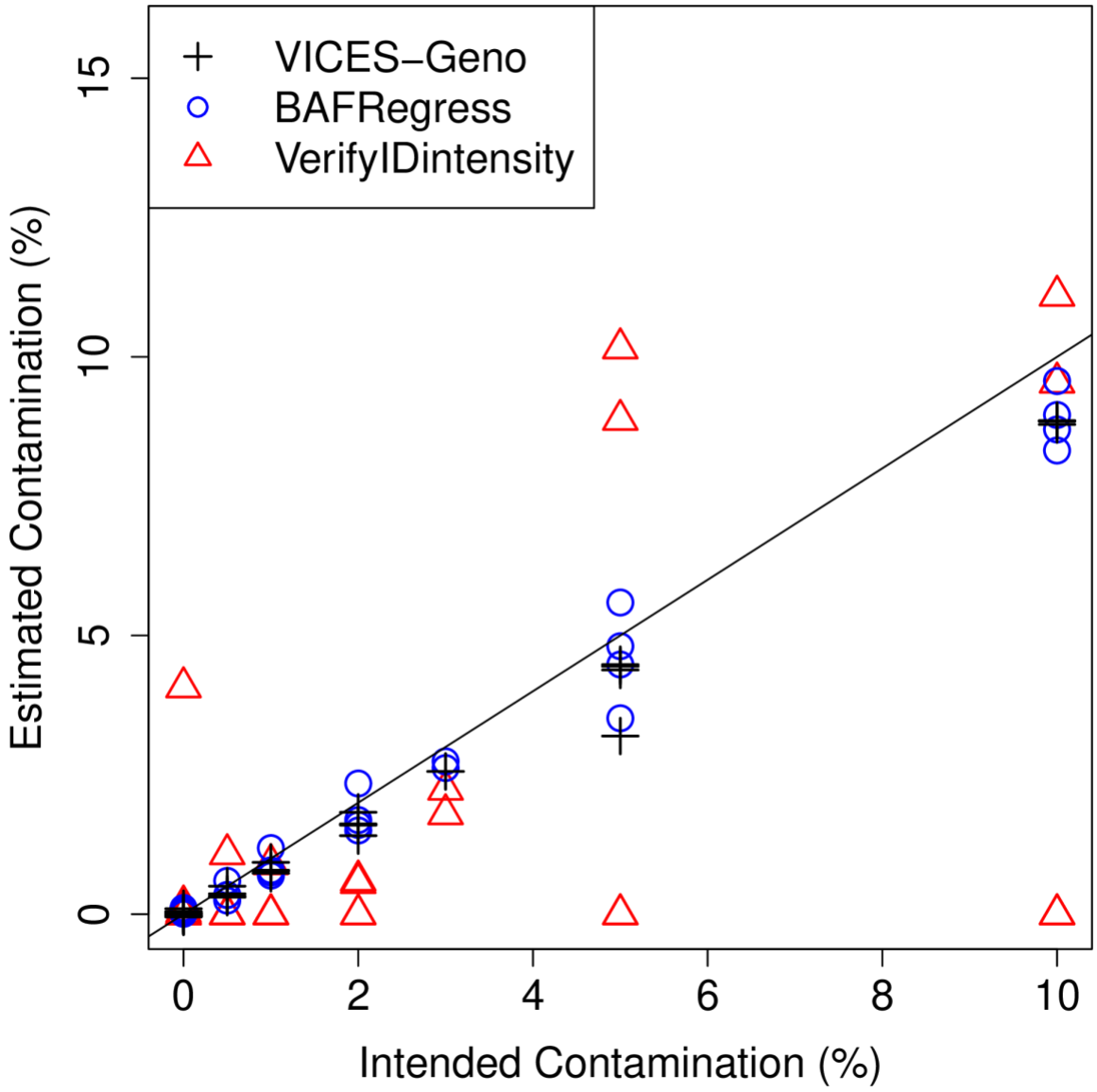


Figure 3-3 Comparison of contamination estimates in HapMap

Comparison of estimates from our method using contaminating sample genotypes, BAFRegress, and VerifyIDintensity on the 34 mixtures of HapMap DNA to the intended contamination proportion.

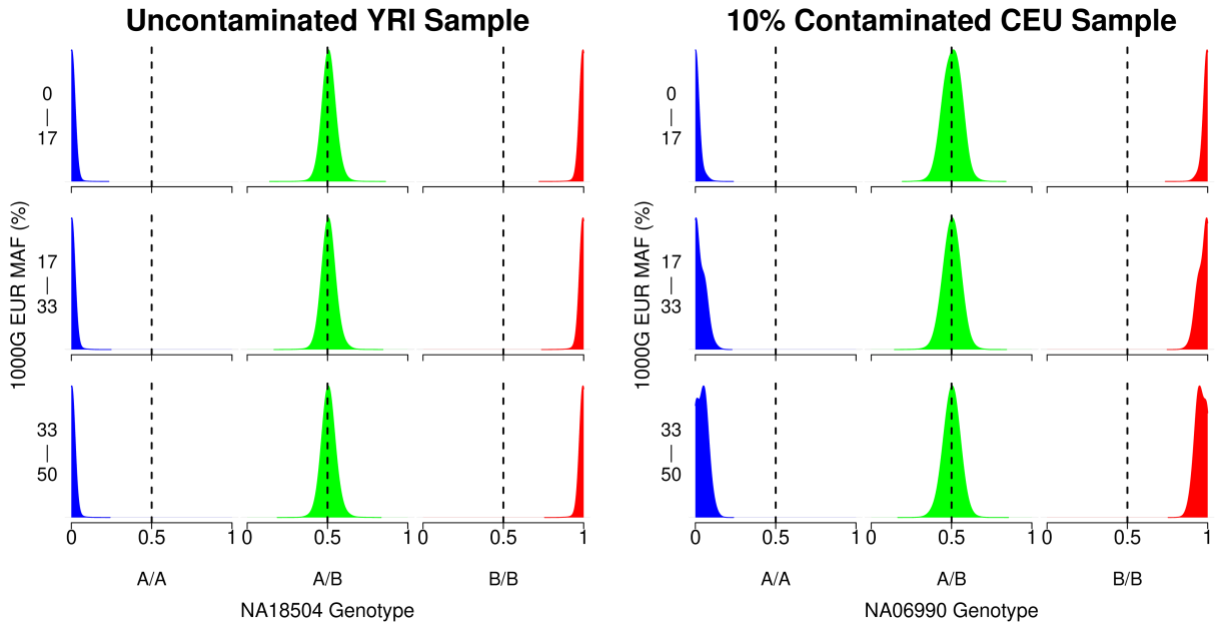


Figure 3-4 Distribution of array probe intensities by correctly-specified MAF

Kernel density plots showing the distribution of array probe intensities for an uncontaminated HapMap Yoruban sample (NA18504, left) and a 10% contaminated HapMap European sample (NA06990, right) at different 1000 Genomes European minor allele frequency (MAF) bins. The sample NA07055 that contributed DNA to the contaminated sample on the right is from the same ancestral population that the MAFs were calculated in, so using the MAFs to estimate contamination with a method like BAFRegress in this case would result in a good estimate for the intended contamination of 10%.

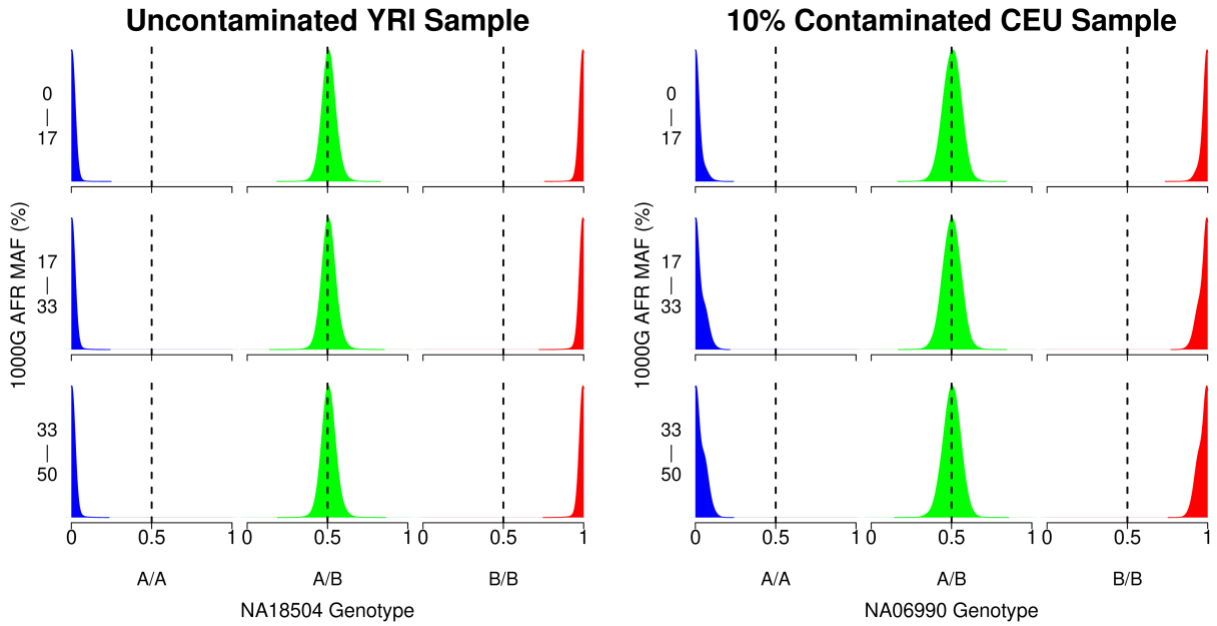


Figure 3-5 Distribution of array probe intensities by misspecified MAF

Kernel density plots showing the distribution of array probe intensities for an uncontaminated HapMap Yoruban sample (NA18504, left) and a 10% contaminated HapMap European sample (NA06990, right) at different 1000 Genomes African minor allele frequency (MAF) bins. The sample NA07055 that contributed DNA to the contaminated sample on the right is European while the MAFs were calculated from African samples, so using the MAFs to estimate contamination with a method like BAFRegress in this case would result in a dramatic underestimate for the intended contamination of 10%.

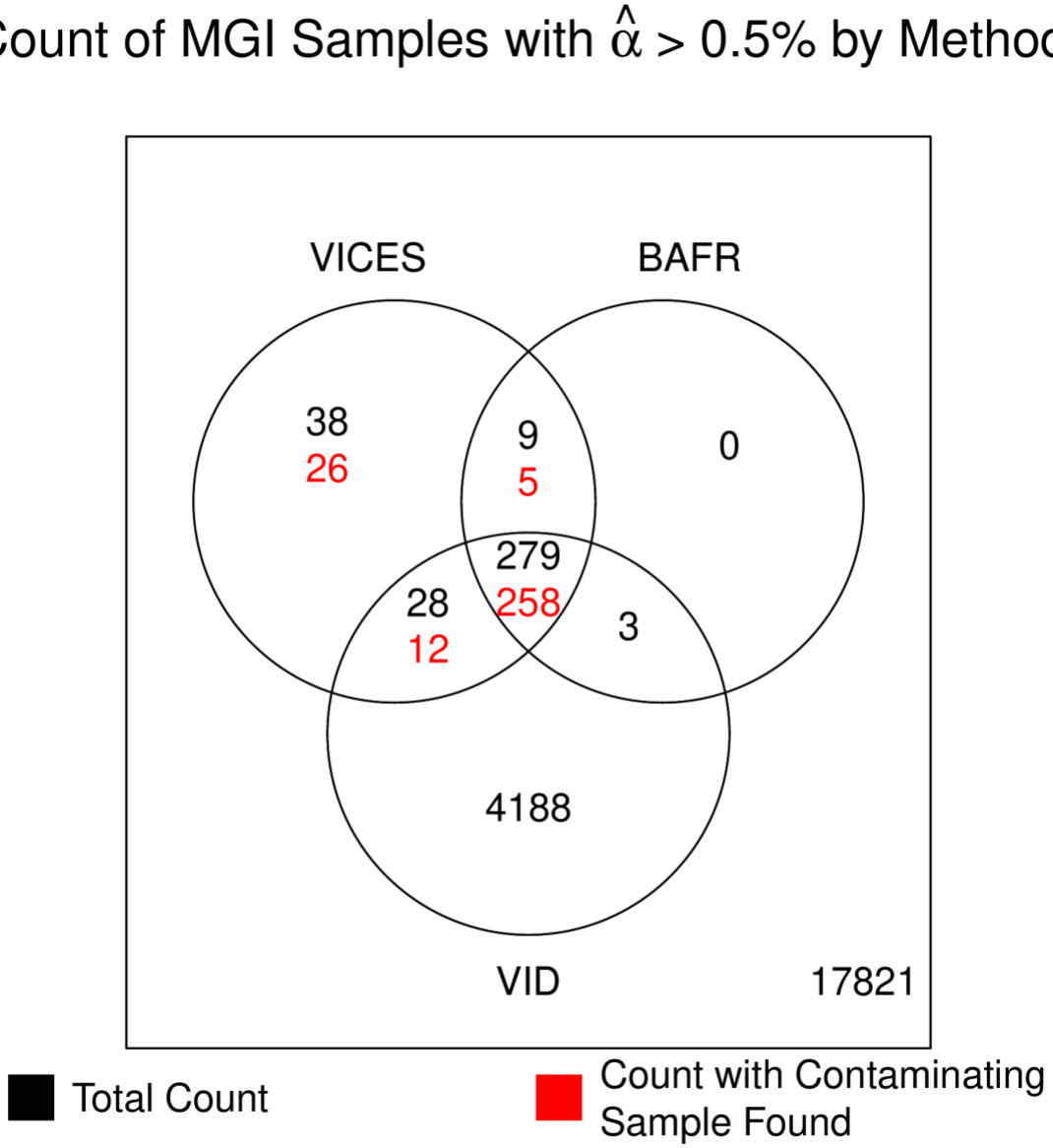


Figure 3-6 Count of contaminated MGI samples by method

Venn diagram showing (black) the count of all Michigan Genomics Initiative samples with estimated contamination greater than 0.5% by VICES, BAFRegress (BAFR), or VerifyIDintensity (VID) or any combination of the three methods, and (red) the count with estimated contamination greater than 0.5% and a contaminating sample found by VICES.

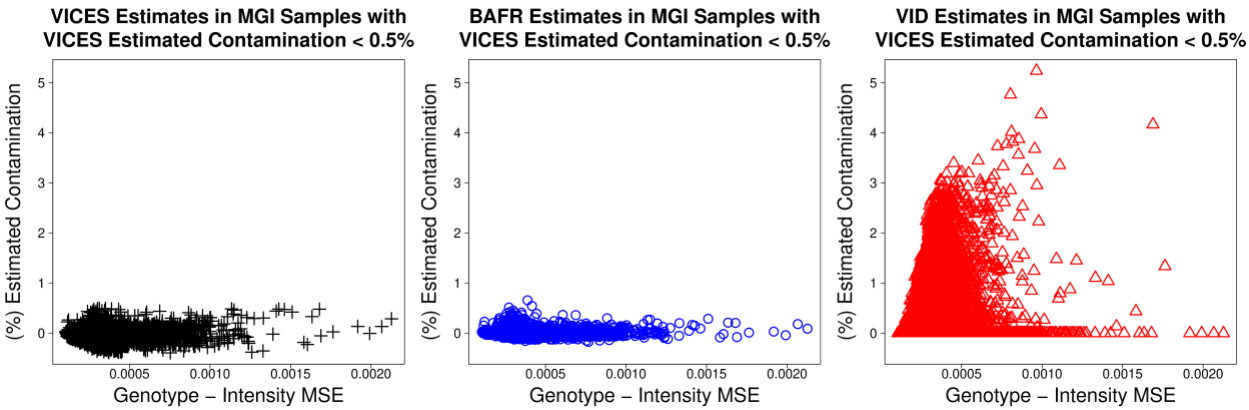


Figure 3-7 VerifyIDIntensity contamination estimates affected by noisy array intensities

Estimated contamination of the three methods as a function of mean-squared-error between intensity and called genotype, in 22,012 Michigan Genomics Initiative samples with contamination < 0.5% as estimated by VICES.

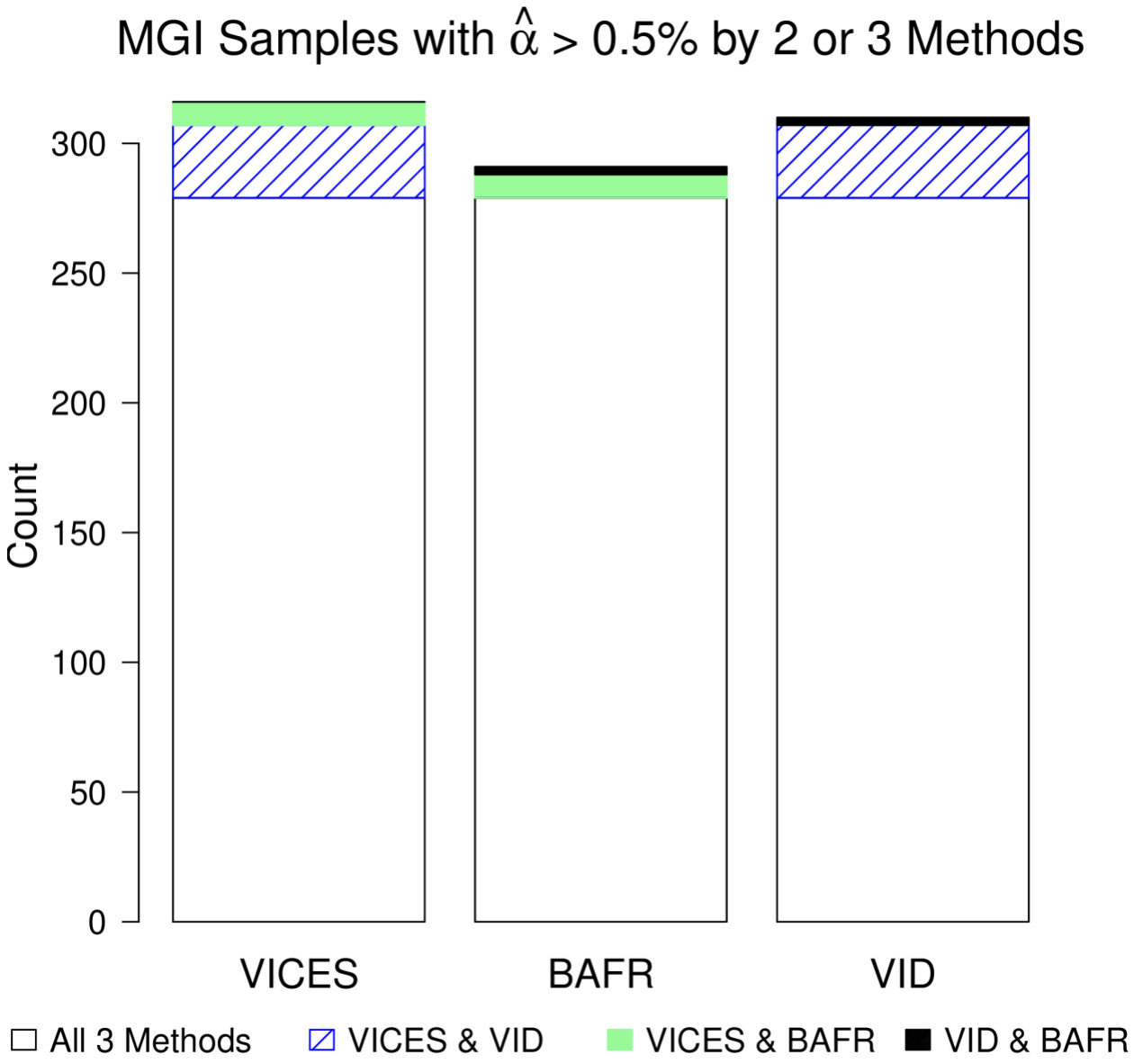


Figure 3-8 Agreement of contamination estimates in MGI by method

Bar plot of the count of Michigan Genomics Initiative samples with estimated contamination greater than 0.5% by VICES, BAFRegress, or VerifyIDintensity and at least one other method.

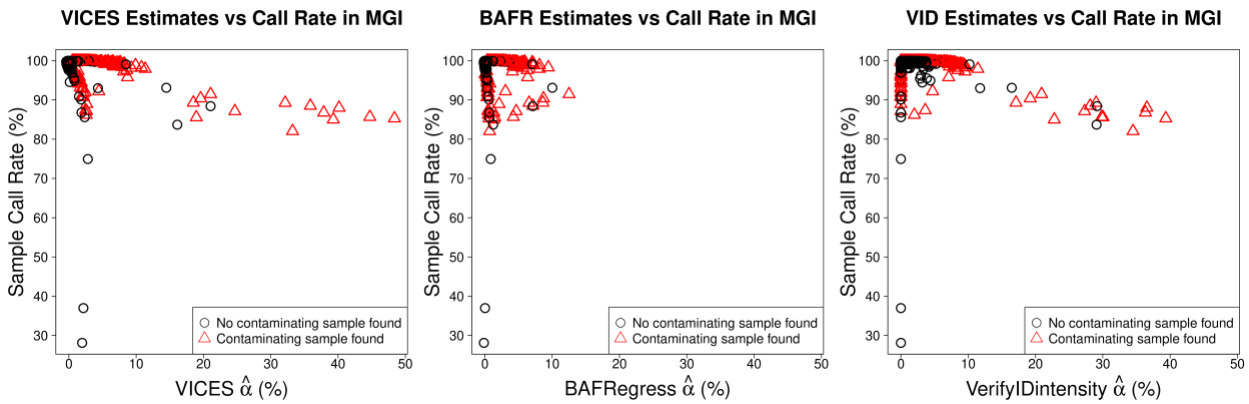


Figure 3-9 Contamination estimates and call rate by method

Comparing estimated contamination in 22,366 Michigan Genomics Initiative samples and their call rates. Left: VICES. Center: BAFRegress. Right: VerifyIDintensity. In all three plots, the red triangles denote the samples that had a contaminating sample detected by our method.

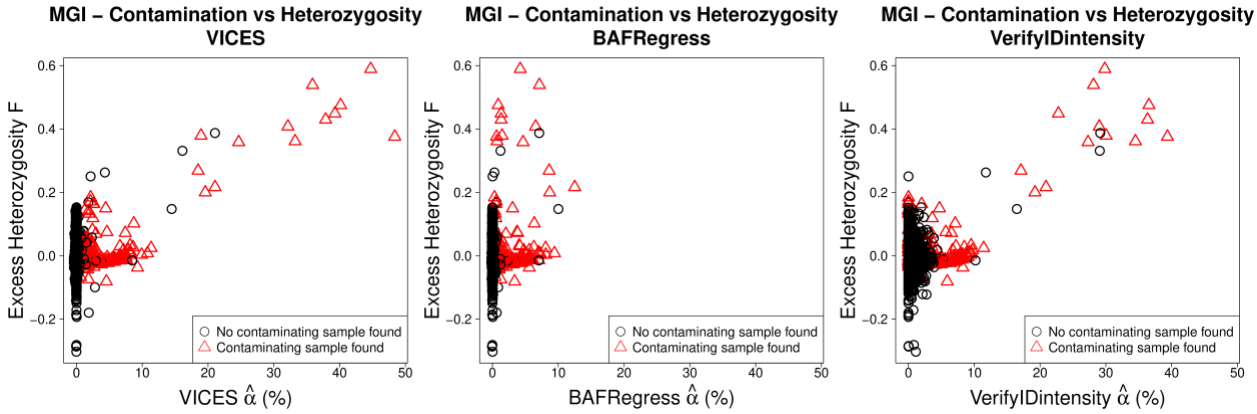


Figure 3-10 Contamination estimates and excess heterozygosity by method

Comparing estimated contamination in 22,366 MGI samples and excess heterozygosity as calculated using Plink 1.9. Left: VICES. Center: BAFRegress. Right: VerifyIDintensity. In all three plots, the red triangles denote the samples that had a contaminating sample detected by our method.

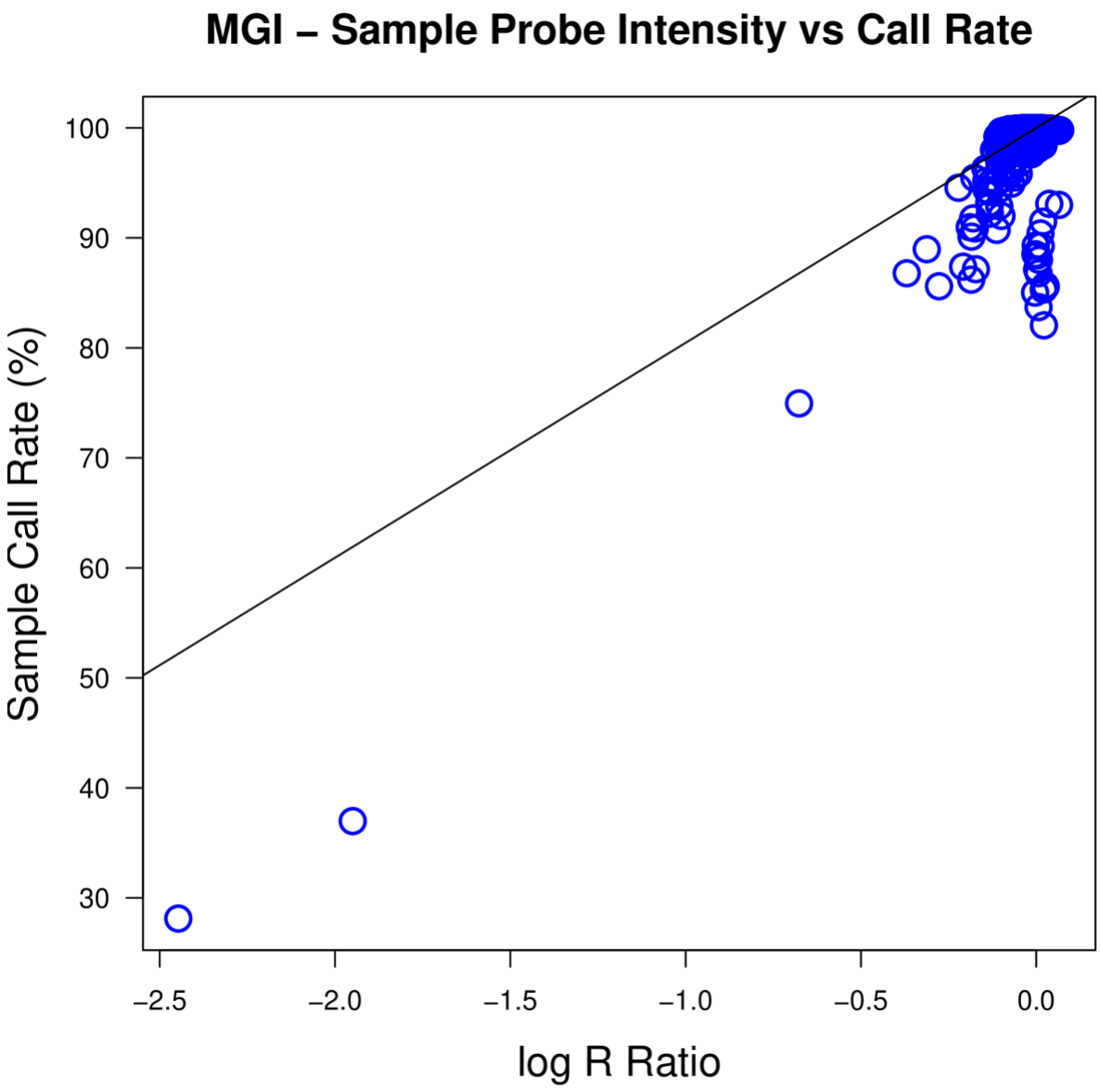


Figure 3-11 Sample probe intensity vs call rate in MGI

Scatterplot of sample call rate for 22,366 genotyped samples from the Michigan Genomics Initiative and average array probe intensity as measured by $\log_2$ R ratio. The black line shows the regression fit to these data.

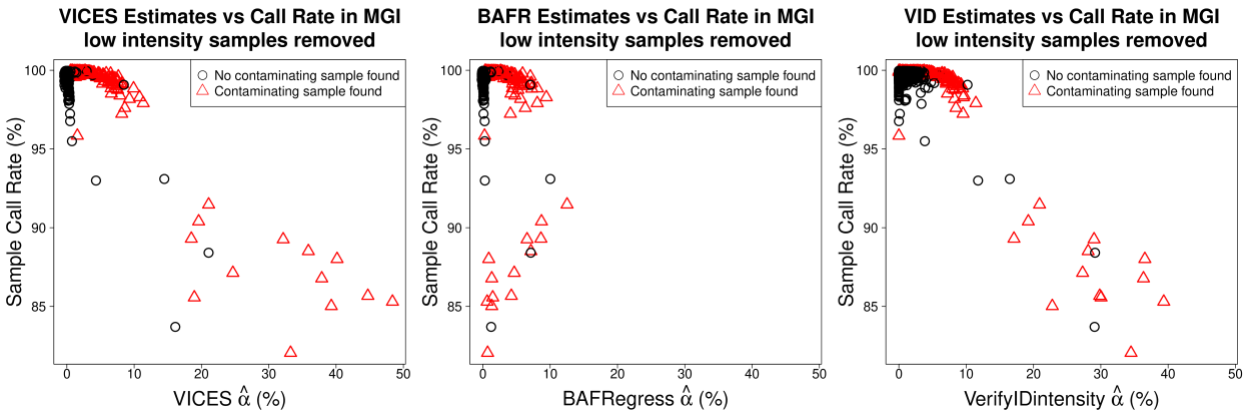


Figure 3-12 Contamination estimates and call rate by method, low intensity samples removed

Comparing estimated contamination against call rates in 22,201 Michigan Genomics Initiative samples that had average array probe intensity (defined as $\log_2$ R ratio) greater than a cutoff set at 2 standard deviations below the mean. Left: VICES. Center: BAFRegress. Right: VerifyIDintensity. In all three plots, the red triangles denote the samples that had a contaminating sample detected by our method.

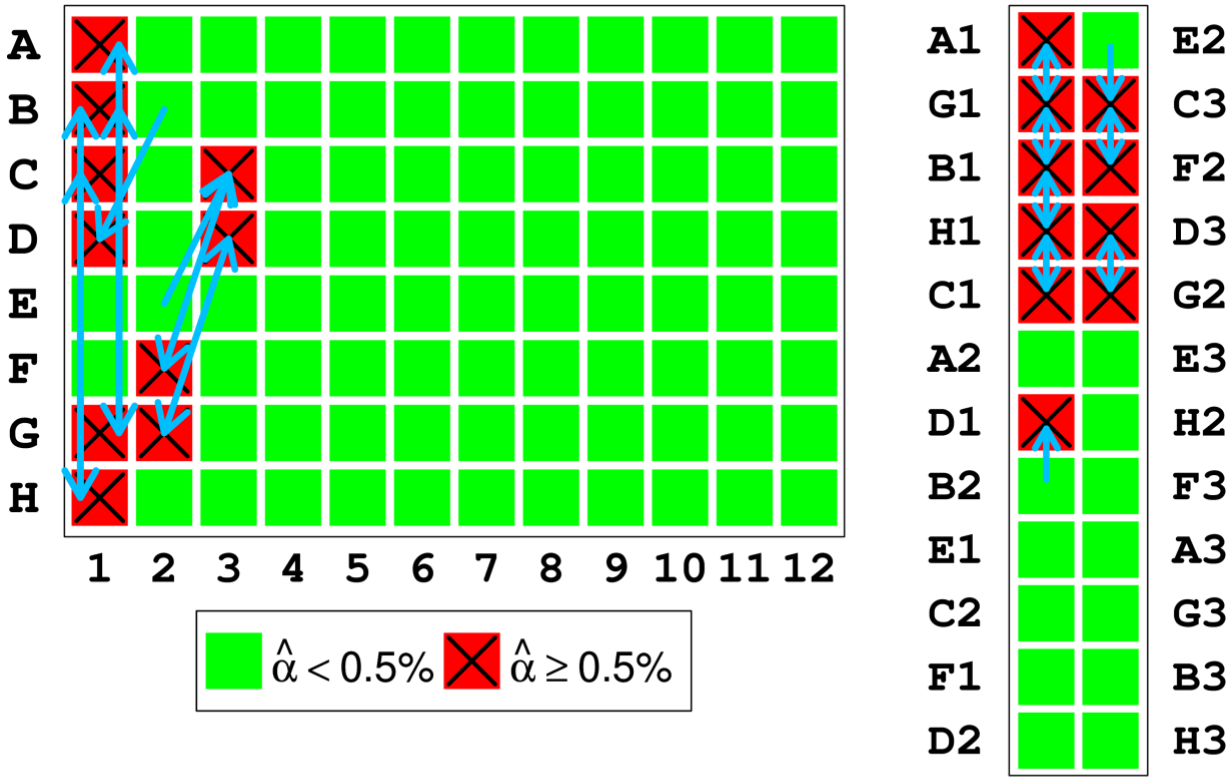


Figure 3-13 Position of contaminated samples in a genotyping experiment in MGI

Left: Eight contaminated samples as they appeared on part of the sample preparation plate. The letters to the left of the plate indicate the rows and numbers below indicate columns. Arrows indicate our method's estimates for which sample contributed DNA to each contaminated sample. **Right:** The position of the same samples on the genotyping array. Letters and numbers indicate the row and column of the plate from which the samples were transferred. Arrows have the same interpretation.
This figure shows that the contaminated samples are adjacent to their contaminating sample on the array, while far apart and without a clear pattern on the processing plate. The relative ease of explaining the pattern of adjacent mixing on the array compared to the processing plate suggests that the DNA mixture occurred on the array itself.

References

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience, 4*, 8. eCollection 2015. doi:10.1186/s13742-015-0047-8 [doi]

Cibulskis, K., McKenna, A., Fennell, T., Banks, E., DePristo, M., & Getz, G. (2011). ContEst: estimating cross-contamination of human samples in next-generation sequencing data. *Bioinformatics (Oxford, England), 27*(18), 2601-2602. doi:10.1093/bioinformatics/btr446 [doi]

Diskin, S. J., Li, M., Hou, C., Yang, S., Glessner, J., Hakonarson, H., . . . Wang, K. (2008). Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic acids research, 36*(19), e126. doi:10.1093/nar/gkn556 [doi]

Flickinger, M., Jun, G., Abecasis, G. R., Boehnke, M., & Kang, H. M. (2015). Correcting for Sample Contamination in Genotype Calling of DNA Sequence Data. *American Journal of Human Genetics, 97*(2), 284-290. doi:10.1016/j.ajhg.2015.07.002 [doi]

Fritsche, L. G., Gruber, S. B., Wu, Z., Schmidt, E. M., Zawistowski, M., Moser, S. E., . . . Mukherjee, B. (2018). Association of Polygenic Risk Scores for Multiple Cancers in a Phenome-wide Study: Results from The Michigan Genomics Initiative. *Am J Hum Genet, 102*(6), 1048-1061. doi:10.1016/j.ajhg.2018.04.001

Genomes Project, C., Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., . . . Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature, 526*(7571), 68-74. doi:10.1038/nature15393

Goes, F. S., McGrath, J., Avramopoulos, D., Wolyniec, P., Pirooznia, M., Ruczinski, I., . . . Pulver, A. E. (2015). Genome-wide association study of schizophrenia in Ashkenazi Jews. *American journal of medical genetics.Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics, 168*(8), 649-659. doi:10.1002/ajmg.b.32349 [doi]

Heiss, J. A., & Just, A. C. (2018). Identifying mislabeled and contaminated DNA methylation microarray data: an extended quality control toolset with examples from GEO. *Clin Epigenetics, 10*, 73. doi:10.1186/s13148-018-0504-1

Hoffmann, T. J., Ehret, G. B., Nandakumar, P., Ranatunga, D., Schaefer, C., Kwok, P. Y., . . . Risch, N. (2017). Genome-wide association analyses using electronic health records identify new loci influencing blood pressure variation. *Nature genetics, 49*(1), 54-64. doi:10.1038/ng.3715 [doi]

Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., . . . Craig, D. W. (2008). Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS genetics, 4*(8), e1000167. doi:10.1371/journal.pgen.1000167 [doi]

Illumina. (2013). *Infinium® HTS Assay Protocol Guide*. San Diego, CA.

Illumina. (2016). *GenomeStudio® Genotyping Module v2.0 Software Guide*. San Diego, CA.

Illumina. (2017). Infinium ® CoreExome-24 v1.2 BeadChip.

International HapMap, C., Altshuler, D. M., Gibbs, R. A., Peltonen, L., Altshuler, D. M., Gibbs, R. A., . . . McEwen, J. E. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature, 467*(7311), 52-58. doi:10.1038/nature09298

Jun, G., Flickinger, M., Hetrick, K. N., Romm, J. M., Doheny, K. F., Abecasis, G. R., . . . Kang, H. M. (2012). Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *American Journal of Human Genetics, 91*(5), 839-848. doi:10.1016/j.ajhg.2012.09.004 [doi]

Kim, W., Gordon, D., Sebat, J., Ye, K. Q., & Finch, S. J. (2008). Computing power and sample size for case-control association studies with copy number polymorphism: application of mixture-based likelihood ratio test. *PloS one, 3*(10), e3475. doi:10.1371/journal.pone.0003475

Li, G. (2016). A new model calling procedure for Illumina BeadArray data. *BMC genetics, 17*(1), 90. doi:10.1186/s12863-016-0398-x

Li, Y., Willer, C. J., Ding, J., Scheet, P., & Abecasis, G. R. (2010). MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic epidemiology, 34*(8), 816-834. doi:10.1002/gepi.20533 [doi]

Liu, J. Z., van Sommeren, S., Huang, H., Ng, S. C., Alberts, R., Takahashi, A., . . . Weersma, R. K. (2015). Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nature genetics, 47*(9), 979-986. doi:10.1038/ng.3359 [doi]

Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., Day, F. R., . . . Speliotes, E. K. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature, 518*(7538), 197-206. doi:10.1038/nature14177 [doi]

Mahajan, A., Go, M. J., Zhang, W., Below, J. E., Gaulton, K. J., Ferreira, T., . . . Morris, A. P. (2014). Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nature genetics, 46*, 234. doi:10.1038/ng.2897 https://www.nature.com/articles/ng.2897#supplementary-information

Marouli, E., Graff, M., Medina-Gomez, C., Lo, K. S., Wood, A. R., Kjaer, T. R., . . . Lettre, G. (2017). Rare and low-frequency coding variants alter human adult height. *Nature, 542*(7640), 186-190. doi:10.1038/nature21039 [doi]

Peiffer, D. A., Le, J. M., Steemers, F. J., Chang, W., Jenniges, T., Garcia, F., . . . Gunderson, K. L. (2006). High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res, 16*(9), 1136-1148. doi:10.1101/gr.5402306

Schmieder, R., & Edwards, R. (2011). Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PloS one, 6*(3), e17288. doi:10.1371/journal.pone.0017288 [doi]

Sobel, E., Papp, J. C., & Lange, K. (2002). Detection and integration of genotyping errors in statistical genetics. *Am J Hum Genet, 70*(2), 496-508. doi:10.1086/338920

Voight, B. F., Kang, H. M., Ding, J., Palmer, C. D., Sidore, C., Chines, P. S., . . . Boehnke, M. (2012). The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet, 8*(8), e1002793. doi:10.1371/journal.pgen.1002793

Chapter 4 A Fast Linkage Method for a Population GWAS Cohort with Related Individuals

Introduction

Linkage analysis jointly models the inheritance of a trait and genetic material in a family. One group of methods developed to study linkage of quantitative traits uses variance-component models to relate identical-by-descent (IBD) sharing to phenotype similarity. Variance components models are statistical models that partition trait variances and co-variances. A typical variance components model for quantitative trait analysis will model the observed trait values using a multivariate normal distribution, and partition trait variances and covariances into shared additive genetic effects and individual specific environmental effects. When used for linkage analysis, the variance components model usually also contains a component for region-specific genetic effects influencing the trait (Amos, 1994). Using this model, it is possible to construct a test for genetic linkage by estimating and testing this variance component for region-specific genetic effects.

A variety of algorithms exist for estimating and testing variance components have been applied to linkage analysis. These include iterative methods for fitting variance components in a general setting like maximum likelihood (Lange & Boehnke, 1983), restricted maximum likelihood (Van Arendonk, Tier, Bink, & Bovenhuis, 1998), and generalized estimating equations (Amos, 1994). In addition to these iterative methods, Haseman and Elston developed a regression-based approach to estimate variance components for the specific application of inferring genetic linkage in sibling pairs (Haseman & Elston, 1972). Haseman-Elston regression fits variance

components by a methods of moments estimator derived from the expectation of either the product, squared sum, or squared difference of pairs of observations (Sham & Purcell, 2001). Their approach has since been generalized to linkage analysis in other relationship types (Sham, Purcell, Cherny, & Abecasis, 2002) and to estimating genetic variance components in unrelated individuals (G. B. Chen, 2014).

The class of linkage methods that exist today was developed when genotype data were relatively sparse and expensive to collect. With the current ability to assay genetic variation at a large number of markers using genotyping arrays or short-read sequencing at much lower costs, genome-wide association scans (GWAS) have become widespread as a method for gene-mapping. Even so, many researchers still use linkage methods either because they collected their data before genotyping arrays were commercially available or to supplement a GWAS analysis of the same individuals (Kathiresan et al., 2007). Some have justified running linkage analysis in parallel with GWAS on the grounds that linkage outperforms GWAS in the presence of population structure or allelic heterogeneity (Minster et al., 2015). Linkage also continues to be a useful tool to associate traits with complex variation that is difficult to genotype like structural variants, copy number variants (Kathiresan et al., 2007), variants in highly repetitive regions (Mousavi, Shleizer-Burko, Yanicky, & Gymrek, 2019), or in loci that exhibit epistatic interaction (Hodge, Hager, & Greenberg, 2016).

As the cost of genotyping has fallen dramatically in recent decades, the development of linkage methods has lagged behind GWAS in its ability to keep up with the size and structure of modern data sets. To illustrate, consider MERLIN, a widely-used implementation of a variance-components linkage method (Abecasis, Cherny, Cookson, & Cardon, 2002). In order to run linkage analysis with MERLIN in an old-order Amish pedigree with 364 individuals and genotypes at

1991 microsatellite markers, the single large pedigree had to be split into nuclear families in order to complete the analysis in a tractable amount of time (Georgi et al., 2014). In addition, existing linkage methods are limited in their ability to model allele sharing between distant relatives when genotypes for intervening relatives are not available (Thompson, 2019).

In addition, existing linkage analysis methods often ignore the cryptic relatedness found even in studies that target unrelated individuals. Though Day-Williams and colleagues (2011) proposed a method to reconstruct pedigrees and perform linkage analysis using genotype data, their approach requires pedigrees which can be uniquely reconstructed from pairwise kinship data. This method could not be applied in a large biobank cohort with many pairs of relatives but few complete families.

A few have taken a more unified approach to linkage analysis and GWAS. The KELVIN method supports both linkage and association analysis in pedigrees, based on a posterior probability of linkage calculation from a Haseman-Elston regression fit (Vieland et al., 2011). This method was successfully used to study autism (Piven et al., 2013) and musical ability (Oikkonen et al., 2015). However, this method requires that families and the genetic relationships in pedigrees be defined prior to analysis and scales poorly beyond nuclear families.

In this paper, we propose Population Linkage, a fast method to perform variance-component linkage analysis on hundreds of traits with arbitrarily related individuals. IBD and kinship estimation only need to be performed once, then a variance components model fit for each trait at each region using Haseman-Elston regression, making the method scalable for studying hundreds of traits in thousands of individuals. The resulting estimates of the trait variance attributable to IBD sharing at a locus and its standard error can then be used to test for linkage and calculate LOD scores, a standard yardstick for genetic linkage signals. Our method uses only the

estimated relatedness and IBD segregation between pairs of individuals. We do not require knowledge of pedigree information or attempt to reconstruct pedigrees using the genetic data.

Material and Methods

Population linkage has four basic steps: 1. preparing the input data, including estimating kinship and identical-by-descent (IBD) regions for the cohort, 2. running diagnostic tests to select the appropriate variance-components model, 3. running the linkage analysis using Haseman-Elston regression, and 4. processing the results. In this section, we will describe these steps in detail together with various improvements in IBD estimation and Haseman-Elston regression that make our method feasible as well as our own innovations to achieve scalability to large datasets. Figure 4-1 is a flow chart that outlines some of the major steps in this method.

Notation

First, we introduce relevant notation. Assume we have genotype information on *n* individuals, and values for a quantitative trait, ***y***. For each pair of individuals, *i*, *j*, their relatedness can be summarized by their kinship ϕ_{ij}, which is the probability that two randomly sampled alleles (one from each individual) are identical-by-descent (IBD). The full kinship matrix for all $n(n-1)/2$ pairs of individuals is denoted as $\mathbf{\Phi}$. Two alternate summaries of the relationships between individuals include p_{ij}, the total proportion of DNA that is shared in IBD segments, and c_{ijl}, the total proportion of chromosome ends that are IBD in the first and last *l* megabases of each chromosome. These two additional summaries are important in ensuring calibration of our method as they enable us to cope with biases of population based IBD estimates. We denote the full matrices of p_{ij} and c_{ijl} as $\mathbf{P}$, and $\mathbf{C}_l$, respectively. The full set of IBD segments are contained in S, where each element s_{ijcse} indicates the IBD status (1 or 2) for individuals *i* and *j*, for the segment

starting on chromosome *c* and starting at position *s* and ending at position *e*. The IBD status (0, 1, or 2) for individuals *i* and *j* at marker *m* can also be indicated as d_{ijm}, with the full matrix of all individuals' IBD sharing at marker m as $\mathbf{D}_m$.

Input data preparation

The first step for Population Linkage is to prepare the input data. These consist of quantitative trait values from the cohort of interest and estimates of genetic relationships and allelic segregation (in the form of IBD estimates) between all pairs of individuals. The structure and requirements of these data will be described in greater detail in the following sub-sections.

Traits

This paper only considers linkage analysis with quantitative traits. The raw values $\boldsymbol{y}^*$ should be prepared for analysis by regressing on relevant covariates like age, sex, medication usage, and principal components, and the residuals inverse-normalized. This not only has the benefit of reducing the effect of extreme observations, but also results in trait values $\boldsymbol{y}$ that are standardized and centered around 0, which will be helpful for obtaining a cross product $\boldsymbol{y}\boldsymbol{y}^T$ that does not depend on the mean and variance of $\boldsymbol{y}^*$.

Kinship

There are a variety of different estimators for kinship $\boldsymbol{\Phi}$ that can be used for Population Linkage to model the contribution of overall genome-wide effects to the variance of a trait. The choice of any such estimator is associated with its own advantages and disadvantages. In this paper, we chose to estimate kinship from observed genotypes rather than reported pedigree relationships because in many large GWAS cohorts pedigree information can be absent, incomplete, or incorrect (Thomson & McWhirter, 2017). In addition, relationships estimated from genotypes can be more

informative than relationships reported from pedigrees since they will reflect how much genetic material is actually shared between a pair of relatives rather than theoretical expectations, which for some relationship types can diverge by a large amount.

In this paper, we use the estimator from Manichaikul et al that is implemented in the KING software package (2010)

$$\widehat{\mathbf{\Phi}}_{ij} = \frac{N_{Aa,Aa} - 2N_{AA,aa}}{2N_{Aa}^{(i)}} + \frac{1}{2} - \frac{1}{4}\frac{N_{Aa}^{(i)} + N_{Aa}^{(j)}}{N_{Aa}^{(i)}} \qquad \text{(Equation 1)}$$

where $N_{Aa,Aa}$ is the number of variants heterozygous in both individuals *i* and *j*, $N_{AA,aa}$ is the number of homozygous discordant variants between individuals *i* and *j*, $N_{Aa}^{(i)}$ is the number of sites heterozygous in individual *i*, and $N_{Aa}^{(j)}$ is the number of sites heterozygous in individual *j*. For *i* = *j*, we set $\widehat{\mathbf{\Phi}}_{ij} = 0.5$, the maximum kinship coefficient in outbred individuals. Using this estimator has several advantages in our context. The first is that this pairwise calculation is extremely computationally efficient in large data sets with tens of thousands of individuals. In addition, Equation 1 does not depend on allele frequencies and is robust in cohorts with population structure, capturing primarily genetic similarity due to recent family relationships as opposed to population-level genetic similarity. This is particularly important for a variance-components linkage analysis like Population Linkage since increased genome-wide genetic sharing for a pair of individuals from belonging to the same ancestral group does not translate to the pair having more co-segregated alleles from a recent ancestor. Finally, while the accuracy of Equation 1 drops off for more distant relationship types compared to close relatives, it is very accurate in close relatives who are likely to share large amounts of genetic material, and relative differences in $\widehat{\mathbf{\Phi}}_{ij}$ are still useful for predicting IBD sharing in more distant relatives.

This kinship estimator differs from what is calculated by default in variance components software like GEMMA and GCTA, which calculate a genetic relationship matrix (GRM) $\mathbf{A}$ using the following equation

$$\mathbf{A} = \frac{1}{M}\mathbf{W}\mathbf{W}^T \quad \text{(Equation 2)}$$

where $\mathbf{W}$ is the matrix of centered, standardized genotypes for the cohort, and M is the number of genetic markers (Yang, Lee, Goddard, & Visscher, 2011; X. Zhou & Stephens, 2012). We chose not to use a GRM estimate for kinship for Population Linkage because of the computational complexity of calculating a GRM in large cohorts, the fact that it captures population structure in addition to relatedness, and that it returns dissimilar $\mathbf{A}_{ij}$ values for the same relationship type.

IBD segments

In addition to kinship estimates which capture genome-wide similarity between pairs of individuals, Population Linkage requires estimates of which genomic segments are shared between relatives in order to map a trait to a particular genomic region. To accomplish this, we chose to use estimates of identical-by-descent (IBD) segments between pairs of individuals, similar to previous variance-components linkage methods for quantitative traits (Amos, 1994).

To estimate the set of all IBD segments S in our cohort, we identify contiguous blocks of at least 64 markers and 2.5 Mb in length between pairs of individuals that are consistent with IBD 2 (identical genotype for both) or IBD 1 (no discordant genotype pairs). We then call IBD 2 segments first since this condition is more stringent, followed by IBD 1. Any remaining genetic material is classified as IBD 0 between the pair. We used a particularly fast implementation of this method in the software package KING to estimate S. From $\hat{S}$, we calculate the matrix $\widehat{\mathbf{D}}_m$ at the

unique endpoints of all IBD segments in $\hat{S}$ where there is a change in IBD status. For the diagonal entries $i = j$, we set $\widehat{\mathbf{D}}_{ijm} = 2$ at all markers m.

We chose to estimate S and **D** with this simple algorithm as opposed to a hidden Markov model method like that described in (Boehnke & Cox, 1997) primarily for its computational simplicity. Because Population Linkage is designed to work without reported pedigree relationships, IBD segments between all pairs of individuals had to be considered to avoid biased results. Even limiting the search to pairs above a modest estimated kinship $\widehat{\mathbf{\Phi}}_{ij}$ would have been problematic because it would have caused any trait with a strong genetic basis to appear to be linked with IBD sharing in a given region, even where no such relationship exists. In contrast, using the above method between all pairs in its KING implementation can practically scale to estimate IBD in GWAS cohorts with hundreds of thousands of individuals.

Proportion of IBD

A single statistic for genome-wide genetic similarity between pairs of individuals can fail to capture all the subtleties and nuances that exist in genetic relationships. For this reason, in addition to estimated kinship, $\widehat{\mathbf{\Phi}}$, we also calculate the genome-wide proportion $\widehat{\mathbf{P}}$ of genetic material estimated to be shared IBD between pairs of individuals from the set of all IBD segments $\hat{S}$ using the following formula:

$$\widehat{\mathbf{P}}_{ij} = \frac{\sum_{\hat{s}_{ijcse} \in \hat{S}_{ij}} \hat{s}_{ijcse}}{2\sum_{c=1}^{N_c} len(c)} \qquad \text{(Equation 3)}$$

where $\hat{S}_{ij}$ is the set of all estimated IBD segments between individuals i and j, N_c indicates the number of chromosomes, c an individual chromosome, $len(c)$ is a function that returns the length of chromosome c in base-pairs, and $\hat{s}_{ijcse}$ indicates the estimated IBD status (1 or 2) for individuals i and j, for the segment on chromosome c starting at position s and ending at position e. In Equation

3, each $\widehat{\mathbf{P}}_{ij}$ is a simple average of the estimated IBD status at every genomic position between individuals i and j. We set $\widehat{\mathbf{P}}_{ij} = 2$ for all $i = j$. $\widehat{\mathbf{P}}$ is more accurate than $\widehat{\mathbf{\Phi}}$ for close relatives, but is biased downward for distant relatives since the above method for estimating S does not consider IBD segments less than 2.5 Mb or 64 markers in length. Empirically, this downward bias in distant relatives appears to be most severe near chromosome ends. Estimates of $\widehat{\mathbf{P}}_{ij}$ in distant relatives who share short IBD segments would therefore be underestimates of the true proportion $\mathbf{P}_{ij}$.

Proportion of IBD chromosome ends

There are additional challenges to estimating IBD segments at the ends of chromosomes. The end of each chromosome physically truncates the length of IBD segments, commercial genotyping arrays typically have lower marker density near telomeres, and the higher recombination rate results in shorter IBD segments that are more difficult to identify. These factors all lead to downward bias in $\widehat{\mathbf{D}}_{ijm}$ toward the ends of chromosomes, and importantly, the shared segments that can be identified tend to be concentrated in closer relatives who often have more similar trait values because of shared non-genetic but familial factors.

For these reasons, for all pairs of individuals we estimate $\mathbf{C}_{ijl}$, the proportion of chromosome ends (of length l) that contain an IBD segment between individuals i and j. We use the following formula:

$$\hat{\mathbf{C}}_{ijl} = \frac{1}{4N_c}\sum_{c=1}^{N_c}\left(\sum_{s<l} s_{ijcse} + \sum_{e>len(c)-l} s_{ijcse}\right) \qquad \text{(Equation 4)}$$

where N_c indicates the number of chromosomes, c an individual chromosome, $len(c)$ is a function that returns the length of chromosome c in base-pairs, and s_{ijcse} indicates the IBD status (1 or 2) for individuals i and j, for the segment starting on chromosome c and starting at position s and ending at position e. For $i = j$, we set $\hat{\mathbf{C}}_{ijl} = 1$, the maximal value.

There are several advantages to estimating $\hat{\mathbf{C}}_l$ with Equation 4. The length parameter l can be tuned according to what works best for a particular data set or phenotype. This tuning process, described in greater detail later in this section, is achieved by testing Population Linkage with different values of l on a subset of the data. The value of l that results in the best performance can then be used to analyze the complete data. Also, since IBD estimates near the ends of chromosomes are inherently unreliable for the aforementioned reasons, Equation 4 does not attempt to estimate the proportion of IBD between pairs of individuals in the ends of chromosomes, a quantity we are not interested in. Instead, Equation 4 only captures the presence of estimated IBD segments to identify the pairs of individuals that are more likely to have an estimated IBD segment there and can therefore be used to account for the bias in IBD estimates these regions.

Statistical model

After the input data has been prepared, there are many options for which inputs can then be used for variance-components estimation and linkage analysis. We begin this section by describing our framework for variance-components estimation and linkage analysis using cross-product Haseman-Elston regression, then by outlining a strategy for users of Population Linkage to decide on a variance-components model for linkage analysis.

Single-VC model

For a model with a single variance component for additive genetic effects, $\sigma^2_{\boldsymbol{K}}$, we model trait variance in the following way:

$$var(\boldsymbol{y}) = \mathbf{K}\sigma^2_{\boldsymbol{K}} + I\sigma^2_e \qquad \text{(Equation 5)}$$

where **K** is a matrix of genetic relationships that has been centered so all rows and columns sum to 0 and scaled by the mean of its diagonal terms, and I is an *n*-by-*n* identity matrix. Centering and

scaling **K** in this way accounts for the scaling and centering of $\boldsymbol{y}$ that took place during its inverse-normal transform and which changes the variance-covariance matrix of $\boldsymbol{y}$. In addition, since **K** is scaled, $\sigma_{\boldsymbol{K}}^2$ can also be interpreted as the proportion of variance explained (PVE) by additive genetic effects and $\sigma_e^2 = 1 - \sigma_{\boldsymbol{K}}^2$ is the proportion of trait variance attributed to environmental effects and individual variability.

We obtain point estimates for $\sigma_{\boldsymbol{K}}^2$ and σ_e^2 by cross-product Haseman-Elston regression using the following formulas from (X. Zhou, 2017)

$$q = \left(\boldsymbol{y}^T\mathbf{K} - \boldsymbol{y}^T\frac{n}{n-1}\right)\boldsymbol{y} \qquad \text{(Equation 6)}$$

$$s = tr(\mathbf{KK}) - \frac{n^2}{n-1} \qquad \text{(Equation 7)}$$

And obtain the estimate of the variance components:

$$\hat{\sigma}_{\boldsymbol{K}}^2 = \frac{q}{s} \qquad \text{(Equation 8)}$$

and

$$\hat{\sigma}_e^2 = 1 - \hat{\sigma}_{\mathbf{K}}^2. \qquad \text{(Equation 9)}$$

Standard errors for $\hat{\sigma}_{\boldsymbol{K}}^2$ and $\hat{\sigma}_e^2$ are obtained using the following formulas from (X. Zhou, 2017)

$$V(q) = 2\left(\boldsymbol{y}^T\mathbf{K} - \boldsymbol{y}^T\frac{n}{n-1}\right)(\hat{\sigma}_{\mathbf{K}}^2\mathbf{K} + \hat{\sigma}_e^2 I)\left(\boldsymbol{y}^T\mathbf{K} - \boldsymbol{y}^T\frac{n}{n-1}\right)^T \qquad \text{(Equation 10)}$$

$$V(\hat{\sigma}_{\boldsymbol{K}}^2) = \frac{V(q)}{s^2} \qquad \text{(Equation 11)}$$

$$SE(\hat{\sigma}_{\boldsymbol{K}}^2) = \sqrt{V(\hat{\sigma}_{\boldsymbol{K}}^2)}. \qquad \text{(Equation 12)}$$

Both the point estimates and standard errors are implemented in the GEMMA software package.

Multi-VC model

To fit a model with $k > 1$ variance components,

$$var(\boldsymbol{y}) = \sum_{a=1}^{k} \mathbf{K}_a \sigma^2_{\boldsymbol{K}_a} + I\sigma_e^2 \quad \text{(Equation 13)}$$

The formulas for the point estimates become

$$\boldsymbol{q} = \begin{pmatrix} \boldsymbol{y}^T\mathbf{K}_1 - \boldsymbol{y}^T\frac{n}{n-1} \\ \vdots \\ \boldsymbol{y}^T\mathbf{K}_k - \boldsymbol{y}^T\frac{n}{n-1} \end{pmatrix}\boldsymbol{y} \quad \text{(Equation 14)}$$

$$\mathbf{S} = \begin{pmatrix} tr(\mathbf{K}_1\mathbf{K}_1) & \cdots & tr(\mathbf{K}_1\mathbf{K}_k) \\ \vdots & \ddots & \vdots \\ tr(\mathbf{K}_k\mathbf{K}_1) & \cdots & tr(\mathbf{K}_k\mathbf{K}_k) \end{pmatrix} - \frac{n^2}{n-1} \quad \text{(Equation 15)}$$

$$\hat{\boldsymbol{\sigma}}^2 = \mathbf{S}^{-1}\boldsymbol{q} \quad \text{(Equation 16)}$$

$$\hat{\sigma}_e^2 = 1 - \sum_{a=1}^{k} \hat{\sigma}^2_{\mathbf{K}_a}. \quad \text{(Equation 17)}$$

And the standard errors:

$$V(\boldsymbol{q}) = 2\begin{pmatrix} \boldsymbol{y}^T\mathbf{K}_1 - \boldsymbol{y}^T\frac{n}{n-1} \\ \vdots \\ \boldsymbol{y}^T\mathbf{K}_k - \boldsymbol{y}^T\frac{n}{n-1} \end{pmatrix}\left(\sum_{a=1}^{k} \mathbf{K}_a\hat{\sigma}^2_{\boldsymbol{K}_a} + I\hat{\sigma}_e^2\right)\begin{pmatrix} \boldsymbol{y}^T\mathbf{K}_1 - \boldsymbol{y}^T\frac{n}{n-1} \\ \vdots \\ \boldsymbol{y}^T\mathbf{K}_k - \boldsymbol{y}^T\frac{n}{n-1} \end{pmatrix}^T \quad \text{(Equation 18)}$$

$$V(\hat{\boldsymbol{\sigma}}^2) = \mathbf{S}^{-1}V(\boldsymbol{q})\mathbf{S}^{-1} \quad \text{(Equation 19)}$$

$$SE\left(\hat{\sigma}^2_{\boldsymbol{K}_a}\right) = \sqrt{V\left(\hat{\sigma}^2_{\boldsymbol{K}_a}\right)}, \forall\ a \in \{1, \ldots, k\}. \quad \text{(Equation 20)}$$

Similar to Equation 10, Equation 18 is an asymptotic approximation for $V(\boldsymbol{q})$ and corresponds to an *n*-fold speedup (complexity $O(k^2n^3)$ to $O(k^2n^2)$) for estimating the standard errors compared to using the expected information matrix (X. Zhou, 2017).

2-VC linkage model

The simplest variance-components model for linkage analysis is one with $k = 2$. Specifically, $\mathbf{K}_1$ will be the centered and scaled version of one of either $\widehat{\mathbf{\Phi}}$, $\widehat{\mathbf{P}}$, or $\widehat{\mathbf{C}}_l$ and $\mathbf{K}_2$ the centered and scaled matrix $\widehat{\mathbf{D}}_m$ at a particular marker m. Three possible options include:

$$var(\boldsymbol{y}) = \widetilde{\mathbf{\Phi}}\sigma^2_{\widetilde{\mathbf{\Phi}}} + \widetilde{\mathbf{D}}_m\sigma^2_{\widetilde{\mathbf{D}}_m} + I\sigma^2_e \quad \text{(Equation 21)}$$

$$var(\boldsymbol{y}) = \widetilde{\mathbf{P}}\sigma^2_{\widetilde{\mathbf{P}}} + \widetilde{\mathbf{D}}_m\sigma^2_{\widetilde{\mathbf{D}}_m} + I\sigma^2_e \quad \text{(Equation 22)}$$

$$var(\boldsymbol{y}) = \widetilde{\mathbf{C}}_l\sigma^2_{\widetilde{\mathbf{C}}_l} + \widetilde{\mathbf{D}}_m\sigma^2_{\widetilde{\mathbf{D}}_m} + I\sigma^2_e. \quad \text{(Equation 23)}$$

The tilde over the matrices reflects that these are the centered and scaled versions of these matrices and not the original estimates. The fitting procedure for these models to obtain point estimates and standard errors of the variance components are the same as Equations 14-20 with $k = 2$. This fit must then be repeated at all marker locations m that will be tested for linkage.

After obtaining $\hat{\sigma}^2_{\widetilde{\mathbf{D}}_m}$ and $SE\left(\hat{\sigma}^2_{\widetilde{\mathbf{D}}_m}\right)$, we calculate a one-sided p-value using the inverse of the standard normal cumulative distribution function (CDF), here denoted as F^{-1}, using the following formula

$$F^{-1}\left(\hat{\sigma}^2_{\widetilde{\mathbf{D}}_m} \Big/ SE\left(\hat{\sigma}^2_{\widetilde{\mathbf{D}}_m}\right)\right). \quad \text{(Equation 24)}$$

Logarithm of odds (LOD) is a traditional statistic for the strength of evidence for (or against) genetic linkage (Morton, 1955). For Population Linkage, we define

$$\text{LOD} = \log_{10} L(\hat{\sigma}^2_{\widetilde{\mathbf{D}}_m}) / L(\sigma^2_{\widetilde{\mathbf{D}}_m} = 0), \quad \text{(Equation 25)}$$

and

$$L\left(\hat{\sigma}^2_{\widetilde{\mathbf{D}}_m}\right) = f\left(\hat{\sigma}^2_{\widetilde{\mathbf{D}}_m} \Big/ SE\left(\hat{\sigma}^2_{\widetilde{\mathbf{D}}_m}\right)\right). \quad \text{(Equation 26)}$$

f here denotes the probability density function (PDF) of the standard normal distribution. From this likelihood, we can derive a simple equation for LOD score

$$\text{LOD} = \begin{cases} 0, & \hat{\sigma}^2_{\widetilde{\mathbf{D}}_m} \leq 0 \\ \frac{\left(\hat{\sigma}^2_{\widetilde{\mathbf{D}}_m} \Big/ SE\left(\hat{\sigma}^2_{\widetilde{\mathbf{D}}_m}\right)\right)^2}{2\log(10)}, & \hat{\sigma}^2_{\widetilde{\mathbf{D}}_m} > 0 \end{cases}. \qquad \text{(Equation 27)}$$

We set $\text{LOD} = 0$ when $\hat{\sigma}^2_{\widetilde{\mathbf{D}}_m} \leq 0$ because negative estimates for variance components do not constitute evidence for linkage. We use the established threshold of $\text{LOD} > 3$ for genome-wide significance which approximately corresponds to a one-sided p-value of 10^{-4} (Risch, 1991).

Multi-VC linkage model

In some settings it may be desirable to run a linkage analysis with $k > 2$, that is, with separate variance-components terms for $\widetilde{\mathbf{\Phi}}$, $\widetilde{\mathbf{P}}$, and $\widetilde{\mathbf{C}}_l$ or some subset of these in addition to $\widetilde{\mathbf{D}}_m$. This can help to control for inflation in LOD scores when a single variance component is not sufficient to capture the effects of genome-wide genetic similarity between pairs of individuals. The main tradeoff to this approach is that the computational complexity of calculating $\hat{\sigma}^2_{\widetilde{\mathbf{D}}_m}$ and $SE\left(\hat{\sigma}^2_{\widetilde{\mathbf{D}}_m}\right)$ increases quadratically with k.

One solution to the increased computational complexity is to fit variance components for $\widetilde{\mathbf{\Phi}}$, $\widetilde{\mathbf{P}}$, and $\widetilde{\mathbf{C}}_l$ or a subset of these jointly without $\widetilde{\mathbf{D}}_m$ and then reweight and combine into a single composite matrix

$$\widetilde{\mathbf{K}} = \left(\widetilde{\mathbf{\Phi}}\sigma^2_{\widetilde{\mathbf{\Phi}}} + \widetilde{\mathbf{P}}\sigma^2_{\widetilde{\mathbf{P}}} + \widetilde{\mathbf{C}}_l\sigma^2_{\widetilde{\mathbf{C}}_l}\right)\frac{\mathbf{1}}{\sigma^2_{\widetilde{\mathbf{\Phi}}} + \sigma^2_{\widetilde{\mathbf{P}}} + \sigma^2_{\widetilde{\mathbf{C}}_l}} \qquad \text{(Equation 28)}$$

to run linkage analysis with $\widetilde{\mathbf{D}}_m$. This approach simplifies the linkage analysis because the matrices $\widetilde{\mathbf{\Phi}}$, $\widetilde{\mathbf{P}}$, and $\widetilde{\mathbf{C}}_l$ do not depend on m and their fit relative to one another only needs to be calculated once rather than repeated at all markers m. The rest of the analysis proceeds identically to the

linkage analysis with $k = 2$ described in the previous subsection. Later, we present results on using $\widetilde{\mathbf{K}}$ for linkage analysis compared to separate variance components for $\widetilde{\mathbf{\Phi}}$, $\widetilde{\mathbf{P}}$, and $\tilde{\mathbf{C}}_l$.

Strategy for model selection

The preceding subsection gives several options for running Population Linkage. Users can choose between $k = 2$ variance component models with one of either $\widetilde{\mathbf{\Phi}}$, $\widetilde{\mathbf{P}}$, or $\tilde{\mathbf{C}}_l$ and $\widetilde{\mathbf{D}}_m$ or $k > 2$ models with a combination of $\widetilde{\mathbf{\Phi}}$, $\widetilde{\mathbf{P}}$, or $\tilde{\mathbf{C}}_l$ and $\widetilde{\mathbf{D}}_m$. In addition, the length l for calculating $\tilde{\mathbf{C}}_l$ is variable and there is a choice whether to combine individual matrices $\widetilde{\mathbf{\Phi}}$, $\widetilde{\mathbf{P}}$, or $\tilde{\mathbf{C}}_l$ or a subset of these into a composite matrix $\widetilde{\mathbf{K}}$ for linkage analysis. This subsection gives some strategies for how to weigh these options and decide on a path forward for analysis.

The first step for deciding on a model for linkage analysis is to establish objective criteria for comparing the different models. Here, we outline some specific model criteria for Population Linkage to use rather than general statistics of model fit like AIC, BIC, or deviance. Genomic control (a.k.a. GC-lambda), a common metric for inflated test statistics in GWAS, can also be calculated from the test statistics generated in a genome-wide linkage scan since these follow the same distribution under the null hypothesis (Devlin & Roeder, 1999). In addition, the proportion of variance explained (PVE) by a variance component, if comparable in size to the other components, is a good indication that it is capturing genetic effects not adequately represented by the other components and is helpful for controlling for the confounding between $\widetilde{\mathbf{D}}_m$ and overall additive genetic effects. Visual inspection of a plot of LOD scores can reveal additional problems with an analysis (for example, inflation at the ends of chromosomes) not captured by GC-lambda. Users of Population Linkage should choose the model with GC-lambda nearest 1 and maximum total PVE, but performance on these criteria must be weighed against constraints on runtime and RAM usage that the user faces. The effects of more complex models on computational cost can be

mitigated by combining individual variance-component matrices into a composite matrix $\widetilde{\mathbf{K}}$, but the accuracy of linkage peaks found with this computational convenience should be checked against the model with the individual matrices.

If a study has a large sample size (for example, $n > 10{,}000$) and multiple traits, then it may be helpful to select a subset of traits and individuals for evaluating the choice of model. Individuals (and corresponding rows and columns of $\widetilde{\mathbf{\Phi}}$, $\widetilde{\mathbf{P}}$, $\tilde{\mathbf{C}}_l$, and $\widetilde{\mathbf{D}}_m$) can be randomly sampled to produce a smaller testing set, and traits can be either randomly selected from a list of those available, chosen to represent traits with different genetic architectures, or simulated to ensure there are no real linkage signals. Such a subsetting approach can vastly reduce the amount of time needed to optimize the linkage analysis for a particular cohort before running it on the full data.

The next step is to test all models with $k = 2$ variance components with a linkage analysis. If none of these appear to be satisfactory in terms of its GC-lambda, PVE, and LOD plot, then the user should test all models with $k > 2$ variance components. If $\tilde{\mathbf{C}}_l$ is included in the final model, then the user should run linkage analyses with different values of l and decide on an optimal value. Finally, if a model with $k > 2$ variance components has outperformed the models with $k = 2$ variance components in terms of better GC-lambda, PVE, and LOD plot compared to the models, then the user should test whether combining variance-component matrices produces comparable results.

Linkage analysis and integration with GWAS

After selecting the appropriate model and other parameters for analysis, the next steps are to run the linkage analysis genome-wide across all available traits at the full sample size and integrate the linkage results with GWAS summary statistics on the same traits. The linkage analysis begins by fitting the static variance components $\widetilde{\mathbf{\Phi}}$, $\widetilde{\mathbf{P}}$, and $\tilde{\mathbf{C}}_l$ for each trait to reweight

and combine these into the composite matrix $\widetilde{\mathbf{K}}$ as in Equation 28. We then run the linkage analysis genome-wide by fitting variance components with $\widetilde{\mathbf{K}}$ and $\widetilde{\mathbf{D}}_m$ at different markers m using Haseman-Elston regression as in Equations 13-20. We proceed with linkage analysis in this way because the variance-components model both provides point estimates and standard errors of $\sigma^2_{\mathbf{D}_m}$ for evaluating evidence of linkage while controlling for genome-wide effects on the trait.

After completing the linkage analysis and calculating LOD scores across the genome, we report which loci show evidence of linkage with the trait of interest. Since the region over which LOD is greater than 3 can be large and span many loci tested with many small increases and decreases in LOD score, we report a specific site as a linkage peak if its LOD is greater than 3 and greater than the LOD scores of the two adjacent sites to the right and the two adjacent sites to the left. This definition effectively results in assigning the linkage signal in a region to the marker with the local maximum of LOD score. We do this because the LOD scores at nearby markers are highly correlated and LOD scores greater than 3 at sites near the peak are likely shadows of the stronger signal. This practice of focusing on the top signal in a region is consistent with previous approaches. This peak marker then becomes the focus for follow up analyses.

Once GWAS results have been generated for the same data used in the linkage analysis, it becomes possible to integrate and compare these with the linkage results. We extract all GWAS variants within 5 Mb of each linkage peak and choose the one with the smallest p-value for comparison. This top GWAS SNP gives a finer-resolution picture for the region that is driving the linkage signal and narrows down the list of candidate genes in that region. We chose to focus our search for GWAS variants within 5 Mb of linkage peaks because the 2.5 Mb minimum length to detect IBD segments limited the resolution of linkage peaks.

Computational approach

The previous sections describe how Population Linkage uses variance components estimated by Haseman-Elston regression to perform a genome-wide linkage analysis on population level data. Despite this improvement in scalability over full-likelihood linkage methods based on the Elston-Stewart or the Lander-Green algorithm, Population Linkage must still deal with several large $n \times n$ matrices as input and iterate over a dense marker map that can make the analysis time-consuming and challenging. This section describes several strategies we implemented to remedy these challenges by limiting the number of sites tested, managing the input data, and re-using several terms in the Haseman-Elston fit while calculating variance components across the genome.

Our first strategy for improving the runtime of Population Linkage is to limit the number of sites tested to the unique endpoints of estimated IBD segments in $\hat{S}$. It is redundant to fit the Haseman-Elston regression at two adjacent markers m_1 and m_2 where $\widehat{\mathbf{D}}_{ijm_1} = \widehat{\mathbf{D}}_{ijm_2}$ for every pair i, j. Instead, we only fit Haseman-Elston regression and test for linkage at all the start and end coordinates in the set of $\hat{S}$. In practice there is a considerable amount of overlap in the start and end points of estimated IBD segments, so this can reduce the number of sites tested from the full number of genotyped markers to only those with distinct patterns of IBD sharing.

Even after removing redundant sites from the linkage analysis, there might still be too many IBD segment endpoints to complete a genome-wide analysis in a reasonable amount of time. To further limit the number of tests, we implemented an option in our software to fit variance components and test for linkage at fixed intervals of physical genomic distance across the genome, performing a specified number M' of equally spaced tests. We perform these tests at fixed physical distance as opposed to genetic map distance because several other key parameters, such as the

minimum length of estimated IBD segments were defined in term of physical distance and we wanted to ensure that low-recombination regions would still receive an adequate number of tests.

Our next strategy for improving the runtime of Population Linkage was to manage how our software processes input data. Because of the large size of $\widehat{\mathbf{\Phi}}$, $\widehat{\mathbf{P}}$, $\hat{\mathbf{C}}_l$, $\widetilde{\mathbf{K}}$ and $\widehat{\mathbf{D}}_m$, reading in the files containing this information at each marker being tested is computationally burdensome and time consuming. Since the matrices $\widetilde{\mathbf{\Phi}}$, $\widetilde{\mathbf{P}}$, $\tilde{\mathbf{C}}_l$, and $\widetilde{\mathbf{K}}$ are identical at all markers, whichever of them are being used for linkage analysis can be read once, kept in RAM, and reused to fit the Haseman-Elston regression at all markers. The first $k-1$ terms in the $\boldsymbol{q}$ vector from Equation 14 and the first $k-1$ rows and columns of the $\mathbf{S}$ matrix from Equation 15, and the intermediate calculation $\boldsymbol{y}^T\mathbf{K}_i - \boldsymbol{y}^T\frac{n}{n-1}, i < k$ only depend on $\widetilde{\mathbf{\Phi}}$, $\widetilde{\mathbf{P}}$, $\tilde{\mathbf{C}}_l$ or $\widetilde{\mathbf{K}}$ and not $\widetilde{\mathbf{D}}_m$, so we compute these terms only once, store the results in RAM and reuse them at all markers *m* to avoid multiplying these large matrices repeatedly.

While high-performance computing systems can generally handle keeping the matrices $\widetilde{\mathbf{\Phi}}$, $\widetilde{\mathbf{P}}$, $\tilde{\mathbf{C}}_l$, or $\widetilde{\mathbf{K}}$ in RAM for repeated use, when n is large and there are a large number of IBD segments in $\hat{S}$, it may not be feasible to store all IBD segments for the entire genome in RAM to calculate $\widetilde{\mathbf{D}}_m$ and fit variance-components at every marker *m*. Even on systems that have enough RAM to store $\widetilde{\mathbf{\Phi}}$, $\widetilde{\mathbf{P}}$, $\tilde{\mathbf{C}}_l$, or $\widetilde{\mathbf{K}}$ and a single matrix $\widetilde{\mathbf{D}}_m$ for Haseman-Elston regression, the collection of all estimated IBD segments can be several orders of magnitude larger. To deal with this issue, we only keep a single uncentered and unscaled $\widehat{\mathbf{D}}_m$ in RAM at any one time. We divide $\hat{S}$ into small files by genomic segments and update $\widehat{\mathbf{D}}_m$ with the IBD changes at each position in $\hat{S}$, as described in the next paragraph. This approach results in identical output compared is much faster.

To accomplish this goal of calculating $\widehat{\mathbf{D}}_m$ at every marker with a low memory footprint, we pre-process the file containing all IBD segments in $\hat{S}$ by storing the chromosome, position, IBD

status and sample IDs for all segment endpoints in separate files that each correspond to one megabase (Mb) of genomic distance. This helps because it avoids sorting the entire file containing possibly billions of IBD segments at once. These IBD segment endpoints are stored in a binary format so the segments can be written and read faster. After all the IBD segments in $\hat{S}$ have been processed this way, we read the first file into RAM, sort it, use the IBD states for all pairs of individuals at the first position to construct $\widehat{\mathbf{D}}_1$, center and scale to obtain $\widetilde{\mathbf{D}}_1$, estimate $\sigma^2_{\mathbf{D}_1}$ using Haseman-Elston regression, and perform the first linkage test. We have saved $\widehat{\mathbf{D}}_1$ prior to centering and scaling so at the next position $m = 2$, most entries can be kept in $\widehat{\mathbf{D}}_2$ and only the entries that represent pairs of individuals that begin or end an IBD segment at $m = 2$ need to be updated. At this point, Haseman-Elston can be fit again to test for linkage at this second marker or skipped if the user chose to limit the number of tests. This process of iteratively updating $\widehat{\mathbf{D}}_m$ and fitting (or skipping) Haseman-Elston can then be repeated until all positions in the file have been exhausted. Then, the next file containing the IBD segment endpoints for the next 1 Mb chunk is read in and the same process repeated. This process continues on each file until the entire genome has been iterated over. Handling the estimated IBD segments this way allows calculating the IBD matrix $\widehat{\mathbf{D}}_m$ and performing Haseman-Elston regression at every genomic position where IBD changes without loading all IBD segments into RAM simultaneously.

Implementation

The method we used for kinship and pairwise IBD estimation described in the section Input data preparation were implemented in KING versions 2.1.2 and higher (W. M. Chen, Manichaikul, Nguyen, Onengut-Gumuscu, & Rich, 2017). We ran KING with default settings, the options "kinship" and "ibdseg" invoked, and on up to 46 CPU cores in parallel. The Haseman-Elston regression in Equations 13-20 was implemented in GEMMA 0.96 (X. Zhou, 2017), which we

modified to directly read in the output from KING, incorporate the computational approach described in this paper, and calculate LOD scores for linkage. This modified version of GEMMA 0.96 for Population Linkage is available for download at https://github.com/gjmzajac/GEMMA-population-linkage.

Experimental data: SardiNIA

To test Population Linkage, our approach for limiting the number of tests and whether we could replicate GWAS associations for lipid traits, we used genotypes and phenotype information from the SardiNIA project (Pilia et al., 2006). SardiNIA is an ongoing genotyping and sequencing study of individuals from a population isolate in the Lanusei valley of the Italian island of Sardinia. Two of the distinctive features of this data set that make it useful for our project are the detailed information collected on a wide range of quantitative traits and the high degree of relatedness between individuals in the sample.

The SardiNIA project data we obtained consists of 6,602 samples with genotype data at 18,754,911 variants. All samples were genotyped on 4 commercial genotyping arrays which together had a total of 890,542 variants. A subset of 2,120 of these were also sequenced and used to impute genotypes at an additional 17.6 million variants in the rest of the samples. Because of the effect of genetic isolation from the rest of Europe in the Sardinia population (Chiang et al., 2018) and strong relatedness in this cohort specifically, the imputed genotypes were of much higher accuracy than would be typical with publicly available imputation panels (Pistis et al., 2015; Sidore et al., 2015). In addition to the genotype data, we also obtained LDL, HDL, and total cholesterol and triglycerides measurements that had been regressed on sex, age and age^2 and the residuals inverse-normal transformed by the SardiNIA study team prior to our analysis. The

SardiNIA study team also shared GWAS summary statistics generated by running EMMAX (Kang et al., 2010) with default settings on these same data.

We wanted to test whether our method could replicate GWAS signals at these lipid traits and also evaluate whether limiting the number of tests affected our ability to detect linkage signals. For this, we estimated marker specific IBD status, pairwise kinship, proportion of IBD in 1 Mb chromosome ends, and genome-wide IBD proportion with KING 2.2 in the full set of variants. After testing the performance of different 2-VC models for linkage, we settled on using genome-wide IBD proportion $\sigma^2_{\mathbf{P}}$ and marker-specific IBD status $\sigma^2_{\mathbf{D}_m}$ in Population Linkage, varying the number of tests performed M' with 1,000, 5,000, 10,000, and 20,000 equally spaced tests. We then integrated the linkage results with the GWAS summary statistics we obtained from the SardiNIA study team and compared these in terms of LOD score and PVE of each linkage peak and the p-value and R^2 of the top GWAS variant.

Experimental data: HUNT

To see how well our method could scale to larger cohorts and reveal additional insights into the genetics of lipid traits, we chose to test for linkage in the HUNT study (Krokstad et al., 2013). Because of the multi-generational time scale and high participation rate in this sparsely populated region of Norway, the HUNT cohort contains a very large number of family relationships that can be inferred using the genetic data and used to test for genetic linkage. HUNT does not collect any reported pedigree information. Genetic samples in the HUNT study totaling 69,716 in number were collected over 24 years from three population-based health surveys of all adults in Nord-Trøndelag County in Norway. All samples were genotyped on a version of the IlluminaHumanCore-24 Exome array with custom content (Illumina, 2017) with a total of about 600,000 genetic markers. We obtained genotypes at 359,432 genetic markers after QC and phasing

with SHAPEIT2 (Delaneau, Zagury, & Marchini, 2013). In addition to the genotype data, we also obtained imputed dosages at 45,453,131 autosomal variants from the Haplotype Reference Consortium as described in (Nielsen et al., 2018) in order to run GWAS analyses.

The first step is to prepare the input data for Population Linkage. For this we estimated IBD and kinship with the 359,432 genotyped variants we obtained from HUNT using KING 2.1.3 and used these to derive the estimated genome-wide IBD proportion, $\widehat{\mathbf{P}}$, and the IBD proportion at the ends of chromosomes, $\widehat{\mathbf{C}}_l$. For all HUNT participants with genetic data we also obtained LDL, HDL, and total cholesterol and triglycerides measurements for our linkage analysis. We also obtained body-mass-index (BMI) measurements as a control phenotype to help with model selection. To analyze all these phenotypes in HUNT in a systematic fashion, we regressed each phenotype on genetic principal components 1-4, genotyping batch, age at time of measurement, and sex. We ran our linkage analysis and GWAS on the inverse-normalized residuals of these phenotypes with the covariates regressed out.

After preparing the input data, we proceeded to select the optimal model for linkage analysis as described in the section, Strategy for model selection. To do this, we randomly sampled 25,000 of our HUNT participants to evaluate the choice of different variance components more quickly than with the full data. Any linkage analyses run on this subset for model selection purposes were also limited to $M' = 1{,}000$ equally spaced tests to further improve computational time. To help with this evaluation, we decided to run a null simulation with no true linkage signals. Since an independent, identically distributed phenotype would not have satisfactorily captured the correlation in any trait from analyzing related individuals, we simulated a correlated phenotype for our cohort with covariance based on the genotype relatedness matrix (GRM) described in Equation 2. To generate the GRM, we used GEMMA 0.96 on the complete set of phased genotype data. To

simulate the correlated phenotype, we performed Cholesky decomposition $\mathbf{A} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$ of our GRM $\mathbf{A}$ and multiplied $((0.25)I + (0.75)\mathbf{L})$ by a vector $\boldsymbol{x}$ of independent, identically distributed values randomly sampled from the standard normal distribution. Using LDL and BMI measurements and these simulated phenotype values across our random subset of 25,000 individuals, we tested using estimated kinship $\widetilde{\boldsymbol{\Phi}}$, genome-wide IBD proportion $\widetilde{\mathbf{P}}$, and proportion of chromosome ends that are IBD $\tilde{\mathbf{C}}_l$ individually and all combinations of these along with IBD estimates $\widetilde{\mathbf{D}}_m$ in a variance-components model for linkage analysis. We evaluated these based on the proportion of variance explained (PVE) by the variance components and genomic-control (GC) lambda values from linkage analysis. After choosing a model, we also tested different lengths of chromosome ends l between 50 kb and 5 Mb for calculating $\tilde{\mathbf{C}}_l$. We also tested the performance of combining $\widetilde{\boldsymbol{\Phi}}$, $\widetilde{\mathbf{P}}$, and $\tilde{\mathbf{C}}_l$ into a single composite matrix $\widetilde{\mathbf{K}}$ to run linkage analysis with $\widetilde{\mathbf{D}}_m$ compared to using separate variance components for all of them.

After running the preceding tests for model selection on the subset with 25,000 individuals, we decided to use all available variance components $\widetilde{\boldsymbol{\Phi}}$, $\widetilde{\mathbf{P}}$, and $\tilde{\mathbf{C}}_l$ with l = 500 kb, reweighted and combined into a single matrix for the linkage analysis of the four lipid traits in the full HUNT cohort. Similar to during model selection, we limited the linkage analysis to $M' = 1{,}000$ equally spaced tests so that computation could complete in less than two weeks on a single CPU core for each phenotype. We also compiled our modified version of GEMMA v0.96 to store all values at float instead of double precision to reduce the memory footprint of the analysis and produce nearly identical results.

After generating the linkage results and determining significant loci for each trait (HDL, LDL, and total cholesterol and triglycerides), we compared these with GWAS results to validate our linkage signals and investigate whether linkage would be advantageous to single-variant tests

for some sites. We performed two GWAS in the HUNT data for each of same four lipid traits as in the linkage analysis for this comparison, one using the 359,432 genotyped variants, and one with 8.6 million imputed variants that remained after filtering out variants with imputation info $R^2 < 0.3$ and minor allele frequency (MAF) < 1%. For both of these we used the SAIGE package (W. Zhou et al., 2018) to perform single-variant tests while controlling for the relatedness in HUNT. In addition to these GWAS in the HUNT data, we also obtained summary statistics from the Global Lipids Genetics Consortium (GLGC) meta-analysis of these four lipid traits in 1.6 million individuals of European ancestry across 75 million variants. For each of our significant linkage peaks, we extracted the variant with the smallest p-value for the same trait within 5Mb from each of these three GWAS data sets as described in the section, Linkage analysis and integration with GWAS. From there, we were able to examine whether each linkage peak was replicated at genome-wide significance ($p < 5\times10^{-8}$) in the HUNT genotyped GWAS, the HUNT imputed GWAS, and the GLGC meta-analysis.

Results

SardiNIA

We examined the overall structure of relatedness and IBD sharing in SardiNIA. Using KING 2.2, we estimated 44,006 relative pairs of 3rd degree or closer and a total of 316,729,132 IBD segments in the set. 97% of individuals had at least one relative of 3rd degree or closer and all but 7 individuals shared at least one IBD segment with another individual. Table 4-1 summarizes the number of relative pairs, total and average number of IBD segments, and total and average length of IBD segments for each relationship type. This shows that the vast majority of estimated IBD segments (98.9%) were in more distant relatives of 3rd degree or greater. These

IBD segments estimated in distant relatives are potentially informative for inferring evidence of linkage but would have been excluded in a linkage analysis that split the SardiNIA cohort into smaller pedigrees to test for linkage using classical methods (Liu, Kirichenko, Axenovich, van Duijn, & Aulchenko, 2008).

We next proceeded to test different variance components models to choose one for linkage analysis with the SardiNIA data. KING 2.2 estimated kinship, genome-wide IBD proportion, and 3,316,301 unique IBD-segment endpoints for all pairs of individuals. We then fit single-variance component models using estimated kinship, genome-wide IBD proportion, and proportion of 1 Mb chromosome ends that were IBD for each of the four lipid traits: high-density lipoprotein (HDL), low-density lipoprotein (LDL), and total cholesterol (TC) and triglycerides (TG) to calculate the proportion of variance explained (PVE) and then ran linkage analyses at 1,000 equally spaced sites with each of these same variance components to calculate genomic-control (GC) lambda values. The full results of these tests are reported in Table 4-2. Using the estimated kinship matrix resulted in the highest PVE for each trait (HDL 40.1%, LDL 32.4%, TC 38.7%, TG 26.3%), but GC lambda values were high for some traits (HDL 2.7, LDL 1.7, TC 1.9, TG 0.9). However, when we used the genome-wide IBD proportion matrix, GC lambda values were lower (HDL 1.2, LDL 1.0, TC 1.1, TG 0.9). These GC lambda values for the IBD proportion matrix were reasonable and visual inspection of the LOD plots (for example, see Figure 4-2) did not reveal any problems, so we did not proceed to test models with additional variance components. We used the genome-wide IBD proportion along with marker-specific IBD status to run linkage analysis of the four lipid traits at larger numbers of equally spaced sites and determine significant linkage peaks for SardiNIA.

Having selected a model for linkage analysis and calculated LOD scores, we checked the results for evidence of linkage in any of these 4 lipid traits. We had two significant peaks, one for

the trait LDL and one for total cholesterol, both on chromosome 19 near the gene *APOE* (LDL LOD 3.9, PVE 5.0%; TC LOD 3.5, PVE 4.7%). These linkage peaks are reported in Table 4-3. The SardiNIA study in 2015 associated the missense variants rs7412 and rs429358 in the gene *APOE* with variation in LDL and total cholesterol and rs429358 with levels of high-sensitivity C-reactive protein (Sidore et al., 2015). rs429358 in particular is well-known as a variant that disrupts *APOE* function in lipid transport and metabolism, influencing risk for Alzheimer's disease, macular degeneration, and other traits (Jiang et al., 2008; Liutkeviciene et al., 2018). The 2015 SardiNIA paper showed that rs7412 and rs429358 were independent signals for LDL and total cholesterol with R^2 values of 2.4% and 0.8% for LDL and 1.7% and 0.5% for total cholesterol. Our linkage analysis estimated that IBD sharing in the *APOE* locus explained 5.0% of the variance in LDL measurements and 4.7% for total cholesterol, higher than the R^2 values for rs7412 and rs429358 found in the 2015 GWAS. This implies that our linkage test is able to capture the effects of both these variants and additional genetic variation in the region (for example in the genes *APOC1* or *APOC2* (Jong, Hofker, & Havekes, 1999)) that influence LDL and total cholesterol levels.

After we had identified these significant linkage peaks for LDL and total cholesterol near the gene *APOE*, we wanted to know how our ability to detect this linkage signal was impacted by our choice of the number of equally spaced genetic markers across the genome at which we test for linkage. To evaluate this, we report the top linkage signal for LDL and total cholesterol that we observed when running Population Linkage at 1,000, 5,000, 10,000, and 20,000 equally spaced genetic markers in Table 4-3. These results show that our method is able to detect evidence for linkage with LOD > 3 for LDL and total cholesterol whether 1,000, 5,000, 10,000, or 20,000 equally spaced genetic markers are tested (LDL LOD: 3.54, 3.74, 3.74, and 3.88; total cholesterol

LOD: 3.05, 3.46, 3.49, 3.54). These results show that while the number of sites tested does impact the largest observed LOD score in a region, running Population Linkage at 1,000 markers in order to improve runtime is sufficient to detect the strongest linkage signals present in a cohort.

We next wanted to compare the performance of Population Linkage to GWAS in lipid genes beyond *APOE*. In Figure 4-2, we illustrate the overlap between GWAS hits and LOD scores from our linkage analysis for LDL. As previously mentioned, our linkage analysis successfully replicates the strongest GWAS signal in the gene *APOE* and also shows elevated LOD scores near other GWAS peaks in *HBB* and *PCSK9*. This result indicates that Population Linkage does capture some of the signal in known lipid genes beyond *APOE* and that rerunning this analysis with a larger sample size might be able to yield additional linkage peaks with LOD > 3 near these genes.

HUNT

Our first step for analyzing the HUNT cohort was to examine the structure of its relatedness and IBD sharing. KING 2.1.3 estimated 6,867,367,662 IBD segments in 341,100,522 pairs of individuals that shared at least one IBD segment. This resulted in a total of 279,100 IBD segment endpoints. Table 4-4 gives a summary of the number of pairs of parent-offspring, full sibling, 2nd degree, 3rd degree and more distant relationships, and the average and total number and length of estimated IBD segments in these relationship types. The vast majority (99.6%) of estimated IBD segments were in more distant relatives of 3rd degree or greater. These IBD segments estimated in distant relatives are potentially informative for inferring evidence of linkage but would have been excluded in a linkage analysis that split the HUNT cohort into smaller pedigrees to test for linkage using classical methods.

We then proceeded to select the optimal variance-components model to run Population Linkage in HUNT. A preliminary analysis using a 2-variance-component model with genome-

wide IBD proportion and marker-specific IBD sharing like we used in SardiNIA revealed severe inflation in the test statistics, particularly at the ends of chromosomes, even with the null simulated phenotype (Figure 4-3, top). Further investigation revealed that IBD estimates were significantly biased toward close relatives near the ends of chromosomes relative to the middle (Figure 4-3, bottom). The closeness of relatives IBD at a locus was almost perfectly correlated with the proportion of variance explained (PVE) by IBD estimated at that locus (Pearson's correlation 0.99, Figure 4-4), illustrating the effect of this confounding.

After this preliminary analysis, we began a systematic evaluation of all potential models for Population Linkage described in the methods section of this paper using our randomly sampled subset of 25,000 samples. We began by fitting low-density lipoprotein cholesterol (LDL), body mass index (BMI), and our simulated trait with single-variance-component models of kinship, pairwise IBD proportion, and average IBD in chromosome ends to assess their relative ability to explain the correlation structure in these traits. We also ran linkage analyses with the same traits and each of the same individual variance components together with the matrix of marker-specific IBD sharing, $\widetilde{\mathbf{D}}_m$ to calculate genomic-control (GC) lambda values. These analyses revealed that while the chromosome ends had the largest PVE (50.2% LDL, 37.0% BMI, and 68.8% Simulation), using the pairwise IBD proportion for linkage resulted in the lowest GC lambda values for LDL and BMI (2.73 and 2.33) when used for linkage and kinship resulted in the best GC lambda for the simulation (3.62, full results in Table 4-5). However, all of these 2-variance-component models had inflated test statistics as shown by these observed GC lambda values much greater than 1.

Since none of the 2 variance component models for linkage could adequately control for inflation, we ran all combinations of estimated kinship, genome-wide IBD proportion, and

proportion of IBD in chromosome ends both with and without marker-specific IBD sharing $\widetilde{\mathbf{D}}_m$ to test if one of these models would mitigate the issue. The results were that running kinship, chromosome ends, and Dm resulted in the highest PVE (50.3% LDL, 37.0% BMI, 68.9% Simulation), while kinship, pairwise IBD proportion, and chromosome ends resulted in the smallest GC lambda (1.06 LDL, 1.15 BMI, 0.91 Simulation). The full results for this analysis are in Table 4-6. Because this model with all variance components appeared to control for inflation best in this subset of 25,000 individuals, we decided to use it moving forward.

After deciding on the appropriate variance components, it was of interest to determine if reweighting estimated kinship, IBD proportion, and IBD in chromosome ends and combining into a single composite matrix $\widetilde{\mathbf{K}}$ for linkage analysis as described in the section Multi-VC linkage model would have any drastic effect on the LOD scores calculated by our method. In our subset with 25,000 individuals, we found that fitting the three static VCs once for a phenotype and recombining into a single $\widetilde{\mathbf{K}}$ matrix based on the estimated VC weights to fit with IBD at all sites resulted in z-scores that were almost perfectly correlated with those from the multi-vc fit ($r > 0.999$ for LDL, BMI, and simulation) and LOD scores were, on average, 0.0103, 0.0098, and 0.0124 lower for LDL, BMI, and the simulation, respectively. Based on these results we concluded that using the combined $\widetilde{\mathbf{K}}$ matrix would result in nearly identical results at a negligible cost in statistical power but at a greater than two-fold savings in computational time and RAM for the linkage analysis.

It was also of interest to know whether the length l extracted from the ends of each chromosome to calculate the matrix of IBD sharing at the ends of chromosomes $\widetilde{\mathbf{C}}_l$ would impact the results. The optimal lengths in terms of GC-lambda were different for LDL (0.3 Mb with GC lambda 1.0), BMI (0.4 Mb with GC lambda 1.1), and the simulated phenotype (0.2 Mb with GC

lambda 0.9), but overall were very similar for any length less than 1 Mb (Table 4-7). We chose to proceed with extracting a length of 0.5 Mb as this seemed to be a conservative choice yet still near the optimal values for each trait.

Once we had decided on this model with estimated kinship, genome-wide IBD proportion and IBD sharing at the ends of chromosomes, we proceeded to run linkage analysis on the 4 lipid traits (high-density lipoprotein (HDL), LDL, and total cholesterol and triglycerides) at the full sample size. The sample sizes of our traits ranged from 67,429 for LDL measurements to 69,479 for triglycerides levels. Running each trait on one CPU core used an average of 133 GB RAM over 11 hours, 35 minutes for reweighting and combining matrices and 81 GB RAM over 11 days, 20 hours for linkage analysis. We observed a total of 25 significant linkage peaks with LOD > 3 across 19 distinct loci for the four traits. HDL had 7 significant linkage peaks, LDL 9, total cholesterol 7, and triglycerides 2. All these peaks and supporting GWAS evidence for them are reported in Table 4-8.

Our strongest signals, both in terms of LOD score and proportion of variance explained (PVE), were between the trait LDL and the region of chromosome 19 near the gene *APOE* (LOD 29.3, PVE 4.0%) and HDL and the region of chromosome 16 near the gene *CETP* (LOD 30.2, PVE 4.3%). These peaks are shown in the LOD plots for HDL and LDL in Figure 4-5 and Figure 4-6. *APOE* and *CETP* are well known genes for lipid regulation (Freeman & Remaley, 2016) and were also supported in the HUNT GWAS of genotyped variants, imputed variants, and the GLGC meta-analysis (Table 4-8). Multiple genetic variants in *APOE* have been associated with differences in LDL, in particular the relatively common missense variants rs7412 and rs429358 which correspond to the *APOE* ε2 and ε4 alleles and which together are estimated to explain between 3.2% and 4.9% of variance in LDL measurements (Burman et al., 2009; Chasman,

Kozlowski, Zee, Kwiatkowski, & Ridker, 2006). The *CETP* locus is also an example of allelic heterogeneity since multiple variants have been shown to be independently associated with differences in HDL levels, including upstream variant rs183130 and missense variant rs5880 (Spirin et al., 2007). These results show that our method can capture the effects of multiple variants in a region that contribute to a trait.

In addition to *CETP* and *APOE*, the majority of our significant linkage peaks were in other established lipid loci that were easily replicated by our GWAS of genotyped and imputed HUNT variants and the GLGC meta-analysis. These peaks were *PCSK9* (LDL LOD 5.9, TC LOD 4.6), *CELSR2* (LDL LOD 5.0, TC LOD 3.5), *APOB* (LDL LOD 6.6, TC LOD 6.3), *GCKR* (TG LOD 4.0), *ABCG8* (LDL LOD 5.6, TC LOD 5.4), *LPA* (LDL LOD 3.8), *ABCA1* (HDL LOD 7.0), *ZNF259* (HDL LOD 4.1, TG LOD 10.3), *SCARB1* (HDL LOD 3.0), *ALDH1A2* (HDL LOD 9.1), *LDLR* (LDL LOD 15.7, TC LOD 10.8). LOD plots showing these peaks are shown in Figure 4-5, Figure 4-6, Figure 4-7, and Figure 4-8. Table 4-8 contains these peaks and p-values from the GWAS. This shows that results from Population Linkage are in line with expectations and are able to highlight the most important lipid genes.

In addition to confirming known lipid genes, it was of interest to know where Population Linkage could provide additional insights beyond GWAS. There were 5 peaks with LOD > 3 which were not replicated at genome-wide significance in the HUNT GWAS of 359,432 genotyped variants. All of them were replicated in either the HUNT GWAS of imputed variants or the GLGC meta-analysis. The first of these peaks was for the trait HDL and the region on chromosome 16 near the gene *GPR139* (LOD 3.1, GLGC p-value 8.3×10^{-13}). The locus near *GPR139* and nearby gene *GPRC5B* had previously been associated with differences in BMI (Pulit et al., 2019) and HDL (Tekola-Ayele, Lee, Workalemahu, & Sánchez-Pozos, 2019). The next peak

for HDL was near the gene *CDH1* (LOD 3.4, imputed p-value 1.5×10^{-42}, GLGC p-value 9.2×10^{-46}). While *CDH1* is better known as a protein that helps form epithelial tissues and as a tumor suppressor gene, the top SNP rs571298027 is also near (11 kb) to the *TANGO6* gene that has been associated with differences in HDL (More et al., 2007; Richardson et al., 2020). The peaks for LDL were near the genes *LLGL1* (LOD 4.2, GLGC p-value 6.4×10^{-13}) and *CEBPA* (LOD 6.0, imputed p-value 3.6×10^{-9}, GLGC p-value 3.3×10^{-12}). *LLGL1* has been previously associated with LDL levels (Klarin et al., 2018), and *CEBPG* is a transcription factor involved in adipogenesis and is part of the *PEPD-CEBPA-CEBPG* locus that has been associated with waist-to-hip ratio (Lotta et al., 2018) and lipid levels (Freeman & Remaley, 2016). The last peak was for total cholesterol and a region containing AC005307.3 and several other non-coding genes (LOD 4.1, imputed p-value 6.8×10^{-9}). AC005307.3 is a pseudogene of *SHCBP1* (Carithers & Moore, 2015) whose function has not been extensively described. Since these 5 peaks did not have a genome-wide significant variant among single-variant tests of 359,432 genotyped variants that were used in the linkage analysis but were confirmed in either the HUNT GWAS of imputed variants or the GLGC meta-analysis, these results confirm that Population Linkage is able to detect linkage signals in ungenotyped variants.

Discussion

We have demonstrated the feasibility of genome-wide linkage analysis on 10,000s of individuals with 100,000s of markers with our method and that this approach is able to replicate known associations in lipid traits. Sample relatedness is a nearly unavoidable situation in population or case-control cohorts of GWAS-scale, prompting the development of an entire class of methods for linear-mixed-model GWAS to correct and control for the effect of sample relatedness (Kang et al., 2010; Kang et al., 2008; W. Zhou et al., 2018; X. Zhou & Stephens, 2012).

In contrast, our method provides the opportunity to use this relatedness to map traits to genetic loci, often with greater variance explained in a region than the top associated SNP in GWAS. Additionally, because Haseman-Elston regression had been successfully extended to binary traits with the phenotype-correlation-genotype-correlation (PCGC) approach (Golan, Lander, & Rosset, 2014), we are hopeful that our method can be similarly adapted to perform linkage on binary traits with thousands of cases and controls.

One limitation of the results we have presented here is that several of our peaks were near our threshold of LOD 3.0 (approximately a one-sided p-value of 10^{-4}) and would no longer be significant after adjusting for testing for linkage in 4 traits. Any multiple-testing adjustment would have to account for the autocorrelation of the test statistics and that many of the separate traits, like LDL and total cholesterol, were not truly independent. Even so, we feel confident the results presented here do not contain a large number of false positives since all 27 peaks we observed with LOD > 3.0 across both HUNT and SardiNIA had a supporting GWAS SNP within 5 Mb.

One of the purported benefits of linkage analysis over GWAS is the ability to test for linkage in untyped genetic variation, as long as IBD segments in the region can be identified. Our results for high-density lipoprotein (HDL), LDL, and total cholesterol from the HUNT study support this claim since we observed 5 linkage signals with LOD > 3 where the evidence for association from the GWAS of genotyped variants was not genome-wide significant but GWAS with additional variants was genome-wide significant. In addition, our results for linkage in the *APOE* region in the SardiNIA study show how a test for linkage, even at a single marker, is able to capture the effects of multiple variants in the region that influence LDL levels. Since the inclusion of an individual SNP has little effect on the IBD estimates calculated in a region, this

feature of linkage analysis would extend to capturing the effects of ungenotyped variants that influence the trait as well.

Since this work only considers applying Population Linkage to genotyping array data, a natural question that arises is how the analysis would differ in a sequencing study. Some of the advantages of working with whole-genome sequencing data include having a denser set of genotypes with which to estimate IBD sharing. This can potentially improve the resolution of IBD segment endpoints and help to estimate shorter IBD segments in more distantly related individuals, which can improve statistical power. Sequencing will also result in having more variants to use for following up on and interpreting significant linkage signals. One tradeoff to using sequencing data is the higher error rate compared to genotyping arrays, which could also impact IBD segment estimation. Possible remedies for this include using an algorithm for IBS segment estimation that is more robust to genotyping errors than the one used here, or one that can determine IBD segments based on genotype probabilities rather than hard genotype calls.

One drawback of using Haseman-Elston regression to estimate variance components, and of methods of moments generators is general, is a loss in statistical efficiency that can affect the accuracy of variance components estimates and power of tests for linkage. While these are legitimate concerns for a method like Population Linkage, our use of Haseman-Elston regression as presented here enables linkage analysis with a number of genetic relationships several orders of magnitude larger than was previously possible with the classical methods. Such a tradeoff between statistical efficiency and the ability to analyze a greater sample size exists in many application areas, and there are many examples of where the benefits of using more data outweigh the statistical concerns.

This work opens up new possibilities for how linkage analysis might be further applied and developed in the future. Scaling linkage analysis up by another order of magnitude than what we presented here, for example to the UK Biobank (Bycroft et al., 2018), will require additional computational simplifications, for example by subdividing the data into smaller groups and meta-analyzing, or reworking the fitting of variance-components to avoid multiplying over all possible pairs of individuals. In addition, recent innovations in finding shorter IBD segments in more distant, apparently unrelated, individuals (Delaneau, Zagury, Robinson, Marchini, & Dermitzakis, 2019) has the potential to increase power and resolution if applied to linkage analysis. Most importantly, a seamless integration of variance components linkage analysis and GWAS into a mixed effects model can together increase the power and utility of both. Such an approach can not only combine the signal from both linkage and association for a more powerful test to find additional novel associations, but also opens up the possibility of testing a variant for association conditioned on the genetic background of individuals in a region. This can help to fine map causal variants in a region that shows evidence for association and further our understanding of how they impact our biology and risk for disease.

Acknowledgements

G.R.A. was supported by HG007022. The authors thank Sarah Gagliano Taliun, Francesco Cucca, David Schlessinger, and Carlo Sidore for use of the SardiNIA data; and Cristen Willer, Kristian Hveem, Sarah Graham, Bjorn Olav Asvold, Ben Brumpton, Anne Heidi, Jonas Billie Nielsen, Wei Zhou, Lars Fritsche, and Maiken Gabrielsen for use of data from the HUNT study. We also acknowledge Xiang Zhou for support with using GEMMA, and Wei-Min Chen for support with using KING and for fixing bugs the authors had found in the course of preparing this paper.

Tables

Table 4-1 Relatedness and IBD sharing statistics in SardiNIA

Degree Relationship	N pairs	Tot. N IBD Segments	Avg. N IBD Segments per Pair	Tot. IBD Segment Length (Tb)	Avg. IBD Segment Length (Mb)
MZ Twins	16	1,455	91	0.1	59
Parent – Child	4,655	255,672	55	13	49
Full Siblings	5,442	906,470	167	15	16
2nd Degree	12,262	959,258	78	18	18
3rd Degree	21,631	1,386,018	64	17	12
> 3rd Degree	21,745,895	313,220,259	14	1,453	4.6

The number of pairs, total number of IBD segments estimated, average number of IBD segments per pair, total length of IBD segments estimated, and average length of IBD segments per pair by relationship type in SardiNIA, as estimated by KING 2.2.

Table 4-2 Choice of fitting single variance components for linkage in SardiNIA

Pheno	N	VC	PVE (%)	GC Lambda
HDL	5,942	Kinship	40.1	2.74
HDL	5,942	IBD prop	36.7	1.16
HDL	5,942	Chr ends	38.5	4.44
LDL	5,937	Kinship	32.4	1.69
LDL	5,937	IBD prop	27.7	1.00
LDL	5,937	Chr ends	30.5	2.96
Total Cholesterol	5,937	Kinship	38.7	1.87
Total Cholesterol	5,937	IBD prop	33.7	1.10
Total Cholesterol	5,937	Chr ends	37.1	3.31
Triglycerides	5,905	Kinship	26.3	0.94
Triglycerides	5,905	IBD prop	20.8	0.91
Triglycerides	5,905	Chr ends	22.9	2.37

A table comparing the impact of different choices of single variance components on the proportion of variance explained (PVE) and genomic-control lambda (GC Lambda) in a linkage analysis of the phenotypes high-density lipoprotein (HDL), low-density lipoprotein (LDL), and total cholesterol and triglycerides measurements. “IBD prop” refers to the proportion of IBD shared genome-wide and “Chr ends” refers to average IBD sharing in the first and last Mb of each chromosome.

Table 4-3 Linkage peaks in SardiNIA (19M SNPs) at different numbers of markers tested

Sites Tested	Trait	Gene	Chr	Pos	PVE (%)	LOD
1,000	LDL	*APOE*	19	47,471,344	4.8	3.54
	Total Cholesterol	*APOE*	19	47,471,344	4.3	3.05
5,000	LDL	*APOE*	19	48,030,170	4.9	3.74
	Total Cholesterol	*APOE*	19	48,030,170	4.7	3.46
10,000	LDL	*APOE*	19	48,030,170	4.9	3.74
	Total Cholesterol	*APOE*	19	48,287,640	4.8	3.49
20,000	LDL	*APOE*	19	47,880,669	5.0	3.88
	Total Cholesterol	*APOE*	19	47,880,669	4.7	3.54

Table of all linkage peaks with LOD > 3 in SardiNIA. The proportion of variance explained (PVE) and LOD for linkage, are reported at 1,000, 5,000, 10,000, and 20,000 equally spaced linkage tests. IBD estimates are from KING 2.2.

Table 4-4 Relatedness and IBD sharing statistics in HUNT

Degree Relationship	N pairs	Tot. N IBD Segments	Avg. N IBD Segments	Tot. IBD Segment Length (Tb)	Avg. IBD Segment Length (Mb)
Parent – Child	47,113	2,042,502	43	125	61
Full Siblings	35,888	4,058,463	113	72	18
2nd Degree	117,478	7,524,029	64	162	22
3rd Degree	251,385	13,031,221	52	187	14
> 3rd Degree	2,429,673,606	6,833,336,344	2.8	3,192	0.5

The number of pairs, total number of IBD segments estimated, average number of IBD segments per pair, total length of IBD segments estimated, and average length of IBD segments per pair by relationship type in HUNT, as estimated by KING 2.1.3.

Table 4-5 Choice of fitting single variance components for linkage in HUNT

Pheno	VC	PVE (%)	GC Lambda
BMI	Kinship	24.0	8.29
BMI	IBD prop	17.3	2.33
BMI	Chr ends	37.0	5.95
LDL	Kinship	30.0	20.29
LDL	IBD prop	26.0	2.73
LDL	Chr ends	50.2	12.61
Simulation	Kinship	62.8	3.62
Simulation	IBD prop	32.9	3.74
Simulation	Chr ends	68.8	15.83

A table comparing the impact of different choices of single variance components on the proportion of variance explained (PVE) and genomic-control lambda (GC Lambda) in a linkage analysis of the phenotypes LDL, BMI, and the simulation. All results are from the random n=25,000 subset. "IBD prop" refers to the proportion of IBD shared genome-wide and "Chr ends" refers to average IBD sharing in the first and last Mb of each chromosome.

Table 4-6 Choice of fitting multiple variance components for linkage

Pheno	VCs	Kinship PVE (%)	IBD prop PVE (%)	Chr ends PVE (%)	Total PVE (%)	GC
BMI	Kinship, IBD prop	20.0	8.1		28.1	1.49
BMI	Kinship, Chr ends	18.6		18.4	37.0	2.70
BMI	IBD prop, Chr ends		8.7	26.8	35.5	1.34
BMI	Kinship, IBD prop, Chr ends	17.7	4.7	13.8	36.2	1.15
LDL	Kinship, IBD prop	22.3	15.7		38.0	1.54
LDL	Kinship, Chr ends	21.7		28.5	50.3	6.56
LDL	IBD prop, Chr ends		15.8	31.6	47.4	1.29
LDL	Kinship, IBD prop, Chr ends	19.3	11.5	17.4	48.2	1.06
Sim	Kinship, IBD prop	60.4	5.0		65.4	0.99
Sim	Kinship, Chr ends	60.3		8.6	68.9	1.95
Sim	IBD prop, Chr ends		17.2	48.7	65.8	1.41
Sim	Kinship, IBD prop, Chr ends	59.6	3.8	4.9	68.3	0.91

A table comparing the impact of different combinations of variance components on the proportion of variance explained (PVE) and genomic-control lambda (GC) in a linkage analysis of the phenotypes LDL, BMI, and the null simulation. All results are from the random n=25,000 subset. "IBD prop" refers to the proportion of IBD shared genome-wide and "Chr ends" refers to average IBD sharing in the first and last Mb of each chromosome.

Table 4-7 Choice of length of chromosome ends extracted

Mb Extracted	LDL PVE (%)	LDL GC	BMI PVE (%)	BMI GC	Sim PVE (%)	Sim GC
0.05	48.62	1.0477	36.47	1.1605	68.84	0.8991
0.1	48.67	1.0476	36.49	1.1573	68.80	0.8976
0.2	48.66	1.0478	36.45	1.1607	68.75	0.8958
0.3	48.58	1.0427	36.61	1.1255	68.68	0.8993
0.4	48.61	1.0432	36.66	1.1194	68.46	0.9090
0.5	48.35	1.0523	36.37	1.1306	68.34	0.9145
2	47.78	1.0536	35.87	1.1476	67.80	0.9221
2	46.54	1.0750	34.42	1.1861	67.48	0.9191
3	44.90	1.1712	33.29	1.2260	66.05	0.9760
4	43.48	1.1614	32.46	1.2356	65.52	0.9878
5	42.51	1.1670	31.52	1.2584	64.87	1.0380

A table comparing the impact of different choices of chromosome length extracted on the proportion of variance explained (PVE) by kinship, proportion of IBD shared, and average IBD sharing at the ends of chromosomes fit as a combined matrix and genomic-control lambda (GC) in a linkage analysis of the phenotypes LDL, BMI, and the null simulation. All results are from the random n=25,000 subset.

Table 4-8 HUNT Linkage peaks

Trait	N	Chr:Mb	PVE IBD (%)	LOD	Geno p-value	Imputed p-value	GLGC p-value	Top SNP rsid	Top SNP Gene
HDL	69,214	9:110	1.18	7.0	3.9×10^{-61}	3.7×10^{-61}	1.3×10^{-420}	rs2740488	*ABCA1*
HDL	69,214	11:118	0.60	4.1	4.8×10^{-30}	4.8×10^{-30}	1.4×10^{-637}	rs964184	*ZNF259*
HDL	69,214	12:126	0.89	3.0	1.9×10^{-18}	3.0×10^{-30}	5.6×10^{-188}	rs61941677	*SCARB1*
HDL	69,214	15:61	1.20	9.1	4.1×10^{-66}	6.1×10^{-67}	2.3×10^{-1162}	rs261290	*ALDH1A2*
HDL	69,214	16:25	0.82	3.1	1.7×10^{-7}	1.1×10^{-7}	8.3×10^{-13}	rs9938120	*GPR139*
HDL	69,214	16:58	4.26	30.2	1.6×10^{-225}	1.1×10^{-234}	2.2×10^{-5270}	rs183130	*CETP*
HDL	69,214	16:74	0.54	3.4	6.4×10^{-6}	1.5×10^{-42}	9.2×10^{-46}	rs571298027	*CDH1*
LDL	67,429	1:57	0.85	5.9	5.3×10^{-72}	2.8×10^{-70}	5.2×10^{-1390}	rs11591147	*PCSK9*
LDL	67,429	1:110	0.81	5.0	1.5×10^{-46}	1.2×10^{-46}	4.7×10^{-1726}	rs12740374	*CELSR2*
LDL	67,429	2:22	0.59	6.6	2.3×10^{-51}	3.0×10^{-48}	1.3×10^{-927}	rs934197	*APOB*
LDL	67,429	2:43	0.79	5.6	6.0×10^{-35}	5.1×10^{-35}	1.4×10^{-470}	rs4299376	*ABCG8*
LDL	67,429	6:162	0.89	3.8	9.8×10^{-13}	5.3×10^{-28}	5.5×10^{-377}	rs10455872	*LPA*
LDL	67,429	17:15	0.55	4.2	3.4×10^{-4}	9.4×10^{-8}	6.4×10^{-13}	rs28811342	*LLGL1*
LDL	67,429	19:10	1.19	15.7	1.8×10^{-86}	2.2×10^{-88}	1.4×10^{-2108}	rs73015024	*LDLR*
LDL	67,429	19:33	0.83	6.0	5.0×10^{-4}	3.6×10^{-9}	3.3×10^{-12}	rs147791730	*CEBPG*
LDL	67,429	19:47	4.05	29.3	0*	0*	5.6×10^{-8411}	rs7412	*APOE*
TC	69,234	1:57	0.70	4.6	3.8×10^{-62}	4.4×10^{-61}	1.5×10^{-1119}	rs11591147	*PCSK9*
TC	69,234	1:110	0.62	3.5	1.5×10^{-33}	1.5×10^{-33}	9.0×10^{-1298}	rs12740374	*CELSR2*
TC	69,234	2:22	0.59	6.3	1.7×10^{-39}	1.1×10^{-36}	2.8×10^{-781}	rs934197	*APOB*
TC	69,234	2:43	0.73	5.4	1.4×10^{-33}	1.1×10^{-33}	8.6×10^{-436}	rs4299376	*ABCG8*
TC	69,234	19:10	0.91	10.8	1.2×10^{-69}	4.2×10^{-70}	1.0×10^{-1679}	rs73015024	*LDLR*
TC	69,234	19:29	0.88	4.1	5.2×10^{-4}	6.8×10^{-9}	1.1×10^{-7}	rs62108075	*AC005307.3*
TC	69,234	19:47	2.70	17.9	0*	0*	8.3×10^{-4123}	rs7412	*APOE*
TG	69,479	2:27	0.35	4.0	4.5×10^{-30}	1.4×10^{-29}	6.6×10^{-1357}	rs1260326	*GCKR*
TG	69,479	11:116	1.21	10.3	1.4×10^{-99}	1.4×10^{-99}	2.5×10^{-3336}	rs964184	*ZNF259*

Table of all linkage peaks with LOD > 3 in HUNT. The PVE and LOD are from our linkage analysis. The smallest observed p-values within 5 Mb of each linkage peak from our GWAS of genotyped variants in HUNT, our GWAS of imputed variants in HUNT, and the GLGC meta-analysis are reported in the table. The rs ids and gene names in the table refer to the SNP with the smallest of these p-values for each linkage peak.

*P-values of 0 are the result of underflow in the software used.

Figures

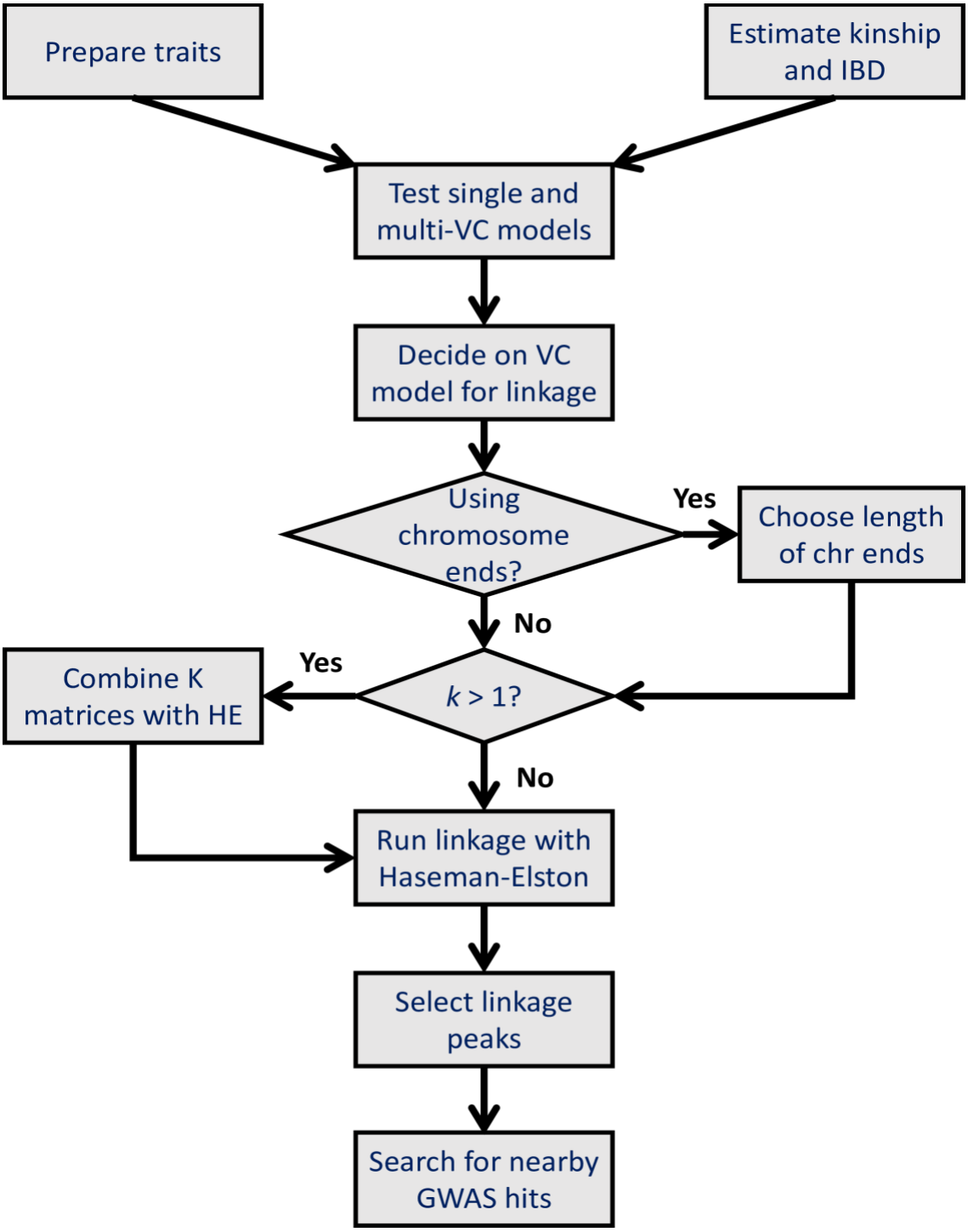


Figure 4-1 A flow chart of Population Linkage

A flow chart showing the basic steps for researchers to run Population Linkage on their own data.

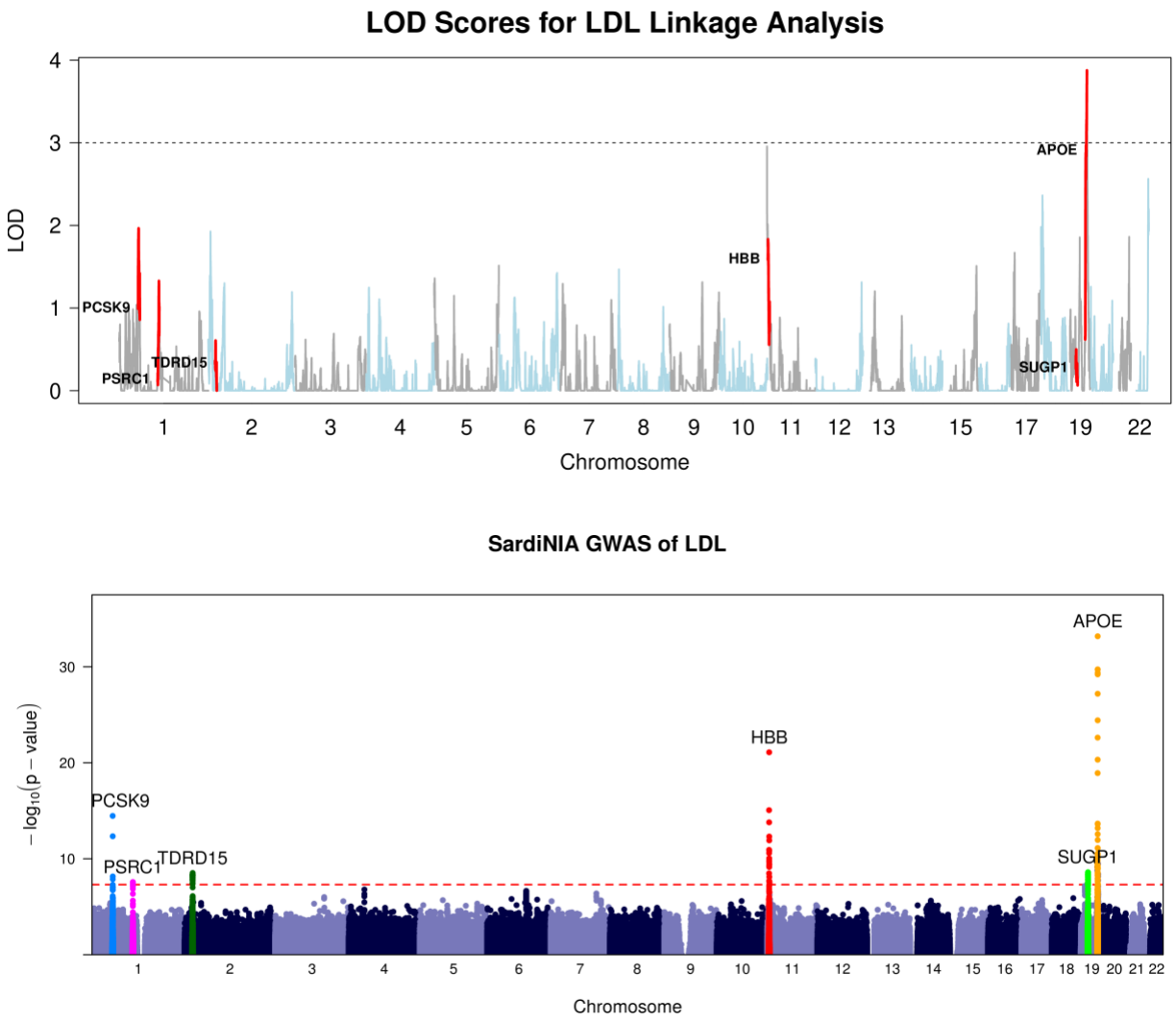


Figure 4-2 Overlap of LOD Scores for LDL with Known Regions

Above: LOD plot for the linkage analysis of LDL in SardiNIA. **Below:** GWAS of LDL in SardiNIA. Highlighted regions are the same in both plots.

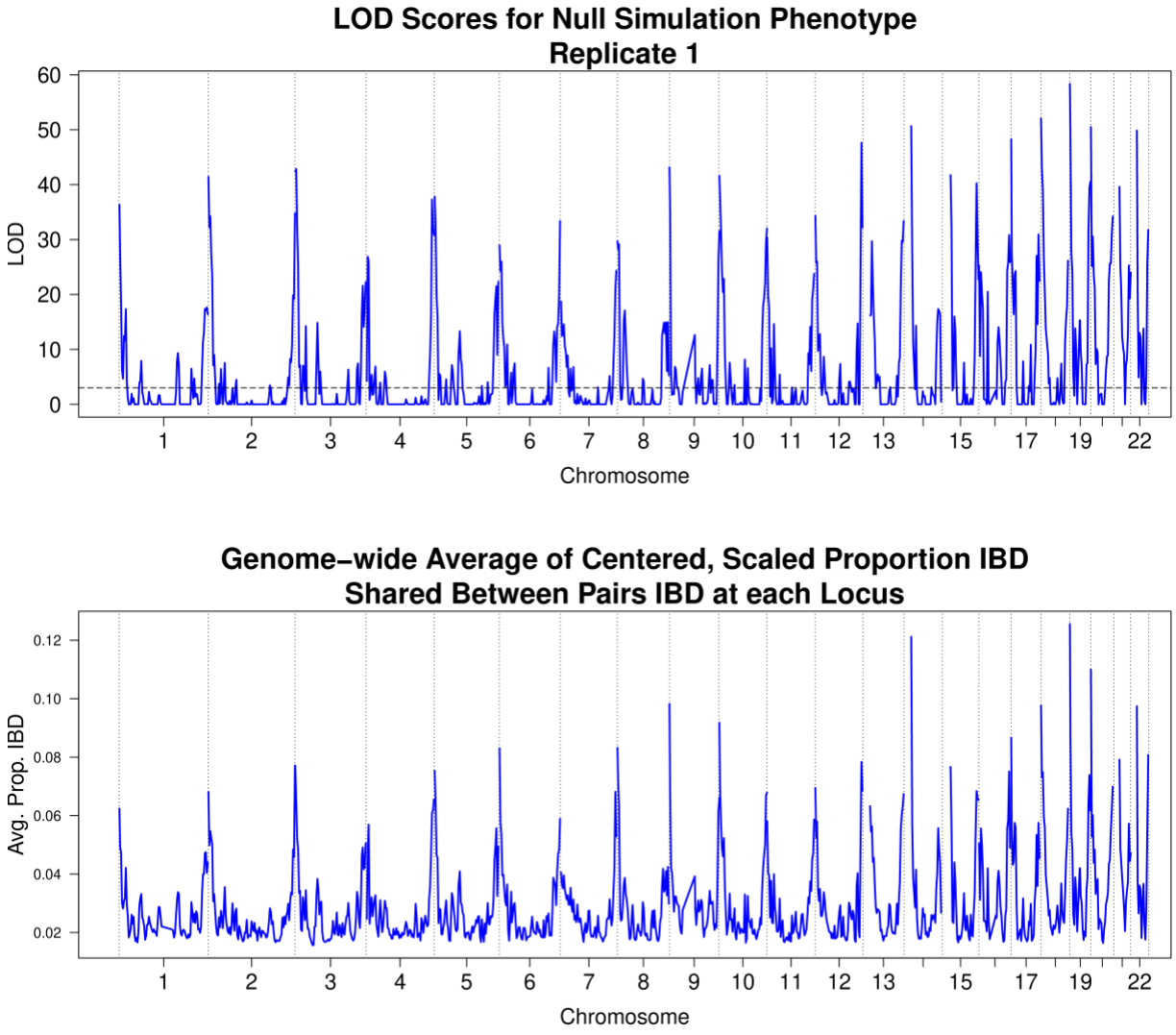


Figure 4-3 HUNT inflation example

Above: LOD plot from linkage analysis of the null simulated phenotype across 69,716 HUNT samples, using genome-wide pairwise IBD proportion and region-specific IBD sharing as variance components. **Below:** The average kinship values for individuals IBD across the genome. The phenotype below is LDL (n = 67,429) so 2,287 samples in the simulation above do not appear in the plot below but otherwise the plots would be identical since the values of kinship and estimated IBD status at each marker do not depend on the phenotype.

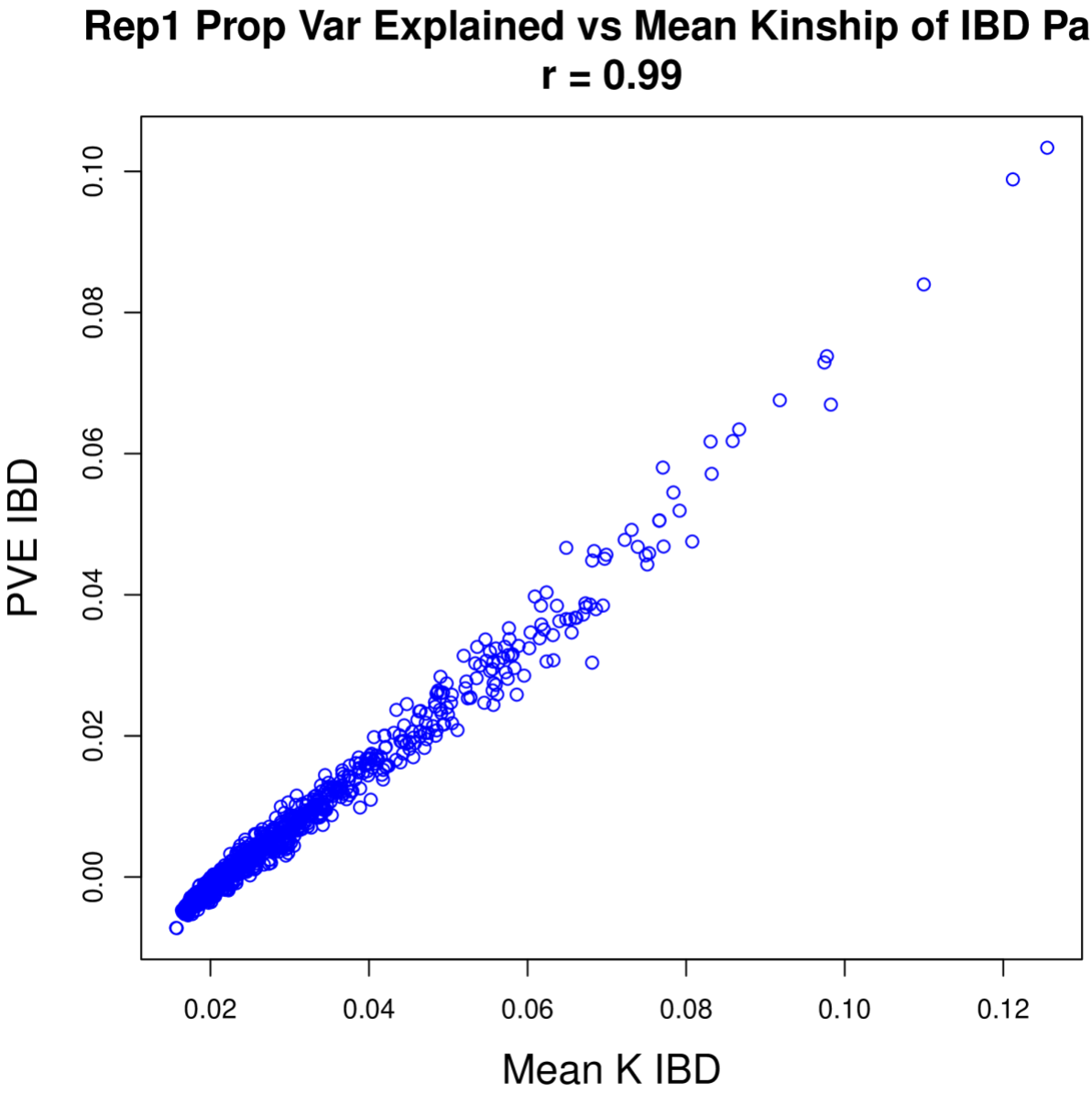


Figure 4-4 Proportion of variance explained vs mean kinship of IBD pairs

The estimated proportion of variance explained by IBD sharing at 1,000 sites from a linkage analysis of the HUNT null simulation phenotype plotted against the average kinship values for individuals IBD at each locus tested.

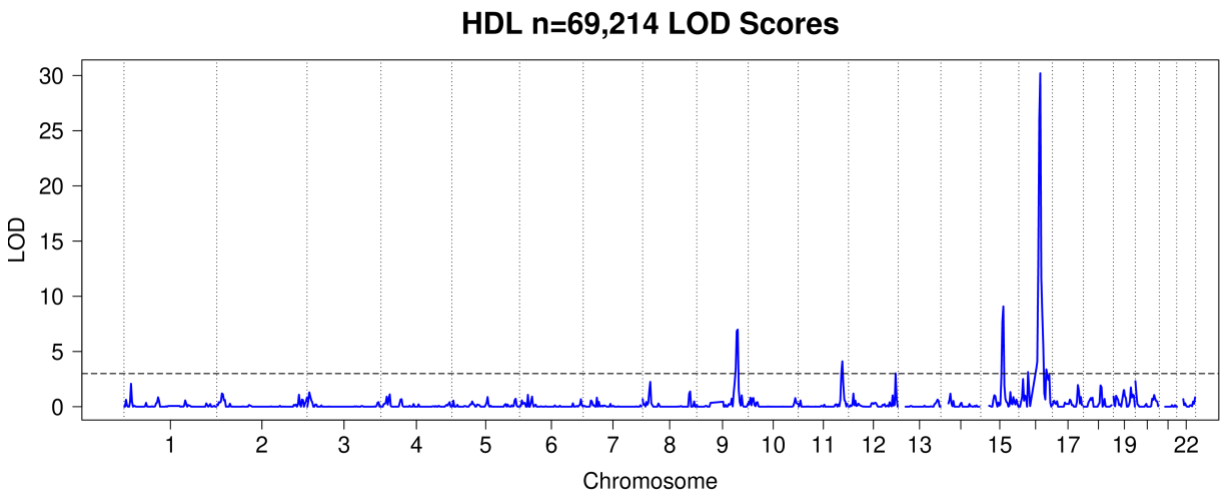


Figure 4-5 HUNT HDL LOD Scores

LOD plot from linkage analysis of high-density lipoprotein cholesterol measurements from HUNT.

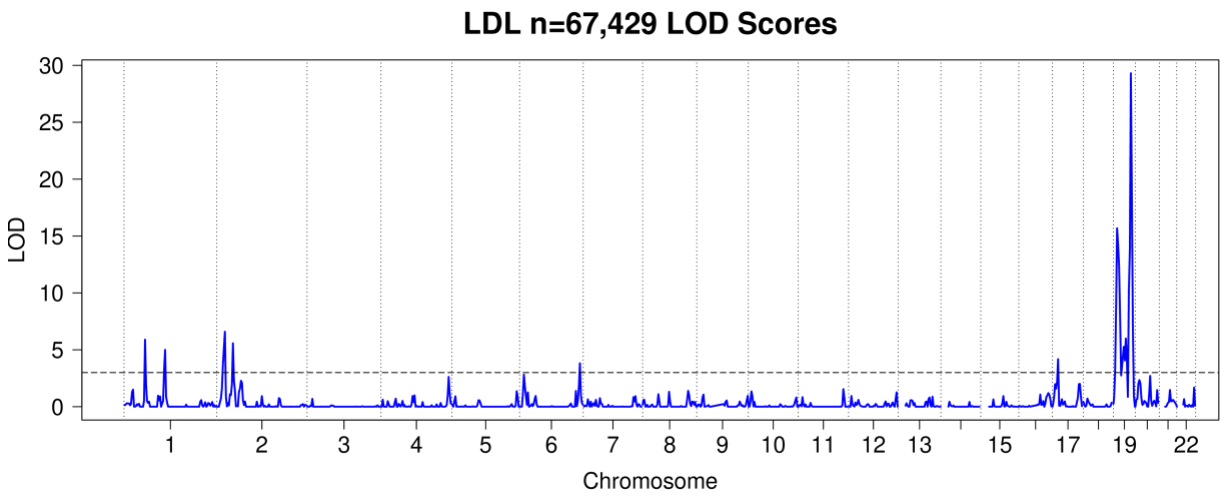


Figure 4-6 HUNT LDL LOD Scores

LOD plot from linkage analysis of low-density lipoprotein cholesterol measurements from HUNT.

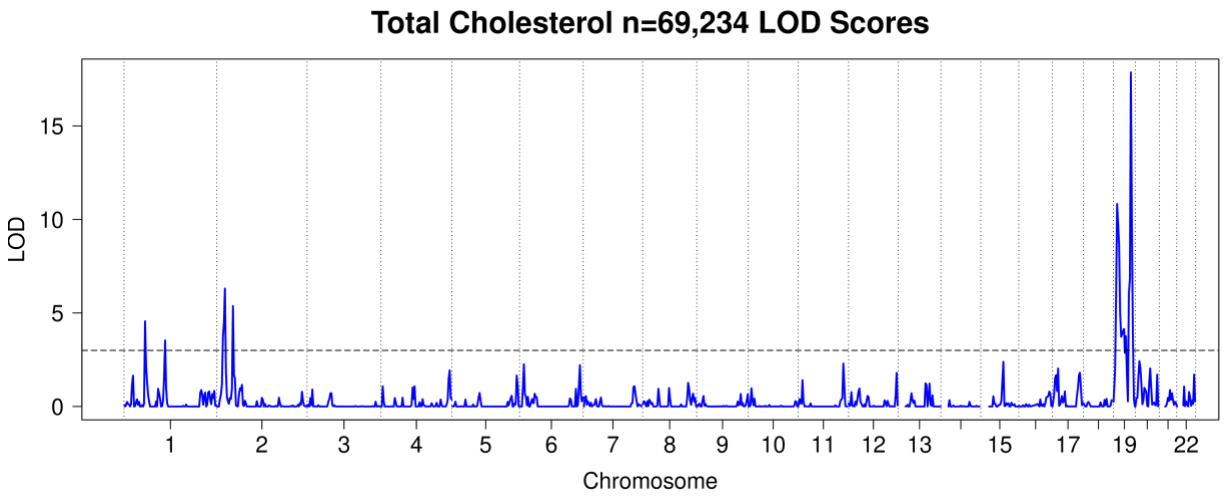


Figure 4-7 HUNT Total Cholesterol LOD Scores

LOD plot from linkage analysis of total cholesterol measurements from HUNT.

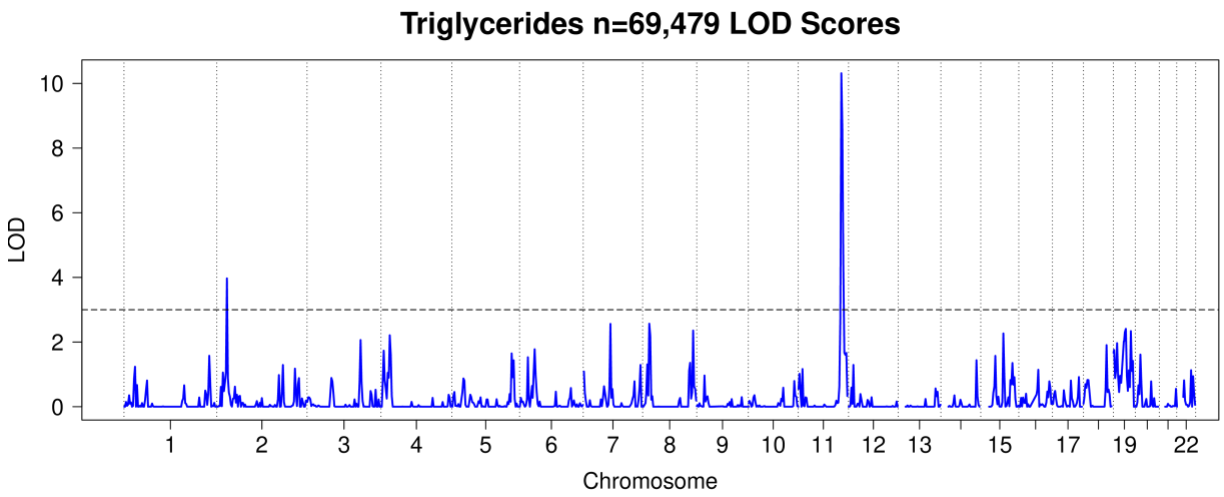


Figure 4-8 HUNT Triglycerides LOD Scores.

LOD plot from linkage analysis of triglycerides measurements from HUNT.

References

Abecasis, G. R., Cherny, S. S., Cookson, W. O., & Cardon, L. R. (2002). Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics, 30*(1), 97-101.

Amos, C. I. (1994). Robust variance-components approach for assessing genetic linkage in pedigrees. *American Journal of Human Genetics, 54*(3), 535-543.

Boehnke, M., & Cox, N. J. (1997). Accurate inference of relationships in sib-pair linkage studies. *American Journal of Human Genetics, 61*(2), 423-429.

Burman, D., Mente, A., Hegele, R. A., Islam, S., Yusuf, S., & Anand, S. S. (2009). Relationship of the ApoE polymorphism to plasma lipid traits among South Asians, Chinese, and Europeans living in Canada. *Atherosclerosis, 203*(1), 192-200. doi:10.1016/j.atherosclerosis.2008.06.007

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., . . . Marchini, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature, 562*(7726), 203-209. doi:10.1038/s41586-018-0579-z

Carithers, L. J., & Moore, H. M. (2015). The Genotype-Tissue Expression (GTEx) Project. *Biopreserv Biobank, 13*(5), 307-308. doi:10.1089/bio.2015.29031.hmm

Chasman, D. I., Kozlowski, P., Zee, R. Y., Kwiatkowski, D. J., & Ridker, P. M. (2006). Qualitative and quantitative effects of APOE genetic variation on plasma C-reactive protein, LDL-cholesterol, and apoE protein. *Genes Immun, 7*(3), 211-219. doi:10.1038/sj.gene.6364289

Chen, G. B. (2014). Estimating heritability of complex traits from genome-wide association studies using IBS-based Haseman-Elston regression. *Front Genet, 5*, 107. doi:10.3389/fgene.2014.00107

Chen, W. M., Manichaikul, A., Nguyen, J., Onengut-Gumuscu, S., & Rich, S. S. (2017). Integrated inference that accurately identifies close relatives in > 1 million samples. *Annual Meeting of the American Society of Human Genetics 2017, Orlando, FL*.

Chiang, C. W. K., Marcus, J. H., Sidore, C., Biddanda, A., Al-Asadi, H., Zoledziewska, M., . . . Novembre, J. (2018). Genomic history of the Sardinian population. *Nat Genet, 50*(10), 1426-1434. doi:10.1038/s41588-018-0215-8

Day-Williams, A. G., Blangero, J., Dyer, T. D., Lange, K., & Sobel, E. M. (2011). Linkage analysis without defined pedigrees. *Genet Epidemiol, 35*(5), 360-370. doi:10.1002/gepi.20584

Delaneau, O., Zagury, J. F., & Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods, 10*(1), 5-6. doi:10.1038/nmeth.2307

Delaneau, O., Zagury, J. F., Robinson, M. R., Marchini, J. L., & Dermitzakis, E. T. (2019). Accurate, scalable and integrative haplotype estimation. *Nat Commun, 10*(1), 5436. doi:10.1038/s41467-019-13225-y

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

Freeman, L. A., & Remaley, A. T. (2016). Chapter 6 - Discovery of High-Density Lipoprotein Gene Targets from Classical Genetics to Genome-Wide Association Studies. In A. Rodriguez-Oquendo (Ed.), *Translational Cardiometabolic Genomic Medicine* (pp. 119-159). Boston: Academic Press.

Georgi, B., Craig, D., Kember, R. L., Liu, W., Lindquist, I., Nasser, S., . . . Bucan, M. (2014). Genomic view of bipolar disorder revealed by whole genome sequencing in a genetic isolate. *PLoS Genet, 10*(3), e1004229. doi:10.1371/journal.pgen.1004229

Golan, D., Lander, E. S., & Rosset, S. (2014). Measuring missing heritability: inferring the contribution of common variants. *Proc Natl Acad Sci U S A, 111*(49), E5272-5281. doi:10.1073/pnas.1419064111

Haseman, J. K., & Elston, R. C. (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behav Genet, 2*(1), 3-19. doi:10.1007/bf01066731

Hodge, S. E., Hager, V. R., & Greenberg, D. A. (2016). Using Linkage Analysis to Detect Gene-Gene Interactions. 2. Improved Reliability and Extension to More-Complex Models. *PLoS One, 11*(1), e0146240. doi:10.1371/journal.pone.0146240

Illumina. (2017). Infinium ® CoreExome- 24 v1.2 BeadChip. https://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet_human_core_exome_beadchip.pdf.

Jiang, Q., Lee, C. Y., Mandrekar, S., Wilkinson, B., Cramer, P., Zelcer, N., . . . Landreth, G. E. (2008). ApoE promotes the proteolytic degradation of Abeta. *Neuron, 58*(5), 681-693. doi:10.1016/j.neuron.2008.04.010

Jong, M. C., Hofker, M. H., & Havekes, L. M. (1999). Role of ApoCs in lipoprotein metabolism: functional differences between ApoC1, ApoC2, and ApoC3. *Arterioscler Thromb Vasc Biol, 19*(3), 472-484. doi:10.1161/01.atv.19.3.472

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S. Y., Freimer, N. B., . . . Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat Genet, 42*(4), 348-354. doi:10.1038/ng.548

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., & Eskin, E. (2008). Efficient control of population structure in model organism association mapping. *Genetics, 178*(3), 1709-1723. doi:178/3/1709 [pii] 10.1534/genetics.107.080101

Kathiresan, S., Manning, A. K., Demissie, S., D'Agostino, R. B., Surti, A., Guiducci, C., . . . Cupples, L. A. (2007). A genome-wide association study for blood lipid phenotypes in the Framingham Heart Study. *BMC Med Genet, 8 Suppl 1*, S17. doi:10.1186/1471-2350-8-S1-S17

Klarin, D., Damrauer, S. M., Cho, K., Sun, Y. V., Teslovich, T. M., Honerlaw, J., . . . Program, V. M. V. (2018). Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. *Nat Genet, 50*(11), 1514-1523. doi:10.1038/s41588-018-0222-9

Krokstad, S., Langhammer, A., Hveem, K., Holmen, T. L., Midthjell, K., Stene, T. R., . . . Holmen, J. (2013). Cohort Profile: the HUNT Study, Norway. *Int J Epidemiol, 42*(4), 968-977. doi:10.1093/ije/dys095

Lange, K., & Boehnke, M. (1983). Extensions to pedigree analysis. IV. Covariance components models for multivariate traits. *American Journal of Medical Genetics, 14*(3), 513-524.

Liu, F., Kirichenko, A., Axenovich, T. I., van Duijn, C. M., & Aulchenko, Y. S. (2008). An approach for cutting large and complex pedigrees for linkage analysis. *Eur J Hum Genet, 16*(7), 854-860. doi:10.1038/ejhg.2008.24

Liutkeviciene, R., Vilkeviciute, A., Smalinskiene, A., Tamosiunas, A., Petkeviciene, J., Zaliuniene, D., & Lesauskaite, V. (2018). The role of apolipoprotein E (rs7412 and rs429358) in age-related macular degeneration. *Ophthalmic Genet, 39*(4), 457-462. doi:10.1080/13816810.2018.1479429

Lotta, L. A., Wittemans, L. B. L., Zuber, V., Stewart, I. D., Sharp, S. J., Luan, J., . . . Langenberg, C. (2018). Association of Genetic Variants Related to Gluteofemoral vs Abdominal Fat Distribution With Type 2 Diabetes, Coronary Disease, and Cardiovascular Risk Factors. *JAMA, 320*(24), 2553-2563. doi:10.1001/jama.2018.19329

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W. M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics, 26*(22), 2867-2873. doi:10.1093/bioinformatics/btq559

Minster, R. L., Sanders, J. L., Singh, J., Kammerer, C. M., Barmada, M. M., Matteini, A. M., . . . Newman, A. B. (2015). Genome-Wide Association Study and Linkage Analysis of the Healthy Aging Index. *J Gerontol A Biol Sci Med Sci, 70*(8), 1003-1008. doi:10.1093/gerona/glv006

More, H., Humar, B., Weber, W., Ward, R., Christian, A., Lintott, C., . . . Guilford, P. (2007). Identification of seven novel germline mutations in the human E-cadherin (CDH1) gene. *Hum Mutat, 28*(2), 203. doi:10.1002/humu.9473

Morton, N. E. (1955). Sequential tests for the detection of linkage. *Am J Hum Genet, 7*(3), 277-318.

Mousavi, N., Shleizer-Burko, S., Yanicky, R., & Gymrek, M. (2019). Profiling the genome-wide landscape of tandem repeat expansions. *Nucleic Acids Res, 47*(15), e90. doi:10.1093/nar/gkz501

Nielsen, J. B., Fritsche, L. G., Zhou, W., Teslovich, T. M., Holmen, O. L., Gustafsson, S., . . . Willer, C. J. (2018). Genome-wide Study of Atrial Fibrillation Identifies Seven Risk Loci and Highlights Biological Pathways and Regulatory Elements Involved in Cardiac Development. *Am J Hum Genet, 102*(1), 103-115. doi:10.1016/j.ajhg.2017.12.003

Oikkonen, J., Huang, Y., Onkamo, P., Ukkola-Vuoti, L., Raijas, P., Karma, K., . . . Järvelä, I. (2015). A genome-wide linkage and association study of musical aptitude identifies loci containing genes related to inner ear development and neurocognitive functions. *Mol Psychiatry, 20*(2), 275-282. doi:10.1038/mp.2014.8

Pilia, G., Chen, W. M., Scuteri, A., Orru, M., Albai, G., Dei, M., . . . Schlessinger, D. (2006). Heritability of cardiovascular and personality traits in 6,148 Sardinians. *PLoS Genet, 2*(8), e132. doi:06-PLGE-RA-0086R2 [pii] 10.1371/journal.pgen.0020132

Pistis, G., Porcu, E., Vrieze, S. I., Sidore, C., Steri, M., Danjou, F., . . . Sanna, S. (2015). Rare variant genotype imputation with thousands of study-specific whole-genome sequences: implications for cost-effective study designs. *Eur J Hum Genet, 23*(7), 975-983. doi:10.1038/ejhg.2014.216

Piven, J., Vieland, V. J., Parlier, M., Thompson, A., O'Conner, I., Woodbury-Smith, M., . . . Szatmari, P. (2013). A molecular genetic study of autism and related phenotypes in extended pedigrees. *J Neurodev Disord, 5*(1), 30. doi:10.1186/1866-1955-5-30

Pulit, S. L., Stoneman, C., Morris, A. P., Wood, A. R., Glastonbury, C. A., Tyrrell, J., . . . Consortium, G. (2019). Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry. *Hum Mol Genet, 28*(1), 166-174. doi:10.1093/hmg/ddy327

Richardson, T. G., Sanderson, E., Palmer, T. M., Ala-Korpela, M., Ference, B. A., Davey Smith, G., & Holmes, M. V. (2020). Evaluating the relationship between circulating lipoprotein lipids and apolipoproteins with risk of coronary heart disease: A multivariable Mendelian randomisation analysis. *PLoS Med, 17*(3), e1003062. doi:10.1371/journal.pmed.1003062

Risch, N. (1991). A note on multiple testing procedures in linkage analysis. *Am J Hum Genet, 48*(6), 1058-1064.

Sham, P. C., & Purcell, S. (2001). Equivalence between Haseman-Elston and variance-components linkage analyses for sib pairs. *Am J Hum Genet, 68*(6), 1527-1532. doi:10.1086/320593

Sham, P. C., Purcell, S., Cherny, S. S., & Abecasis, G. R. (2002). Powerful regression-based quantitative-trait linkage analysis of general pedigrees. *American Journal of Human Genetics, 71*(2), 238-253.

Sidore, C., Busonero, F., Maschio, A., Porcu, E., Naitza, S., Zoledziewska, M., . . . Abecasis, G. R. (2015). Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. *Nat Genet, 47*(11), 1272-1281. doi:10.1038/ng.3368

Spirin, V., Schmidt, S., Pertsemlidis, A., Cooper, R. S., Cohen, J. C., & Sunyaev, S. R. (2007). Common single-nucleotide polymorphisms act in concert to affect plasma levels of high-density lipoprotein cholesterol. *Am J Hum Genet, 81*(6), 1298-1303. doi:10.1086/522497

Tekola-Ayele, F., Lee, A., Workalemahu, T., & Sánchez-Pozos, K. (2019). Shared genetic underpinnings of childhood obesity and adult cardiometabolic diseases. *Hum Genomics, 13*(1), 17. doi:10.1186/s40246-019-0202-x

Thompson, E. A. (2019). Descent-Based Gene Mapping in Pedigrees and Populations. *Handbook of Statistical Genomics: Two Volume Set*, 573-596.

Thomson, R., & McWhirter, R. (2017). Adjusting for Familial Relatedness in the Analysis of GWAS Data. *Methods Mol Biol, 1526*, 175-190. doi:10.1007/978-1-4939-6613-4_10

Van Arendonk, J. A., Tier, B., Bink, M. C., & Bovenhuis, H. (1998). Restricted maximum likelihood analysis of linkage between genetic markers and quantitative trait loci for a granddaughter design. *J Dairy Sci, 81 Suppl 2*, 76-84. doi:10.3168/jds.s0022-0302(98)70156-0

Vieland, V. J., Huang, Y., Seok, S. C., Burian, J., Catalyurek, U., O'Connell, J., . . . Valentine-Cooper, W. (2011). KELVIN: a software package for rigorous measurement of statistical evidence in human genetics. *Hum Hered, 72*(4), 276-288. doi:10.1159/000330634

Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011). GCTA: a tool for genome-wide complex trait analysis. *American Journal of Human Genetics, 88*(1), 76-82. doi:10.1016/j.ajhg.2010.11.011

Zhou, W., Nielsen, J. B., Fritsche, L. G., Dey, R., Gabrielsen, M. E., Wolford, B. N., . . . Lee, S. (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat Genet, 50*(9), 1335-1341. doi:10.1038/s41588-018-0184-y

Zhou, X. (2017). A Unified Framework for Variance Component Estimation with Summary Statistics in Genome-Wide Association Studies. *Ann Appl Stat, 11*(4), 2027-2051. doi:10.1214/17-AOAS1052

Zhou, X., & Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics, 44*(7), 821-824. doi:10.1038/ng.2310

Chapter 5 Conclusion

In this dissertation, we have 1. implemented a platform for collecting human genetic data and phenotypes at a national scale, 2. described a novel method for joint estimation of DNA contamination and its sources in a genotyping study, and 3. extended a well-known method for linkage analysis to population-scale data.

The Genes for Good platform described in Chapter 2 has had the most easily measurable impact thus far. To date the study has recruited participants resembling the US distribution for income and chronic health indicators in all 50 states, received 2.9 million survey responses, and genotyped and returned results to 27,000 participants. We have contributed data to 4 studies and meta-analyses (Jiang et al., 2018; Liu et al., 2019; Sanchez-Roige et al., 2017; Savage et al., 2018). Most recently, we contacted participants to take our survey on COVID-19 and study the genetics of this new disease.

In addition to Genes for Good, there are now several academic or publicly-funded genotyping efforts that have a component of direct interaction with participants and return of results. The Healthy Nevada Project has partnered with 23andMe and Helix to genotype and return results to their study population and has successfully recruited and genotyped an ethnically and geographically diverse cohort of over 26,000 participants (Joseph J. Grzymski et al., 2018; J. J. Grzymski et al., 2020). All of Us is a particularly large study of health and epidemiology that aims to recruit over 1 million individuals from health centers and clinics all over the United States which places special emphasis on providing participants with access to their data, including returning

data on clinically actionable genetic variants and offering genetic counseling services (Denny et al., 2019). DNA.land is another academic effort that operates differently from these others in that it does not recruit participants for genotyping, but offers them genetic interpretation services including genetic ancestry inference and trait prediction for individuals who share genotypes they have already obtained from a direct-to-consumer genotyping company (Yuan et al., 2018). Use of DNA.land services is also not conditional on providing research data to the study.

The VICES method described in Chapter 3 has facilitated more accurate and more robust contamination estimation while yielding more useful results to researchers. This method has been incorporated into quality control pipelines for both the Michigan Genomics Initiative and Genes for Good. In the latter, we were able to conclude that contamination had occurred on the genotyping array for about 30 samples and were thus able to regenotype DNA from an earlier step to obtain clean results without collecting new DNA from participants. Using these same results, we were also able to communicate with Illumina about the contamination issues with their arrays and receive a commitment from them to address this issue in future products.

Since the publication of Chapter 3 of this dissertation and distribution of VICES, there have been additional developments in the area of contamination estimation that have also addressed some of the issues raised in that chapter. Specifically, Zhang et al (2020) introduced verifyBamID2, which performs joint estimation of sample genetic ancestry and contamination in sequencing reads. This approach, though distinct from VICES, also addresses the dependence of previous contamination methods on correctly specified allele frequencies to effectively estimate contamination and produces more robust results in diverse settings.

In Chapter 4, which introduced Population Linkage, we have enabled linkage analysis to be run on larger sample sizes and to run faster than ever before. In particular, we have showed how

Haseman-Elston regression—originally developed to test for genetic linkage in sibling pairs only—can be used as a general, method-of-moments approximation to variance components estimation with pairwise relationship and identical-by-descent estimates. While less powerful than the classical, full-likelihood linkage methods, Population Linkage allows applying this method to larger data. Such a tradeoff is a common theme in statistical genetics (Howie, Fuchsberger, Stephens, Marchini, & Abecasis, 2012). This work has also opened up new potential for running linkage and GWAS in tandem on large data sets and using the two to complement and strengthen one another.

While working on Population Linkage, a new algorithm for fitting genetic variance components with Haseman-Elston regression was introduced (Hou et al., 2019; Pazokitoroudi et al., 2020) that claims to achieve a speedup of several orders of magnitude over the method from Zhou (2017) used in Chapter 4. The method, called RHE-mc, achieves such speed by multiplying vectors of random subsets of genotypes to estimate a term in the Haseman-Elston regression that is normally calculated by multiplying genotype relatedness matrices together. By taking this approach, RHE-mc skips these costly calculations that depend on large matrices. While the method itself is not directly applicable to linkage since it uses individual SNP genotypes and not estimated IBD segments, several ideas from RHE-mc can be taken as inspiration for how to improve linkage analysis of large-scale genotyping data using Haseman-Elston regression. Possible benefits of developing a similar strategy for estimating variance components from IBD segments would include faster runtime, the ability to analyze even greater numbers of samples, and also the potential to estimate variance components for different genomic regions jointly from a single Haseman-Elston fit rather than from individual fits at each region being tested.

As statistical genetics continues to advance into the 2020s, we can expect to see more progress with respect to the topics presented in this dissertation. More genetic data from a variety of platforms will continue to be generated and from increasingly diverse groups of individuals. This will motivate new needs for statistical methodology in terms of participant recruitment, more nuanced approaches to quality checking, and new needs in terms of downstream analysis. These needs will become particularly relevant as long-read sequencing and trans-omics approaches become more ubiquitous. Even while we can expect a great deal of change for statistical genetics in the years ahead, the drive for discovery and the shared mission to learn what shapes us as human beings and how we can improve our health will doubtless remain as strong as ever.

References

Denny, J. C., Rutter, J. L., Goldstein, D. B., Philippakis, A., Smoller, J. W., Jenkins, G., . . . Investigators, A. o. U. R. P. (2019). The "All of Us" Research Program. *N Engl J Med, 381*(7), 668-676. doi: 10.1056/NEJMsr1809937

Grzymski, J. J., Coppes, M. J., Metcalf, J., Galanopoulos, C., Rowan, C., Henderson, M., . . . Slonim, A. (2018). The Healthy Nevada Project: rapid recruitment for population health study. *bioRxiv*, 250274. doi: 10.1101/250274

Grzymski, J. J., Elhanan, G., Morales Rosado, J. A., Smith, E., Schlauch, K. A., Read, R., . . . Lu, J. T. (2020). Population genetic screening efficiently identifies carriers of autosomal dominant diseases. *Nat Med, 26*(8), 1235-1239. doi: 10.1038/s41591-020-0982-5

Hou, K., Burch, K. S., Majumdar, A., Shi, H., Mancuso, N., Wu, Y., . . . Pasaniuc, B. (2019). Accurate estimation of SNP-heritability from biobank-scale data irrespective of genetic architecture. *Nat Genet, 51*(8), 1244-1251. doi: 10.1038/s41588-019-0465-0

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., & Abecasis, G. R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet, 44*(8), 955-959. doi: 10.1038/ng.2354

Jiang, Y., Chen, S., McGuire, D., Chen, F., Liu, M., Iacono, W. G., . . . Liu, D. J. (2018). Proper conditional analysis in the presence of missing data: Application to large scale meta-analysis of tobacco use phenotypes. *PLoS Genet, 14*(7), e1007452. doi: 10.1371/journal.pgen.1007452

Liu, M., Jiang, Y., Wedow, R., Li, Y., Brazel, D. M., Chen, F., . . . Psychiatry, H. A.-I. (2019). Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nat Genet, 51*(2), 237-244. doi: 10.1038/s41588-018-0307-5

Pazokitoroudi, A., Wu, Y., Burch, K. S., Hou, K., Zhou, A., Pasaniuc, B., & Sankararaman, S. (2020). Efficient variance components analysis across millions of genomes. *Nat Commun, 11*(1), 4020. doi: 10.1038/s41467-020-17576-9

Sanchez-Roige, S., Fontanillas, P., Elson, S. L., the 23andMe Research, T., Pandit, A., Schmidt, E. M., . . . Palmer, A. A. (2017). Genome-wide association study of delay discounting in 23,217 adult research participants of European ancestry. *Nature Neuroscience*.

Savage, J. E., Jansen, P. R., Stringer, S., Watanabe, K., Bryois, J., de Leeuw, C. A., . . . Posthuma, D. (2018). Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nat Genet, 50*(7), 912-919. doi: 10.1038/s41588-018-0152-6

Yuan, J., Gordon, A., Speyer, D., Aufrichtig, R., Zielinski, D., Pickrell, J., & Erlich, Y. (2018). DNA.Land is a framework to collect genomes and phenomes in the era of abundant genetic information. *Nat Genet, 50*(2), 160-165. doi: 10.1038/s41588-017-0021-8

Zhang, F., Flickinger, M., Taliun, S. A. G., Abecasis, G. R., Scott, L. J., McCaroll, S. A., . . . Consortium, I. P. G. (2020). Ancestry-agnostic estimation of DNA sample contamination from sequence reads. *Genome Res, 30*(2), 185-194. doi: 10.1101/gr.246934.118

Zhou, X. (2017). A UNIFIED FRAMEWORK FOR VARIANCE COMPONENT ESTIMATION WITH SUMMARY STATISTICS IN GENOME-WIDE ASSOCIATION STUDIES. *Ann Appl Stat, 11*(4), 2027-2051. doi: 10.1214/17-AOAS1052

Appendix

In an uncontaminated sample, the following probability distribution relates array intensity for sample i at marker j, I_{ij}, to the genotype G_{ij}:

$$\Pr(I_{ij} = x|G_{ij}) = \begin{cases} \Phi\left(-\frac{G_{ij}}{\sigma}\right) & if\ x = 0 \\ 1 - \Phi\left(\frac{1-G_{ij}}{\sigma}\right) & if\ x = 1 \\ x \sim N(G_{ij}, \sigma^2) & if\ 0 < x < 1 \\ 0 & o.w. \end{cases}.$$

Under this model, intensities are normally distributed around the genotype G_{ij} with additional point masses reflecting the truncation at boundaries $I_{ij} = 0$ and $I_{ij} = 1$. σ^2 represents the naturally-occurring variability in intensity values.

For a contaminated sample, I_{ij} is instead distributed around a linear combination of the sample's own genotype and the genotypes of each contaminating sample, which we denote as μ_{ij}. Let α_i be the total proportion of contaminating DNA in sample i and α_{ik} the proportion of DNA mixture from sample k. Then, we define

$$\mu_{ij} = (1 - \alpha_i)G_{ij} + \sum_k \alpha_{ik} G_{kj}$$

and the distribution of I_{ij} in the presence of contamination now becomes

$$\Pr(I_{ij} = x|\mu_{ij}) = \begin{cases} \Phi\left(-\frac{\mu_{ij}}{\sigma}\right) & if\ x = 0 \\ 1 - \Phi\left(\frac{1-\mu_{ij}}{\sigma}\right) & if\ x = 1 \\ x \sim N(\mu_{ij}, \sigma^2) & if\ 0 < x < 1 \\ 0 & o.w. \end{cases}.$$