Statistical and Computational Methods for Analyzing and Visualizing Large-Scale Genomic Datasets

by

Alan M. Kwong

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
Kari Branham
Associate Professor Hyun Min Kang
Professor Jun Z. Li
Associate Professor Xiaoquan Wen

Alan M. Kwong

amkwong@umich.edu

ORCID iD: 0000-0002-0779-7126

DEDICATION

To my family,
beacons of kindness and generosity
who continue to inspire me
to never stop striving to become a better person

ACKNOWLEDGMENTS

I am grateful to everyone who helped me along in this lengthy journey. I would like to thank, first and foremost, Gonçalo Abecasis, both for the brilliantly effortless way he drills down to the core of all topics, and his infinite patience in explaining those topics to me when I inevitably fail to understand some crucial detail; Hyun Min Kang, who has been so generous with his time in providing me with guidance on matters big and small; Lars Fritsche, always offering great advice and good humor; Michael Boehnke, whose empathy and care is only exceeded by the fullness of his schedule; Andy Boughton, with invaluable help on programming and also baked goods; and many more generous souls in the University of Michigan Department of Biostatistics.

I am also grateful to my friends and colleagues, who made this journey worthwhile. Paul Imbriano, with whom I can discuss any topic and inevitably discover interesting new insights; Sai Dharmarajan, with an inspirational appreciation for all that life has to offer in all its majesty and comedy; Daniel Muenz, a stalwart friend who always sees the good in all things, a force of positivity; Sayantan Das, with an ambitious appetite for all things and an intellect to match; Greg Zajac, always generous and in good cheer, who will find a way to help anyone in need; and too many others to list.

Finally, thank you to my parents and my sister, who have supported me through all these years in graduate school, without whom I would not be who I am today.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Advances in large-scale genomic data production have led to a need for better methods to process, interpret, and organize this data. Starting with raw sequencing data, generating results requires many complex data processing steps, from quality control, alignment, and variant calling to genome wide association studies (GWAS) and characterization of expression quantitative trait loci (eQTL). In this dissertation, I present methods to address issues faced when working with large-scale genomic datasets.

In Chapter 2, I present an analysis of 4,787 whole genomes sequenced for the study of age-related macular degeneration (AMD) as a follow-up fine-mapping study to previous work from the International AMD Genomics Consortium (IAMDGC). Through whole genome sequencing, we comprehensively characterized genetic variants associated with AMD in known loci to provide additional insights on the variants potentially responsible for the disease by leveraging 60,706 additional controls. Our study improved the understanding of loci associated with AMD and demonstrated the advantages and disadvantages of different approaches for fine-mapping studies with sequence-based genotypes.

In Chapter 3, I describe a novel method and a software tool to perform Hardy-Weinberg equilibrium (HWE) tests for structured populations. In sequence-based genetic studies, HWE test statistics are important quality metrics to distinguish true

genetic variants from artifactual ones, but it becomes much less informative when it is applied to a heterogeneous and/or structured population. As next generation sequencing studies contain samples from increasingly diverse ancestries, we developed a new HWE test which addresses both the statistical and computational challenges of modern large-scale sequencing data and implemented the method in a publicly available software tool. Moreover, we extensively evaluated our proposed method with alternative methods to test HWE in both simulated and real datasets. Our method has been successfully applied to the latest variant calling QC pipeline in the TOPMed project.

In Chapter 4, I describe *PheGET*, a web application to interactively visualize Expression Quantitative Trait Loci (eQTLs) across tissues, genes, and regions to aid functional interpretations of regulatory variants. Tissue-specific expression has become increasingly important for understanding the links between genetic variation and disease. To address this need, the Genotype-Tissue Expression (GTEx) project collected and analyzed a treasure trove of expression data. However, effectively navigating this wealth of data to find signals relevant to researchers has become a major challenge. I demonstrate the functionalities of *PheGET* using the newest GTEx data on our eQTL browser website at https://eqtl.pheweb.org/, allowing the user to 1) view all cis-eQTLs for a single variant; 2) view and compare single-tissue, single-gene associations within any genomic region; 3) find the best eQTL signal in any given genomic region or gene; and 4) customize the plotted data in real time. *PheGET* is designed to handle and display the kind of complex multidimensional data often seen in our post-GWAS era, such as multi-tissue expression data, in an intuitive and convenient interface, giving researchers an additional tool to better understand the links between genetics and disease.

Chapter 1

Introduction

1.1 Background

Genome-wide association studies (GWAS), in which variants across the whole genome are surveyed for connections to phenotypes, are widely used to study diseases (FRITSCHE *et al.* 2016; XUE *et al.* 2018), traits (WOOD *et al.* 2014; LOCKE *et al.* 2015), and many other measurable phenotypes, including biomarkers, biomedical images, and survey results (BYCROFT *et al.* 2018). In 1996, Risch and Merikangas (1996) argued that association could be more powerful than linkage analysis with larger sample sizes, which would be feasible with developments in genomic technology. In 2005, the first successful genome wide association study for age-related macular degeneration (KLEIN *et al.* 2005) contained 103,611 variants in 96 cases and 50 controls. As genotyping arrays and next-generation sequencing technologies improved and became much cheaper over time, studies rapidly increased in size and diversity: from thousands of samples of mainly European ancestry (HAKONARSON *et al.* 2007; SLADEK *et al.* 2007) to hundreds of thousands of samples (LOCKE *et al.* 2015; BYCROFT *et al.* 2018; TALIUN *et al.* 2019), for hundreds of millions of variants from samples across a range of ancestries. With this comes an increase in power to discover associations, with the largest studies finding dozens to hundreds of independent genetic loci for complex traits. While GWAS has successfully discovered thousands of associated loci, we are still in the early days of connecting them to biological mechanisms underlying many disease-related traits (FARASHI *et al.* 2019; ORMEL *et al.* 2019; CLAUSSNITZER *et al.* 2020). One of the next steps in unraveling the role of genetics in biology lies in understanding tissue-specific gene expression's biological consequences, the next major focus for human genomic studies (THE GTEX CONSORTIUM 2015). Genetic studies now encompass a much larger range of data, including genotypes, gene expression values (MCDERMAID *et al.* 2019), methylation and other epigenetic modifications (LAM *et al.* 2016; BAKUSIC *et al.* 2017), and somatic mutations (LEIJA-SALAZAR *et al.* 2018), for bulk tissue and single-cell analyses. The size and nature of genetic data in the current post-GWAS era promises vast new opportunities for biological discoveries, but also comes with many technical and scientific challenges.

Challenges for genetic studies in the post-GWAS era

As GWAS cohorts continue to grow in size, so does the number of associated loci. In single-variant association tests, the top signals will often not fall within exons, promoter, or enhancer regions of genes, with well-defined functional effects on the transcription and translation of the protein product; instead, the most significant associated variants often lie in intronic regions, and are assumed to be in high linkage disequilibrium with true causal variants (WANG *et al.* 2010). While group-based association tests can provide additional insight, they require additional assumptions and contain pitfalls which must be carefully avoided. Properly interpreting associated loci and identifying causal variants have become enormous challenges of their own.

As genetic studies become both larger and more diverse, previous quality control methods for variant calling are no longer sufficient both statistically and computationally, prompting a need for an improved method to deal with the more diverse nature of the data and a better implementation of that method to handle the larger data files found in modern genetic studies.

With the growing popularity of multi-tissue expression studies, the resulting expression quantitative trait loci (eQTL) data has become much more common. This kind of data involves millions of variants, each affecting up to dozens of genes in multiple tissues. The multidimensional nature of eQTL data makes them difficult to display using traditional tools, such as GWAS or phenome-wide association study (PheWAS) plots, just when visualization becomes a fundamental tool for understanding the structure of such data. With three different data dimensions—variant position, affected gene, and tissue—visualization must necessarily hold one dimension constant to display the other two, but this will limit our ability to find correlation patterns within our data for that dimension, i.e. for nearby variants, proximal genes, and biologically similar tissues. With an increasing number of publicly available genetic resources, it is even more difficult to get a more complete picture of known information for any particular eQTL. There is a need for an eQTL browser which allows for convenient and intuitive navigation across data dimensions, while also cross-referencing data from other existing databases, to give researchers a more comprehensive understanding of the results, facilitating the characterization of disease-related traits to generate new hypotheses and the exploration of biological mechanisms underlying those traits.

Purpose

In this dissertation, we develop and apply statistical methods and approaches to evaluate single-variant and set-based approaches for case-control genome wide association for sequencing data using data with an emphasis on the interpretation of significant associated loci and rare loss-of-function variants, develop and implement a robust and unified Hardy-Weinberg test for quality which can handle both the increasing size and diversity of genetic studies, and develop an interactive, intuitive, and convenient web-based application for the browsing of multi-tissue eQTL data, designed to facilitate the interpretation of expression association signals.

1.2 Quality control for variant calls in diverse genetic data

With large genetic studies comes the need for reliable quality control of genetic data. Accurate genetic association analyses require high quality genotypes, but genomic technologies are susceptible to errors. Wall et al. (2014) estimated genotype error rates ranging from 0.1% to 6%, depending on allele frequency and sequencing platform. Sample contamination can further increase errors in genotype calls (JUN *et al.* 2012; FLICKINGER *et al.* 2015; ZAJAC *et al.* 2019). Filtering erroneously called variants from downstream analysis is necessary to prevent spurious association signals.

Partly due to concerns about inflated Type I errors caused by population stratification (MARCHINI *et al.* 2004; NEED AND GOLDSTEIN 2009), early genetic studies often used only samples from participants of European ancestry. With samples of largely homogeneous ancestry, simple tests of HWE proved effective as a quality control metric for genotyped markers (GOMES *et al.* 1999; ANDERSON *et al.* 2010). HWE tests have since become one of the most common metrics for variant quality and are still widely used to this day.

For many years, the vast majority of participants in genetic research were of European ancestry. One study found that by 2016, 81% of studied samples were of European ancestry, and another 14% were of Asian ancestry (POPEJOY AND FULLERTON 2016). Unfortunately, this lack of diversity in genetic research has directly led disparities in translational medicine. For example, using results from genetic research based only on samples of European ancestry led to genetic misdiagnoses for patients of non-European ancestries (MANRAI *et al.* 2016). This lack of diversity has become a major hurdle for precision medicine (LANDRY *et al.* 2018), such as in the use of polygenic risk scores to gauge genetic risk for patients with non-European ancestry (DUNCAN *et al.* 2019; MARTIN *et al.* 2019; SIRUGO *et al.* 2019). The field of genetics is at risk of generating results that are only useful for improving health outcomes in Europeans and Asians (WANG *et al.* 2018). To address these concerns, recent studies such as the Trans-Omics Precision Medicine (TOPMed) project (TALIUN *et al.* 2019) and the All of Us Research Program (2019) made an effort to collect and sequence genetic samples from considerably more diverse populations.

However, the size and diversity of these recent genetic studies present new challenges for data processing and quality control. As genetic studies participants became both more numerous and more diverse, the same HWE tests were used for variant quality control, but with more stringent P-value thresholds (LOCKE *et al.* 2015; FRITSCHE *et al.* 2016; TALIUN *et al.* 2019). Though the traditional HWE test will still identify variants which wildly depart from HWE, it will fail to distinguish variants with significant errors (variants we want to filter) from those under the effect of population stratification (variants which are otherwise high-quality and which we want to keep for downstream analysis). A better HWE-based method for variant quality control is needed to ensure high quality genotype data is available for downstream analysis.

In Chapter 3, we propose and implement a robust and unified Hardy-Weinberg test for variant calling quality control in sequencing data. Our method is designed to handle the diverse samples found in modern genetic studies by leveraging ancestry information from raw sequencing data to adjust for population stratification, with the ability to directly process commonly-used file formats for genotype data with a computationally- and memory-efficient implementation.

1.3 Variant interpretation and causal variant discovery with whole genome sequence data

With our ability to discover large numbers of associated loci, we must now confront the gap between association and biological function: first, within each locus, identifying causal variants responsible for the association; and second, pinpointing the genes affected by these causal variants (GALLAGHER AND CHEN-PLOTKIN 2018). The early successes of GWAS on age-related macular degeneration (AMD), which identified the complement system as a key component of disease, were encouraging. However, as genetic studies exploded in popularity in subsequent years, the gap widened for estimates of the proportions of phenotypic variance explained by genetic differences (commonly called heritability) between GWAS-based and traditional epidemiological methods, especially for complex traits such as human height (MAHER 2008; MANOLIO *et al.* 2009), highlighting the limitations of GWAS in identifying the biological mechanisms behind association signals. Larger sample sizes and more variants, from denser genotyping and next-generation sequencing technologies, can help close part of that gap.

Foreseeing the increasing availability of genomic data, Cooper and Shendure (2011) argued that the main roadblock for finding causal variants will not simply be identifying more and more loci, but will instead be the interpretation of association results. As studies increased in both sample size and variant count, more associated loci were indeed identified, from dozens (LOCKE *et al.* 2015; SANDERS *et al.* 2015; FRITSCHE *et al.* 2016; OKBAY *et al.* 2016) to over a hundred (WILLER *et al.* 2013; WOOD *et al.* 2014; SCOTT *et al.* 2017; YENGO *et al.* 2018). With so many associated loci, connecting association signals to underlying biological mechanisms became correspondingly complicated. Recent approaches include using kernel methods for aggregated association testing (LARSON *et al.* 2019), constructing polygenic risk models using machine learning algorithms (OH *et al.* 2017), and using network and pathway-based methods to characterize function and biochemical pathway enrichment (HU *et al.* 2017).

There is some disagreement about the assumptions underlying genetic research design. Chakravarti and Turner (2016) emphasized the importance of understanding gene regulatory networks to properly interpret GWAS results, using Hirschsprung disease as an example of how complex traits can be affected by cis-regulatory elements in multiple genes. Boyle et al. (2017) argued for an “omnigenic” model, in which genes are divided into a small set of “core” genes and a large set of “peripheral” genes, and variants across the entire genome can affect phenotypic traits, such that peripheral genes are generally responsible for the majority of the genetic effect on phenotypes, if we assume gene regulatory networks have structures resembling the highly-connected "small world" network model (WATTS AND STROGATZ 1998). On the other hand, Wray et al. (2018) thought that the core genes assumption "may underestimate the true biological complexity" of common diseases, instead emphasized increasing sample sizes to maximize the discovery of common associated variants. Despite their differences, all parties largely agreed that cell-specific gene regulatory networks remain an important target for genomic research, and that identifying significant associations in genetic studies is only the first step in elucidating the underlying biology of the associated traits.

In Chapter 2, we evaluate different methods and approaches to analyze genetic association data in a case-control study of a complex disease, identifying the strengths and weaknesses of single-variant and grouped association tests in the context of trying to understand the underlying biology of age-related macular degeneration. We also evaluated the feasibility of leveraging publicly available genetic data in improving the ability to find association signals for rare loss-of-function variants in our effort to identify potential core genes for the disease. Finally, we give recommendations for how to interpret significantly associated loci with results from different association tests.

1.4 Visualization of multi-tissue expression data and the future of genetic studies

It has long been understood that discovering associated loci is just the first step towards disentangling the biological links between genes and disease (MANOLIO *et al.* 2009). Since all somatic cells in the human body generally share the same DNA sequence, differences between tissues were long assumed to be caused by tissue-specific expression (Britten and Davidson 1969), decades before the technology existed to accurately map and quantify mRNA. With the development of more affordable DNA technology, it was possible to finally put this hypothesis to the test for various diseases. While some Mendelian diseases, such as Huntington's Disease and phenylketonuria, have well-understood loss-of-function and missense risk variants which directly modify translated protein products, complex traits and diseases tend to be much less straightforward: very often, the most significant variants within each locus in a GWAS are either in linkage disequilibrium with rarer exonic variants which directly affect the translated protein product, or play a part in modifying the regulatory behavior of the gene. Group-based association tests in GWAS can test for the first case, but gene expression studies are needed in the second case.

The study of gene expression requires the quantification of RNA. Adapting the same technology used for reading DNA, RNA quantification initially used array-based and Sanger sequencing methods. Early expression arrays suffered from non-specific hybridization, leading to spurious results (Okoniewski and Miller 2006), while Sanger sequencing-based methods were expensive, preventing them from being used widely. Eventually, the development of RNA-seq (Wang *et al.* 2009) provided a method for high-throughput, accurate, and affordable expression quantification, sparking a prodigious increase in the amount of expression data generated in genetic studies. For example, in a study of gene expression in immune cells from 91 subjects (Schmiedel *et al.* 2018), cis-eQTLs were identified for over 12,000 unique genes, 41% of which with strong cis-associations in only a single cell type. The largest project to catalogue expression data is the GTEx project, with expression data for up to ~700 subjects in 49 different tissues in the latest data freeze (AGUET *et al.* 2019).

This new bounty of expression data, in which a single variant can be associated with multiple genes and tissues, creates new challenges for data visualization. Existing tools for showing genome-wide (Manhattan plots and GWAS locus plots) and phenome-wide association data (PheWAS plots) can only properly show one tissue or gene at a time, and are insufficient for showing a more complete picture of the relationships between a variant and all affected genes and tissues. Notably, the GTEx Portal (gtexportal.org) hosts a suite of services designed to display their latest data set. While comprehensive, the information is divided between multiple different and separate visualizations which cannot be easily viewed together, and it can be difficult to navigate between the different views.

At the same time, with the increasing availability of publicly-available genomic data on a variety of platforms—for example, dbSNP (SHERRY *et al.* 1999), UCSC (HAEUSSLER *et al.* 2019), GTEx Portal (AGUET *et al.* 2019), gnomAD (KARCZEWSKI *et al.* 2020), UK Biobank (BYCROFT *et al.* 2018), and the Expression Atlas (PETRYSZAK *et al.* 2016)—the basic task of retrieving all available information for a given variant or gene has become increasingly difficult. Together, these serve as unnecessary speed bumps for researchers for effectively interpreting tissue-specific eQTLs, in order to understand their effects on disease-related traits and the biological mechanisms underlying those effects.

In Chapter 4, we propose and implement *PheGET*, a browser for eQTLs designed to be intuitive, convenient, and flexible for displaying complicated multidimensional data, such as that found in multi-tissue cis-eQTL GTEx data. Using extended functionality from the popular visualization tool LocusZoom (PRUIM *et al.* 2010), the browser provides a simple way to navigate to any variant to view all available associated cis-eQTL information, or browse to any genomic region for gene-specific eQTLs in any of the 49 tissues found in GTEx. Our browser also offers links to other large public genetic databases, making it easier to cross-reference all available information on the variant or gene in question. The underlying tools and technology allow anyone with expression data to set up their own eQTL browser, making it easier to create and share visualizations of eQTLs of interest.

1.5 References

Aguet, F., A. N. Barbeira, R. Bonazzola, A. Brown, S. E. Castel *et al.*, 2019 The GTEx Consortium atlas of genetic regulatory effects across human tissues. bioRxiv.

All of Us Research Program, I., J. C. Denny, J. L. Rutter, D. B. Goldstein, A. Philippakis *et al.*, 2019 The "All of Us" Research Program. N Engl J Med 381**:** 668-676.

Anderson, C. A., F. H. Pettersson, G. M. Clarke, L. R. Cardon, A. P. Morris *et al.*, 2010 Data quality control in genetic case-control association studies. Nat Protoc 5**:** 1564-1573.

Bakusic, J., W. Schaufeli, S. Claes and L. Godderis, 2017 Stress, burnout and depression: A systematic review on DNA methylation mechanisms. J Psychosom Res 92**:** 34-44.

Boyle, E. A., Y. I. Li and J. K. Pritchard, 2017 An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell 169**:** 1177-1186.

Britten, R. J., and E. H. Davidson, 1969 Gene regulation for higher cells: a theory. Science 165**:** 349-357.

Bycroft, C., C. Freeman, D. Petkova, G. Band, L. T. Elliott *et al.*, 2018 The UK Biobank resource with deep phenotyping and genomic data. Nature 562**:** 203-209.

Chakravarti, A., and T. N. Turner, 2016 Revealing rate-limiting steps in complex disease biology: The crucial importance of studying rare, extreme-phenotype families. Bioessays 38**:** 578-586.

Claussnitzer, M., J. H. Cho, R. Collins, N. J. Cox, E. T. Dermitzakis *et al.*, 2020 A brief history of human disease genetics. Nature 577**:** 179-189.

Cooper, G. M., and J. Shendure, 2011 Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat Rev Genet 12**:** 628-640.

Duncan, L., H. Shen, B. Gelaye, J. Meijsen, K. Ressler *et al.*, 2019 Analysis of polygenic risk score usage and performance in diverse human populations. Nat Commun 10**:** 3328.

Farashi, S., T. Kryza, J. Clements and J. Batra, 2019 Post-GWAS in prostate cancer: from genetic association to biological contribution. Nat Rev Cancer 19**:** 46-59.

Flickinger, M., G. Jun, G. R. Abecasis, M. Boehnke and H. M. Kang, 2015 Correcting for Sample Contamination in Genotype Calling of DNA Sequence Data. Am J Hum Genet 97**:** 284-290.

Fritsche, L. G., W. Igl, J. N. Bailey, F. Grassmann, S. Sengupta *et al.*, 2016 A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. Nat Genet 48**:** 134-143.

Gallagher, M. D., and A. S. Chen-Plotkin, 2018 The Post-GWAS Era: From Association to Function. Am J Hum Genet 102**:** 717-730.

Gomes, I., A. Collins, C. Lonjou, N. S. Thomas, J. Wilkinson *et al.*, 1999 Hardy-Weinberg quality control. Ann Hum Genet 63**:** 535-538.

Haeussler, M., A. S. Zweig, C. Tyner, M. L. Speir, K. R. Rosenbloom *et al.*, 2019 The UCSC Genome Browser database: 2019 update. Nucleic Acids Res 47**:** D853-D858.

Hakonarson, H., S. F. Grant, J. P. Bradfield, L. Marchand, C. E. Kim *et al.*, 2007 A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene. Nature 448**:** 591-594.

Hu, Y. S., J. Xin, Y. Hu, L. Zhang and J. Wang, 2017 Analyzing the genes related to Alzheimer's disease via a network and pathway-based approach. Alzheimers Res Ther 9**:** 29.

Jun, G., M. Flickinger, K. N. Hetrick, J. M. Romm, K. F. Doheny *et al.*, 2012 Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. Am J Hum Genet 91**:** 839-848.

Karczewski, K. J., L. C. Francioli, G. Tiao, B. B. Cummings, J. Alföldi *et al.*, 2020 The mutational constraint spectrum quantified from variation in 141,456 humans. bioRxiv.

Klein, R. J., C. Zeiss, E. Y. Chew, J. Y. Tsai, R. S. Sackler *et al.*, 2005 Complement factor H polymorphism in age-related macular degeneration. Science 308**:** 385-389.

Lam, K., K. Pan, J. F. Linnekamp, J. P. Medema and R. Kandimalla, 2016 DNA methylation based biomarkers in colorectal cancer: A systematic review. Biochim Biophys Acta 1866**:** 106-120.

Landry, L. G., N. Ali, D. R. Williams, H. L. Rehm and V. L. Bonham, 2018 Lack Of Diversity In Genomic Databases Is A Barrier To Translating Precision Medicine Research Into Practice. Health Aff (Millwood) 37**:** 780-785.

Larson, N. B., J. Chen and D. J. Schaid, 2019 A review of kernel methods for genetic association studies. Genet Epidemiol 43**:** 122-136.

Leija-Salazar, M., C. Piette and C. Proukakis, 2018 Review: Somatic mutations in neurodegeneration. Neuropathol Appl Neurobiol 44**:** 267-285.

Locke, A. E., B. Kahali, S. I. Berndt, A. E. Justice, T. H. Pers *et al.*, 2015 Genetic studies of body mass index yield new insights for obesity biology. Nature 518**:** 197-206.

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

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff *et al.*, 2009 Finding the missing heritability of complex diseases. Nature 461**:** 747-753.

Manrai, A. K., B. H. Funke, H. L. Rehm, M. S. Olesen, B. A. Maron *et al.*, 2016 Genetic Misdiagnoses and the Potential for Health Disparities. N Engl J Med 375**:** 655-665.

Marchini, J., L. R. Cardon, M. S. Phillips and P. Donnelly, 2004 The effects of human population structure on large genetic association studies. Nat Genet 36**:** 512-517.

Martin, A. R., M. Kanai, Y. Kamatani, Y. Okada, B. M. Neale *et al.*, 2019 Clinical use of current polygenic risk scores may exacerbate health disparities. Nat Genet 51**:** 584-591.

McDermaid, A., B. Monier, J. Zhao, B. Liu and Q. Ma, 2019 Interpretation of differential gene expression results of RNA-seq data: review and integration. Brief Bioinform 20**:** 2044-2054.

Need, A. C., and D. B. Goldstein, 2009 Next generation disparities in human genomics: concerns and remedies. Trends Genet 25**:** 489-494.

Oh, J. H., S. Kerns, H. Ostrer, S. N. Powell, B. Rosenstein *et al.*, 2017 Computational methods using genome-wide association studies to predict radiotherapy complications and to identify correlative molecular processes. Sci Rep 7**:** 43381.

Okbay, A., J. P. Beauchamp, M. A. Fontana, J. J. Lee, T. H. Pers *et al.*, 2016 Genome-wide association study identifies 74 loci associated with educational attainment. Nature 533**:** 539-542.

Okoniewski, M. J., and C. J. Miller, 2006 Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. BMC Bioinformatics 7**:** 276.

Ormel, J., C. A. Hartman and H. Snieder, 2019 The genetics of depression: successful genome-wide association studies introduce new challenges. Transl Psychiatry 9**:** 114.

Petryszak, R., M. Keays, Y. A. Tang, N. A. Fonseca, E. Barrera *et al.*, 2016 Expression Atlas update--an integrated database of gene and protein expression in humans, animals and plants. Nucleic Acids Res 44**:** D746-752.

Popejoy, A. B., and S. M. Fullerton, 2016 Genomics is failing on diversity. Nature 538: 161-164.

Pruim, R. J., R. P. Welch, S. Sanna, T. M. Teslovich, P. S. Chines *et al.*, 2010 LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics 26: 2336-2337.

Risch, N., and K. Merikangas, 1996 The future of genetic studies of complex human diseases. Science 273: 1516-1517.

Sanders, S. J., X. He, A. J. Willsey, A. G. Ercan-Sencicek, K. E. Samocha *et al.*, 2015 Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. Neuron 87: 1215-1233.

Schmiedel, B. J., D. Singh, A. Madrigal, A. G. Valdovino-Gonzalez, B. M. White *et al.*, 2018 Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression. Cell 175: 1701-1715 e1716.

Scott, R. A., L. J. Scott, R. Magi, L. Marullo, K. J. Gaulton *et al.*, 2017 An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans. Diabetes 66: 2888-2902.

Sherry, S. T., M. Ward and K. Sirotkin, 1999 dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. Genome Res 9: 677-679.

Sirugo, G., S. M. Williams and S. A. Tishkoff, 2019 The Missing Diversity in Human Genetic Studies. Cell 177: 26-31.

Sladek, R., G. Rocheleau, J. Rung, C. Dina, L. Shen *et al.*, 2007 A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 445: 881-885.

Taliun, D., D. N. Harris, M. D. Kessler, J. Carlson, Z. A. Szpiech *et al.*, 2019 Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. bioRxiv.

The GTEx Consortium, 2015 Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 348: 648-660.

Wall, J. D., L. F. Tang, B. Zerbe, M. N. Kvale, P. Y. Kwok *et al.*, 2014 Estimating genotype error rates from high-coverage next-generation sequence data. Genome Res 24: 1734-1739.

Wang, K., S. P. Dickson, C. A. Stolle, I. D. Krantz, D. B. Goldstein *et al.*, 2010 Interpretation of association signals and identification of causal variants from genome-wide association studies. Am J Hum Genet 86: 730-742.

Wang, S., F. Qian, Y. Zheng, T. Ogundiran, O. Ojengbede *et al.*, 2018 Genetic variants demonstrating flip-flop phenomenon and breast cancer risk prediction among women of African ancestry. Breast Cancer Res Treat 168**:** 703-712.

Wang, Z., M. Gerstein and M. Snyder, 2009 RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10**:** 57-63.

Watts, D. J., and S. H. Strogatz, 1998 Collective dynamics of 'small-world' networks. Nature 393**:** 440-442.

Willer, C. J., E. M. Schmidt, S. Sengupta, G. M. Peloso, S. Gustafsson *et al.*, 2013 Discovery and refinement of loci associated with lipid levels. Nat Genet 45**:** 1274-1283.

Wood, A. R., T. Esko, J. Yang, S. Vedantam, T. H. Pers *et al.*, 2014 Defining the role of common variation in the genomic and biological architecture of adult human height. Nat Genet 46**:** 1173-1186.

Wray, N. R., C. Wijmenga, P. F. Sullivan, J. Yang and P. M. Visscher, 2018 Common Disease Is More Complex Than Implied by the Core Gene Omnigenic Model. Cell 173**:** 1573-1580.

Xue, A., Y. Wu, Z. Zhu, F. Zhang, K. E. Kemper *et al.*, 2018 Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. Nat Commun 9**:** 2941.

Yengo, L., J. Sidorenko, K. E. Kemper, Z. Zheng, A. R. Wood *et al.*, 2018 Meta-analysis of genome-wide association studies for height and body mass index in approximately 700000 individuals of European ancestry. Hum Mol Genet 27**:** 3641-3649.

Zajac, G. J. M., L. G. Fritsche, J. S. Weinstock, S. L. Dagenais, R. H. Lyons *et al.*, 2019 Estimation of DNA contamination and its sources in genotyped samples. Genet Epidemiol 43**:** 980-995.

Chapter 2

Whole Genome Sequencing and Analysis of 4,787 Individuals for Age-Related Macular Degeneration

A paper covering most of the material in this chapter is in preparation, with myself as first author.

2.1 Introduction

Age-related macular degeneration, one of the leading causes of blindness in the elderly (FRIEDMAN *et al.* 2004), has been the target of many genetic association studies, leading to ~30 significantly associated loci (MCKAY *et al.* 2011; YU *et al.* 2011; FRITSCHE *et al.* 2013; FRITSCHE *et al.* 2016). In most cases, the specific genetic and molecular functions affected by associated variants are not yet understood (FERRARA AND SEDDON 2015; HUANG *et al.* 2015; KAVANAGH *et al.* 2015). The most significant signals in many of the loci come from common single-nucleotide polymorphisms (SNPs) with no known function (FRITSCHE *et al.* 2014). Recent studies have explored the contribution of rare coding variants to disease, especially those variants found in genes in the complement system (SEDDON *et al.* 2013; ZHAN *et al.* 2013; FERRARA AND SEDDON 2015; KAVANAGH *et al.* 2015; TRIEBWASSER *et al.* 2015). A common theme among past studies is that the molecular mechanisms linking these variants to disease is complex (DU *et al.* 2016), and there remains a gulf between genome association results and biological interpretation.

So far, most studies either focused on genome wide scans of common variants (FRITSCHE *et al.* 2013; MIYAKE *et al.* 2015) or targeted explorations of specific loci (SEDDON *et al.* 2013; RATNAPRIYA *et al.* 2014; HUANG *et al.* 2015; DUVVARI *et al.* 2016). The largest previous study from Fritsche et al. (FRITSCHE *et al.* 2016) used genotyping arrays on 439,350 markers and used an existing population variant panel (1000 Genomes) for imputation to fill in the gaps, adding an additional 11,584,480 variants. Though this approach yields very good results for common variants, it can only find variants present in the array or the reference panel, potentially missing many variants which were not present in the reference panel or genotyping array because they are rare in the general population. Larger reference panels and reference panels that are specifically enriched for disease cases are both expected to improve power for genomewide association studies that rely on imputation.

We set out to more fully characterize AMD-associated variants, including those that could not be studied using the standard array-and-imputation approach. To accomplish this, here we use whole-genome sequencing to enable more comprehensive studies of genetic variation, including both single-variant and gene-based association tests. This allows us to systematically assess both common and rare variants across the genome, with the potential to identify strongly associated rare variants with large predicted effects, such as strongly associated loss-of-function (pLoF) variants (TRIEBWASSER *et al.* 2015; FRITSCHE *et al.* 2016), and to identify sets of candidate functional variants for each common variant signal. By sequencing AMD cases, we improve our chances of finding AMD-associated rare variants not found in previous reference panels and have the opportunity, through imputation, to study these variants in additional genotyped samples. To further increase power to discover gene-based associations due to very rare variants, we compare our variant list and allele counts with those from the Exome Aggregation Consortium (ExAC), (Lek *et al.* 2016), the Genome Aggregation Database (gnomAD) (Karczewski *et al.* 2020), and TOPMed (Taliun *et al.* 2019). Our study shows the advantages and limitations of a low-coverage sequencing study of a complex disease and examines the power of several different approaches to discover genetic association signals.

2.2 Methods

2.2.1 Sample characteristics

Our primary data consisted of whole genome sequence samples for 2,394 AMD cases and 2,393 controls, matched by age and sex, with an average depth of 6x. Their ages ranged from 50 to 101, with an average of 75, with 55% females and 45% males. There were no significant differences in age or sex between cases and controls (**Table 2.1**).

We obtained samples of European ancestry with advanced AMD (defined as subjects diagnosed as having large drusen, choroidal neovascularization [CNV], geographic atrophy [GA], or mixed CNV/GA in at least one eye) from the University of Michigan Kellogg Eye Center, the AREDS1 and AREDS2 studies from the National Eye Institute, and the Scheie Eye Institute at the University of Pennsylvania Perelman School of Medicine. The AREDS1 study included their own matched controls, while additional matched control samples for the other studies were obtained from the Michigan Genomics Initiative (MGI; Table 2.1) as follows: first, we calculated propensity scores for all available samples, using age and sex as covariates and treatment group as outcome; then, we used a 1-to-1 greedy matching algorithm to match each unmatched case to the best-matching MGI control sample (Parsons 2001). The matched pair was removed from the selection pool, and the matching continued iteratively until all cases were matched. Samples from MGI used for matching were restricted to subjects who were not diagnosed with AMD at the time of sample collection. Study participants provided informed consent, and all protocols were reviewed and approved by the University of Michigan Institutional Review Board.

2.2.2 Whole genome sequencing data processing and quality control

We used BWA-MEM to align the genomes using the NCBI RefSeq hg19 human genome reference assembly, GotCloud to call single nucleotide polymorphisms (SNPs) and short insertions and deletions (indels) (Jun *et al.* 2015), and beagle4 (Browning and Browning 2007) for phasing and genotype refinement. We used the called genotypes for all summaries and downstream analyses. To annotate the variants, we used Variant Effect Predictor (VEP), build 84 (Yates *et al.* 2016), using the combined transcripts from both RefSeq and Ensembl as reference.

For quality control, we used LASER (Chaolong Wang *et al.* 2014) to project principal components (PCs) from our variants onto reference samples from the Human Genome Diversity Project (HGDP) (Cann *et al.* 2002) to estimate ancestry. To filter out samples of non-European ancestry, we calculated an acceptance region for samples with European ancestry as follows: first, using the first two PCs, we calculated the mean coordinates for European, African, East Asian, and all HGDP samples; next, we defined a circular region using the mean European HGDP coordinates as the center, with a radius equal to one-quarter the distance from the mean European HGDP coordinates and the mean all-samples HGDP coordinates (Figure 2.1). Thirty-two samples outside of this circular region were removed from downstream analyses. Furthermore, we used VerifyBamID (Jun *et al.* 2012) to estimate sample-level contamination, and excluded 42 samples with estimated contamination above 3%. A subset of our sequenced samples had array genotypes available, which we used to identify sample labeling errors in our sequence data, identifying 4 pairs of sample swaps, one duplicated sample, and 19 other samples with mismatches between sequencing and array data. We relabeled the swapped samples, combined the duplicated samples, and removed the mismatched samples.

2.2.3 Association analyses

We defined a single binary outcome, “advanced AMD”, which included samples with large drusen, geographic atrophy (GA), and/or choroidal neovascularization (CNV). We used the most likely called genotypes in our association tests. Because we used age- and sex- matched case-control samples, we did not include covariates in our association models. We used $P \leq 5 \times 10^{-8}$ as our genome wide significance threshold for all single-variant tests. We tested the advanced AMD cases against control samples using Firth bias-corrected logistic regression using a likelihood ratio test (Ma *et al.* 2013), as implemented in EPACTS (Kang 2012) to perform single-variant GWAS.

Next, to identify statistically independent loci, we used a chromosome-based sequential forward selection approach: first, we identified the most significantly associated variant by *P*-value in each chromosome; if that variant was genome wide significant (*P*-value < 5 x 10^{-8}), then we performed a conditional analysis on all variants in that chromosome by adding the genotypes at that variant as a covariate. If the most significant variant in this analysis was also genome wide significant, then we added it as an additional covariate and repeated the analyses of all variants in that chromosome until no significant variants remained. Each of the variants selected in this sequential procedure was matched to the nearest locus among the 34 AMD-associated loci described by Fritsche et al.(all variants were within 434 kb of one of these previously reported loci).

For gene-based tests, we used SKAT-O (LEE *et al.* 2012) for all predicted nonsynonymous variants with allele frequency less than 5% in each gene. Predicted nonsynonymous variants were defined as any variant annotated as missense, nonsense, frameshift, or essential splice site. We applied the Bonferroni correction for 22,502 tests to get a significance threshold of 0.05 / 22,502 = 2.22 x 10^{-6}.

2.2.4 External data sets

We used several other data sets for gene-based pLoF comparison (detailed below). For external controls, we used pLoF allele count data from ExAC (Lek et al. 2016), consisting of 60,706 exome-sequenced samples; gnomAD (KARCZEWSKI et al. 2020), with 125,748 sequenced exomes and 15,708 sequenced genomes; and TOPMed (TALIUN et al. 2019) (via the BRAVO browser (NHLBI 2018)) with 62,784 sequenced genomes. We also obtained summary data for pLoF allele counts in five genes (CFH, CFI, ORMDL2, SLC16A8, and TIMP3) from Regeneron, with 1,714 AMD cases and 7,356 controls, for a follow-up replication analysis.

2.2.5 Gene-level comparison of rare pLoF variants with the Exome Aggregation Consortium

We annotated both the ExAC variant list and our own variant list using Variant Effect Predictor (VEP) build 84, using the merged RefSeq/Ensembl human database as reference. We generated a subset of rare pLoF variants, defined as variants with an allele frequency of less than 0.1% in our samples and annotated as stop-gained, frameshift, splice donor, or splice acceptor, from our sequenced data. Rare pLoF variants are especially informative for genetic association studies, because they can point to very specific effector genes and disease mechanisms. To increase power, we compared frequencies of rare pLoFs in our case and controls to variant frequencies in ExAC. We have N_{ctrl} = 2,393 total controls, N_{case} = 2,394 total cases, and N_{ExAC} = 60,706 total ExAC samples. One complication is that the ExAC data consists of variant-level allele summaries with no individual-level data. Thus, to tally the number of rare pLoF carriers in each gene, we summed the number of pLoF carriers for each variant in the gene. In this model, we assumed that all alternate alleles were present in different individuals in ExAC due to the rarity of the alleles, so that the number of rare pLoF variants represented our approximation of the number of pLoF carriers. This number could produce a small overestimate, since for some genes the same individual might carry multiple pLoF variants, leading to a larger estimated pLoF carrier count than the number of individuals with a pLoF in those genes. For the *i*th variant in the *j*th gene, we estimated the number of pLoF variant carriers by obtaining allele counts for all rare pLoF variants in three sets of samples—our controls ($A_{ctrl,i,j}$), our cases ($A_{case,i,j}$), and ExAC ($A_{ExAC,i,j}$); next, we summed up the total number of alleles in each of our three cohorts as our estimate of the number of pLoF carriers (c) for the *j*th gene:

$$c_{cohort,j} = \sum_{i} A_{cohort,i,j} \ for\ cohort \in \{ctrl, case, ExAC\}$$

We filtered out all genes for which $c_{cohort,j} = 0$ in all cohorts. We then calculated the number of non-carriers (w) in the *j*th gene for each cohort:

$$w_{cohort,j} = N_{cohort} - \ c_{cohort,j}\ for\ cohort \in \{ctrl, case, ExAC\}$$

With these estimates, we created two 2x2 contingency tables for allele carriers vs. non-carriers: controls vs. ExAC ($w_{ctrl,j}$ and $c_{ctrl,j}$ vs. $w_{ExAC,j}$ and $c_{ExAC,j}$) and cases vs. ExAC ($w_{case,j}$ and $c_{case,j}$ vs. $w_{ExAC,j}$ and $c_{ExAC,j}$).

Since our sequenced data had a lower average coverage than ExAC, we expected fewer rare pLoF alleles per sample to be called in our data compared to ExAC, leading to a lower pLoF carrier frequency. Therefore, while a higher pLoF carrier frequency in our cases compared to ExAC would provide evidence to suggest an association, we do not attempt to interpret situations where our sample shows fewer pLoF carriers than ExAC. We calculated *P*-values for these contingency tables using a one-sided Fisher's Exact Test ($PFET_{A\text{-}vs\text{-}B,j}$), representing the probability of our tested cohort (AMD controls or AMD cases) having a lower pLoF carrier frequency compared to ExAC, given a fixed total number of pLoF carriers between the tested cohort and ExAC. A significant *P*-value is therefore evidence supporting a higher pLoF carrier frequency in our tested cohort compared to ExAC. Due to differences in sequencing method, sequencing depth, variant calling algorithms, and experimental conditions, we restricted our comparison to only those genes for which the distribution of pLoF alleles in our sequenced controls was not significantly higher than that in ExAC.

We performed a simulation study by varying effect size, carrier frequency, and *P*-value threshold to characterize the balance between false discovery rate and power under different thresholds. We found that the less stringent threshold of 0.01 had good power to discover gene-level associations across a variety of allele frequencies and effect sizes while maintaining a low false discovery rate (Table 2.2). We retained the subset of genes found in IAMDGC risk loci for which $PFET_{CtrlExAC,j} > 0.01$, representing the set of genes in which the pLoF carrier frequencies in our control samples were not statistically significantly higher than the pLoF carrier frequencies in ExAC. Following these results, we filtered the gene subset for those with $PFET_{CaseExAC,j} < 0.01$.

To test the validity of using pLoF allele counts as an estimate of pLoF carrier counts, we estimated both the proportion of genes in which at least one sample had multiple pLoF variants and the proportion of samples containing multiple rare pLoF variants in a single gene in our sequencing data. Starting with 21,351 genes, we identified a subset of 8,222 genes in which our AMD cases or AMD controls had at least one rare pLoF variant, where “rare” was defined as any given variant having an allele frequency of less than 0.1% (in our samples, this translated to having an allele count of 9 or fewer out of 4,787 x 2 = 9,574 total alleles). Of these 8,222 genes, we observed two or more rare pLoF variants in the same individual in 59 out of 4,787 samples (1.2%), distributed across 26 genes (0.3% of genes). None of these genes are found in our list of potentially associated genes.

2.2.6 Replication analysis of gene-level rare pLoF alleles from Regeneron samples

We repeated our rare pLoF association study using samples from Regeneron for the genes in which we found a difference between cases and controls in our ExAC comparison, along with genes which contained previously-discovered large-effect pLoF variants, totaling five genes: *CFH, CFI, ORMDL2, SLC16A8,* and *TIMP3*. This data set contained 1,714 cases and 7,356 controls. The AMD phenotypes for these samples were predicted using the eMERGE Network EMR Phenotype Algorithm(WEI AND DENNY 2015). We used Fisher's Exact Test to compare rare pLoF variants in the cases vs. controls within the Regeneron samples, where rare was defined as an allele frequency less than 1%. We also compared the rare pLoF variants in Regeneron cases and controls to ExAC in the same way we compared our AMD WGS sequencing samples to ExAC above: for each gene, we compared controls to ExAC first to determine whether the two samples were different, then compared cases to ExAC if the controls were not statistically different from ExAC.

2.3 Results

2.3.1 Variant calling

From our initial data set of 4,869 samples, we removed samples with over 3% contamination (20 samples), samples of non-European ancestry (32 samples), samples which did not match array genotypes (11 samples), and samples which did not conform to our disease criteria (19 samples). Our final data set contained 4,787 samples, with 2,394 cases and 2,393 controls. We were successful in finding a large number of rare variants: our final variant call set contained 46,946,619 variants in all, of which 43,596,300 were SNPs and 3,350,319 were indels. About three-quarters of variants had an allele frequency below 0.5%, and about half of all variants were either singletons or doubletons (among SNPs: 17,157,068 singletons, 5,129,082 doubletons, and 11,320,032 others with allele frequencies less than 0.5%; among indels, 692,506 singletons, 267,158 doubletons, and 974,289 others with allele frequencies less than 0.5%; **Table 2.3**). We discovered 27,352,890 variants not found in dbSNP build 138 (SHERRY *et al.* 2001). Among these, over 98% had an allele frequency of less than 0.5%, and over 73% were either singletons or doubletons. This substantial increase in rare variants compared to the previous largest AMD GWAS (FRITSCHE *et al.* 2016) was due to our use of whole genome sequencing: of our 46.9 million variants, 37,090,168 (79%) had an allele frequency of <1%; of the 12,023,830 variants in IAMDGC, 3,050,013 (25%) had an allele frequency of <1%. Our study discovered 31,584,812 autosomal variants not found by IAMDGC: 13,627 pLoF, 174,611 nonsynonymous, 112,050 synonymous, and 31,284,524 intronic and intergenic variants.

Our final data set contained 301,288 nonsynonymous variants, 165,161 synonymous variants, 19,276,755 intronic variants, and 27,203,395 variants that belonged to other categories (including intergenic, upstream, downstream, and untranslated region variants). Of our exonic variants, 282,834 were missense, 7,043 were nonsense, 6,805 were frameshifts, and 4,606 were essential splice site variants (Table 2.4). Most of these were very rare: 92.2% of stop gains and 81.7% of other pLoF variants had an allele frequency of less than 0.5% (Table 2.5).

2.3.2 Single-variant association tests

Using a genome wide significance threshold of 5 x 10^{-8}, we found 1,854 significant variants in our single-variant tests, located in four known loci with very strong association signals in previous studies (*CFH*, *C2/CFB/SKIV2L*, *ARMS2/HTRA1*, and *C3*; Table 2.6 and Figure 2.2). There was no evidence of population stratification (genomic control λ_{GC} = 1.021, Figure 2.3). Sequential forward selection led to 4 significant independent signals in the *CFH* locus, 2 in *C2/CFB/SKIV2L*, 2 in *C3*, and 1 in *ARMS2/HTRA1* (Table 2.7).

As was found in past studies, the most significant signals we discovered in each locus were generally common variants not predicted to be functionally disruptive in the translated protein, except in *C3*: our top variant in the *C3* locus was the previously-known nonsynonymous top variant (rs2230199, 19:6718387 G>C, *C3* Arg102Gly, case allele frequency = 0.28, control allele frequency = 0.20, odds ratio = 1.55, P-value 2.6 x 10^{-2}). This was not true for the other 3 loci: in the *CFH* locus, the most significant signal was rs6688272 (1:196684392 G>T, case allele frequency = 0.20, control allele frequency = 0.42, odds ratio = 0.36, P-value = 4.6 x 10^{-114}), an intronic variant in very high linkage disequilibrium (LD) with the most significant signal in the same locus from Fritsche et al (rs10922109) in our samples (R^2 = 0.995; Table 2.6). Similarly, the top variant in the *ARMS2/HTRA1* locus, rs144224550 (10:124214600 G>GGT, case allele frequency = 0.41, control allele frequency = 0.22, odds ratio = 2.44, P-value = 2.1 x 10^{-92}), was also in high LD with the top variant for the same locus in Fritsche et al. (rs3750846, R^2 = 0.994).

While prior results show that signals in the *C2/CFB/SKIV2L* locus could be explained by two coding variants in *CFB*, our more complete dataset highlighted an independent signal in *C2*. Our top variant in the *C2/CFB/SKIV2L* locus, rs556679 (6:31894355 C>T, case allele frequency = 0.051, control allele frequency = 0.11, odds ratio = 0.48, P-value = 4.0 x 10^{-25}), which was previously identified by Zhan et al. as exhibiting strong evidence of association, was only in moderate LD (R^2 = 0.659) with the known top variant in Fritsche et al. (rs116503776, 6:31930462 G>A, intronic in *SKIV2L*). Whereas previous studies suggested that variant associations in *C2* could be explained by LD with causal variants in *CFB* (GOLD *et al.* 2006; MALLER *et al.* 2006; FRITSCHE *et al.* 2016), our haplotype analysis showed that rs556679 was associated with AMD independent of the two known coding variants in *CFB* (Table 2.8).

We tested the robustness of our results by repeating our single-variant association tests using the first two principal components of each sample as covariates, analyzing only GA or CNV samples, and performing the analysis on the subset of 2,300 pairs of matched samples only. The results largely agreed with our original single-variant analysis and did not change the discovered loci or significant association signals (results not shown).

2.3.3 Gene-based association tests

We used SKAT-O to test for evidence of an excess of rare nonsynonymous alleles in our samples. Grouping together nonsynonymous variants with allele frequencies below 5% and using a Bonferroni correction for 22,502 tests to get a *P*-value significance threshold of 0.05/22,502 = 2.22 x 10^{-6}, 2 genes in the *CFH* locus, 4 genes in the *C2/CFB/SKIV2L* locus, and 2 genes in the *C3* locus, including *NRTN,* showed significant associations (**Table 2.9**). Thus, our results show how, through linkage disequilibrium, multiple genes in the same locus can show a rare variant association signal. After conditioning on the top single variant signals in *CFH* and *C2/CFB/SKIV2L*, the SKAT-O signals for these genes disappeared. The two signals in *C3* remained after conditioning on the top variant in that locus, and the signal for *NRTN* remained after conditioning on the top two variants, indicating that the *NRTN* signal might be independent of the signal in *C3.* In Fritsche et al., one of the three independent signals in the *C3* locus (rs12019136, 19:5835677 G>A, intronic) was found to be near *NRTN/FUT6*, and their nonsynonymous variant enrichment analysis yielded a 95% credible set for this locus, which included a variant in *NRTN* (rs79744308, 19:5827765 G>A, p.Ala59Thr) as one of two signals with over 5% posterior probability of being causal (along with a variant in *FUT6*). This same variant had the strongest signal in our single-variant conditional analysis in the *C3* locus after conditioning on the first two independent signals, though it did not pass the threshold for genome wide significance (conditional $P = 5.85 \times 10^{-7}$, OR = 0.59, case AF = 0.031, control AF = 0.051). The SKAT-O signal for *NRTN* disappeared when we conditioned on the *NRTN* nonsynonymous variant rs79744308 ($P = 2.8 \times 10^{-7}$ when not conditioning on rs79744308; $P = 0.32$ when conditioning on rs79744308), suggesting the variant is a major contributor to the signal.

To verify our SKAT-O results, we repeated our gene-based analysis using the variable threshold (VT) burden test (PRICE *et al.* 2010) as implemented in EPACTS using a 5% threshold. Only two genes remained significantly associated with AMD by the Bonferroni-corrected *P*-value threshold of $0.05/22{,}502 = 2.22 \times 10^{-6}$: *C2* ($P = 7.0 \times 10^{-7}$) and *NRTN* ($P = 4.0 \times 10^{-7}$). To test the sensitivity of the analysis with respect to the variant threshold, we repeated the SKAT-O and VT analysis at the AF<1% level.

The only gene that remained significant in the 1% SKAT-O analysis was *C3* (*P* = 1.7 x 10^{-11}); no genes were significant in the 1% VT analysis.

2.3.4 Rare pLoF rate comparison with a large external panel (ExAC)

To further test our rarer variants for enrichment in cases, we performed an analysis restricted to very rare (allele frequency < 0.1% in our samples) pLoF variants and aggregated results at the gene level. We performed this analysis by comparing our data against the variant list from the Exome Aggregation Consortium (ExAC) (LEK *et al.* 2016), with 60,706 population-based sequenced samples.

Using estimated rare pLoF carrier counts, we generated contingency tables using pLoF carrier counts and case-control status. First, restricting our analysis to genes found in the 34 known risk loci from IAMDGC removed 20,538 out of 22,502 genes, narrowing the list down to 845 genes. Next, we removed genes with no rare pLoF alleles in both AMD cases and AMD controls, which accounted for an additional 538 genes, leaving us with 307. We kept only genes for which the control samples and ExAC samples were not significantly different, using *P* > 0.01 as the criterion, which removed 249 additional genes, with 289 remaining. We narrowed this list down further to include only genes in which the case samples and ExAC samples were significantly different, using P < 0.01 as the criterion, removing 279 genes, leaving 10 genes. We filtered this list of genes for those in which the case pLoF allele frequency was greater than the control pLoF allele frequency in our sequenced samples, which suggested an association of deleterious rare pLoF alleles with AMD. This removed 7 more genes, leaving us with 3 genes with evidence of association (Table 2.10).

Two of these three genes (*CFH* and *CFI*) had been identified in previous studies to contain rare coding variants (including missense and pLoF variants) associated with AMD (SEDDON *et al.* 2013; FERRARA AND SEDDON 2015; KAVANAGH *et al.* 2015; TRIEBWASSER *et al.* 2015). In our data, five out of 2,394 case samples each carried a pLoF allele in *CFH* (a pLoF rate of about 1 in 500), distributed across four different variants, while none of our 2,393 control samples carried any alternate alleles in the same variants. Similarly, five case samples each carried a pLoF allele in *CFI*, distributed across three different variants (about 1 in 500), while no control samples carried any alternate allele. In comparison, ExAC contains nine pLoF allele carriers in *CFH* (about 1 in 7,000) and 30 in *CFI* (about 1 in 2,000). The third gene, *ORMDL2*, was found in the *RDH5/CD63* locus, and no previous study had examined the effect of rare variants in this locus on AMD. The credible set analysis for the *RDH5/CD63* locus in Fritsche et al. did not find any nonsynonymous variant likely to be causal within this locus, with the lead variant (rs3138141) lying downstream of *BLOC1S1*, and the only exonic variant being a synonymous SNP in *RDH5* (rs3138142, p.Ile14=) (FRITSCHE *et al.* 2016). There was a significant difference in pLoF allele frequencies in *ORMDL2* between our cases and the large population-based samples from ExAC based on pLoF variants, but not for controls vs. ExAC (*P*-value = 3.9 x 10^{-3} and 0.79, respectively): three of our case samples carried a pLoF allele in *ORMDL2* (about 1 in 800) with none found in our control samples, while ExAC contains six pLoF alleles in the same gene (about 1 in 2,000).

2.3.5 Replication study of rare pLoF carrier frequencies using Regeneron samples

Comparing 1,714 cases with 7,356 controls in the Regeneron samples, and comparing pLoF carrier rates for variants with an allele frequency <1% in these samples, we replicated our rare pLoF signal in *CFH*, with a significant difference between cases and controls in *CFH* (case allele frequency = 4.1 x 10^{-3}, control allele frequency = 2.7 x 10^{-4}, *P*-value = 2.1 x 10^{-4}, OR = 15.1, 95% CI = 2.9-148.7; Table 2.11). Only two of the five tested genes (*CFH* and *SLC16A8*) contained rare pLoF variants in the Regeneron samples (7 and 39 case carriers, and 2 and 139 control carriers, in *CFH* and *SLC16A8*, respectively). We did not find evidence for a significant difference between cases and controls in the other four genes (*CFI*, *ORMDL2, SLC16A8,* and *TIMP3*) in the Regeneron samples.

2.4 Discussion

Our sequencing sample represents the largest whole-genome sequencing study for age-related macular degeneration to date. We were able to discover a large number of novel rare variants across the whole genome in a set of samples enriched for AMD cases, providing a valuable resource for future AMD studies. Because our data spanned the entire genome, it may be useful for future studies involving regions outside the exome: for example, past studies have found regulatory roles for non-coding RNA in the complement pathway (LUKIW *et al.* 2012; ZOU *et al.* 2016). Using our disease-specific panel to accurately impute rare non-exonic variants may help in the discovery of regulation-specific variants in future AMD association studies. A subset of our samples was submitted to the Haplotype Reference Consortium (MCCARTHY *et al.* 2016), allowing future GWA studies to impute AMD-specific variants into their samples using a web-based imputation service (DAS *et al.* 2016). We have contributed rare variant data to other AMD studies (AL-KHERSAN *et al.* 2018; PIETRASZKIEWICZ *et al.* 2018).

In our single variant association results, we replicated the most significant signals from IAMDGC. We found evidence supporting results from past studies which found clinically significant variants in both *CFB* and *C2* (SUN *et al.* 2012; THAKKINSTIAN *et al.* 2012), and that most significant signals in both of these genes were often in high linkage disequilibrium with each other (SPENCER *et al.* 2007).

Our results from grouped association tests generally recapitulated previous results. The elimination of our SKAT-O signals in *CFH* and *C2/CFB/SKIV2L* after conditioning on the top variants in each locus provides support for the hypothesis that the top single-variant signals served to tag a larger set of rare nonsynonymous variants in linkage disequilibrium with those top variants. Similarly, in *NRTN*, the elimination of our SKAT-O signal after conditioning on rs79744308 (*NRTN* p.Ala59Thr) indicates that the SKAT-O signal was largely driven by this variant. The nonsynonymous nature of this variant, along with previous mouse studies which found functional relationships between *NTRN* and neuron development and activity in the retina (JOMARY *et al.* 1999; HARADA *et al.* 2003; SONG *et al.* 2003; JOMARY *et al.* 2004; BRANTLEY *et al.* 2008; HOOVER *et al.* 2014), suggest that rs79744308 in *NRTN* may have a protective effect on AMD independent of the effects of variants in *C3*.

Using a large publicly available data set as an external panel, we were able to further improve our association analysis. Our analysis of very rare pLoF variants provides us with association evidence not found by other methods. We found confirmatory signals in *CFH* and *CFI*. The signal in *CFI* was especially interesting, since we could not find significant signals in this gene otherwise in either single-variant or grouped association tests. These results agree with previous studies, which found pLoF variants in *CFH* to be significantly associated with early-onset macular drusen, a severe AMD subtype (Taylor *et al.* 2019), while pLoF variants in *CFI* were associated with advanced AMD (Kavanagh *et al.* 2015). Additionally, we found a similar signal in *ORMDL2*, whose function is still not well understood, though a study using knockout mice suggests that the *Ormdl* protein family may play an important part in preventing damage to the nervous system (Clarke *et al.* 2019), making *ORMDL2* a potentially interesting new gene for future AMD studies.

Our findings provide several insights for designing and performing future sequencing studies. First, the use of well-matched cases and controls can greatly reduce the potential problem of population stratification. In our study, we were aggressive about matching cases and controls on age, sex, and ancestry, which led to a relatively low genomic control ($\lambda_{GC} = 1.021$). The consistency of our results after replicated analyses using traditional controls for stratification showed that our strategy helped us greatly reduce noise from population structure.

Second, the limitations of traditional association methods highlight the importance of a more holistic approach in interpreting GWAS results. For example, our SKAT-O test of nonsynonymous variants showed that there was no additional significant gene-based signal in the *C2/CFB/SKIV2L* locus after conditioning on the top variant, rs556679, yet our haplotype analysis of the same locus showed that there was evidence of additional haplotype effect on disease beyond the top variant even after controlling for two known nonsynonymous variants. This suggests the real risk factors may involve more complex interactions between different variants, including but not limited to nonlinear interaction effects between multiple variants, including enhancers, promoters, and variants not called by the sequencing pipeline. As another example, the significant SKAT-O signal in *NRTN* remained after conditioning on the only significant ($p < 5 \times 10^{-8}$) independent variants in the *C3* locus. The signal disappeared after conditioning on rs79744308, a nonsynonymous variant in *NRTN* with a single-variant association *p*-value of 9.3×10^{-7}. Though this variant was not genome wide significant, it was responsible for driving the entire *NRTN* SKAT-O signal, suggesting a possible biological connection. Identification of this variant as potentially interesting was only possible after combining the results from both the single-variant and grouped tests.

Third, to supplement traditional association techniques, we leveraged of a large, publicly available, population-based variant list to supplement association tests within our sequenced samples. In studies with modest samples sizes, both single-variant tests and gene-based tests will be underpowered to find associations. This will be especially true for rare variants. To partially mitigate this problem, we chose to prioritize genes within known risk loci and focused on extremely rare predicted loss-of-function variants, ones with major, well-defined effects on the translated products, with a better chance of leading to biological explanations for genetic associations. Though individual genotypes usually cannot be shared due to privacy concerns, variant lists with allele counts are often available to the public. The availability of the ExAC variant list and our focus on rare pLoF variants led to our discovery of a difference in loss-of-function allele frequencies between our cases and the samples in ExAC in the *ORMDL2* gene, suggesting a possible functional role for this gene in AMD. Additionally, though we had very few rare pLoF alleles in each gene, by targeting genes in known loci and using a less stringent FDR threshold, our method was able to recover the known *CFI* signal that SKAT-O was underpowered to detect. To test the robustness of our approach, we repeated our pLoF analysis with the larger gnomAD dataset (Karczewski et al. 2020). The 141,456 gnomAD samples had 32 pLoF alleles in *CFH* (~1 in 4,400), 100 in *CFI* (~1 in 1,400), and 28 in *ORMDL2* (~1 in 5,000), all lower than in our AMD cases (~1 in 500 in *CFH*, ~1 in 500 in *CFI*, and ~1 in 800 in *ORMDL2*). The difference in *CFH* remained significant (P = 3.6 x 10^{-4} using the Exact Test), while the differences in *CFI* and *ORMDL2* were attenuated (P = 0.031 and 0.015, respectively). The same comparison with pLoF alleles in TOPMed data led to similar results (~1 in 4,000 pLoF alleles in *CFH*, ~1 in 900 in *CFI,* and ~1 in 2,700 in *ORMDL2*). We note that our use of low-pass sequencing limited our ability to discover and call rare pLoF variants in our samples, likely leading to an underestimation of pLoF counts in our cases. Moreover, our ability to replicate the *CFH* signal in an independent data set from Regeneron was an encouraging sign of the viability of our approach. As the number of sequencing studies increases, our new group-based approach will be able to leverage new sources of data to uncover associations that could not be detected using traditional methods.

Finally, our study highlights the importance of high-quality phenotypes for genomic studies. The 2,394 cases in our study had diagnoses from ophthalmologists for advanced AMD, and our single-variant association tests confirmed four previously discovered loci with highly significant results. In comparison, the UK Biobank (SUDLOW *et al.* 2015) contained a phenotype with similar power (2,524 cases and 106,293 controls), “6148_5”, which consisted of all respondents who answered “Eye problems/disorders: Macular degeneration” to the question “Has a doctor told you that you have any of the following problems with your eyes?”, obtained via a self-reported touchscreen questionnaires. The association tests for this phenotype was only able to confirm the top two loci, with greatly reduced significance for the top signals in each locus (p_{CFH} = 4.6 x 10^{-114} vs. 1 x 10^{-18} and p_{ARMS2} = 2.1 x 10^{-92} vs. 1.9 x 10^{-24} for our study vs. UK Biobank, respectively). The high quality of our phenotypes, along with age- and sex-matching cases and controls and filtering for ancestral background, may have improved our study’s ability to discover associations.

Our study had a few limitations which decreased our ability to find associations. First, our sample size was relatively modest compared to IAMDGC, decreasing our power to discover associations. Second, low sequencing depth (~6x) means we may have been unable to call millions of very rare variants from the sequence data. Third, using only samples with European ancestry means our results are not easily generalizable to other populations. Fourth, the allele frequency threshold in our replication study using samples from Regeneron was higher than that used in our initial analysis with ExAC (1% instead of 0.1%), which might lead to overestimating the number of carriers, leading to larger Type I errors in gene-level carrier tests and adding noise to the results. Finally, the differences in sequencing depths, variant calling algorithms, and phenotype definition between our case-control data, ExAC, and Regeneron samples increased the difficulty to jointly analyze variants in these data sets. Despite these limitations, we were able to expand on results from previous studies and highlight potentially novel gene-level associations.

Table 2.1 Sample summary

In our association analysis, the cases have been combined for comparisons against controls. P-value for age was obtained via a two-sample t-test; P-value for male proportion was obtained using Pearson's Chi-squared test with Yates' continuity correction

AMD Status	Controls	Cases	*P*-value
None	2,393		-
Large Drusen		583	-
Geographic Atrophy (GA)		419	-
Choroidal Neovascularization (CNV)		1,122	-
Mixed GA+CNV		270	-
Total	2,393	2,394	-
Age, mean (range)	74.9 (50.0-94.2)	75.1 (50.4-101.0)	0.49
Males (%)	45.2	44.9	0.86

Table 2.2 Power simulation for Fisher's exact test using external controls

Studies were simulated with pLoF carrier frequencies of 0.0001, 0.0005, and 0.001; effect sizes (odds ratios, "OR") of 1 (no effect), 5, 10, and 30; with 2,394 cases, and either 2,394 (1:1 matched) or 60,706 (external ExAC) controls ("Case-Control" and "Case-ExAC", respectively), with 1,000 simulated studies of 845 genes for each combination of carrier frequency and odds ratio. Powers were calculated either with a Bonferroni-corrected threshold of P < 0.05 / 845 = 5.9 x 10^{-5} (for testing 845 genes in the 34 known IAMDGC loci), or with P < 0.01. We also present examples of genes with pLoF carrier frequencies close to our simulation values, as estimated from pLoF allele counts in ExAC. There is substantial power gain by using external controls when the carrier frequency is low but the effect size is high, as would be expected for loss-of-function variants

		P < 5.9 x 10^{-5}		*P* < 0.01		Example Gene
pLoF carrier freq.	OR	Case-Control	Case-ExAC	Case-Control	Case-ExAC	(ExAC pLoF carrier freq)
1.0 x 10^{-4}	1	0	1.5 x 10^{-5}	0	0.0032	*CFH* (7.4 x 10^{-5})
	5	0	0.008	2.8 x 10^{-5}	0.13	
	10	0	0.087	0.0025	0.42	
	30	0.0024	0.82	0.36	0.97	
5.0 x 10^{-4}	1	0	1.2 x 10^{-5}	9.1 x 10^{-5}	0.0049	*ARMS2* (4.6 x 10^{-4})
	5	1.3 x 10^{-4}	0.19	0.094	0.64	
	10	0.06	0.86	0.62	0.99	
	30	0.996	1.00	1.00	1.00	
0.001	1	0	2.4 x 10^{-5}	0.0013	0.0056	*MMP9* (8.1 x 10^{-3})
	5	0.024	0.57	0.40	0.92	
	10	0.62	0.998	0.97	1.00	
	30	1.00	1.00	1.00	1.00	

Table 2.3 Variant counts across all samples by allele frequency

About half of all called variants were singletons or doubletons, and about three-quarters had an allele frequency of less than 0.5%. A majority of our variants that are in dbSNP (build 138) were common: 54.5% of them had an allele frequency of 0.5% or higher, and 83.8% of them had at least three copies of the allele in our samples. Meanwhile, the vast majority of our variants outside of dbSNP are very rare: over 98% of them have an allele frequency of less than 0.5%, and about 73.4% of them are either singletons or doubletons.

	All Variants		SNPs		Indels		in dbSNP		outside dbSNP	
Allele frequency	Count	(%)	Count	(%)	Count	(%)	Count	(%)	Count	(%)
Singletons	17,849,574	38.0	17,157,068	39.4	692,506	20.7	1,970,011	10.2	15,879,563	58.1
Doubletons	5,396,240	11.5	5,129,082	11.8	267,158	8.0	1,189,381	6.1	4,206,859	15.4
[Tripleton, 0.5%)	12,494,321	26.6	11,520,032	26.4	974,289	29.1	5,751,081	29.7	6,743,240	24.7
[0.5%, 1.0%)	1,350,033	2.9	1,195,594	2.7	154,439	4.6	1,165,205	6.0	184,828	0.7
[1.0%, 5.0%)	2,873,308	6.1	2,415,580	5.5	457,728	13.7	2,654,177	13.7	219,131	0.8
[5.0%, 100%)	6,983,143	14.9	6,178,944	14.2	804,199	24.0	6,863,874	35.4	119,269	0.4
Total	**46,946,619**	**100**	**43,596,300**	**100**	**3,350,319**	**100**	**19,393,729**	**100**	**27,352,890**	**100**

Table 2.4 All variants discovered, by annotation and allele frequency

Variants were annotated using Variant Effect Predictor (VEP) build 84, using the merged RefSeq/Ensembl reference for humans. For variants with multiple annotations, the most severe consequence was used. Allele frequencies were defined as allele counts divided by the total number of alleles.

Allele frequency	Nonsense	Frameshift	E.Splice [a]	Nonsyn [b]	Synonymous	Intron	Intergenic and others [c]
Singletons	3,987	2,747	2,420	131,532	65,331	7,403,632	10,239,925
Doubletons	967	777	552	37,264	20,515	2,230,800	3,105,365
[Tripleton, 0.5%)	1,537	1,804	1,027	74,223	45,331	5,157,955	7,212,444
[0.5%, 1.0%)	109	199	107	6,773	4,665	556,154	782,026
[1.0%, 5.0%)	172	485	164	11,810	8,945	1,180,033	1,671,699
[5.0%, 100%)	271	793	336	21,232	20,374	2,748,201	4,191,936
Total	7,043	6,805	4,606	282,834	165,161	19,276,775	27,203,395

[a] splice acceptor and splice donor; [b] missense, start loss, stop loss, inframe deletion, and inframe insertion; [c] includes 5-prime UTR, 3-prime UTR, upstream, downstream, and other non-coding variants that do not belong in the other listed categories

Table 2.5 Nonsynonymous variants by consequence type and allele frequency

Annotated consequences were generated by Variant Effect Predictor (build 84). The majority of the nonsynonymous variants were very rare: 85.9% of all nonsynonymous variants (92.2% of all stop gains, 81.7% of other pLoF, and 85.9% of other nonsynonymous variants) had an allele frequency of less than 0.5%.

Allele frequency	All nonsyn.	(%)	Stop gain	(%)	Other pLoF [a]	(%)	Other nonsyn. [b]	(%)
Singleton	140,686	46.7	3,987	56.6	5,167	45.3	131,532	46.5
Doubleton	39,560	13.1	967	13.7	1,329	11.6	37,264	13.2
[Tripleton, 0.5%)	78,591	26.1	1,537	21.8	2,831	24.8	74,223	26.2
[0.5%,1.0%)	7,188	2.4	109	1.5	306	2.7	6,773	2.4
[1.0%,5.0%)	12,631	4.2	172	2.4	649	5.7	11,810	4.2
[5.0%, 100%)	22,632	7.5	271	3.8	1,129	9.9	21,232	7.5
Total	301,288	100.0	7,043	100.0	11,411	100.0	282,834	100.0

[a] Frameshift variants and essential splice variants. [b] Missense variants, in-frame insertions, in-frame deletions, start-loss variants, and stop-loss variants.

Table 2.6 Single variant association tests for advanced AMD

We compared 2,394 cases of AMD—including large drusen, CNV, GA, and mixed GA/CNV—to 2,393 controls using Firth bias-corrected logistic regression. Alleles indicate the major and minor alleles for the given variants. P-values were obtained from Firth bias-corrected logistic regression. IAMDGC top variant indicates the top variant in the given locus in Fritsche et al. (Fritsche et al. 2016). R^2 with IAMDGC indicates the genotype R^2 between our top variant and the top variant for the same locus in IAMDGC, as found in our data. Our most significant signal in the *C2* locus is only in moderate linkage disequilibrium with the top variant found in IAMDGC: in the previous study, the top variant (rs116503776) was found in a haplotype analysis to tag two previously-described *CFB* missense variants, which appear to be the risk-carrying variants (Supplementary Figure 4 of Fritsche et al. 2016). In the *C3* locus, our most significant signal was rs2230199 (c.304C>G, p.Arg102Gly), the same one found in IAMDGC and previous studies.

Chr	Top var. pos.	Alleles	Locus	*P*-value	Annotation	IAMDGC top variant	R^2 w/ IAMDGC
1	196,684,392	G/T	*CFH*	4.62×10^{-114}	Intron-*CFH*	196,704,632_C/A (rs10922109)	0.995
6	31,894,355	C/T	*C2/CFB/SKIV2L*	4.04×10^{-25}	Intron-*C2*	31,930,462_G/A (rs116503776)	0.659
10	124,214,600	G/GGT	*ARMS2/HTRA1*	2.13×10^{-92}	Intron-*ARMS2*	124,215,565_T/C (rs3750846)	0.994
19	6,718,387	G/C	*C3*	2.55×10^{-20}	Nonsyn-*C3*	6,718,387_G/C (rs2230199)	-

Table 2.7 Sequential forward selection single variant association tests for advanced AMD

Sequential forward selection results for 2,394 AMD cases and 2,393 controls. Firth bias-corrected single-variant association. For association results in each locus, the top-most variant did not use any variants as covariates; each subsequent variant uses all variants above it in the same locus as covariates. The variants below represent statistically independent signals across the entire genome. Locus names correspond to names given to the equivalent loci in IAMDGC.

Lead variant			Major/minor		Allele Frequencies		Association results	
RS number	Chr	Position	Alleles	Locus name	Cases	Controls	OR	Cond. P-value
rs6688272	1	196,684,392	G/T	*CFH*	0.201	0.418	0.36	2.90×10^{-113}
rs10922094	1	196,661,505	G/C		0.388	0.619	0.6	1.03×10^{-20}
-	1	196,024,122	TG/T		0.00497	0.000209	32.8	2.06×10^{-10}
rs79436252	1	196,358,288	A/G		0.0655	0.0224	1.95	5.72×10^{-9}
rs556679	6	31,894,355	C/T	*C2/CFB/SKIV2L*	0.051	0.109	0.48	2.05×10^{-25}
rs28383438	6	32,609,038	C/T		0.166	0.123	1.52	5.81×10^{-13}
rs144224550	10	124,214,600	G/GGT	*ARMS2/HTRA1*	0.414	0.217	2.44	5.08×10^{-91}
rs2230199	19	6,718,387	G/C	*C3*	0.281	0.2	1.55	2.58×10^{-20}
rs181290250	19	6,722,565	C/T		0.0157	0.00251	6.9	3.46×10^{-14}

Table 2.8 Haplotype analysis in the *CFB* locus

We examined six different haplotypes that included our top variant in the *CFB* locus (31894355 C>T) along with two known nonsynonymous variants in the *CFB* gene to determine whether our observed variant had an effect on AMD independent of known variants. The odds ratio for haplotype H4, which contained the alternate allele for our top variant but the reference allele for both nonsynonymous variants, had attenuated effect compared to haplotypes with the alternate allele in either of the nonsynonymous variants, but was still significantly different from the null, suggesting that our top variant may be in high linkage disequilibrium with other variants causal for AMD. Odds ratios and *P*-values were calculated by comparing the case and control haplotype counts for H1 against the case and control haplotype counts for each of the alternate haplotypes (H2-H6) using Fisher's exact test.

	Haplotype			Haplotype Frequency (%)			
	rs641153	rs4151667	rs556679				
Haplotype No.	p.Arg32Gln	p.Leu9His	31894355_C/T	Cases	Controls	OR	*P*-value
H1	G	T	C	92.4	83.9	Reference	
H2	A	T	T	2.6	5.5	0.43	2.55E-15
H3	A	T	C	1.7	3.7	0.42	4.46E-11
H4	G	T	T	1.4	2.6	0.51	7.11E-06
H5	G	A	T	1.1	2.8	0.36	3.74E-11
H6	G	A	C	0.8	1.5	0.47	0.000167

Table 2.9 Gene-based tests on nonsynonymous variants for advanced AMD using SKAT-O

Significant signals from gene-based associations of 2,394 AMD cases vs 2,393 controls from SKAT-O were found in multiple genes in three of the four significant loci found in single-variant association tests. We used a Bonferroni correction for 22,502 tests to obtain a significance threshold of 2.22 x 10^{-6}. No significant gene-based signal was found in the *ARMS2/HTRA1* locus. Conditioning on the top variants in *CFH* (rs6688272, 1:196684392 G/T) and *C2/CFB/SKIV2L* (rs556679, 6:31894355 C/T) eliminated the SKAT-O signals in those loci, while conditioning on the top two independent variants in *C3* (rs2230199, 19:6718387 G/C and rs181290250, 19:6722565 C/T, respectively) eliminated the signal in *C3* but not *NRTN*. The *NRTN* signal disappeared after conditioning on the relatively common nonsynonymous variant rs79744308 (19:5827765 G/A, p.Ala59Thr, allele frequency = 4.1% in our samples), indicating that the *NRTN* signal was largely driven by this variant.

Chr	Gene	Locus	Frac. with rare vars [a]	# of variants	# of singletons	P-value	SKAT-O Rho [b]	Number of conditioning variants
1	*CFH*	*CFH*	0.069	94	56	2.33 x 10^{-10}	0	0
1	*CFHR2*	*CFH*	0.12	16	2	4.24 x 10^{-9}	0	0
6	*CFB*	*C2/CFB/SKIV2L*	0.23	80	47	3.90 x 10^{-11}	0	0
6	*C2*	*C2/CFB/SKIV2L*	0.11	54	38	8.29 x 10^{-10}	0	0
6	*NOTCH4*	*C2/CFB/SKIV2L*	0.23	114	68	7.15 x 10^{-9}	0	0
19	*C3*	*C3*	0.036	50	36	1.70 x 10^{-11}	0	0
19	*NRTN*	*C3*	0.082	12	8	5.67 x 10^{-7}	0.4	0
19	*C3*	*C3*	0.036	50	36	7.37 x 10^{-13}	0	1
19	*NRTN*	*C3*	0.082	12	8	1.34 x 10^{-7}	0.5	1
19	*NRTN*	*C3*	0.082	12	8	2.83 x 10^{-7}	0.5	2
19	*NRTN*	*C3*	0.082	12	8	0.318	1	3

[a] The fraction of samples with rare variants, defined as having an allele frequency of less than 5%. [b] The estimated weight parameter for SKAT-O, indicating the proportion of the weight assigned to the burden test. A Rho of 0 means the SKAT-O test is equivalent to a SKAT test, while a Rho of 1 means the SKAT-O test is equivalent to a burden test.

Table 2.10 Comparing rare pLoF allele carriers in AMD cases and controls with ExAC

The 2x2 contingency tables between case-control/ExAC status and pLoF carrier status were constructed as follows: allele carriers = pLoF count; allele non-carriers = N – pLoF count, where N = the number of samples in each data source (cases, controls, or ExAC); then two 2x2 tables can be constructed, one between cases and ExAC and one between controls and ExAC. (For example, the 2x2 table for *CFH* had 5 carriers and 2,389 non-carriers in cases, and 9 carriers and 60,697 non-carriers in ExAC.) We used one-sided Fisher's Exact Tests with null hypotheses that the odds ratio for the effect of ExAC vs. the tested cohort was less than 1 (that is, the gene had a lower pLoF frequency in the tested cohort than in ExAC). We analyzed a subset of genes for which the Case-ExAC *P*-value were significant (*p*-value < 0.01) but the Control-ExAC *P*-value were not significant (*p*-value > 0.01), and which lay within known AMD risk loci, totaling 3 genes. These genes had a higher case carrier frequency than control carrier frequency, consistent with what would be expected from rare deleterious loss-of-function variants.

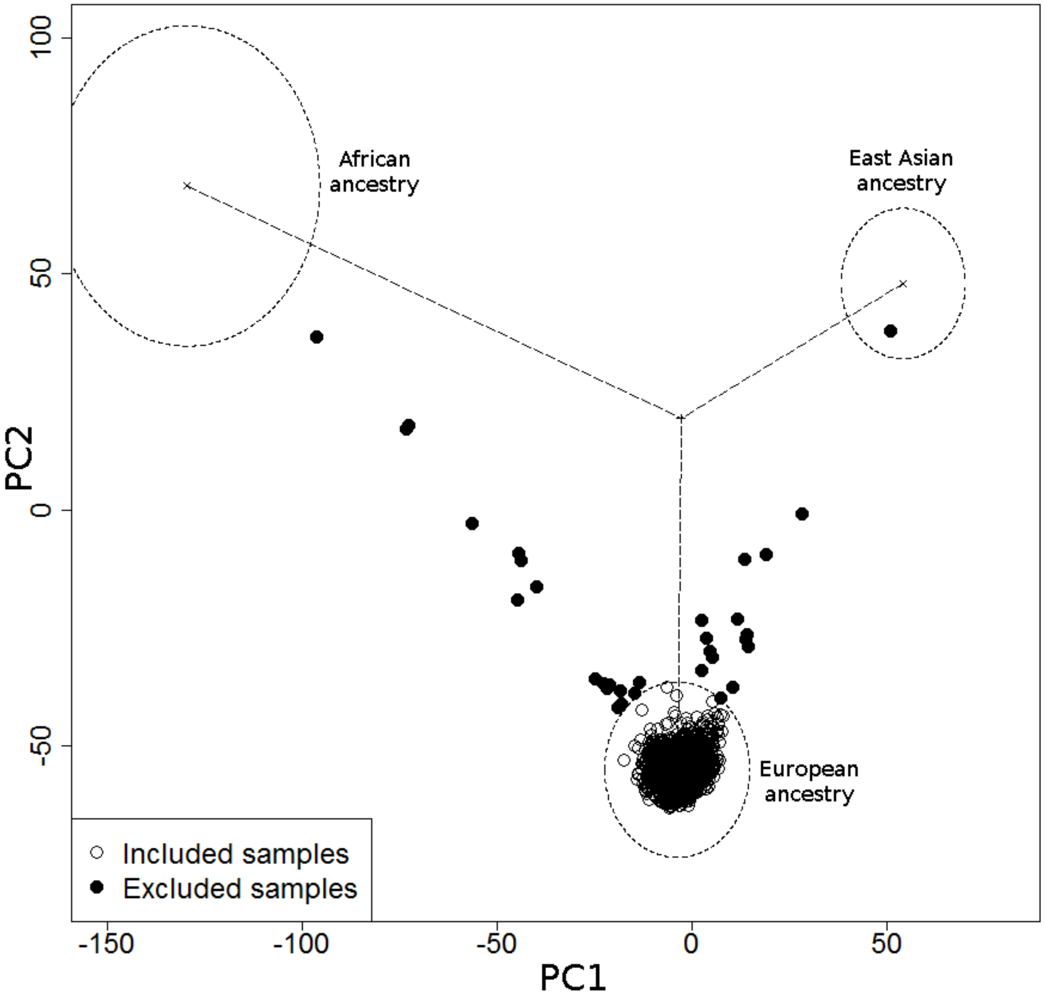
Gene	Case carriers	Control carriers	ExAC carriers	Case vs. ExAC	
	(N=2,394)	(N=2,393)	(N=60,706)	*P*-value	OR
CFH	5	0	9	1.2×10^{-4}	14.1
ORMDL2	3	0	6	3.9×10^{-3}	12.7
CFI	5	0	30	9.8×10^{-3}	4.2

Table 2.11 Replication gene-level association study of rare pLoF variants using Regeneron samples

Predicted LoF variants in 1,714 cases and 7,356 controls with allele frequency <1% were used to estimate rare pLoF carrier counts in five genes. Only two of the five genes (*CFH* and *SLC16A8*) contained rare pLoF variants in the Regeneron samples. We were able to replicate the signal in *CFH*. The pLoF frequency in *SLC16A8* was significantly different between Regeneron controls and ExAC, indicating that the significant *SLC16A8* result for case vs. ExAC could be a false positive, possibly arising from the differences in sequencing depth, variant calling algorithms, or population structure between the Regeneron and ExAC samples.

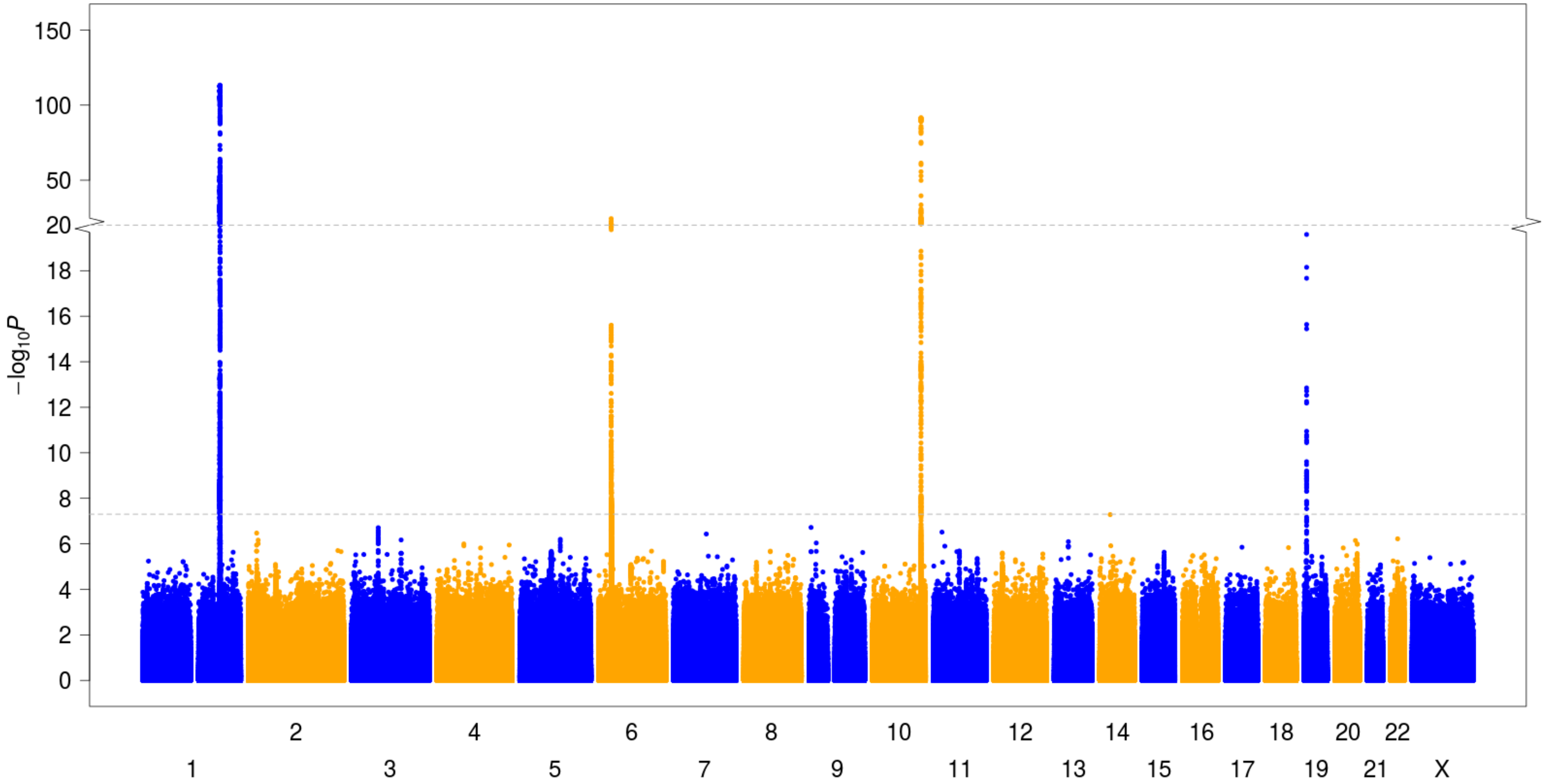
Gene	Case alleles	Control alleles	ExAC alleles	Case vs. Control		Case vs. ExAC		Control vs. ExAC	
	(N=1,714)	(N=7,356)	(N=60,706)	P-value	OR	P-value	OR	P-value	OR
CFH	**7**	**2**	**9**	**2.14×10^{-4}**	15.1	**1.07×10^{-7}**	27.7	0.34	1.8
CFI	**0**	**0**	**30**	1	n/a	1	n/a	1	n/a
ORMDL2	**0**	**0**	**6**	1	n/a	1	n/a	1	n/a
SLC16A8	**39**	**139**	**159**	0.29	1.21	**$<2.2 \times 10^{-16}$**	6.1	**$<2.2 \times 10^{-16}$**	7.3
TIMP3	**0**	**0**	**33**	1	n/a	1	n/a	1	n/a

Figure 2.1 Filtering samples with non-European ancestry with LASER and samples from the Human Genome Diversity Project



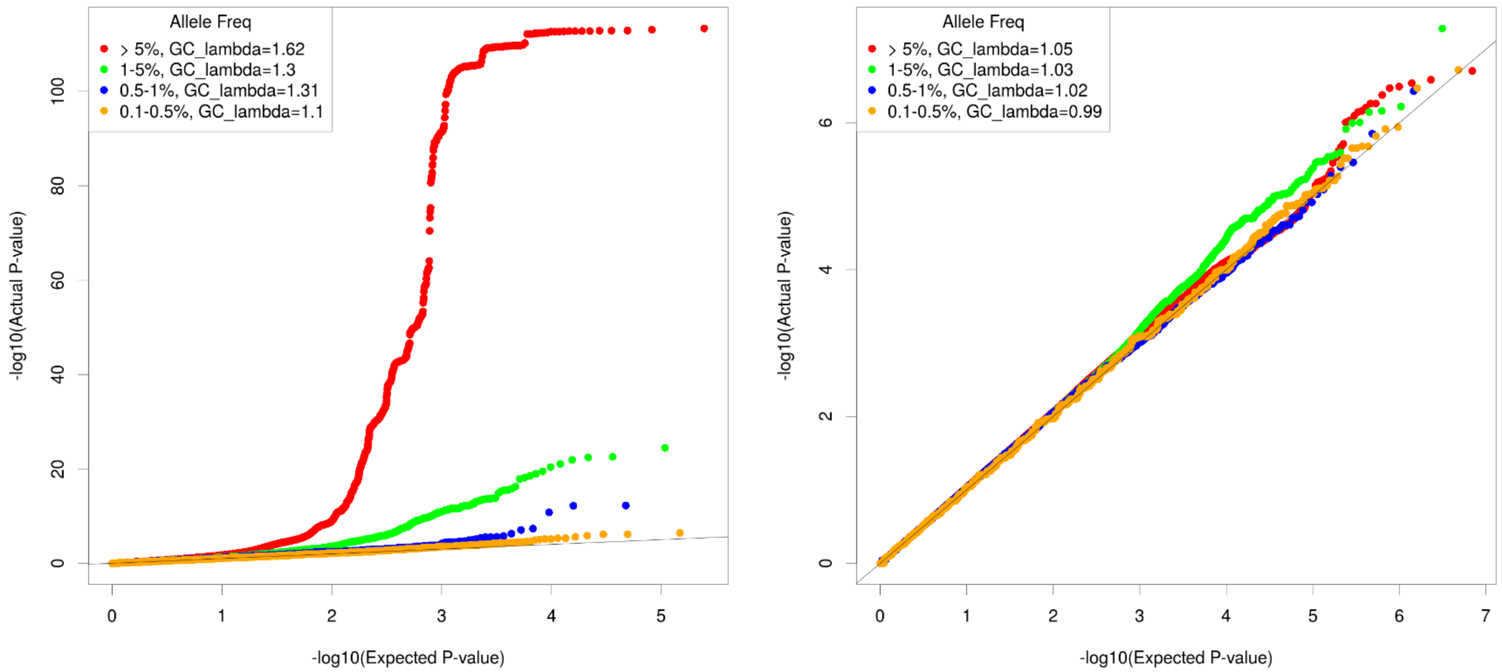
First, we calculated the average coordinates for the first two PCs of the European, African, East Asian, and all samples. Next, we defined the radius of the acceptance region for European samples to be one-quarter of the distance from the coordinates of the average European samples and the average coordinates of all samples. This led to the exclusion of 32 samples from our analysis.

Figure 2.2 Manhattan plot of AMD case-control single-variant genome wide association



Association signals can be found in *CFH* (chromosome 1), *C2/CFB/SKIV2L* (chromosome 6), *ARMS2/HTRA1* (chromosome 10), and *C3* (chromosome 19) using the genome wide significance threshold 5 x 10^{-8}.

Figure 2.3 QQ plots of single variant association P-values for variants inside and outside of known loci



(a) QQ Plot of single-variant association *P*-values for variants within known AMD risk loci. All significant signals (*P*-value < 5 x 10^{-8}) were found within these previously-discovered AMD risk loci. The most significant signals were all common variants (AF > 5%). (b) QQ Plot of single-variant association *P*-values for variants outside known AMD risk loci. There was no evidence of population stratification when we considered all variants with AF > 0.1% outside of known loci (λ_{GC} = 1.021).

2.5 References

Al-Khersan, H., A. Kwong and M. A. Grassi, 2018 Mutations in MERTK are not associated with age-related macular degeneration. Int Ophthalmol.

Brantley, M. A., Jr., S. Jain, E. E. Barr, E. M. Johnson, Jr. and J. Milbrandt, 2008 Neurturin-mediated ret activation is required for retinal function. J Neurosci 28**:** 4123-4135.

Browning, S. R., and B. L. Browning, 2007 Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet 81**:** 1084-1097.

Cann, H. M., C. de Toma, L. Cazes, M. F. Legrand, V. Morel *et al.*, 2002 A human genome diversity cell line panel. Science 296**:** 261-262.

Chaolong Wang, X. Z., Jennifer Bragg-Gresham, Hyun Min Kang, Dwight Stambolian,, K. E. B. Emily Y Chew, John Heckenlively, The FUSION Study, Robert Fulton, Richard K Wilson, and X. L. Elaine R Mardis, Anand Swaroop, Sebastian Zöllner, and Gonçalo R Abecasis, 2014 Ancestry estimation and control of population stratification for sequence-based association studies. Nature Genetics 46**:** 6.

Clarke, B. A., S. Majumder, H. Zhu, Y. T. Lee, M. Kono *et al.*, 2019 The Ormdl genes regulate the sphingolipid synthesis pathway to ensure proper myelination and neurologic function in mice. Elife 8.

Das, S., L. Forer, S. Schonherr, C. Sidore, A. E. Locke *et al.*, 2016 Next-generation genotype imputation service and methods. Nat Genet 48**:** 1284-1287.

Du, H., X. Xiao, T. Stiles, C. Douglas, D. Ho *et al.*, 2016 Novel Mechanistic Interplay between Products of Oxidative Stress and Components of the Complement System in AMD Pathogenesis. Open J Ophthalmol 6**:** 43-50.

Duvvari, M. R., J. P. van de Ven, M. J. Geerlings, N. T. Saksens, B. Bakker *et al.*, 2016 Whole Exome Sequencing in Patients with the Cuticular Drusen Subtype of Age-Related Macular Degeneration. PLoS One 11**:** e0152047.

Ferrara, D., and J. M. Seddon, 2015 Phenotypic Characterization of Complement Factor H R1210C Rare Genetic Variant in Age-Related Macular Degeneration. JAMA Ophthalmol 133**:** 785-791.

Friedman, D. S., B. J. O'Colmain, B. Munoz, S. C. Tomany, C. McCarty *et al.*, 2004 Prevalence of age-related macular degeneration in the United States. Arch Ophthalmol 122**:** 564-572.

Fritsche, L. G., W. Chen, M. Schu, B. L. Yaspan, Y. Yu *et al.*, 2013 Seven new loci associated with age-related macular degeneration. Nat Genet 45**:** 433-439, 439e431-432.

Fritsche, L. G., R. N. Fariss, D. Stambolian, G. R. Abecasis, C. A. Curcio *et al.*, 2014 Age-related macular degeneration: genetics and biology coming together. Annu Rev Genomics Hum Genet 15**:** 151-171.

Fritsche, L. G., W. Igl, J. N. Bailey, F. Grassmann, S. Sengupta *et al.*, 2016 A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. Nat Genet 48**:** 134-143.

Harada, C., T. Harada, H. M. Quah, F. Maekawa, K. Yoshida *et al.*, 2003 Potential role of glial cell line-derived neurotrophic factor receptors in Muller glial cells during light-induced retinal degeneration. Neuroscience 122**:** 229-235.

Hoover, J. L., C. E. Bond, D. B. Hoover and D. M. Defoe, 2014 Effect of neurturin deficiency on cholinergic and catecholaminergic innervation of the murine eye. Exp Eye Res 122**:** 32-39.

Huang, L. Z., Y. J. Li, X. F. Xie, J. J. Zhang, C. Y. Cheng *et al.*, 2015 Whole-exome sequencing implicates UBE3D in age-related macular degeneration in East Asian populations. Nat Commun 6**:** 6687.

Jomary, C., R. M. Darrow, P. Wong, D. T. Organisciak and S. E. Jones, 2004 Expression of neurturin, glial cell line-derived neurotrophic factor, and their receptor components in light-induced retinal degeneration. Invest Ophthalmol Vis Sci 45**:** 1240-1246.

Jomary, C., M. Thomas, J. Grist, J. Milbrandt, M. J. Neal *et al.*, 1999 Expression patterns of neurturin and its receptor components in developing and degenerative mouse retina. Invest Ophthalmol Vis Sci 40**:** 568-574.

Jun, G., M. Flickinger, K. N. Hetrick, J. M. Romm, K. F. Doheny *et al.*, 2012 Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. Am J Hum Genet 91**:** 839-848.

Jun, G., M. K. Wing, G. R. Abecasis and H. M. Kang, 2015 An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. Genome Res 25**:** 918-925.

Kang, H. M., 2012 EPACTS: efficient and parallelizable association container toolbox, pp.

Karczewski, K. J., L. C. Francioli, G. Tiao, B. B. Cummings, J. Alfoldi *et al.*, 2020 The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 581**:** 434-443.

Kavanagh, D., Y. Yu, E. C. Schramm, M. Triebwasser, E. K. Wagner *et al.*, 2015 Rare genetic variants in the CFI gene are associated with advanced age-related macular degeneration and commonly result in reduced serum factor I levels. Hum Mol Genet 24**:** 3861-3870.

Lee, S., M. C. Wu and X. Lin, 2012 Optimal tests for rare variant effects in sequencing association studies. Biostatistics 13**:** 762-775.

Lek, M., K. J. Karczewski, E. V. Minikel, K. E. Samocha, E. Banks *et al.*, 2016 Analysis of protein-coding genetic variation in 60,706 humans. Nature 536**:** 285-291.

Lukiw, W. J., B. Surjyadipta, P. Dua and P. N. Alexandrov, 2012 Common micro RNAs (miRNAs) target complement factor H (CFH) regulation in Alzheimer's disease (AD) and in age-related macular degeneration (AMD). Int J Biochem Mol Biol 3**:** 105-116.

Ma, C., T. Blackwell, M. Boehnke, L. J. Scott and T. D. i. Go, 2013 Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. Genet Epidemiol 37**:** 539-550.

McCarthy, S., S. Das, W. Kretzschmar, O. Delaneau, A. R. Wood *et al.*, 2016 A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet 48**:** 1279-1283.

McKay, G. J., C. C. Patterson, U. Chakravarthy, S. Dasari, C. C. Klaver *et al.*, 2011 Evidence of association of APOE with age-related macular degeneration: a pooled analysis of 15 studies. Hum Mutat 32**:** 1407-1416.

Miyake, M., K. Yamashiro, H. Tamura, K. Kumagai, M. Saito *et al.*, 2015 The Contribution of Genetic Architecture to the 10-Year Incidence of Age-Related Macular Degeneration in the Fellow Eye. Invest Ophthalmol Vis Sci 56**:** 5353-5361.

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

Parsons, L. S., 2001 Reducing Bias in a Propensity Score Matched-Pair Sample Using Greedy Matching Techniques. SAS Global Users Group 26**:** 214-226.

Pietraszkiewicz, A., F. van Asten, A. Kwong, R. Ratnapriya, G. Abecasis *et al.*, 2018 Association of Rare Predicted Loss-of-Function Variants in Cellular Pathways with Sub-Phenotypes in Age-Related Macular Degeneration. Ophthalmology 125**:** 398-406.

Price, A. L., G. V. Kryukov, P. I. de Bakker, S. M. Purcell, J. Staples *et al.*, 2010 Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet 86**:** 832-838.

Ratnapriya, R., X. Zhan, R. N. Fariss, K. E. Branham, D. Zipprer *et al.*, 2014 Rare and common variants in extracellular matrix gene Fibrillin 2 (FBN2) are associated with macular degeneration. Hum Mol Genet 23**:** 5827-5837.

Seddon, J. M., Y. Yu, E. C. Miller, R. Reynolds, P. L. Tan *et al.*, 2013 Rare variants in CFI, C3 and C9 are associated with high risk of advanced age-related macular degeneration. Nat Genet 45**:** 1366-1370.

Sherry, S. T., M. H. Ward, M. Kholodov, J. Baker, L. Phan *et al.*, 2001 dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29**:** 308-311.

Song, X. J., D. Q. Li, W. Farley, L. H. Luo, R. O. Heuckeroth *et al.*, 2003 Neurturin-deficient mice develop dry eye and keratoconjunctivitis sicca. Invest Ophthalmol Vis Sci 44**:** 4223-4229.

Spencer, K. L., M. A. Hauser, L. M. Olson, S. Schmidt, W. K. Scott *et al.*, 2007 Protective effect of complement factor B and complement component 2 variants in age-related macular degeneration. Hum Mol Genet 16**:** 1986-1992.

Sudlow, C., J. Gallacher, N. Allen, V. Beral, P. Burton *et al.*, 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**:** e1001779.

Sun, C., M. Zhao and X. Li, 2012 CFB/C2 gene polymorphisms and risk of age-related macular degeneration: a systematic review and meta-analysis. Curr Eye Res 37**:** 259-271.

Taliun, D., D. N. Harris, M. D. Kessler, J. Carlson, Z. A. Szpiech *et al.*, 2019 Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. bioRxiv.

Taylor, R. L., J. A. Poulter, S. M. Downes, M. McKibbin, K. N. Khan *et al.*, 2019 Loss-of-Function Mutations in the CFH Gene Affecting Alternatively Encoded Factor H-like 1 Protein Cause Dominant Early-Onset Macular Drusen. Ophthalmology 126**:** 1410-1421.

Thakkinstian, A., M. McEvoy, U. Chakravarthy, S. Chakrabarti, G. J. McKay *et al.*, 2012 The association between complement component 2/complement factor B polymorphisms and age-related macular degeneration: a HuGE review and meta-analysis. Am J Epidemiol 176**:** 361-372.

Triebwasser, M. P., E. D. Roberson, Y. Yu, E. C. Schramm, E. K. Wagner *et al.*, 2015 Rare Variants in the Functional Domains of Complement Factor H Are Associated With Age-Related Macular Degeneration. Invest Ophthalmol Vis Sci 56**:** 6873-6878.

Wei, W. Q., and J. C. Denny, 2015 Extracting research-quality phenotypes from electronic health records to support precision medicine. Genome Med 7**:** 41.

Yates, A., W. Akanni, M. R. Amode, D. Barrell, K. Billis *et al.*, 2016 Ensembl 2016. Nucleic Acids Res 44**:** D710-716.

Yu, Y., T. R. Bhangale, J. Fagerness, S. Ripke, G. Thorleifsson *et al.*, 2011 Common variants near FRK/COL10A1 and VEGFA are associated with advanced age-related macular degeneration. Hum Mol Genet 20**:** 3699-3709.

Zhan, X., D. E. Larson, C. Wang, D. C. Koboldt, Y. V. Sergeev *et al.*, 2013 Identification of a rare coding variant in complement 3 associated with age-related macular degeneration. Nat Genet 45**:** 1375-1379.

Zou, L., Y. Feng, G. Xu, W. Jian and W. Chao, 2016 Splenic RNA and MicroRNA Mimics Promote Complement Factor B Production and Alternative Pathway Activation via Innate Immune Signaling. J Immunol 196**:** 2788-2798.

Chapter 3

Robust, Flexible, and Scalable Tests for Hardy-Weinberg Equilibrium Across Diverse Ancestries

A paper covering most of the material in this chapter is in preparation, with myself as first author

3.1 Introduction

Hardy-Weinberg equilibrium (HWE) is a fundamental theorem of population genetics and has been one of the key mathematical principles to understand the characteristics of genetic variation in a population for more than a century (HARDY 1908; WEINBERG 1908). HWE describes a remarkably simple relationship between allele frequencies and genotype frequencies which is constant across generations in a homogeneous, random-mating populations. Genetic variants in a homogeneous population typically follow HWE except for unusual deviations due by very strong case-control association and enrichment (NIELSEN *et al.* 1998), sex linkage, or non-random sampling (WAPLES 2015).

HWE tests are often used to assess the quality of microsatellite (VAN OOSTERHOUT *et al.* 2004), SNP-array (WIGGINTON *et al.* 2005), and sequence-based (DANECEK *et al.* 2011) genotypes. Testing for HWE may reveal technical artifacts in sequence or genotype data, such as high rates of genotyping error and/or missingness or sequencing/alignment errors (NIELSEN *et al.* 2011). HWE testing may also be used to identify structural variants in which hemizygotes are incorrectly called as homozygotes, especially when multiple variants deviating from HWE occur close together (McCarroll *et al.* 2006). Quality control for array-based or sequence-based genotypes typically includes a HWE test to detect and filter out artifactual or poorly genotyped variants (Laurie *et al.* 2010; Nielsen *et al.* 2011).

While HWE tests are commonly and reliably used for variant quality control in samples from homogenous populations, applying them to more diverse samples remains challenging. When analyzing individuals from a heterogeneous population, the standard HWE tests may falsely flag real, well-genotyped variants, unnecessarily filtering them out for downstream analyses (Hao and Storey 2019). This problem is important since genetic studies increasingly collect data from heterogeneous populations. In principle, HWE tests in these structured populations can be performed on smaller cohorts with homogenous backgrounds (Bycroft *et al.* 2018), and the test statistics combined using Fisher's or Stouffer's method (Mosteller and Fisher 1948; Stouffer 1949). However, such a procedure requires much more effort than using a single HWE test across all samples and information that may be imperfect or unavailable.

Here, we describe RUTH (Robust Unified Test for Hardy-Weinberg Equilibrium) which tests for HWE under heterogeneous population structure. Our primary motivation for developing RUTH is to robustly filter out artifactual or poorly genotyped variants using HWE test statistics. RUTH is (1) computationally efficient, (2) robust against various degrees of population structure, and (3) flexible in accepting key representations of sequence-based genotypes including best-guess genotypes and genotype likelihoods. We perform systematic evaluations of RUTH and alternative methods for HWE testing using simulated and real data to explore the advantages and disadvantages of these methods for samples of diverse ancestries.

3.2 Materials and Methods

3.2.1 Unadjusted HWE tests

Consider a study of n participants with true (unobserved) genotypes $g_1, g_2, \cdots, g_n$ at a bi-allelic variant coded as 0 (reference homozygote), 1 (heterozygote), or 2 (alternate homozygote). Represent the best-guess/hard-call (observed) genotypes as $\hat{g}_1, \hat{g}_2, \cdots, \hat{g}_n$. A simple HWE test uses the chi-squared statistic to compare the expected and observed genotype counts assuming no population structure and no genotype uncertainty. The chi-squared HWE test statistic is defined as $T_{\chi^2} = \sum_{k=0}^{2} \frac{(c_k - \hat{c}_k)^2}{\hat{c}_k}$ where $c_j = \sum_{i=0}^{n} I(\hat{g}_i = j)$ (ignoring missing genotypes), $\hat{p} = \frac{c_1 + 2c_2}{2n}, \hat{q} = 1 - \hat{p}$, $\hat{c}_0 = n\hat{q}^2, \hat{c}_1 = 2n\hat{p}\hat{q}, \text{and } \hat{c}_2 = n\hat{p}^2$. Under HWE, the asymptotic distribution of T_{χ^2} is usually assumed to follow χ_1^2 (ROHLFS AND WEIR 2008). An exact test is known to be more accurate for finite samples, particularly for rare variants (WIGGINTON *et al.* 2005). HWE tests stratified by case-control status are known to prevent an inflation of Type I errors for disease-associated variants (LI AND LI 2008). Widely used software tools such as PLINK (PURCELL *et al.* 2007) and VCFTools (DANECEK *et al.* 2011) implement an exact HWE test based on best-guess genotypes. We will refer to the exact test as the unadjusted test.

3.2.2 Existing HWE tests accounting for structured populations

The unadjusted HWE test assumes that the population is homogeneous. If a study is comprised of a set of discrete structured subpopulations, a straightforward extension of the unadjusted test is to (1) stratify each study participant into exactly one of the subpopulations, (2) perform the unadjusted HWE test for each subpopulation separately, and (3) meta-analyze test statistics across subpopulations to obtain a combined p-value using Stouffer's method (STOUFFER *et al.* 1949). More specifically, let $z_1, z_2, \cdots, z_s$ be the z-scores from HWE test statistics for *s* distinct subpopulations with sample sizes $n_1, n_2, \cdots, n_s$. A combined meta-analysis HWE test statistic across the subpopulations is then $T_{meta} = \frac{\sum_{i=1}^{s} z_i \sqrt{n_i}}{\sqrt{\sum_{i=1}^{s} n_i}}$, which asymptotically follows a standard normal distribution when each subpopulation follows HWE.

When the population cannot be easily stratified into distinct subpopulations (e.g. intra-continental diversity or an admixed population), a quantitative representation of genetic ancestry, such as principal component (PC) coordinates or fractional mixture over subpopulations, can be more useful for representing each study participant's genetic diversity (ROSENBERG *et al.* 2002; PRICE *et al.* 2006). HWES takes PCs as additional input to perform HWE tests under population structure with logistic regression (SHA AND ZHANG 2011), and a similar idea was suggested by Hao and colleagues (2016). However, existing implementations do not support for sequence-based genotypes (where genotype uncertainty may remain when sequencing depth is low or moderate) or commonly used formats for genetic array data. A recent method, PCAngsd estimates PCs from uncertain genotypes represented as genotype likelihoods (MEISNER AND ALBRECHTSEN 2019) and uses these estimates to perform a likelihood ratio test (LRT) for HWE, which is similar to the LRT version of RUTH with differences in computational performance (see below).

3.2.3 Robust HWE testing with RUTH

Here we describe RUTH (Robust and Unified Test for Hardy-Weinberg equilibrium) to enable HWE testing under structured populations, which is especially useful for large sequencing studies. We developed RUTH to produce HWE test statistics to allow quality control of sequence-based variant callsets from increasingly diverse samples. RUTH models the uncertainty encoded in sequence-based genotypes to robustly distinguish true and artifactual variants in the presence of population structure, and seamlessly scales to millions of individuals and genetic variants.

We assume the observed genotype for individual i can be represented as a genotype likelihood (GL) $L_i^{(G)} = \Pr(Data_i | g_i = G)$, where $Data_i$ represents observed data (e.g. sequence or array), and $g_i \in \{0,1,2\}$ the true (unobserved) genotype. For example, GLs for sequence-based genotypes can be represented as $L_i^{(G)} = \prod_{j=1}^{d_i} \Pr(r_{ij} | g_i = G; q_{ij})$ where d_i is the sequencing depth, r_{ij} is the observed read, and q_{ij} is the corresponding quality score (EWING AND GREEN 1998; JUN *et al.* 2012). We model GLs for best-guess genotypes $\hat{g}_i$ from SNP arrays as $L_i^{(G)} = (1 - e_i)^2$, $2e_i(1 - e_i)$, e_i^2 for $\hat{g}_i = 2, 1, 0$ where e_i is assumed per-allele error rate. Imputed genotypes may also be approximately modeled using this framework, but the current implementation requires creating a pseudo-genotype likelihood to describe this uncertainty (see Discussion).

3.2.4 Accounting for population structure with individual-specific allele frequencies

We account for population structure by modeling individual-specific allele frequencies from quantitative coordinates of genetic ancestry such as PCs, similar to the model (HAO *et al.* 2016). For any given variant, instead of assuming that genotypes follow HWE with a single universal allele frequency across all individuals, we assume that genotypes follow HWE with heterogeneous allele frequencies specific to each individual, modeled as a function of genetic ancestry. Let $\boldsymbol{x_i} \in \mathbb{R}^k$ represent the genetic ancestry of individual i, where k is the number of PCs used. We estimate individual-specific allele frequency p as a bounded linear function of genetic ancestry

$$p(\boldsymbol{x}_i;\boldsymbol{\beta}) = \begin{cases} \boldsymbol{\beta^T x}_i & \varepsilon \le \boldsymbol{\beta^T x}_i \le 1-\varepsilon \\ \varepsilon & \boldsymbol{\beta^T x}_i < \varepsilon \\ 1-\varepsilon & \boldsymbol{\beta^T x}_i > 1-\varepsilon \end{cases},$$

where ε is the minimum frequency threshold. We used $\varepsilon = \frac{1}{4n}$ in our evaluation. Even though we used a linear model for $p(\boldsymbol{x}_i;\boldsymbol{\beta})$ for computational efficiency, it is straightforward to apply a logistic model, which is arguably better (YANG *et al.* 2012; HAO *et al.* 2016).

Let $p_i = p(\boldsymbol{x}_i;\boldsymbol{\beta})$ and $q_i = 1 - p_i$ be the individual specific allele frequencies of the non-reference and reference alleles for individual $i.$ Under the null hypothesis of HWE, the frequencies of genotypes (0, 1, 2) are $[q_i^2,\ 2p_iq_i,\ p_i^2]$. Under the alternative hypothesis, we assume these frequencies are $[q_i^2 + \theta p_iq_i,\ 2p_iq_i(1-\theta),\ p_i^2 + \theta p_iq_i]$ where θ is the inbreeding coefficient. This model is a straightforward extension of a fully general model where p_i, q_i is identical across all samples. Then the log-likelihood across all study participants is

$$l(\boldsymbol{\beta}, \theta) = \sum_{i=1}^{n} \log\left[L_i^{(0)}(q_i^2 + \theta p_i q_i) + L_i^{(1)}\, 2p_i q_i(1-\theta) + L_i^{(2)}(p_i^2 + \theta p_i q_i)\right]$$

Under both the null ($\theta = 0$) and alternative ($\theta \neq 0$) hypotheses, we maximize the log-likelihood using an Expectation-Maximization (E-M) algorithm (DEMPSTER *et al.* 1977). As we empirically observed quick convergence within several iterations in most cases, we used a fixed (n=20) number of iterations in our implementation.

3.2.5 RUTH Score Test

The score function of the log-likelihood is

$$U(\theta) = \sum_{i=1}^{n} \frac{p_i q_i\left[L_i^{(0)} - 2L_i^{(1)} + L_i^{(2)}\right]}{L_i^{(0)}(q_i^2 + \theta p_i q_i) + L_i^{(1)}\ 2p_i q_i(1-\theta) + L_i^{(2)}(p_i^2 + \theta p_i q_i)} = \sum_{i=1}^{n} u_i(\theta)$$

Since $u_i'(\theta) = -u_i^2(\theta)$, we construct a score test statistic of $H_0\colon \theta = 0$ vs $H_1\colon \theta \neq 0$ as:

$$T_{score} = \frac{[U(0)]^2}{I(0)} = \frac{[\sum_{i=1}^{n} u_i(0)]^2}{\sum_{i=1}^{n} u_i^2(0)}$$

where *I*(0) is the Fisher information under the null hypothesis. Under the null, T_{score} has an asymptotic chi-squared distribution with one degree of freedom, i.e. $T_{score} \sim \chi_1^2$. We estimate $\widehat{\boldsymbol{\beta}}$ with an E-M algorithm.

3.2.6 RUTH Likelihood Ratio Test

The log-likelihood function $l(\boldsymbol{\beta}, \theta)$ can also be used to calculate a likelihood ratio test statistic:

$$T_{LRT} = 2\left[\max_{\boldsymbol{\beta},\theta} l(\boldsymbol{\beta}, \theta) - \max_{\boldsymbol{\beta}} l(\boldsymbol{\beta}, 0)\right].$$

Like the score test, we estimate MLE parameters $\boldsymbol{\beta}, \theta$ iteratively using an E-M algorithm to test $H_0: \theta = 0$ vs $H_1: \theta \neq 0$. Under the null hypothesis, the asymptotic distribution of T_{LRT} is expected to follow χ_1^2. This test is very similar to the likelihood-ratio test proposed by PCAngsd (MEISNER AND ALBRECHTSEN 2019), except PCAngsd does not re-estimate $\boldsymbol{\beta}$ under the alternative hypothesis. In principle, the RUTH LRT should be slightly more powerful due to this difference; we expect the practical difference in power to be small, as deviations from HWE usually do not change the estimates of $\boldsymbol{\beta}$ substantially.

3.2.7 Simulation of genotypes and sequence reads under population structure

We simulated sequence-based genotypes under population structure using the following procedure. First, for each variant, we simulated an ancestral allele frequency and population-specific allele frequencies. Second, we sampled unobserved (true) genotypes based on these allele frequencies. Third, we sampled sequence reads based on the unobserved genotypes. Fourth, we generated genotype likelihoods and best-guess genotypes based on sequence reads.

To simulate ancestral and population-specific allele frequencies, we followed the BALDING AND NICHOLS (1995) procedure, except we sampled ancestral allele frequencies from $p \sim Uniform(0,1)$ instead of $p \sim \mathrm{Uniform}(0.1, 0.9)$ to include rare variants. For each of $K \in \{1, 2, 5, 10\}$ populations, we sampled population-specific allele frequencies from $p_k \sim Beta\left(\frac{p(1-F_{st})}{F_{st}}, \frac{(1-p)(1-F_{st})}{F_{st}}\right)$, where $k \in \{1, \cdots, K\}$, and $F_{st} \in \{.01, .02, .03, .05, .10\}$ was the fixation index to quantify the differentiation between the populations, as suggested by Holsinger (HOLSINGER 1999) and implemented in previous studies (HOLSINGER *et al.* 2002; BALDING 2003). Because p_k no longer follows the uniform distribution, we used rejection sampling to ensure that $\bar{p} = \frac{1}{K}\sum_{k=1}^{K} p_k$ is uniformly distributed across 100 bins across simulations to avoid artifacts caused by systematic differences in allele frequencies.

The unobserved genotype $G_i \in \{0,1,2\}$ for individual $i \in \{1, \cdots, n_k\}$, belonging to population k with sample size n_k, was simulated from genotype frequencies $(q_k^2 + \theta\, p_k q_k, 2p_k q_k(1-\theta), p_k^2 + \theta\, p_k q_k)$, where $q_k = 1 - p_k$ and $\theta \in \left[-\min\left(\frac{q_k}{p_k}, \frac{p_k}{q_k}\right), 1\right]$ quantifies deviation from HWE; $\theta = 0$ represents HWE, while $\theta < 0$ and $\theta > 0$ represent excess heterozygosity and homozygosity compared to HWE expectation, respectively. In our experiments, we evaluated $\theta \in \{0, \pm.01, \pm.05, \pm.1, \pm.5\}$. When θ was smaller than the minimum possible value for a specific population, we replaced it with the minimum value.

We simulated sequence reads based on unobserved genotypes, sequence depths, and base call error rates. To reflect the variation of sequence depths between individuals, we simulated the mean depth of each sequenced sample to be distributed as $\mu_i \sim Uniform(1, 2D-1)$, where D is the expected depth and $D = 5$ and $D = 30$ representing low-coverage and deep sequencing, respectively. For each sequenced sample and variant site, we sampled the sequence depth from $d_i \sim Poisson(\mu_i)$. Each sequence read carried either of the possible unobserved (true) alleles $r_{ij} \in \{0,1\}$, where $j \in \{1, \cdots, d_i\}$. Given unobserved genotype G_i, we generated $r_{ij} \sim Bernoulli\left(\frac{G_i}{2}\right)$, with observed allele $o_{ij} = \left(1 - e_{ij}\right)r_{ij} + e_{ij}\left(1 - r_{ij}\right)$ flipping to the other allele when a sequencing error occurs with probability $e_{ij} \sim Bernoulli(\epsilon)$. We used $\epsilon = 0.01$ throughout our simulations (which corresponds to phred-scale base quality of 20) and assumed that all base calling errors switched between reference and alternate alleles.

We then generated genotype likelihoods and best-guess genotypes from the simulated alleles. Let $t_i = \sum_{j=1}^{d_i} o_{ij}$ be the observed alternate allele count. The GLs for the three possible genotypes are $L_i^{(0)} = (1 - \epsilon)^{d_i - t_i}\,(\epsilon)^{t_i},\ L_v^{(1)} = 0.5^{d_i}$, $L_i^{(2)} = (\epsilon)^{d_i - t_i}\,(1 - \epsilon)^{t_i}$. We called best-guess genotypes by using the overall ancestral allele frequency $\bar{p}$ for a given variant as the prior, then calling the genotype corresponding to the highest posterior probability among $\left(L_i^{(0)}(1 - \bar{p})^2,\ 2L_i^{(1)}\bar{p}(1 - \bar{p})^2,\ L_i^{(2)}\bar{p}^2\right)$ for each sample. For each possible combination of F_{st}, K, and θ, we generated 50,000 independent variants across a set of $n = 5{,}000$ samples with per-ancestry samples sizes $n_k = \frac{n}{K}$.

3.2.8 Evaluation of Type I error and statistical power

We used different p-value thresholds, F_{st} values, number of ancestry groups K, and average sequencing depth D to determine the number of variants significantly deviating from HWE. To evaluate Type I error, we simulated sequence reads under HWE ($\theta = 0$) and calculated the proportion of significant variants at each p-value threshold. In RUTH tests, we assumed PCs were accurately estimated using true genotypes unless indicated otherwise. For real data, we summarized ancestral information by projecting PCs estimated from their full genomes onto the reference PC space of the Human Genome Diversity Panel (HGDP) (LI *et al.* 2008) using verifyBamID2 (ZHANG *et al.* 2020), similar to the procedure for variant calling in the TOPMed Project, which has already integrated RUTH as part of its quality control pipeline (https://github.com/statgen/topmed_variant_calling).

3.2.9 Real data

We used variants from two different real data sets in our evaluations: the 1000 Genomes Project (1000G) (THE 1000 GENOMES PROJECT CONSORTIUM *et al.* 2015) and the Trans-Omics Precision Medicine (TOPMed) project. We restricted our test to variants on chromosome 20. The 1000G data consists of 2,504 individuals from 26 populations, sequenced at an average depth of 6x, with 5,041,480 total variants. The TOPMed data consists of variants from 53,831 individuals from the TOPMed sequencing study (TALIUN *et al.* 2019), sequenced at an average depth of 37x, with . 12,983,576 total variants.

3.2.10 Application to 1000 Genomes data

We evaluated all tests using 1000G data. Unlike the simulated data, variants in 1000G are not clearly classified into “true” or “artifactual”, so evaluation of false positives and power is less straightforward. We focused on two specific subsets of variants in chromosome 20. We selected 17,740 variable sites found in both the Illumina Infinium Omni2.5 genotyping array and in HapMap3 (THE INTERNATIONAL HAPMAP CONSORTIUM *et al.* 2010), which we expect to be “high-quality” (HQ) variants that mostly follow HWE after controlling for ancestry. Similarly, we selected 10,966 variants that displayed high discordance between duplicates or Mendelian inconsistencies within family members in TOPMed sequencing study as “low quality” (LQ) variants that show be enriched for deviations from HWE, even after accounting for ancestry. Among 329,699 LQ variants from TOPMed in chromosome 20, we found that only 10,966 overlap with 1000 Genome samples because likely artifactual variants were stringently filtered prior to haplotype phasing. We suspect that a substantial fraction of these 10,966 LQ variants are true variants since they passed all of the 1000G Project’s quality filters. Nevertheless, we still expect a much larger fraction of these LQ variants to deviate from HWE compared to HQ variants.

We evaluated multiple representations of sequence-based genotypes from 1000G. As 1000G samples were sequenced at relatively low-coverage of $6\times$ on average, best-guess genotypes inferred only from sequence reads (raw GT) tend to have poor accuracy. Therefore, the officially released best-guess genotypes in 1000G were estimated by combining genotype likelihoods (GL), calculated based on sequence reads, with haplotype information from nearby variants through linkage-disequilibrium (LD)-aware genotype refinement using SHAPEIT2 (DELANEAU *et al.* 2013). This procedure resulted in more accurate genotypes (LD-aware GT), but it implicitly assumed HWE during refinement. As different representations of sequence genotypes may result in different performance in HWE tests, we evaluated all three different representations—raw GT, LD-aware GT, and GL. In all tests of RUTH using hard genotype calls, we assumed the error rate for GT-based genotypes to be 0.5%, which is representative of a typical non-reference genotype error rate for SNP arrays. We restricted our analyses to biallelic variants. The positions and alleles of 1000G and TOPMed variants were matched using the liftOver software tool (KUHN *et al.* 2013).

We evaluated all tests as described above. For meta-analysis with Stouffer's method, we divided the samples into 5 strata, using the five 1000G super population code labels – African (AFR), Admixed American (AMR), East Asian (EAS), European (EUR), and South Asian (SAS). To obtain PC coordinates for 1000G samples, we estimated 4 PCs from the aligned sequence reads (BAM) with verifyBamID2 (ZHANG *et al.* 2020), using PCs from 936 samples from the Human Genome Diversity Project (HGDP) panel as reference coordinates. The RUTH score test and LRT used these PCs as inputs, along with genotypes in raw GT, LD-aware GT, and GL formats. For PCAngsd, we used GLs from all variants tested as the input. We limited the analysis to a single chromosome due to the heavy computational requirements of PCAngsd.

3.2.11 Application to TOPMed data

We evaluated all tests using TOPMed samples, which came from multiple studies from a diverse spectrum of ancestries, leading to substantial population structure. Using the same criteria as our 1000G analysis, we identified 17,524 high-quality variants and 329,699 low-quality variants across chromosome 20. Since TOPMed genomes were deeply sequenced at $37.2\times(\pm 4.5\times)$, LD-aware genotype refinement was not necessary to obtain accurate genotypes. Therefore, we used two genotype representations – raw GT and GL – in our evaluations.

Similar to 1000G, for best-guess genotypes (raw GT), we used PLINK for the unadjusted test. For meta-analysis, we assigned each sample to one of the five 1000G super populations as follows. First, we summarized the genetic ancestries of aligned sequenced genomes with verifyBamID2 by estimating 4 PCs using HGDP as reference. Second, we used Procrustes analysis (DRYDEN AND MARDIA 1998; WANG *et al.* 2010) to align the PC coordinates of HGDP panels (to account for different genome builds) so that the PC coordinates were compatible between TOPMed and 1000G samples. Third, for each TOPMed sample, we identified the 10 closest corresponding individuals from 1000G using the first 4 PC coordinates with a weighted voting system (assigning the closest individual a score of 10, next closest a score of 9, and so on until the 10th closest individual is assigned a score of 1, then adding up the scores for each super population) to determine the super population code that had the highest sum of scores, and therefore best described that sample. In this way, we classified 15,580 samples as AFR, 4,836 as AMR, 29,943 as EUR, 2,960 as EAS, and 716 as SAS. Among these samples, 94.5% had the same super population code for all 10 nearest 1000G neighbors. To evaluate the RUTH score test and LRT for both raw GT and GL, we used 4 PCs estimated by verifyBamID2 (ZHANG *et al.* 2020), consistent with the method applied for the 1000G data.

3.2.12 Impact of ancestry estimates on adjusted HWE tests

We examined the effect of changing the number of PCs used as input for RUTH tests by using 2 PCs as opposed to 4 PCs. We also evaluated the impact of using different approaches to classify ancestry when adjusting for population structure with meta-analysis. By default, our analysis classified the 1000 Genomes subjects into 5 continental super populations based on published information (THE 1000 GENOMES PROJECT CONSORTIUM *et al.* 2015). For TOPMed, the best-matching 1000 Genomes continental ancestry was carefully determined using the PCA-based matching strategy described above. However, in practice, ancestry classification may be performed with a coarser resolution (JIN *et al.* 2019). To mimic such a setting, we used k-means clustering on the first 2 PCs of our samples to divide individuals into 3 distinct groups, and performed meta-analyses based on this coarse classification for both 1000G and TOPMed data.

3.2.13 Evaluation of sensitivity and specificity

In all datasets, we evaluated the tradeoff between Type I Error and power for each method using precision-recall curves (PRCs) and receiver-operator characteristic curves (ROCs). In simulated data, we considered variants with $\theta = 0$ to be true negatives and variants with $\theta = -0.05$ to be true positives. In both our 1000G and TOPMed data, we labeled HQ variants as negative and LQ variants as positive.

3.3 Results

3.3.1 Simulation: Effect of genotype uncertainty

To evaluate the impact of genotype uncertainty, we first compared tests in the absence of population structure (i.e. single ancestry). For the unadjusted test, we used only best-guess genotypes (GTs). For PCAngsd, we used only genotype likelihoods (GLs). For RUTH score and likelihood ratio tests, we used both.

Using GLs over GTs substantially reduced Type I errors in HWE tests, especially in low-coverage data (Figure 3.1A-C). For example, the standard HWE test based on GTs resulted in a 229-fold inflation (22.9%) at $p < .001$ (Figure 3.1B, Supplemental

Table 3.1), which is a threshold to evaluate Type I error with a reasonable precision with 50,000 variants (50 expected false positives under the null). GT-based RUTH-Score and RUTH-LRT tests showed similar inflations. When GLs were used instead of best-guess genotypes, RUTH-Score and RUTH-LRT had Type I errors close to the null expectation (.0010 and .0011, respectively). PCAngsd, which also accounts for genotype uncertainty (MEISNER AND ALBRECHTSEN 2019), had similar performance. The severely inflated Type I errors with best-guess genotypes can largely be attributed to high uncertainty and bias towards homozygote reference genotypes in single site calls from low-coverage sequence data, resulting in apparent deviations from HWE. For high-coverage sequence data, inflation of Type I error with GTs was substantially attenuated (.0040 and .0021 for RUTH-Score and RUTH-LRT, respectively); inflation nearly disappeared when using GLs (.0014 and .0010 for RUTH-Score and RUTH-LRT, respectively; Figure 3.1D-F).

Next, we evaluated the power to identify variants truly deviating from HWE at various levels of inbreeding coefficient (θ). For low-coverage sequence data, we skip interpretation of power of GT-based tests owing to their extremely inflated false positive rates. All GL-based tests behaved similarly, achieving ~19-21% power at $p < .001$ with moderate excess heterozygosity ($\theta = -0.05$) (Figure 3.2B, Supplementary Table 3.1). For high-coverage sequence data, the power of GL-based tests at the same p-value threshold increased to ~56-60%, comparable to corresponding GT-based tests. Interestingly, the unadjusted GT-based test showed much lower power than RUTH and PCAngsd tests under excess heterozygosity ($\theta < 0$) while demonstrating much higher power with excess homozygosity ($\theta > 0$). Upon further investigation, we observed that the tests behave very differently for rare variants for which an asymptotic approximation performs poorly.

We also generated precision-recall curves (PRC) and receiver-operator characteristic (ROC) curves to better understand the tradeoff between the Type I errors and power under moderate excess heterozygosity ($\theta = -.05$) (Supplementary Figure 3.1C-D). Again, accounting for genotype uncertainty resulted in better empirical power and Type I error, especially for low-coverage data, for which, at an empirical false positive rate of 1%, GL-based tests had 41-45% power, as opposed to 4-10% for GT-based tests. For high-coverage data, GL-based tests had 1-2% greater power than GT-based tests at the same false positive rate. These results suggest that ignoring genotype uncertainty in HWE tests is reasonable for high-coverage sequence data.

3.3.2 Simulation: Impact of population structure on HWE test statistics

As expected, the unadjusted HWE test had substantially inflated Type I errors under population structure based on the Balding-Nichols (1995) model (Figure 3.1, Supplementary Table 3.1). Even for an intra-continental level of population differentiation ($F_{ST} = .01$), the Type I errors at $p < .001$ were inflated 13.5-fold even for high-coverage data. With an inter-continental level of differentiation ($F_{ST} = .1$), we observed orders of magnitude more Type I errors across different simulation conditions. This inflation is expected to increase with larger sample sizes, suggesting that adjustment for population structure is important even if a study focuses on a single continental population.

One simple approach to account for population structure is to stratify individuals into distinct subpopulations to apply HWE tests separately (BYCROFT *et al.* 2018), and meta-analyze the results (Figure 3.3B). Type I errors were appropriately controlled with this approach in high-coverage but not low-coverage data, likely due to unmodeled genotype uncertainty (Figure 3.1, Supplementary Table 3.1). Instead of classifying individuals into distinct subpopulations, RUTH incorporates PCs to jointly perform HWE tests (Figure 3.3C). For both low- or high-coverage data, GL-based RUTH tests and PCAngsd showed well-controlled Type I errors, while GT-based tests showed slight (high-coverage) or severe (low-coverage) inflation.

Although meta-analysis resulted in well-controlled Type I errors for high-coverage data, it was considerably less powerful than RUTH. For example, with moderate excess heterozygosity (θ = -.05) across five ancestries (F_{ST} = .1), RUTH tests identified 20-27% more variants as significant at p < .001 (Figure 3.2, Supplementary Table 3.1) compared to meta-analysis. PRCs also clearly showed better operating characteristics for RUTH and PCAngsd compared to meta-analysis (Supplementary Figure 3.2). For example, at an empirical false positive rate of 1%, RUTH showed much greater power (66-68%) than meta-analysis (43%), even though the simulation scenario favors meta-analysis because samples were perfectly classified into distinct subpopulations.

3.3.3 Application to 1000 Genomes WGS data

Next, we evaluated the performance of various HWE tests in low-coverage (~6x) sequence data from the 1000 Genomes Project. We evaluated three representations of genotypes—(1) raw GT, (2) LD-aware GT, and (3) GL, as described in Materials and

Methods. Among chromosome 20 variants, we selected 17,740 high-quality (HQ) variants that are polymorphic in GWAS arrays, and 10,966 low-quality (LQ) variants enriched for genotype discordance in duplicates and trios. Unlike simulation studies, not all LQ variants are necessarily expected to violate HWE, so we consider the proportion of significant LQ variants as a lower bound on the sensitivity to identify significant variants. Similarly, not all HQ variants are necessarily expected to follow HWE, although we expect most to do so, so that the proportion of significant HQ variants serves as an upper bound for the false positive rate.

Consistent with our simulation results, all tests based on raw GTs generated from low-coverage sequence data had severe inflation of false positives (Figure 3.4A, Table 3.1). This was true even for HQ variants, presumably due to genotyping error and bias in raw GTs. Standard HWE tests, which model neither genotype uncertainty nor population structure, showed the highest inflation of false positives at 44% for $p < 10^{-6}$, a threshold commonly used for HWE testing in large genetic studies (LOCKE *et al.* 2015; FRITSCHE *et al.* 2016). Modeling population structure substantially reduced inflation, with RUTH tests showing fewer false positives (0.7-1.0% at $p < 10^{-6}$) than meta-analysis (2.0% at $p < 10^{-6}$). False positives were inflated across all methods when using raw GTs.

Consistent with our simulation studies, GL-based RUTH tests reduced false positives even further (0.034% at $p < 10^{-6}$). In contrast to our simulations, PCAngsd demonstrated considerably higher false positives than RUTH (2.1% at $p < 10^{-6}$), likely because PCAngsd estimates PCs from the input data without the ability to use externally provided PCs (see Discussion). The sensitivity for detecting significant LQ variants was also consistent with our simulations (Figure 3.4B, Table 3.1). GL-based tests, which showed better control of false positives, identified 22-25% of LQ variants as significant at $p < 10^{-6}$.

Strikingly, while using LD-aware GTs reduced false positives with adjusted tests, it was at the expense of substantially reduced sensitivity to detect LQ variants. The false positive rates of any adjusted test with LD-aware GTs were uniformly lower than those of any GL- and raw GT-based tests across all p-value thresholds (Figure 3.4A). However, sensitivity was also substantially reduced with LD-aware genotypes (Figure 3.4B). For example, at $p < 10^{-6}$, GL-based RUTH tests identified 22-23% of LQ variants significant, while using LD-aware GTs halved the proportions. Running meta-analysis with LD-aware GTs reduced sensitivity even further, likely because the implicit HWE assumption in the LD-aware genotype refinement algorithms may have further reduced false positives and sensitivity by altering the LD-aware genotypes to conform to HWE.

We evaluated PRCs between HQ and LQ variants to further evaluate this tradeoff. The results clearly demonstrated that HWE tests using LD-aware GTs are substantially less robust than tests on other genotype representations (Supplementary Table 3.2, Supplementary Figure 3.3A). For example, for the RUTH score test, when LD-aware GTs identified 0.1% of HQ variants as significant, 17% of LQ variants were identified as significant. However, with raw GT and GL, 24~27% were identified as significant at the same threshold. Even fewer were significant in meta-analysis with LD-aware GTs (13%). Similar trends were observed across all thresholds, suggesting that using LD-aware GTs results in substantially poorer operating characteristics than other genotype representations. As more accurate genotyping in LD-aware genotype refinement is expected to improve the performance of QC metrics compared to raw GTs, these results are quite striking, and highlight a potential oversight in using LD-aware genotypes in various QC metrics for sequence-based genotypes.

3.3.4 Application to TOPMed deep WGS data

We evaluated the various HWE tests on a subset of the Freeze 5 variant calls from the high-coverage (~37×) whole genome sequence (WGS) data in the TOPMed Project (TALIUN *et al.* 2019). We identified 17,524 HQ variants and 329,699 LQ variants using the same criteria used for 1000G variants and evaluated raw GTs and GLs. We did not evaluate PCAngsd due to excessive computational time (see “Evaluation of computational cost” below).

We first evaluated the false positive rates of different HWE tests indirectly by using HQ variants. With a >20-fold larger sample size than 1000G, we identified more significant HQ variants, while the false positive rates were still reasonable with adjusted tests. At $p < 10^{-6}$, 74% of HQ variants were significant with unadjusted tests, while the adjusted GL-based tests identified ~0.3% at $p < 10^{-6}$ (Figure 3.4C-D, Table 3.2). Adjusted GT-based tests had only slightly higher levels of false positives at $p < 10^{-6}$. However, inflation was more noticeable at less stringent p-value thresholds suggesting that GL-based tests may be needed for larger sample sizes.

Next, we evaluated the proportions of LQ variants found to be significant by different tests to indirectly evaluate their statistical power. GT- and GL-based RUTH tests showed similar power, while meta-analysis showed considerably lower power. For example, at $p < 10^{-6}$, meta-analysis identified 47% of LQ variants as significant, while

RUTH tests identified 54-58%. This pattern was similar across different p-value thresholds (Figure 3.4C-D) or choices of LQ variants (Supplementary Table 3.3, Supplementary Figure 3.4). Our results suggest that GL-based RUTH tests are suitable for testing HWE for tens of thousands of deeply sequenced genomes with diverse ancestries, but that using raw GTs will also result in a comparable performance at typically used HWE p-value thresholds (e.g. $p < 10^{-6}$) when performing QC without access to GLs.

We used PRCs to evaluate the tradeoff between empirical false positive rates and power. Consistent with previous results, the GL-based RUTH test showed the best tradeoff between false positives and power, while the GT-based RUTH test and meta-analysis were slightly less robust but largely comparable (Supplementary Figure 3.3). Notably, when we evaluated the different methods at an empirical false positive rate of 0.1%, RUTH score tests had ~4% higher power than RUTH LRT for both raw GTs and GLs (Supplementary Figures 3.5 and 3.6).

3.3.5 Impact of ancestry estimation accuracy on HWE tests

So far, our evaluations relied on genetic ancestry estimates carefully determined with sophisticated methods (see Materials and Methods). However, simpler approaches may be used instead during the variant QC step, which may affect the performance of adjusted HWE tests. We evaluated whether the number of PC coordinates affected the performance of RUTH tests by comparing the performance of RUTH tests when using 2 PCs to using 4 PCs (default). The results from both simulated and real datasets consistently demonstrated that using 4 PCs led to substantially reduced Type I errors compared to using 2 PCs at a similar level of power (Supplementary Table 3.2 and 3.4, Supplementary Figure 3.7). PRCs also clearly showed that using 4 PCs was more robust against population structure across both simulated and real datasets (Supplementary Figure 3.8).

We also evaluated whether the classification accuracy of subpopulations affected the performance of meta-analysis. Instead of assigning 1000 Genomes individuals into five continental populations, we used the k-means algorithm on those samples' top 2 PCs to classify them into 3 crude subpopulations (Supplementary Figure 3.9). This led to a much higher false positive rate with virtually no increase in true positives (Supplementary Figure 3.10, Supplementary Table 3.2). We saw the same pattern in simulated data (Supplementary Figure 3.8, Supplementary Table 3.5).

3.3.6 Computational cost

We compared the computational costs of RUTH and PCAngsd for simulated and real data. RUTH has linear time complexity to sample size, while PCAngsd appears to have quadratic time complexity (Table 3.3, Supplementary Table 3.6). RUTH also has low memory requirement compared to PCAngsd (for example, 14 MB vs 2 GB for 1000 Genomes data). Extrapolating our results to the whole genome scale, analyzing 1000 Genomes (i.e. 80 million variants) is expected to take 120 CPU-hours for RUTH, and 3,200 CPU-hours for PCAngsd (with >1 TB memory consumption). Additionally, RUTH can be parallelized into smaller regions in a straightforward manner.

3.4 Discussion

RUTH is a unified, flexible, and robust approach to incorporate genetic ancestry and genotype uncertainty for testing Hardy-Weinberg Equilibrium capable of handling large amounts of genotype data with structured populations. Sha and Zhang (2011) proposed HWES, an HWE test for structured populations, to address some of these challenges, but it has not been widely used due to the lack of an implementation that supports widely used genotype data formats (e.g. PED, BED, VCF, or BCF) and inability to handle imputed or uncertain genotypes. Hao and colleagues (2016) proposed sHWE which can only handle best-guess (hard call) genotypes (i.e. 0, 1, or 2 for biallelic variants) and does not account for genotype uncertainty. MEISNER AND ALBRECHTSEN (2019) proposed PCAngsd to address some of these issues, but it does not support the standard VCF/BCF formats for sequence-based genotypes, and its current implementation scales poorly with genome-wide analyses of large samples.

Similar to previous studies (SHA AND ZHANG 2011; HAO *et al.* 2016), our proposed framework uses individual-specific allele frequencies rather than allele frequencies pooled across all samples to systematically account for population structure in HWE tests. Unlike previous studies, we model genotype uncertainty in sequence-based genotypes in a likelihood-based framework. We implemented two RUTH tests – a score test and a likelihood ratio test (LRT) – to test for HWE under population structure for genotypes with uncertainty. While RUTH LRT is similar to the independently developed PCAngsd, the software implementation of RUTH is more flexible, scales much better to large studies, and supports the standard VCF format.

We provide a comprehensive evaluation of various approaches for testing HWE using simulated and real data. Our results demonstrated that modeling population stratification is necessary for HWE tests on heterogenous populations. We showed that accounting for genotype uncertainty via genotype likelihoods performs substantially better than testing HWE with best-guess genotypes, especially for low-coverage sequenced genomes. Importantly, we included the evaluations for an unpublished but commonly used approach – meta-analysis across stratified subpopulations, cohorts, or batches. Our results demonstrate that meta-analysis may be effective in reducing false positives, but at the expense of substantially reduced power compared to RUTH.

We observed that the current implementation of PCAngsd does not scale well to large-scale sequencing data, though in principle it can be implemented more efficiently, because the underlying HWE test itself is similar to RUTH LRT. PCAngsd requires loading all genotypes into memory, which is often infeasible for large sequencing studies. For example, loading all of 1000 Genomes will require ~4.8 TB of memory. In our evaluation of 1000G chromosome 20 variants, the inability of PCAngsd to estimate PCs from the whole genome may have contributed to the observed difference in results from RUTH compared to our simulation studies.

Although our 1000G experiments demonstrated the unexpected result that using raw GTs had better sensitivity than using LD-aware GTs at the same empirical false positive rates for low-coverage data, we do not advocate using raw GTs for low-coverage sequence data. First, the results for raw GTs were still consistently less robust than GL-based RUTH tests. Moreover, it would be tricky to determine an appropriate p-value threshold when the false positives are severely inflated. Therefore, we strongly advocate using GL-based RUTH tests for robust HWE tests with low-coverage sequence data. For the now more typical high-coverage sequence data, GL-based tests are still preferred, but GT-based RUTH tests should be acceptable for cases in which genotype likelihoods are unavailable.

Our experiment compared using 2 vs 4 PCs only because *verifyBamID2* software tool estimated up to 4 PCs projected onto HGDP panel by default (ZHANG *et al.* 2020). Because our method focuses on testing HWE during the QC steps in sequence-based variant calls, a curated version of PCs, estimated from sequenced cohort themselves, may not be readily available at the time of HWE test. However, it is possible to use a larger number of PCs (e.g. >10 PCs) if available at the time of HWE test. We expect that a larger number of PCs will account for finer-grained population structure and may benefit the performance of HWE test, but additional experiments are needed to quantify the impact of using larger number of PCs.

Our results demonstrate that RUTH score and LRT tests perform similarly in simulated and experimental datasets. Overall, the RUTH-LRT was slightly more powerful than the RUTH-score test at the expense of slightly greater false positive rates, although this tendency was not consistent. We observed that the RUTH tests tended to be slightly more powerful in identifying deviation from HWE in the direction of excess heterozygosity than excess homozygosity when compared to adjusted meta-analysis. These results might be caused by the difference between our model-based asymptotic tests compared to the exact test used in meta-analysis.

We did not evaluate our methods on imputed genotypes in this manuscript. Because imputed genotypes implicitly assume HWE, we suspect that HWE tests based on imputed genotypes may have reduced power compared to directly genotyped variants. It is possible to use approximate genotype likelihoods instead of best-guess genotypes for imputed genotypes, but this requires genotype probabilities, not just the genotype dosages. If genotype probabilities $\Pr\,(g_i = G|Data_i)$ are available, they can be converted to genotype likelihoods $L_i^{(G)} = \Pr\,(Data_i|g_i = G)$ using Bayes' rule by modeling $\Pr(g_i = G)$ as a binomial distribution based on allele frequencies (which implicitly assumes HWE). However, similar to LD-aware genotypes in low-coverage sequencing, the power of HWE tests with imputed genotypes may be poor. Further evaluation is needed to understand how useful this approximation will be compared to alternative methods including the use of best-guess imputed genotypes.

Our methods have room for further improvement. First, we used a truncated linear model for individual-specific allele frequencies for computational efficiency. Although such an approximation was demonstrated to be effective in practice (ZHANG *et al.* 2020), applying a logistic model or some other more sophisticated model may be more effective in improving the precision and recall of RUTH tests. Second, we did not attempt to model or evaluate the effect of admixture in our method. Because HWE is reached in two generations with random mating, accounting for admixed individuals may only have marginal impact. However, systematic evaluations focusing on admixed populations are needed to ensure that RUTH works robustly on such samples. Third, RUTH tests do not account for family structure. We suspect that the apparent inflation of Type I error for the TOPMed data was partially due to sample relatedness. Accounting for family structure in other ways, for example using variance components models, will require much longer computational times and may not be feasible for large-scale datasets. Fourth, RUTH currently does not directly support imputed genotypes or genotype dosages. In principle, it is possible to convert posterior probabilities for imputed genotypes into genotype likelihoods to account for genotype uncertainty (by using individual-specific allele frequencies). However, because most genotype imputation methods implicitly assume HWE, we suspect that HWE tests on imputed genotypes will be underpowered, similar to our observations with LD-aware genotypes in the 1000 Genomes dataset, even though explicitly modeling posterior probabilities may slightly mitigate this reduction in power.

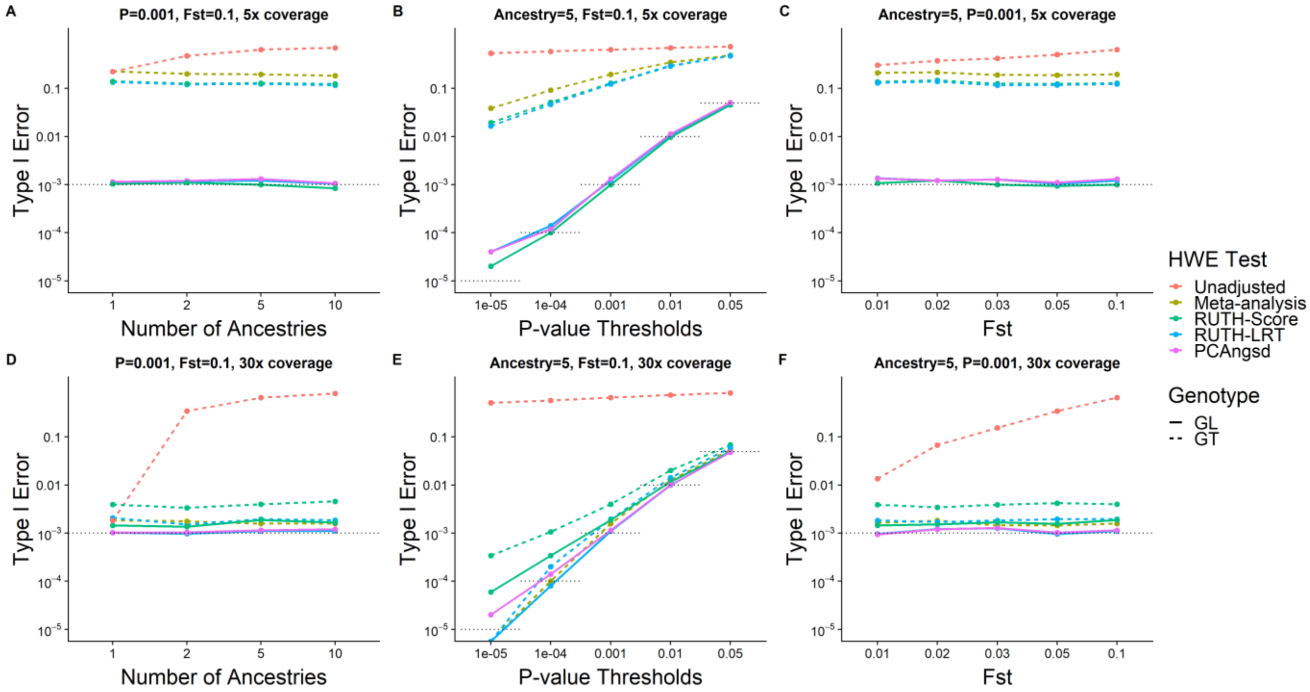
In summary, we have developed and implemented robust and rapid methods and software tools to enable HWE tests that account for population structure and genotype uncertainty. We performed comprehensive evaluations of both our methods and alternative approaches. Our tools can be used to evaluate variant quality in very large-scale genetic data sets, with the ability to handle standard VCF formats for storing sequence-based genotypes. Our software tools are publicly available at http://github.com/statgen/ruth.

3.5 Acknowledgements

This work was supported by NIH grants HL137182 (from NHLBI), HG009976 (from NHGRI), HG007022 (from NHGRI), DA037904 (from NIDA), HL117626-05-S2 (from NHLBI), and MH105653 (from NIMH). Molecular data for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). Core support including centralized genomic read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I). Core support including phenotype harmonization, data management, sample-identity QC, and general program coordination were provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN268201800001I). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed.

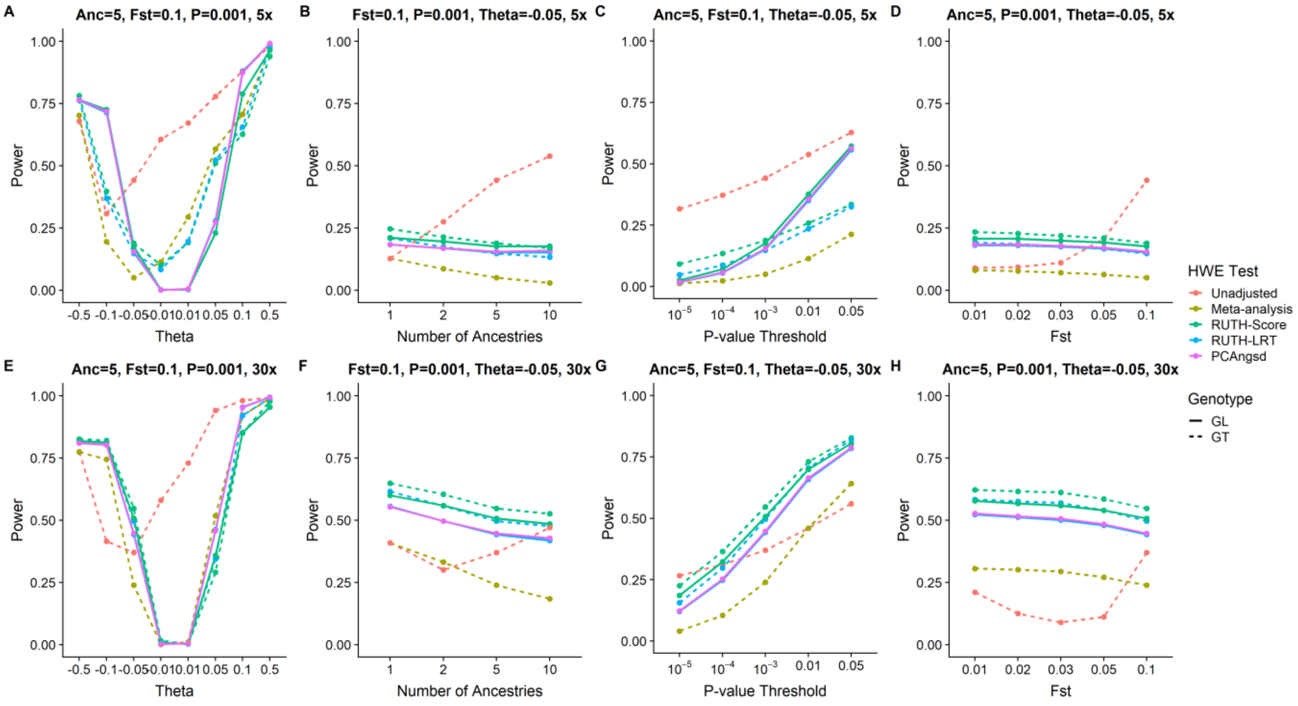
TOPMed source studies and sample counts are described in Supplementary Table 3.7. Acknowledgements for TOPMed omics support are detailed in Supplementary Table 3.8. Full TOPMed study acknowledgements are listed in Supplementary Text 3.1.

Figure 3.1 Evaluation of Type I errors between variant HWE tests on simulated genotypes



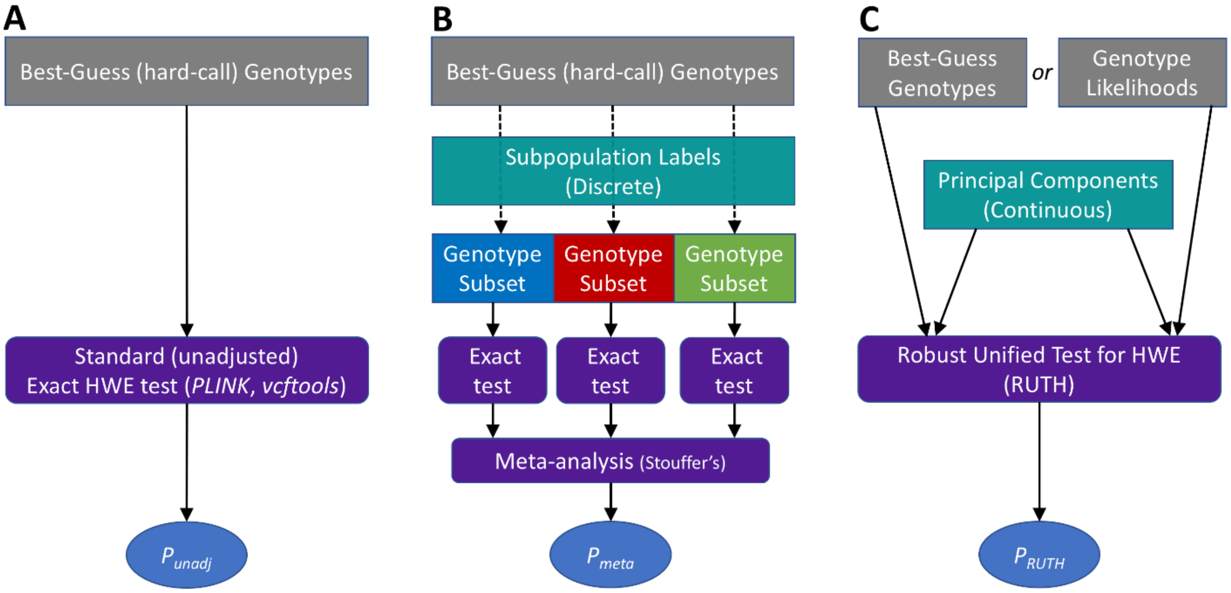
Under each combination of simulation conditions (number of ancestries, sequencing coverage, and fixation index), we simulated 5,000 samples with 50,000 variants that follow HWE within each of the subpopulations and determined the Type I error performances of different HWE tests based on the proportion of variants labeled as having significant p-values. Five HWE tests—(1) Unadjusted HWE test (WIGGINTON *et al.* 2005) implemented in PLINK-1.9 (PURCELL *et al.* 2007) using hard genotypes, (2) meta-analysis using Stouffer's method across ancestries using hard genotypes (GT), (3) RUTH test using hard genotypes, (4) RUTH test using phred-scale likelihood (GL) computed from simulated sequence reads, and (5) PCAngsd (MEISNER AND ALBRECHTSEN 2019)—were tested under HWE with various parameter settings. Gray dotted lines indicate targeted Type I Error rates. Top panels (A-C) represent results from shallow sequencing (5x), and the bottom panels (D-F) represent results from deep sequencing (30x). Using GL-based genotypes resulted in Type I Error rates closer to the targeted rate than using GT-based genotypes across different numbers of ancestries (A, D), P-value thresholds (B, E), and fixation indices (C, F). The difference is especially large for low-coverage genotypes.

Figure 3.2 Evaluation of power between different HWE tests on simulated genotypes



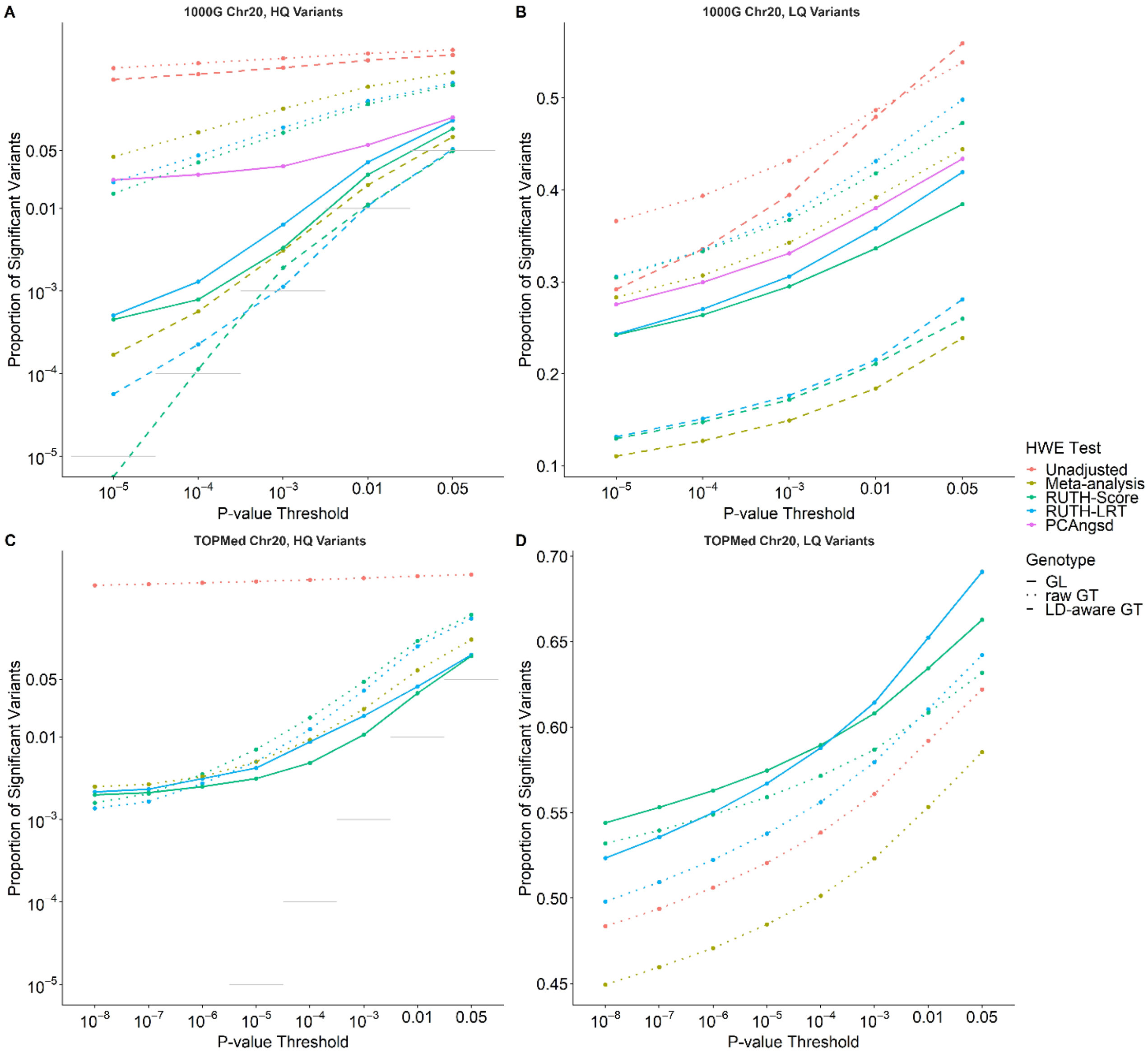
Under each combination of simulation conditions (number of ancestries, sequencing coverage, fixation index, and deviation from HWE), we simulated 50,000 variants for 5,000 samples and evaluated the ability of different HWE tests to find the variants significant. Unless otherwise specified, the default simulation parameters are 5 ancestries, with F_{ST}=.1, P-value threshold=.001, and Theta=-0.05. Tests that can find a larger proportion of significant variants are considered more powerful. Five HWE tests— (1) Unadjusted HWE test (WIGGINTON *et al.* 2005) implemented in PLINK-1.9 using hard genotypes (2) RUTH test using hard genotypes, (3) RUTH test using phred-scale likelihood (PL) computed from simulated sequence reads, (4) meta-analysis using Stouffer's method across ancestries using hard genotypes, and (5) PCAngsd (MEISNER AND ALBRECHTSEN 2019)—were tested for variants deviating from HWE with various parameter settings, for low coverage (A-D) and high coverage (E-H) data. (A, E) Theta controls the degree of deviation from HWE, with negative values indicating excess heterozygosity and positive values indicating heterozygote depletion. The high Type I Error rates in GT-based tests (Figure 2) lead to those methods appearing to have higher power in some scenarios. The unadjusted test suffers from this problem the most. GL-based methods have slightly lower powers than GT-based methods in exchange for a much better controlled Type I error rate. This pattern mostly holds across different numbers of ancestries (B, F), p-value thresholds (C, G), and fixation indices (D, H). Meta-analysis had the lowest power in the presence of excess heterozygosity.

Figure 3.3 Schematic diagrams of different methods to test HWE under population structure



Three different methods to test HWE under population structure are described. (A) In the standard (unadjusted) HWE test, all samples are tested together using best-guess genotypes. This test does not adjust for sample ancestry. (B) In a meta-analysis of stratified HWE tests, the samples must first be categorized into discrete subpopulations, determined a priori based on their genotypes or self-reported ancestries. Next, standard HWE tests (based on best-guess genotypes) are performed on each of these subpopulations. Then, the resulting HWE statistics are converted into Z-scores and combined in a meta-analysis using Stouffer's method, with the sample sizes of the subpopulations as weights. (C) In our proposed method (RUTH), either best-guess genotypes or genotype likelihoods can be used as input for HWE test. We assume that the genetic ancestries of each sample are estimated a priori, typically as principal components (PCs). We combine the genotypes and PCs to perform either a score test or a likelihood ratio test to obtain a joint ancestry adjusted HWE statistic for each variant across all samples.

Figure 3.4 Evaluation of different HWE tests on 1000 Genomes and TOPMed variants



In 1000 Genomes data (A, B), we identified 17,740 “high quality” (HQ) variants and 10,966 “low quality” (LQ) variants in chromosome 20. In TOPMed data (C, D), we identified 17,524 HQ variants and 329,699 LQ variants in chromosome 20. A well-behaved HWE test should maximize the proportion of significant LQ variants while controlling the false positive rate for HQ variants. Dotted gray lines represent targeted Type I error levels if we assume all HQ variants follow HWE. (A) Both the unadjusted test and PCAngsd found substantially more significant variants than expected in the 1000G HQ variant set, while both RUTH and meta-analysis were more conservative. Methods that used raw GTs showed substantial false positive rates, while methods that used GLs and LD-aware GTs had much better control of false positives. (B) In 1000G LQ variants, meta-analysis lagged behind RUTH and the unadjusted test in discovering significant deviation from HWE. RUTH behaved well for HQ variants while having more power to find low-quality variants significantly deviating from HWE. (C) In TOPMed data, the unadjusted test resulted in an excess of false positives. Tests using GL-based genotypes outperformed tests using GT-based genotypes. (D) Methods using GL-based genotypes were able to discover more LQ variants than methods using GT-based genotypes, demonstrating the advantage of accounting for genotype uncertainty in HWE tests.

Table 3.1 Performance of the unadjusted test, meta-analysis, RUTH, and PCAngsd on 1000 Genomes chromosome 20 variants.

Variant Category	Genotype Format	HWE Test	Proportion of Significant Variants					Total Variant Count
			$P < 10^{-2}$	$P < 10^{-3}$	$P < 10^{-4}$	$P < 10^{-5}$	$P < 10^{-6}$	
LQ Variants	raw GT	Unadjusted	0.487	0.432	0.394	0.366	0.339	10,966
		Meta-analysis	0.392	0.343	0.307	0.283	0.262	10,966
		RUTH-Score	0.418	0.367	0.333	0.305	0.284	10,966
		RUTH-LRT	0.431	0.373	0.335	0.305	0.280	10,966
	LD-aware GT	Unadjusted	0.479	0.395	0.336	0.292	0.259	10,966
		Meta-analysis	0.184	0.149	0.127	0.111	0.098	10,966
		RUTH-Score	0.211	0.172	0.147	0.130	0.112	10,966
		RUTH-LRT	0.215	0.177	0.151	0.131	0.115	10,966
	GL	RUTH-Score	0.336	0.295	0.264	0.242	0.223	10,966
		RUTH-LRT	0.358	0.306	0.270	0.243	0.225	10,966
		PCAngsd	0.380	0.331	0.300	0.275	0.255	10,920
HQ Variants	raw GT	Unadjusted	0.755	0.657	0.573	0.501	0.443	17,740
		Meta-analysis	0.298	0.161	0.084	0.042	0.020	17,740
		RUTH-Score	0.183	0.083	0.036	0.015	7.4 x 10^{-3}	17,740
		RUTH-LRT	0.200	0.095	0.044	0.021	0.010	17,740
	LD-aware GT	Unadjusted	0.623	0.507	0.422	0.361	0.311	17,740
		Meta-analysis	0.019	3.1 x 10^{-3}	5.6 x 10^{-4}	1.7 x 10^{-4}	1.1 x 10^{-4}	17,740
		RUTH-Score	0.011	1.9 x 10^{-3}	1.1 x 10^{-4}	0	0	17,740
		RUTH-LRT	0.011	1.1 x 10^{-3}	2.3 x 10^{-4}	5.6 x 10^{-5}	0	17,740
	GL	RUTH-Score	0.026	3.3 x 10^{-3}	7.9 x 10^{-4}	4.5 x 10^{-4}	3.4 x 10^{-4}	17,740
		RUTH-LRT	0.036	6.4 x 10^{-3}	1.3 x 10^{-3}	5.1 x 10^{-4}	3.4 x 10^{-4}	17,740
		PCAngsd	0.059	0.032	0.026	0.022	0.021	17,740

The numbers within cells represent the proportions of significant variants under the corresponding testing conditions at the given P-value threshold. We expect our LQ variants to violate HWE at a higher rate than our HQ variants. A well-behaved test is expected to find a high proportion of LQ variants to be significant while maintaining the targeted Type I Error rate in HQ variants. The unadjusted test consistently shows the highest false positive rate among all the tests. HWE tests that rely on raw GTs also show much higher false positive rates than tests that use other genotype representations. RUTH tests were the best at controlling false positives while still maintaining comparable power to the other methods. PCAngsd had a much higher false positive rate than RUTH-based methods, especially at more stringent p-value thresholds.

Table 3.2 Performance of the unadjusted test, meta-analysis, and RUTH on TOPMed freeze 5 chromosome 20 variants

Variant set	Genotype Format	HWE Test	Proportion of Significant Variants					Total Variant Count
			$P < 10^{-2}$	$P < 10^{-3}$	$P < 10^{-4}$	$P < 10^{-5}$	$P < 10^{-6}$	
LQ Variants	raw GT	Unadjusted	0.592	0.561	0.539	0.521	0.506	329,699
	raw GT	Meta-analysis	0.554	0.524	0.502	0.485	0.471	329,699
	raw GT	RUTH-Score	0.608	0.587	0.572	0.559	0.549	329,699
	GL	RUTH-Score	0.635	0.608	0.590	0.575	0.563	329,699
	raw GT	RUTH-LRT	0.610	0.580	0.556	0.538	0.522	329,699
	GL	RUTH-LRT	0.653	0.615	0.588	0.567	0.550	329,699
HQ Variants	raw GT	Unadjusted	0.890	0.842	0.800	0.766	0.736	17,524
	raw GT	Meta-analysis	0.065	0.022	9.0 x 10^{-3}	4.8 x 10^{-3}	3.3 x 10^{-3}	17,524
	raw GT	RUTH-Score	0.145	0.047	0.172	7.1 x 10^{-3}	3.5 x 10^{-3}	17,524
	GL	RUTH-Score	0.034	0.011	4.9 x 10^{-3}	3.1 x 10^{-3}	2.5 x 10^{-3}	17,524
	raw GT	RUTH-LRT	0.125	0.036	0.012	5.0 x 10^{-3}	2.7 x 10^{-3}	17,524
	GL	RUTH-LRT	0.041	0.018	8.5 x 10^{-3}	4.3 x 10^{-3}	3.1 x 10^{-3}	17,524

The numbers within cells represent the proportions of significant variants under the corresponding testing conditions at the given P-value threshold. These results are based on tests that used likelihood-based genotype representations as input. A well-behaved test should reduce the number of significant high-quality (HQ) variants while increasing the number of significant low-quality (LQ) variants. The unadjusted test had a greatly inflated false positive rate for HQ variants while showing a lower true positive rate for LQ variants. While meta-analysis performed better for HQ variants, it had reduced power to find LQ variants to be significant. RUTH performed the best, with fewer false positives (significant HQ variants) compared to both the unadjusted test and meta-analysis, while at the same time finding more true positives (significant LQ variants).

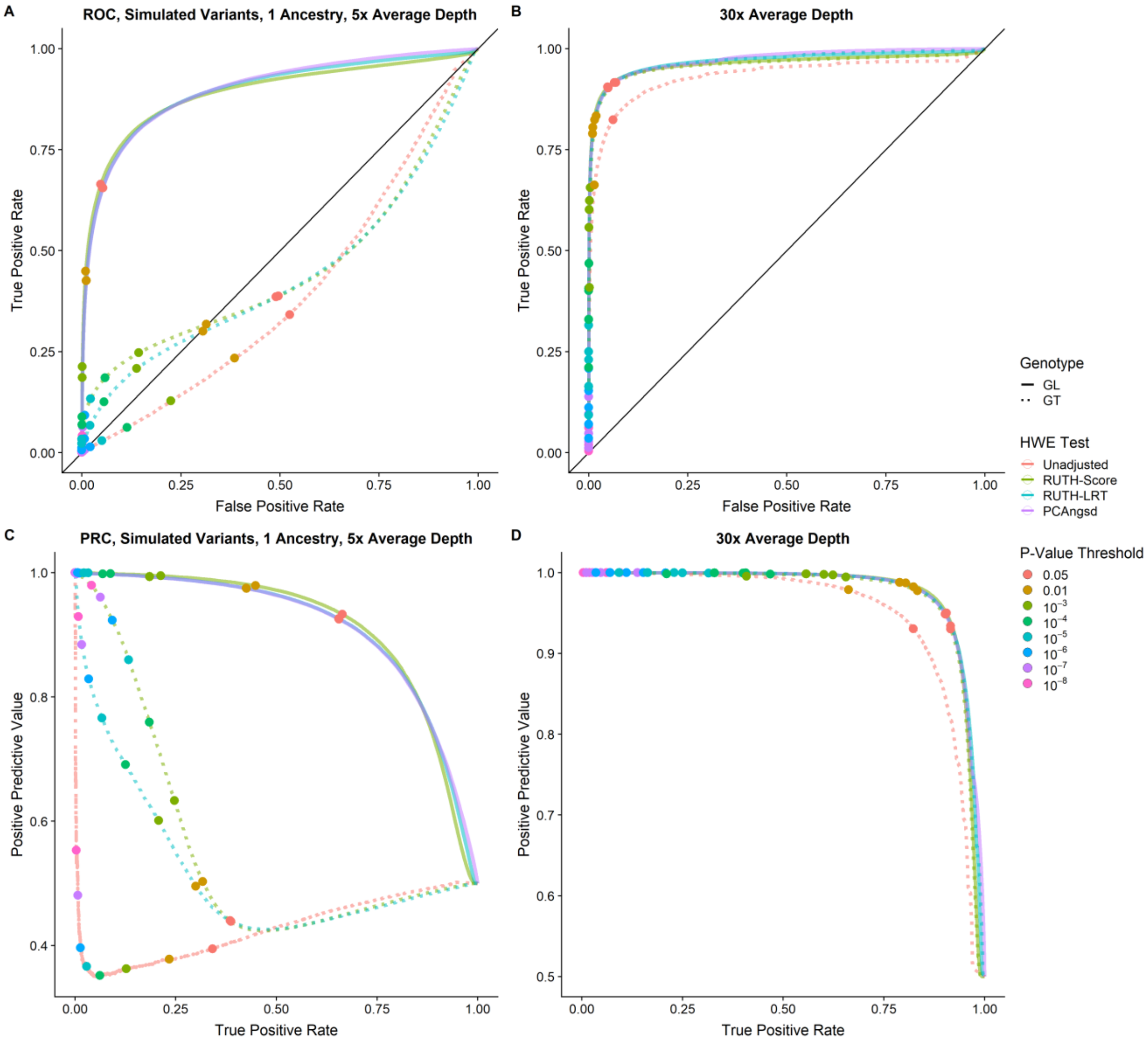
Table 3.3 Runtimes for RUTH and PCAngsd on simulated data

Sample Size	Wall Time (s)			User Time (s)		
	RUTH-LRT	RUTH-	PCAngsd	RUTH-LRT	RUTH-	PCAngsd
1,000	16.21	27.24	173.11	16.16	27.09	172.37
2,000	32.19	54.63	347.10	31.94	54.51	345.58
5,000	82.80	136.44	1,124.83	81.81	136.20	1,102.85
10,000	165.48	273.67	7,396.00	163.88	273.27	7,235.91
20,000	336.75	553.92	38,807.67	332.06	553.05	37,338.69
50,000	902.81	1,438.32	461,971.33	886.67	1,435.87	403,296.5

We simulated 10,000 genotype likelihood-based variants for varying numbers of samples. Wall time indicates total runtime, while user time is the amount of time the CPUs spent running each program. All programs were run in single-threaded mode. System processes make up the difference between the two values, with a majority consisting of file I/O. We used VCF files with GL fields in RUTH and converted them to Beagle3 format for PCAngsd. The RUTH likelihood ratio test (LRT) was the fastest method, with the score test about 60% slower. PCAngsd was about 10 times slower than RUTH-LRT with the smallest sample sizes and over 400 times slower with our largest tested size of 50,000 samples.

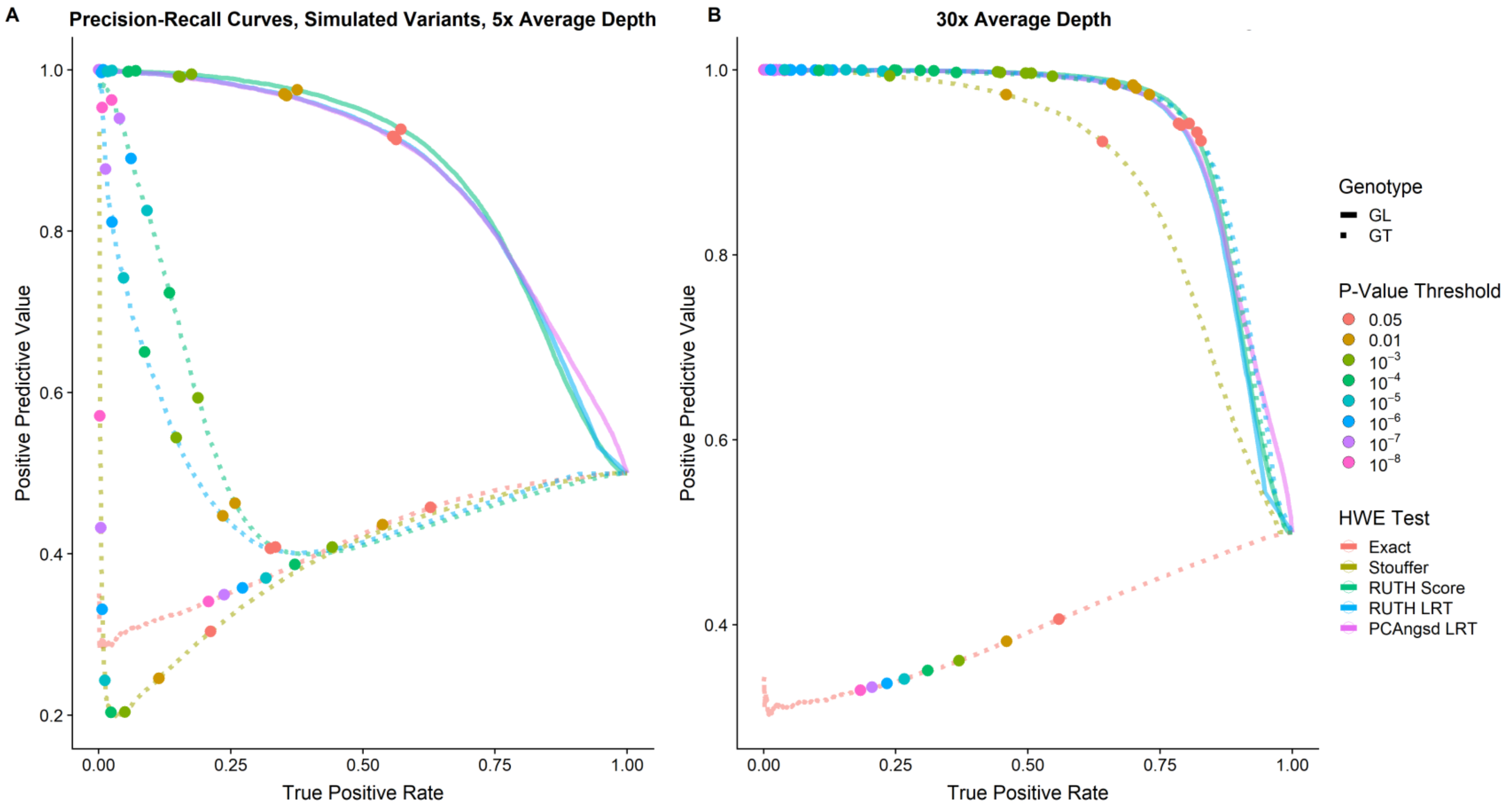
3.6 Appendix: Supplementary Figures and Tables

Supplementary Figure 3.1 ROC and PRC for simulated single-ancestry data



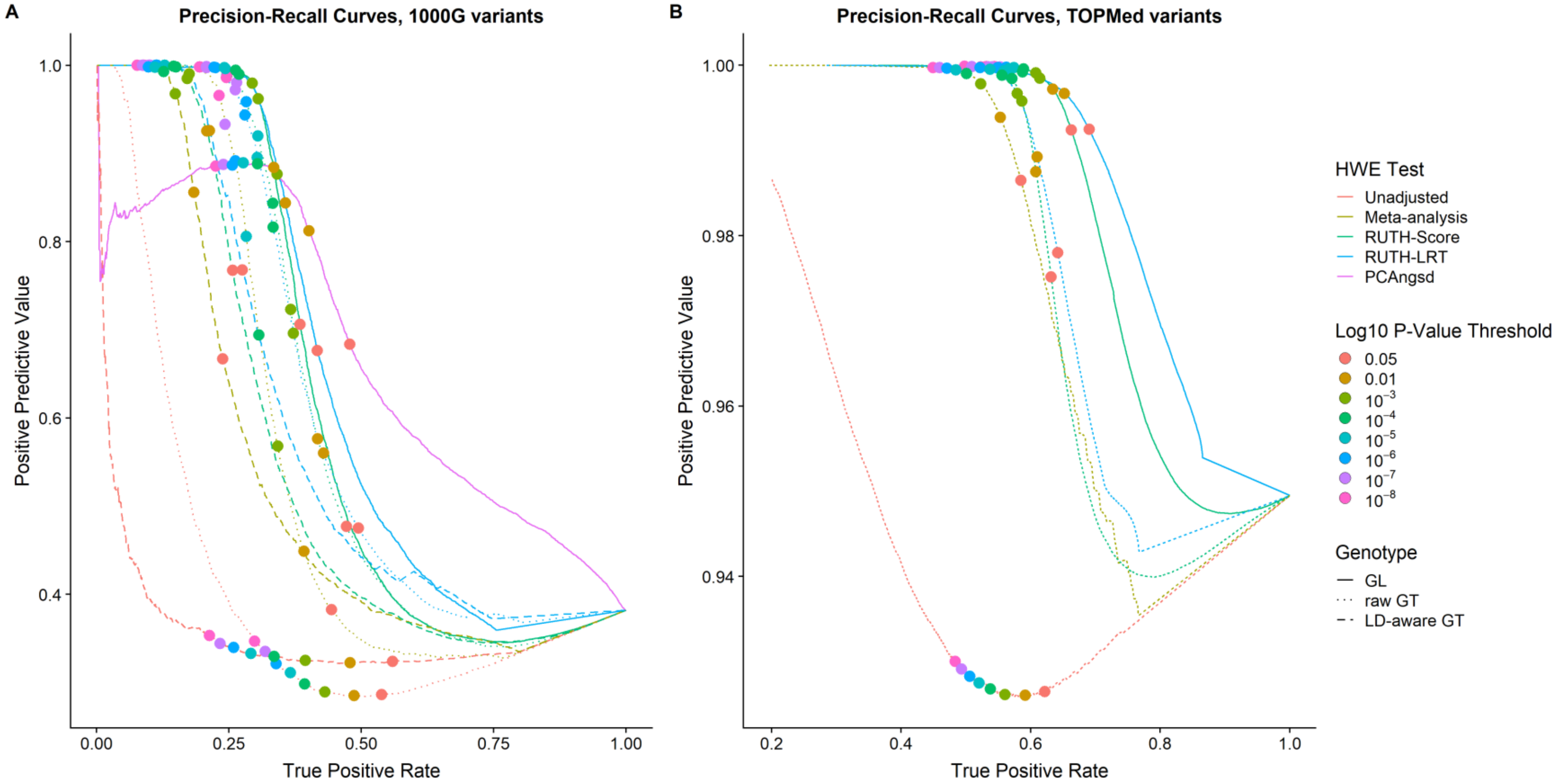
For both low coverage (A, C) and high coverage (B, D) settings, 500,000 variants were generated from 5,000 samples arising from a single ancestry, with half of the variants as true positives (θ = -0.05) and half of the variants as true negatives (θ = 0). The colors of the lines correspond to the different HWE tests, while the colors of the points correspond to different P-value thresholds. In all cases, the unadjusted test performed the worst. For low-coverage data, tests using GT-based genotypes performed poorly due to their inability to capture the effects of genotype uncertainty, whereas tests using GL-based genotypes performed much better. The difference was negligible in high-coverage genotype data.

Supplementary Figure 3.2 Precision-recall curves for simulated data with multiple ancestries



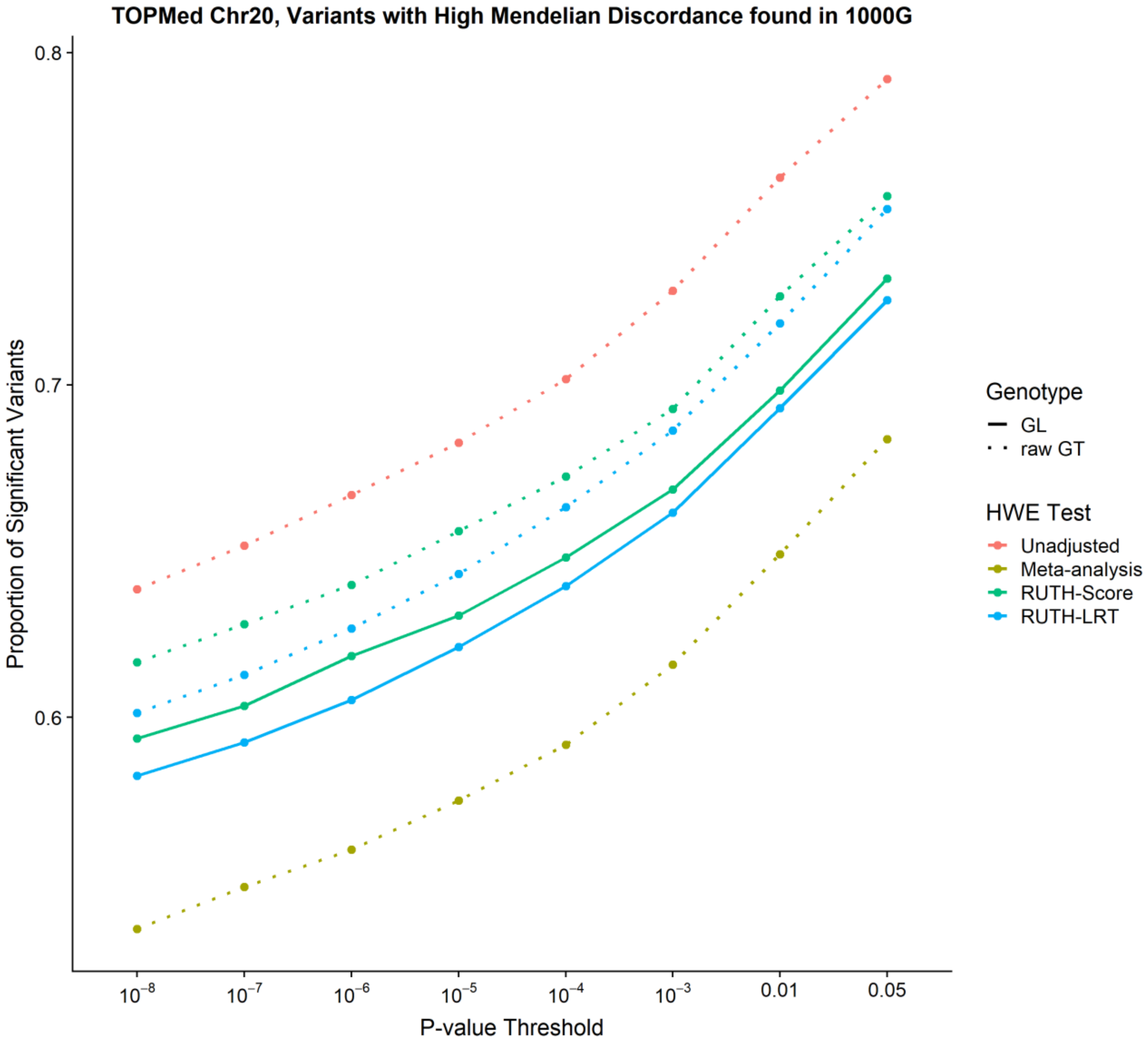
We generated Precision-recall curves to evaluate the tradeoff between the different HWE tests' ability to identify true positive variants while minimizing the misidentification of true negative variants as significantly departing from HWE. We analyzed 50,000 true positive and 50,000 true negative variants in 5,000 samples arising from 5 different ancestries with an average simulated depth of (A) 5x and (B) 30x. True negative variants are defined as variants with the HWE deviation parameter θ = 0. True positives are defined as variants with θ = -0.05. The True Positive Rate (TPR) is defined to be the proportion of variants with θ = -0.05 that are significant at a given P-value threshold, while the Positive Predictive Value (PPV) is defined as the proportion of significant variants with θ = -0.05 at the same P-value threshold. Selected p-value thresholds are indicated with colored circles. For low-depth genotypes, in the presence of high genotype uncertainty, GL-based HWE tests performed relatively well, while GT-based tests performed poorly. For high-depth genotypes, with low genotype uncertainty, all methods adjusting for population structure performed relatively well.

Supplementary Figure 3.3 Precision-recall curves for 1000G and TOPMed variants



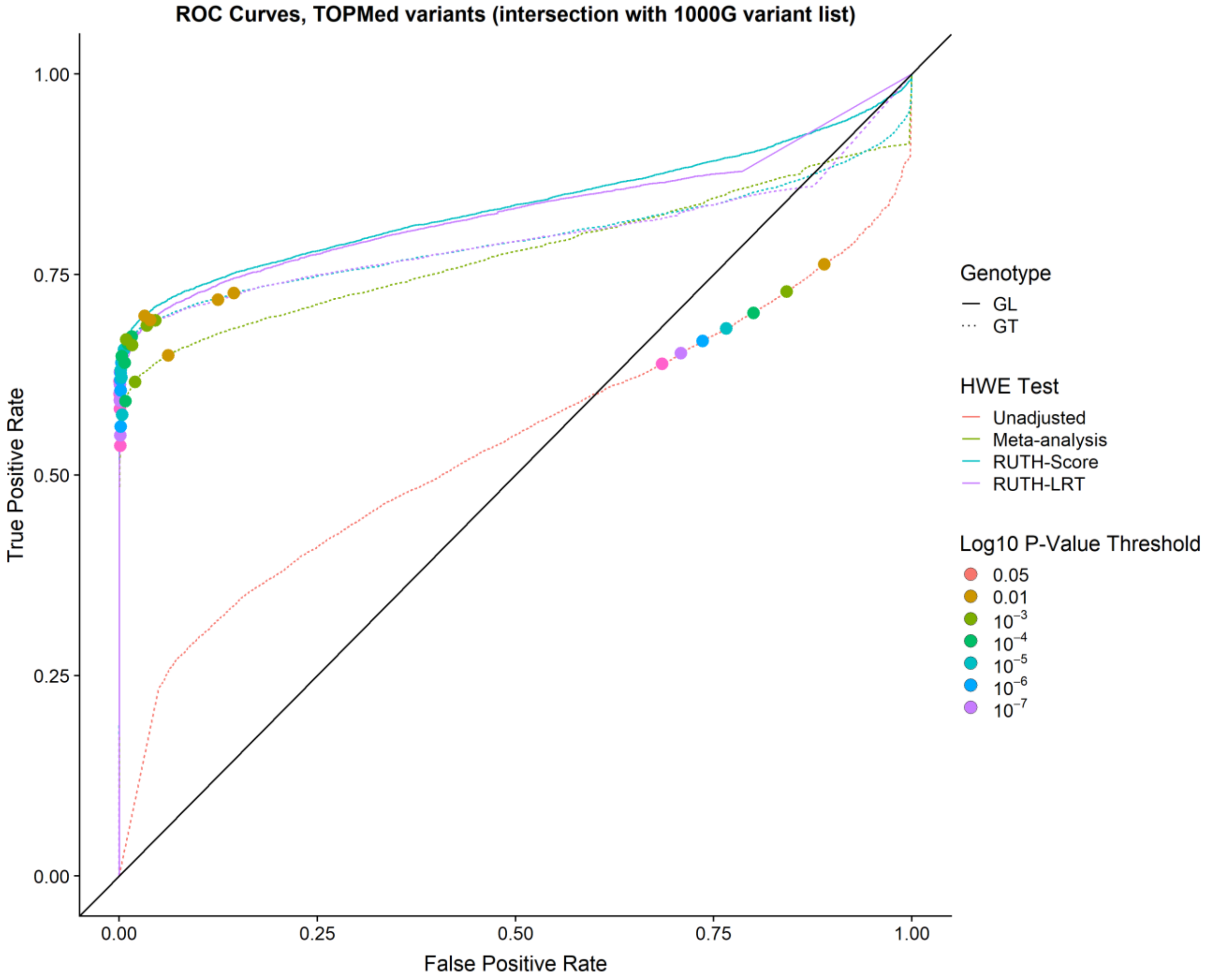
We defined positive variants as those with a high level of Mendelian inconsistency in family-based TOPMed data, and negative variants as those found in the intersection of the Illumina Omni2.5 and HapMap3 variant site lists. (A) For low-coverage sequence data found in 1000G, tests using GL-based genotypes (solid lines) generally performed better than tests using any GT-based genotypes (dotted and dashed lines). Both the unadjusted test and meta-analysis performed much worse than all other methods. (B) For high-coverage sequence data found in TOPMed, tests using GL-based genotypes retained their improved performance over tests using GT-based genotypes.

Supplementary Figure 3.4 Results of testing TOPMed variants found in 1000G variant list



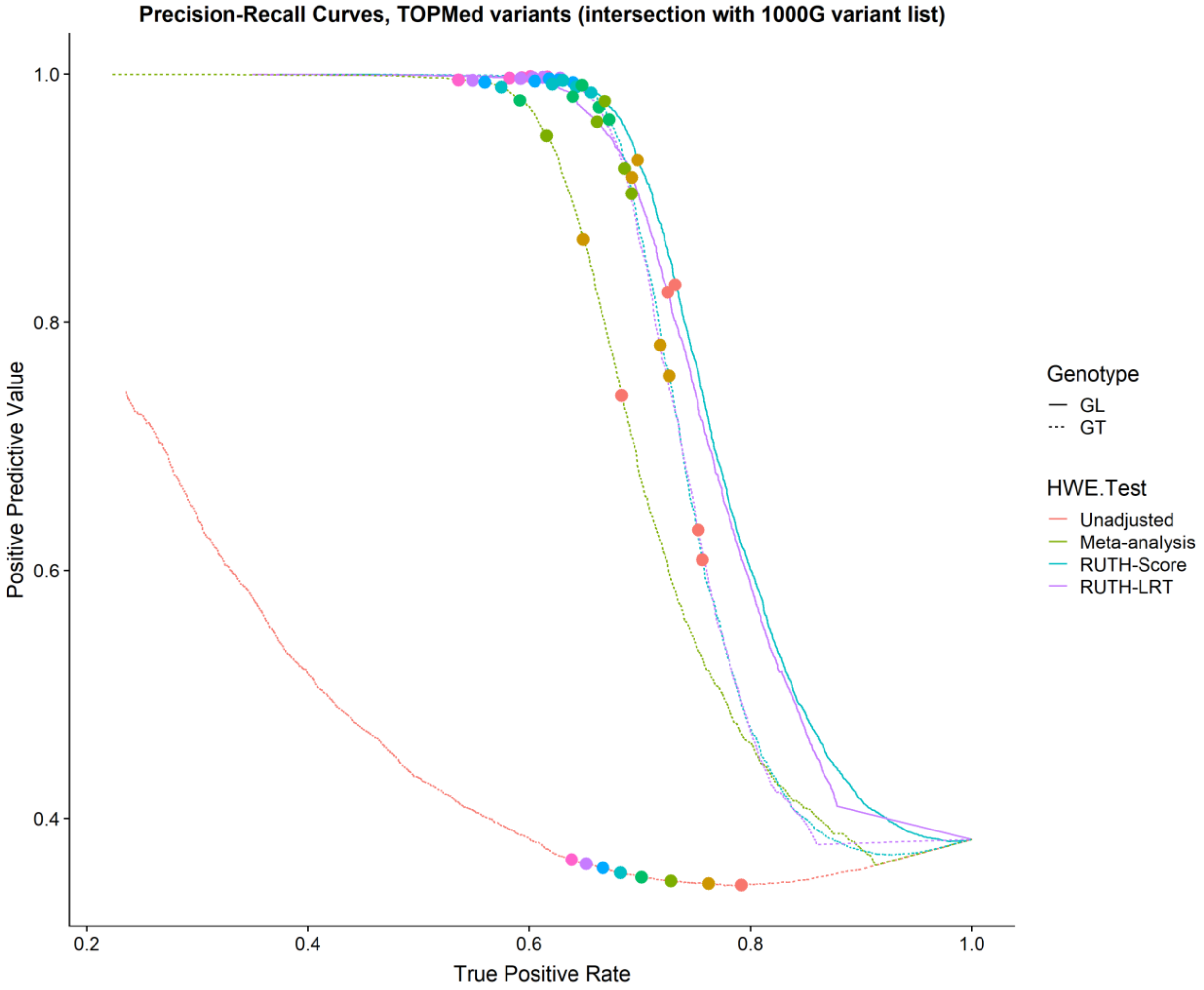
This analysis contains 10,966 TOPMed variants found to be discordant in TOPMed family data and overlapping with 1000G discordant variants, as opposed to all 329,699 discordant TOPMed variants (as seen in Figure 3.4D). Our results are similar to those for 1000G discordant variants (Figure 3.4B), suggesting that the differences between the patterns observed in 1000G and TOPMed results may have been caused by the difference in allele frequency distributions in the two data sets (Supplementary Table 1).

Supplementary Figure 3.5 ROC curves for TOPMed variants found in 1000G variant list



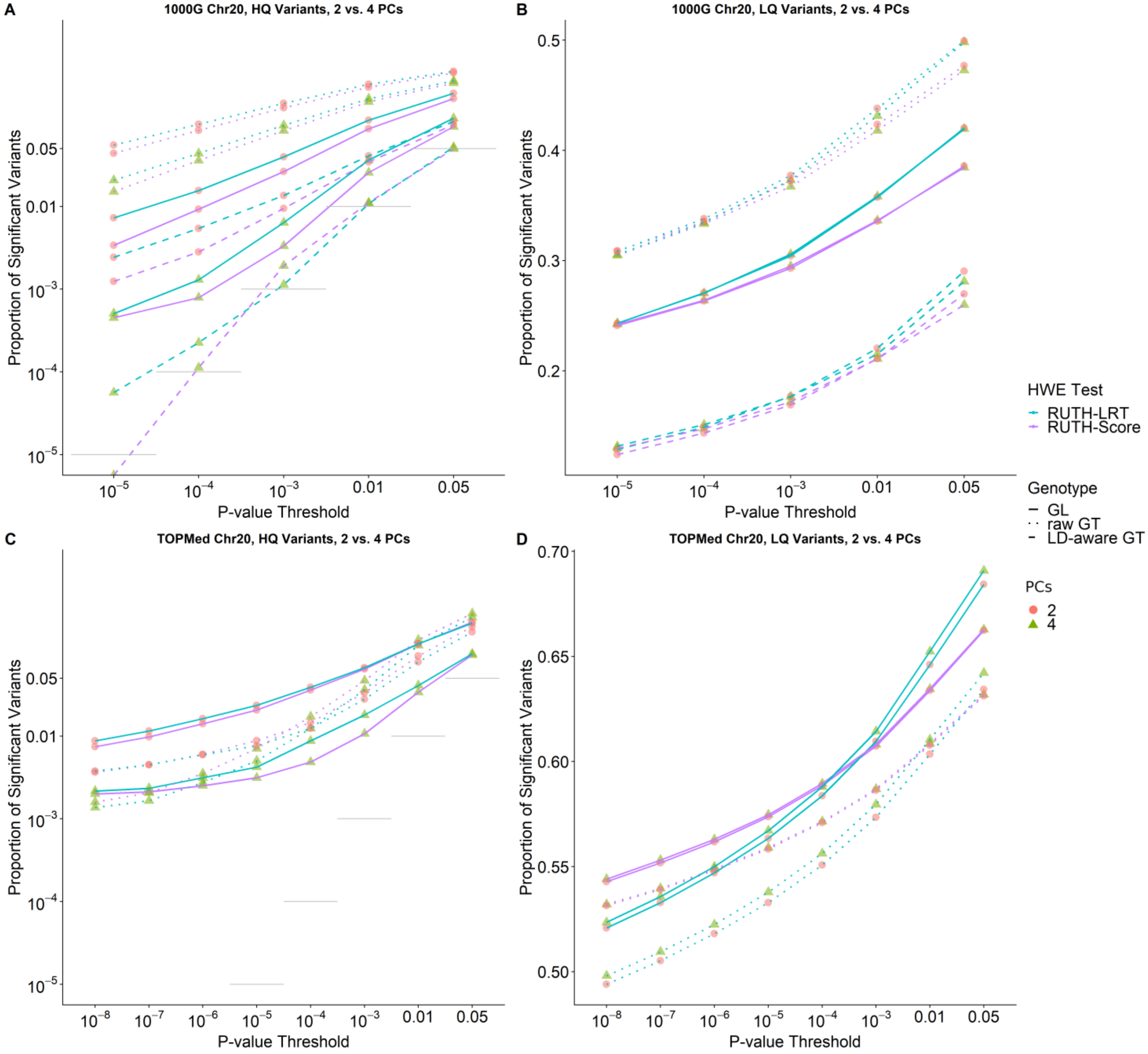
GL-based tests have the best overall performance among the different methods.

Supplementary Figure 3.6 PRC curves for TOPMed variants found in 1000G variant list



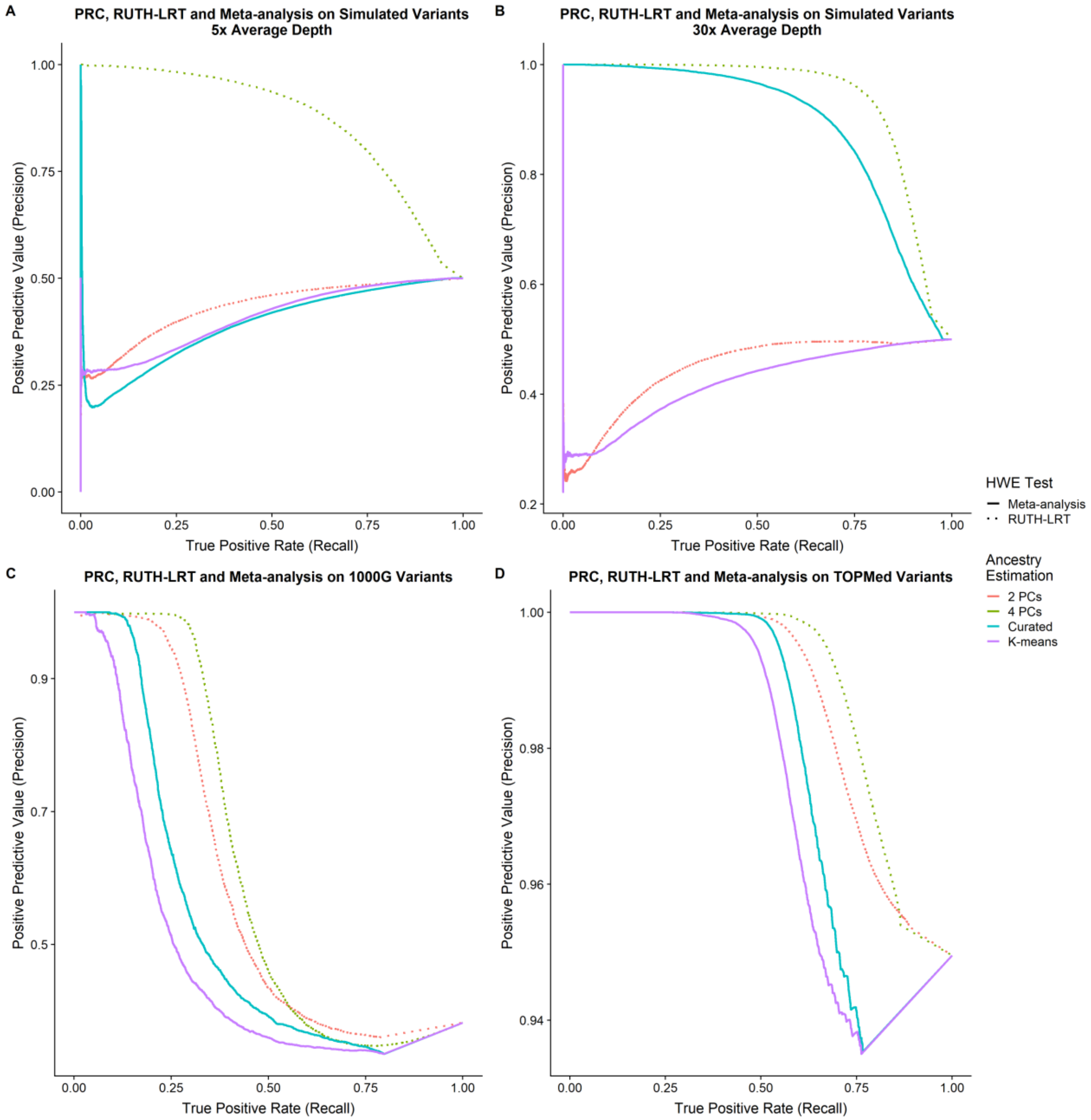
RUTH tests using GLs offer the best balance between finding true positives and maximizing positive predictive value.

Supplementary Figure 3.7 Results of testing 1000G and TOPMed variants with RUTH using two vs. four PCs



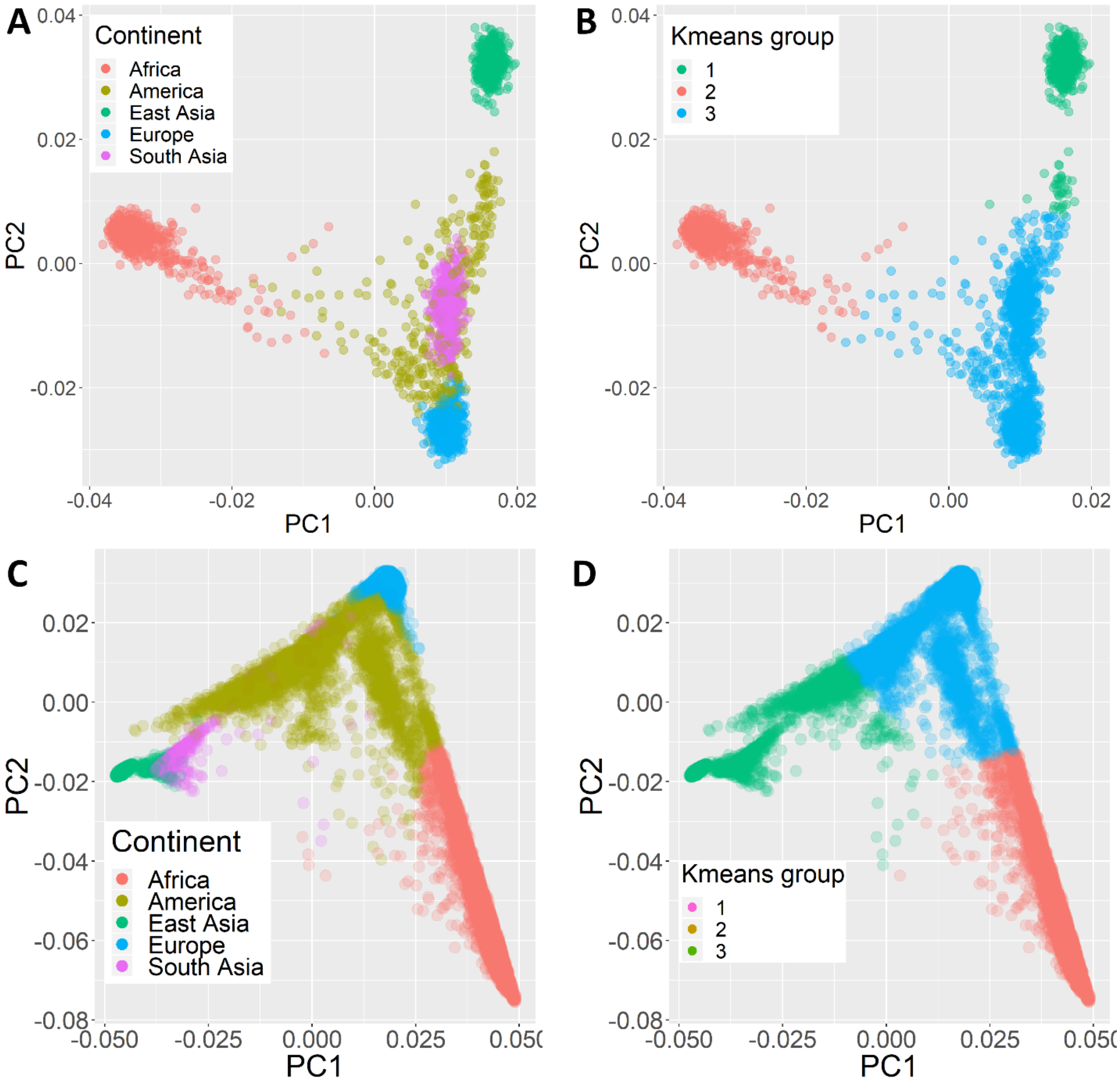
Using only 2 PCs lead to noticeably worse performance, especially for GL-based tests. (A) In 1000 Genomes data, using only 2 PCs leads to much higher false positives in HQ variants for both RUTH-Score and RUTH-LRT compared to using 4 PCs. (B) Tests on LQ variants with 2 PCs appear to have modestly higher power than tests using 4 PCs, but this is mainly due to the much higher false positive rate. (C) For HQ variants in TOPMed, tests using only 2 PCs have substantially higher false positive rate than tests using 4 PCs for GL-based tests, while GT-based tests are comparable. (D) Surprisingly, GL-based tests using 4 PCs discovered more significant LQ variants compared to GL-based tests using 2 PCs, even though GL-based tests using 2 PCs had a higher false positive rate in HQ variants.

Supplementary Figure 3.8 Effect of ancestry estimation accuracy on Precision-Recall Curves



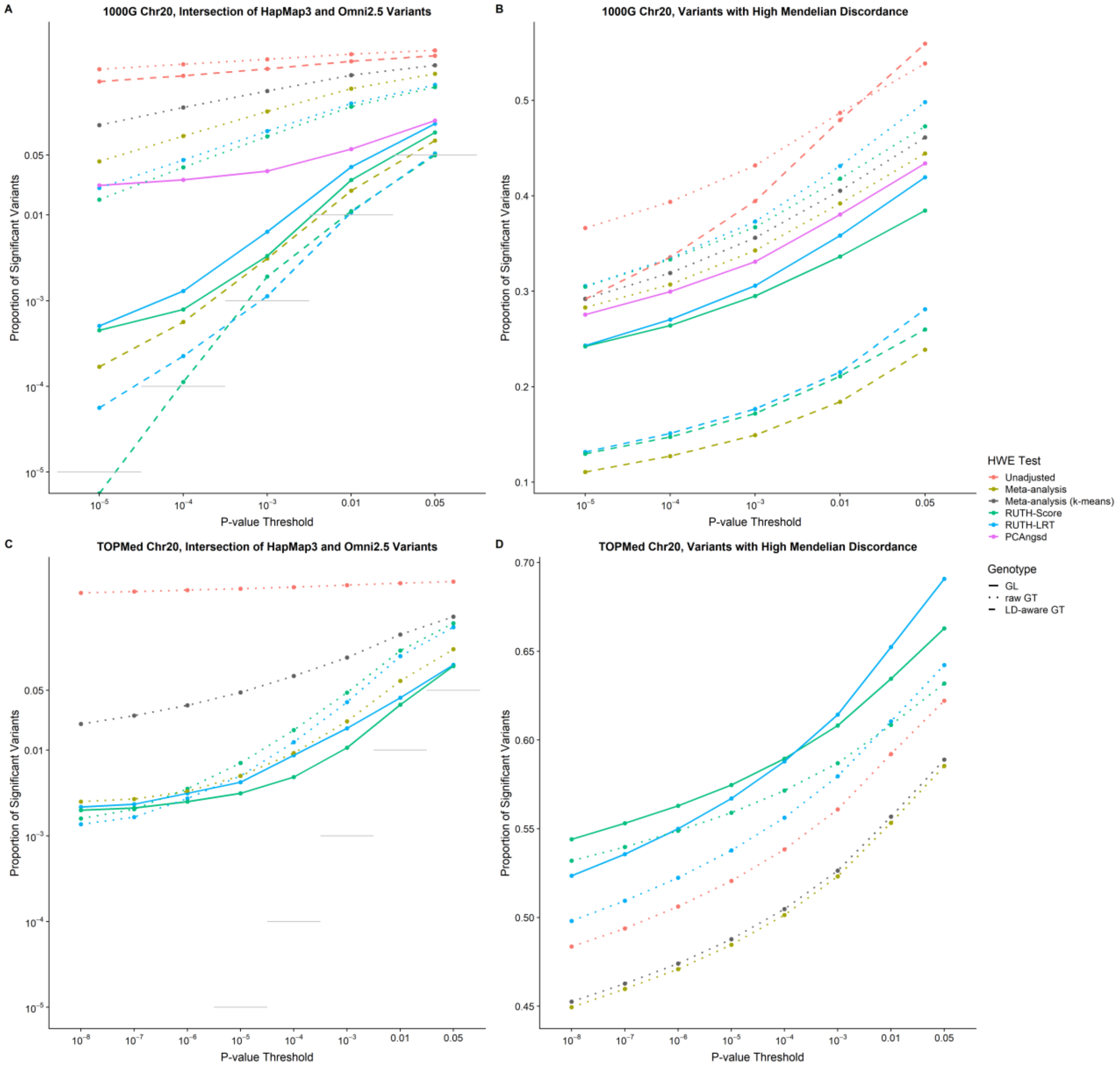
We evaluated the effect of using 2 vs. 4 principal components on the performance of RUTH-LRT, and the effect of using our nearest-neighbor algorithm (“curated”) vs. k-means for subpopulation classification of samples on the performance of meta-analysis on (A) low-depth simulated data, (B) high-depth simulated data, (C) 1000G variants, and (D) TOPMed variants. We simulated null variants with θ = 0 and alternative variants with θ = -0.05, with a fixation index of 0.1 for 5,000 samples from 5 ancestries (1,000 samples each). RUTH-LRT used GL-based genotypes, and meta-analysis used raw GT-based genotypes. K-means classification for simulated data was performed assuming 3 subpopulation clusters.

Supplementary Figure 3.9 Principal component plots and group assignments for 1000 Genomes and TOPMed samples



Ancestry group assignments for samples in 1000G (A, B) and TOPMed (C, D) samples used either a high-quality ancestry estimation method (A, C) or a crude k-means based method (B, D). In meta-analysis, samples within a group were first analyzed together using the unadjusted test. Then, the group-level results were combined using Stouffer's method. Meta-analyses using the cruder k-means groupings performed much worse than those using the high-quality ancestry estimates due to population stratification within the cruder groups.

Supplementary Figure 3.10 Results of testing 1000G and TOPMed variants with meta-analysis using K-means to generate ancestry groups



We generated three subpopulations for 1000G and TOPMed separately by applying k-means to the first two principal components of each group. Next, we calculated subpopulation-specific HWE statistics, which were converted to Z-scores and combined using Stouffer's method, using each subpopulation's size as the weights. (A) K-means-based meta-analysis had much higher false positive rates in 1000G compared to meta-analysis that used more accurate population labels, which (B) confounds its seemingly higher power to discover true positives. (C) We see the same increased false positive rate in K-means-based meta-analysis in TOPMed, but surprisingly (D) it also reduced the power to discover true positives in TOPMed. High-quality ancestry groups can substantially improve the performance of ancestry-based meta-analysis.

Supplementary Table 3.1 Simulation results for the unadjusted test, meta-analysis, RUTH, and PCAngsd for HWE

This table can be found at the following link:
https://docs.google.com/spreadsheets/d/1zdn7jOWgOMG_wwqwgDD4b1i0a2clGlyNFKmI5xR_DoE/edit?usp=sharing

Results from various HWE tests for simulations with 50,000 variants for 5,000 samples. Samples were generated using a population fixation index (F_{ST}) between .01 and .1. “GL” indicates a method using genotype likelihoods, while “GT” indicates a method using best-guess genotypes. Theta denotes deviation from HWE: Theta = 0 indicates no deviation from HWE, Theta < 0 indicates excess heterozygosity, and Theta > 0 indicates heterozygote depletion. When the samples were generated from a single ancestry, meta-analysis and the unadjusted test were identical. *Combined F_{ST} indicates the combined results for F_{ST}=.01, .02, .03, .05, and .1. This is available only when the number of ancestries is 1, because F_{ST} should not affect the results with single ancestry, so the results may be combined.

Supplementary Table 3.2 Results from using lower quality ancestry estimations on meta-analysis and RUTH

Data set	Variant set	Genotype Format	HWE Test	PCs	Proportion of Significant Variants					Total Variant Count
					$P < 0.01$	$P < 10^{-3}$	$P < 10^{-4}$	$P < 10^{-5}$	$P < 10^{-6}$	
1000G	LQ	raw GT	Meta-analysis	n/a	0.392	0.343	0.307	0.283	0.262	10,966
			Meta-analysis (k-means)	n/a	0.405	0.356	0.319	0.292	0.269	10,966
		LD-aware GT	Meta-analysis	n/a	0.184	0.149	0.127	0.111	0.098	10,966
			Meta-analysis (k-means)	n/a	0.221	0.169	0.136	0.116	0.102	10,966
	HQ	raw GT	Meta-analysis	n/a	0.298	0.161	0.084	0.042	0.020	17,740
			Meta-analysis (k-means)	n/a	0.427	0.279	0.180	0.112	0.067	17,740
		LD-aware GT	Meta-analysis	n/a	0.019	3.1 x 10^{-3}	5.6 x 10^{-4}	1.7 x 10^{-4}	1.1 x 10^{-4}	17,740
			Meta-analysis (k-means)	n/a	0.107	0.043	0.020	9.5x10^{-3}	5.0 x 10^{-3}	17,740
TOPMed	LQ	GT	Meta-analysis	n/a	0.553	0.523	0.501	0.485	0.471	329,699
			Meta-analysis (k-means)	n/a	0.557	0.526	0.505	0.488	0.474	329,699
	HQ		Meta-analysis	n/a	0.064	0.022	9.2 x 10^{-3}	5.0 x 10^{-3}	3.3 x 10^{-3}	17,524
			Meta-analysis (k-means)	n/a	0.224	0.121	0.074	0.047	0.033	17,524
1000G	LQ	GL	RUTH-LRT	2	0.357	0.304	0.271	0.243	0.224	10,966
				4	0.358	0.306	0.270	0.243	0.225	10,966
			RUTH-Score	2	0.336	0.293	0.263	0.241	0.221	10,966
				4	0.336	0.295	0.264	0.242	0.223	10,966
		LD-aware GT	RUTH-LRT	2	0.220	0.177	0.149	0.128	0.113	10,966
				4	0.215	0.177	0.151	0.131	0.115	10,966
			RUTH-Score	2	0.211	0.169	0.143	0.124	0.109	10,966
				4	0.211	0.172	0.147	0.130	0.112	10,966
		raw GT	RUTH-LRT	2	0.438	0.377	0.338	0.308	0.284	10,966
				4	0.431	0.373	0.335	0.305	0.28	10,966
			RUTH-Score	2	0.424	0.372	0.335	0.309	0.286	10,966
				4	0.418	0.367	0.333	0.305	0.284	10,966
	HQ	GL	RUTH-LRT	2	0.110	0.040	0.016	7.3 x 10^{-3}	3.3 x 10^{-3}	17,740
				4	0.036	6.4 x 10^{-3}	1.3 x 10^{-3}	5.1 x 10^{-4}	3.4 x 10^{-4}	17,740
			RUTH-Score	2	0.087	0.026	9.2 x 10^{-3}	3.4 x 10^{-3}	1.6 x 10^{-3}	17,740
				4	0.026	3.3 x 10^{-3}	7.9 x 10^{-4}	4.5 x 10^{-4}	3.4 x 10^{-4}	17,740
			RUTH-LRT	2	0.041	0.014	5.4 x 10^{-3}	2.4 x 10^{-3}	1.4 x 10^{-3}	17,740

		LD-aware GT		4	0.011	1.1×10^{-3}	2.3×10^{-4}	5.6×10^{-5}	0	17,740
			RUTH-Score	2	0.034	9.5×10^{-3}	2.8×10^{-3}	1.2×10^{-3}	5.1×10^{-4}	17,740
				4	0.011	1.9×10^{-3}	1.1×10^{-4}	0	0	17,740
		raw GT	RUTH-LRT	2	0.299	0.176	0.098	0.055	0.03	17,740
				4	0.200	0.095	0.044	0.021	9.7×10^{-3}	17,740
			RUTH-Score	2	0.276	0.155	0.083	0.044	0.023	17,740
				4	0.183	0.083	0.036	0.015	7.4×10^{-3}	17,740
TOPMed	LQ	GL	RUTH-LRT	2	0.646	0.610	0.584	0.563	0.547	329,699
				4	0.652	0.614	0.588	0.567	0.55	329,699
			RUTH-Score	2	0.634	0.607	0.589	0.574	0.562	329,699
				4	0.635	0.608	0.590	0.575	0.562	329,699
		GT	RUTH-LRT	2	0.603	0.573	0.551	0.533	0.518	329,699
				4	0.610	0.580	0.556	0.538	0.552	329,699
			RUTH-Score	2	0.608	0.586	0.571	0.558	0.548	329,699
				4	0.608	0.587	0.572	0.559	0.549	329,699
	HQ	GL	RUTH-LRT	2	0.130	0.067	0.039	0.024	0.016	17,524
				4	0.041	0.018	8.7×10^{-3}	4.2×10^{-3}	3.1×10^{-3}	17,524
			RUTH-Score	2	0.130	0.065	0.036	0.021	0.014	17,524
				4	0.034	0.011	4.9×10^{-3}	3.1×10^{-3}	2.5×10^{-3}	17,524
		GT	RUTH-LRT	2	0.079	0.028	0.012	7.6×10^{-3}	5.9×10^{-3}	17,524
				4	0.125	0.036	0.012	5.0×10^{-3}	2.7×10^{-3}	17,524
			RUTH-Score	2	0.093	0.033	0.015	8.8×10^{-3}	6.0×10^{-3}	17,524
				4	0.145	0.047	0.017	7.1×10^{-3}	3.5×10^{-3}	17,524

In both 1000G and TOPMed, the false positive rate was much higher when k-means-based groupings were used for meta-analysis, compared to when high quality ancestry groupings were used. Similarly, the false positive rate was much higher when only 2 PCs were used, compared to when 4 PCs were used. Surprisingly, in TOPMed, using 4 PCs led to both a lower false positive rate and higher true positive rate when compared to using 2 PCs.

Supplementary Table 3.3 Performance of the unadjusted test, meta-analysis, and RUTH on the subset of TOPMed freeze 5 chromosome 20 variants that are also found in 1000G

Variant set	Genotype Format	HWE Test	Proportion of Significant Variants					Total Variant Count
			$P < 10^{-2}$	$P < 10^{-3}$	$P < 10^{-4}$	$P < 10^{-5}$	$P < 10^{-6}$	
HQ Variants	raw GT	Unadjusted	0.890	0.842	0.800	0.766	0.736	16,924
	raw GT	Meta-analysis	0.062	0.020	$8.0x10^{-3}$	$3.8x10^{-3}$	$2.3x10^{-3}$	16,924
	raw GT	RUTH-Score	0.145	0.046	0.016	$6.3x10^{-3}$	$2.8x10^{-3}$	16,924
	GL	RUTH-Score	0.032	$9.3x10^{-3}$	$3.7x10^{-3}$	$2.0x10^{-3}$	$1.5x10^{-3}$	16,924
	raw GT	RUTH-LRT	0.125	0.035	0.011	$4.2x10^{-3}$	$1.9x10^{-3}$	16,924
	GL	RUTH-LRT	0.039	0.016	$7.4x10^{-3}$	$3.1x10^{-3}$	$2.2x10^{-3}$	16,924
LQ Variants	raw GT	Unadjusted	0.762	0.728	0.702	0.683	0.667	10,513
	raw GT	Meta-analysis	0.649	0.616	0.592	0.575	0.560	10,513
	raw GT	RUTH-Score	0.727	0.693	0.673	0.656	0.640	10,513
	GL	RUTH-Score	0.698	0.669	0.648	0.631	0.618	10,513
	raw GT	RUTH-LRT	0.719	0.686	0.663	0.643	0.627	10,513
	GL	RUTH-LRT	0.693	0.662	0.639	0.621	0.605	10,513

For HQ variants, GL-based HWE tests had much better control of false positives than GT-based tests. Conversely, for LQ variants, GT-based HWE tests had a slightly better true positive rate than GL-based tests. Overall, GL-based tests had the best performance when considering the tradeoff between false positives and true positives (Supplementary Figure 3.5-3.6).

Supplementary Table 3.4 Simulation results for RUTH tests using 2 vs 4 principal components

This table can be found at the following link:
https://docs.google.com/spreadsheets/d/1Ac9rveZax5Y8NlKQ47wBaJNELqeJkFuNUpa1sNgnsno/edit?usp=sharing
We tested the effect of using different numbers of PCs in RUTH on Type I Error (θ = 0) and power (θ ≠ 0) for simulated samples with different numbers of ancestries, fixation indices, sequencing depths, and genotype representations. We simulated 50,000 variants for each combination of simulation parameters.

Supplementary Table 3.5 The effect of high vs. low quality subpopulation classification on meta-analysis in simulated samples

Grouping	Depth	Theta	Proportion of significant variants				
			$P < 10^{-6}$	$P < 10^{-5}$	$P < 10^{-4}$	$P < 10^{-3}$	$P < 0.01$
True ancestry labels	5	-0.05	0.0073	0.0125	0.0235	0.05	0.1145
		0	0.0147	0.0388	0.0919	0.1955	0.3519
	30	-0.05	0.0139	0.04	0.1048	0.2389	0.4594
		0	0	0	0.0001	0.0016	0.0127
k-means (3 groups)	5	-0.05	0.1201	0.149	0.19	0.2509	0.3513
		0	0.2907	0.3496	0.4195	0.4977	0.5826
	30	-0.05	0.0919	0.1122	0.1447	0.2017	0.3097
		0	0.2183	0.2553	0.3054	0.3734	0.4747

We simulated 50,000 variants in 5,000 samples arising from 5 distinct subpopulations (1,000 samples each), at low (5x) and high (30x) depth, with no deviation from HWE ($\theta = 0$) and moderate excess heterozygosity ($\theta = -0.05$). We used one of two different groupings for our samples: for high-quality labels, we used the original true ancestry labels from which we simulated our data; for low-quality labels, we ran k-means classification on the first 2 principal components of genetic variation for all our samples to generate 3 groups. We meta-analyzed all data sets using Stouffer's method. Type I error rates for low-depth samples were greatly inflated. For high-depth samples, when we used the true ancestry labels, Type I errors were well-controlled, with reasonable power to discover deviations from HWE, while when we used the crude k-means labels, Type I errors were greatly inflated, with surprisingly less power to discover deviations from HWE at less stringent P-value thresholds. These results highlight the importance of high-quality subpopulation classification for meta-analysis.

Supplementary Table 3.6 Comparison of runtimes and memory requirements for RUTH and PCAngsd in simulated and 1000G data

Data set	Genotype Format	Software	Test	N	Total Variant Count	Runtime (s)	Memory requirement (MB)
Simulated	GT	PLINK	Unadjusted	5,000	50,000	22	10
	GT	RUTH	RUTH LRT	5,000	50,000	348	15
	GL	RUTH	RUTH LRT	5,000	50,000	341	15
	GT	RUTH	RUTH Score	5,000	50,000	460	15
	GL	RUTH	RUTH Score	5,000	50,000	469	15
Simulated (5x)	GL	PCAngsd	PCAngsd	5,000	50,000	6,068	6,946
Simulated (30x)	GL	PCAngsd	PCAngsd	5,000	50,000	5,337	6,872
1000G	GT	PLINK	Unadjusted	2,504	28,706	2	8
	GL	RUTH	RUTH LRT	2,504	28,706	147	14
	GT	RUTH	RUTH LRT	2,504	28,706	96	13
	GL	RUTH	RUTH Score	2,504	28,706	216	14
	GT	RUTH	RUTH Score	2,504	28,706	177	13
	GL	PCAngsd	PCAngsd	2,504	28,660	4,105	2,073
TOPMed	GT	RUTH	RUTH LRT	53,831	347,223	158,731	57
	GL	RUTH	RUTH LRT	53,831	347,223	196,169	57

Simulation runtimes for PLINK and RUTH are averaged over 360 runs, across combinations of different simulation parameters. Simulation results for PCAngsd are averaged over 66 runs each for 5x and 30x coverage data. The higher uncertainty in low depth simulated data appears to have led to slower convergence in PCAngsd. All results for 1000G were from single runs. The listed TOPMed runtimes and memory requirements are for single-threaded analyses of all variants.

Supplementary Table 3.7 Sample contributions from each of the participating TOPMed studies

TOPMed Study Name	TOPMed Accession	Sample Size
Genetics of Cardiometabolic Health in the Amish	phs000956	1,025
Trans-Omics for Precision Medicine Whole Genome Sequencing Project: ARIC	phs001211	3,585
The Genetics and Epidemiology of Asthma in Barbados	phs001143	944
Cleveland Clinic Atrial Fibrillation Study	phs001189	328
The Cleveland Family Study (WGS)	phs000954	919
Cardiovascular Health Study	phs001368	69
Genetic Epidemiology of COPD (COPDGene) in theTOPMed Program	phs000951	8,733
The Genetic Epidemiology of Asthma in Costa Rica	phs000988	1,040
Diabetes Heart Study African American Coronary Artery Calcification (AA CAC)	phs001412	322
Whole Genome Sequencing and Related Phenotypes in the Framingham Heart Study	phs000974	3,725
Genes-environments and Admixture in Latino Asthmatics (GALA II) Study	phs000920	912
GeneSTAR (Genetic Study of Atherosclerosis Risk)	phs001218	1,633
Genetic Epidemiology Network of Arteriopathy (GENOA)	phs001345	1,069
Genetic Epidemiology Network of Salt Sensitivity (GenSalt)	phs001217	1,680
Genetics of Lipid Lowering Drugs and Diet Network (GOLDN)	phs001359	892
Heart and Vascular Health Study (HVH)	phs000993	64
HyperGEN - Genetics of Left Ventricular (LV) Hypertrophy	phs001293	1,752
Jackson Heart Study	phs000964	3,074
Whole Genome Sequencing of Venous Thromboembolism (WGS of VTE)	phs001402	1,250
MESA and MESA Family AA-CAC	phs001416	4,804
MGH Atrial Fibrillation Study	phs001062	916
Partners HealthCare Biobank	phs001024	109
San Antonio Family Heart Study (WGS)	phs001215	1,478
Study of African Americans, Asthma, Genes and Environment (SAGE) Study	phs000921	450
African American Sarcoidosis Genetics Resource	phs001207	606
Genome-wide Association Study of Adiposity in Samoans	phs000972	1,198
The Vanderbilt AF Ablation Registry	phs000997	154
The Vanderbilt Atrial Fibrillation Registry	phs001032	1016
Novel Risk Factors for the Development of Atrial Fibrillation in Women	phs001040	97
Women's Health Initiative (WHI)	phs001237	9,984
Total		**53,831**

Supplementary Table 3.8 TOPMed acknowledgements for omics support

TOPMed Accession #	TOPMed Project	Parent Study	TOPMed Phase	Omics Center	Omics Support
phs000956	Amish	Amish	1	Broad Genomics	3R01HL121007-01S1
phs001211	AFGen	ARIC AFGen	1	Broad Genomics	3R01HL092577-06S1
phs001211	VTE	ARIC	2	Baylor	3U54HG003273-12S2 / HHSN268201500015C
phs001143	BAGS	BAGS	1	Illumina	3R01HL104608-04S1
phs001189	AFGen	CCAF	1	Broad Genomics	3R01HL092577-06S1
phs000954	CFS	CFS	1	NWGC	3R01HL098433-05S1
phs000954	CFS	CFS	3.5	NWGC	HHSN268201600032I
phs001368	CHS	CHS	3	Baylor	HHSN268201600033I
phs001368	VTE	CHS VTE	2	Baylor	3U54HG003273-12S2 / HHSN268201500015C
phs000951	COPD	COPDGene	1	NWGC	3R01HL089856-08S1
phs000951	COPD	COPDGene	2	Broad Genomics	HHSN268201500014C
phs000951	COPD	COPDGene	2.5	Broad Genomics	HHSN268201500014C
phs000988	CRA_CAMP	CRA	1	NWGC	3R37HL066289-13S1
phs000988	CRA_CAMP	CRA	3	NWGC	HHSN268201600032I
phs001412	AA_CAC	DHS	2	Broad Genomics	HHSN268201500014C
phs000974	AFGen	FHS AFGen	1	Broad Genomics	3R01HL092577-06S1
phs000974	FHS	FHS	1	Broad Genomics	3U54HG003067-12S2
phs000920	ATGC	GALAII ATGC	3	NWGC	HHSN268201600032I
phs000920	PGX_Asthma	GALAII	1	NYGC	3R01HL117004-02S3
phs001218	AA_CAC	GeneSTAR AA_CAC	2	Broad Genomics	HHSN268201500014C
phs001218	GeneSTAR	GeneSTAR	legacy	Illumina	R01HL112064
phs001218	GeneSTAR	GeneSTAR	2	Psomagen	3R01HL112064-04S1
phs001345	HyperGEN_GENOA	GENOA	2	NWGC	3R01HL055673-18S1
phs001345	AA_CAC	GENOA AA_CAC	2	Broad Genomics	HHSN268201500014C
phs001217	GenSalt	GenSalt	2	Baylor	HHSN268201500015C
phs001359	GOLDN	GOLDN	2	NWGC	3R01HL104135-04S1
phs000993	AFGen	HVH	1	Broad Genomics	3R01HL092577-06S1
phs000993	VTE	HVH VTE	2	Baylor	3U54HG003273-12S2 / HHSN268201500015C
phs001293	HyperGEN_GENOA	HyperGEN	2	NWGC	3R01HL055673-18S1
phs000964	JHS	JHS	1	NWGC	HHSN268201100037C
phs001402	VTE	Mayo_VTE	2	Baylor	3U54HG003273-12S2 / HHSN268201500015C
phs001416	AA_CAC	MESA AA_CAC	2	Broad Genomics	HHSN268201500014C
phs001416	MESA	MESA	2	Broad Genomics	3U54HG003067-13S1
phs001062	AFGen	MGH_AF	1.4; 1.5; 2.4	Broad Genomics	3U54HG003067-12S2 / 3U54HG003067-13S1; 3U54HG003067-12S2 / 3U54HG003067-13S1; 3UM1HG008895-01S2
phs001062	AFGen	MGH_AF	1	Broad Genomics	3R01HL092577-06S1
phs001024	AFGen	Partners	1	Broad Genomics	3R01HL092577-06S1
phs001215	SAFS	SAFS	1	Illumina	3R01HL113323-03S1
phs001215	SAFS	SAFS	legacy	Illumina	R01HL113322
phs000921	ATGC	SAGE ATGC	3	NWGC	HHSN268201600032I
phs000921	PGX_Asthma	SAGE	1	NYGC	3R01HL117004-02S3
phs000972	Samoan	Samoan	1	NWGC	HHSN268201100037C
phs000972	Samoan	Samoan	2	NYGC	HHSN268201500016C
phs001207	Sarcoidosis	Sarcoidosis	2	Baylor	3R01HL113326-04S1
phs001207	Sarcoidosis	Sarcoidosis	3.5	NWGC	HHSN268201600032I

phs000997	AFGen	VAFAR	1.5; 2.4; 5.3	Broad Genomics	3U54HG003067-12S2 / 3U54HG003067-13S1; 3UM1HG008895-01S2; 3UM1HG008895-01S2
phs000997	AFGen	VAFAR	1	Broad Genomics	3R01HL092577-06S1
phs001032	AFGen	VU_AF	1	Broad Genomics	3R01HL092577-06S1
phs001040	AFGen	WGHS	1	Broad Genomics	3R01HL092577-06S1
phs001237	WHI	WHI	2	Broad Genomics	HHSN268201500014C

Supplementary Text 3.1

TOPMed Study Acknowledgements

NHLBI TOPMed: Genetics of Cardiometabolic Health in the Amish

The Amish studies upon which these data are based were supported by NIH grants R01 AG18728, U01 HL072515, R01 HL088119, R01 HL121007, and P30 DK072488. See publication: PMID: 18440328

NHLBI TOPMed: Trans-Omics for Precision Medicine Whole Genome Sequencing Project: ARIC

Genome Sequencing for “NHLBI TOPMed: Atherosclerosis Risk in Communities (ARIC)” (phs001211) was performed at the Baylor College of Medicine Human Genome Sequencing Center (HHSN268201500015C and 3U54HG003273-12S2) and the Broad Institute of MIT and Harvard (3R01HL092577-06S1).

The Atherosclerosis Risk in Communities study has been funded in whole or in part with Federal funds from the National Heart, Lung, and Blood Institute, National Institutes of Health, Department of Health and Human Services (contract numbers HHSN268201700001I, HHSN268201700002I, HHSN268201700003I, HHSN268201700004I and HHSN268201700005I). The authors thank the staff and participants of the ARIC study for their important contributions.

NHLBI TOPMed: The Genetics and Epidemiology of Asthma in Barbados

The Genetics and Epidemiology of Asthma in Barbados is supported by National Institutes of Health (NIH) National Heart, Lung, Blood Institute TOPMed (R01 HL104608-S1) and: R01 AI20059, K23 HL076322, and RC2 HL101651. For the specific cohort descriptions and descriptions regarding the collection of phenotype data can be found at: https://www.nhlbiwgs.org/group/bags-asthma. The authors wish to give special recognition to the individual study participants who provided biological samples and or data, without their support in research none of this would be possible.

NHLBI TOPMed: Cleveland Clinic Atrial Fibrillation Study

The research reported in this article was supported by grants from the National Institutes of Health (NIH) National Heart, Lung, and Blood Institute grants R01 HL090620 and R01 HL111314, the NIH National Center for Research Resources for Case Western Reserve University and the Cleveland Clinic Clinical and Translational Science Award (CTSA) UL1-RR024989, the Department of Cardiovascular Medicine philanthropic research fund, Heart and Vascular Institute, Cleveland Clinic, the Fondation Leducq grant 07-CVD 03, and The Atrial Fibrillation Innovation Center, State of Ohio.

NHLBI TOPMed: The Cleveland Family Study (WGS)

Support for the Cleveland Family Study was provided by NHLBI grant numbers R01 HL46380, R01 HL113338 and R35 HL135818.

NHLBI TOPMed: Cardiovascular Health Study

This research was supported by contracts HHSN268201200036C, HHSN268200800007C, HHSN268201800001C, N01-HC85079, N01-HC-85080, N01-HC-85081, N01-HC-85082, N01-HC-85083, N01-HC-85084, N01-HC-85085, N01-HC-85086, N01-HC-35129, N01-HC-15103, N01-HC-55222, N01-HC-75150, N01-HC-45133, and N01-HC-85239; grant numbers U01 HL080295, U01 HL130114 and R01 HL059367 from the National Heart, Lung, and Blood Institute, and R01 AG023629 from the National Institute on Aging, with additional contributions from the National Institute of Neurological Disorders and Stroke. A full list of principal CHS investigators and institutions can be found at https://chs-nhlbi.org/pi. Its content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

NHLBI TOPMed: Genetic Epidemiology of COPD (COPDGene) in the TOPMed Program

This research used data generated by the COPDGene study, which was supported by NIH Award Number U01 HL089897 and Award Number U01 HL089856 from the National Heart, Lung, and Blood Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Heart, Lung, and Blood Institute or the National Institutes of Health.

The COPDGene project is also supported by the COPD Foundation through contributions made to an Industry Advisory Board comprised of AstraZeneca, Boehringer Ingelheim, GlaxoSmithKline, Novartis, Pfizer, Siemens and Sunovion.

NHLBI TOPMed: The Genetic Epidemiology of Asthma in Costa Rica

This study was supported by NHLBI grants R37 HL066289 and P01 HL132825. We wish to acknowledge the investigators at the Channing Division of Network Medicine at Brigham and Women’s Hospital, the investigators at the Hospital Nacional de Niños in San José, Costa Rica and the study subjects and their extended family members who contributed samples and genotypes to the study, and the NIH/NHLBI for its support in making this project possible.

NHLBI TOPMed: Diabetes Heart Study African American Coronary Artery Calcification (AA CAC)

This work was supported by R01 HL92301, R01 HL67348, R01 NS058700, R01 AR48797, R01 DK071891, R01 AG058921, the General Clinical Research Center of the Wake Forest University School of Medicine (M01 RR07122, F32 HL085989), the American Diabetes Association, and a pilot grant from the Claude Pepper Older Americans Independence Center of Wake Forest University Health Sciences (P60 AG10484).

NHLBI TOPMed: Whole Genome Sequencing and Related Phenotypes in the Framingham Heart Study

The Framingham Heart Study (FHS) is a prospective cohort study of 3 generations of subjects who have been followed up to 65 years to evaluate risk factors for cardiovascular disease.13-16 Its large sample of ~15,000 men and women who have been extensively phenotyped with repeated examinations make it ideal for the study of genetic associations with cardiovascular disease risk factors and outcomes. DNA samples have been collected and immortalized since the mid-1990s and are available on ~8000 study participants in 1037 families. These samples have been used for collection of GWAS array data and exome chip data in nearly all with DNA samples, and for targeted sequencing, deep exome sequencing and light coverage whole genome sequencing in limited numbers. Additionally, mRNA and miRNA expression data, DNA methylation data, metabolomics and other ‘omics data are available on a sizable portion of study participants. This project will focus on deep whole genome sequencing (mean 30X coverage) in ~4100 subjects and imputed to all with GWAS array data to more fully understand the genetic contributions to cardiovascular, lung, blood and sleep disorders.

The FHS acknowledges the support of contracts NO1-HC-25195 and HHSN268201500001I from the National Heart, Lung, and Blood Institute and grant supplement R01 HL092577-06S1 for this research. We also acknowledge the dedication of the FHS study participants without whom this research would not be possible.

NHLBI TOPMed: Genes-environments and Admixture in Latino Asthmatics (GALA II) Study

The Genes-environments and Admixture in Latino Americans (GALA II) Study was supported by the National Heart, Lung, and Blood Institute of the National Institute of Health (NIH) grants R01HL117004 and X01HL134589; study enrollment supported by the Sandler Family Foundation, the American Asthma Foundation, the RWJF Amos Medical Faculty Development Program, Harry Wm. And Diana V. Hind Distinguished Professor in Pharmaceutical Sciences II and the National Institute of Environmental Health Sciences grant R01ES015794.

The GALA II study collaborators include Shannon Thyne, UCSF; Harold J. Farber, Texas Children’s Hospital; Denise Serebrisky, Jacobi Medical Center; Rajesh Kumar, Lurie Children’s Hospital of Chicago; Emerita Brigino-Buenaventura, Kaiser Permanente; Michael A. LeNoir, Bay Area Pediatrics; Kelley Meade, UCSF Benioff Children’s Hospital, Oakland; William Rodriguez-Cintron, VA Hospital, Puerto Rico; Pedro C. Avila, Northwestern University; Jose R. Rodriguez-Santana, Centro de Neumologia Pediatrica; Luisa N. Borrell, City University of New York; Adam Davis, UCSF Benioff Children’s Hospital, Oakland; Saunak Sen, University of Tennessee and Fred Lurmann, Sonoma Technologies, Inc.

The authors acknowledge the families and patients for their participation and thank the numerous health care providers and community clinics for their support and participation in GALA II. In particular, the authors thank study coordinator Sandra Salazar; the recruiters who obtained the data: Duanny Alva, MD, Gaby Ayala-Rodriguez, Lisa Caine, Elizabeth Castellanos, Jaime Colon, Denise DeJesus, Blanca Lopez, Brenda Lopez, MD, Louis Martos, Vivian Medina, Juana Olivo, Mario Peralta, Esther Pomares, MD, Jihan Quraishi, Johanna Rodriguez, Shahdad Saeedi, Dean Soto, Ana Taveras; and the lab researcher Celeste Eng who processed the biospecimens.

NHLBI TOPMed: Genetic Epidemiology Network of Arteriopathy (GENOA)

Support for GENOA was provided by the National Heart, Lung and Blood Institute (HL054457, HL054464, HL054481, HL119443, and HL087660) of the National Institutes of Health. WGS for “NHLBI TOPMed: Genetic Epidemiology Network of Arteriopathy” (phs001345) was performed at the Mayo Clinic Genotyping Core, the DNA Sequencing and Gene Analysis Center at the University of Washington (3R01HL055673-18S1), and the Broad Institute (HHSN268201500014C) for their genotyping and sequencing services. We would like to thank the GENOA participants.

NHLBI TOPMed: Genetic Epidemiology Network of Salt Sensitivity (GenSalt)

The Genetic Epidemiology Network of Salt-Sensitivity (GenSalt) was supported by research grants (U01HL072507, R01HL087263, and R01HL090682) from the National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD.

NHLBI TOPMed: Genetics of Lipid Lowering Drugs and Diet Network (GOLDN)

GOLDN biospecimens, baseline phenotype data, and intervention phenotype data were collected with funding from the National Heart, Lung and Blood Institute (NHLBI) grant U01 HL072524. Whole-genome sequencing in GOLDN was funded by NHLBI grant R01 HL104135 and supplement R01 HL104135-04S1.

NHLBI TOPMed: Heart and Vascular Health Study (HVH)

The research reported in this article was supported by grants HL068986, HL085251, HL095080, and HL073410 from the National Heart, Lung, and Blood Institute.

NHLBI TOPMed: Hypertension Genetic Epidemiology Network (HyperGEN)

The HyperGEN Study is part of the National Heart, Lung, and Blood Institute (NHLBI) Family Blood Pressure Program; collection of the data represented here was supported by grants U01 HL054472 (MN Lab), U01 HL054473 (DCC), U01 HL054495 (AL FC), and U01 HL054509 (NC FC). The HyperGEN: Genetics of Left Ventricular Hypertrophy Study was supported by NHLBI grant R01 HL055673 with whole-genome sequencing made possible by supplement -18S1.

NHLBI TOPMed: The Jackson Heart Study

The Jackson Heart Study (JHS) is supported and conducted in collaboration with Jackson State University (HHSN268201800013I), Tougaloo College (HHSN268201800014I), the Mississippi State Department of Health (HHSN268201800015I/HHSN26800001) and the University of Mississippi Medical Center (HHSN268201800010I, HHSN268201800011I and HHSN268201800012I) contracts from the National Heart, Lung, and Blood Institute (NHLBI) and the National Institute for Minority Health and Health Disparities (NIMHD). The authors also wish to thank the staffs and participants of the JHS.

NHLBI TOPMed: Multi-Ethnic Study of Atherosclerosis

MESA and the MESA SHARe projects are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts 75N92020D00001, HHSN268201500003I, N01-HC-95159, 75N92020D00005, N01-HC-95160, 75N92020D00002, N01-HC-95161, 75N92020D00003, N01-HC-95162, 75N92020D00006, N01-HC-95163, 75N92020D00004, N01-HC-95164, 75N92020D00007, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-000040, UL1-TR-001079, and UL1-TR-001420. Also supported by the National Center for Advancing Translational

Sciences, CTSI grant UL1TR001881, and the National Institute of Diabetes and Digestive and Kidney Disease Diabetes Research Center (DRC) grant DK063491 to the Southern California Diabetes Endocrinology Research Center.

NHLBI TOPMed: Whole Genome Sequencing of Venous Thromboembolism (WGS of VTE) Funded in part by grants from the National Institutes of Health, National Heart, Lung, and Blood Institute (HL66216 and HL83141) and the National Human Genome Research Institute (HG04735).

NHLBI TOPMed: MGH Atrial Fibrillation Study

This work was supported by the Fondation Leducq (14CVD01), and by grants from the National Institutes of Health to Dr. Ellinor (1RO1HL092577, R01HL128914, K24HL105780). This work was also supported by a grant from the American Heart Association to Dr. Ellinor (18SFRN34110082). Dr. Lubitz is supported by NIH grant 1R01HL139731 and AHA 18SFRN34250007.

NHLBI TOPMed: Partners HealthCare Biobank

We thank the Broad Institute for generating high-quality sequence data supported by the NHLBI grant 3R01HL092577-06S1 to Dr. Patrick Ellinor. The datasets used in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs001024.

NHLBI TOPMed: Study of African Americans, Asthma, Genes and Environment (SAGE)

The Study of African Americans, Asthma, Genes and Environments (SAGE) was supported by the National Heart, Lung, and Blood Institute of the National Institute of Health (NIH) grants R01HL117004 and X01HL134589; study enrollment supported by the Sandler Family Foundation, the American Asthma Foundation, the RWJF Amos Medical Faculty Development Program, Harry Wm. And Diana V. Hind Distinguished Professor in Pharmaceutical Sciences II.

The SAGE study collaborators include Harold J. Farber, Texas Children's Hospital; Emerita Brigino-Buenaventura, Kaiser Permanente; Michael A. LeNoir, Bay Area Pediatrics; Kelley Meade, UCSF Benioff Children's Hospital, Oakland; Luisa N. Borrell, City University of New York; Adam Davis, UCSF Benioff Children's Hospital, Oakland and Fred Lurmann, Sonoma Technologies, Inc.

The authors acknowledge the families and patients for their participation and thank the numerous health care providers and community clinics for their support and participation in SAGE. In particular, the authors thank study coordinator Sandra Salazar; the recruiters who obtained the data: Lisa Caine, Elizabeth Castellanos, Brenda Lopez, MD, Shahdad Saeedi; and the lab researcher Celeste Eng who processed the biospecimens.

E.G.B was supported by National Heart, Lung, and Blood Institute (NHLBI): U01HL138626, R01HL117004, R01HL128439, R01HL135156, X01HL134589, R01HL141992, R01HL141845; the National Human Genome Research Institute (NHGRI): U01HG009080; the National Institute of Environmental Health Sciences (NIEHS): R01ES015794, R21ES24844; the National Institute on Minority Health and Health Disparities (NIMHD): P60MD006902, R01MD010443, RL5GM118984,R56MD013312; the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD): R01HD085993; and the Tobacco-Related Disease Research Program (TRDRP): 24RT-0025 and 27IR-0030.

NHLBI TOPMed: San Antonio Family Heart Study (WGS)

Collection of the San Antonio Family Study data was supported in part by National Institutes of Health (NIH) grants R01 HL045522, MH078143, MH078111 and MH083824; and whole genome sequencing of SAFS subjects was supported by U01 DK085524 and R01 HL113323. We are very grateful to the participants of the San Antonio Family Study for their continued involvement in our research programs.

NHLBI TOPMed: The Samoan Obesity, Lifestyle and Genetic Adaptations Study (OlaGA) Group

Financial support for the Samoan Obesity, Lifestyle and Genetic Adaptations Study (OlaGA) Group comes from the U.S. National Institutes of Health Grant R01-HL093093 and R01-HL133040. We acknowledge the assistance of the Samoa Ministry of Health and the Samoa Bureau of Statistics for their guidance and support in the conduct of this study. We thank the local village officials for their help and the participants for their generosity. The following publication describes the origin of the dataset: Hawley NL, Minster RL, Weeks DE, Viali S, Reupena MS, Sun G, Cheng H, Deka R, McGarvey ST. Prevalence of Adiposity and Associated Cardiometabolic Risk Factors in the Samoan Genome-Wide Association Study. Am J Human Biol 2014. 26: 491-501. DOI: 10.1002/jhb.22553. PMID: 24799123.

Our study name: ‘The Samoan Obesity, Lifestyle and Genetic Adaptations Study (OlaGA) Group’.

Ranjan Deka, Department of Environmental and Public Health Sciences, College of Medicine, University of Cincinnati, Cincinnati, OH 45267-0056. Email: dekar@uc.edu.

Nicola L Hawley, Department of Epidemiology (Chronic Disease), School of Public Health, Yale University, New Haven, CT 06520-0834. Email: nicola.hawley@yale.edu.

Stephen T McGarvey, International Health Institute, Department of Epidemiology, School of Public Health, and Department of Anthropology, Brown University. 02912. Email: stephen_mcgarvey@brown.edu.

Ryan L Minster, Department of Human Genetics and Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA 15261. Email: rminster@pitt.edu.

Take Naseri, Ministry of Health, Government of Samoa, Apia, Samoa. Email: taken@health.gov.ws.

Muagututi‘a Sefuiva Reupena, Lutia I Puava Ae Mapu I Fagalele, Apia, Samoa. Email: smuagututia51@gmail.com.

Daniel E Weeks, Department of Human Genetics and Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA 15261. Email: weeks@pitt.edu.

NHLBI TOPMed: The Vanderbilt AF Ablation Registry

The research reported in this article was supported by grants from the American Heart Association to Dr. Shoemaker (11CRP742009), Dr. Darbar (EIA 0940116N), and grants from the National Institutes of Health (NIH) to Dr. Darbar (R01 HL092217), and Dr. Roden (U19 HL65962, and UL1 RR024975). The project was also supported by a CTSA award (UL1 TR00045) from the National Center for Advancing Translational Sciences. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Center for Advancing Translational Sciences or the NIH.

NHLBI TOPMed: The Vanderbilt Atrial Fibrillation Registry

The research reported in this article was supported by grants from the American Heart Association to Dr. Darbar (EIA 0940116N), and grants from the National Institutes of Health (NIH) to Dr. Darbar (HL092217), and Dr. Roden (U19 HL65962, and UL1 RR024975). This project was also supported by CTSA award (UL1TR000445) from the National Center for Advancing Translational Sciences. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Center for Advancing Translational Sciences of the NIH.

NHLBI TOPMed: Novel Risk Factors for the Development of Atrial Fibrillation in Women

The Women’s Genome Health Study (WGHS) is supported by HL 043851 and HL099355 from the National Heart, Lung, and Blood Institute and CA 047988 from the National Cancer Institute, the Donald W. Reynolds Foundation with collaborative scientific support and funding for genotyping provided by Amgen. AF endpoint confirmation was supported by HL-093613 and a grant from the Harris Family Foundation and Watkin’s Foundation.

NHLBI TOPMed: Women’s Health Initiative (WHI)

The WHI program is funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services through contracts HHSN268201600018C, HHSN268201600001C, HHSN268201600002C, HHSN268201600003C, and HHSN268201600004C.

NHLBI TOPMed: GeneSTAR (Genetic Study of Atherosclerosis Risk)

The Johns Hopkins Genetic Study of Atherosclerosis Risk (GeneSTAR) was supported by grants from the National Institutes of Health through the National Heart, Lung, and Blood Institute (U01HL72518, HL087698, HL112064) and by a grant from the National Center for Research Resources (M01-RR000052) to the Johns Hopkins General Clinical

Research Center. We would like to thank the participants and families of GeneSTAR and our dedicated staff for all their sacrifices.

NHLBI TOPMed: Genetics of Sarcoidosis in African Americans (Sarcoidosis)

National Institutes of Health (R01HL113326, P30 GM110766-01)

3.7 References

Balding, D. J., 2003 Likelihood-based inference for genetic correlation coefficients. Theor Popul Biol 63**:** 221-230.

Balding, D. J., and R. A. Nichols, 1995 A Method for Quantifying Differentiation between Populations at Multi-Allelic Loci and Its Implications for Investigating Identity and Paternity. Genetica 96**:** 3-12.

Bycroft, C., C. Freeman, D. Petkova, G. Band, L. T. Elliott *et al.*, 2018 The UK Biobank resource with deep phenotyping and genomic data. Nature 562**:** 203-209.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks *et al.*, 2011 The variant call format and VCFtools. Bioinformatics 27**:** 2156-2158.

Delaneau, O., J. F. Zagury and J. Marchini, 2013 Improved whole-chromosome phasing for disease and population genetic studies. Nat Methods 10**:** 5-6.

Dempster, A. P., N. M. Laird and D. B. Rubin, 1977 Maximum Likelihood from Incomplete Data Via the EM Algorithm. Journal of the Royal Statistical Society: Series B (Methodological) 39**:** 1-22.

Dryden, I. L., and K. V. Mardia, 1998 *Statistical shape analysis*. John Wiley & Sons, Chichester ; New York.

Ewing, B., and P. Green, 1998 Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8**:** 186-194.

Fritsche, L. G., W. Igl, J. N. Bailey, F. Grassmann, S. Sengupta *et al.*, 2016 A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. Nat Genet 48**:** 134-143.

Hao, W., M. Song and J. D. Storey, 2016 Probabilistic models of genetic variation in structured populations applied to global human studies. Bioinformatics 32**:** 713-721.

Hao, W., and J. D. Storey, 2019 Extending Tests of Hardy-Weinberg Equilibrium to Structured Populations. Genetics 213**:** 759-770.

Hardy, G. H., 1908 Mendelian Proportions in a Mixed Population. Science 28**:** 49-50.

Holsinger, K. E., 1999 Analysis of Genetic Diversity in Geographically Structured Populations: A Bayesian Perspective. Hereditas 130**:** 245-255.

Holsinger, K. E., P. O. Lewis and D. K. Dey, 2002 A Bayesian approach to inferring population structure from dominant markers. Mol Ecol 11**:** 1157-1164.

Jin, Y., A. A. Schaffer, M. Feolo, J. B. Holmes and B. L. Kattman, 2019 GRAF-pop: A Fast Distance-Based Method To Infer Subject Ancestry from Multiple Genotype Datasets Without Principal Components Analysis. G3 (Bethesda) 9**:** 2447-2461.

Jun, G., M. Flickinger, K. N. Hetrick, J. M. Romm, K. F. Doheny *et al.*, 2012 Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. Am J Hum Genet 91**:** 839-848.

Kuhn, R. M., D. Haussler and W. J. Kent, 2013 The UCSC genome browser and associated tools. Brief Bioinform 14**:** 144-161.

Laurie, C. C., K. F. Doheny, D. B. Mirel, E. W. Pugh, L. J. Bierut *et al.*, 2010 Quality control and quality assurance in genotypic data for genome-wide association studies. Genet Epidemiol 34**:** 591-602.

Li, J. Z., D. M. Absher, H. Tang, A. M. Southwick, A. M. Casto *et al.*, 2008 Worldwide human relationships inferred from genome-wide patterns of variation. Science 319**:** 1100-1104.

Li, M., and C. Li, 2008 Assessing departure from Hardy-Weinberg equilibrium in the presence of disease association. Genet Epidemiol 32**:** 589-599.

Locke, A. E., B. Kahali, S. I. Berndt, A. E. Justice, T. H. Pers *et al.*, 2015 Genetic studies of body mass index yield new insights for obesity biology. Nature 518**:** 197-206.

McCarroll, S. A., T. N. Hadnott, G. H. Perry, P. C. Sabeti, M. C. Zody *et al.*, 2006 Common deletion polymorphisms in the human genome. Nat Genet 38**:** 86-92.

Meisner, J., and A. Albrechtsen, 2019 Testing for Hardy-Weinberg Equilibrium in Structured Populations using Genotype or Low-Depth NGS Data. Mol Ecol Resour.

Mosteller, F., and R. A. Fisher, 1948 Questions and Answers. The American Statistician 2**:** 30-31.

Nielsen, D. M., M. G. Ehm and B. S. Weir, 1998 Detecting marker-disease association by testing for Hardy-Weinberg disequilibrium at a marker locus. Am J Hum Genet 63**:** 1531-1540.

Nielsen, R., J. S. Paul, A. Albrechtsen and Y. S. Song, 2011 Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet 12**:** 443-451.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick *et al.*, 2006 Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38**:** 904-909.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira *et al.*, 2007 PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81**:** 559-575.

Rohlfs, R. V., and B. S. Weir, 2008 Distributions of Hardy-Weinberg equilibrium test statistics. Genetics 180**:** 1609-1616.

Rosenberg, N. A., J. K. Pritchard, J. L. Weber, H. M. Cann, K. K. Kidd *et al.*, 2002 Genetic structure of human populations. Science 298**:** 2381-2385.

Sha, Q., and S. Zhang, 2011 A test of Hardy-Weinberg equilibrium in structured populations. Genet Epidemiol 35**:** 671-678.

Stouffer, S. A., 1949 *The American soldier*. Princeton University Press, Princeton,.

Stouffer, S. A., E. A. Suchman, L. C. DeVinney, S. A. Star and R. M. Williams Jr, 1949 The American soldier: Adjustment during army life.(Studies in social psychology in World War II), Vol. 1.

Taliun, D., D. N. Harris, M. D. Kessler, J. Carlson, Z. A. Szpiech *et al.*, 2019 Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. bioRxiv.

The 1000 Genomes Project Consortium, A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison *et al.*, 2015 A global reference for human genetic variation. Nature 526**:** 68-74.

The International HapMap Consortium, D. M. Altshuler, R. A. Gibbs, L. Peltonen, D. M. Altshuler *et al.*, 2010 Integrating common and rare genetic variation in diverse human populations. Nature 467**:** 52-58.

Van Oosterhout, C., W. F. Hutchinson, D. P. M. Wills and P. Shipley, 2004 MICRO-CHECKER: software for identifying and correcting genotyping errors in microsatellite data. Molecular Ecology Notes 4**:** 535-538.

Wang, C., Z. A. Szpiech, J. H. Degnan, M. Jakobsson, T. J. Pemberton *et al.*, 2010 Comparing spatial maps of human population-genetic variation using Procrustes analysis. Stat Appl Genet Mol Biol 9**:** Article 13.

Waples, R. S., 2015 Testing for Hardy-Weinberg proportions: have we lost the plot? J Hered 106**:** 1-19.

Weinberg, W., 1908 Uber den nachweis der vererbung beim menschen. Jh. Ver. vaterl. Naturk. Wurttemb. 64**:** 369-382.

Wigginton, J. E., D. J. Cutler and G. R. Abecasis, 2005 A note on exact tests of Hardy-Weinberg equilibrium. Am J Hum Genet 76**:** 887-893.

Yang, W. Y., J. Novembre, E. Eskin and E. Halperin, 2012 A model-based approach for analysis of spatial structure in genetic data. Nat Genet 44**:** 725-731.

Zhang, F., M. Flickinger, S. A. G. Taliun, P. P. G. C. In, G. R. Abecasis *et al.*, 2020 Ancestry-agnostic estimation of DNA sample contamination from sequence reads. Genome Res 30**:** 185-194.

Chapter 4

PheGET: An Interactive Multi-Tissue eQTL Browser

A paper covering most of the material in this chapter is in preparation, with myself as first author

4.1 Introduction

Expression quantitative trait loci (eQTLs) are an important piece of the puzzle for understanding the regulatory mechanisms underlying genetic associations (GALLAGHER AND CHEN-PLOTKIN 2018). The continuing advances in genomic technology have allowed researchers to generate enormous amounts of molecular profiles across many individuals and tissues. For example, the Genotype Tissue Expression (GTEx) Consortium analyzed transcriptomic profiles of 49 different tissues across 838 samples and identified >4 million eQTLs across >10 million genetic variants (AGUET *et al.* 2019). The sheer number of eQTLs produced in datasets such as this require scalable, custom-designed visualization tools as aids for interpretation and analysis. Such eQTL resources allow the exploration of a wide range of clinically relevant hypotheses, such as interpreting potential regulatory mechanisms in individual GWAS signals (ROSELLI *et al.* 2018; YENGO *et al.* 2018), understanding tissue-specific epigenetic architecture of complex traits (EHRLICH *et al.* 2019), and pinpointing likely causal variants by colocalizing GWAS and eQTL signals (LIU *et al.* 2018; WU *et al.* 2019).

Functional interpretation of disease-associated regulatory variants can be facilitated by the interactive exploration of large transcriptomic profiles. However, existing web tools provide only a limited range of information that requires additional effort to address scientific questions elucidating the tissue-specific relationship between the variants and genes. For example, the GTEx Portal (https://gtexportal.org) allows us to visualize expression levels or list significant cis-eQTLs across tissues to understand the regulatory landscape of the gene. However, the GTEx Portal does not readily address many questions relevant to functional interpretation of regulatory variants. For example, the marginal p-values and effect sizes of eQTLs do not directly inform whether a trait-associated genetic variant also likely regulate the expression of a gene in a particular tissue. Other online resources, such as PheWeb (GAGLIANO TALIUN *et al.* 2020), BRAVO (NHLBI 2018), gnomAD (KARCZEWSKI *et al.* 2020), and the UCSC browser (HAEUSSLER *et al.* 2019) can also provide clues for the functional interpretation, so it is crucial to connect these resources in the context of tissue-specific regulation of disease-associated variants.

To facilitate functional interpretation of regulatory variants from population-scale transcriptomic resources like GTEx, we developed PheGET, an eQTL-focused web application that leverages the widely used tools LocusZoom (PRUIM *et al.* 2010), PheWeb (https://github.com/statgen/pheweb/), and LD server (https://github.com/statgen/LDServer). PheGET visualizes the genomic landscape of cis-eQTLs across multiple tissues, focusing on a variant, a gene, or a genomic locus. PheGET is designed to aid the interpretation of the regulatory function of genetic variants by providing answers to functionally relevant questions, for example, (1) how likely is a specific genetic variant causal for a cis-eQTL; (2) is a cis-eQTL is tissue-specific or shared across tissues; (3) what is the linkage disequilibrium (LD) structure around the variant or gene; (4) which nearby genes are likely co-regulated by the variant and in which tissues; (5) are there additional information from other resources, such as biobank-based PheWAS (phenome-wide association) results, variant databases, or regulatory genomic resources, that corroborate the functional interpretation. PheGET provides interactive visualizations of cis-eQTLs with relevant context and connects to relevant external resources. PheGET complements existing transcriptomic resources such as the GTEx Portal and serves as a centralized tool for interpreting regulatory variants.

4.2 Key Features

PheGET allows investigators to query an eQTL database for a variant, gene, or locus to interactively visualize multi-tissue cis-eQTLs from various viewpoints, either in a comprehensive single-variant view showing all its cis-eQTLs, or in a LocusZoom-based region view specific to a gene and a tissue. When a variant is queried for, PheGET visualizes the cis-regulation landscape of all nearby genes across all available tissues; PheGET also identifies the likely regulatory effects of the variant. When a locus is queried for, PheGET offers multi-tissue and/or multi-gene LocusZoom visualization of strongly associated cis-eQTLs. These visualizations support exploratory analysis to help investigators interpret the regulatory functions of the associated variants.

To illustrate the capabilities of PheGET, we will use cardiovascular disease as a motivating example. Low-density lipoprotein cholesterol (LDL) level in blood has been strongly associated with cardiovascular disease, with hundreds of significantly associated loci (KLARIN *et al.* 2018). For example, in a UK Biobank (UKB) genome-wide association study (GWAS) of total cholesterol in LDL, the strongest trait-marker associations are found near *APOE, PCSK9, LDLR,* and *APOB* (Figure S1A-B). The mechanism behind the *APOB* association is currently unclear (NIU *et al.* 2017) We use PheGET to explore this relationship using the GTEx (v8) data, and provide two examples with step-by-step instructions (Figure S2-S4).

PheGET's variant-centric visualization helps users understand the potentially causative role of a variant in tissue-specific regulation beyond the marginal summary statistics. For example, if a user searches for rs934197, the top (normalized) LDL-associated variant in UKB (NRM_LDL_C) near *APOB*, PheGET provides a single variant view similar to PheWAS, in which the variant is plotted against multiple traits. In this view, the traits are grouped by proximal (<1Mb) genes and tissues (Figure 1A, S5-6). It is clear that rs934197 is strongly associated with *APOB* expression levels in several tissues, including subcutaneous adipose (p = 4.0 x 10^{-17}), tibial artery (p = 7.9 x 10^{-11}), and esophagus gastroesophageal junction (p = 6.9 x 10^{-9}). However, it is unclear whether the variant is causal to these eQTL signals or a shadow of other causal variants via LD. Using posterior inclusion probabilities (PIPs) under the DAP-G model (Wen et al. 2017), PheGET shows that the strongest eQTL in subcutaneous adipose is not included in the credible set of putative causal variants (Figure S7). Instead, it can be best explained as a shadow signal of another cis-eQTL, most likely rs4665178 (p = 1.4x10^{-28}, PIP=0.35), located 54kb upstream. Meanwhile, rs934197 has the strongest PIPs in esophagus gastroesophageal junction (0.90), esophagus muscularis (0.71), tibial artery (0.83), and sigmoid colon (0.73) tissues even though marginal p-values were weaker than subcutaneous adipose. These results provide a more complete picture of tissue-specific cis-regulation potentially caused by a specific variant (Figure S8-10).

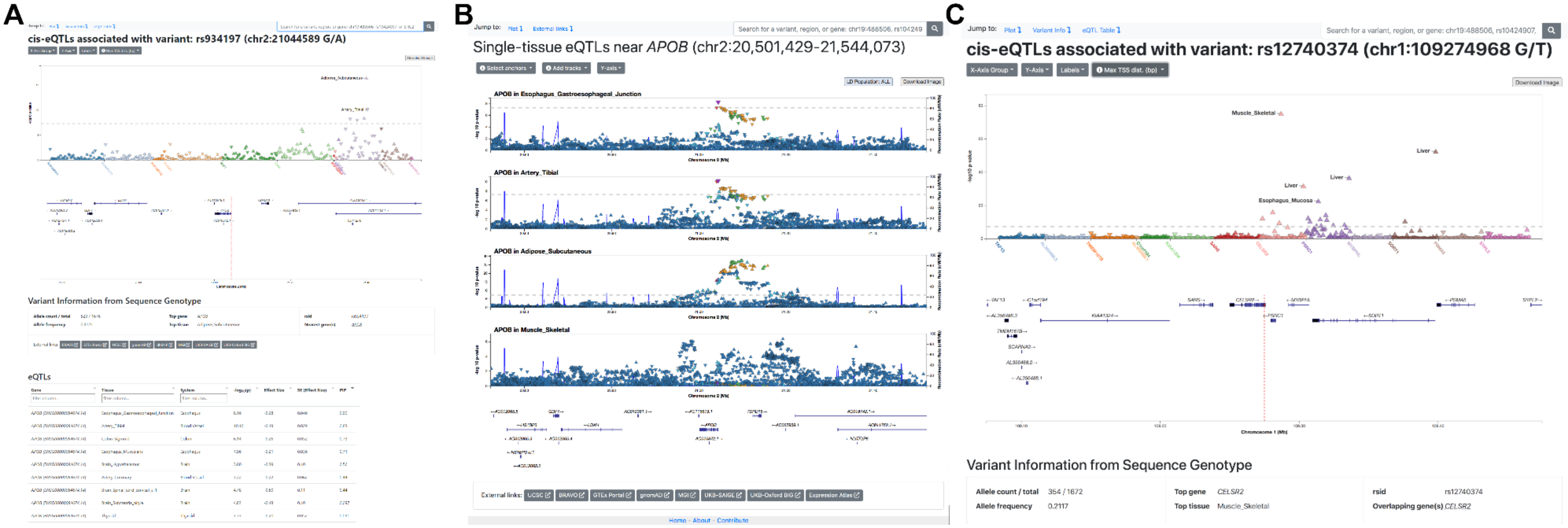
PheGET can display multiple parallel LocusZoom plots to visualize results across tissues and/or expression across multiple nearby genes to visualize the complex structure of gene regulation entangled with LD. For example, searching for *APOB* in PheGET displays a LocusZoom view for *APOB* expression in the tissues with the strongest PIPs for variants with $p < 10^{-6}$ (Figure S11). PheGET also provides users with a sortable table of all cis-eQTLs associated with the gene with PIP > 10^{-5}. Users can add other tissues of interest, or proximal genes on demand (Figure 1B, S12). When we set rs934197 as the index variant in our LocusZoom view of *APOB*, we can easily see that this variant is the top signal in four different tissues in the gastrointestinal or lower circulatory systems, while other tissues—heart (rs661665), adipose (rs4665178), skin (rs579826), and muscle (rs56327713)—feature different nearby top variants, with only rs934197 co-localized as a peak association signal with any phenotype in UKB PheWAS. Together, these observations suggest that the lipid association signal near *APOB* may be explained by gene regulation in specific gastrointestinal and/or circulatory systems.

PheGET also allows users to identify and visualize nearby genes sharing a cis-eQTL. For example, one of the peak signals associated with self-reported high cholesterol in the UK Biobank (biobank_20002_1473) is rs12740374 near the *SORT1-PSRC1-CELSR2* locus (Figure S3A). Querying this variant in PheGET demonstrates that the variant is also a strong cis-eQTL, specifically in liver (Figure S14-15). Interestingly, all three genes (*SORT1, PSRC1, CELSR2*) regulated by this variant have high PIPs (Figure 1C), with the association appearing to be highly liver-specific for *SORT1* and *PSRC1* and shared across tissues for *CELSR2* (Figure S3B-D). This is presumably due to tissue-specific regulatory elements shared between these genes (SCHADT *et al.* 2008; MUSUNURU *et al.* 2010; WANG *et al.* 2018) and it provides an important insight to understand the functional mechanism underlying the association between the variants and lipid traits. As shown in our examples, PheGET provides investigators intuitive and interactive representations of expression data and posterior probabilities across genes and tissues, largely complementary to the GTEx Portal and other online resources to help with functional interpretation of eQTLs and trait-associated variants.

4.3 Discussion

Understanding the function of trait-associated non-coding variants is becoming increasingly important as more genomes, transcriptomes, and epigenomes are sequenced. Gene regulation is believed to be involved in a large fraction of such associations, but there are limited resources for trait experts to generate hypotheses to explain regulatory mechanisms underlying the association signals. PheGET offers new interactive ways to visualize and summarize eQTLs in a tissue-specific manner by combining key features from LocusZoom and PheWeb, focusing on putative causal eQTLs through PIPs (Figure S4). We plan to implement additional useful features in the near future, including the abilities to import new data generated by individual investigators, to perform conditional analysis on the fly, to integrate with biobank-driven PheWeb resources more seamlessly, to visualize isoform-aware eQTLs, and to include chromatin-accessibility QTLs and other epigenetic resources. We expect PheGET will aid with translating GWAS associations into underlying regulatory mechanisms by enabling the exploration of plausible hypotheses through our intuitive and practical user interface. As more online resources like PheGET become available to address tailored scientific questions on functional variants, precise and integrative translation of genomic findings will be more accessible to broader scientific community.

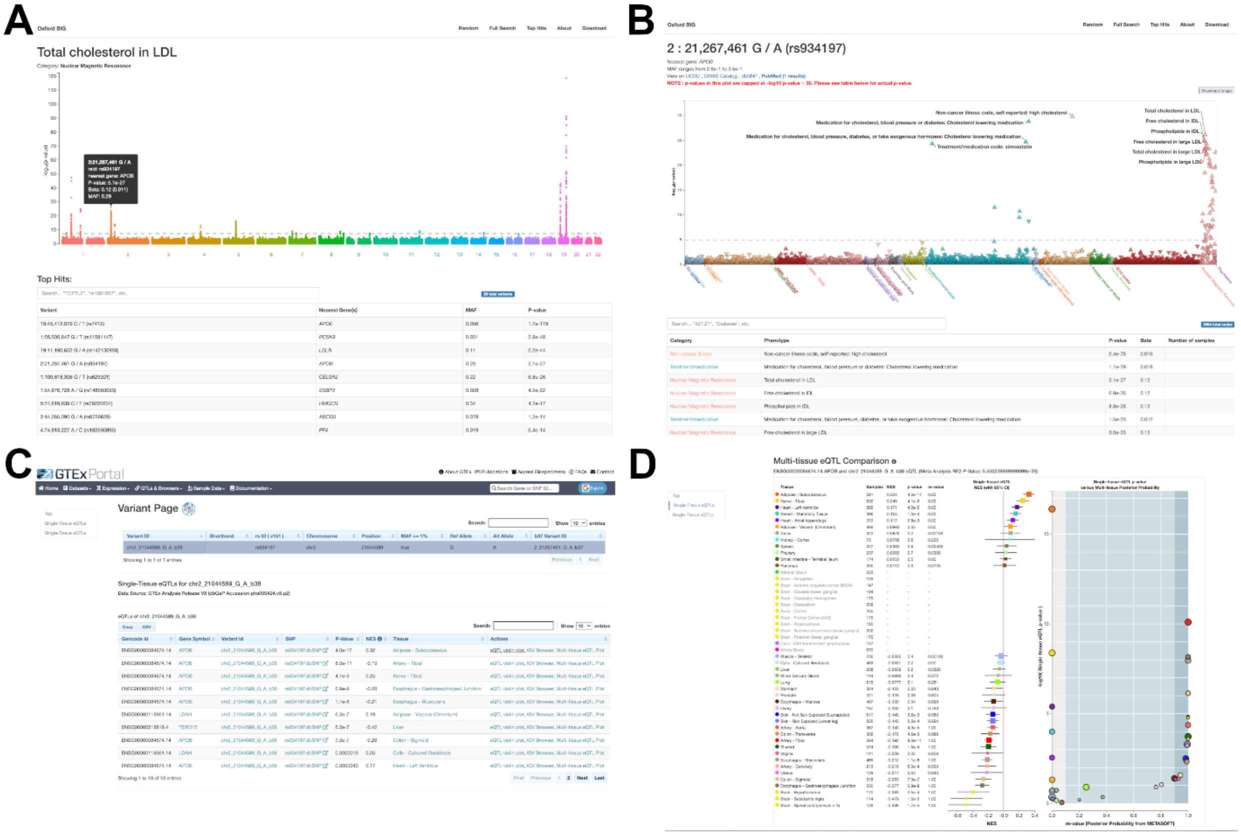
Figure 4.1 Examples of PheGET views



(A) A variant-centric view of PheGET as an outcome of searching for rs934197, the most strongly associated variant with LDL cholesterol near APOB in UK Biobank. In the top panel, the x-axis is ordered by genes overlapping with 1Mb window by genomic coordinates, each representing an individual tissue. The y-axis represents p-values in log-scale, which can be toggled between effect sizes and PIPs. The gene panel annotate the genomic location of genes and exon, with the position of the queried variant marked with a red dotted line. In the middle panel, the basic information of the variant is shown with external links to easily navigate to online resources relevant to the variant. The table in the bottom summarizes the strongest cis-eQTLs. (B) A locus-centric view of PheGET when querying *APOB* and adding relevant tissues using the dropdown menu at the top. The cis-eQTLs for different tissues near *APOB* are shown using LocusZoom. The first two tissues share the same peak cis-eQTLs, while the other two tissues do not. (C) Another variant-centric view of PheGET when querying rs12740374, the most strongly associated variant with self-reported high cholesterol level in UK Biobank near the *SORT1* locus. The results show that the variant is a strong cis-eQTL regulating multiple genes (*SORT1, PSRC1, CELSR2*), particularly in liver tissue, illustrating the benefit of PheGET to identify co-regulation of proximal genes.

4.4 Appendix: Supplementary Figures and Table

Supplementary Figure 4.1 Exploring LDL-associated eQTLs in the *APOB* locus


A
B
C
D
Oxford BIG
Random
Full Search
Top Hits
About
Download
Total cholesterol in LDL
Category: Nuclear Magnetic Resonance
2:21,267,461 G / A
rsid: rs934197
nearest gene: APOB
P-value: 5.1e-27
Beta: 0.12 (0.011)
MAF: 0.29
2 : 21,267,461 G / A (rs934197)
Total cholesterol in LDL
Free cholesterol in IDL
Phospholipids in IDL
Free cholesterol in large LDL
Total cholesterol in large LDL
Phospholipids in large LDL
Non-cancer illness code, self-reported: high cholesterol
Medication for cholesterol, blood pressure or diabetes: Cholesterol lowering medication
Medication for cholesterol, blood pressure, diabetes, or take exogenous hormones: Cholesterol lowering medication
Treatment/medication code: simvastatin
GTEx Portal
Variant Page
Multi-tissue eQTL Comparison
Single-tissue eQTL NES (with 95% CI)
Single-tissue eQTL p-value versus Multi-tissue Posterior Probability
NES
m-value (Posterior Probability from METASOFT)

Top Hits:

Variant	Nearest Gene(s)	MAF	P-value
19:45,412,079 C / T (rs7412)	*APOE*	0.066	1.7e-119
1:55,505,647 G / T (rs11591147)	*PCSK9*	0.031	2.8e-48
19:11,190,632 G / A (rs142130958)	*LDLR*	0.11	6.2e-44
2:21,267,461 G / A (rs934197)	*APOB*	0.29	5.1e-27
1:109,818,306 G / T (rs629301)	*CELSR2*	0.22	6.8e-26
1:54,876,728 A / G (rs148560033)	*SSBP3*	0.028	4.5e-22
5:74,618,628 C / T (rs79225634)	*HMGCR*	0.34	4.7e-17
2:44,065,090 G / A (rs6756629)	*ABCG5*	0.078	1.2e-14
4:74,813,227 A / C (rs182695896)	*PF4*	0.019	6.4e-14

Category	Phenotype	P-value	Beta	Number of samples
Non-cancer illness	Non-cancer illness code, self-reported: high cholesterol	2.4e-76	0.016	
Treatment/medication	Medication for cholesterol, blood pressure or diabetes: Cholesterol lowering medication	1.1e-29	0.018	
Nuclear Magnetic Resonance	Total cholesterol in LDL	5.1e-27	0.12	
Nuclear Magnetic Resonance	Free cholesterol in IDL	6.8e-26	0.12	
Nuclear Magnetic Resonance	Phospholipids in IDL	8.8e-26	0.12	
Treatment/medication	Medication for cholesterol, blood pressure, diabetes, or take exogenous hormones: Cholesterol lowering medication	1.5e-25	0.012	
Nuclear Magnetic Resonance	Free cholesterol in large LDL	3.0e-25	0.12	

Variant ID	Shorthand	rs ID (v151)	Chromosome	Position	MAF >= 1%	Ref Allele	Alt Allele	b37 Variant ID
chr2_21044589_G_A_b38		rs934197	chr2	21044589	true	G	A	2_21267461_G_A_b37

Showing 1 to 1 of 1 entries

Single-Tissue eQTLs for chr2_21044589_G_A_b38

Data Source: GTEx Analysis Release V8 (dbGaP Accession phs000424.v8.p2)

eQTLs of chr2_21044589_G_A_b38

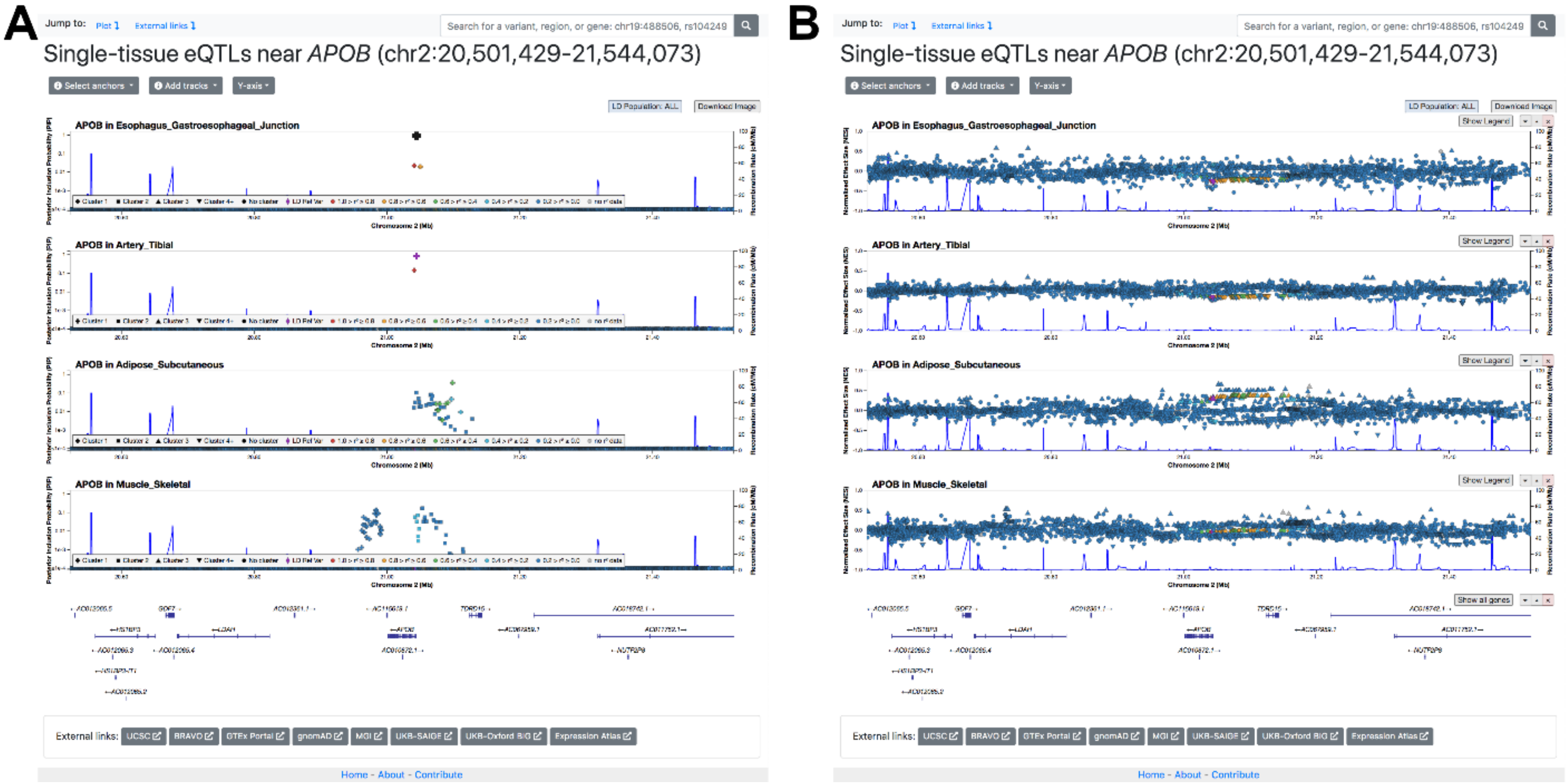
Gencode Id	Gene Symbol	Variant Id	SNP	P-Value	NES	Tissue	Actions
ENSG00000084674.14	APOB	chr2_21044589_G_A_b38	rs934197 dbSNP	4.0e-17	0.32	Adipose - Subcutaneous	eQTL violin plot, IGV Browser, Multi-tissue eQTL Plot
ENSG00000084674.14	APOB	chr2_21044589_G_A_b38	rs934197 dbSNP	8.0e-11	-0.19	Artery - Tibial	eQTL violin plot, IGV Browser, Multi-tissue eQTL Plot
ENSG00000084674.14	APOB	chr2_21044589_G_A_b38	rs934197 dbSNP	4.1e-9	0.26	Nerve - Tibial	eQTL violin plot, IGV Browser, Multi-tissue eQTL Plot
ENSG00000084674.14	APOB	chr2_21044589_G_A_b38	rs934197 dbSNP	6.9e-9	-0.20	Esophagus - Gastroesophageal Junction	eQTL violin plot, IGV Browser, Multi-tissue eQTL Plot
ENSG00000084674.14	APOB	chr2_21044589_G_A_b38	rs934197 dbSNP	1.1e-8	-0.21	Esophagus - Muscularis	eQTL violin plot, IGV Browser, Multi-tissue eQTL Plot
ENSG00000118961.14	LDAH	chr2_21044589_G_A_b38	rs934197 dbSNP	5.3e-7	0.19	Adipose - Visceral (Omentum)	eQTL violin plot, IGV Browser, Multi-tissue eQTL Plot
ENSG00000218819.4	TDRD15	chr2_21044589_G_A_b38	rs934197 dbSNP	5.5e-7	-0.42	Liver	eQTL violin plot, IGV Browser, Multi-tissue eQTL Plot
ENSG00000084674.14	APOB	chr2_21044589_G_A_b38	rs934197 dbSNP	7.3e-7	-0.28	Colon - Sigmoid	eQTL violin plot, IGV Browser, Multi-tissue eQTL Plot
ENSG00000118961.14	LDAH	chr2_21044589_G_A_b38	rs934197 dbSNP	0.0000019	0.26	Cells - Cultured fibroblasts	eQTL violin plot, IGV Browser, Multi-tissue eQTL Plot
ENSG00000084674.14	APOB	chr2_21044589_G_A_b38	rs934197 dbSNP	0.0000040	0.17	Heart - Left Ventricle	eQTL violin plot, IGV Browser, Multi-tissue eQTL Plot

Showing 1 to 10 of 13 entries

ENSG00000084674.14 APOB and chr2_21044589_G_A_b38 eQTL (Meta Analysis RE2 P-Value: 5.5502389999999999e-76)

(A) Using an external resource (the Oxford Brain Imaging Genetics Server) to search for variants associated with total cholesterol in LDL, the variant with the strongest association on chromosome 2 is rs934197, in the *APOB* locus. (B) A PheWAS view of this variant reveals multiple strong association signals, with both traits and medication use directly related to LDL cholesterol. (C) When searching for this variant in the GTEx Portal, we are given a list of eQTLs sorted by P-value, but with no context about whether this variant is the strongest eQTL for each tissue and gene, or whether it is in linkage disequilibrium (LD) with a stronger nearby eQTL. (D) The multi-tissue comparison view in the GTEx Portal provides m-values, which evaluates cross-tissue effects for this variant, but provides no context about whether the signals are confounded by the LD structure of the genomic region.

Supplementary Figure 4.2 PheGET provides genomic context for tissue-specific expression



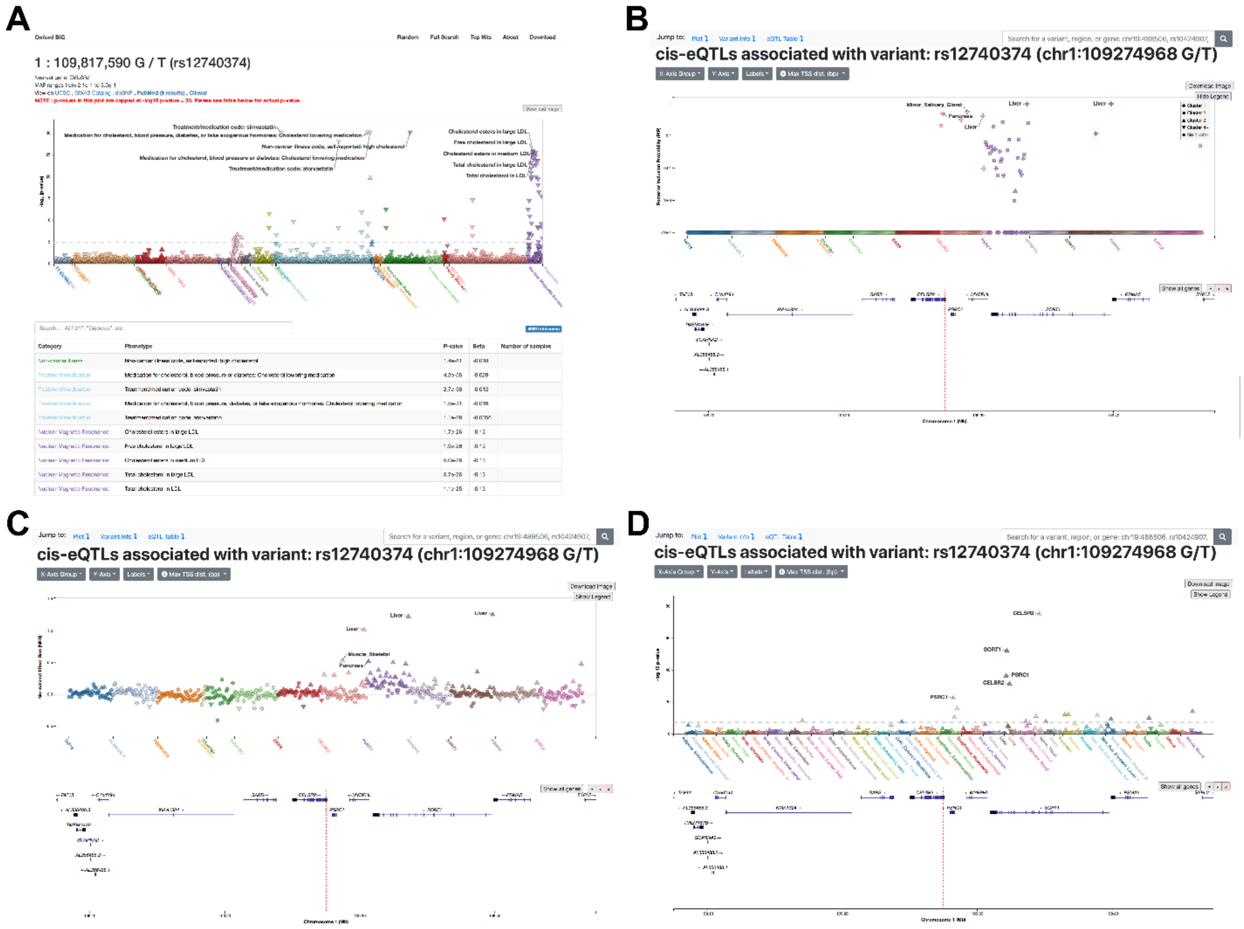
Single-tissue eQTLs near *APOB* (chr2:20,501,429-21,544,073)

Single-tissue eQTLs near *APOB* (chr2:20,501,429-21,544,073)

A
Jump to: Plot External links
Search for a variant, region, or gene: chr19:488506, rs104249
Select anchors
Add tracks
Y-axis
LD Population: ALL
Download Image
APOB in Esophagus_Gastroesophageal_Junction
APOB in Artery_Tibial
APOB in Adipose_Subcutaneous
APOB in Muscle_Skeletal
Posterior Inclusion Probability (PIP)
Recombination Rate (cM/Mb)
Chromosome 2 (Mb)
Cluster 1
Cluster 2
Cluster 3
Cluster 4+
No cluster
LD Ref Var
External links: UCSC BRAVO GTEx Portal gnomAD MGI UKB-SAIGE UKB-Oxford BIG Expression Atlas
Home - About - Contribute
B
Jump to: Plot External links
Search for a variant, region, or gene: chr19:488506, rs104249
Select anchors
Add tracks
Y-axis
LD Population: ALL
Download Image
Show Legend
APOB in Esophagus_Gastroesophageal_Junction
APOB in Artery_Tibial
APOB in Adipose_Subcutaneous
APOB in Muscle_Skeletal
Normalized Effect Size (NES)
Recombination Rate (cM/Mb)
Chromosome 2 (Mb)
Show all genes
External links: UCSC BRAVO GTEx Portal gnomAD MGI UKB-SAIGE UKB-Oxford BIG Expression Atlas
Home - About - Contribute

(A) When we view PIPs for different tissues in the *APOB* locus in PheGET's region view, and use the top signal in esophagus - gastroesophageal junction (EGJ) tissue as the reference variant to compare with signals in other tissues, we see that it is also the top variant in tibial arterial tissue, but is distinct from the signal cluster in subcutaneous adipose and skeletal muscle, showing tissue-specific expression differences associated with distinct LD blocks. (B) When we view by effect sizes, we see that variants associated with lower *APOB* expression in EGJ and tibial artery tissues, but higher expression in subcutaneous adipose tissue, highlighting the heterogeneous effects one variant can have on the same gene in different tissues.

Supplementary Figure 4.3 PheGET shows multiple genes associated with a single variant


A
Oxford BIG
Random
Full Search
Top Hits
About
Download
Download Image
Treatment/medication code: simvastatin
Medication for cholesterol, blood pressure, diabetes, or take exogenous hormones: Cholesterol lowering medication
Non-cancer illness code, self-reported: high cholesterol
Medication for cholesterol, blood pressure or diabetes: Cholesterol lowering medication
Treatment/medication code: atorvastatin
Cholesterol esters in large LDL
Free cholesterol in large LDL
Cholesterol esters in medium LDL
Total cholesterol in large LDL
Total cholesterol in LDL
-log10(p-value)
B
C
D
Jump to: Plot Variant Info eQTL Table
X-Axis Group
Y-Axis
Labels
Max TSS dist. (bp)
Download Image
Hide Legend
Show Legend
Show all genes
Minor_Salivary_Gland
Pancreas
Liver
Muscle_Skeletal
SORT1
PSRC1
CELSR2
Chromosome 1 (Mb)

1 : 109,817,590 G / T (rs12740374)

Nearest gene: CELSR2
MAF ranges from 2.1e-1 to 3.0e-1
View on UCSC , GWAS Catalog , dbSNP , PubMed (9 results) , Clinvar
NOTE : p-values in this plot are capped at -log10 p-value = 30. Please see table below for actual p-value.

Search... "427.21", "Diabetes", etc.

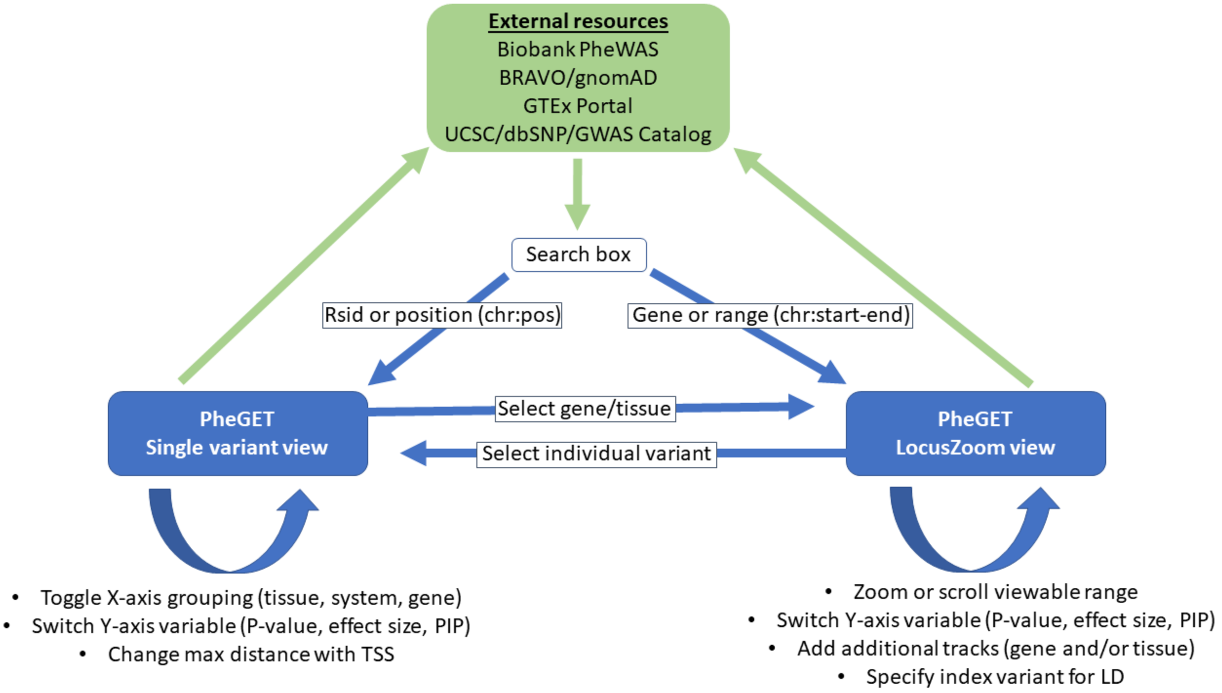
Category	Phenotype	P-value	Beta	Number of samples
Non-cancer illness	Non-cancer illness code, self-reported: high cholesterol	1.4e-41	-0.018	
Treatment/medication	Medication for cholesterol, blood pressure or diabetes: Cholesterol lowering medication	4.3e-35	-0.029	
Treatment/medication	Treatment/medication code: simvastatin	2.7e-38	-0.012	
Treatment/medication	Medication for cholesterol, blood pressure, diabetes, or take exogenous hormones: Cholesterol lowering medication	1.6e-31	-0.018	
Treatment/medication	Treatment/medication code: atorvastatin	1.1e-30	-0.0035	
Nuclear Magnetic Resonance	Cholesterol esters in large LDL	1.7e-26	-0.13	
Nuclear Magnetic Resonance	Free cholesterol in large LDL	1.9e-28	-0.13	
Nuclear Magnetic Resonance	Cholesterol esters in medium LDL	5.0e-26	-0.13	
Nuclear Magnetic Resonance	Total cholesterol in large LDL	6.7e-26	-0.13	
Nuclear Magnetic Resonance	Total cholesterol in LDL	1.1e-25	-0.13	

cis-eQTLs associated with variant: rs12740374 (chr1:109274968 G/T)

Search for a variant, region, or gene: chr19:488506, rs10424907,

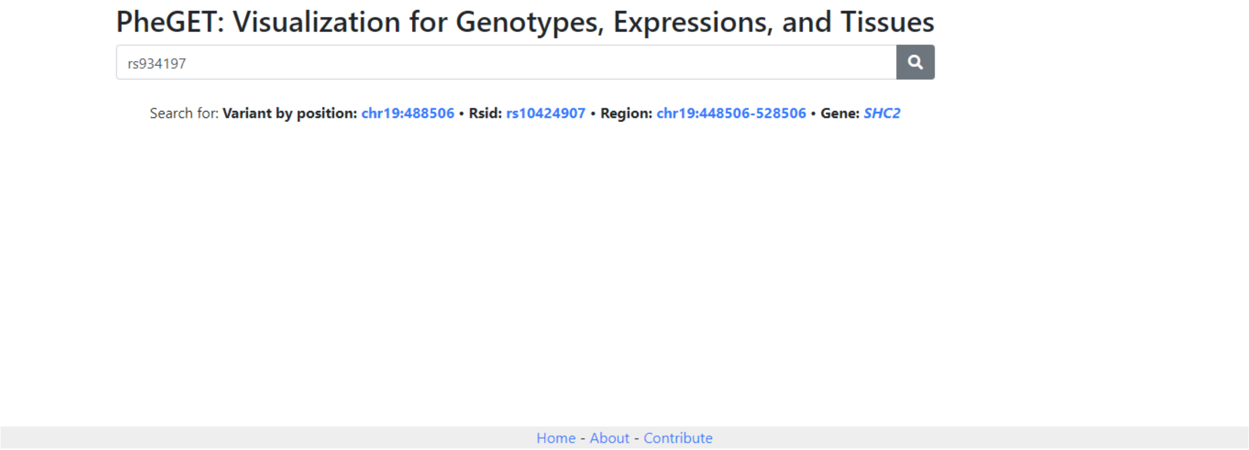
(A) Using UK Biobank as a reference, we searched for variants strongly associated with cholesterol-related traits. The variant rs12740374, located in the *SORT1-PSRC1-CELSR2* locus, has strong associations with multiple cholesterol-related traits. (B) In PheGET's single variant view with PIPs on the y-axis, we can see strong associations with liver tissue for all three genes, along with multiple other tissues, evidence for both tissue-specific and cross-tissue regulatory effects. (C) The pattern of effect sizes of the associations indicate that this variant has strong upregulation effects on all three genes in liver tissue. (D) The strong P-values in liver across all three tissues provide additional evidence for liver-specific regulatory mechanisms associated with this variant.

Supplementary Figure 4.4. Navigating PheGET.


External resources
Biobank PheWAS
BRAVO/gnomAD
GTEx Portal
UCSC/dbSNP/GWAS Catalog
Search box
Rsid or position (chr:pos)
Gene or range (chr:start-end)
PheGET
Single variant view
Select gene/tissue
Select individual variant
PheGET
LocusZoom view
• Toggle X-axis grouping (tissue, system, gene)
• Switch Y-axis variable (P-value, effect size, PIP)
• Change max distance with TSS
• Zoom or scroll viewable range
• Switch Y-axis variable (P-value, effect size, PIP)
• Add additional tracks (gene and/or tissue)
• Specify index variant for LD

An illustration of navigating eQTL data in PheGET. A researcher may wish to learn more about gene regulation related to a variant, gene, or region of interest. A variant query will send the user to a single variant view with extensive information about all eQTL information related to one variant, while a gene or region query will send the user to a LocusZoom view of the gene or region for the tissue with the strongest eQTL association. The user can manipulate the displayed data within each view in real time to show different metrics of association: P-values for evidence of association, effect sizes for strength of regulatory effect, or PIPs for a modeled probability of a variant being causal for an eQTL.

Supplementary Figure 4.5. Step-by-step tutorial on reproducing examples in the PheGET main text.



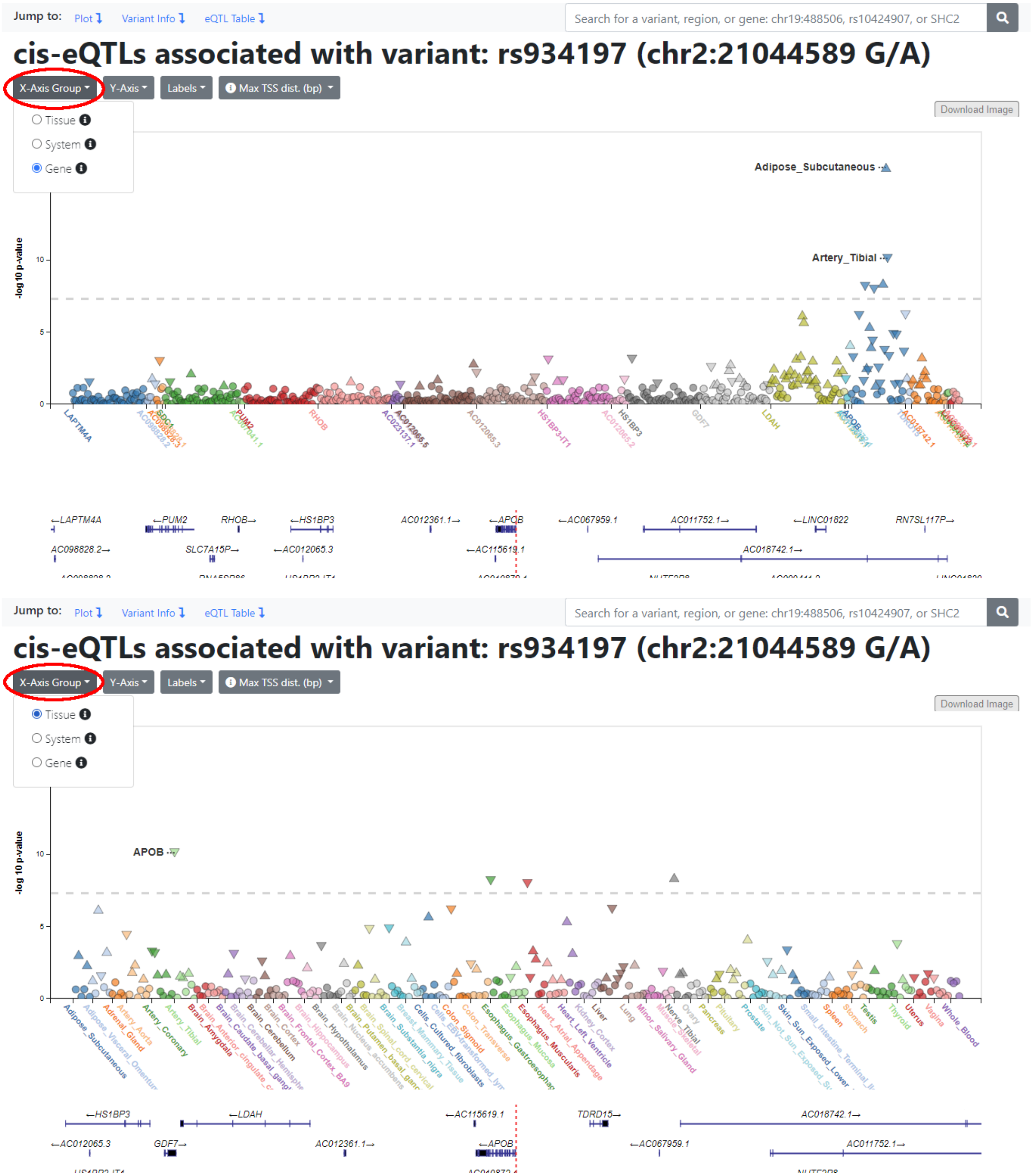
PheGET: Visualization for Genotypes, Expressions, and Tissues

rs934197

Search for: **Variant by position:** chr19:488506 • **Rsid:** rs10424907 • **Region:** chr19:448506-528506 • **Gene:** *SHC2*

We will first enter ‘rs934197’, the Rsid for the top LDL-associated variant located ~500 bp upstream of *APOB*, in the search box. PheGET will send us to a single variant view.

Supplementary Figure 4.6. Dynamically change the grouping variable on the x-axis.



Jump to: Plot Variant Info eQTL Table

Search for a variant, region, or gene: chr19:488506, rs10424907, or SHC2

cis-eQTLs associated with variant: rs934197 (chr2:21044589 G/A)

X-Axis Group

○ Tissue
○ System
● Gene

Jump to: Plot Variant Info eQTL Table

Search for a variant, region, or gene: chr19:488506, rs10424907, or SHC2

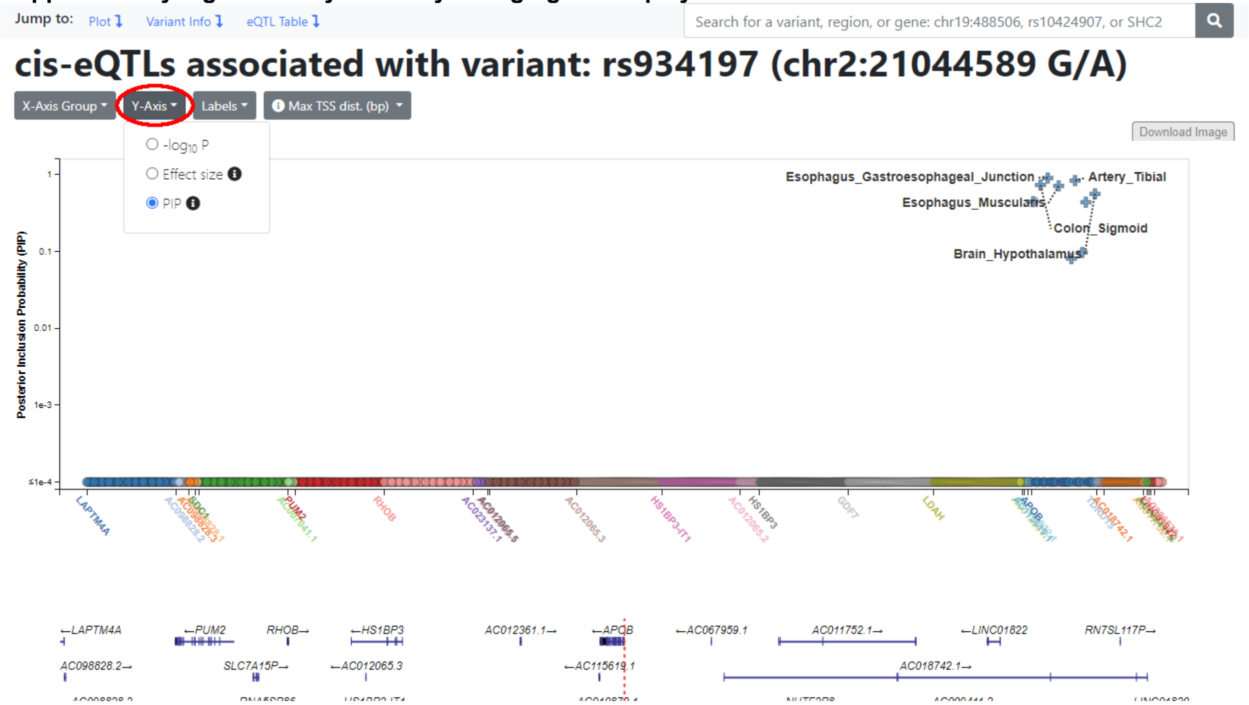
cis-eQTLs associated with variant: rs934197 (chr2:21044589 G/A)

X-Axis Group

● Tissue
○ System
○ Gene

The first dropdown menu gives us the ability to change the x-axis grouping of the data in real time. Currently, the points are grouped by gene, arranged by the genomic positions of their TSS. Grouping by gene will make it easier to see multi-tissue regulation of specific genes, while grouping by tissue will make it easier to see multi-gene regulation in specific tissues. Grouping by system will further group tissues for a more general overview of gene regulation in different tissue systems.

Supplementary Figure 4.7. Dynamically changing the displayed Y-axis variable.



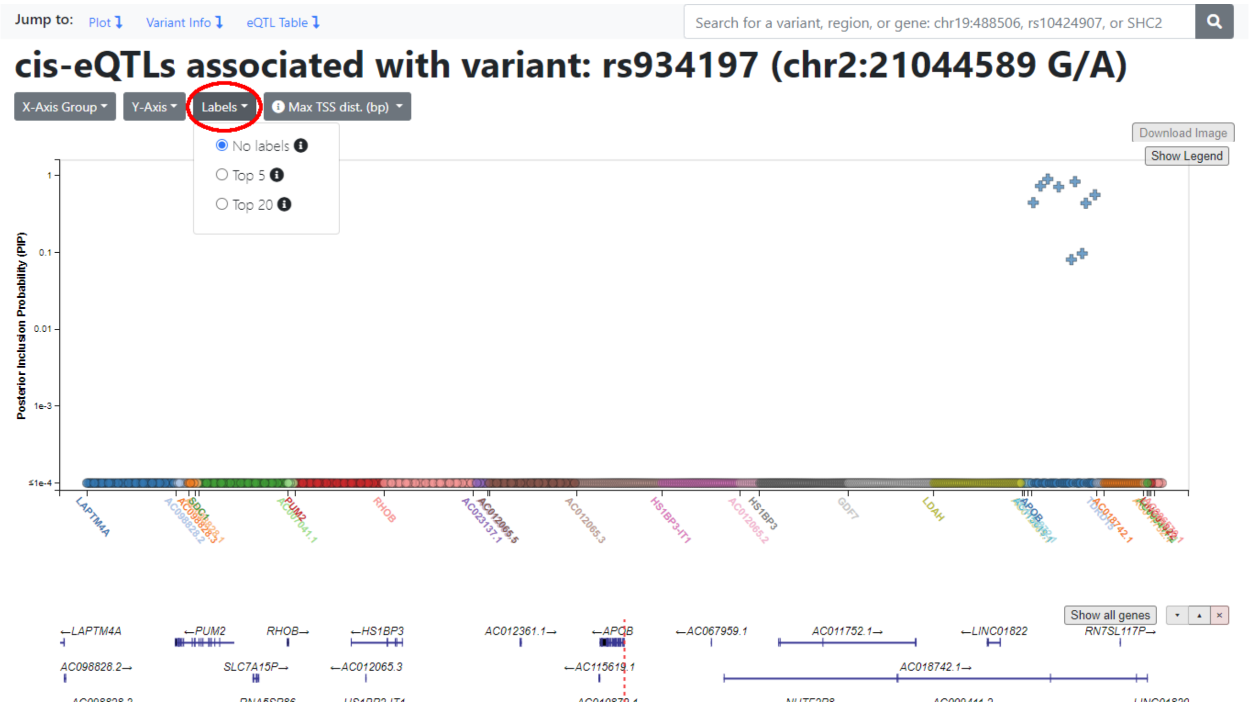
Jump to: Plot Variant Info eQTL Table

cis-eQTLs associated with variant: rs934197 (chr2:21044589 G/A)

○ $-\log_{10}$ P
○ Effect size
● PIP

The second dropdown menu allows us to show y-axis variables other than P-values. For example, we can view PIPs, which take into account the LD structure around a variant to calculate a posterior probability that a variant is causal for an eQTL. For example, subcutaneous adipose, which has a highly significant P-value for its association with the expression of *APOB*, does not have a corresponding PIP signal.

Supplementary Figure 4.8. Toggling labels for the strongest signals.

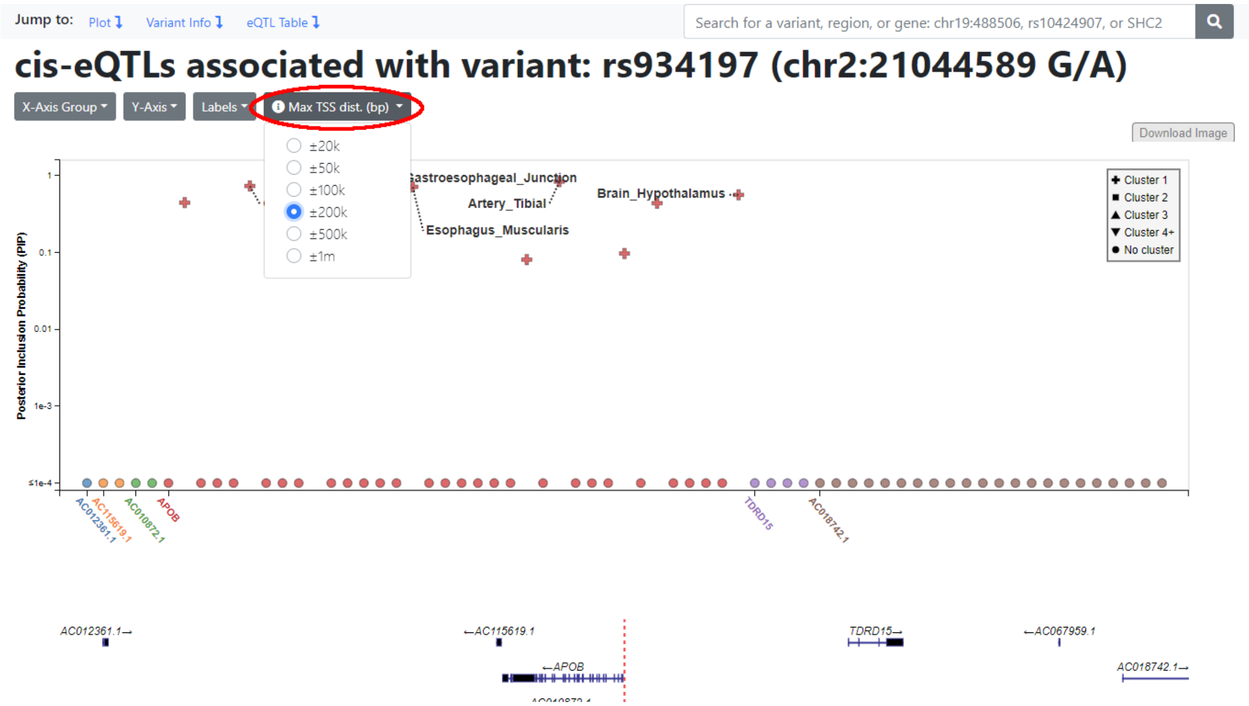


cis-eQTLs associated with variant: rs934197 (chr2:21044589 G/A)

Labels:
● No labels
○ Top 5
○ Top 20

The third dropdown menu allows us to toggle labels for significant signals. Here we turned off the labels to give us a better look at the PIP signals in *APOB*. When the results are grouped on the x-axis by gene, labels will show the tissues for the data points; when the results are grouped by tissue or system, labels will show genes instead.

Supplementary Figure 4.9. Changing the viewing window using maximum TSS distance.



Jump to: Plot ↓ Variant Info ↓ eQTL Table ↓

Search for a variant, region, or gene: chr19:488506, rs10424907, or SHC2

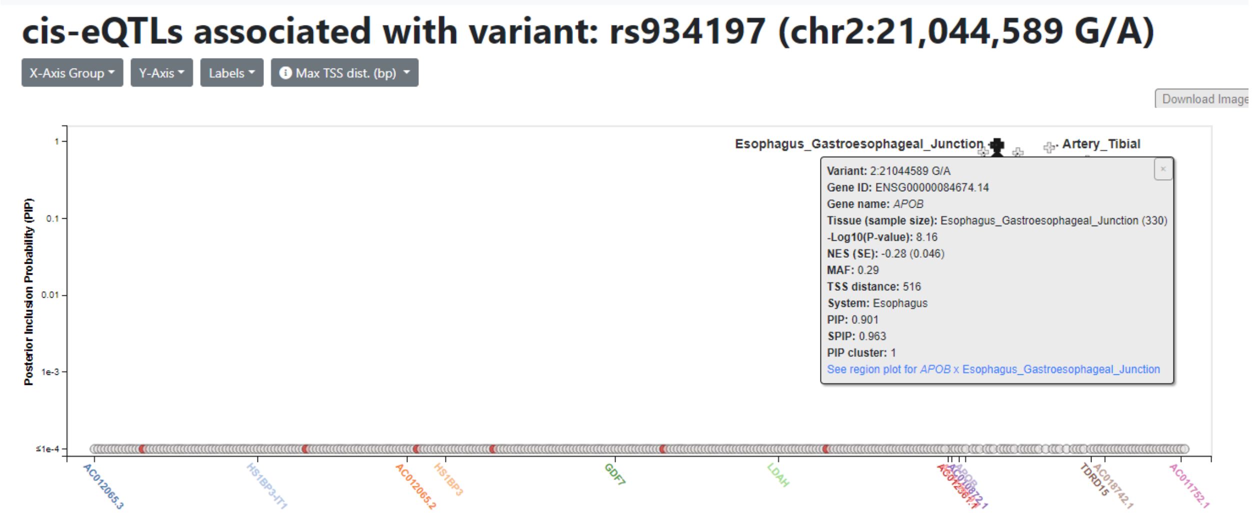
cis-eQTLs associated with variant: rs934197 (chr2:21044589 G/A)

X-Axis Group ▾ Y-Axis ▾ Labels ▾ Max TSS dist. (bp) ▾

○ ±20k
○ ±50k
○ ±100k
● ±200k
○ ±500k
○ ±1m

The fourth dropdown menu allows us to change the amount of data displayed by setting a maximum transcription start site (TSS) distance from the current variant. For example, setting this to ±200k means displaying only the eQTLs for genes with a TSS within 200kbp of the current variant.

Supplementary Figure 4.10. Dynamic highlighting of points with a shared attribute.

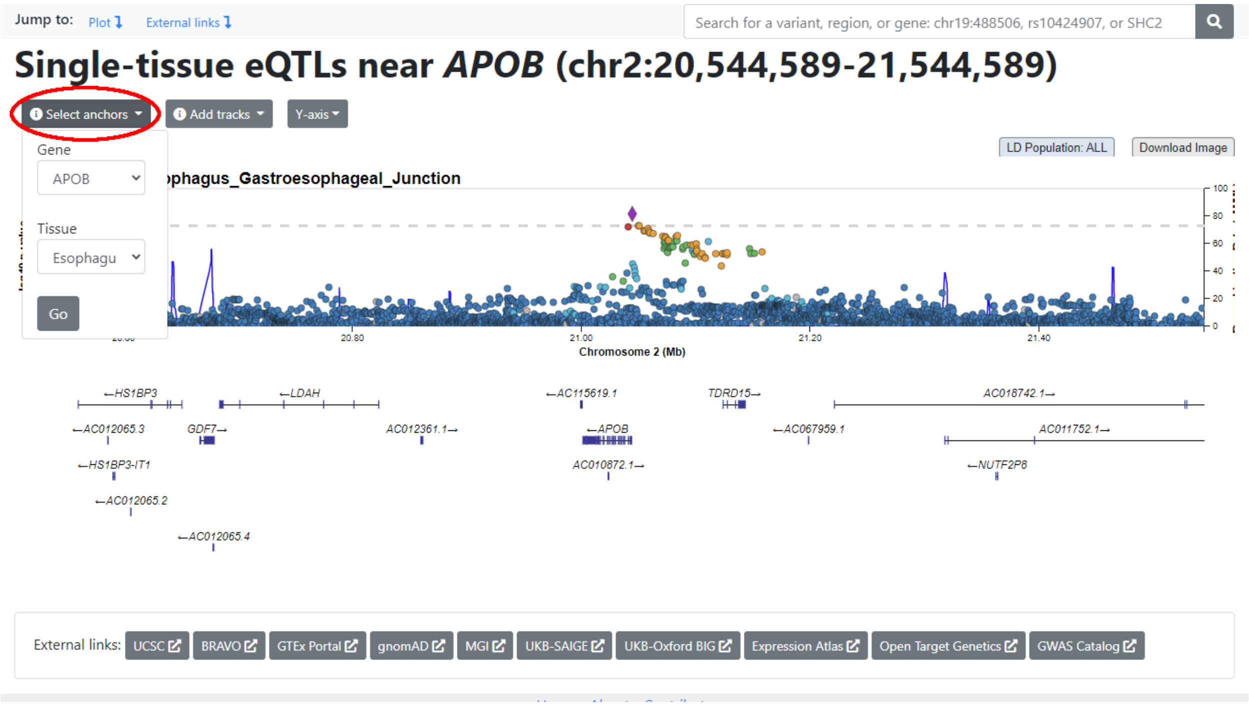


cis-eQTLs associated with variant: rs934197 (chr2:21,044,589 G/A)

Variant: 2:21044589 G/A
Gene ID: ENSG00000084674.14
Gene name: *APOB*
Tissue (sample size): Esophagus_Gastroesophageal_Junction (330)
-Log10(P-value): 8.16
NES (SE): -0.28 (0.046)
MAF: 0.29
TSS distance: 516
System: Esophagus
PIP: 0.901
SPIP: 0.963
PIP cluster: 1
See region plot for *APOB* x Esophagus_Gastroesophageal_Junction

Clicking on any point highlights all other points which share the same labeled attribute. Here, we see all other eQTLs in subcutaneous adipose tissue highlighted in red. It also brings up a tooltip window which displays detailed information about the eQTL, with a link to a region view around this signal. We will follow the link to a LocusZoom view of the region around rs934197 in gastroesophageal junction tissue.

Supplementary Figure 4.11. A LocusZoom view is defined by an anchor gene and tissue.



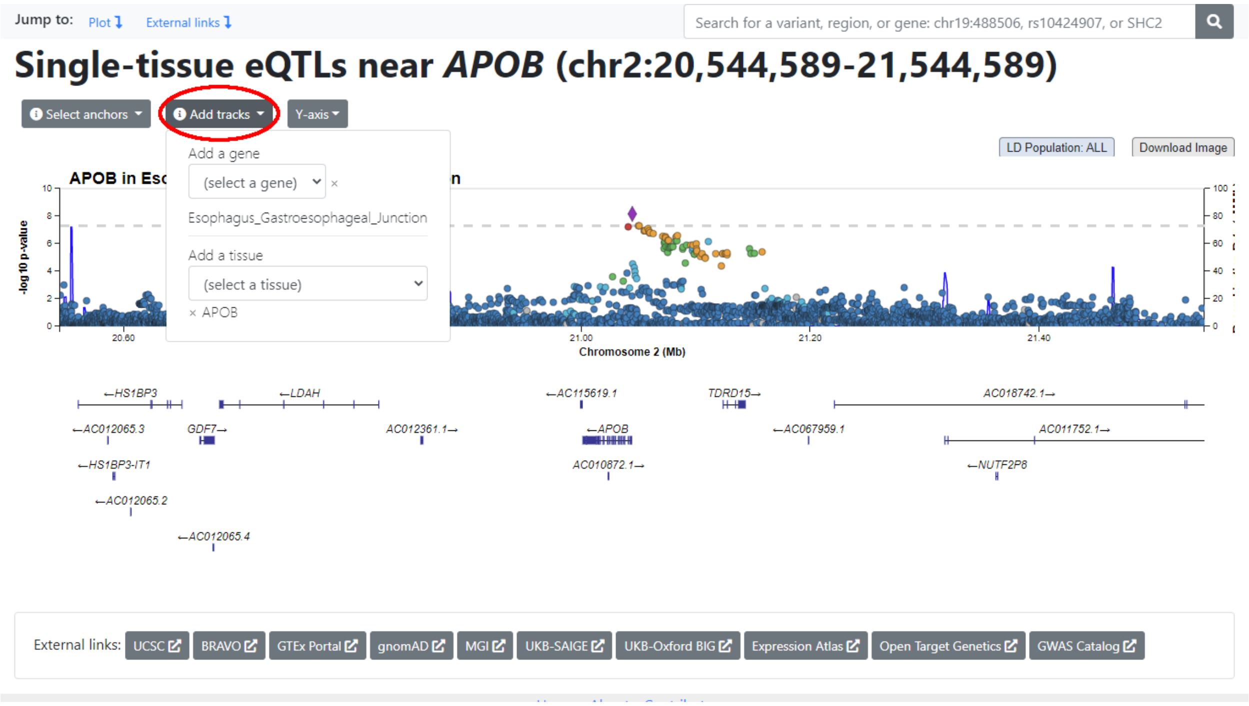
Single-tissue eQTLs near *APOB* (chr2:20,544,589-21,544,589)

Gene: APOB

Tissue: Esophagu

The LocusZoom view around our variant is “anchored” around one gene and one tissue. In our example, the anchors are *APOB* and Esophagus – Gastroesophageal Junction, respectively. Anchors give us stable pivot points around which we can explore additional genes and tissues. We can change our anchor gene and tissue via the first pulldown menu.

Supplementary Figure 4.12. Adding additional tracks in LocusZoom view to facilitate comparison.



Single-tissue eQTLs near *APOB* (chr2:20,544,589-21,544,589)

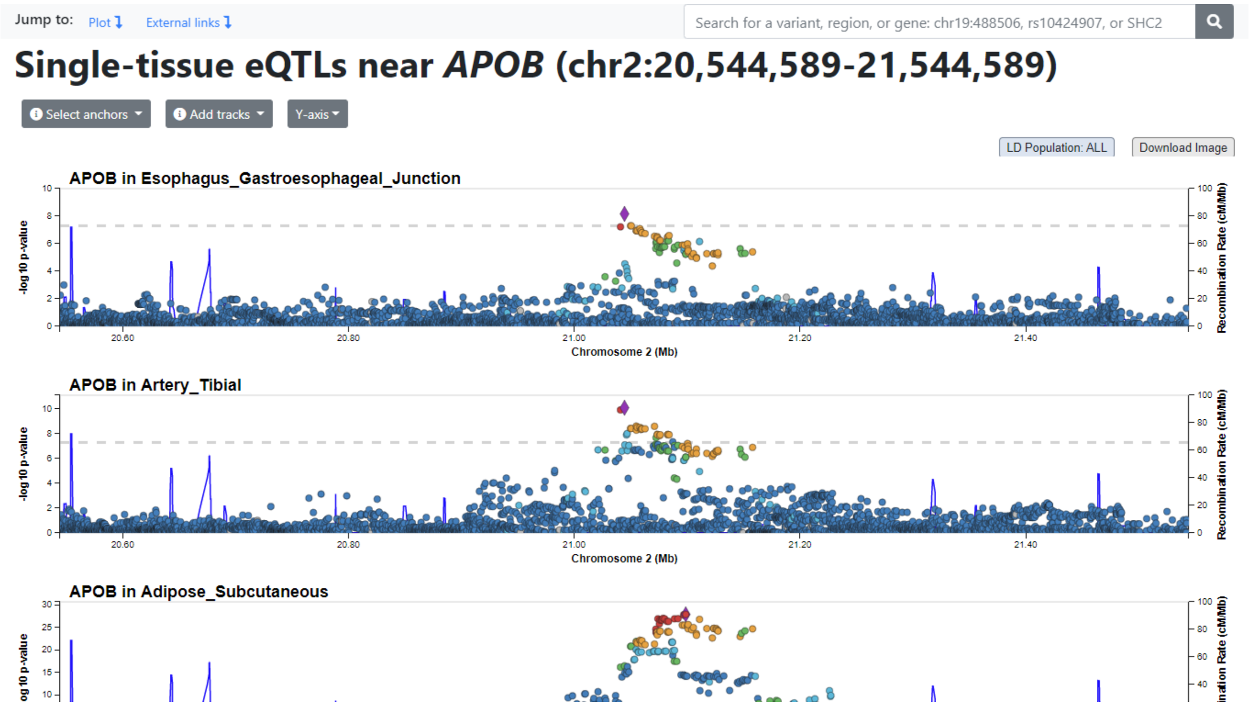
Search for a variant, region, or gene: chr19:488506, rs10424907, or SHC2

Add a gene: (select a gene)

Esophagus_Gastroesophageal_Junction

Add a tissue: (select a tissue)

× APOB

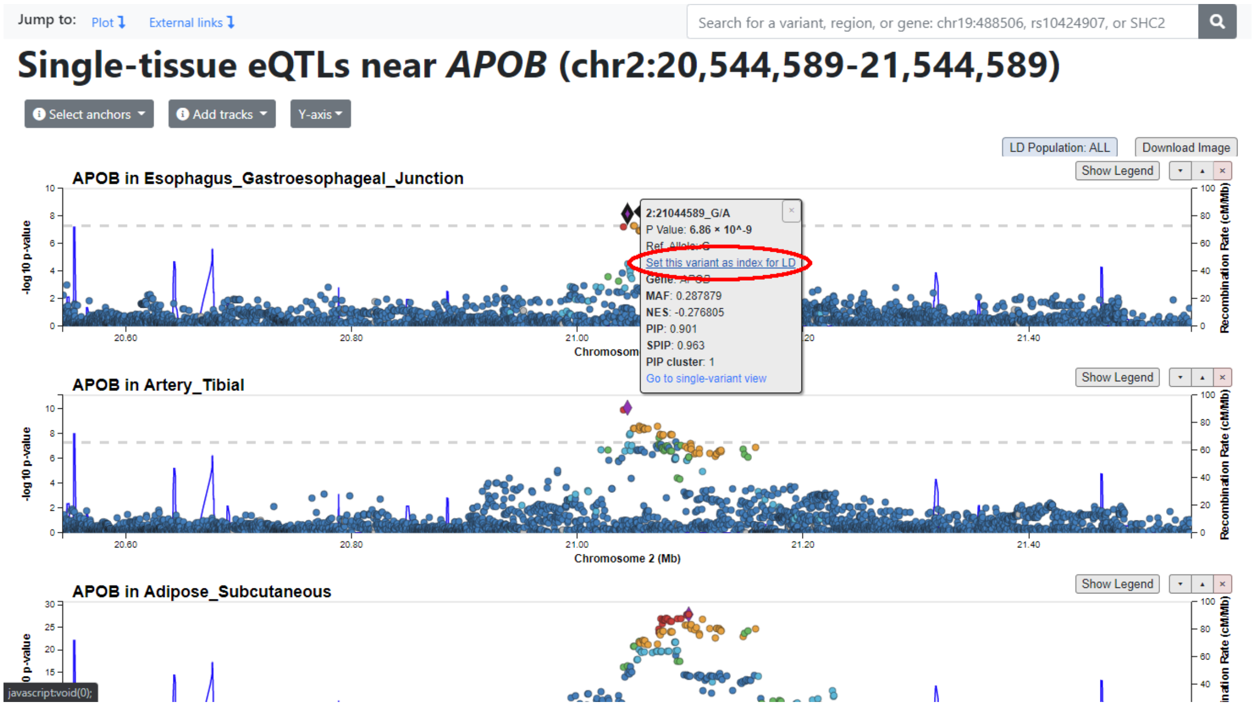


Single-tissue eQTLs near *APOB* (chr2:20,544,589-21,544,589)

Search for a variant, region, or gene: chr19:488506, rs10424907, or SHC2

The second pulldown menu allows us to add genes or tissues with respect to one of the anchors. For example, we can add a track to see eQTLs for *APOB*, our anchor gene, in a different tissue. Similarly, we can add a track to see eQTLs for a different gene in our anchor tissue. We will add tibial artery and subcutaneous adipose as additional tissue tracks.

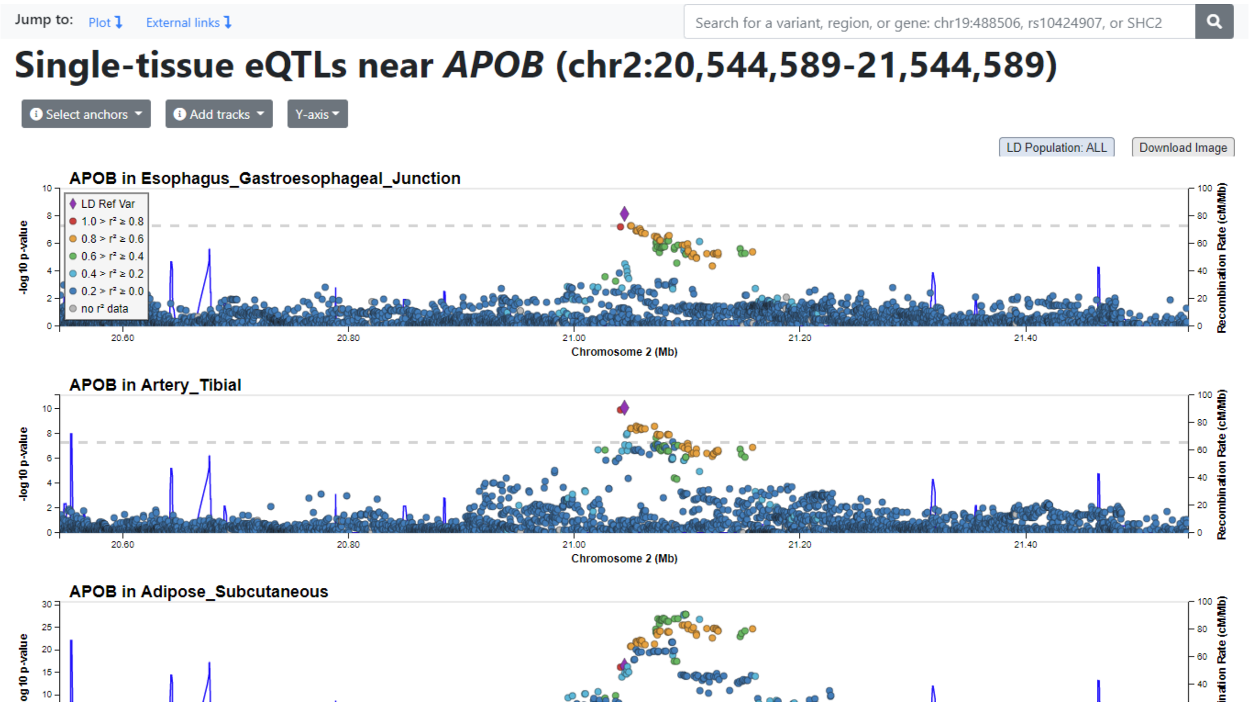
Supplementary Figure 4.13. Setting a reference variant to obtain linkage disequilibrium information.



Single-tissue eQTLs near *APOB* (chr2:20,544,589-21,544,589)

2:21044589_G/A
P Value: 6.86 × 10^-9
Ref. Allele: G
Set this variant as index for LD
Gene: APOB
MAF: 0.287879
NES: -0.276805
PIP: 0.901
SPIP: 0.963
PIP cluster: 1
Go to single-variant view

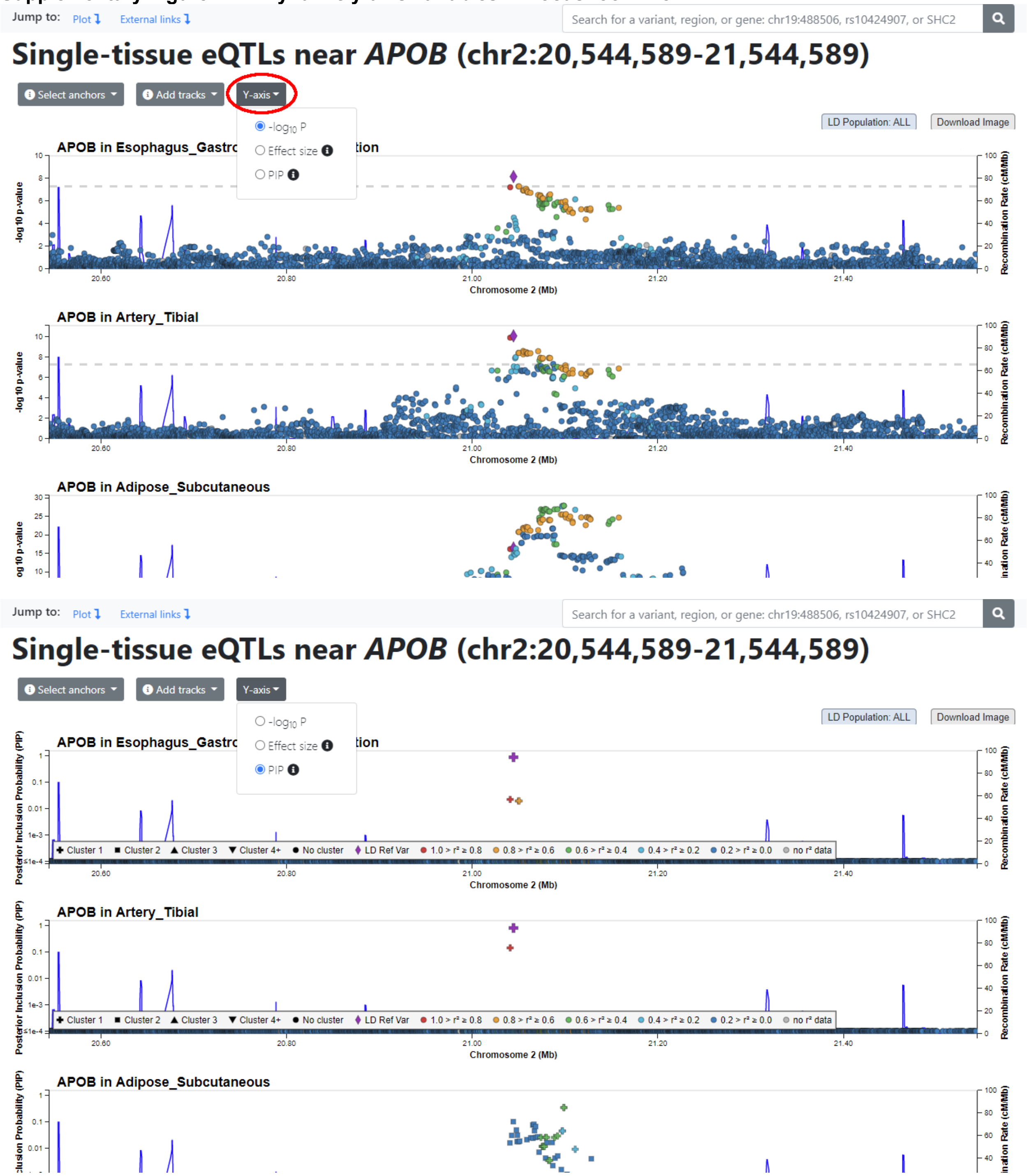
Jump to: Plot External links
Search for a variant, region, or gene: chr19:488506, rs10424907, or SHC2
Select anchors
Add tracks
Y-axis
LD Population: ALL
Download Image
Show Legend
APOB in Esophagus_Gastroesophageal_Junction
APOB in Artery_Tibial
APOB in Adipose_Subcutaneous
-log 10 p-value
Recombination Rate (cM/Mb)
Chromosome 2 (Mb)
20.60
20.80
21.00
21.20
21.40
javascript:void(0);



Single-tissue eQTLs near *APOB* (chr2:20,544,589-21,544,589)

The tooltip window in region view allows us to set a variant as the index for linkage disequilibrium (LD), which will recolor the points in all the displayed plots to reflect their LD with the index. LD information is based on data from the 1000 Genomes Project. The index variant is indicated by a purple diamond. We see that our index variant is the top signal in esophageal and arterial tissues, but is a shadow of a stronger signal in adipose.

Supplementary Figure 4.14. Dynamic y-axis variables in LocusZoom view.



Jump to: Plot External links

Search for a variant, region, or gene: chr19:488506, rs10424907, or SHC2

Single-tissue eQTLs near *APOB* (chr2:20,544,589-21,544,589)

● $-\log_{10}$ P
○ Effect size
○ PIP

APOB in Esophagus_Gastr… tion

APOB in Artery_Tibial

APOB in Adipose_Subcutaneous

Jump to: Plot External links

Search for a variant, region, or gene: chr19:488506, rs10424907, or SHC2

Single-tissue eQTLs near *APOB* (chr2:20,544,589-21,544,589)

○ $-\log_{10}$ P
○ Effect size
● PIP

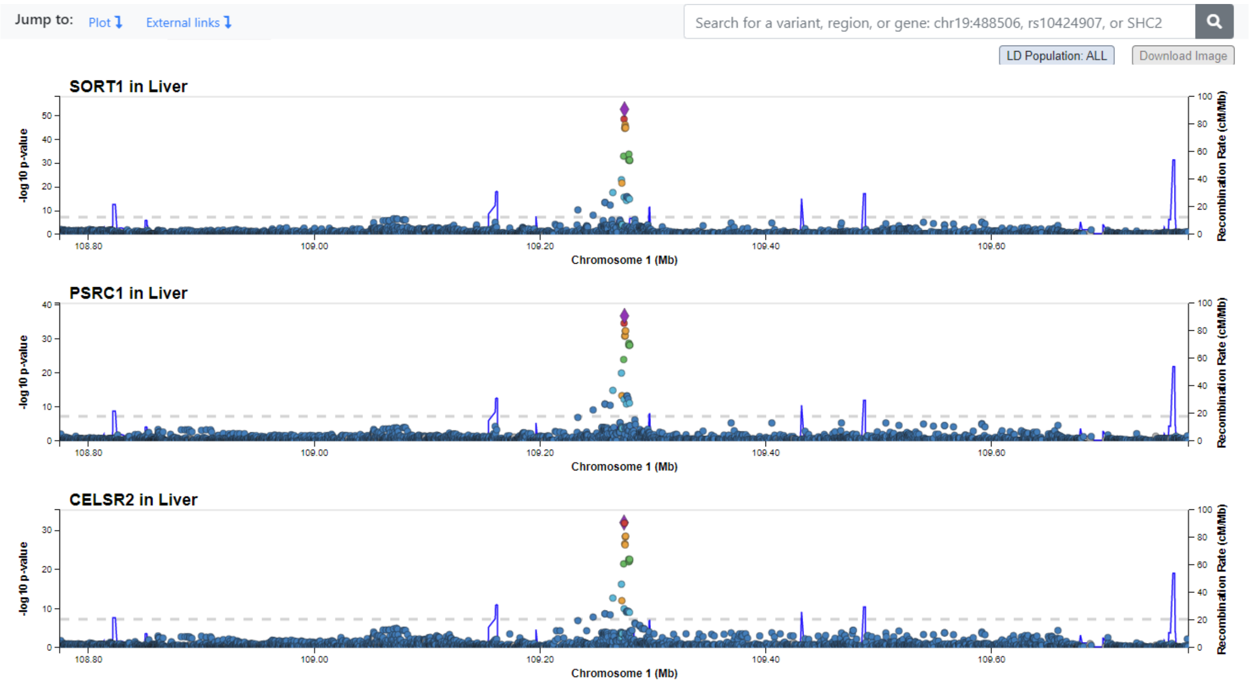
APOB in Esophagus_Gastr… tion

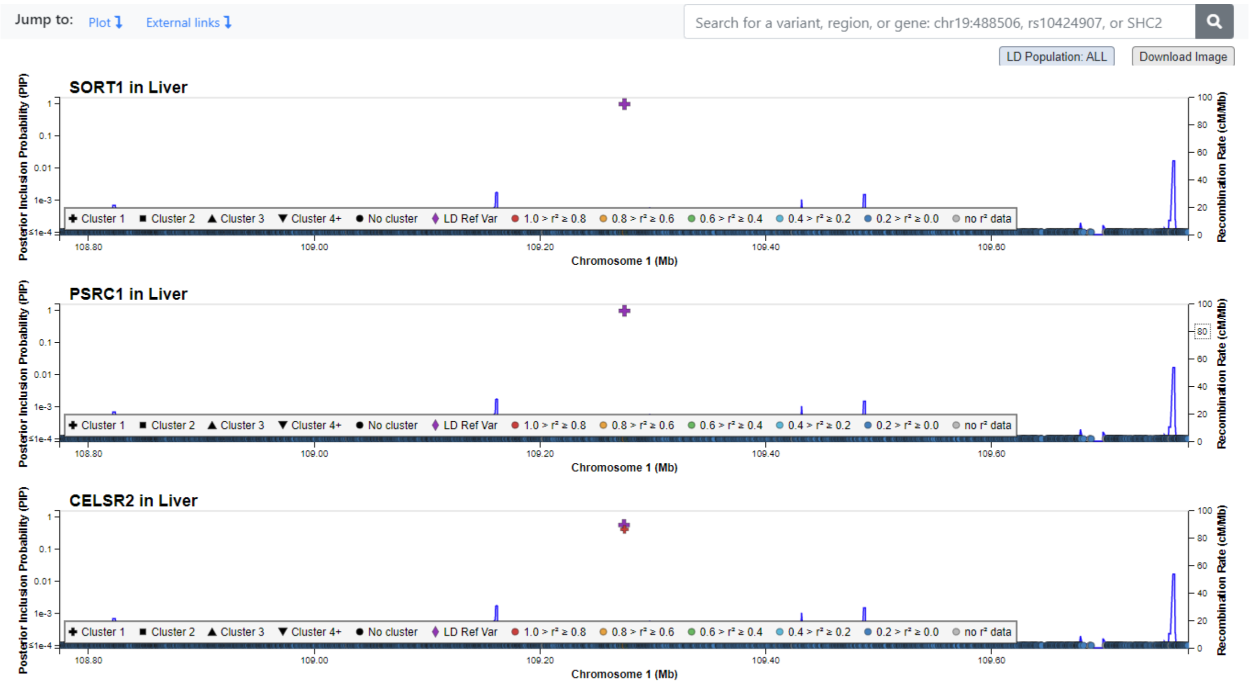
APOB in Artery_Tibial

APOB in Adipose_Subcutaneous

The third pulldown menu changes the Y-axis variable just like in our single variant view. In PIP view, we see that in esophageal and arterial tissues, there is only one signal cluster and our index variant is the strongest contributor. However, in adipose, the PIP signal is distributed more evenly between two different variant clusters, indicating more uncertainty about the potentially causal variant for eQTLs in this tissue.

Supplementary Figure 4.15. Comparing expressions of multiples genes in one tissue via LocusZoom view.




Jump to: Plot External links
Search for a variant, region, or gene: chr19:488506, rs10424907, or SHC2
LD Population: ALL
Download Image
SORT1 in Liver
PSRC1 in Liver
CELSR2 in Liver
Posterior Inclusion Probability (PIP)
Recombination Rate (cM/Mb)
1
0.1
0.01
1e-3
≤1e-4
✚ Cluster 1 ■ Cluster 2 ▲ Cluster 3 ▼ Cluster 4+ ● No cluster ♦ LD Ref Var ● 1.0 > r² ≥ 0.8 ● 0.8 > r² ≥ 0.6 ● 0.6 > r² ≥ 0.4 ● 0.4 > r² ≥ 0.2 ● 0.2 > r² ≥ 0.0 ● no r² data
108.80
109.00
109.20
109.40
109.60
Chromosome 1 (Mb)

We can also compare the eQTL signals for multiple genes in the same tissue. Here, we searched for the top variant for cholesterol-related association signals in the *SORT1-PSRC1-CELSR2* locus, rs12740374, and navigated to a region view. Using liver tissue and *SORT1* as anchors, we added the other two genes to compare the eQTL signals between them. We see that rs12740374 is the strongest signal for these three genes, both for P-values and PIPs, suggesting a regulatory pathway involving all three.

Supplementary Table 4.1 Comparison of top PIP variants for *APOB* across different tissues

Tissue	Max-PIP Variant	Max PIP	Min p-value
Skin – Sun-Exposed Lower leg	rs579826	0.451	4.6E-58
Skin – Not Sun-Exposed Suprapubic	rs579826	0.342	7.9E-35
Heart – Left Ventricle	rs661665	0.748	1.1E-30
Adipose – Subcutaneous	rs4665178	0.354	1.4E-28
Nerve – Tibial	rs66984774	0.412	1.5E-15
Heart – Atrial Appendage	rs661665	0.748	1.1E-11
Artery – Tibial	rs934197	0.833	8.0E-11
Esophagus – Gastroesophageal Junction	rs934197	0.901	6.9E-09
Esophagus– Muscularis	rs934197	0.711	1.1E-08
Colon – Sigmoid	rs934197	0.730	7.3E-07
Muscle – Skeletal	rs56327713	0.138	8.5E-07

Tibial artery, gastroesophageal junction, sigmoid colon, and skeletal muscle tissues share the same top PIP signal (rs934197) for affecting the expression of *APOB*. Both skin tissues share a different variant as the top signal (rs579826), while subcutaneous adipose tissue has a third variant (rs4665178) with the strongest PIP. This highlights differences in regulatory mechanisms between different tissues for the same gene, providing clues for a better understanding of tissue-specific gene expression.

4.5 References

Aguet, F., A. N. Barbeira, R. Bonazzola, A. Brown, S. E. Castel *et al.*, 2019 The GTEx Consortium atlas of genetic regulatory effects across human tissues. bioRxiv.

Ehrlich, K. C., M. Lacey and M. Ehrlich, 2019 Tissue-specific epigenetics of atherosclerosis-related ANGPT and ANGPTL genes. Epigenomics 11**:** 169-186.

Gagliano Taliun, S. A., P. VandeHaar, A. P. Boughton, R. P. Welch, D. Taliun *et al.*, 2020 Exploring and visualizing large-scale genetic associations by using PheWeb. Nat Genet 52**:** 550-552.

Gallagher, M. D., and A. S. Chen-Plotkin, 2018 The Post-GWAS Era: From Association to Function. Am J Hum Genet 102**:** 717-730.

Haeussler, M., A. S. Zweig, C. Tyner, M. L. Speir, K. R. Rosenbloom *et al.*, 2019 The UCSC Genome Browser database: 2019 update. Nucleic Acids Res 47**:** D853-D858.

Karczewski, K. J., L. C. Francioli, G. Tiao, B. B. Cummings, J. Alföldi *et al.*, 2020 The mutational constraint spectrum quantified from variation in 141,456 humans. bioRxiv.

Klarin, D., S. M. Damrauer, K. Cho, Y. V. Sun, T. M. Teslovich *et al.*, 2018 Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. Nat Genet 50**:** 1514-1523.

Liu, B., M. Pjanic, T. Wang, T. Nguyen, M. Gloudemans *et al.*, 2018 Genetic Regulatory Mechanisms of Smooth Muscle Cells Map to Coronary Artery Disease Risk Loci. Am J Hum Genet 103**:** 377-388.

Musunuru, K., A. Strong, M. Frank-Kamenetsky, N. E. Lee, T. Ahfeldt *et al.*, 2010 From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature 466**:** 714-719.

The NHLBI Trans-Omics for Precision Medicine (TOPMed) Whole Genome Sequencing Program, 2018 BRAVO variant browser: University of Michigan and NHLBI. Available from: https://bravo.sph.umich.edu/freeze5/hg38/

Niu, C., Z. Luo, L. Yu, Y. Yang, Y. Chen *et al.*, 2017 Associations of the APOB rs693 and rs17240441 polymorphisms with plasma APOB and lipid levels: a meta-analysis. Lipids Health Dis 16**:** 166.

Pruim, R. J., R. P. Welch, S. Sanna, T. M. Teslovich, P. S. Chines *et al.*, 2010 LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics 26**:** 2336-2337.

Roselli, C., M. D. Chaffin, L. C. Weng, S. Aeschbacher, G. Ahlberg *et al.*, 2018 Multi-ethnic genome-wide association study for atrial fibrillation. Nat Genet 50**:** 1225-1233.

Schadt, E. E., C. Molony, E. Chudin, K. Hao, X. Yang *et al.*, 2008 Mapping the genetic architecture of gene expression in human liver. PLoS Biol 6**:** e107.

Wang, X., A. Raghavan, D. T. Peters, E. E. Pashos, D. J. Rader *et al.*, 2018 Interrogation of the Atherosclerosis-Associated SORT1 (Sortilin 1) Locus With Primary Human Hepatocytes, Induced Pluripotent Stem Cell-Hepatocytes, and Locus-Humanized Mice. Arterioscler Thromb Vasc Biol 38**:** 76-82.

Wen, X., R. Pique-Regi and F. Luca, 2017 Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. PLoS Genet 13**:** e1006646.

Wu, Y., K. A. Broadaway, C. K. Raulerson, L. J. Scott, C. Pan *et al.*, 2019 Colocalization of GWAS and eQTL signals at loci with multiple signals identifies additional candidate genes for body fat distribution. Hum Mol Genet 28**:** 4161-4172.

Yengo, L., J. Sidorenko, K. E. Kemper, Z. Zheng, A. R. Wood *et al.*, 2018 Meta-analysis of genome-wide association studies for height and body mass index in approximately 700000 individuals of European ancestry. Hum Mol Genet 27**:** 3641-3649.

Chapter 5

Conclusion

5.1 Summary

In this dissertation, we assessed the strengths and weaknesses of different approaches in genome wide association studies, developed and implemented a robust Hardy-Weinberg test adjusting for sample ancestry to aid in variant calling quality control, and created a browser for eQTL data designed for clarity and ease of use. Here we review these works, discuss their limitations, placing them in the context of the current trends in genetic studies, and explore future directions for research.

5.2 The analysis and interpretation of GWAS

In Chapter 2, we performed a genome wide association study for age-related macular degeneration and assessed different approaches in analyzing and interpreting association results. Matching cases and controls by age and restricting our samples to those of European ancestry helped mitigate problems with population stratification. From single-variant association tests, taking into account the difference in sample size of our study when compared to larger studies (FRITSCHE *et al.* 2016), we were still able to replicate some of the strongest known loci. Using group-based association tests, we were able to assess the importance of loss-of-function variants in several genes. We were also able to identify an additional risk variant in the *C2/CFB/SKIV2L* locus by comparing the haplotypes within the region with respect to two known risk variants. We explored approaches to incorporate external data to improve the power of discovering very rare loss-of-function variants. Finally, we followed up on existing studies on likely causal variants within the *C3* locus, providing evidence of association for a missense variant in *NRTN*.

Our approach in analysis was informed by previous studies which identified a larger set of risk loci, most importantly the analysis from the International AMD Genomics Consortium from (FRITSCHE *et al.* 2016). With these studies serving as the vanguard in defining a broad set of associated loci, we were then able to both systematically scan the whole genome to confirm previous findings while aiming follow-up association tests to a narrower and more targeted set of potential risk loci. We were able to use several different approaches—single-variant associations, group-based associations, and an augmented Fisher's exact test—to confirm association signals in genes belonging to the complement system, long known to be important factors for AMD (GEERLINGS *et al.* 2017). By disentangling the signals between *C3*, which has been extensively studied, and *NRTN*, which has received much less attention, we were then able to provide additional insight into the possible functional role of the latter with the help from the results of a mouse-based study, highlighting the importance of integrating the results from multiple types of association tests with functional studies in living cells or organisms to provide a better understanding of the observed signals. In the current era of large consortium-based GWAS, with decreasing per-sample sequencing costs and larger sample sizes, the discovery of loci has become a less daunting issue, supplanted by the much more complex task of interpreting the association signals to generate testable hypotheses about gene functions, which will be useful for guiding future biological studies.

While there is some existing work on using external control samples to boost association power (LEE *et al.* 2017; HENDRICKS *et al.* 2018), the problem of how to identify true differences between cases and controls due to disease association while controlling for batch- or ancestry-based differences between internal and external samples has proven to be quite difficult. Our Fisher's exact test-based approach requires case and control samples to have matched ancestries, and for internal and external samples to not significantly differ in the frequency of loss-of-function variants, assumptions which can be relaxed with the development of better ways to adjust for ancestry and batch effects for external samples. While a statistically complicated problem, using external genetic samples to improve power for very rare variant associations could prove to be very powerful in generating new insights from existing genetic studies, and would be especially valuable for moderately-sized disease-specific sequencing studies which are generally underpowered to find associations for loss-of-function variants with large effect sizes, which are typically very rare due to negative selection (MACARTHUR *et al.* 2012).

Future studies of age-related macular degeneration will require much more focus on functional fine-mapping, such as family-based linkage studies (RATNAPRIYA *et al.* 2020) and expression studies (MENON *et al.* 2019; STRUNZ *et al.* 2020). To close the gap between genetics and disease will require a holistic approach, combining results from different kinds of studies to provide testable hypotheses for in vitro and in vivo genetic studies. With the development of CRISPR-based gene editing technologies (JINEK *et al.* 2012), much more rapid and targeted laboratory-based genetic studies are now possible, finally allowing researchers to bridge the gap between association results and biological function.

5.3 Quality control for diverse genetic data

In Chapter 3, we developed a robust and unified test for Hardy-Weinberg equilibrium (HWE) with a computationally efficient implementation capable of processing tens of thousands of samples for millions of variants and evaluated its performance along with existing methods. With larger and more diverse samples for increasing numbers of variants found in modern genetic studies, the traditional HWE test, in its use as a quality control metric for variant calling, has been stretched to its limit. A common strategy used in large genetic studies is to perform HWE tests either within processed batches or within each contributing cohort, setting a strict p-value threshold (typically 10^{-6}), and using the minimum p-value for a given variant across all batches or cohorts as the filtering criterion. If substantial population structure exists within a tested batch or cohort, then this procedure will remove otherwise high-quality variants from downstream analysis. Moreover, if a cohort contains both population stratification (which generally decreases heterozygosity) and technical errors (which generally increases heterozygosity), then some low-quality variants may not be significant at the chosen p-value threshold. Our proposed method addresses these issues by explicitly modeling and adjusting for the deviation from HWE caused by differences in genetic ancestry, for a more reliable metric for the quality of variant calls. In addition, our software implementation can directly process commonly used file formats, with the ability to handle the large sample sizes and variant counts typical in modern genetic studies. Our software was designed to be easy to integrate into pipelines for variant calling quality control in large-scale genetic studies and is currently used in the TOPMed variant calling pipeline.

Our method can be improved in a few ways. Our model assumptions may be further refined to improve performance (for example, a more accurate model for the relationship between ancestry summary statistics and individual-specific allele frequencies can improve the accuracy of our model), Also, an efficient method to account for family structure for closely related samples can further help reduce false positives. In principle, it is possible for our method to support genotype dosages obtained from imputation, though in practice both genotype scaffolds (target panels) and imputation reference panels should undergo strict QC before genotype imputation.

It may be possible to relax our model assumptions to allow for more flexibility in our data. For example, our method currently assumes a constant inbreeding coefficient across all samples after adjusting for global ancestry. A more general model could allow for individual-level differences in inbreeding coefficients to better model the data, though it would require modifying the statistical tests, because the null hypothesis would now involve an individual-level inbreeding estimate instead of a global inbreeding coefficient.

5.4 Visualization of eQTLs in multiple tissues

In Chapter 4, we developed a web-based browser for displaying eQTL data for multiple tissues and genes, with a focus on providing a convenient and intuitive navigation interface designed to provide useful information clearly. Using the latest GTEx data as a proof of principle, the eQTL browser has two main visualization modes: a single variant view and a region view. For the single variant view, we provide a comprehensive picture of the queried variant's effect on the expression of all surrounding genes in all tissues, while for the region view, the user can explore the effects of variants in a region on the expression of a given gene within any tissue, and compare different tissues or genes. We also provide an easy way to navigate between the two views. Additionally, for each view, the user can customize the value displayed on the plot, between p-values, effect sizes, and posterior inclusion probabilities (PIP) from DAP-G (WEN *et al.* 2017). Finally, we provide a unified search box capable of handling variant, region, or gene queries, facilitating the retrieval of eQTL information of interest to their particular field of research.

While the current implementation of the browser (https://eqtl.pheweb.org) was designed with GTEx as the data source, the underlying technology can be used to display any multi-tissue gene expression dataset. The same technology can be adapted to show other multi-dimensional genomic data, for which typical GWAS or PheWAS plots are insufficient. For example, single-cell expression data can be shown using the same views as our browser, with predicted clusters in place of tissues and boxplots in place of data points, to allow for convenient comparison of expression levels between different clusters for any given gene.

The visualization elements of the browser can be improved in several ways. For example, a new genome-wide tissue-focused view, in the form of a Manhattan plot of the most significant signals across all genes in any given tissue, can provide another intuitive starting point for data exploration, especially for researchers interested in a specific tissue. Other related data—for example, raw or normalized gene expression levels—can be incorporated into existing views, or in new views designed to best display that particular data. Additionally, to help in identifying colocalization of GWAS and eQTL signals, a helpful new feature could support the uploading of GWAS and eQTLs results, showing both in a comparative region view, to simplify the task of identifying variants with significant associations in both GWAS and eQTLs.

A natural extension of *PheGET*'s functionality is the ability to perform real-time conditional analysis using linkage disequilibrium (LD) information. Given eQTL information within a genomic region for any gene and tissue, alongside LD information between an index variant and other variants within the region, we can calculate the conditional expression associations within the region via linear regression, with the LD value as an additional covariate. This will allow researchers to quantify variant-specific LD effects on other signals within the same locus, making it easier to identify different LD blocks within the same region, as a step towards identifying the biological mechanisms underlying eQTL signals.

5.5 Closing remarks

The development of genomic technology in the last two decades has been nothing short of revolutionary, enabling DNA and RNA studies on a scale near unimaginable back in 2003, when the Human Genome Project released the finished draft of the first sequenced human genome at a cost of $2.7 billion. Seventeen years later, the sequencing of a whole genome costs about $1,000 per sample. Genetic studies have gone from dozens of samples to over a million. We have gone from studying thousands of variants to hundreds of millions. In addition to DNA extracted from blood or cell culture, we can now perform transcriptome-wide RNA sequencing on multiple tissues, and even for single cells. By any measure, this is an astonishing amount of progress in a very short amount of time. With the growing size and complexity of genomic data—which promises to continue apace—statistical methods and tools that were once essential now face major challenges in both theoretical and practical performance. The need for better methods for data processing and analysis has become an essential and inextricable part of genomics research.

The work presented in this dissertation addresses a few of these issues, from variant calling quality control in data generation to data interpretation and visualization for variant and expression data in downstream analysis. With the continuing development of genomic technologies, genetics research will be able to explore hypotheses previously thought impossible to test. It is a privilege to contribute to improvements in statistical genetics methods and techniques, in pursuit of our common mission of improving human health by unlocking the mysteries of genomics, one discovery at a time.

5.6 References

Fritsche, L. G., W. Igl, J. N. Bailey, F. Grassmann, S. Sengupta *et al.*, 2016 A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. Nat Genet 48**:** 134-143.

Geerlings, M. J., E. K. de Jong and A. I. den Hollander, 2017 The complement system in age-related macular degeneration: A review of rare genetic variants and implications for personalized treatment. Mol Immunol 84**:** 65-76.

Hendricks, A. E., S. C. Billups, H. N. C. Pike, I. S. Farooqi, E. Zeggini *et al.*, 2018 ProxECAT: Proxy External Controls Association Test. A new case-control gene region association test using allele frequencies from public controls. PLoS Genet 14**:** e1007591.

Jinek, M., K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna *et al.*, 2012 A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science 337**:** 816-821.

Lee, S., S. Kim and C. Fuchsberger, 2017 Improving power for rare-variant tests by integrating external controls. Genet Epidemiol 41**:** 610-619.

MacArthur, D. G., S. Balasubramanian, A. Frankish, N. Huang, J. Morris *et al.*, 2012 A systematic survey of loss-of-function variants in human protein-coding genes. Science 335**:** 823-828.

Menon, M., S. Mohammadi, J. Davila-Velderrain, B. A. Goods, T. D. Cadwell *et al.*, 2019 Single-cell transcriptomic atlas of the human retina identifies cell types associated with age-related macular degeneration. Nat Commun 10**:** 4902.

Ratnapriya, R., I. E. Acar, M. J. Geerlings, K. Branham, A. Kwong *et al.*, 2020 Family-based exome sequencing identifies rare coding variants in age-related macular degeneration. Hum Mol Genet.

Strunz, T., S. Lauwen, C. Kiel, A. M. D. G. C. International, A. D. Hollander *et al.*, 2020 A transcriptome-wide association study based on 27 tissues identifies 106 genes potentially relevant for disease pathology in age-related macular degeneration. Sci Rep 10**:** 1584.

Wen, X., R. Pique-Regi and F. Luca, 2017 Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. PLoS Genet 13**:** e1006646.