Statistical Methods for Aggregation of Sequence Data and Multiple Testing Correction in Common and Rare Variant Analysis

by

Zhongsheng Chen

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

Doctoral Committee:

Professor Michael Boehnke, Co-Chair
Professor Bhramar Mukherjee, Co-Chair
Research Professor Laura Scott
Associate Professor Xiaoquan Wen
Associate Professor Cristen Willer
Associate Professor Xiang Zhou

Zhongsheng Chen

zhongshc@umich.edu

ORCID iD: 0000-0002-0828-2044

DEDICATION

To my wife who has supported me throughout this journey.

ACKOWLEDGMENTS

I’d like to thank my committee members for their guidance during my dissertation. I am especially grateful for my faculty advisors who has provided me with support and encouragements over the years.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Over the last fifteen years, there have been substantial improvements in how we study the association between trait and genetic variations in the human genome. Genome-wide association studies (GWAS) now routinely test millions of variants in hundreds of thousands of individuals and the advance of genome sequencing technology allows us to examine the role of genetic variants across the full allele-frequency spectrum. However, with these changes come new challenges in analyzing and interpreting genetic results. In this dissertation, we present methods to aggregate sequence data and identify significant associations in common and rare variant analysis.

In chapter two, we compare two strategies to aggregate sequence data from multiple studies: joint variant calling of all samples together versus calling each study individually and then combining the results using meta-analysis. Although joint calling is the gold standard, single-study calling can be more appealing due to fewer privacy restrictions and smaller computational burden. We use deep- and low-coverage sequence data on 2,250 samples from the GoT2D study to compare the two strategies in terms of variant detection sensitivity, genotype accuracy, and association power. We show single-study calling to be a viable alternative to joint calling for deep-coverage sequence data but show them to have noticeable discrepancies in rare variant calling and association results for low-coverage sequence data.

In chapter three, we revisit the common variant P-value significance threshold of 5×10^{-8} and explore the rates of true and false discoveries that can be expected using less restrictive P-value thresholds and three other multiple testing procedures: Benjamini-Hochberg (BH) and Benjamini-Yekutieli (BY) for controlling false discovery rate (FDR), and Bayesian false discovery probability for controlling Bayesian FDR. Using data from the Global Lipids and GIANT consortia, we show for large sample common variant GWAS that using a less stringent P-value threshold of 5×10^{-7} or use of the BH procedure at target FDR threshold of 5% substantially increases the number of true positive discoveries while only modestly increasing false positive discoveries compared with the 5×10^{-8} threshold. The latter threshold remains appropriate for modest-sized studies or for resource-intensive follow-ups such as constructing animal models where a stringently curated list of significant loci is desired from GWAS.

In the chapter four, we propose a Bayesian method for multiple testing correction in rare variant studies that calculates the posterior probabilities using an approximation of the Bayes factor and estimates prior parameters from summary statistics using an Expectation-Maximization algorithm. Using simulations analyses of ~400,000 individuals and ~107 million variants from the TOPMed-imputed UK Biobank study, we show that our Bayesian method discovers more true positive loci than P-value-based methods such as the P-value threshold, BH, and BY procedures at equivalent false positive rates. In addition, we show that the Bayesian method controls empirical FDR among discovered loci. Finally, we estimate the genome-wide significant P-value threshold for testing ~107 million variants from the TOPMed imputation reference panel to be 1×10^{-9}.

Chapter 1

Introduction

1.1 From then to now

We have come a long way in understanding our own genetic makeup. What started thirty years ago as an ambitious project to completely sequence a single human genome (International Human Genome Sequencing Consortium, 2001) has now evolved into efforts such as the UK Biobank (UKBB) and the Trans-Omics for Precision Medicine (TOPMed) programs that routinely sequence tens of thousands of individuals with unprecedented accuracy (Hout et al., 2019; Taliun et al., 2019). The goal, however, remains the same: to use our understanding of the human genome to improve the lives of people worldwide.

1.2 Understanding impact of genetic variations

It is not enough to simply identify the DNA bases that constitute our genome, we need to understand how variations in the genome between individuals affect the incidence or severity of common diseases. This process was facilitated by the development of genome-wide association studies (GWAS) that test for links between millions of genetic variants and diseases or traits (Wellcome Trust Case Control Consortium, 2007). However, there are several challenges in GWAS ranging from issues in experimental design to difficulties in interpreting the results (M. I. McCarthy et al., 2008).

1.3 Aggregating data from different studies

Genetic associations with complex diseases often have small effect size which makes it difficult to identify and replicate them in a single GWAS with limited power (Burton et al., 2009). A solution is to increase sample size by combining data from multiple genetic studies using meta-analysis (Zeggini & Ioannidis, 2009). This method has been remarkably successful in increasing the number of discovered loci for diseases such as type 2 diabetes (Fuchsberger et al., 2016; Morris et al., 2012) and obesity (Yengo et al., 2018).

With the remarkable improvement in sequencing technology, it has become possible for studies to conduct whole-genome sequencing of their samples. Just like array-based GWAS, individual sequencing studies face issues of small sample size that can be alleviated by aggregating sequence data across multiple studies. However, unlike array-based meta-analysis where participating cohorts typically perform genotype imputation (Y. Li et al., 2009a) on the same reference panel, sequencing is performed individually by each cohort, often using different technology platforms (Fox et al., 2014) and variant calling pipelines (Jun et al., 2015; McKenna et al., 2010). This raises the question of whether association results from meta-analysis of sequencing studies can achieve the same power as the gold standard joint analysis like it did for array-based meta-analysis (D. Y. Lin & Zeng, 2010).

1.4 Correcting for multiple testing in GWAS

Prior to GWAS, genetic association studies faced issues of irreproducibility in their reported discoveries (Hirschhorn et al., 2002) which could be partly attributed to a lax significance threshold that did not adequately account for the large testing burden explicitly or implicitly present in such studies (Risch & Merikangas, 1996). To address this, the genetics community adopted a stringent P-value threshold of 5×10^{-8} to define genome-wide significant associations by controlling the family-wise error rate at 5% based on the work of several groups (Dudbridge & Gusnanto, 2008a; Pe'er et al., 2008a). The success of this method in producing tens of thousands of largely reproducible results (Buniello et al., 2019) has likely limited interest in alternative multiple testing methods such as the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995). Such methods based on the false discovery rate (FDR) have the potential to increase true positive discoveries at the cost of a theoretically controlled number of false positives. However, there are complications (Brzyski et al., 2017) with adapting FDR-controlling methods to GWAS, chief among them the need to account for the complex linkage-disequilibrium structures present among the genetic variants included in genetic studies (The International HapMap Consortium, 2007).

1.5 More variants more problems

The development of sequencing technology along with larger imputation reference panels such as those from the 1000 Genomes phase 3 study (The 1000 Genomes Project Consortium et al., 2015), Haplotype Reference Consortium (S. McCarthy et al., 2016), and TOPMed (Taliun et al., 2019) has allowed us to accurately identify and study low-frequency (minor allele frequency [MAF] 0.5-5%) and rare variants (MAF < 0.5%). Exploring the full allele frequency spectrum can potentially explain a greater proportion of the genetic contribution to complex diseases (Seunggeung Lee et al., 2014) than common variants (MAF > 5%) alone and provide us with a better understanding of the genetic factors that influence disease risk and variability in quantitative traits. Of course, more variants in our association studies also means an increased testing burden that must be corrected for to properly control false positives (Pulit et al., 2017). Here, the relatively recent emergence of rare variant studies provides the opportunity to develop alternative multiple testing methods in a field where the P-value threshold is less clearly established. These methods include the aforementioned FDR-controlling procedures (Benjamini & Hochberg, 1995; Benjamini & Yekutieli, 2001a; Brzyski et al., 2017) as well as Bayesian approaches (Wakefield, 2007a; Whittemore, 2007) that use posterior probabilities instead of P-values.

1.6 P-value versus Bayes factor and connection to multiple testing

In GWAS, we typically use P-values to assess the evidence for association between tested variants and a trait of interest. Although a P-value is theoretically simple and well defined -- the probability of obtaining a result at least as extreme as what was observed under the assumption that the null hypothesis is true -- it is often misconstrued (Wasserstein & Lazar, 2016). A frequent misinterpretation is that P-values measure the probability of the null hypothesis given the data (Held & Ott, 2018) but such direct inferences can only be achieved using Bayes factors (BF) which is the ratio of likelihoods of the observed data under the null versus alternative hypotheses. Past works (Berger & Sellke, 1987; Sellke et al., 2001) have shown that P-values can be interpreted in a Bayesian manner through calibration using the function $B(p) = -\mathrm{e}\, p\log(p)$. For P-values $(p) < 1/\mathrm{e}$, this function acts as a lower bound on the BF and can further be used to calculate the posterior probability of the null hypothesis using the equation $(1 + B(p)^{-1})^{-1}$ under the assumption that the prior probability of the null hypothesis is equal to $1/2$.

For multiple testing in GWAS, P-value-based methods such as the P-value threshold or the Benjamini-Hochberg procedure are the easiest to use because most publicly available association results are presented in terms of P-values. Calculating the BF, and subsequently the posterior probability of the null hypothesis given the data, requires access to individual-level genotype data that may not be publicly shared due to privacy and other concerns (Paltoo et al., 2014). As an alternative, Johnson (2005) proposed to calculate the BF using the observed test statistics instead of the original data. Wakefield (2007a) proposed a version of this test-based BF for GWAS:

$$ABF(Z, W) = \sqrt{\frac{V + W}{V}} \exp\left[-\frac{Z^2}{2}\left(\frac{W}{V + W}\right)\right]$$

which can be calculated using just the test statistic (Z), testing variance (V), and the prior variance (W) of variant effect. Although this approximate BF (ABF) is calculated using the P-value (from the test statistic Z), these two measures do not provide the same ranking of tested variants and thus will declare different sets of variants to be significantly associated with the trait. However, these two approaches can be reconciled using an alternative formulation of the ABF (Wakefield, 2009) that eliminates the prior and testing variance:

$$ABF(Z) = \sqrt{1 + K} \exp\left[-\frac{Z^2}{2}\left(\frac{K}{1 + K}\right)\right]$$

where $K = W/V$ and does not depend on the data. This version of the ABF only depends on the data through Z^2 and produces the same ranking of tested variants as P-values.

1.7 Overview

In this dissertation, I develop statistical methods to address the challenges in analyzing and interpreting genetic results. A list of the datasets that used throughout this dissertation and their public availability is shown in Table 1.1. In Chapter 2, I present a protocol to aggregate sequence data from multiple studies considering sequence coverage and power of subsequent association analyses. In Chapter 3, I investigate the performance of less restrictive P-value thresholds and alternative multiple testing methods in common variant GWAS and present a procedure for calculating their true and false positive rates in empirical studies by leveraging the sequential nature of meta-analyses. In Chapter 4, I propose a Bayesian method for multiple testing correction that increases true positive discovery in rare variant studies compared with the P-value threshold while still maintaining control of FDR.

Table 1.1: Description of datasets

Chapter	Description	Sample size	Availability
2	Low-coverage whole-genome sequence data from GoT2D	2,250	European Genome-phenome Archive and dbGAP
2	Deep-coverage whole-exome sequence data from GoT2D	2,250	European Genome-phenome Archive and dbGAP
3	Meta-analysis of lipid traits by GLGC (2009)	19,840	http://csg.sph.umich.edu/willer/public/lipids2008/
3	Meta-analysis of lipid traits by GLGC (2013)	188,577	http://csg.sph.umich.edu/willer/public/lipids2013/
3	Meta-analysis of height from GIANT (2010)	133,653	https://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files
3	Meta-analysis of BMI from GIANT (2010)	123,865	https://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files
3	Meta-analysis of height and BMI from GIANT (2018)	694,529	https://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files
4	Whole-genome sequence data from UKBB	487,409	http://biobank.ctsu.ox.ac.uk/crystal/label.cgi?id=263
4	Whole-exome sequence data from UKBB	49,960	http://biobank.ctsu.ox.ac.uk/crystal/label.cgi?id=263

Chapter 2

Combining Sequence Data from Multiple Studies: Impact of Analysis Strategies on Rare Variant Calling and Association Results

2.1 Introduction

Genome-wide association studies (GWAS) based on genotype arrays have identified thousands of common (minor allele frequency [MAF] >5%) genetic variants associated with a wide range of human diseases and traits (Hindorff et al., 2012). However, these common variants comprise only 10% of the ~84 million variant sites discovered in the human genome by the 1000 Genomes Project (The 1000 Genomes Project Consortium et al., 2015) with the rest being low-frequency (MAF 0.5-5%; ~14%) and rare (MAF < 0.5%; ~76%) variants that are less well captured by genotype arrays and subsequent genotype imputation (Zuk et al., 2014). With the advance of genome sequencing technology, we can now directly study the role of variants across the full allele-frequency spectrum. Although sequencing studies to date have reaffirmed and expanded on the common variant associations of array-based GWAS, the modest sample sizes of most sequencing studies to date have limited the discovery of rare and low-frequency variant associations (Auer et al., 2016; Fuchsberger et al., 2016).

To increase sample size, researchers often aggregate sequence data across multiple studies. To combine sequence data across studies, the gold standard strategy is to jointly call all samples together (Auer et al., 2016). This joint calling strategy increases the quality of variant calls and minimizes batch effects such as those due to different sequencing centers or platforms (Auer et al., 2016). However, joint calling for sequence data can be difficult to implement due to restrictions on data sharing (Jiang et al., 2014; Paltoo et al., 2014) and the potentially heavy computation burden (Exome Aggregation Consortium et al., 2016). An alternative strategy that adheres to privacy rules and mitigates computing load is single-study calling (Okada et al., 2018) which variants are identified and genotypes called separately within each study and then combined through meta-analysis of study-level association statistics or joint analysis of pooled individual-level data (i.e. mega-analysis). Although single-study calling is easier to implement than the gold standard joint calling, there is a need to quantify the difference in calling results between these two strategies and assess how it affects downstream association analysis.

Past research has shown that meta-analysis of study-level association results is as statistically efficient as joint analysis of individual-level data for combining common-variant GWAS (D. Y. Lin & Zeng, 2010). More recent research has extended methods for meta-analysis to sequencing studies for rare variants (Z.-Z. Tang & Lin, 2015). However, this research only analyzes the relative power of joint and meta-analysis under a single-study calling strategy and does not consider the impact of joint calling on association results. In addition, sequencing studies often differ in sequencing coverage depending on project needs and goals. For example, deep-coverage sequencing results in improved genotyping accuracy, particularly for rare variants (Seunggeung Lee et al., 2014; Chao Xu et al., 2017), while low-coverage sequencing results in more sequenced samples at the same cost (Y. Li et al., 2011). Thus, there is also a need to compare rare variant association tests for joint and single-study calling under different sequencing coverage.

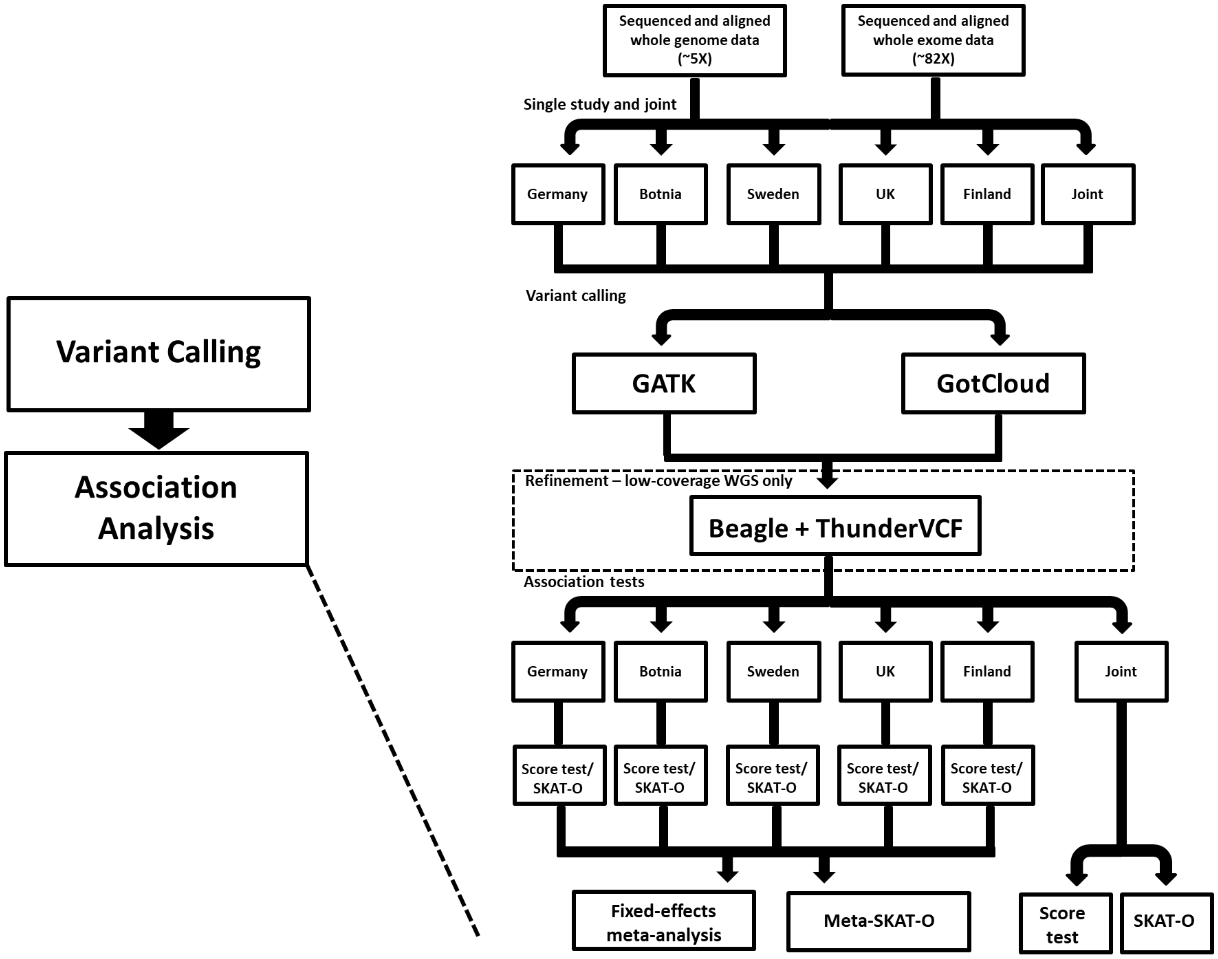
In this paper, we aim to quantify the difference between the gold standard joint calling and the alternative single-study calling strategies and assess their impact on association testing of rare single nucleotide variants (SNVs) in deep and low-coverage sequence data. Specifically, we compare variant detection and genotyping accuracy for joint and single-study callsets on deep-coverage whole exome sequence (WES) and low-coverage whole genome sequence (WGS) dataset from the Genetics of Type 2 Diabetes (GoT2D) study (Fuchsberger et al., 2016) using the GotCloud variant calling pipelines (Jun et al., 2015) at default settings. Then for each data type, we compare single-variant and gene-based association test results for rare SNVs between three types of joint and single-study strategies: 1) joint calling with joint analysis, 2) single-study calling with meta-analysis, and 3) single-study calling with mega-analysis.

2.2 Methods

2.2.1 Data description

We analyzed data on 2,250 individuals from the GoT2D study (Fuchsberger et al., 2016) for whom deep-coverage whole exome sequence (mean depth 82X), low-coverage whole genome sequence (mean depth 5X), and Illumina HumanOmni 2.5M array data were all available. Study participants came from five geographical regions: (1) Augsburg, Germany (n=193; KORA study), (2) the Botnia region of western Finland (n=303; DGI study), (3) Sweden (n=391; DGI study), (4) the United Kingdom (n=473; UKT2D study), and (5) Finland (n=890; FUSION study). For clarity, we will refer to the sample of 2,250 individuals as the "joint" cohort and the five subsets as the "single-study" cohorts (Figure 2.1).

Figure 2.1: Workflow for variant calling and association analysis.



Sequencing and alignment are described in Fuchsberger et al., 2016. Haplotype-based refinement was only applied to low-coverage whole genome sequence data.

2.2.2 DNA sample preparation and sequencing

DNA samples were processed at the Broad Institute (FUSION and DGI), Wellcome Trust Centre for Human Genetics (UKT2D), and Helmholtz Zentrum München (KORA). DNA samples were genome and exome sequenced using the Illumina GAII or HiSeq 2000 sequencers. Sequence data were aligned to human reference genome version 19 (hg19) using Picard (DePristo et al., 2011) and BWA (H. Li & Durbin, 2009). Further details on data generation, processing, and quality control can be found in Fuchsberger et al. (2016).

Processed and filtered sequence reads for the joint and single-study cohorts were analyzed by the GotCloud and GATK (McKenna et al., 2010; Van der Auwera et al., 2013) variant calling pipelines according to the best practice workflows recommended by their developers at default settings. We restricted our analyses to chromosome 2 (~8% of the human genome) to reduce computational burden.

2.2.3 Whole-genome and exome sequence data processing: GotCloud and GATK pipeline

We called SNVs with GotCloud at default settings using processed BAM files (Figure 2.1). We used SAMtools pileup and glfFlex to generate genotype likelihoods for all samples in 5 Mb chromosomal segments. We then used a support vector machine classifier to filter out likely false-positive variant sites (Jun et al., 2015).

Adhering to the recommended GATK workflow, we "hard called" every variable site in each sample for the number of non-reference alleles (0, 1, or 2) using HaplotypeCaller in GVCF mode. To parallelize this step, we divided chromosome 2 into 5 Mb segments with 100 bp overlap and simultaneously carried out hard-calls within each segment. We merged intermediate genomic VCF (gVCF) files from each sample into batches of 100 samples with CombineGVCFs and then jointly genotyped them with GenotypeGVCFs. We used the GATK CatVariants tool to concatenate variant sets from all genomic regions to form a combined callset. We identified a set of high-quality variant calls from the raw variant callset using the Variant Quality Score Recalibration (VQSR) method which applies machine learning algorithms to score each variant call and filter them at a desired level of sensitivity. We used GATK VariantRecalibrator and ApplyRecalibration to filter the raw variant callset at the recommended tranche threshold of 99.9% which provides high sensitivity while maintaining a reasonable level of specificity. Finally, we removed indels from the filtered variant callset in keeping with our settings for the GotCloud pipeline and to focus on SNVs in subsequent analyses.

We used haplotype-based refinement to improve genotype and haplotype quality for whole genome genotype calls from both pipelines (Figure 2.1). Specifically, we used Beagle (Browning & Yu, 2009) to phase the genotype data in chunks of 10,000 SNVs with 1,000 SNVs overlaps and refined the phased sequences using Thunder (Jun et al., 2015) with 300 states.

We ran whole exome sequence reads through the GotCloud and GATK discovery pipelines under the same settings as the whole-genome data. We did not apply any refinement steps to the exome calls, consistent with standard practice for both pipelines for deep-coverage sequence data.

The final dataset for each of the four combinations of sequencing coverage (genome and exome) and pipeline (GotCloud and GATK) consists of a joint callset for all 2,250 samples, five separate single-study callsets for the geographically subdivided cohorts, and a union callset which merges the five single-study callsets. Since comparing the joint callset to five single-study callsets individually is difficult because detection of rare SNVs is heavily dependent on sample size and the results would be potentially skewed by the considerable sample size differences between cohorts, we use the union callset as an overall representation of single-study calling to provide a more apt comparison with the joint callset. For the union callset, we set genotype calls for SNVs not found in one or more of the single-study callset(s) as missing.

2.2.4 Non-reference genotype accuracy

For both pipelines, we assessed the accuracy of whole genome calls by comparing the Thunder-refined non-reference genotypes against a set of 192,322 variants of highly accurate ("high-confidence") genotypes determined through joint statistical analysis of deep-coverage (~82X) exome sequence and Illumina HumanOmni 2.5 array data in the GoT2D whole genome sequencing study (Fuchsberger et al., 2016). We assessed the accuracy of exome calls by comparing unrefined non-reference genotypes against the set of high-confidence genotypes from Illumina HumanOmni 2.5 array data.

2.2.5 Single-variant association analysis

We evaluated the impact of joint and single-study calling on single-variant association tests by comparing -$\log_{10}$p-values from joint analysis of the joint callset against those from meta-analysis of single-study summary statistics and joint analysis of the union callset (i.e. mega-analysis). In each single-study callset, we used the logistic score test to test for T2D association under an additive genetic model with the top two principal components as covariates (Figure 2.1). For meta-analysis, we combined summary-level results from the single-study callsets with fixed-effects sample-size weighted meta-analysis using METAL (Willer et al., 2010) and with trans-ethnic meta-analysis using MR-MEGA software (Mägi et al., 2017).

2.2.6 Gene-based association analysis

We used SKAT-O to test for association with multiple rare and low-frequency SNVs within coding regions of the genome. We prepared four lists of SNVs ("masks") based on MAF and functional annotation. For the creation of the masks, we considered a SNV to have MAF < 1% if its MAF in every one of the single-study callsets is <1%. Mask 1 contained SNVs predicted to be protein-truncating, Mask 2 included all SNVs from Mask 1 together with missense SNVs with MAF < 1%, Mask 3 included all SNVs from Mask 1 and those predicted to be deleterious by all five algorithms applied (Polyphen2-HumDiv, PolyPhen2-HumVar, LRT, Mutation Taster, and SIFT), and Mask 4 included all SNVs from Mask 1 and those predicted to be deleterious by at least one algorithm with MAF < 1%.

We performed SKAT-O (Seunggeun Lee et al., 2012) analysis on the four masks separately within each single-study callset (Figure 2.1). We combined SKAT-O results from each single-study callset using Meta-SKAT-O test in the MetaSKAT R package (Seunggeun Lee et al., 2013) once assuming homogeneous genetic effects across single-study cohorts and again assuming heterogeneous genetic effects.

2.3 Results

2.3.1 Overview

We evaluated the utility of single-study calling as an alternative to the gold standard joint calling by comparing these methods in terms of variant detection, genotype accuracy, and impact on power of association tests for different sequencing coverage. For our analysis (restricted to chromosome 2 due to computational burden), we focus on the gold standard *joint callset*, which are calls from analyzing all 2,250 samples together (the "joint" cohort), the five *single-study callsets,* which are calls from the five geographically subdivided cohorts (the "single-study" cohorts: Germany, Botnia, Sweden, UK, Finland), and the *union callset*, which pools calls from the five single-study callsets. There are 25,689 deep-coverage WES SNVs and 2,101,401 (15,344 when restricted to coding regions) low-coverage WGS SNVs in the joint callset and 26,364 deep-coverage WES SNVs and 2,249,181 (16,457) low-coverage WGS SNVs in the union callset. We present only GotCloud results as we found choice of software pipelines (GotCloud or GATK) to have no meaningful impact on variant calling and association results.

2.3.2 Union callset

The union callset pools calling results from the five single-study cohorts by merging their SNV calls. For SNV sites found in only a subset of the studies, we assign missing genotypes for studies in which the SNV site was not called. Using the union callset, we examine the overlap in variant detection between single-study cohorts. For deep-coverage data, 78% of all rare SNVs detected by single-study calling (i.e. those in the union callset) are "study specific" (Table 2.1), meaning they were found in only one of the single-study callsets and missing in all others, compared with 1.2% of low-frequency SNVs and 0.05% of common SNVs (Table 2.1). Conversely, only 2.3% of rare SNVs in the union callset are found in all five studies (Table 2.1) compared with 80% of low-frequency and 99% of common SNVs (Table 2.1). Similar numbers are seen for low-coverage data (restricted to coding regions) (Table 2.1). Overall, there are three possible reasons for a missing SNV site in a study: 1) the SNV was monomorphic in the study sample; 2) the variant caller did not have confidence to declare the SNV site; or 3) the SNV site was identified but removed by quality control as likely false-positive. However, for single-study calling, we are unable to differentiate between the three types of missingness because of privacy restrictions for individual-level data such as BAM files and calling results.

Table 2.1: Overlap in variant detection for the union callset

Data type	**Variants detected by only one study**	**Variants detected by 2 to 4 studies**	**Variants detected by all 5 studies**
Deep-coverage			
Rare (MAF <0.5%)	17,128 (78.0%)	4,316 (20%)	507 (2.3%)
Low-frequency (MAF 0.5-5%)	28 (1.2%)	435 (19%)	1,873 (80%)
Common (MAF >5%)	1 (0.05%)	26 (1.3%)	2,050 (99%)
Low-coverage (coding regions)			
Rare (MAF <0.5%)	9,262 (77%)	2,563 (21%)	160 (1.4%)
Low-frequency (MAF 0.5-5%)	38 (1.6%)	890 (38%)	1,432 (61%)
Common (MAF >5%)	5 (0.24%)	123 (5.8%)	1,984 (94%)

Note. The union callset pools variant calling results from the five single-study cohorts. Numbers in table refers to SNVs from chromosome 2 in deep-coverage (~82X) exome sequence data and low-coverage (~5X) genome sequence data restricted to coding regions.

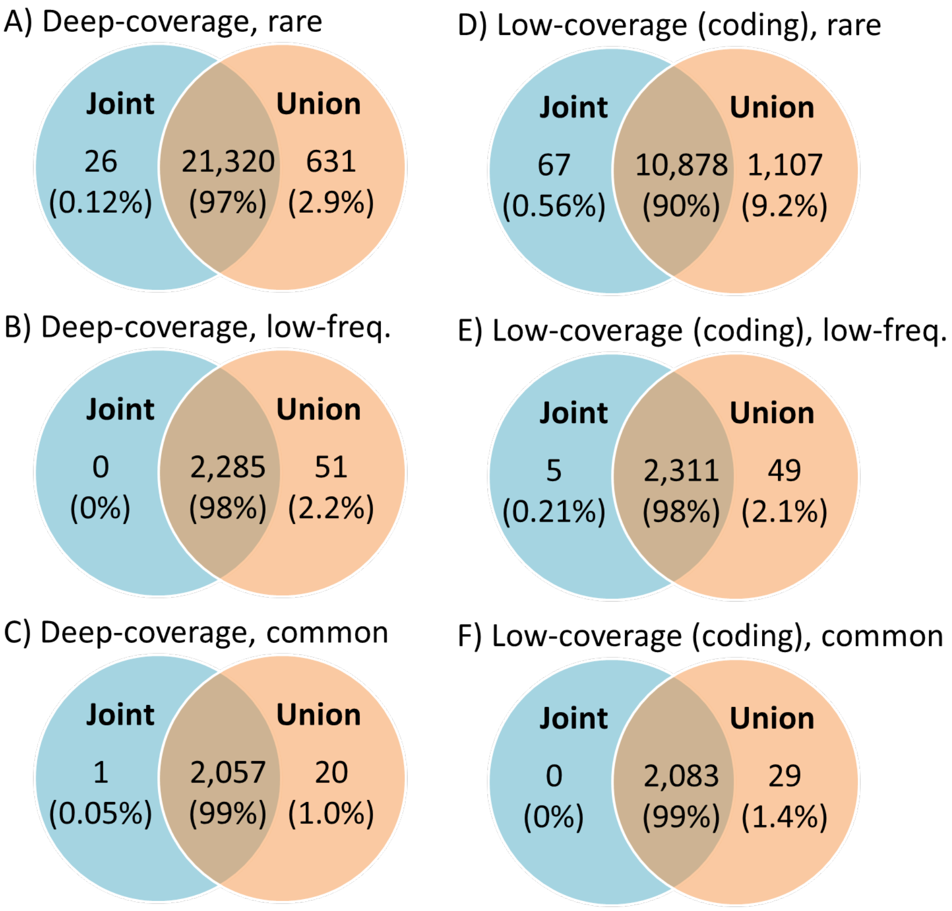
2.3.3 Variant detection: callset size

We evaluated variant detection for joint and single-study strategies by comparing the joint and union callsets across a range of MAFs. For low-frequency and common SNVs in both deep-coverage exome and low-coverage genome (restricted to coding regions) sequence data, there is almost complete overlap between the joint and union callsets (Figure 2.2C-F). However, for rare SNVs, there are noticeable discrepancies between the two callsets as described below.

The overwhelming majority of rare SNVs detected in deep-coverage data are found in both the joint and union callsets (97% of all rare SNVs) with the remaining SNVs found exclusively in the joint (0.1%) and union (2.9%) callsets (Figure 2.2A). Contrary to expectations, the union callset is larger than the joint callset, mainly due to inconsistencies in variant filtering. Of the 631 rare SNVs exclusive to the union callset, 540 of them were filtered out during joint calling and excluded from the final joint callset. SNVs in joint calling go through variant filters once whereas SNVs in single-study calling have one chance per study to pass filters and be included in the union callset. In this scenario, a lack of consistent variant filtering between joint and single-study calling can lead to the differences seen here.

For rare SNVs in low-coverage data (Figure 2.2B), we observed a similar pattern of variant detection as for deep-coverage data. However, inconsistencies in variant filtering only accounts for a small fraction of differences between the joint and union callsets. Only 128 of the 1,107 rare SNVs exclusive to the union callset were filtered out during joint calling.

Figure 2.2: Comparison of variant detection between joint and single study calling strategies



Comparison for rare (MAF<0.5%), low-frequency (MAF 0.5-5%), and common (MAF>5%) SNVs in deep-coverage (~82X) exome sequence data and low-coverage (~5X) genome sequence data restricted to coding regions.

2.3.4 Variant detection: genotype calls

In addition to comparing the number of SNVs detected by joint and single-study calling, we also compared the genotype calls made by the two strategies at different sequencing coverage. We show in Tables 2.2 and 2.3 the comparison of genotype calls between joint and the single-study calling for 9,096 rare SNVs found in the joint and union callsets from deep-coverage exome as well as from low-coverage genome (restricted to coding regions) sequence data. Genotype comparisons for 2,127 low-frequency and 2,027 common SNVs are shown in Supplementary Tables 2.1-2.4. Excluding missing calls, overall genotype discordance between joint and single-study calling is lower in deep-coverage data than in low-coverage data. Furthermore, for rare SNVs, 64% of all genotype calls from single-study calling in deep-coverage data (Table 2.2) are missing compared with 70% for low-coverage data (Table 2.3). Breaking down rare SNVs further by minor allele count (MAC), we observe this missingness to be a function of MAC in both types of sequencing data with the rarest categories most affected. In deep-coverage data, we can attribute almost all missing calls for rare SNVs to monomorphic SNVs in the single-study cohort(s) since 13,093,060 of the 13,093,128 missing single-study calls were called as homozygous reference by joint calling (Table 2.2). Using the GATK pipeline, it is possible to identify monomorphic SNVs in gVCFs and assign homozygous reference genotypes to the 13,093,060 missing calls. However, we were unable to do this for the GotCloud pipeline since it does not support gVCFs. In low-coverage data, 6,365 of 14,246,613 missing single-study calls were called as non-reference by joint calling (Table 2.3) compared with 68 non-reference calls for deep-coverage data (Table 2.2). Since rare SNVs naturally have low allele counts to begin with, any small change to their overall allele counts will have a noticeable impact on association testing and other downstream analyses. Finally, the missingness appears to be mostly localized to rare SNVs as we observe only a slight number of missing genotype calls in low-frequency SNVs (4.3% in deep-coverage data, 9.2% in low-coverage data; Supplementary Tables 2.1 and 2.2) and a negligible number in common SNVs (0.21% and 0.78%; Supplementary Tables 2.3 and 2.4).

Table 2.2: Comparison of genotype calls for rare SNVs from deep-coverage exome sequence data

Single-study variant calling (union callset)	**Joint variant calling (joint callset)**				
	Missing	**Homozygous reference**	**Heterozygous**	**Homozygous alternate**	**Total**
Missing	0	13,093,060 (64%)	68 (0.00033%)	0	13,093,128 (64%)
Hom. ref.	0	**7,135,459 (35%)**	9 (0.000044%)	0	7,135,468 (35%)
Heterozygous	0	31 (0.00015%)	**25,862 (0.13%)**	0	25,893 (0.13%)
Hom. alt.	0	0	4 (0.000020%)	**211,507 (1.0%)**	211,511 (1.0%)
Total	0	20,228,550 (99%)	25,943 (0.13%)	211,507 (1.0%)	20,466,000 (100%)

Note. Genotype calls from joint (horizontal axis) and single-study (vertical axis) calling strategies for 9,096 rare (MAF <0.5%) SNVs from chromosome 2 in deep-coverage (~82X) exome sequence data. Concordant calls between the two strategies are highlighted in bold.

Table 2.3: Comparison of genotype calls for rare SNVs from low-coverage genome sequence data (coding regions)

Single-study variant calling (union callset)	**Joint variant calling (joint callset)**				
	Missing	**Homozygous reference**	**Heterozygous**	**Homozygous alternate**	**Total**
Missing	0	14,240,248 (70%)	5,966 (0.029%)	399 (0.002%)	14,246,613 (70%)
Hom. ref.	0	**5,981,638 (29%)**	1,855 (0.009%)	2 (0.000010%)	5,983,495 (29%)
Heterozygous	0	3,687 (0.02%)	**21,073 (0.10%)**	99 (0.00048%)	24,859 (0.12%)
Hom. alt.	0	0	37 (0.00018%)	**210,996 (1.0%)**	211,033 (1.0%)
Total	0	20,225,573 (99%)	28,931 (0.14%)	211,496 (1.0%)	20,466,000 (100%)

Note. Genotype calls from joint (horizontal axis) and single-study (vertical axis) calling strategies for 9,096 rare (MAF <0.5%) SNVs from chromosome 2 in low-coverage (~5X) genome sequence data restricted to coding regions. Concordant calls between the two strategies are highlighted in bold.

2.3.5 Genotype concordance

We assessed non-reference genotype accuracy (hereafter referred to as "genotype concordance") of joint and single-study calling in deep-coverage exome sequence data by comparing non-reference calls for SNVs found in both the joint and union callsets against a "truth" set of high confidence genotypes from Illumina HumanOmni 2.5 array data (Fuchsberger et al., 2016). The joint and union callsets have nearly identical genotype concordance with the truth set for SNVs of all MAFs and negligible differences in raw counts (Table 2.4).

Next, we assessed genotype concordance for SNVs in low-coverage genome sequence data (not restricted to coding regions to preserve a meaningful number of comparisons) by comparing against high confidence genotypes from Illumina HumanOmni 2.5 array data and/or from deep (~82X) exome sequence in the GoT2D integrated panel (Fuchsberger et al., 2016). The joint callset correctly calls 0.4% more genotypes than the union callset for rare SNVs, 0.5% more for low-frequency SNVs, and 0.2% more for common SNVs (Table 2.4). Compared with deep-coverage data, here we observe a larger difference in genotype concordance with the truth set between the joint and union callsets. For example, the joint callset calls 13,322 more genotypes correctly (out of 3,575,402 total comparisons) than the union callset for rare SNVs in low-coverage data while it only calls 1 more genotype correctly (out of 91,756) for rare SNVs in deep-coverage data. As expected, the improvements to calling accuracy offered by larger sample sizes in the joint strategy are more pronounced when the average read coverage is low.

Table 2.4: Non-reference genotype accuracy for joint and single-study calling strategies

Data type	**Genotype concordance for joint callset**	**Genotype concordance for union callset**
Deep-coverage		
Rare (MAF <0.5%)	99.7% (91,457/91,756)	99.7% (91,456/91,756)
Low-frequency (MAF 0.5-5%)	99.3% (171,939/173,131)	99.3% (171,930/173,131)
Common (MAF >5%)	99.3% (1,712,741/1,72,4873)	99.2% (1,711,385/1,724,873)
Low-coverage (all regions)		
Rare (MAF <0.5%)	99.7% (3,563,500/3,575,402)	99.3% (3,550,178/3,575,402)
Low-frequency (MAF 0.5-5%)	99.6% (6,837,310/6,866,584)	99.1% (6,807,530/6,866,584)
Common (MAF >5%)	99.6% (112,966,946/113,401,131)	99.4% (112,694,329/113,401,131)

Note. Genotype concordance for joint and single-study calling strategies in deep-coverage (~82X) exome and low-coverage (~5X) genome sequence data. The "truth" set of high confidence genotypes being compared against comes from Illumina HumanOmni 2.5 array data and deep exome sequence in the GoT2D integrated panel. Raw genotype counts are displayed in parentheses.

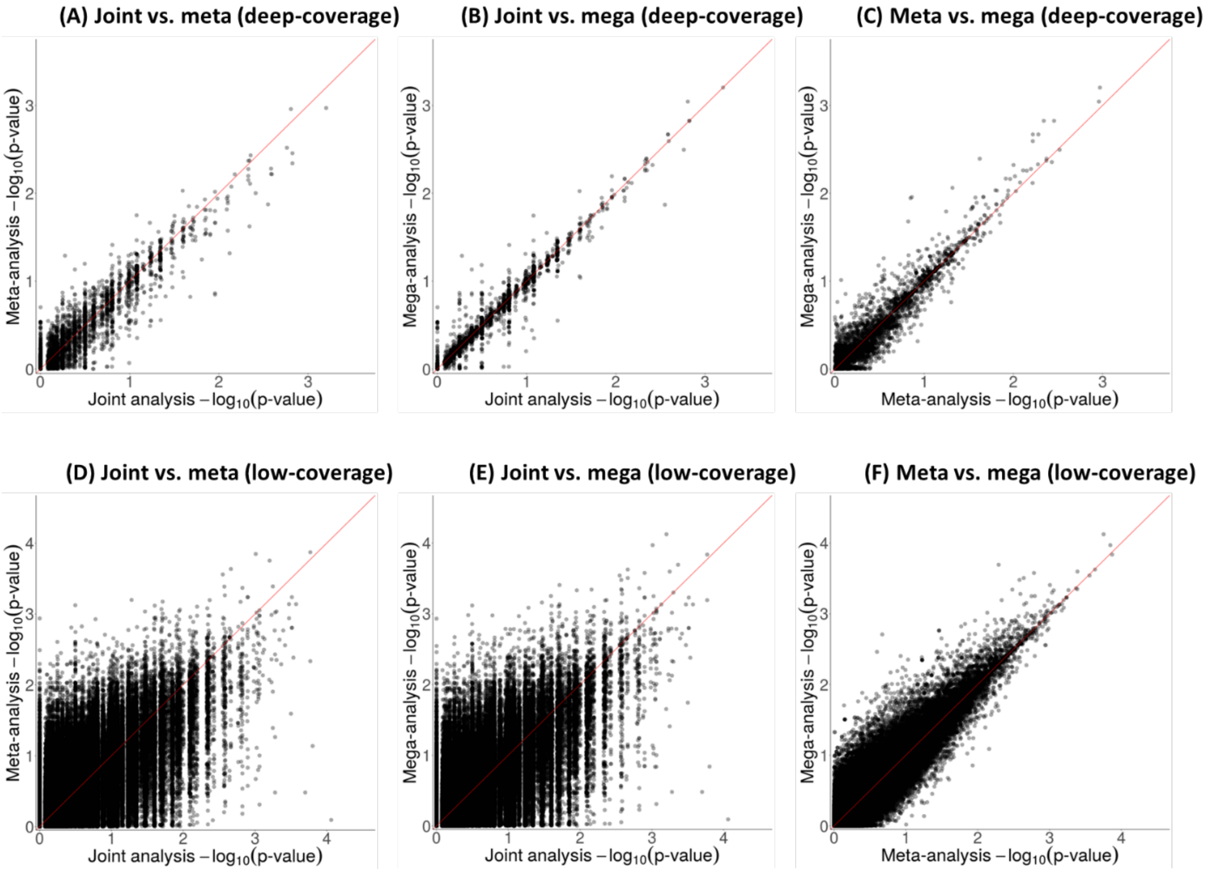
2.3.6 Effect of GC bias on genotype concordance

It is a well-known that sequencing read coverage tends to be lower in high GC-content regions. To investigate the effect of this GC bias on joint and single-study calling, we compared genotype concordance between the joint and union callset in regions of low GC-content (<60% of base pairs are GC) and in regions of high GC-content (≥60%) in chromosome 2. In low GC-content regions, we observe similar genotype concordance between the joint and union callset in both deep- and low-coverage sequence data (Supplementary Table 2.5). In high GC-content regions, we observe similar genotype concordance between the two callsets in deep-coverage data but notice larger differences in low-coverage data where the joint callset correctly calls 0.7% more genotypes than the union callset for rare and low-frequency SNVs (Supplementary Table 2.6). The performance of the two calling strategies in high GC-content regions are nearly equal in deep-coverage data but single-study calling can be slightly less accurate than joint calling in low-coverage data.

2.3.7 Association analysis

Overall, we observe similar p-values between joint analysis of the joint callset, fixed-effects meta-analysis of single-study summary statistics, and joint analysis of the union callset (mega-analysis) for rare SNVs in deep-coverage data (Figure 2.3A-C). This is due to almost perfect concordance in genotype calls between joint and single-study calling and the fact that missing variant calls for rare SNVs from single-study calling were almost all called as homozygous reference in the joint callset. However, for low-coverage data, we observe large discrepancies in p-values between joint and meta-analysis (Figure 2.3D) as well as between joint and mega-analysis for rare SNVs (Figure 2.3E). These differences in association results is caused by a combination of lower concordance in genotype calls between the two calling strategies for low-coverage data and an increase in the number of missing single-study calls being called as non-reference in the joint callset. Since both meta-analysis and mega-analysis use single-study calling, their association results are more similar (Figure 2.3F).

Figure 2.3: Comparison of single-variant association test p-values between joint and single study calling strategies for rare SNVs



Comparison for rare (MAF<0.5%) SNVs in (A-C) deep-coverage (~82X) exome sequence data and (D-F) low-coverage (~5X) genome sequence data. *Joint* refers to joint analysis of the joint callset, *meta* refers to meta-analysis of single-study summary statistics, and *mega* refers to joint analysis of the union callset (mega-analysis).

We evaluated association power between joint and single-study calling for gene-based tests by comparing -$\log_{10}$p-values from SKAT-O test of the joint callset versus those from meta-analysis of single-study SKAT-O test results assuming homogeneous genetic effects. For all masks, SKAT-O based joint analysis and Meta-SKAT-O based meta-analysis produce similar p-values (Supplementary Figure 2.3).

2.3.8 Heterogeneity between single-study cohorts

To address possible heterogeneity in genetic effects between our single-study cohorts, we combined single-study summary statistics using a trans-ethnic meta-analysis implemented in MR-MEGA and combined single-study SKAT-O test results using Meta-SKAT-O assuming heterogeneous genetic effects. For single-variant tests, we observe that trans-ethnic meta-analysis had slightly greater power to detect variants whose heterogeneity in genetic effects were correlated with ancestry compared with fixed-effects meta-analysis (Supplementary Figure 2.4). However, none of these variants are close to reaching genome-wide significance (p-value$<5 \times 10^{-8}$) while those that are have more significant p-values under a fixed-effects meta-analysis. For gene-based tests, we observe slight variations in p-values between homogeneous and heterogeneous effect meta-analyses for Masks 1 and 3 but much greater p-value variability for Masks 2 and 4 (Supplementary Figure 2.5).

2.4 Discussion

Although jointly calling all samples together is the gold standard strategy for analyzing rare SNVs in sequencing studies, single-study calling is more appealing due to fewer privacy restrictions and smaller computation burden. In this study, we compared joint and single-study calling in terms of variant detection, non-reference genotype concordance, and their impact on association power as a function of sequencing coverage.

For single-study calling, we found that low overlap in variant detection among single-study cohorts for rare SNVs results in an abundance of “missing” genotype calls where we lose information for variant sites in cohorts where they were not detected. We show that for deep-coverage data, the impact of missing genotype calls on association testing of rare SNVs from single-study calling is minimal because almost all of this missingness is due to monomorphic SNVs, as evident by corresponding homozygous reference calls in the joint callset. However, for low-coverage data, average read depth is low and thus, a portion of the missing genotype calls may be due to lack of coverage at the variant sites (Xu et al., 2017). Indeed, we show that a fraction amount of missing single-study calls for rare SNVs in low-coverage data have corresponding non-reference calls in the joint callset, resulting in lower than expected allele counts and reduced power for association testing of these SNVs. In addition, these missing calls can have a negative impact on gene-based aggregation tests, which will be underpowered if too many variant sites within a gene have missing genotype calls, and genotype-based callbacks, since the majority of loss-of-function SNVs are rare. A possible, but resource-intensive solution is to generate a list of SNV sites based on the union callset and then go back and genotype these sites within each single-study cohort. With parallel computation for each sample and every 5 Mb chromosomal segment, this process takes on average one hour CPU-time per sample per cohort with a maximum memory usage of approximately 0.5 GB to re-call 1 to 1.2 million variants in chromosome 2.

Although the low overlap in variant detection among single-study cohorts for rare SNVs can arise naturally due to sample population differences between cohorts, another contributing factor is the inconsistency of variant calling filters (i.e. false-positive screening). In our analysis, rare SNVs that were filtered out during joint calling may pass filters during calling in some single-study cohorts while being filtered out in others. This increases the possibility of introducing false-positive SNVs to downstream analyses since they only need to pass filters in one of the single-study cohorts to be included in association tests.

2.4.1 Recommendations

For deep-coverage data, single-study calling and either meta-analysis or mega-analysis can be recommended as a viable alternative to joint calling and analysis for rare SNVs based on almost perfect concordance of genotype calls between the two calling strategies, comparable non-reference genotype concordance with an external truth set, and comparable association results. Furthermore, missing genotype calls in single-study calling for deep-coverage data can be assumed to be homozygous reference and attributed to monomorphic variant due to a matching homozygous reference call for their counterparts in the joint callset. When combining many smaller single studies, meta-analysis can be more conservative and less powerful than mega-analysis (Ma et al., 2013).

For low-coverage data or low-coverage regions in deep data, single-study calling cannot be recommended as a viable alternative to joint calling for rare SNVs. Discordance in genotype calls between the two calling strategies is approximately 150 times higher than that in deep-coverage data (0.09% versus 0.0006%) and combined with a sizable number of genotype calls in single-study calling being missing due to lack of coverage at variant sites, we observe large discrepancies in association results between the two calling strategies.

In general, for studying low-frequency and common SNVs, single-study calling can be used as an alternative to joint calling in both deep-coverage and low-coverage data (Supplementary Figures 2.1 and 2.2). The only exception is for studying low-frequency SNVs in low-coverage data (Supplementary Figure 2.1D-F) where there remain noticeable discrepancies in association results between joint and meta/mega-analysis, although less than that seen for rare SNVs in low-coverage data.

2.4.2 Comparison with GATK pipeline

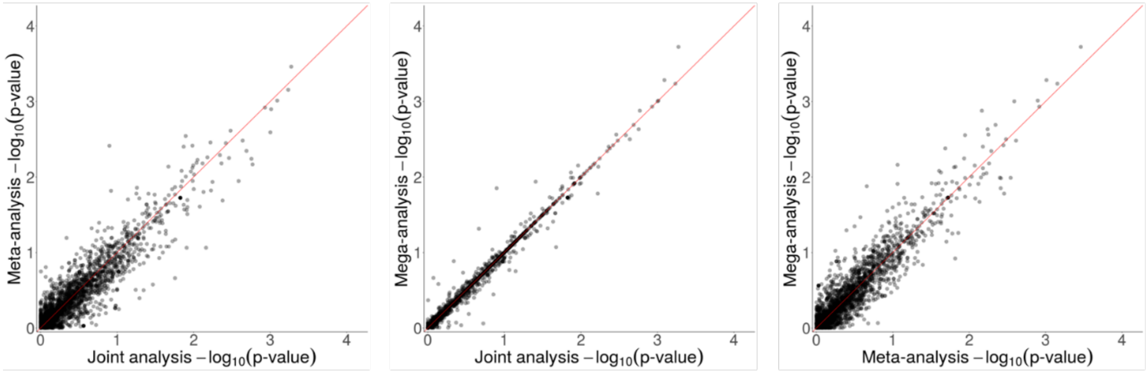
In addition to the GotCloud pipeline, we ran our analyses with the widely used GATK pipeline at default settings. Choice of software pipeline had a limited impact on variant detection and genotype accuracy (Supplementary Table 2.7) with little to no impact on association results (data not shown). There is more overlap in detected SNVs between joint and single-study calling for the GotCloud pipeline in deep-coverage data and vice versa for the GATK pipeline in low-coverage data. The GotCloud pipeline was slightly more accurate in calling common and low-frequency SNVs; however, on average this difference amounts to less than 1.5% more correctly called non-reference genotypes.

2.4.3 Summary

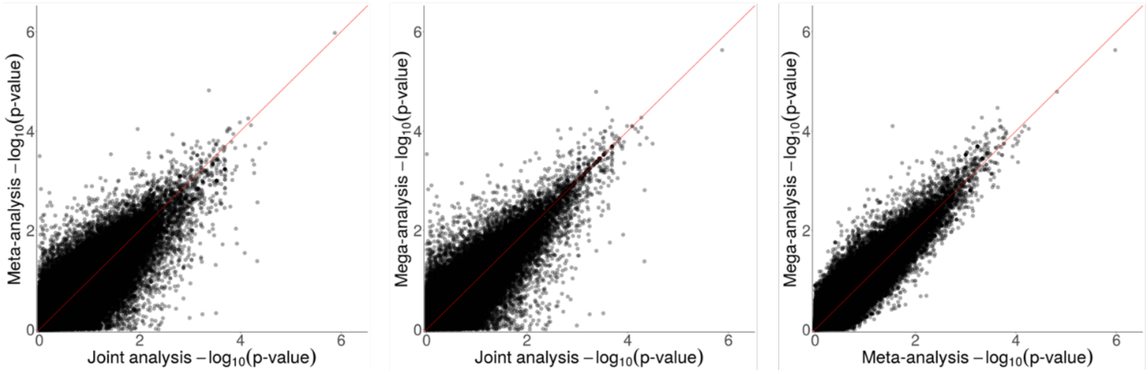
We show single-study calling to be a viable alternative to joint calling for deep-coverage sequence data but show them to have noticeable discrepancies in rare variant calling and association results for low-coverage sequence data.

2.5 Supplementary figures

Supplementary Figure 2.1: Comparison of single-variant association test p-values between joint and single study calling strategies for low-frequency SNVs



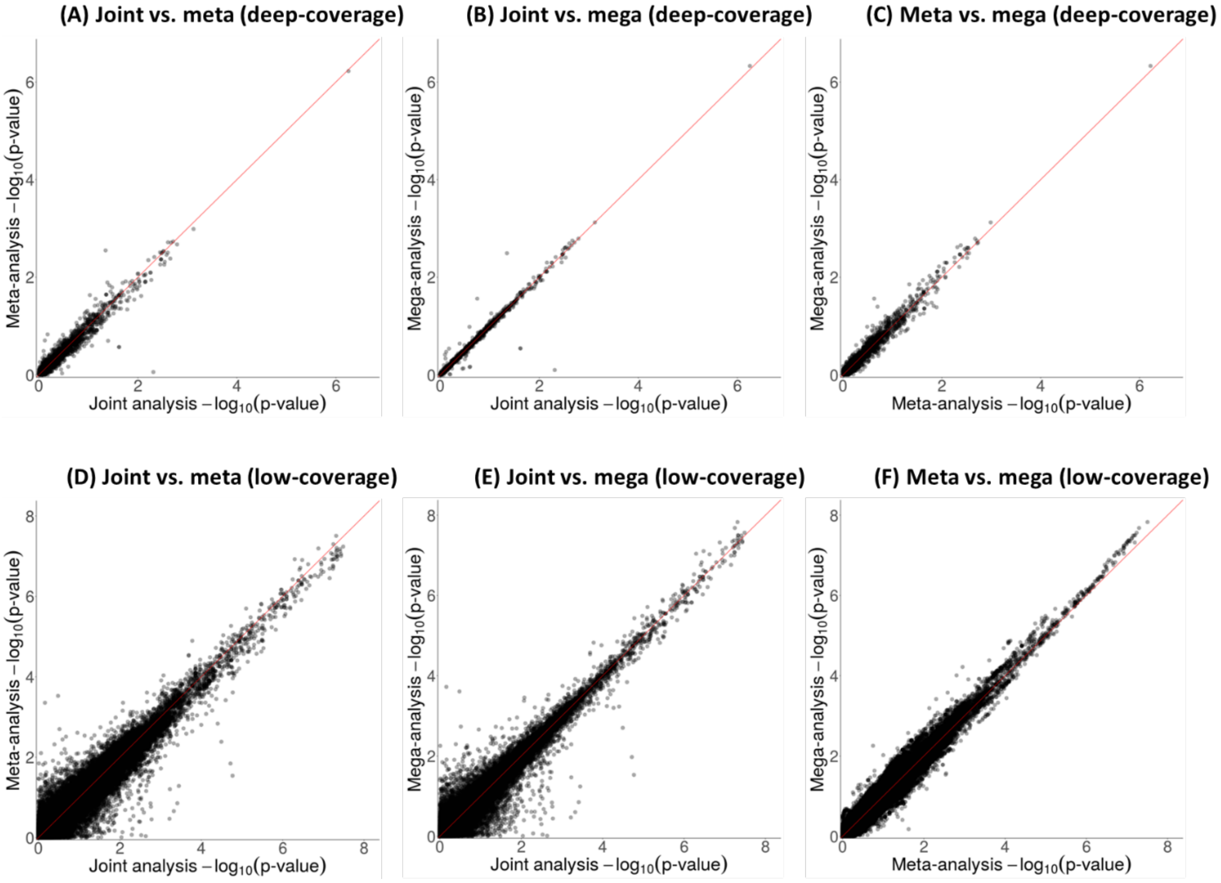
(A) Joint vs. meta (deep-coverage) **(B) Joint vs. mega (deep-coverage)** **(C) Meta vs. mega (deep-coverage)**



(D) Joint vs. meta (low-coverage) **(E) Joint vs. mega (low-coverage)** **(F) Meta vs. mega (low-coverage)**

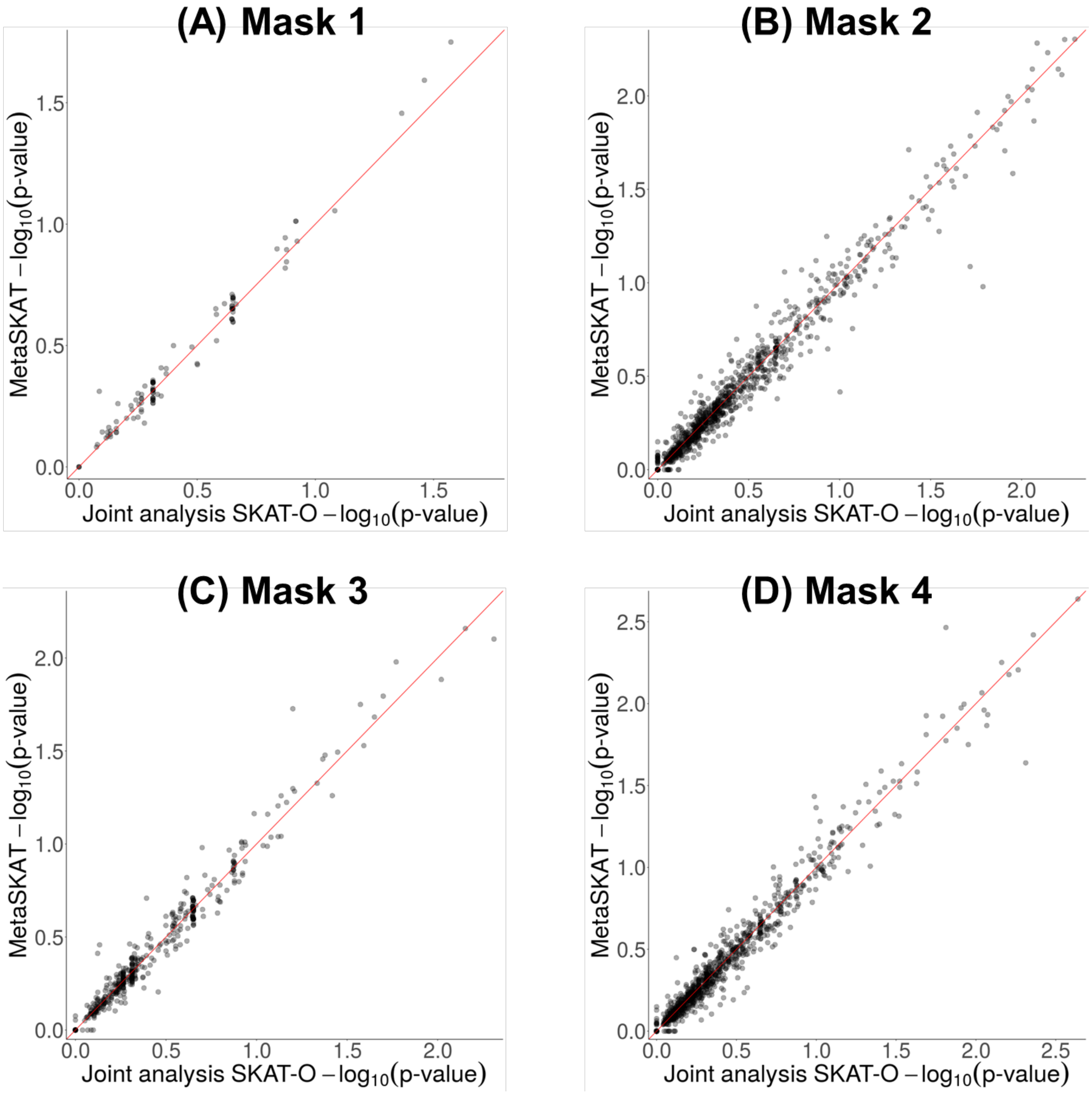
Comparison for low-frequency (MAF 0.5-5%) SNVs in (A-C) deep-coverage (~82X) exome sequence data and (D-F) low-coverage (~5X) genome sequence data. *Joint* refers to joint analysis of the joint callset, *meta* refers to meta-analysis of single-study summary statistics, and *mega* refers to joint analysis of the union callset (mega-analysis).

Supplementary Figure 2.2: Comparison of single-variant association test p-values between joint and single study calling strategies for common SNVs



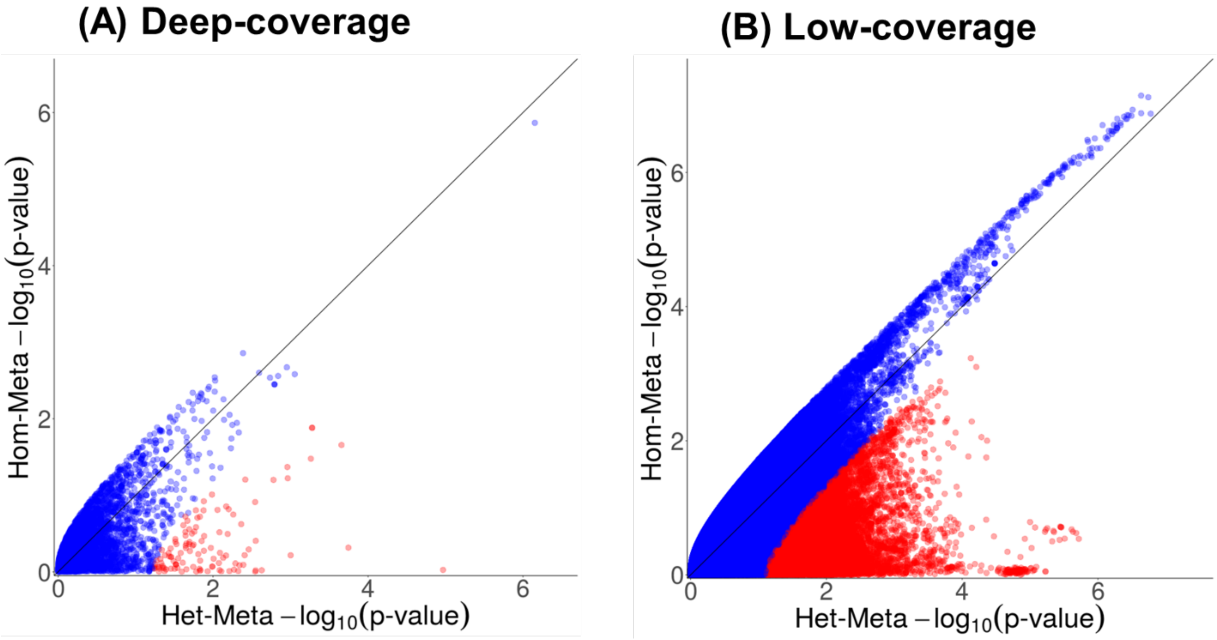
Comparison for common (MAF >5%) SNVs in (A-C) deep-coverage (~82X) exome sequence data and (D-F) low-coverage (~5X) genome sequence data. *Joint* refers to joint analysis of the joint callset, *meta* refers to meta-analysis of single-study summary statistics, and *mega* refers to joint analysis of the union callset (mega-analysis).

Supplementary Figure 2.3: Comparison of gene-based association test p-values between joint and single study calling strategies in deep-coverage exome sequence data



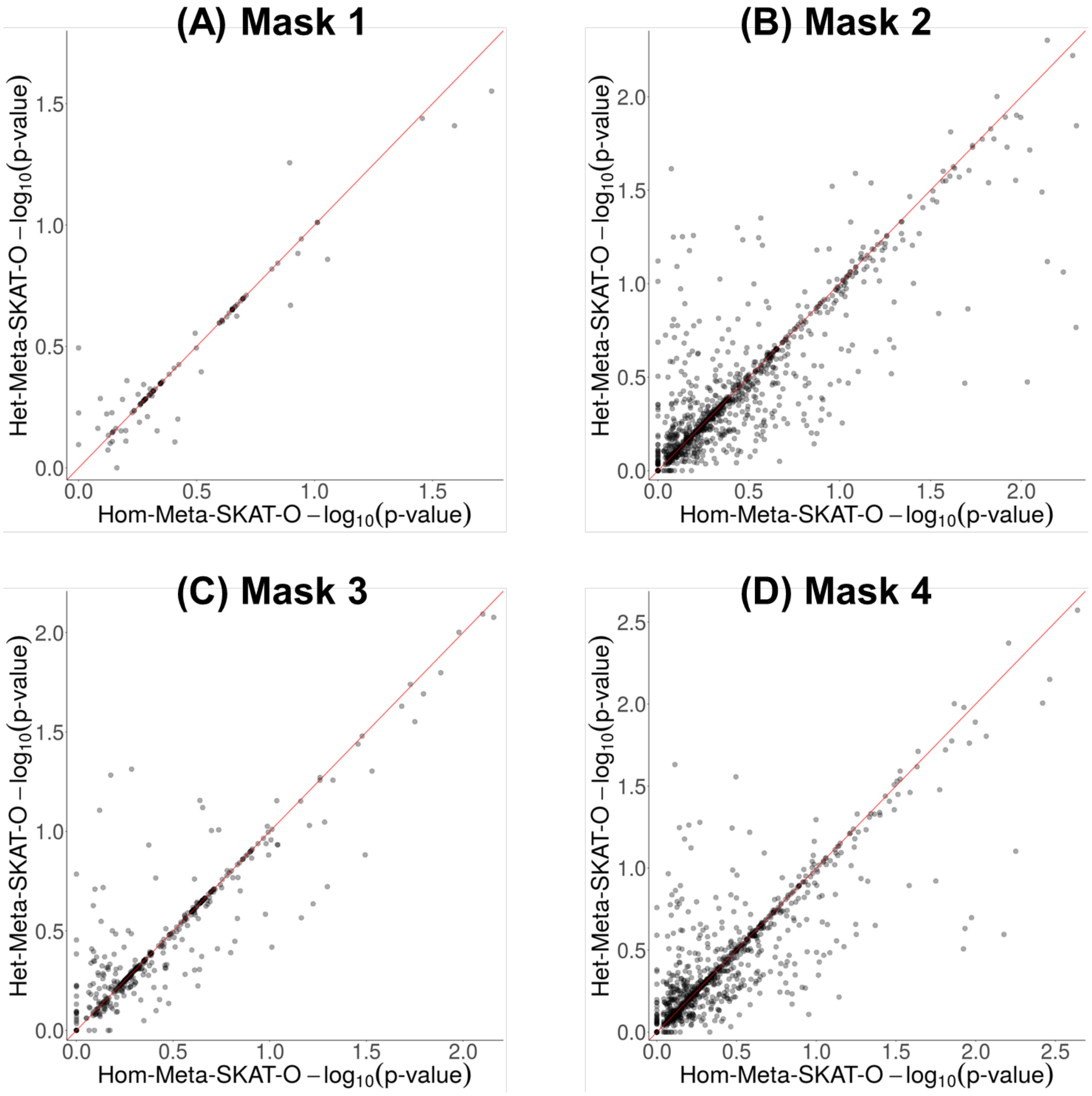
Mask 1: protein-truncating SNVs; Mask 2: Mask1+missense SNVs with MAF<1%; Mask 3: Mask1+SNVs predicted deleterious by all algorithms (Polyphen2-HumDiv, PolyPhen2-HumVar, LRT, Mutation Taster, and SIFT); Mask 4: Mask1+SNVs with MAF<1% predicted deleterious by at least one algorithm.

Supplementary Figure 2.4: Comparison of trans-ethnic meta-analysis and fixed effects meta-analysis



Comparison of trans-ethnic meta-analysis (Het-Meta) using MR-MEGA and fixed effects meta-analysis (Hom-Meta) using METAL for (A) deep-coverage (~82X) exome sequence data and (B) low-coverage (~5X) genome sequence data. Red points denote variants whose heterogeneity in genetic effects is correlated with ancestry (p-value<0.05) while blue points denote variants whose heterogeneity is not correlated with ancestry (p-value≥0.05).

Supplementary Figure 2.5: Comparison between gene-based meta-analysis assuming homogenous genetic effects between single-study cohorts versus heterogenous genetic effects in deep-coverage exome sequence data



Comparison between gene-based meta-analysis assuming homogenous genetic effects between single-study cohorts (Hom-Meta-SKAT-O) and gene-based meta-analysis assuming heterogenous genetic effects (Het-Meta-SKAT-O) in deep-coverage (~82X) exome sequence data. Mask 1: protein-truncating SNVs; Mask 2: Mask1+missense SNVs with MAF<1%; Mask 3: Mask1+SNVs predicted deleterious by all algorithms (Polyphen2-HumDiv, PolyPhen2-HumVar, LRT, Mutation Taster, and SIFT); Mask 4: Mask1+SNVs with MAF<1% predicted deleterious by at least one algorithm.

2.6 Supplementary Tables

Supplementary Table 2.1: Comparison of genotype calls for low-frequency SNVs from deep-coverage exome sequence data

Single-study variant calling (union callset)	**Joint variant calling (joint callset)**				
	Missing	**Homozygous reference**	**Heterozygous**	**Homozygous alternate**	**Total**
Missing	**0**	205,517 (4.3%)	241 (0.005%)	1 (<0.000%)	205,759 (4.3%)
Hom. ref.	0	**4,292,576 (90%)**	145 (0.003%)	0	4,292,721 (90%)
Heterozygous	0	134 (0.003%)	**168,488 (3.5%)**	6 (<0.000%)	168,628 (3.5%)
Hom. alt.	0	1 (<0.000%)	8 (<0.000%)	**118,633 (2.5%)**	118,642 (2.5%)
Total	0	4,498,228 (94%)	168,882 (3.5%)	118,640 (2.5%)	**4,785,750 (100%)**

Note. Genotype calls from joint (horizontal axis) and single-study (vertical axis) calling strategies for 2,127 low-frequency (MAF 0.5-5%) SNVs from chromosome 2 in deep-coverage (~82X) exome sequence data. Concordant calls between the two strategies are highlighted in bold.

Supplementary Table 2.2: Comparison of genotype calls for low-frequency SNVs from low-coverage genome sequence data (coding regions)

Single-study variant calling (union callset)	**Joint variant calling (joint callset)**				
	Missing	**Homozygous reference**	**Heterozygous**	**Homozygous alternate**	**Total**
Missing	**0**	435,208 (9.1%)	4,551 (0.095%)	460 (0.010%)	440,219 (9.2%)
Hom. ref.	0	**4,051,957 (85%)**	3,976 (0.083%)	20 (<0.000%)	4,055,953 (85%)
Heterozygous	0	6,244 (0.13%)	**164,676 (3.4%)**	446 (0.009%)	171,366 (3.5%)
Hom. alt.	0	35 (<0.000%)	348 (0.007%)	**117,829 (2.5%)**	118,212 (2.5%)
Total	0	4,493,444 (94%)	173,551 (3.5%)	118,755 (2.5%)	**4,785,750 (100%)**

Note. Genotype calls from joint (horizontal axis) and single-study (vertical axis) calling strategies for 2,127 low-frequency (MAF 0.5-5%) SNVs from chromosome 2 in low-coverage (~5X) exome sequence data. Concordant calls between the two strategies are highlighted in bold.

Supplementary Table 2.3: Comparison of genotype calls for common SNVs from deep-coverage exome sequence data

Single-study variant calling (union callset)	Joint variant calling (joint callset)				
	Missing	Homozygous reference	Heterozygous	Homozygous alternate	Total
Missing	**0**	5,418 (0.12%)	2,216 (0.049%)	1,876 (0.041%)	9,510 (0.21%)
Hom. ref.	0	**2,344,309 (51%)**	913 (0.020%)	62 (0.001%)	2,345,284 (51%)
Heterozygous	0	2,288 (0.050%)	**1,454,893 (32%)**	817 (0.018%)	1,457,998 (32%)
Hom. alt.	0	34 (<0.000%)	930 (0.020%)	**746,994 (16%)**	747,958 (16%)
Total	0	2,352,049 (51%)	1,458,952 (32%)	749,749 (16%)	**4,560,750 (100%)**

Note. Genotype calls from joint (horizontal axis) and single-study (vertical axis) calling strategies for 2,027 common (MAF >5%) SNVs from chromosome 2 in deep-coverage (~82X) exome sequence data. Concordant calls between the two strategies are highlighted in bold.

Supplementary Table 2.4: Comparison of genotype calls for common SNVs from low-coverage exome sequence data (coding regions)

Single-study variant calling (union callset)	Joint variant calling (joint callset)				
	Missing	Homozygous reference	Heterozygous	Homozygous alternate	Total
Missing	**0**	18,544 (0.41%)	12,430 (0.27%)	4,791 (0.11%)	35,765 (0.78%)
Hom. ref.	0	**2,318,025 (51%)**	6,935 (0.15%)	104 (0.0023%)	2,325,064 (51%)
Heterozygous	0	7,321 (0.26%)	**1,443,179 (32%)**	4,258 (0.093%)	1,454,758 (32%)
Hom. alt.	0	105 (0.002%)	4,340 (0.095%)	**740,718 (16%)**	745,163 (16%)
Total	0	2,343,995 (52%)	1,466,884 (32%)	749,871 (16%)	**4,560,750 (100%)**

Note. Genotype calls from joint (horizontal axis) and single-study (vertical axis) calling strategies for 2,027 common (MAF >5%) SNVs from chromosome 2 in low-coverage (~5X) exome sequence data. Concordant calls between the two strategies are highlighted in bold.

Supplementary Table 2.5: Non-reference genotype accuracy for joint and single-study calling strategies in low GC-content regions

Data type	**Genotype concordance for joint callset**	**Genotype concordance for union callset**
Deep-coverage		
Rare (MAF <0.5%)	99.8% (56,820/56,930)	99.8% (56,819/56,930)
Low-frequency (MAF 0.5-5%)	99.6% (83,993/84,362)	99.6% (83,990/84,362)
Common (MAF >5%)	99.6% (717,566/720,794)	99.5% (717,430/720,794)
Low-coverage (all regions)		
Rare (MAF <0.5%)	99.8% (2,371,854/2,377,331)	99.6% (2,366,887/2,377,331)
Low-frequency (MAF 0.5-5%)	99.7% (4,145,660/4,159,713)	99.4% (4,133,671/4,159,713)
Common (MAF >5%)	99.6% (65,993,499/66,235,206)	99.4% (65,868,693/66,235,206)

Note. Low GC-content regions denote regions of chr2 with <60% of base pairs as GC in data downloaded from UCSC Genome Browser (http://hgdownload.cse.ucsc.edu/downloads.html). Genotype concordance is for joint and single-study calling strategies in deep-coverage (~82X) exome and low-coverage (~5X) genome sequence data. The “truth” set of high confidence genotypes being compared against comes from Illumina HumanOmni 2.5 array data and deep exome sequence in the GoT2D integrated panel. Raw genotype counts are displayed in parentheses.

Supplementary Table 2.6: Non-reference genotype accuracy for joint and single-study calling strategies in high GC-content regions

Data type	**Genotype concordance for joint callset**	**Genotype concordance for union callset**
Deep-coverage		
Rare (MAF <0.5%)	99.5% (34,637/34,826)	99.5% (34,637/34,826)
Low-frequency (MAF 0.5-5%)	99.1% (87,946/88,769)	99.1% (87,940/88,769)
Common (MAF >5%)	99.1% (995,175/1,004,079)	99.0% (993,955/1,004,079)
Low-coverage (all regions)		
Rare (MAF <0.5%)	99.5% (1,189,066/1,195,441)	98.8% (1,180,848/1,195,441)
Low-frequency (MAF 0.5-5%)	99.4% (2,700,173/2,717,315)	98.7% (2,682,205/2,717,315)
Common (MAF >5%)	99.6% (46,967,504/47,158,111)	99.3% (46,819,733/47,158,111)

Note. High GC-content regions denote regions of chr2 with ≥60% of base pairs as GC in data downloaded from UCSC Genome Browser (http://hgdownload.cse.ucsc.edu/downloads.html). Genotype concordance is for joint and single-study calling strategies in deep-coverage (~82X) exome and low-coverage (~5X) genome sequence data. The "truth" set of high confidence genotypes being compared against comes from Illumina HumanOmni 2.5 array data and deep exome sequence in the GoT2D integrated panel. Raw genotype counts are displayed in parentheses.

Supplementary Table 2.7: Non-reference genotype accuracy for joint and single-study calling strategies using the GATK pipeline

Data type	**Genotype concordance for joint callset**	**Genotype concordance for union callset**
Deep-coverage		
Rare (MAF <0.5%)	99.6% (91,166/91,538)	99.6% (91,164/91,538)
Low-frequency (MAF 0.5-5%)	98.7% (170,338/172,514)	98.7% (170,232/172,514)
Common (MAF >5%)	98.5% (1,688,315/1,714,336)	98.3% (1,684,449/1,714,336)
Low-coverage (all regions)		
Rare (MAF <0.5%)	99.4% (3,542,984/3,564,234)	99.0% (3,530,150/3,564,234)
Low-frequency (MAF 0.5-5%)	99.0% (6,753,260/6,823,574)	98.1% (6,695,430/6,823,574)
Common (MAF >5%)	99.3% (111,765,967/112,550,367)	98.9% (111,319,059/112,550,367)

Note. Genotype concordance for joint and single-study calling strategies in deep-coverage (~82X) exome and low-coverage (~5X) genome sequence data. The "truth" set of high confidence genotypes being compared against comes from Illumina HumanOmni 2.5 array data and deep exome sequence in the GoT2D integrated panel. Raw genotype counts are displayed in parentheses.

Chapter 3

Revisiting the Genome-wide Significance Threshold for Common Variant GWAS

3.1 Introduction

There has been recent discussion in the statistical community on changing the standard P-value significance threshold for a single test from 0.05 to 0.005 (Amrhein et al., 2019; Benjamin et al., 2018; Wasserstein et al., 2019). Although the authors of the corresponding paper (Benjamin et al., 2018) commended human geneticists for using very stringent P-value thresholds to help ensure reproducibility, the cost of this strategy in current genetic studies is that many true genetic signals are not identified. The benefit is, of course, rigorous control of false positives.

To account for multiple testing in genome-wide association studies (GWAS), a fixed P-value threshold of 5×10^{-8} is widely used to identify association between a common genetic variant and a trait of interest. Risch and Merikangas (1996) suggested this strict P-value threshold for studying the genetics of complex diseases due to the many false positive discoveries reported by candidate gene studies at that time. Later, the International HapMap Consortium (Altshuler et al., 2005), Dudbridge and Gusnanto (Dudbridge & Gusnanto, 2008b), and Pe'er *et al.* (Pe'er et al., 2008b) independently suggested near-identical thresholds for common variant (minor allele frequency [MAF] > 5%) GWAS. Each group of investigators sought to control the family-wise error rate (FWER) through Bonferroni correction for the effective number of independent tests given the linkage disequilibrium (LD) structure of the genome; they used different approaches to estimate the effective number of independent tests. Based on these studies and reinforced by widespread use, the $P = 5\times10^{-8}$ threshold soon became standard for common variant GWAS. Using this threshold has been remarkably successful in limiting false positive association findings, leading to robust and reproducible results in a field that prior to GWAS had reported many nonreplicable results.

Since the acceptance of the $P = 5\times10^{-8}$ threshold a decade ago, there have been substantial experimental and methodological advances that have allowed study of many more common variants in much larger samples. The construction of denser genotype arrays (Burdick et al., 2006), development of genotype imputation (Y. Li et al., 2009b, 2010), and increasing sizes of imputation reference panels (S. McCarthy et al., 2016) now allow assay of nearly all common human genetic variants. Development of tools for meta-analysis (Willer et al., 2010; Winkler et al., 2014) has facilitated the aggregation of results across GWAS and contributed to the increasing sample sizes of genetic studies. With this changing landscape, it is worthwhile to revisit (Panagiotou et al., 2012) the common variant genome-wide threshold of $P = 5\times10^{-8}$ considering the knowledge and data acquired in the last decade.

Instead of controlling the FWER, an inherently conservative metric, an alternative approach to multiple testing corrections is to use adjusted P-values to control the false discovery rate (FDR) or to use posterior probabilities to control the Bayesian FDR (Efron et al., 2001). Although using the Benjamini-Hochberg (B-H) procedure (Benjamini & Hochberg, 1995) is the standard practice in expression quantitative trait locus (eQTL) studies and a Bayesian counterpart has also been proposed (Wen, 2017), FDR-controlling procedures have not been widely used in GWAS. In the case of B-H, this may be due to concerns about inflated estimates of FDR under the LD structure observed in genetic data (Schwartzman & Lin, 2011). Recently, Brzyski *et al.* (Brzyski et al., 2017) proposed a blocking strategy that groups tested variants into clusters based on LD before applying B-H and showed that this adapted procedure controlled the FDR at their target threshold of 5%. However, their analysis was limited to 364,590 variants in 5,402 samples and it is unclear how their procedure applies to meta-analysis where LD structures can vary across studies. There is a need to evaluate this adapted B-H and the more conservative Benjamini-Yekutieli (B-Y) procedure (Benjamini & Yekutieli, 2001b) as well as other procedures that control the Bayesian FDR over a broad range of FDR thresholds at the current scale of common variant GWAS with larger samples and millions of variants.

Here, we use knowledge gathered from current studies to re-evaluate earlier common variant GWAS meta-analyses and assess the impact of different multiple testing procedures on true and false positive rate. Along with varying the P-value threshold which controls the FWER, we evaluate the B-H and B-Y procedures to control the FDR, and the Bayesian false discovery probability (BFDP) procedure to control the Bayesian FDR. We apply the multiple testing procedures to earlier common variant meta-analyses from the Global Lipids (GLGC) and GIANT GWAS consortia on several frequently studied traits: lipid levels, height, and body mass index (BMI). Since the true set of causal variants for each trait is unknown, we use the latest and largest meta-analyses for each trait as the approximate "truth" to evaluate the performance of the multiple testing procedures in our empirical datasets. We supplement this analysis with simulation studies where the truth is known. Our results demonstrate that the standard 5×10^{-8} P-value threshold is the best multiple testing procedure for limiting false positives and is appropriate for modest-sized studies or for resource-intensive follow-ups such as constructing animal models where the cost of follow-up for each locus is high. In contrast, a less stringent P-value threshold of 5×10^{-7} (as first suggested by the Wellcome Trust Case Control Consortium (Wellcome Trust Case Control Consortium, 2007)) or the adapted B-H procedure at target FDR thresholds of 5% increases power to detect true positives in large studies and can be viable for follow-ups where the cost of including a modestly greater set of false positives is low, such as gene set enrichment, pathway analysis, or high-throughput functional follow-ups. This in-depth examination provides useful guidance to investigators who are currently conducting GWAS.

3.2 Methods

3.2.1 Introduction

We first consider an additive genetic model for a single continuous trait Y and the genotype G_j at variant $j = 1, \dots, m$

$$Y = X^T\beta + G_j\theta_j + \varepsilon_j \qquad (1)$$

where X is a $p \times 1$ vector of covariates including the intercept, β is a $p \times 1$ vector of covariate effects, θ_j is the effect of variant j, and ε_j is the normally distributed error with mean 0 and variance σ^2_j. This model can easily be extended to binary traits using a logit link function.

In a sample of n individuals, we wish to test the null hypotheses $H_{0,j}: \theta_j = 0$ against the alternatives $H_{1,j}: \theta_j \neq 0$ for each variant j. Table 3.1 summarizes the possible outcomes for the m tests of which m_0 null hypotheses are true. For studying multiple testing procedures, we focus on the first row of the table: R is the total number of rejected null hypotheses, V the number of null hypotheses incorrectly rejected (false positives), and S the number of null hypotheses correctly rejected (true positives). The proportion of false positives Q among all rejected hypotheses is then equal to V/R for $R > 0$ and set to 0 for $R = 0$.

Several procedures can be used to address the issue of controlling false positives when testing multiple hypotheses. In the remainder of this section we describe four such procedures, their extension to joint analysis of multiple traits, and application and assessment of these procedures in empirical and simulation studies in the context of common variant GWAS.

Table 3.1: Outcomes for testing multiple hypotheses

		True hypothesis		Total
		H$_0$	H$_1$	
Test	H$_0$ rejected	V	S	R
	H$_0$ not rejected	U	T	m-R
Total		m_0	m-m_0	m

3.2.2 FWER control

The standard procedure to correct for multiple testing in GWAS is to control the FWER, the probability of rejecting at least one true null hypothesis:

$$FWER = \mathrm{P}(V > 0) = \mathrm{P}(Q > 0)$$

Fixed P-value thresholds often control the FWER by using the Bonferroni procedure which provides control of FWER at level α by rejecting any null hypothesis $H_{0,j}$ for variant $j = 1, \dots, m$ with P-value

$$p_j \leq \frac{\alpha}{m}$$

When the variants are in LD and the corresponding test statistics are correlated, this procedure is conservative. One can increase the power of the Bonferroni procedure by adjusting for the effective number of independent tests (Altshuler et al., 2005; Dudbridge & Gusnanto, 2008b; Pe'er et al., 2008b) $m' \leq m$ that takes into account LD.

3.2.3 FDR control

While FWER procedures control the probability of incorrectly rejecting at least one true null hypothesis, FDR procedures control the expected proportion of incorrectly rejected true null hypotheses. At equivalent values of α, FDR is a less conservative error rate than FWER (Goeman & Solari, 2014). In the context of Table 3.1,

$$FDR = E[Q] = \begin{cases} E\left[\frac{V}{R}\right] & if\ R > 0 \\ 0 & if\ R = 0 \end{cases}$$

The Benjamini-Yekutieli (B-Y) procedure controls the FDR at level α under any dependency structure by ordering the P-values for the m variants from smallest to largest: $p_{(1)}, \dots, p_{(m)}$ and rejecting all null hypotheses $H_{0,j}$, $j = 1, \dots, k$ where k is the largest value for which

$$p_{(k)} \leq \frac{k}{m}\left(\frac{\alpha}{c(m)}\right)$$

and

$$c(m) = \sum_{i=1}^{m} \frac{1}{i}$$

The Benjamini-Hochberg (B-H) procedure, a commonly used FDR procedure that is valid when test statistics are positively correlated, is a special case of B-Y where *c(m)* is defined to be equal to 1.

Applying the B-H or B-Y procedure to GWAS can be challenging because discoveries are counted in units of loci (clusters of nearby variants that are correlated due to LD) rather than by each individual variant. Thus, FDR-controlling procedures need to control for a subset of tested variants, typically the most strongly associated (lead) variant at each locus. Since FDR-control does not extend to a subset of the rejected null hypotheses (Goeman & Solari, 2014), we adapt the B-H and B-Y procedures to GWAS by applying an approach similar to that proposed by Brzyski *et al.* (Brzyski et al., 2017) We first cluster the m null hypotheses into $m^* < m$ loci by performing LD clumping on the m variants using a LD threshold of $r^2 > 0.1$ and a maximal variant distance of 1Mb (e.g. Fritsche *et al.* 2019). We then form a set of m^* test statistics using the lead variant from each locus and apply the B-H or B-Y procedures on these m^* test statistics.

3.2.4 Bayesian approach to multiple testing

A Bayesian approach to multiple testing involves calculating the posterior probability of the null hypotheses of no association given the data. For a single variant j, let the probability of the observed data $D = (Y, X, G_j)$ given the null hypothesis $H_{0,j}$ be $\mathrm{P}(D|H_{0,j})$. Then by Bayes' theorem, the probability of the null hypothesis given the data is

$$\mathrm{P}(\,H_{0,j}|\,D) = \frac{\mathrm{P}(D|H_{0,j})\mathrm{P}(H_{0,j})}{\mathrm{P}(D|H_{0,j})\mathrm{P}(H_{0,j}) + \mathrm{P}(D|H_{1,j})\left(1 - \mathrm{P}(H_{0,j})\right)} = \frac{BF \times PO}{BF \times PO + 1}$$

where $BF = P(D|H_{0,j})/P(D|H_{1,j})$ is the Bayes factor and $PO = \mathrm{P}(H_{0,j})/\left(1 - \mathrm{P}(H_{0,j})\right)$ is the prior odds of no association. Here, we make the commonly accepted exchangeability assumption that every tested variant has the same prior probability of being associated with the trait, i.e. $1 - \mathrm{P}(H_{0,j}) = \pi_1$ and then conservatively estimate π_1 as the proportion of tested variants with $P < 5\times10^{-8}$ in the observed summary statistics. This assumption can be easily relaxed, allowing for different priors among tested variants based on their functional annotations (H. Yang & Wang, 2015).

For calculating the posterior probability, Wakefield (Wakefield, 2007b) proposed using an approximate Bayes Factor (ABF) based on the maximum likelihood estimator (MLE) $\widehat{\theta_j}$ of the variant effect θ_j as a succinct summary of the observed data D. Following Wakefield, we approximate the BF by $P(\widehat{\theta_j}|H_{0,j})/P(\widehat{\theta_j}|H_{1,j})$. Further assuming the sampling distribution of $\widehat{\theta_j}$ is normal with mean θ_j and variance V_j and that θ_j has a prior normal distribution with mean 0 and variance W_j, we calculate the ABF as a ratio of prior predictive densities $\hat{\theta}_j | H_{0,j} \sim N(0, V_j)$ and $\hat{\theta}_j | H_{1,j} \sim N(0, V_j + W_j)$ and use it to approximate the Bayesian false discovery probability (BFDP):

$$ABF_j = \frac{1}{\sqrt{1 - r_j}} \exp\left[-\frac{Z_j^2}{2} r_j\right]$$

$$BFDP_j = \frac{ABF_j \times PO}{ABF_j \times PO + 1} \qquad (2)$$

Here, $r_j = W_j/(V_j + W_j)$ is the ratio of the prior variance to the total variance and Z_j is the test statistic for variant j. Calculating the approximate BFDP requires effect size or standard error estimates. These may not be included in publicly available GWAS results, which often are limited to P-values and/or Z statistics. If necessary, we can reliably estimate the effect size and standard error for each variant from its Z statistic and estimated MAF (Zhu et al., 2016b).

To control the Bayesian FDR (Müller et al., 2004; Wen, 2017) in multiple hypotheses testing at level α, we order the BFDPs for m variants from smallest to largest: $BFDP_{(1)}, \ldots, BFDP_{(m)}$ and reject all null hypotheses $H_{0,j}$, $j = 1, \ldots, k$ where k is the largest value for which

$$\frac{\sum_{i=1}^{k} BFDP_i}{k} \leq \alpha$$

As with the B-H and B-Y procedures, we use the BFDPs to first cluster the tested variants into loci and then apply the BFDP procedure on the lead variant for each locus.

3.2.5 Joint analysis of multiple traits

In studies of L correlated traits, there is potentially more power to detect association if the traits are analyzed together (Diggle et al., 2002). One approach is to conduct L parallel univariate tests and correct for testing multiple traits simultaneously (i.e. divide the P-value thresholds by L); an alternative is to jointly analyze the L traits using multivariate test statistics and then apply the usual multiple testing procedures to the m resulting tests.

Consider joint testing of the association between genetic variant j and the L traits under an extension of **(1)**:

$$Y_{1\times L} = X_{1\times p}^{T}\beta_{p\times L} + G_j\theta_{j,1\times L} + \varepsilon_{j,1\times L} \tag{3}$$

where ε_j is normally distributed with mean $0_{1\times L}$ and variance $\Sigma_{L\times L}$ representing the covariance matrix of the trait residuals. In **(3)**, we test the m null hypotheses of no association with any trait: $H_{0,j}: \theta_{1,j} = \ldots = \theta_{L,j} = 0$ for each variant $j = 1, \ldots, m$.

For Bonferroni and B-H/B-Y, we jointly analyzed all traits with metaMANOVA (Bolormaa et al., 2014; Ray & Boehnke, 2018) using the test statistic

$$t_{metaMANOVA} = Z'\hat{\Omega}^{-1}Z$$

Here Z is the vector of test statistics for the L traits, $\hat{\Omega}$ is the estimated correlation matrix for the L traits, and $t_{metaMANOVA}$ follows an apoproximate chi-squared distribution with L degrees of freedom. We then apply the Bonferroni and B-Y procedures to the multivariate test statistics using the same approach as for the univariate study. To control BFDP, we use an extension (Wakefield, 2007b) of the ABF in **(2)** to multiple traits (Appendix 3.1).

3.2.6 Empirical studies

We evaluated the performance of the multiple testing procedures in the context of common variant GWAS by using publicly available meta-analysis results from the GLGC and the GIANT consortia. For each procedure, we calculated the empirical false discovery rate (eFDR) as the number of false positive loci in the test set (V in Table 3.1) divided by the total number of significant loci identified in the test set (R in Table 3.1). Since V is unknown as we do not know the truth, we assume that the largest, most recent GWAS represents "truth". We clustered variants declared significant by each procedure into loci using LD clumping. First, we ordered the significant variants by P-values and then using the variant with the smallest P-value (i.e. most significant) as the lead variant, we grouped all other variants that had LD threshold of $r^2 > 0.1$ with the lead variant and within ±1Mb of the lead variant into one locus. Next, we repeated this step on the remaining ungrouped variants until all significant variants were clustered into loci. In the test set, we labeled loci whose lead variants had $r^2 > 0.80$ with a variant in the truth set with $P < 5\times10^{-8}$ as true positives; the remaining loci we considered false positives.

Out of four GWAS meta-analyses (Kathiresan et al., 2009; Teslovich et al., 2010; Willer et al., 2008, 2013) sequentially carried out for plasma high-density lipoprotein cholesterol (HDL), lowdensity lipoprotein cholesterol (LDL), and triglycerides (TG) levels, we picked the largest meta-analysis (Willer et al., 2013) with n = 188,577 to serve as the truth set and the second smallest meta-analysis (Kathiresan et al., 2009) with n = 19,840 to serve as the test set. We do not present results for the other two meta-analyses in the main text because one (Willer et al., 2008) (n = 8,816) had limited power and detected few significant variants and the other (Teslovich et al., 2010) (n = 100,184) had very substantial overlap in samples with the truth set so that there was insufficient sample size differences for the truth set to well approximate the truth. Of the 2,373,282 variants analyzed in both the truth and test sets, we analyzed the 2,120,069 (89%) with MAF > 5%.

To evaluate the multiple testing procedures over a wider range of sample sizes and genetic architectures, we also applied the procedures to meta-analyses for height and body mass index (BMI) from the GIANT consortium (Lango Allen et al., 2010; Speliotes et al., 2010; Yengo et al., 2018). We present results for these meta-analyses from a larger set of sequential meta-analyses (Lango Allen et al., 2010; Locke et al., 2015; Speliotes et al., 2010; Wood et al., 2014; Yengo et al., 2018) using the same rationale as described above for GLGC: the largest, most recent meta-analyses (Yengo et al., 2018) for height and BMI (n = 694,529 and n = 681,275, respectively) served as the truth sets and the smallest meta-analyses (Lango Allen et al., 2010; Speliotes et al., 2010) for each trait (n = 133,653 and n = 123,865) served as the test sets. Of the 2,282,242 variants analyzed in both meta-analyses for height, we analyzed the 2,036,404 (89%) with MAF > 5%. Of the 2,282,195 variants in both meta-analyses for BMI, we analyzed the 2,035,656 (89%) with MAF > 5%.

For univariate analysis of each lipid and anthropometric trait, we used published meta-analyses results. Detailed descriptions of the statistical analyses for each of the results can be found in their respective papers (Kathiresan et al., 2009; Lango Allen et al., 2010; Speliotes et al., 2010; Willer et al., 2013; Yengo et al., 2018). For multivariate analysis of the three lipid traits together, we combined the univariate results using the appropriate multivariate extension for each of the procedures as described above.

3.2.7 Simulation studies

To evaluate the multiple testing procedures when truth is known, we generated 1,000 replicate datasets based on the empirical association structures observed in the latest GWAS for each of the five traits.

To mimic the GLGC test set which consisted of European cohorts, we randomly sampled 19,840 individuals from 276,791 unrelated individuals of white British ancestry in the UK BioBank dataset. For each replicate, we used the genotypes of these individuals to generate outcomes on $n = 19{,}840$ individuals for each lipid trait following model **(1)**. We assumed the trait value Y is inverse normalized to maintain consistency with the empirical studies, we estimated the causal variant effect sizes θ from the latest GLGC GWAS (the truth set), and the error term is normally distributed with mean 0 and variance equal to 1 minus the proportion of trait variance explained by the simulated causal variants. We ran association analysis with each replicate dataset using a linear regression model with no additional covariates. We took a similar approach for simulating height and BMI based on the GIANT dataset using separate generation models for the two traits.

3.3 Results

We applied the multiple testing procedures to HDL, LDL, TG, height, and BMI to assess their performances for different sample sizes and genetic structures.

3.3.1 P-value threshold

Applying various fixed P-value thresholds to the empirical GLGC and GIANT test sets, we observed as expected that the empirical false discovery rate (eFDR) generally increased as the P-value threshold increased (Table 3.2). The lone exception (for HDL) likely reflected statistical noise due to the small number of identified loci.

Table 3.2: Empirical and simulation results for P-value thresholds

Trait	Threshold (P-value)	Empirical				Simulation			
		Positives		eFDR[b]	Δ in #of sig. loci (% true)	Positives		eFDR	Δ in #of sig. loci (% true)
		False	True[a]			False	True		
HDL (n_{test} = 19,840 n_{truth} = 188,577)	5×10^{-8}	1	16	5.9%	-	0.28	9.5	2.9%	-
	5×10^{-7}	1	18	5.3%	+2 (100%)	0.89	12	6.7%	+3.6 (83%)
	5×10^{-6}	8	21	28%	+10 (30%)	4.2	17	20%	+8.2 (60%)
LDL (n_{test} = 19,840 n_{truth} = 188,577)	5×10^{-8}	0	14	0%	-	0.19	13	1.5%	-
	5×10^{-7}	3	16	16%	+5 (40%)	0.71	16	4.2%	+3.9 (87%)
	5×10^{-6}	10	19	34%	+10 (30%)	4.6	22	17%	+9.1 (58%)
TG (n_{test} = 19,840 n_{truth} = 188,577)	5×10^{-8}	1	8	11%	-	0.11	9.0	1.2%	-
	5×10^{-7}	2	10	17%	+3 (67%)	0.54	10	4.9%	+1.9 (77%)
	5×10^{-6}	6	15	29%	+9 (56%)	3.8	13	22%	+6.1 (47%)
Height (n_{test} = 133,653 n_{truth}= 693,529)	5×10^{-8}	0	157	0%	-	1.6	181	0.89%	-
	5×10^{-7}	1	217	0.46%	+61 (98%)	4.9	223	2.2%	+46 (93%)
	5×10^{-6}	2	312	0.64%	+96 (99%)	16	283	5.4%	+72 (84%)
BMI (n_{test} = 123,865 n_{truth} = 681,275)	5×10^{-8}	0	22	0%	-	0.62	39	1.6%	-
	5×10^{-7}	0	37	0%	+15 (100%)	2.7	58	4.4%	+22 (90%)
	5×10^{-6}	1	55	1.8%	+19 (95%)	11	90	11%	+41 (79%)

Note: [a] Number of loci in truth set for HDL: 89, LDL: 72, TG: 60, height: 1100, BMI: 724.
[b] eFDR is calculated as number of false positives divided by sum of true and false positives.

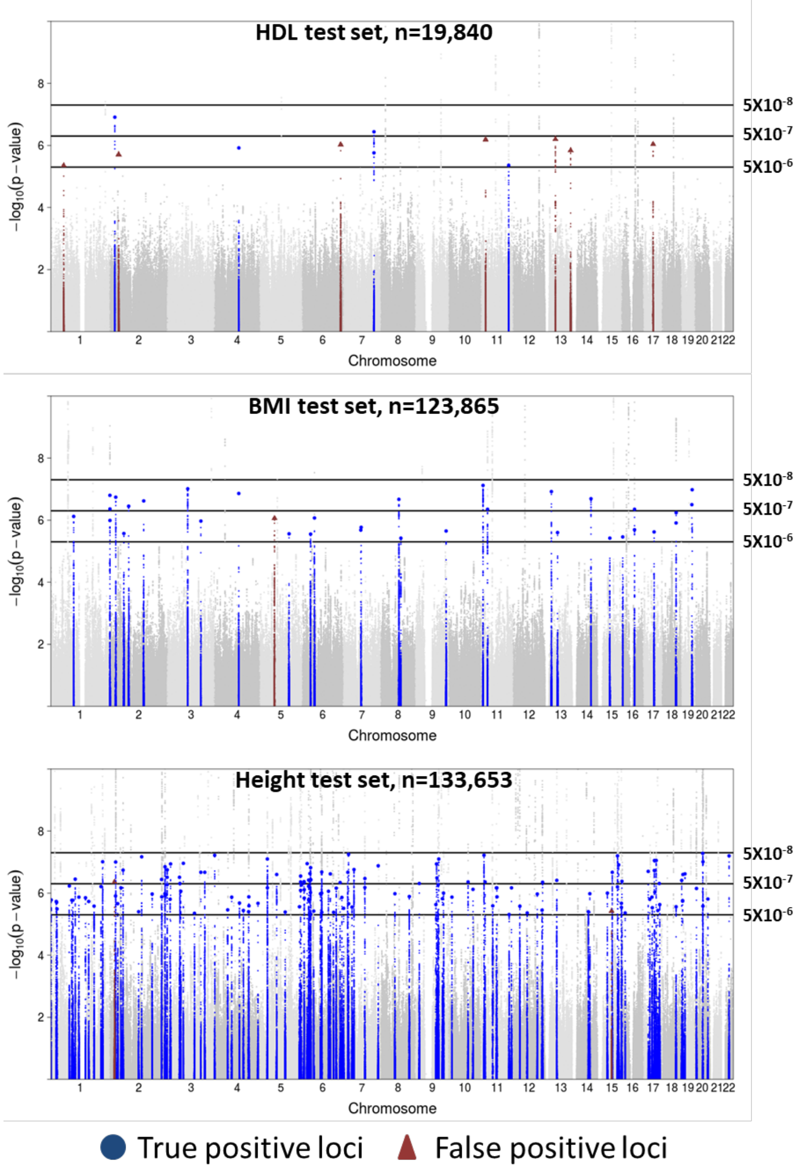
For height and BMI, we identified substantially more loci by relaxing the threshold from $P = 5\times10^{-8}$ to $P = 5\times10^{-7}$ with nearly all these new loci being true positives (Table 3.2 and Figure 3.1): 60 of 61 (98%) for height; 15 of 15 (100%) for BMI. Further relaxing the threshold from $P = 5\times10^{-7}$ to $P = 5\times10^{-6}$ maintained high proportions of true positives among the additional loci: 95 of 96 (99%) for height, 18 of 19 (95%) for BMI. For the lipid traits in the GLGC test set, relaxing the threshold from $P = 5\times10^{-8}$ to $P = 5\times10^{-7}$ resulted in HDL, LDL, and TG gaining 2, 5, and 3 loci with 2, 2, and 2 (100%, 40%, and 67%) being true positives. Further relaxing the threshold from $P = 5\times10^{-7}$ to $P = 5\times10^{-6}$ resulted in ≤ 56% of the additional loci being true positives for the lipid traits.

We observed in the GLGC- and GIANT-based simulated datasets that the average eFDR increased as the P-value threshold increased for all traits (Table 3.2); the inconsistency described before for the empirical HDL test set disappeared when we averaged over 1,000 simulation replicates.

Consistent with the empirical results, there was a clear difference in the proportion of true positives between the lipid and anthropometric traits in the simulated results (Table 3.2). Relaxing the threshold from $P = 5\times10^{-8}$ to $P = 5\times10^{-7}$ in the simulated datasets resulted in an average of 77% to

87% of the additional loci being true positives for the lipid traits and 93% and 90% for height and BMI. Further relaxing the threshold from $P = 5\times10^{-7}$ to $P = 5\times10^{-6}$ resulted in 47% to 60% of the additional loci being true positives for lipids, and 84% and 79% for height and BMI.

Figure 3.1: Manhattan plot of empirical P-value thresholds



Plots of different P-value thresholds applied to empirical test sets for HDL, BMI, and height. Colored variants depict true positive loci (blue) and false positive loci (red) for variants with P-value $\geq 5\text{x}10^{-8}$. Lead variants for true and false positive loci are represented by large blue circles and large triangles, respectively.

To address whether the higher rates of true positives we observed when relaxing the P-value threshold for height and BMI compared to those for lipids were the result of differences in sample sizes, we simulated test sets for height and BMI at the same sample sizes (n=8,816 and n = 19,840) as the GLGC meta-analyses. For both traits, an increase in sample size generally led to higher proportion of true positives gained from relaxing the P-value threshold (Table 3.3), suggesting a better yield of true positives by using relaxed thresholds in larger samples than in smaller ones.

Table 3.3: Effect of sample size on simulation results for P-value thresholds

Trait	Threshold (P-value)	n = 8,816				n = 19,840			
		Positives		eFDR[b]	Δ sig. loci (% True positive)	Positives		eFDR	Δ sig. loci (% True positive)
		False	True[a]			False	True		
Height	5×10^{-8}	0.04	0.90	4.3%	-	0.03	11	0.27%	-
	5×10^{-7}	0.40	2.2	15%	+1.7 (78%)	0.32	18	1.7%	+7.4 (96%)
	5×10^{-6}	3.0	5.9	34%	+6.3 (58%)	3.6	30	11%	+15 (79%)
BMI	5×10^{-8}	0.04	0.20	17%	-	0.09	1.5	5.7%	-
	5×10^{-7}	0.34	0.41	45%	+0.51 (41%)	0.46	2.4	16%	+1.3 (72%)
	5×10^{-6}	3.1	1.2	73%	+3.5 (21%)	3.2	4.6	41%	+5.0 (44%)

Note: [a] Number of loci in truth set for HDL: 89, LDL: 72, TG: 60, height: 1100, BMI: 724.
[b] eFDR is calculated as number of false positives divided by sum of true and false positives.

3.3.2 Benjamini-Hochberg and Benjamini-Yekutieli procedures

As expected, empirical results for the two FDR controlling procedures showed B-Y was conservative, resulting in eFDR far below the target FDR threshold for all traits at commonly used (5-15%; Table 3.4) and more extreme (1-25%; Supplementary Table 3.1) thresholds. B-H controlled the eFDR at the target thresholds (Table 3.5 and Supplementary Table 3.2) for height and BMI but not lipid traits, likely because the number of lipid trait discoveries was modest (≤ 26 loci for B-H) so that even a small change in numbers of true and false positives substantially influenced estimated eFDR.

Simulation results for B-Y were consistent with empirical results in showing that B-Y is overly conservative for all five traits and all target FDR thresholds (Table 3.4 and Supplementary Table 3.1). For example, the observed eFDR for target threshold of 15% is < 3.4% for all traits. Compared to the empirical results, B-H did a better job of controlling eFDR at the commonly used thresholds (Table 3.5) for all traits; only for height at 5% and BMI at 5% did B-H show noticeable inflation in eFDR (6.8% for height, 7.9% for BMI). When we relaxed our criterion for defining a true positive (see below), inflations for height and BMI decreased (eFDR = 5.5% and 5.2%). eFDR was well-controlled at high thresholds 20% and 25% for all five traits but poorly-controlled at low thresholds 1% and 3% (Supplementary Table 3.2).

Table 3.4: Empirical and simulation results for Benjamini-Yekutieli procedure

Trait	Threshold (FDR)	Empirical			Simulation		
		Positives			Positives		
		False	True[a]	eFDR[b]	False	True	eFDR
HDL (n_{test} = 19,840 n_{truth} = 188,577)	5%	0	14	0%	0.17	8.2	2.0%
	10%	1	16	5.9%	0.25	9.1	2.7%
	15%	1	16	5.9%	0.31	9.6	3.1%
LDL (n_{test} = 19,840 n_{truth} = 188,577)	5%	0	14	0%	0.13	12	1.1%
	10%	0	14	0%	0.19	13	1.5%
	15%	0	15	0%	0.26	13	1.9%
TG (n_{test} = 19,840 n_{truth} = 188,577)	5%	0	8	0%	0.05	8.5	0.58%
	10%	0	8	0%	0.06	8.7	0.68%
	15%	1	8	11%	0.11	9.0	1.2%
Height (n_{test} = 133,653 n_{truth} = 693,529)	5%	0	197	0%	4.3	217	2.0%
	10%	1	234	0.43%	6.3	235	2.6%
	15%	1	249	0.40%	7.9	246	3.1%
BMI (n_{test} = 123,865 n_{truth} = 681,275)	5%	0	20	0%	0.83	41	2.0%
	10%	0	22	0%	1.4	47	2.9%
	15%	0	26	0%	1.8	52	3.4%

Note: [a] Number of loci in truth set for HDL: 89, LDL: 72, TG: 60, height: 1100, BMI: 724.
[b] eFDR is calculated as number of false positives divided by sum of true and false positives.

Table 3.5: Empirical and simulation results for Benjamini-Hochberg procedure

Trait	Threshold (FDR)	Empirical			Simulation		
		Positives			Positives		
		False	True[a]	eFDR[b]	False	True	eFDR
HDL (n_{test} = 19,840 n_{truth} = 188,577)	5%	1	18	5.3%	0.70	12	5.6%
	10%	5	18	22%	1.3	13	8.5%
	15%	6	20	23%	1.7	14	11%
LDL (n_{test} = 19,840 n_{truth} = 188,577)	5%	3	16	16%	0.67	16	4.0%
	10%	4	16	20%	1.1	18	6.0%
	15%	7	17	29%	1.7	18	8.6%
TG (n_{test} = 19,840 n_{truth} = 188,577)	5%	2	9	18%	0.32	10	3.1%
	10%	5	10	33%	0.63	11	5.6%
	15%	5	10	33%	0.96	11	8.0%
Height (n_{test} = 133,653 n_{truth} = 693,529)	5%	2	351	0.57%	22	301	6.8%
	10%	4	421	0.94%	37	331	10%
	15%	8	468	1.7%	50	351	13%
BMI (n_{test} = 123,865 n_{truth} = 681,275)	5%	1	41	2.4%	6.6	77	7.9%
	10%	1	47	2.1%	11	91	11%
	15%	1	55	1.8%	16	102	14%

Note: [a] Number of loci in truth set for HDL: 89, LDL: 72, TG: 60, height: 1100, BMI: 724.

[b] eFDR is calculated as number of false positives divided by sum of true and false positives.

We investigated whether FDR control for B-H and B-Y extended across sample sizes by using simulated datasets for height and BMI at n = 8,816, n = 19,840 and n = 133,653 (height) or 123,865 (BMI). Both procedures controlled eFDR at the target FDR thresholds 5-15% for height (Supplementary Table 3.4 and 3.5); BMI showed inflation under B-H for all test sets which disappeared under the relaxed definition of true positives.

3.3.3 BFDP

For the BFDP procedure, we estimated the prior probability of association at a variant site (π_1) separately for each test set using the proportion of tested variants with $P < 5\times10^{-8}$. Empirical results showed that eFDR was well controlled for height and BMI at target Bayesian FDR thresholds 1-25% but poorly controlled for lipid traits (Table 3.6 and Supplementary Table 3.3), again likely due to the smaller number of discoveries for lipid traits (≤ 24 loci for BFDP).

Simulation results for BFDP showed that eFDR was generally well controlled at target Bayesian FDR thresholds 5-15% (Table 3.6) for all traits except height (eFDR = 8.1%, 13%, and 17%). For more extreme thresholds (Supplementary Table 3.3), eFDR was controlled at 1 and 3% for lipid traits, albeit with inflation for HDL at 1%; eFDR was controlled at 20% and 25% for all traits.

Table 3.6: Empirical and simulation results for BFDP procedure

Trait	Threshold (Bayesian FDR)	Empirical				Simulation			
		$\widehat{\pi_1}$[a]	Positives False	Positives True[b]	eFDR[c]	$\widehat{\pi_1}$[d]	Positives False	Positives True	eFDR
HDL (n_{test} = 19,840 n_{truth} = 188,577)	5%	1.3×10^{-4}	1	17	5.6%	8.7×10^{-5}	0.41	10	4.0%
	10%		4	17	19%		0.76	12	6.1%
	15%		6	18	25%		1.2	13	8.4%
LDL (n_{test} = 19,840 n_{truth} = 188,577)	5%	1.3×10^{-4}	2	17	11%	9.6×10^{-5}	0.37	14	2.5%
	10%		5	17	23%		0.83	16	4.9%
	15%		6	18	25%		1.3	18	7.0%
TG (n_{test} = 19,840 n_{truth} = 188,577)	5%	2.1×10^{-4}	1	10	9.1%	1.6×10^{-4}	0.36	9.8	3.6%
	10%		4	10	29%		1.0	11	8.4%
	15%		4	12	25%		1.6	12	12%
Height (n_{test} = 133,653 n_{truth} = 693,529)	5%	2.0×10^{-3}	2	338	0.59%	2.9×10^{-3}	28	317	8.1%
	10%		7	406	1.7%		51	356	13%
	15%		9	468	1.9%		76	385	17%
BMI (n_{test} = 123,865 n_{truth} = 681,275)	5%	3.6×10^{-4}	0	35	0%	5.2×10^{-4}	3.9	67	5.4%
	10%		0	43	0%		7.2	82	8.0%
	15%		0	50	0%		11	93	10%

Note: [a] $\widehat{\pi_1}$ is the estimated prior probability of association at a variant site equal to the proportion of tested variants with P-value $< 5\times10^{-8}$.

[b] Number of loci in truth set for HDL: 89, LDL: 72, TG: 60, height: 1100, BMI: 724.

[c] eFDR is calculated as number of false positives divided by sum of true and false positives.

[d] Average $\widehat{\pi_1}$ in 1,000 replicate datasets.

3.3.4 Multi-trait analysis results for lipids

In empirical results (Supplementary Table 3.6), the $P = 5\times10^{-8}$ threshold had the lowest eFDR for the parallel univariate tests both adjusted (Bonferroni corrected threshold of 1.67×10^{-8}) and unadjusted (5×10^{-8}) for testing three traits. For the multivariate tests, the $P = 5\times10^{-8}$ and $P = 5\times10^{-7}$ thresholds had identical eFDR of 0%. Between the three sets of thresholds, the multivariate analysis had the lowest eFDR as well as the highest proportion of true positive discoveries when relaxing the P-value thresholds. For both the B-H and BFDP procedures, multivariate tests had lower eFDR than the univariate tests but only the multivariate B-H procedure controlled the eFDR at target thresholds 5-15%.

In simulation results (Supplementary Table 3.7), the $P = 5\times10^{-8}$ threshold had the lowest eFDR for all three sets of tests. Consistent with empirical results, multivariate tests had the lowest eFDR at all three P-value thresholds and better true positive rate for relaxing thresholds compared with the univariate tests. For the B-H and BFDP procedure, both univariate and multivariate tests controlled the eFDR at target thresholds 5-15%.

3.3.5 Sensitivity analyses

We defined true positives in the test set strictly as loci whose lead variants had LD threshold of $r^2 > 0.80$ with a genome-wide significant (P-value $< 5\times10^{-8}$) variant in the truth set. We chose this strict criterion to avoid underestimating the number of false positives in our analysis but it likely led to overestimation of eFDR. To assess the impact of this, we repeated our simulation analyses using a relaxed definition of true positives by lowering the LD threshold requirement to 0.60 and the P-value requirement to 5×10^{-7} (Supplementary Table 3.8). As expected, under this relaxed definition we observed fewer false positives and occurrences of inflated eFDR largely disappeared as well. For example, simulation results for height using the BFDP procedure at FDR thresholds of 10% and 15% showed eFDR of 13% and 17% under the strict definition and 10% and 14% under the relaxed definition.

In addition to our LD-based definitions, we used physical distance to define loci and true positives. We grouped variants within ±1Mb of the lead variants into loci and defined true positives as loci whose lead variants were within ±50kb of a genome-wide significant variant in the truth set. The analyses results (Supplementary Table 3.9) showed that the distance-based definitions led to smaller numbers of true and false positives for all traits and multiple testing procedures.

3.4 Discussion

In this paper, we leverage the sequentially growing nature of GWAS meta-analyses to evaluate true and false positive rate of P-value thresholds and other multiple testing procedures. Although the standard procedure for identifying significant associations in common variant GWAS is to use a P-value threshold of 5×10^{-8}, relaxing the significance criteria, whether through use of less stringent P-value thresholds or controlling for alternative error rate measures such as FDR (depending on the target threshold) increases the number of identified loci. We demonstrated that a substantial proportion of the additional loci identified are true positives, with larger proportions of true positives in analysis of larger samples.

3.4.1 Application to downstream analyses

GWAS identify trait-associated variants and loci based on association analysis of millions of variants. The identified loci are often further validated in replication studies before being used for statistical and functional analyses to identify causal genes, variants, and mechanisms. Although relaxed P-value thresholds are often used to generate the list of loci for replication, the expected true and false discovery rates under different thresholds have not been quantified. We showed by simulation for common variant GWAS with sample size > 100,000 that 90-93% of additional discoveries with P-values between 5×10^{-8} and 5×10^{-7} were true positives, representing true associations that would be lost under a more stringent threshold. However, for more modest sample sizes (~20,000), our simulation showed that only 77-87% of additionally discovered loci with P-values between 5×10^{-8} and 5×10^{-7} were true positives. Here, investigators should exercise caution when relaxing the significance threshold for replication studies as the increase to replicated associations may not outweigh the inflated false discovery rate.

For follow-up studies such as constructing animal models where the per-locus cost of follow-up is high, a stringent P-value threshold of 5×10^{-8} is ideal in both large and modest-sized studies to generate a highly accurate list of associated loci. However, such threshold may be unhelpfully conservative for analyses where including (many) more true loci at the cost of (a few) more false positives is acceptable such as gene-set enrichment or pathway analysis. In these situations, a relaxed threshold of 5×10^{-7} or even 5×10^{-6} may be better served to prioritize GWAS results for downstream analyses. The utility of these relaxed thresholds can be seen in the DEPICT (Pers et al., 2015) software designed for gene prioritization, gene set enrichment analysis, and identifying enriched tissue or cell types at significant loci discovered by GWAS. Here, the authors recommend using DEPICT on all GWAS loci with P-value < 1×10^{-5} to improve discovery of causal gene sets for direct functional follow-ups.

3.4.2 FDR- and Bayesian FDR-control

FDR-control is an appropriate choice for practitioners who are willing to tolerate some proportion of false positive discoveries as long as it can be controlled below a target threshold. At equivalent thresholds, controlling the FDR is less conservative than controlling the FWER and thus expands the GWAS-identified set of associated loci for downstream analysis, especially for highly polygenic traits. We showed that the B-H procedure adapted for GWAS (see Methods) provided approximate control of the empirical estimate of FDR (eFDR) for the tested traits and samples at target thresholds 5-25%. The B-Y procedure is far too conservative in GWAS as the correction factor which removes assumptions on the dependency structure of test statistics is unnecessary under the adapted B-H procedure which forms independent test statistics using the lead variants from each locus.

For BFDP, a Bayesian alternative to B-H, we estimated the proportion of trait-associated variants π_1 using the proportion of tested variants with P-values less than 5×10^{-8} and found the Bayesian FDR to be reasonably well controlled at thresholds of 5-25%. For comparison, when we estimated π_1 as the number of loci with lead variant $P < 5\times10^{-8}$ divided by 1 million (an estimate for the total number of independent common variants in the genome (Altshuler et al., 2005; Pe'er et al., 2008b)), the resulting lower π_1 estimates led to conservative results (Supplementary Table 3.10).

3.4.3 Comparison between procedures

A P-value threshold has the advantages of familiarity, simplicity, and ease of implementation while B-H and BFDP control the eFDR across a range of sample sizes. In simulations, both B-H and BFDP controlled the eFDR for height and HDL, two traits with different genetic architectures and for which we analyzed very different sample sizes (n = 133,653 and n = 19,840). In contrast, 95% of discoveries at a P-value threshold of 5×10^{-6} were true positives for height while only 80% were true positives for HDL. A stringent P-value threshold is needed if our primary goal is to limit the number of false positives as both B-H and BFDP struggled to control the eFDR at low target thresholds 1% and 3%.

3.4.4 Summary

In this study, we evaluated the performance of four procedures for multiple testing corrections in the context of common variant GWAS: P-value thresholds, B-H and B-Y for FDR control, and BFDP for Bayesian FDR control. We have shown that for studies based on large samples, using a less stringent P-value threshold of 5×10^{-7} or use of FDR-controlling procedure (B-H) at target threshold of 5% substantially increases the number of true positive discoveries that can be used in downstream analyses while only modestly increasing false positives compared with the commonly used 5×10^{-8} P-value threshold. The latter threshold remains the preferred choice for modest-sized studies or when a stringently curated list of loci is desired. Finally, we show that FDR-control extends across sample sizes and FDR-controlling procedures can be similarly applied to large and modest-sized studies.

3.5 Supplementary materials

Appendix 3.1: Multivariate BFDP

Consider joint testing of the association between a genetic variant and L traits under model **(3)**. To match our analysis, we set $L = 3$ for the rest of this section but this multivariate extension can be applied to any number of traits.

As in the univariate case described in Methods and in Wakefield (2007),[23] we approximate the multivariate Bayes' factor by $P(\widehat{\boldsymbol{\theta}}|H_0)/P(\widehat{\boldsymbol{\theta}}|H_1)$ where $\widehat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3)$ is the estimated vector of variant effect sizes for the three traits. We assume the sampling distribution of $\widehat{\boldsymbol{\theta}}$ is multivariate normal with mean $\boldsymbol{\theta}$ and variance $\boldsymbol{V}$ and that $\boldsymbol{\theta}$ has a prior multivariate normal distribution with mean $\mathbf{0}$ and variance $\boldsymbol{W}$. If $\boldsymbol{\rho}$ is the 3×3 matrix of correlation between the traits and ρ_{ij} is the correlation between traits i and j, then

$$\boldsymbol{V} = \begin{bmatrix} V_1 & \rho_{12}\sqrt{V_1V_2} & \rho_{13}\sqrt{V_1V_3} \\ \rho_{21}\sqrt{V_2V_1} & V_2 & \rho_{23}\sqrt{V_2V_3} \\ \rho_{31}\sqrt{V_3V_1} & \rho_{32}\sqrt{V_3V_2} & V_3 \end{bmatrix}$$

$$\boldsymbol{W} = \begin{bmatrix} W_1 & \rho_{12}\sqrt{W_1W_2} & \rho_{13}\sqrt{W_1W_3} \\ \rho_{21}\sqrt{W_2W_1} & W_2 & \rho_{23}\sqrt{W_2W_3} \\ \rho_{31}\sqrt{W_3W_1} & \rho_{32}\sqrt{W_3W_3} & W_3 \end{bmatrix}$$

where V_k is the sample variance and W_k is the prior variance for trait k. Finally, the multivariate approximate Bayes' factor (ABF) can be calculated as a ratio of prior predictive densities $\widehat{\boldsymbol{\theta}}|H_0 \sim MVN(\mathbf{0}, \boldsymbol{V})$ and $\widehat{\boldsymbol{\theta}}|H_1 \sim MVN(\mathbf{0}, \boldsymbol{V} + \boldsymbol{W})$ and used to approximate the BFDP:

$$ABF_{multi} = |\boldsymbol{V}|^{-\frac{1}{2}} |\boldsymbol{V} + \boldsymbol{W}|^{\frac{1}{2}} \exp\left[\frac{\widehat{\boldsymbol{\theta}}^T(-\boldsymbol{V}^{-1} + (\boldsymbol{V} + \boldsymbol{W})^{-1})\widehat{\boldsymbol{\theta}}}{2}\right]$$

$$BFDP_{multi} = \frac{ABF_{multi} \times PO}{ABF_{multi} \times PO + 1}$$

For the prior odds of no association, we estimate the prior probability of being associated with the three traits ($\widehat{\pi_1}_{multi}$) as the average of the $\widehat{\pi_1}$'s from each trait which is calculated as described in Methods.

3.6 Supplementary Tables

Supplementary Table 3.1: Benjamini-Yekutieli results for extreme thresholds

Trait	Threshold (FDR)	Empirical			Simulation		
		Positives		eFDR[b]	Positives		eFDR
		False	True[a]		False	True	
HDL (n_{test}=19,840 n_{truth}=188,577)	1%	0	14	0%	0.08	7.0	1.1%
	3%	0	14	0%	0.13	7.9	1.6%
	20%	1	16	5.9%	0.34	10	3.3%
	25%	1	17	5.6%	0.35	10	3.3%
LDL (n_{test}=19,840 n_{truth}=188,577)	1%	0	14	0%	0.06	10	0.59%
	3%	0	14	0%	0.08	11	0.72%
	20%	0	15	0%	0.27	14	1.9%
	25%	0	15	0%	0.30	14	2.1%
TG (n_{test}=19,840 n_{truth}=188,577)	1%	0	8	0%	0.02	7.8	0.26%
	3%	0	8	0%	0.03	8.3	0.36%
	20%	1	8	11%	0.13	9.1	1.4%
	25%	2	8	20%	0.14	9.2	1.5%
Height (n_{test}=133,653 n_{truth}693,529)	1%	0	157	0%	1.8	185	0.95%
	3%	0	180	0%	3.3	206	1.6%
	20%	1	272	0.37%	9.3	254	3.5%
	25%	1	281	0.35%	11	261	4.0%
BMI (n_{test}=123,865 n_{truth}=681,275)	1%	0	19	0%	0.22	31	0.71%
	3%	0	20	0%	0.51	37	1.4%
	20%	0	29	0%	2.2	55	3.9%
	25%	0	32	0%	2.7	58	4.4%

Note: [a] Number of loci in truth set for HDL: 89, LDL: 72, TG: 60, height: 1100, BMI: 724.

[b] eFDR is calculated as number of false positives divided by sum of true and false positives.

Supplementary Table 3.2: Benjamini-Hochberg results for extreme thresholds

Trait	Threshold (FDR)	Empirical			Simulation		
		Positives		eFDR[b]	Positives		eFDR
		False	True[a]		False	True	
HDL (n_{test}=19,840 n_{truth}=188,577)	1%	1	16	5.9%	0.31	9.6	3.1%
	3%	1	17	5.6%	0.50	11	4.3%
	20%	7	20	26%	2.2	15	13%
	25%	7	20	26%	2.7	16	15%
LDL (n_{test}=19,840 n_{truth}=188,577)	1%	0	15	0%	0.26	13	1.9%
	3%	0	15	0%	0.45	15	2.9%
	20%	7	18	28%	2.2	19	10%
	25%	7	18	28%	2.8	20	12%
TG (n_{test}=19,840 n_{truth}=188,577)	1%	1	8	11%	0.11	9.0	1.2%
	3%	2	8	20%	0.24	9.6	2.5%
	20%	5	10	33%	1.3	12	10%
	25%	5	10	33%	1.6	12	12%
Height (n_{test}=133,653 n_{truth}693,529)	1%	1	249	0.40%	7.90	246	3.1%
	3%	2	309	0.64%	15	281	5.2%
	20%	10	496	2.0%	63	368	15%
	25%	16	540	2.9%	76	382	17%
BMI (n_{test}=123,865 n_{truth}=681,275)	1%	0	26	0%	1.8	52	3.4%
	3%	0	37	0%	4.3	68	6.0%
	20%	3	61	4.7%	22	111	16%
	25%	3	66	4.3%	27	118	19%

Note: [a] Number of loci in truth set for HDL: 89, LDL: 72, TG: 60, height: 1100, BMI: 724.

[b] eFDR is calculated as number of false positives divided by sum of true and false positives.

Supplementary Table 3.3: BFDP results for extreme thresholds

Trait	Threshold (Bayesian FDR)	Empirical				Simulation			
		$\widehat{\pi_1}$[a]	Positives		eFDR[c]	$\widehat{\pi_1}$[d]	Positives		eFDR
			False	True[b]			False	True	
HDL (n_{test}=19,840 n_{truth}=188,577)	1%	1.3×10^{-4}	0	14	0%	8.7×10^{-5}	0.18	7.5	2.3%
	3%		1	16	5.9%		0.30	9.0	3.2%
	20%		8	18	31%		1.8	14	11%
	25%		9	19	32%		2.5	15	14%
LDL (n_{test}=19,840 n_{truth}=188,577)	1%	1.3×10^{-4}	0	15	0%	9.6×10^{-5}	0.12	11	1.1%
	3%		1	16	5.9%		0.26	13	1.9%
	20%		7	19	27%		2.2	19	10%
	25%		9	20	31%		3.2	20	13%
TG (n_{test}=19,840 n_{truth}=188,577)	1%	2.1×10^{-4}	0	8	0%	1.6×10^{-4}	0.07	8.1	0.85%
	3%		1	9	10%		0.20	9.2	2.1%
	20%		4	14	22%		2.6	12	17%
	25%		6	14	30%		3.6	13	22%
Height (n_{test}=133,653 n_{truth}693,529)	1%	2.0×10^{-3}	1	234	0.43%	2.9×10^{-3}	8.7	254	3.3%
	3%		2	299	0.66%		18	294	5.9%
	20%		16	523	3.0%		106	410	21%
	25%		20	584	3.3%		141	433	25%
BMI (n_{test}=123,865 n_{truth}=681,275)	1%	$3.6\ \times10^{-4}$	0	25	0%	5.2×10^{-4}	1.1	46	2.2%
	3%		0	31	0%		2.6	59	4.1%
	20%		1	57	1.7%		15	103	13%
	25%		1	65	1.5%		21	112	16%

Note: [a] $\widehat{\pi_1}$ is the estimated prior probability of association at a variant site equal to the proportion of tested variants with P-value less than 5×10^{-8}.

[b] Number of loci in truth set for HDL: 89, LDL: 72, TG: 60, height: 1100, BMI: 724.

[c] eFDR is calculated as number of false positives divided by sum of true and false positives.

[d] Average $\widehat{\pi_1}$ in 1,000 replicate datasets.

Supplementary Table 3.4: Effect of sample size on simulation results for Benjamini-Yekutieli

Trait	Threshold (FDR)	n=8,816			n=19,840			n=133,653 or 123,865		
		Positives		eFDR[b]	Positives		eFDR	Positives		eFDR
		False	True[a]		False	True		False	True	
Height	5%	0	0.22	0%	0.01	9.0	0.11%	4.3	217	2.0%
	10%	0	0.30	0%	0.01	11	0.10%	6.3	235	2.6%
	15%	0	0.35	0%	0.06	12	0.51%	7.9	246	3.1%
BMI	5%	0.01	0.05	17%	0.01	0.99	1.0%	0.83	41	2.0%
	10%	0.01	0.07	13%	0.01	1.1	0.88%	1.4	47	2.9%
	15%	0.01	0.08	11%	0.01	1.2	0.81%	1.8	52	3.4%

Note: [a] Number of loci in truth set for HDL: 89, LDL: 72, TG: 60, height: 1100, BMI: 724.
[b] eFDR is calculated as number of false positives divided by sum of true and false positives.

Supplementary Table 3.5: Effect of sample size on simulation results for Benjamini-Hochberg

Trait	Threshold (FDR)	n=8,816			n=19,840			n=133,653 or 123,865		
		Positives		eFDR[b]	Positives		eFDR	Positives		eFDR
		False	True[a]		False	True		False	True	
Height	5%	0.05	0.87	5.4%	0.27	18	1.5%	22	301	6.8%
	10%	0.12	1.2	9.0%	0.85	22	3.7%	37	331	10%
	15%	0.18	1.5	11%	1.5	25	5.6%	50	351	13%
BMI	5%	0.01	0.17	5.6%	0.11	1.5	6.7%	6.6	77	7.9%
	10%	0.05	0.22	19%	0.23	1.8	11%	11	91	11%
	15%	0.08	0.25	24%	0.33	2.1	14%	16	102	14%

Note: [a] Number of loci in truth set for HDL: 89, LDL: 72, TG: 60, height: 1100, BMI: 724.
[b] eFDR is calculated as number of false positives divided by sum of true and false positives.

Supplementary Table 3.6. Combined univariate and multivariate empirical results for lipids

	Metric		Threshold	Combined univariate			Multivariate		
				Positives		eFDR[b]	Positives		eFDR
				False	True[a]		False	True	
HDL,LDL,TG (n_{test}=19,840 n_{truth}=188,577)	P-value	Unadjusted	5×10^{-8}	2	38	5.0%	0	41	0%
			5×10^{-7}	6	47	11%	0	45	0%
			5×10^{-6}	24	58	29%	2	52	3.7%
		Adjusted	1.67×10^{-8}	0	35	0%	*Not applicable*		
			1.67×10^{-7}	2	40	4.8%			
			1.67×10^{-6}	13	53	20%			
	B-H		5%	6	53	10%	2	52	3.7%
			10%	14	60	19%	5	52	8.8%
			15%	18	60	23%	8	53	13%
	BFDP[c]		5%	7	57	11%	5	58	7.9%
			10%	15	61	20%	10	61	14%
			15%	19	66	22%	14	65	18%

Note: [a] Number of loci in truth set for all three lipids: 139 (non-overlapping), height: 1100, BMI: 724.
[b] eFDR is calculated as number of false positives divided by sum of true and false positives.
[c] $\widehat{\pi_1}$'s for univariate analyses is 1.3×10^{-4} for HDL, 1.3×10^{-4} for LDL, and 2.1×10^{-4} for TG. $\widehat{\pi_1}_{multi}$ is 1.7×10^{-4}, the average of the $\widehat{\pi_1}$'s for the three lipid traits.

Supplementary Table 3.7. Combined univariate and multivariate simulation results for lipids

	Metric		Threshold	Combined univariate			Multivariate		
				Positives		eFDR[b]	Positives		eFDR
				False	True[a]		False	True	
HDL,LDL,TG (n_{test}=19,840 n_{truth}=188,577)	P-value	Unadjusted	5×10^{-8}	0.43	22	1.9%	0.57	30	1.9%
			5×10^{-7}	1.7	28	5.7%	1.6	37	4.1%
			5×10^{-6}	10	36	22%	6.4	47	12%
		Adjusted	1.67×10^{-8}	0.26	20	1.3%	*Not applicable*		
			1.67×10^{-7}	0.78	25	3.0%			
			1.67×10^{-6}	4.4	32	12%			
	B-H		5%	1.1	32	3.3%	2.4	40	5.7%
			10%	2.2	35	5.9%	3.9	43	8.3%
			15%	3.2	37	8.0%	5.2	46	10%
	BFDP[c]		5%	1.2	31	3.7%	1.6	34	4.5%
			10%	2.6	35	6.9%	3.1	38	7.5%
			15%	4.1	38	9.7%	4.8	40	11%

Note: [a] Number of loci in truth set for all three lipids: 139 (non-overlapping), height: 1100, BMI: 724.
[b] eFDR is calculated as number of false positives divided by sum of true and false positives.
[c] Average $\widehat{\pi_1}$'s for univariate analyses is 7.7×10^{-5} for HDL, 8.7×10^{-5} for LDL, and 1.4×10^{-4} for TG. $\widehat{\pi_1}_{multi}$ is 9.9×10^{-5}, the average of the $\widehat{\pi_1}$'s for the three lipid traits.

Supplementary Table 3.8: Simulation results for P-value thresholds under relaxed definition of true positives

	Threshold P-value | FDR	P-value threshold			B-H			BFDP		
		Positives		eFDR	Positives		eFDR	Positives		eFDR
		False	True[a]		False	True		False	True	
HDL (n_{test} = 19,840 n_{truth} = 188,577)	5×10^{-8} | 5%	0.13	9.6	1.3%	0.42	12	3.3%	0.16	10	1.5%
	5×10^{-7} | 10%	0.58	13	4.4%	0.92	14	6.3%	0.44	12	3.6%
	5×10^{-6} | 15%	3.6	18	17%	1.3	15	8.0%	0.80	13	5.7%
LDL (n_{test} = 19,840 n_{truth} = 188,577)	5×10^{-8} | 5%	0.11	13	0.84%	0.51	16	3.1%	0.27	14	1.8%
	5×10^{-7} | 10%	0.55	16	3.2%	0.93	18	5.0%	0.67	16	3.9%
	5×10^{-6} | 15%	4.0	22	15%	1.5	19	7.2%	1.1	18	5.9%
TG (n_{test} = 19,840 n_{truth} = 188,577)	5×10^{-8} | 5%	0.06	9.2	0.66%	0.23	10	2.2%	0.25	9.9	2.5%
	5×10^{-7} | 10%	0.43	11	3.9%	0.50	11	4.5%	0.79	11	6.6%
	5×10^{-6} | 15%	3.4	14	20%	0.80	11	6.7%	1.4	12	10%
Height (n_{test} = 133,653 n_{truth} = 693,529)	5×10^{-8} | 5%	1.1	181	0.62%	18	306	5.5%	23	322	6.6%
	5×10^{-7} | 10%	3.7	225	1.6%	30	337	8.3%	43	364	10%
	5×10^{-6} | 15%	13	289	4.3%	42	360	11%	65	396	14%
BMI (n_{test} = 123,865 n_{truth} = 681,275)	5×10^{-8} | 5%	0.37	39	0.94%	4.3	79	5.2%	2.4	69	3.4%
	5×10^{-7} | 10%	1.5	59	2.5%	8.2	94	8.0%	4.7	84	5.3%
	5×10^{-6} | 15%	8.0	93	7.9%	12	106	10%	7.6	96	7.3%

Note: [a] True positive defined as a locus whose lead variant had $r^2 > 0.60$ with a variant in the truth set with P-value < 5×10^{-7}.

Supplementary Table 3.9: Simulation results for P-value threshold results under distance-based definitions

	Threshold P-value | FDR	P-value threshold			B-H			BFDP		
		Positives		eFDR	Positives		eFDR	Positives		eFDR
		False	True[a]		False	True		False	True	
HDL (n_{test} = 19,840 n_{truth} = 188,577)	5×10^{-8} | 5%	0.07	7.8	0.89%	0.20	9.4	2.1%	0.06	8.0	0.75%
	5×10^{-7} | 10%	0.35	10	3.4%	0.41	10	3.8%	0.19	9.3	2.0%
	5×10^{-6} | 15%	2.9	14	17%	0.69	11	5.9%	0.46	10	4.2%
LDL (n_{test} = 19,840 n_{truth} = 188,577)	5×10^{-8} | 5%	0.02	12	0.17%	0.24	13	1.8%	0.12	12	0.96%
	5×10^{-7} | 10%	0.33	14	2.3%	0.53	14	3.5%	0.43	14	3.0%
	5×10^{-6} | 15%	3.2	17	16%	0.85	15	5.4%	0.90	15	5.6%
TG (n_{test} = 19,840 n_{truth} = 188,577)	5×10^{-8} | 5%	0.05	8.7	0.57%	0.18	9.5	1.9%	0.17	9.2	1.8%
	5×10^{-7} | 10%	0.32	9.8	3.2%	0.35	9.9	3.4%	0.58	10	5.4%
	5×10^{-6} | 15%	3.0	12	20%	0.58	10	5.4%	1.3	11	10%
Height (n_{test} = 133,653 n_{truth} = 693,529)	5×10^{-8} | 5%	0.08	172	0.05%	3.7	263	1.4%	8.7	291	2.9%
	5×10^{-7} | 10%	0.49	207	0.24%	7.3	284	2.5%	20	324	5.8%
	5×10^{-6} | 15%	2.8	255	1.1%	11	297	3.6%	34	348	9.0%
BMI (n_{test} = 123,865 n_{truth} = 681,275)	5×10^{-8} | 5%	0.14	38	0.37%	2.5	75	3.2%	1.3	67	1.9%
	5×10^{-7} | 10%	0.81	57	1.4%	4.7	90	5.0%	2.9	82	3.4%
	5×10^{-6} | 15%	4.9	90	5.1%	6.9	101	6.4%	4.8	94	4.8%

Note: [a] Locus defined as variants within ±1Mb of the lead variant

[b] True positive defined as a locus whose lead variant was within ±50kb of a variant with $P < 5\times10^{-8}$ in the truth set.

Supplementary Table 3.10: BFDP results using alternative estimation of prior

Trait	Threshold (Bayesian FDR)	Empirical				Simulation			
		$\widehat{\pi_1}$[a]	Positives		eFDR[c]	$\widehat{\pi_1}$[d]	Positives		eFDR
			False	True[b]			False	True	
HDL (n_{test}=19,840 n_{truth}=188,577)	5%	2.0×10^{-5}	0	17	0%	9.9×10^{-6}	0.14	7.2	1.9%
	10%		1	19	5.0%		0.28	8.2	3.3%
	15%		1	22	4.3%		0.35	9.2	3.7%
LDL (n_{test}=19,840 n_{truth}=188,577)	5%	1.5×10^{-5}	0	14	0%	1.1×10^{-5}	0.02	9.2	0.22%
	10%		0	16	0%		0.10	10	0.95%
	15%		1	16	5.9%		0.19	11	1.6%
TG (n_{test}=19,840 n_{truth}=188,577)	5%	1.6×10^{-5}	0	14	0%	9.7×10^{-6}	0.02	7.6	0.26%
	10%		0	16	0%		0.10	8.4	1.2%
	15%		0	18	0%		0.17	9.1	1.8%
Height (n_{test}=133,653 n_{truth}693,529)	5%	1.4×10^{-4}	1	170	0.58%	1.7×10^{-4}	4.3	218	2.0%
	10%		1	190	0.52%		7.6	241	3.1%
	15%		1	207	0.48%		12	260	4.3%
BMI (n_{test}=123,865 n_{truth}=681,275)	5%	8.8×10^{-6}	0	22	0%	4.0×10^{-5}	0.62	39	1.6%
	10%		0	22	0%		1.1	47	2.3%
	15%		0	28	0%		1.8	53	3.2%

Note: [a] $\widehat{\pi_1}$ is the estimated prior probability of association at a variant site equal to the number of loci with lead variant P-value less than 5×10^{-8} divided by 1,000,000 (estimated total number of independent common variants in genome).

[b] Number of loci in truth set for HDL: 89, LDL: 72, TG: 60, height: 1100, BMI: 724.

[c] eFDR is calculated as number of false positives divided by sum of true and false positives.

[d] Average $\widehat{\pi_1}$ in 1,000 replicate datasets.

Chapter 4

Multiple Testing Correction in Rare Variant Analysis

4.1 Introduction

There is a huge multiple testing burden in genetic association studies which must be addressed to control false positive discoveries. This burden was first quantified by Risch and Merikangas (1996) for theoretically testing one million alleles and have rapidly increased over time as technological and methodological advances allowed us to study a broader range of variants including indels and structural variations over a wider allele frequency spectrum. A recent study by the Trans-Omics for Precision Medicine (TOPMed) program (Taliun et al., 2019) identified more than 410 million genetic variants from 53,581 sequenced individuals. Even after excluding singletons (46%) and variants that do not pass quality control filters, there remain more than 120 million variants to be tested.

To account for multiple testing, researchers typically use a genome-wide P-value threshold to identify significant associations. For common variant (minor allele frequency [MAF] > 5%) studies, this threshold has been set (Dudbridge & Gusnanto, 2008a; Pe'er et al., 2008a) to 5×10^{-8} for controlling the family-wise error rate (FWER) at $\alpha=5\%$. Studies including low-frequency (MAF 0.5-5%) and rare variants (MAF < 0.5%) have a much greater testing burden than their common variant counterparts which necessitates even more stringent P-value thresholds. Recent work (D. Lin, 2019; Pulit et al., 2017) has estimated a threshold of approximately 5×10^{-9} for testing ~27 million variants from the Haplotype Reference Consortium (HRC) imputation reference panel (S. McCarthy et al., 2016). However, this number has already been surpassed by UK Biobank (UKBB) studies using the TOPMed reference panel (Taliun et al., 2019) with more than 120 million variants available for testing. There is a need to re-estimate the genome-wide significance threshold to account for the increased testing burden.

The Benjamini-Hochberg (Benjamini & Hochberg, 1995) and Benjamini-Yekutieli (Benjamini & Yekutieli, 2001a) procedures are alternatives to the P-value threshold and controls the false discovery rate (FDR), the expected proportion of false discoveries over a large number of hypothetically repeated experiments (Wen, 2016). However, to date, these FDR-controlling methods have been only infrequently applied, specifically to association studies limited to common and low-frequency variants (Nielsen et al., 2018) or a subset of tests in exome studies (Locke et al., 2019). This may be due in part to concerns about FDR control in genetic studies where test statistics are strongly correlated due to linkage disequilibrium (LD) between variants, and the testing unit and analysis unit may differ (Brzyski et al., 2017; Peterson et al., 2016; Siegmund et al., 2011). For example, researchers typically test each variant individually for association with a trait but interpret the association results in groups of correlated variants known as loci. Although BH and BY control the "global" FDR among all tested variants, this control does not extend to a subset of the testing unit such as lead variants from loci (Goeman & Solari, 2014). Modified versions of the BH and BY procedures for genetic studies have been proposed (Brzyski et al., 2017) but their implementation depends on accurately modeling the LD structure among tested variants which can be challenging to estimate for rare variants.

Although BH, BY, and P-value thresholds control different error rates, they act on the same set of P-values from association results. Developing a Bayesian method for multiple testing correction offers a different approach based on analyzing posterior probabilities instead of P-values. A Bayesian method controls the Bayesian FDR, the proportion of false positives among all discoveries conditional on the observed data (Wen, 2016; Whittemore, 2007). Several methods (Bogdan et al., 2008; Y. Tang et al., 2007; Wakefield, 2007b) have been proposed with differing levels of complexity to arrive at the same end goal of calculating the posterior probabilities. Two such methods (Bogdan et al., 2008; Y. Tang et al., 2007) formulate a Dirichlet mixture framework for modeling the P-value distribution under the alternative hypothesis and estimate the posterior probabilities using a Markov chain Monte Carlo (MCMC) algorithm. However, these Bayesian methods were developed for quantitative trait loci (QTL) studies involving several hundred simultaneous tests and may not scale to testing tens or hundreds of millions of variants in rare variant GWAS. This is an issue because an important advantage of a Bayesian method is that it can incorporate prior knowledge about the effect size distribution of the tested variants to improve discovery which is especially useful for rare variants that may have large effect sizes but for which statistical power for association is still small due to low MAF. Wakefield (2007b) proposed a less computationally-intensive method to calculate the posterior probability by combining an approximation of the Bayes factor using observed test statistics and user-specified prior parameters. The challenge with using this method is that misspecification of the prior can impact properties of the Bayesian FDR.

In this study, we propose a Bayesian method to correct for multiple testing in rare variant studies that calculates posterior probabilities using an approximation of the Bayes factor and estimates prior parameters from observed summary statistics using an E-M algorithm. An important advantage of our Bayesian method is that it only requires GWAS summary statistics (P-values and MAF) and so does not require access to individual-level data. We compare the ability of our Bayesian method to accurately identify true positives with that of the P-value threshold, BH, and BY procedures in simulated datasets based on empirical association structures observed for three traits (waist-hip-ratio (WHR), body-mass-index (BMI) , and high-density lipoprotein (HDL) cholesterol levels) in the latest genome sequence datasets from the UKBB imputed using TOPMed imputation reference panel. In addition, we assess FDR control for our Bayesian method, BH, and BY procedures. Finally, we extend the multiple testing methods to gene-based tests and apply them to real datasets from the UKBB.

4.2 Methods

Consider an additive genetic model for a single continuous trait *Y* and the genotype G_j at autosomal variant $j = 1, \dots, m$:

$$Y = X^T\theta + G_j\beta_j + \varepsilon_j \tag{1}$$

where X is a $p \times 1$ vector of covariates including the intercept, θ is a $p \times 1$ vector of covariate effects, β_j is the effect of variant j, and ε_j is the normally distributed error with mean 0 and variance σ_j^2.

For single-variant analysis, we wish to test the null hypotheses $H_{0,j}: \beta_j = 0$ against the alternatives $H_{1,j}: \beta_j \neq 0$ for each variant j. After applying multiple testing correction, we declare R variants to be statistically significant. Of these R discoveries, V of them are false positives and S of them are true positives. The goal of multiple testing methods is to facilitate the discovery of true positives while controlling the number of false positives.

4.2.1 FWER control

One way to control false positives in association tests is through the family-wise error rate (FWER):

$$FWER = \mathrm{P}(V > 0)$$

which denotes the probability of discovering at least one false positive. FWER is typically controlled by P-value thresholds using the Bonferroni procedure or modifications that seek to account for dependence of the tests. In a set of null hypotheses $H_{0,j}$ for variants $j = 1, \dots, m$, we declare variants with P-values less than the threshold α/m to be significant, where α is the acceptable FWER. When variants are in linkage disequilibrium (LD), the corresponding test statistics are correlated and the Bonferroni procedure is conservative. Here, one can increase the power by adjusting for the effective number of independent tests m' instead of m (Altshuler et al., 2005; Dudbridge & Gusnanto, 2008a; Pe'er et al., 2008a).

4.2.2 FDR control

An alternative way to control false positives is through the false discovery rate (FDR):

$$FDR = E\left[\frac{V}{R}\right]$$

which is the expected proportion of incorrectly rejected true null hypotheses. When the number of causal variants $m_1 = 0$, FDR and FWER are identical (Benjamini & Hochberg, 1995). When $m_1 > 0$, FDR is less conservative than FWER at equivalent α. This implies that every FWER-controlling method such as the P-value threshold also controls the FDR at that same level (Goeman & Solari, 2014).

The Benjamini-Hochberg (BH) procedure (Benjamini & Hochberg, 1995) controls the FDR at level α by ordering the P-values for tested variants from smallest to largest: $p_{(1)}, \ldots, p_{(m)}$ and rejecting all null hypotheses $H_{0,j}, j = 1, \ldots, k$ where k is the largest value for which

$$p_{(k)} \leq \frac{k}{m}\alpha$$

This procedure requires an assumption of *positive regression dependence on a subset* (PRDS) among the test statistics (Benjamini & Yekutieli, 2001a; Goeman & Solari, 2014). In this context, this means that variants with more significant (i.e. smaller) P-values need to have a higher probability of being a true causal variant than variants with less significant P-values. To control the FDR under any arbitrary dependency structure of the test statistics, one can use the Benjamini-Yekutieli (BY) procedure (Benjamini & Yekutieli, 2001a). This procedure is similar to BH except we increase the stringency of our rejection threshold by a factor of $c(m)$:

$$p_{(k)} \leq \frac{k}{m}\left(\frac{\alpha}{c(m)}\right)$$

where

$$c(m) = \sum_{i=1}^{m} \frac{1}{i}$$

BH is a special case of BY where $c(m) = 1$.

In genetic studies, we typically test each of the m variants individually but count the discoveries in units of loci (clusters of nearby variants that are correlated due to LD). This has important implications for BH and BY because FDR lacks a subsetting property (Goeman & Solari, 2014), meaning that FDR-controlling procedures only guarantees control for the set of R discoveries from m tests but not for any subset of R. Consider a scenario where a researcher applies BH or BY to their association results at an α=5%. The researcher is not interested in controlling the expected proportion of false positives in their single-variant discoveries at 5% (because many of these discoveries are in LD) but instead is interested in controlling the proportion of falsely discovered loci at 5%. In this scenario, the classic BH and BY procedures do not provide the type of FDR-control that the researcher desire. To address this issue, Brzyski et al. (Brzyski et al., 2017) proposed a modified version of the BH procedure called BH$_S$ which adds an initial screening step that filters the m tested variants into $m^* < m$ loci before applying the classic BH procedure on the lead variant from each locus. Here, the testing and discovery units are both counted in terms of loci.

4.2.3 Bayesian FDR control

A Bayesian approach to multiple testing correction seeks to control the Bayesian FDR (Whittemore, 2007) using posterior probabilities of the null hypotheses given the observed data $D = (Y, X, G_j)$. Bayesian FDR is the expected proportion of false positives among all discoveries conditional on the observed data while the traditional FDR is the average Bayesian FDR over many hypothetically repeated experiments (Wen, 2016).

For a single variant j, let the probability of the observed data D given the null hypothesis $H_{0,j}$ be $\mathrm{P}(D|H_{0,j})$. Then by Bayes' theorem, the probability of the null hypothesis given the data is

$$\mathrm{P}(\, H_{0,j} |\, D) = \frac{\mathrm{P}(D|H_{0,j})\mathrm{P}(H_{0,j})}{\mathrm{P}(D|H_{0,j})\mathrm{P}(H_{0,j}) + \mathrm{P}(D|H_{1,j})\left(1 - \mathrm{P}(H_{0,j})\right)} = \frac{BF \times PO}{BF \times PO + 1}$$

where $BF = P(D|H_{0,j})/P(D|H_{1,j})$ is the Bayes factor and $PO = \mathrm{P}(H_{0,j})/\left(1 - \mathrm{P}(H_{0,j})\right)$ is the prior odds of no association. For quantitative traits, previous work (Servin & Stephens, 2007) has derived an exact calculation of the BF:

$$BF = \frac{(n)^{0.5}}{W_j^{0.5}\, det(\Omega_j)^{0.5}} \left(\frac{Y^TY - Y^T\bar{G}_J\Omega_j^{-1}\bar{G}_J^{\,T}Y}{Y^TY - n\bar{Y}^2}\right)^{-\frac{n}{2}}$$

where $\bar{G}_J$ is a *n* x 2 matrix with first column all 1s and second column genotype dosages, W_j is the prior variance of variant effect, and $\Omega_j = \begin{pmatrix} 0 & 0 \\ 0 & \alpha \end{pmatrix} + \bar{G}_J^{\,T}\bar{G}_J$. However, this exact calculation of BF requires individual-level data which are frequently unavailable due to data use restrictions. Several methods (Bogdan et al., 2008; Y. Tang et al., 2007) have been proposed to estimate the posterior probability using a MCMC algorithm but such approach does not scale to rare variant GWAS with tens to hundreds of millions of tests. A less computationally-intensive calculation of the BF that only requires summary statistics is the approximate Bayes factor (ABF) based on the maximum likelihood estimator of the variant effect β as a succinct summary of the observed data D (Wakefield, 2007a):

$$ABF\left(Z_j, W_j\right) = \sqrt{\frac{V_j + W_j}{V_j}} \exp\left[-\frac{Z_j^2}{2}\left(\frac{W_j}{V_j + W_j}\right)\right] \tag{2}$$

where Z_j is the test statistic and V_j is test variance.

To control (Müller et al., 2004) the Bayesian FDR at level α, we order the posterior probabilities $\mathrm{P}(D|H_0)$ ($abbr.\ PP$) for the m tested variants from smallest to largest: $PP_{(1)}, \dots, PP_{(m)}$ and declare R variants to be significant where R is the largest value for which:

$$\frac{\sum_{i=1}^{R} PP_{(i)}}{R} \leq \alpha$$

This is equivalent to saying that we want to discover the largest number of variants such that the average posterior probability of the discoveries is less than or equal to α.

Calculating the posterior probability for a variant j requires two prior parameters: (1) $\mathrm{P}(H_0, j) = \pi_j$, the prior probability of no association and (2) W_j, the prior variance of variant effect β_j. Previous work has suggested assigning a fixed or range (Wakefield, 2007a) of values to the prior parameters. However, this can be rather arbitrary and may be prone to misspecification, particularly in the case of a single value. In this study, we take an empirical Bayes approach to estimate the prior parameters from observed summary statistics using an Expectation-Maximization (E-M) algorithm as described below.

4.2.4 Prior distribution for effect size

First, we assume a prior distribution for the effect sizes of our tested variants that follows a mixture of three zero-mean normal distributions with a point mass at zero:

$$\beta_j \sim \pi_0 \delta_0 + \pi_1 N(0, W_1) + \pi_2 N(0, W_2) + \pi_3 N(0, W_3) \tag{3}$$

Each variant j has probability π_0 of being null (i.e. $H_j = 0$) with an effect size of 0 (denoted by δ_0, a point mass at 0) and a probability π_k of being in one of the non-null groups $k = 1, 2,$ or 3 with effect size from a $N(0, W_k)$ distribution, where $\pi_0 + \pi_1 + \pi_2 + \pi_3 = 1$. Under this model, variants from the same group k will share the same prior parameters π_k and W_k.

We chose to model the causal variants using a mixture of three normal distributions because it has been shown (Park et al., 2010; Zhang et al., 2018) that any single parametric distribution (e.g. normal) does not adequately model the long tails typically present in the effect size distribution of complex traits. These tails are the result of rare causal variants that have on average larger effect sizes than those of common and low-frequency variants. We treat the membership of each tested variant in one of the null or non-null groups as latent variables whose expectations are maximized in the E-M algorithm described in the next section. We show later in Results our rationale for choosing three non-null groups and how they correspond to MAF bins for common+low-frequency (MAF > 0.5%), rare (MAF 0.005-0.5%), and very rare (MAF < 0.005%) variants.

4.2.5 E-M algorithm

The goal of our E-M algorithm is to estimate the unknown parameters $\boldsymbol{\pi} = \{\pi_1, \dots, \pi_3\}$ and $\boldsymbol{W} = \{W_1, \dots, W_3\}$ from the observed summary statistics which is composed of: (1) test statistics $\mathbf{Z} = \{Z_1, \dots, Z_m\}$, (2) estimated effect sizes $\widehat{\boldsymbol{\beta}} = \{\hat{\beta}_1, \dots, \hat{\beta}_m\}$, and (3) testing variances $\boldsymbol{V} = \{V_1, \dots, V_m\}$. If necessary, we can estimate the effect size and testing variance for each variant from its test statistic and MAF (Zhu et al., 2016a). The latent variables are indicators C_{jk} for variants $j = 1, \dots, m$ and for membership in groups $k = 1,2,3$.

For the E-step, we update the probabilities $\hat{C}_{jk}^{(i)}$ for each variant and MAF group:

$$\hat{C}_{jk}^{(i)} = E\left[C_{jk} \middle| Z_j, \pi_k^{(i)}, W_k^{(i)}\right] = \frac{\pi_k^{(i)} ABF\left(Z_j, W_k^{(i)}\right)}{\pi_0^{(i)} + \sum_{l=1}^{K} \pi_l^{(i)} ABF\left(Z_j, W_l^{(i)}\right)}$$

where the formula for *ABF* is from (2).

For the M-step, we update the prior probabilities $\pi_k^{(i)}$ and prior variances $W_k^{(i)}$:

$$\pi_k^{(i+1)} = \frac{1}{m} \sum_{j=1}^{m} \hat{C}_{jk}^{(i)}$$

$$W_k^{(i+1)} = \frac{\sum_{j=1}^{m} \hat{C}_{jk}^{(i)} \left(\hat{\beta}_j\right)^2}{\sum_{j=1}^{m} \hat{C}_{jk}^{(i)}}$$

In practice, we found this version of the E-M algorithm tended to overestimate the π_k's, the prior probabilities of association. To address this issue, we modified the update formula for the prior probabilities $\pi_k^{(i)}$ by adding a shrinkage estimator $I\left(\hat{C}_{jk}^{(i)} > t_k^{(i)}\right)$ where $t_k^{(i)}$ is the median of $\left\{\hat{C}_{1k}^{(i)}, \dots, \hat{C}_{mk}^{(i)}\right\}$. This estimator ensures that we are only counting likely causal variants in the estimation of π_k by "shrinking" very small values of membership probabilities ($\hat{C}_{jk}^{(i)}$) belonging to non-causal variants to 0. The update formula for $\pi_k^{(i)}$ in the i^{th} M-step of the modified E-M is:

$$\pi_k^{(i+1)} = \frac{1}{m} \sum_{j=1}^{m} \hat{C}_{jk}^{(i)} \, I\left(\hat{C}_{jk}^{(i)} > t_k^{(i)}\right)$$

4.2.6 Empirical analyses

We applied our Bayesian method, BH, BY, and the P-value threshold to whole-genome sequence datasets from the UKBB with genotypes imputed using TOPMed reference panel for analysis of three quantitative traits: waist-hip-ratio (WHR), body-mass-index (BMI), and high-density lipoprotein (HDL) cholesterol levels.

We conducted single-variant association testes between each trait and ~107 million variants (9.7 common, 41 rare, and 57 very rare) found in the TOPMed-imputed genotypes with imputation $R^2 > 0.1$ and minor allele count (MAC) ≥ 3. Analyses were run using linear mixed models implemented by SAIGE (W. Zhou et al., 2018). We then ran conditional analysis on the association results with P-value < 5×10^{-8} using GCTA (J. Yang et al., 2011) to obtain sets of near-independent, significant variants.

Our empirical analyses served to demonstrate the performance of multiple testing methods in applied real-data setting and to provide empirical association structures to guide our simulations.

4.2.7 Simulation

To evaluate the ability of our Bayesian method to identify true positive discoveries and compare it with the other multiple testing methods, we simulated 20 replicates for each of the traits WHR, BMI, and HDL based on their empirical association structures observed in our empirical analyses.

For each trait, we generate phenotypes for *n* individuals according to the following model:

$$\boldsymbol{Y_i} = \sum_{\boldsymbol{j=1}}^{\boldsymbol{L}} \boldsymbol{X_j \beta_j} + \boldsymbol{\varepsilon_i} \tag{4}$$

for $i = 1, .., n$. Here, X_j is the genotype dosage for the j^{th} causal variant obtained from UKBB TOPMed-imputed genotypes, β_j is the effect size for the j^{th} causal variant obtained from conditional analysis, and $\varepsilon_i \sim N(0, \tau)$ where τ is the proportion of phenotypic variance not explained by the causal variants. For each replicate dataset, we ran single-variant association tests using SAIGE.

4.2.8 Defining true and false positives in simulation

After applying the multiple testing methods to association results in each replicate dataset, we need to know how many of the discoveries made by each method were true and false positives. This process is complicated by the fact that not only do causal variants show up as significant discoveries in our replicate dataset but also variants in LD with them. To distinguish between true and false positives, we create 99% credible sets (The Wellcome Trust Case Control Consortium et al., 2012) for each variant discovery to capture the causal variant responsible for that signal.

Consider a set of R variants declared significant by a given multiple testing method (e.g. P-value threshold). For each variant $j = 1, \dots, R$, we cluster together all other tested variants within ±1Mb of variant j. Within each cluster, we calculate the posterior probability for each of the m_j variants in the cluster as:

$$pp_i = \frac{ABF_i}{\sum_{i=1}^{m_j} ABF_i}$$

where ABF is calculated using (2). We then form a 99% credible set for variant j as the smallest set of variants from the cluster such that 99% of the posterior probability is accounted for. If the credible set for variant j contains a causal variant or a variant with LD $r^2 > 0.80$ with a causal variant, we consider variant j to be a true positive. Conversely, if neither of those two criteria are met, then we consider variant j to be a false positive.

4.2.9 Gene-based tests

Single-variant tests often lack power to detect rare variant associations due to their low allele frequencies (Asimit & Zeggini, 2010) and massive testing burden which requires a stringent significance threshold (Seunggeung Lee et al., 2014). An alternative to testing each variant individually is to combine the effects of multiple variants within a gene or region and test for association between this cumulative effect and a trait of interest. Compared with single-variant tests, these gene- or region-based tests have a smaller testing burden and are more powerful if multiple variants within the gene or region are associated with the trait (Seunggeung Lee et al., 2014). In this section, we describe the extension of the multiple testing methods specifically for gene-based tests, but these methods will be equally applicable to the more general region-based tests.

Suppose that the m variants from model **(1)** are from a single gene γ. For the burden test (Asimit & Zeggini, 2010; B. Li & Leal, 2008), we aggregate all m variants into a single genetic score:

$$C_\gamma = \sum_{i=1}^{m} w_i G_i$$

using weights (w_i) based on MAF and the genotype dosage (G_i). We can then test for association between the aggregate score and a trait. This is equivalent to testing a single null hypothesis $H_0{:}\,\beta = 0$ in **(1)** for no association between the aggregate score and the trait using the score statistic:

$$Q_{burden} = \left(\sum\nolimits_{i=1}^{m_\gamma} w_i S_i\right)^2$$

where S_i is the score statistic from single-variant analysis.

For the sequence kernel association test (SKAT) (Wu et al., 2011), we assume that β_j from **(1)** comes from a distribution with mean 0 and variance $w_j^2\tau$ and test the single null hypothesis $H_0{:}\,\tau = 0$ using a variance component score test:

$$Q_{SKAT} = \sum\nolimits_{i=1}^{m_\gamma} w_i^2 S_i^2$$

The burden test is more powerful than SKAT if the gene contains a high proportion of causal variants with the same direction of effects while SKAT is more powerful if there is a sizeable number of noncausal variants in the gene or if the causal variants have different direction of effects (Seunggeung Lee et al., 2014). SKAT-O (Seunggeun Lee et al., 2012) is an omnibus tests that seeks to maximize power across both scenarios by optimally combing the burden and SKAT test using an adaptive procedure based on the observed data.

4.2.10 Multiple testing for gene-based tests

Typically, we would conduct a set of gene-based tests for each of the L genes in our dataset. Here, we need to correct for multiple testing just like we did for single-variant tests.

To control FWER in the set of L genes at level α, we use the Bonferroni procedure and declare genes with P-values less than α/L to be significant. Similar to single-variant tests, we can increase power by adjusting for the effective number of independent tests L' instead of L. However, past work (D. Lin, 2019) has found only a slight difference between these two number of tests, likely due to weak LD among rare variants and across genes. Thus, it may be preferable to use the total number of genes L in the Bonferroni procedure to avoid the computation cost and limited gain from estimating the number of independent tests L'.

To control FDR among the tested genes, we directly apply the BH and BY procedure described in section 4.2.2 to the P-values from gene-based tests. Since both our testing and analysis are conducted in units of genes, we do not need to apply the modifications proposed by Brzyski et al. (Brzyski et al., 2017) as we did for single-variant testing.

Extending our Bayesian method to gene-based tests, we calculate a gene-level BF (Wilson et al., 2010) to test the null hypothesis of no association between any variants in gene γ and the trait versus the alternative hypothesis of at least one association. The gene-level BF is:

$$BF_\gamma = \frac{1}{m}\sum_{j=1}^{m} BF_j$$

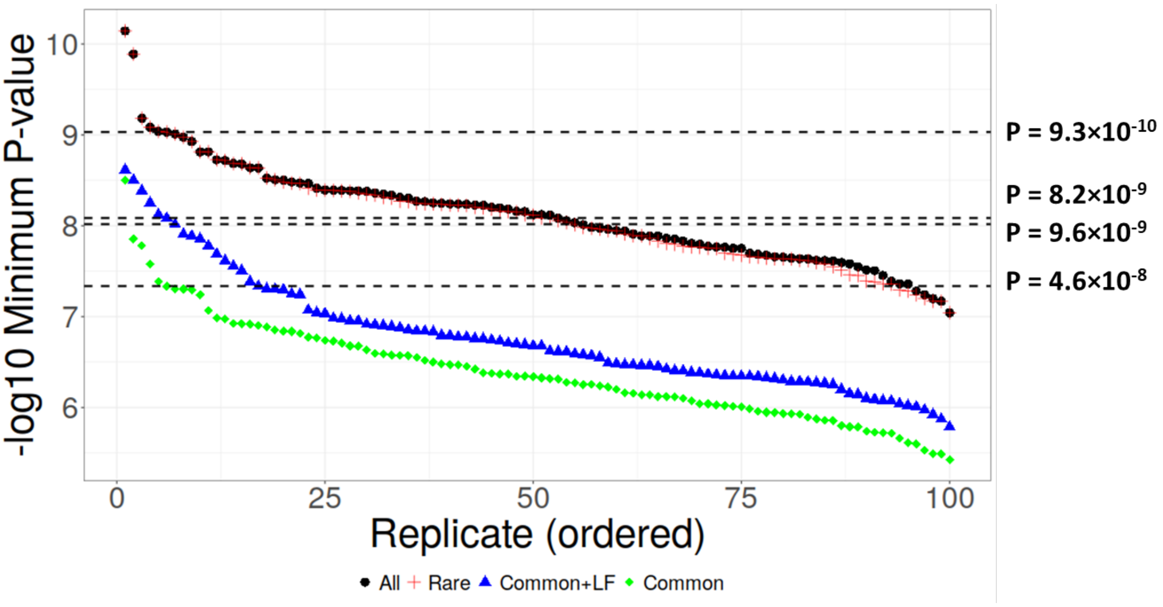
where the BF_j can be approximated using equation **(2)** in the absence of individual-level data. To simplify calculations, we make the exchangeability assumption that every tested gene has the same prior probability π_γ of being associated with the trait and then conservatively estimate π_γ as the proportion of tested genes with P-values less than α/L. This assumption can be relaxed, allowing for different priors among tested genes based on the functional annotations of variants within the genes (H. Yang & Wang, 2015). We then calculate the posterior probability of association for each of the tested genes and control the Bayesian FDR using the procedure described in section 4.2.3.

4.3 Results

4.3.1 Null simulations

To estimate the genome-wide significant P-value threshold for rare variant studies, we simulated association results under the global null hypothesis of no association to any genetic variant using genotypes for ~107 million variants from the n=487,409 individuals in the TOPMed-imputed UKBB dataset and phenotypes randomly generated from a standard normal distribution independent of genotypes. Figure 4.1 shows the distribution of minimum P-values from 100 genome-wide simulations partitioned by MAF. The 5% quantile of the minimum P-values for each MAF bin represents the P-value threshold to control FWER at α=5% for testing variants in that bin.

Figure 4.1: Minimum P-value distribution



Minimum P-value distribution for 100 simulations under the global null hypothesis. Horizontal dotted lines represent the 5% quantile of minimum P-values in each MAF bin. Rare: MAF≤0.5%, Common+LF: MAF>0.5%, Common: MAF>5%.

We estimated a P-value threshold of 9.3×10^{-10} for testing ~107 million total variants with minor allele count (MAC) ≥ 3 in our dataset (Supplementary Table 4.1). This is equivalent to testing 53.7 million independent variants ($0.05/9.3\times10^{-10}$). We also estimated a threshold of 8.2×10^{-9} to test 9.7 million common (MAF > 5%) and low-frequency variants (MAF 0.5-5%), 9.6×10^{-9} to test 3.8 million low-frequency (LF) variants, and 4.6×10^{-8} to test 5.9 million common variants. Interestingly, our estimated threshold to test 98 million rare (MAF < 0.5%) variants was the same as our estimated threshold to test all 107 million variants. This is partly due to a lack of precision in the estimates arising from only simulating 100 null replicates but also is consistent with the fact that the vast majority of the testing burden can be attributed to rare variants.

4.3.2 Non-null simulations

We generated three sets of 20 simulation replicates using model **(2)** to assess the true and false positive rate of the multiple testing methods. Table 4.1 summarizes the empirical association results for the three traits on which we based our three sets of simulation: waist-hip-ratio-based (WHR_{sim}), body-mass-index-based (BMI_{sim}), and high-density lipoprotein cholesterol levels-based

(HDL_{sim}). For each trait we used the conditionally independent variants with MAC ≥ 3 and conditional P-value < 5×10^{-8} as causal variants in our simulation models.

Table 4.1: Description of empirical datasets and association results used to inform simulations

Trait	Sample size	MAF bin	Total # variants (millions)	# conditionally independent significant[1] variants
WHRsim	407,399	Common+LF (>0.5%)	9.7	179
		Rare (0.005-0.5%)	41	20
		Very rare (<0.005%)	57	40
BMI_{sim}	406,860	Common+LF (>0.5%)	9.7	313
		Rare (0.005-0.5%)	41	26
		Very rare (<0.005%)	57	0
HDL_{sim}	356,103	Common+LF (>0.5%)	9.7	384
		Rare (0.005-0.5%)	41	120
		Very rare (<0.005%)	55	0

[1] Variants with conditional P-values<5×10^{-8}

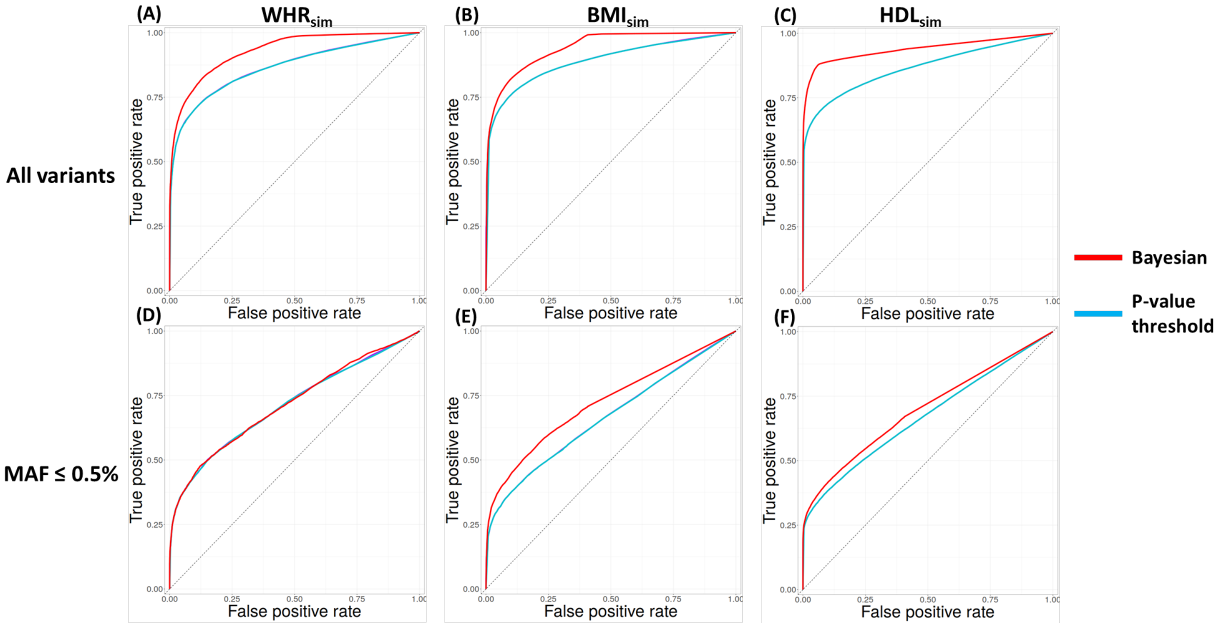
The simulated traits have different genetic architectures with WHR_{sim} having the fewest causal variants (239) and HDL_{sim} having the most (504). In addition, WHR_{sim} is the only simulated trait with causal variants that have MAF < 0.005%.

In the effect size distribution of causal variants for BMI_{sim} and HDL_{sim} (Supplementary Figure 4.1D and 4.1F) we observe that the effect sizes for causal rare variants (blue) are noticeably elevated compared to the effect sizes of causal common+LF variants (red). However, we observe in WHR_{sim} (Supplementary Figure 4.1B) the presence of very rare (MAF < 0.005%) variants (purple) which have much larger effect sizes than the detected rare and common+LF variants. We believe these very rare causal variants should be modeled differently than rare or common+LF variants, hence the mixture of three normal distributions. We observe in the application of our EM algorithm for all three traits that membership into the non-null groups among the tested variants with small (i.e. significant) P-values is largely in accordance with MAF (Supplementary Figure 4.2). Almost all of the common variants are placed into group $k = 1$ with a small estimated prior variance while rare and very rare variants are placed into groups $k = 2$ and $k = 3$ with larger prior variances.

4.3.3 True and false positive rate of multiple testing methods

We compared the abilities of the four multiple testing methods: P-value threshold, BH, BY, and our proposed Bayesian method to accurately classify true and false positive variants in our three sets of simulations. We divide the four methods into two categories: (1) methods that classify true and false positives using posterior probabilities and (2) methods that classify true and false positives using P-values. Our Bayesian method falls into the first category while the other three methods fall into the second. Since methods within each category produce identical true positive rates (TPR) at the same false positive rate (FPR), we only present results for our Bayesian method and the P-value threshold.

Figure 4.2: Classifying true and false positive variants



Receiver operating characteristic (ROC) curves for the Bayesian method and P-value threshold in (A-C) all variants and (D-F) rare+very rare variants (MAF<0.5%) in our simulated datasets based on the waist-hip-ratio (WHR), body-mass-index (BMI), and high-density lipoprotein (HDL) cholesterol level traits.

Figure 4.2 shows the receiver operating characteristic (ROC) curves for our Bayesian method and P-value-based methods. For variants across the full allele frequency spectrum in all three sets of simulations, our Bayesian method identifies more true positives than the P-value threshold at the same false positive rates (Figure 4.2A-C). We can quantify this difference using the area-under-curve (AUC), a measure of discriminatory power. The average AUC for our Bayesian method in WHR_{sim} is 93% compared with 87% for the P-value threshold (Table 4.2). We see a similar pattern for the other two simulated traits where the average AUC for our Bayesian method is again 5-7% higher than that of the P-value threshold.

Table 4.2: AUC values for multiple testing methods

Trait	Method	All variants		MAF < 0.5%	
		AUC	95% CI	AUC	95% CI
WHR_{sim}	P-value	0.866	(0.860, 0.871)	0.713	(0.693, 0.733)
	Bayesian	0.929	(0.924, 0.934)	0.716	(0.695, 0.738)
BMI_{sim}	P-value	0.892	(0.888, 0.896)	0.666	(0.648, 0.684)
	Bayesian	0.942	(0.937,0.947)	0.723	(0.709, 0.738)
HDL_{sim}	P-value	0.867	(0.865, 0.870)	0.671	(0.667, 0.674)
	Bayesian	0.939	(0.938, 0.940)	0.699	(0.694, 0.703)

To investigate this surprisingly large increase in power, we calculated the TPR for both methods at an identical FPR of 1% (Supplementary Table 4.2) and found that our Bayesian method has a more relaxed threshold for discovery compared to the P-value threshold. For example, in WHR_{sim}, the average maximum P-value of the discoveries made by our Bayesian method is 8.1×10^{-8} compared with 2.6×10^{-8} for the P-value threshold. On average, there is a total of 857 variants between these two P-values in the simulated replicates for WHR_{sim} and 850 (99%) of them were true positives. Thus, our Bayesian method was making more discoveries within a set of variants which were almost all true positives, substantially increasing the number of true positives identified compared to the P-value threshold which did not consider any of these variants as significant discoveries. We saw similar patterns for BMI_{sim} and HDL_{sim}. We continue to explore this unexpectedly large difference.

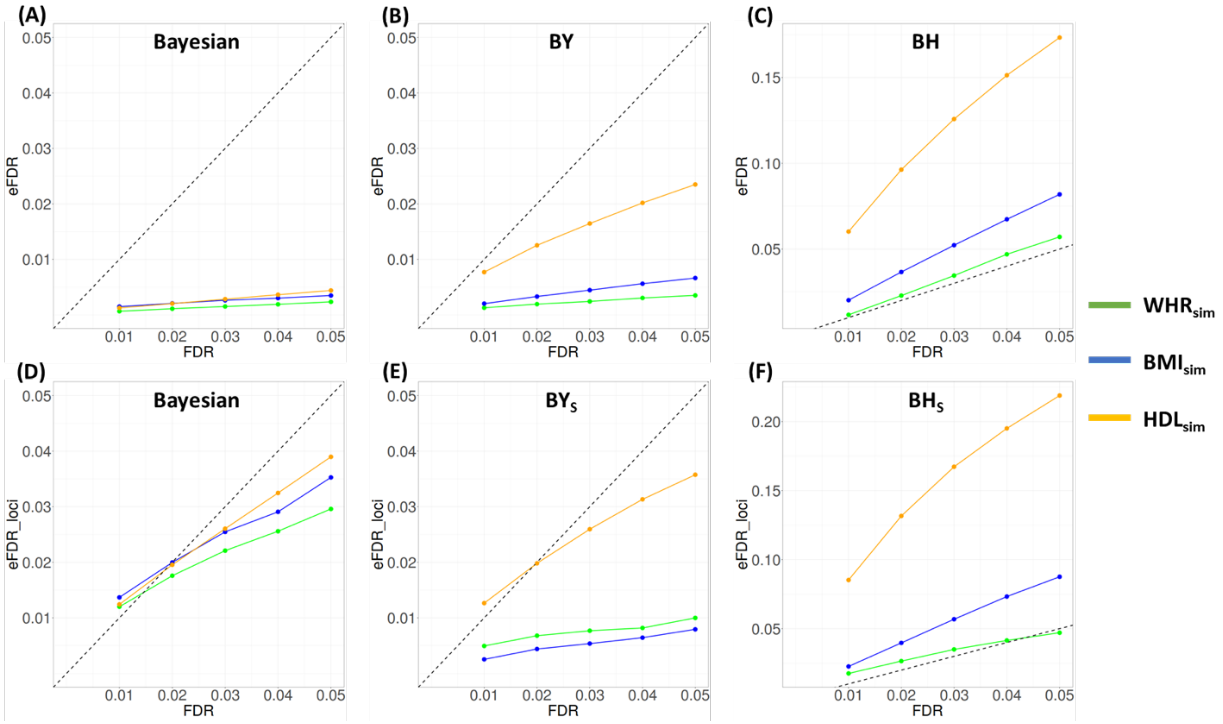
4.3.4 True and false positive rate of multiple testing methods for rare variants

Considering only rare and very rare variants (Figure 4.2D-F), our Bayesian method has approximately the same power as the P-value-based methods for WHR_{sim} but greater power for BMI_{sim} and HDL_{sim}. This can be seen in the average AUC values for our Bayesian method which were similar between the three sets of simulations (70%-72%) while the average AUC for the P-value threshold dropped from 71% in WHR_{sim} to 67% in BMI_{sim} and HDL_{sim} (Table 4.2). This may be due to the larger proportion of rare and very rare false positive variants in the latter two traits. Our Bayesian method limited its discovery of these false positives through smaller estimates of the prior probabilities for these two MAF bins (Supplementary Table 4.3) but the P-value threshold is unable to use information other than the test statistic in classifying (true and false) positives.

4.3.5 FDR control for all discoveries versus a subset of discoveries

As described in Methods, it is important to consider the testing unit when assessing FDR control for the BH and BY procedures. These procedures are designed to control the "global" FDR among all discovered variants but do not guarantee control if the variants are clustered into loci later in the analysis. When we wish to control the FDR among loci, it is more appropriate to use the modified BH and BY procedures (BH_{S} and BY_{S}) since they are applied directly on lead variants from loci.

Figure 4.3: Control of FDR



Control of FDR among (A-C) all discoveries and (D-F) a subset of all discoveries formed using lead variants from discovered loci. For the BH and BY procedures, we show the plot of FDR control in loci for the modified version of the methods, BH_S and BY_S, due to large inflations observed in the classic versions.

Figure 4.3A-C shows the control of global FDR at 1-5% for our Bayesian method, BH, and BY procedures. Both our Bayesian method and BY procedure controlled the empirical FDR (eFDR) at or below the theoretical FDR threshold for all simulated traits, albeit very conservatively. For example, the average eFDR at thresholds 1-5% in WHR_sim is 0.064%-0.23% for our Bayesian method and 0.13%-0.35% for BY. BH showed slightly inflated eFDR for WHR_sim (1.2-5.7%), moderately inflated eFDR for BMI_sim (2.0-8.2%), and very inflated eFDR for HDL_sim (6.0-17%). This is likely due to the correlation between test statistics that violates the PRDS assumption. As the number of causal variants increase, the number of correlated test statistics also increases, resulting in further departure from a positive dependency structure. This likely explains the greater inflation for BMI_sim and HDL_sim than for WHR_sim.

Figure 4.3D-F shows the control of FDR for the lead variants from discovered loci, (a subset of the discovered variants. Both our Bayesian method and the modified BY procedure (BY_S) control the eFDR in loci at theoretical thresholds 1-5% with a slight inflation at 1% for the Bayesian method in all three simulated traits (1.2%, 1.4%, and 1.2%) and for the BY_S procedure in HDL_{sim} (1.3%). Our Bayesian method's control of eFDR is much less conservative in loci (Figure 4.3D) than in all discovered variants (Figure 4.3A). However, BY_S control of eFDR in loci for WHR_{sim} and BMI_{sim} remained conservative (Figure 4.3E). The classic BH and BY procedures did not control the eFDR in loci at any threshold from 1-5% (Supplementary Figure 4.3), consistent with the lack of subsetting property for FDR-controlling procedures.

4.3.6 Empirical analyses

To compare the multiple testing methods in real datasets, we applied the P-value threshold, BY_S, and our Bayesian method to the actual TOPMed-imputed UKBB data for BMI, WHR, and HDL (Table 4.1). We excluded BH, BH_S, and BY from our empirical analyses because of their poor eFDR controls in loci in our simulations.

To estimate the proportion of true positives among the common+LF discoveries in WHR and BMI, we compared our results with those from a meta-analysis (Pulit et al., 2019) of 485,486 individuals from the UKBB study and 212,248 individuals from the Genetic Investigation of Anthropometric traits (GIANT) consortium (combined n=697,734) on the same traits. We considered any of our discoveries that were within the locus (defined as ±1Mb physical distance from the index variant) of a genome-wide significant variant (P-value$<5\times10^{-8}$) in the meta-analysis to be true positives. Since the UKBB study was used in both our empirical analyses and the meta-analysis, there are significant overlaps in samples between the two set of results. However, considering the large sample size differences between the two analyses, we believe the meta-analysis can still serve as an approximate "truth set" for our empirical results. We were unable to repeat this process for the HDL trait and rare/very rare variants due to a lack of comparable truth sets for them.

Table 4.3: Empirical results for selected multiple testing methods

			Conditionally independent variants		
Trait	Method	Threshold	Common+LF MAF>0.5% (Likely TP[1])	Rare MAF 0.005-0.5%	Very rare MAF<0.005%
WHR	P-value	$P=1\times10^{-9}$	102 (102)	4	12
	BY_S	FDR=2%	117 (115)	6	20
	Bayesian	BFDR=2%	162 (159)	14	45
BMI	P-value	$P=1\times10^{-9}$	152 (152)	6	0
	BY_S	FDR=2%	196 (195)	7	0
	Bayesian	BFDR=2%	293 (288)	27	0
HDL	P-value	$P=1\times10^{-9}$	217	56	0
	BY_S	FDR=2%	259	83	0
	Bayesian	BFDR=2%	366	242	0

[1] True positives defined as within ±1Mb physical distance from variant with P-value$<5\times10^{-8}$ in truth set

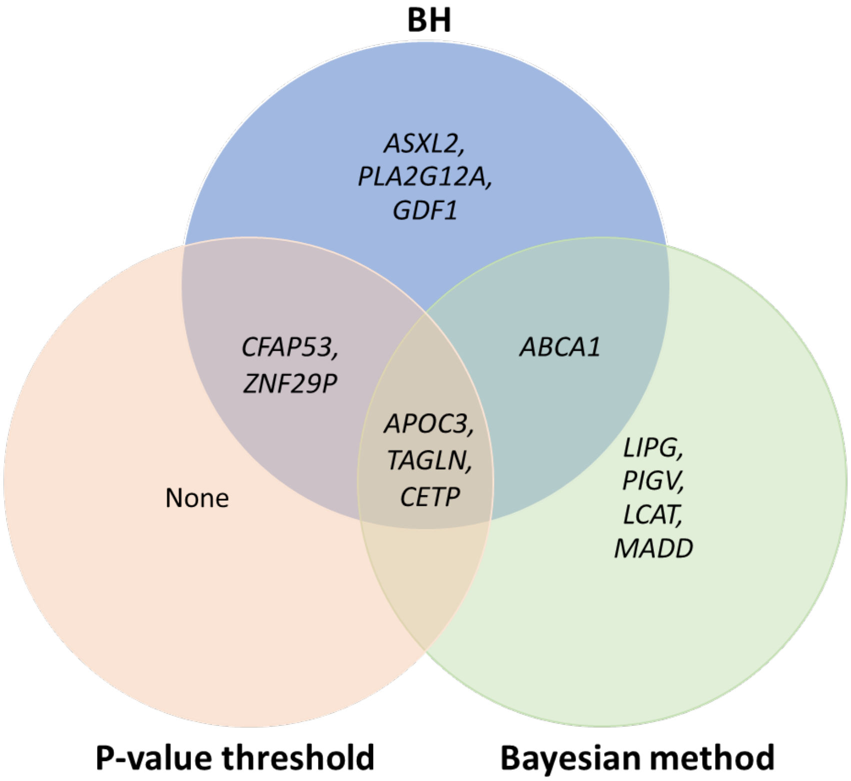
Table 4.3 shows the empirical results with the number of conditionally independent variants discovered by each method partitioned by MAF. All conditionally independent common+LF discoveries made by the P-value threshold for WHR and BMI were found in the truth set. For the BY_S procedure, 115 out of 117 (eFDR=1.7%) common+LF discoveries were found in the truth set for WHR and 195 of 196 (eFDR=0.5%) for BMI. For the Bayesian method, we observed similar levels of eFDR between the two traits for this MAF bin (1.9% for WHR, 1.7% for BMI).

4.3.7 Results for gene-based tests

To assess the performance of the multiple testing methods in gene-based tests, we applied the methods to whole-exome sequence data on 49,960 individuals from the UKBB (Hout et al., 2019). We conducted single-variant and gene-based tests (SKAT-O) on a subset of 36,364 individuals from the UKBB whole-exome dataset with phenotype values on the HDL trait. In total, there are ~5.23 million variants with MAF between 0.0014-1% in the dataset which were grouped into 17,795 genes. Due to the heavy computation cost, we restricted our analysis to the HDL trait.

To correct for multiple testing among the 17.795 tested genes, we used a P-value significance threshold of 2.8×10^{-6} (0.05/17795) and thresholds for BH and the Bayesian method corresponding to theoretical FDR of 5%. A Venn diagram of the discoveries for the P-value threshold, BH procedure, and Bayesian method can be seen in Figure 4.4. We did not include results for the BY procedure because its discoveries were a proper subset of the discoveries made by BH.

Figure 4.4: Results for gene-based tests



Comparison of discoveries made by different multiples testing methods for gene-based tests.

In total, there were 13 unique genes declared significant by the three multiple testing methods (Figure 4.4). *APOC3*, *TAGLN*, and *CETP* were discovered by all three methods and are protein encoding genes that have been found to be associated with HDL in previous studies (Jeong et al., 2014; Luo et al., 2017; Willer et al., 2013). Comparing the two P-value-based methods, we observed that the BH procedure discovered four more genes (*ABCA1*, *ASXL2*, *GDF1*, *PLA2G12A*) than the P-value threshold. *ABCA1* is a protein coding gene that has been previously found to be associated with lipid metabolism in human subjects (Willer et al., 2008) while the other three genes are potential novel discoveries that have only been previously found to affect lipid metabolism (Izawa et al., 2015; Shen et al., 2009) or HDL particle size (Strunz et al., 2020) in mice. Finally, the Bayesian method discovered four genes (*LCAT*, *LIPG*, *MADD*, *PIGV*) that were not discovered by the two P-value-based methods, all of which have been previously shown to be associated with HDL (Stanley et al., 2017; Willer et al., 2013).

4.4 Discussion

In this study, we proposed a Bayesian method that correct for multiple testing in rare variant association studies. We assessed the ability of our Bayesian method to discover true positive associations while controlling the number of false positives in simulated replicates for three traits with different genetic architecture and compared it to using the P-value threshold, and the BH and BY procedures.

4.4.1 Bayesian vs. P-value-based methods for multiple testing

Our proposed Bayesian method controls the proportion of false positives among all discoveries and does not require individual-level nor arbitrary specification of priors. It calculates the posterior probabilities using an approximation of the Bayes factor derived from the test statistic and estimates the prior parameters from observed association results using an E-M algorithm. Across three simulated traits we considered, our Bayesian method better distinguished between true and false positives than the three P-value-based methods: the P-value threshold, and the BH and BY procedures. On average, the AUC for our Bayesian method was 5-7% higher than that of the P-value-based methods among all tested variants and 1-5% higher among rare (MAF 0.5-0.005%) and very rare (MAF < 0.005%) variants. This is likely because our Bayesian method uses extra information about the effect size distribution of tested variants (i.e. prior probabilities of different MAF bins) to improve its true positive discovery. This information is accurate because it is estimated directly from the data in a principled way.

4.4.2 FDR control for genetic loci

In genetic studies, we typically test for association between the trait and each individual variant but analyze our results in clusters of correlated variants called loci. This means that FDR-controlling methods such as the Bayesian method, and the BH and BY procedures need to control the eFDR in both the full set of discovered variants and, more importantly, in a subset composed of lead variants from discovered loci. We found that only the Bayesian method and the modified BY procedure (BY_S) controlled the eFDR in both sets of discoveries at theoretical thresholds of 1-5% for all three simulated traits; the classic BH, BY, and modified BH (BH_S) procedures showed inflated eFDR ranging from slight inflation for WHR_{sim} to heavy inflation for HDL_{sim}.

4.4.3 Stratified FDR method

There are similarities between our proposed Bayesian method and the stratified FDR procedure (Sun et al., 2006) that divides all tested variants into strata and estimate a different proportion of null hypothesis (π_0) for each stratum. These strata can be defined based on external information (e.g. MAF) and the estimated π_0's are then incorporated into the BH or BY procedure applied separately to each stratum. If the π_0's are truly variable across the different strata, then the stratified FDR procedure will have increased power to detect true associations compared with the non-stratified approach (ChangJiang Xu et al., 2014). However, the stratified FDR procedure is designed to control FDR in the set of all discovered variants but does not guarantee the control will hold for the subset of discovered loci. As we have shown in our results, this is an important property that is required for FDR-controlling procedures in GWAS. It may be possible to modify the stratified FDR procedure similar to BH_S (Brzyski et al., 2017) but such adjustment is beyond the scope of this work.

4.4.4 Genome-wide significant P-value threshold

We used null simulations to estimate a genome-wide significant P-value threshold for testing our set of ~107 million (98 million with MAF < 0.5%) variants. Although previous work (D. Lin, 2019; Pulit et al., 2017) have estimated this threshold to be 5×10^{-9} for testing ~27 million variants found in the imputation reference panel from the 1000 Genomes phase 3 study (The 1000 Genomes Project Consortium et al., 2015), we have a much larger testing burden in our dataset which requires a more stringent threshold. Indeed, we estimated a P-value threshold of 1×10^{-9} for testing 107 million variants from the TOPMed imputation reference panel (Taliun et al., 2019), equivalent to 53.7 million independent tests. As expected, we found our testing burden to be mainly attributed to the large number of rare and very rare variants.

In addition, we estimated a P-value threshold of 5×10^{-8} for testing 5.9 million common variants, equivalent to 1.1 million independent tests. Compared with common variants, the smaller difference between the total number of tests and estimated number of independent tests for rare variants suggests a lower impact of LD on rare variants in our dataset.

4.4.5 Choosing between multiple testing methods

In this study, we presented our assessment of four different multiple testing methods for three simulated traits. When choosing between the methods, it is important to consider the goal of the study. If the goal is to generate a carefully curated list of loci for expensive downstream analyses (e.g. building animal models), then the P-value threshold is the most appropriate multiple testing method due to its stringent control of even a single false positive discovery. However, if the goal is to instead generate a large number of true positive loci for less expensive, high-throughput downstream analyses (e.g. high throughput bioinformatics or functional follow-up) while controlling the proportion of false discoveries, then FDR-controlling methods may be more appropriate. In this case, our Bayesian method using posterior probabilities demonstrated better ability to discover true positives than the BH or BY procedures that rely on P-values while also controlling the eFDR among discovered loci at FDR thresholds of 1-5%. The Bayesian method may be particularly attractive for testing rare variants where there are likely to be a large number of true signals but P-value thresholds lack power to detect them due to low allele counts.

4.4.6 Limitations

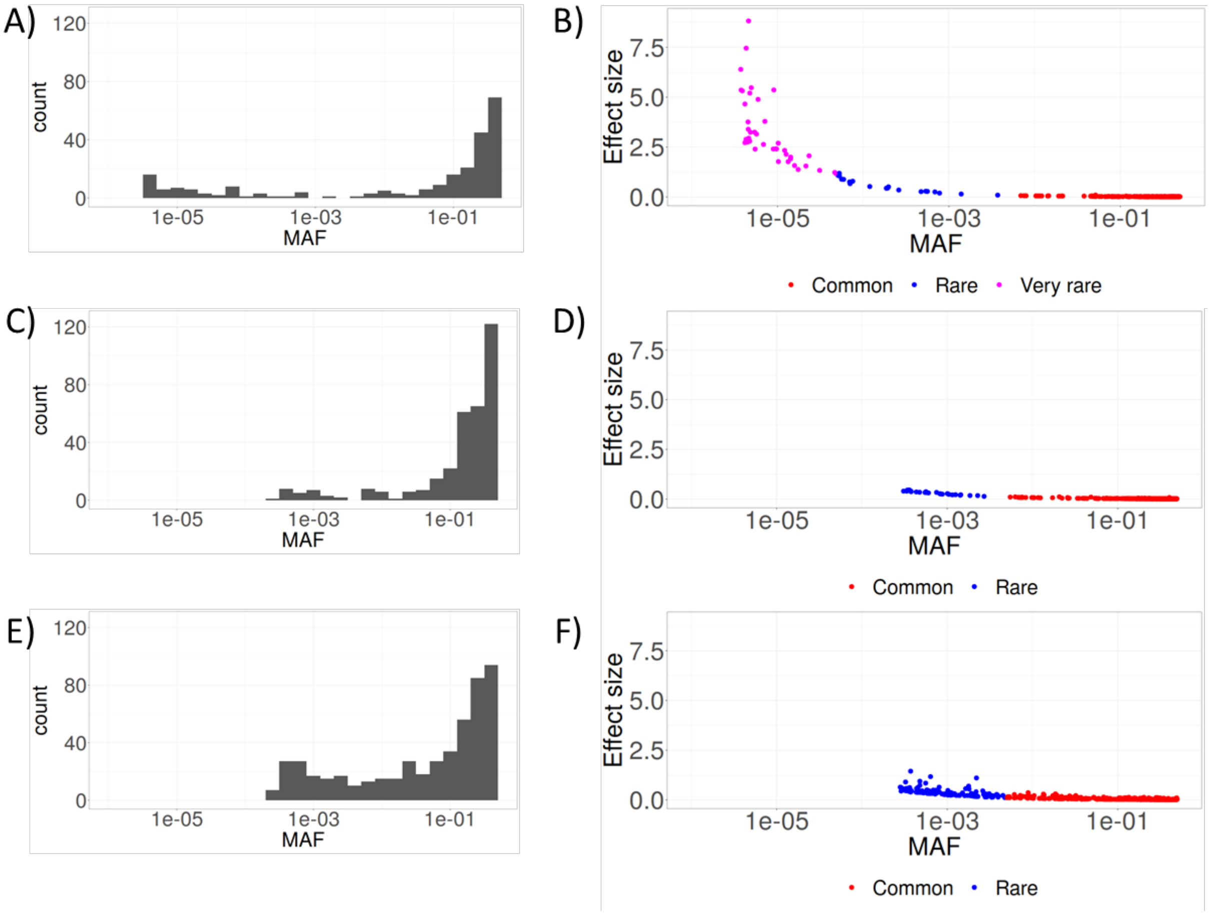
In this study, we did not model the LD between tested variants in our E-M algorithm for the Bayesian method. This is due to the heavy computation cost of estimating the LD structure for ~107 million variants in our dataset. As a results, groups of significant variants correlated with the causal variant can inflate the estimated proportion of causal variants for common and low-frequency variants which will affect the performance of the Bayesian method for those MAF bins. This inflation is likely mitigated for rare variants due to weak LD between those variants. An alternative to estimating the LD structure is to use a LD-pruned set of independent variants for the E-M algorithm. The effect of this choice on the estimated priors for the Bayesian method requires further exploration.

4.4.7 Summary

We showed that our Bayesian method have more true positive discoveries than other multiple testing methods at similar false positive rates while maintaining control of FDR. We estimated a genome-wide significant P-value threshold of 1×10^{-9} to test ~107 million variants from the TOPMed imputation panel which is 5× more stringent than the currently used 5×10^{-9} threshold for testing ~27 million variants from the 1000 Genomes phase 3 imputation panel.

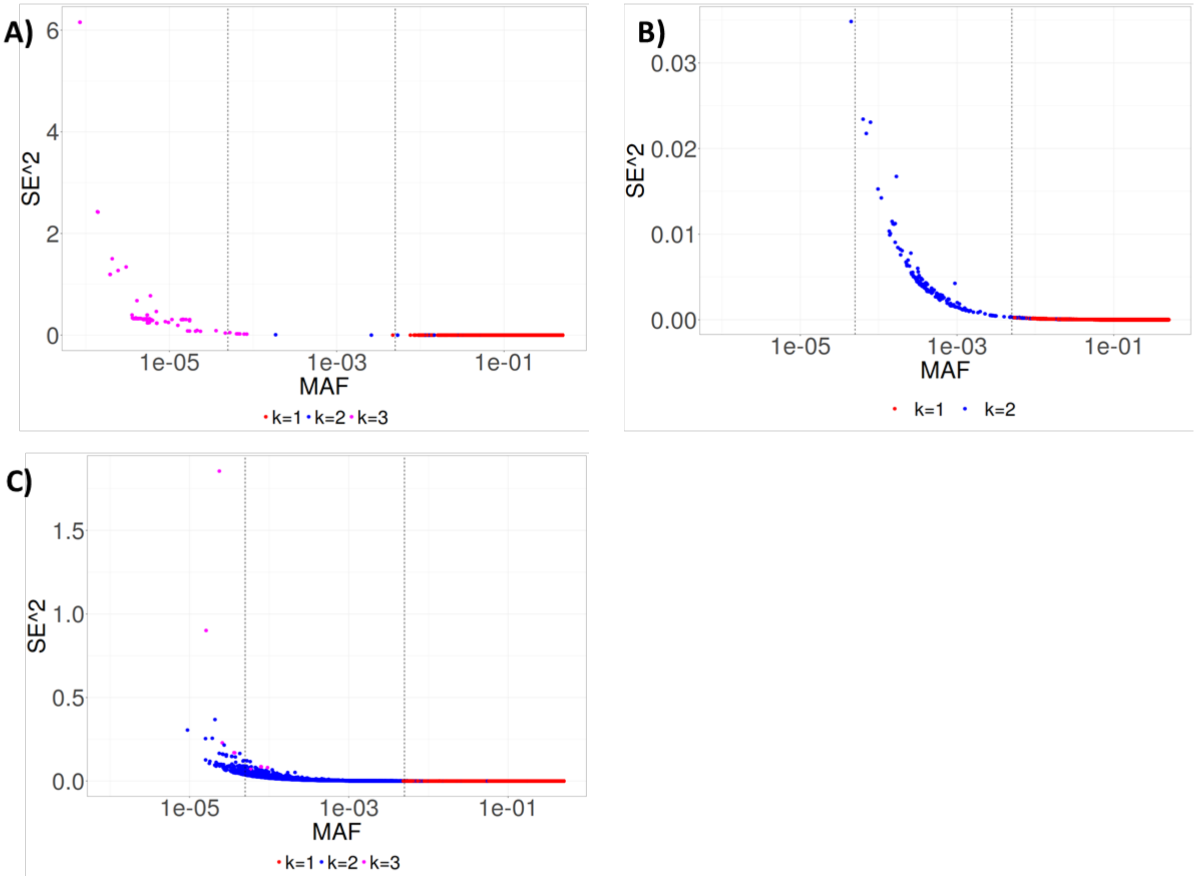
4.5 Supplementary Figures

Supplementary Figure 4.1: Allele frequency and effect size distributions for empirical dataset



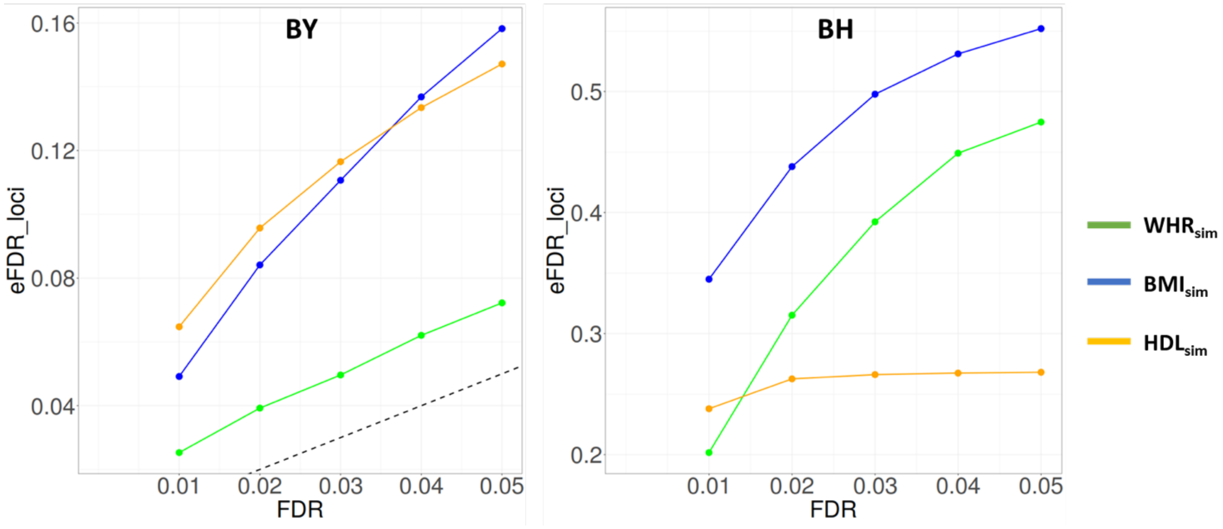
Minor allele frequency (MAF) and effect size distributions for (A-B) waist-hip-ratio, (C-D) body-mass-index, and (E-F) high-density lipoprotein (HDL) cholesterol levels in real data from the UK Biobank study with TOPMed-imputed genotypes.

Supplementary Figure 4.2: Non-null group membership for significant variants in simulation



Membership into non-null groups $k = 1, 2, 3$ for significant variants as determined by E-M algorithm for simulated traits (A) waist-hip-ratio, (B) body-mass-index, and (c) high-density lipoprotein cholesterol levels. Vertical dotted lines denote MAF cutoffs for common+LF (MAF > 0.5%), rare (MAF 0.005-0.5%), and very rare (MAF < 0.005%) varaints.

Supplementary Figure 4.3: Control of eFDR in loci for the classic BH and BY procedures.



Control of eFDR in discovered loci at thresholds 1-5% for simulated traits.

4.6 Supplementary Tables

Supplementary Table 4.1: Estimated genome-wide significant P-value thresholds from 100 null simulation

MAF bin	P_{GWS} at α=5%	Total # of variants (millions)	Total # of independent tests (millions)
All	9.3×10^{-10}	107	53.7
Rare only (<0.5%)	9.3×10^{-10}	97.7	53.7
Common+LF(>0.5%)	8.2×10^{-9}	9.68	6.10
Common only (>5%)	4.6×10^{-8}	5.92	1.09

Supplementary Table 4.2: Numbers of true and false positives identified at false positive of 1%

Trait	Method	FP	TP	Max P-value	Total variants between max P-values	
					FP	TP (%)
WHR_{sim}	P-value	6.2	4822	2.6×10^{-8}	-	-
	Bayesian	6.3	5574	8.1×10^{-8}	6.9	850 (99%)
BMI_{sim}	P-value	10	19823	2.4×10^{-8}	-	-
	Bayesian	11	20897	7.7×10^{-8}	19	2065 (99%)
HDL_{sim}	P-value	32	30080	8.7×10^{-9}	-	-
	Bayesian	32	37819	6.4×10^{-7}	661	10387 (94%)

Supplementary Table 4.3: E-M estimates of prior probabilities for simulated traits

Trait	MAF bin	E-M estimates of prior probability (95% CI)	Conservative[1] estimates of prior probability (95% CI)
WHR_{sim}	Common+LF (>0.5%)	$1.7\ (1.7\text{-}1.8)\times10^{-4}$	$3.3\ (3.2\text{-}3.4)\times10^{-5}$
	Rare (0.005-0.5%)	$0.99\ (0.69\text{-}1.3)\times10^{-6}$	$7.5\ (5.4\text{-}9.6)\times10^{-8}$
	Very rare (<0.005%)	$0.95\ (0.75\text{-}1.2)\times10^{-6}$	$3.0\ (2.1\text{-}3.8)\times10^{-7}$
BMI_{sim}	Common+LF (>0.5%)	$6.6\ (6.5\text{-}6.7)\times10^{-4}$	$1.5\ (1.5\text{-}1.5)\times10^{-4}$
	Rare (0.005-0.5%)	$1.7\ (1.6\text{-}1.9)\times10^{-5}$	$7.2\ (5.6\text{-}8.8)\times10^{-7}$
	Very rare (<0.005%)	$0.91\ (0\text{-}2.7)\times10^{-9}$	$1.4\ (0\text{-}2.9)\times10^{-9}$
HDL_{sim}	Common+LF (>0.5%)	$5.5\ (5.4\text{-}5.6)\times10^{-4}$	$2.3\ (2.3\text{-}2.3)\times10^{-4}$
	Rare (0.005-0.5%)	$1.9\ (1.9\text{-}2.0)\times10^{-4}$	$1.9\ (1.9\text{-}1.9)\times10^{-5}$
	Very rare (<0.005%)	$3.2\ (2.7\text{-}3.7)\times10^{-7}$	$6.5\ (6.0\text{-}7.0)\times10^{-7}$

Conservative estimate = {# of variants with P-value<1×10^{-9}}/{Total # of tested variants}

Chapter 5

Discussion

5.1 Summary

In this dissertation, I provided recommendations on aggregating sequence data from multiple studies, assessed the performance of different multiple testing methods in common variant GWAS leveraging sequential meta-analyses to approximate a truth set, and developed a Bayesian method for multiple testing correction in rare variant studies. Here, I review these projects, discuss their limitations, and provide directions for future research.

5.2 Viability of single-study variant calling strategy for rare variants

In Chapter 2, we compared the gold standard joint variant calling strategy to the more computationally efficient single-study strategy in terms of variant detection, genotype accuracy, and power of downstream association analyses. Due to computation time and burden, we limited our study to variants in chromosome 2. Additional variants from analyzing more chromosomes would be helpful in comparing association power between joint and single-study strategies for genome-wide significant (P-value$<5\times10^{-8}$) rare variants. Currently, our single-variant and gene-based analysis of rare variants are centered on those with P-values$\geq5\times10^{-5}$ with limited information on rare variants near the genome-wide significance threshold. Further analyses in sequence studies with larger sample sizes can reveal whether the single-study strategy maintains similar power to detect significant rare association signals as the joint calling strategy.

5.3 Evaluation of multiple testing methods in real datasets

In Chapter 3, we assessed the performance of the several multiple testing methods in terms of true and false positive discovery for array-based common variant GWAS. In our empirical analyses, we defined true and false positive discoveries using the largest, most recent common variant GWAS which we called the truth sets. Although this did not guarantee the (unknown) list of loci truly associated with each tested trait, the truth sets served as reasonable approximations when there were considerable sample size differences between the truth and test sets. However, some bias likely remains in our analysis due to sample overlap between our truth and test sets. This can be mitigated in future analyses by using truth sets from independent studies on the same traits.

Our empirical analyses in Chapter 3 serves as a template for assessing multiple testing methods in an applied real-data setting. Although we conducted extension simulation in Chapter 4, we were unable to verify the performance of our Bayesian method (as well as the other multiple testing methods) for rare variants in an applied setting due to a lack of truth sets for the TOPMed-imputed UKBB datasets. Ongoing projects by the Global Lipids Genetic Consortium (GLGC) to study variant associations with blood lipid levels in whole-genomes sequencing and large imputation-based cohorts serve as useful truth sets for our high-density lipoprotein (HDL) cholesterol level trait.

5.4 Alternative models for effect size distribution

In Chapter 4, we proposed a Bayesian method for multiple testing that estimated the prior parameters from summary statistics using an E-M algorithm. A key component of this method is accurately modeling the true effect size distribution of our tested variants. There has been extensive work in this area for polygenic risk prediction models (Chatterjee et al., 2016) that assume the true effect size distribution is a mixture of normal distributions with (Guan & Stephens, 2011; Moser et al., 2015) and without (X. Zhou et al., 2013) a point mass at 0, a double exponential distribution (Yi & Xu, 2008), or generalization to a normal exponential gamma distribution (Hoggart et al., 2008). In addition, GENESIS (Zhang et al., 2018) is a publicly available software that estimates the effect size distribution from GWAS summary statistics. This approach is similar to our E-M algorithm but GENESIS is limited to analyzing common variants from the HapMap3 reference panel.

For our Bayesian method, we chose a mixture model of three normal distributions along with a point mass at 0 to represent the observed effect size distribution in our empirical analyses where the magnitude of effect differed for three distinct groups based on MAF: common (MAF>5%) and low-frequency (MAF 0.5-5%), rare (MAF 0.005-0.5%), and very rare (MAF<0.005%) variants. In addition, the choice of normal distributions aligns well with the approximate Bayes factor (Wakefield, 2007a) used in our method and simplifies calculation of the prior variance of variant effect. Although our Bayesian method showed strong performances using the selected model, there needs to be consideration of alternative models with different numbers of MAF bins and different types of distributions. The goal is for the E-M algorithm to accurately estimate the prior parameters across traits with different levels of polygenicity. We plan to compare the performance of our Bayesian method for different prior effect size models in both simulated and empirical datasets.

5.5 Closing remarks

Genetic association studies are the first step in the long journey from discovering associated variants or loci to identifying potentially causal genes and finally to understanding the biological mechanism or pathway in which genetics affect diseases or traits. At a time when new sequencing technology and larger imputation panels allow us unprecedented access to rare variant genotypes, we must balance the need to limit costly false discoveries with the need to promote true discoveries in order to better understand our genetic makeup. To do this, we must continue to develop new statistical methods and update existing ones to deal with the challenges in analyzing a rapidly increasing amount of genetic data.

References

Altshuler, D., Donnelly, P., & The International HapMap Consortium. (2005). A haplotype map of the human genome. *Nature*, *437*(7063), 1299–1320. https://doi.org/10.1038/nature04226

Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, *567*(7748), 305. https://doi.org/10.1038/d41586-019-00857-9

Asimit, J., & Zeggini, E. (2010). Rare Variant Association Analysis Methods for Complex Traits. *Annual Review of Genetics*, *44*(1), 293–308. https://doi.org/10.1146/annurev-genet-102209-163421

Auer, P. L., Reiner, A. P., Wang, G., Kang, H. M., Abecasis, G. R., Altshuler, D., Bamshad, M. J., Nickerson, D. A., Tracy, R. P., Rich, S. S., NHLBI GO Exome Sequencing Project, & Leal, S. M. (2016). Guidelines for Large-Scale Sequence-Based Complex Trait Association Studies: Lessons Learned from the NHLBI Exome Sequencing Project. *American Journal of Human Genetics*, *99*(4), 791–801. https://doi.org/10.1016/j.ajhg.2016.08.012

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., … Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10. https://doi.org/10.1038/s41562-017-0189-z

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300. JSTOR.

Benjamini, Y., & Yekutieli, D. (2001a). The Control of the False Discovery Rate in Multiple Testing under Dependency. *The Annals of Statistics*, *29*(4), 1165–1188. JSTOR.

Benjamini, Y., & Yekutieli, D. (2001b). The Control of the False Discovery Rate in Multiple Testing under Dependency. *The Annals of Statistics*, *29*(4), 1165–1188.

Berger, J. O., & Sellke, T. (1987). Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence. *Journal of the American Statistical Association*, *82*(397), 112–122. https://doi.org/10.1080/01621459.1987.10478397

Bogdan, M., Ghosh, J. K., & Tokdar, S. T. (2008). A comparison of the Benjamini-Hochberg procedure with some Bayesian rules for multiple testing. *ArXiv:0805.2479 [Math, Stat]*, 211–230. https://doi.org/10.1214/193940307000000158

Bolormaa, S., Pryce, J. E., Reverter, A., Zhang, Y., Barendse, W., Kemper, K., Tier, B., Savin, K., Hayes, B. J., & Goddard, M. E. (2014). A Multi-Trait, Meta-analysis for Detecting Pleiotropic Polymorphisms for Stature, Fatness and Reproduction in Beef Cattle. *PLOS Genetics*, *10*(3), e1004198. https://doi.org/10.1371/journal.pgen.1004198

Browning, B. L., & Yu, Z. (2009). Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. *American Journal of Human Genetics*, *85*(6), 847–861. https://doi.org/10.1016/j.ajhg.2009.11.004

Brzyski, D., Peterson, C. B., Sobczyk, P., Candès, E. J., Bogdan, M., & Sabatti, C. (2017). Controlling the Rate of GWAS False Discoveries. *Genetics*, *205*(1), 61–75. https://doi.org/10.1534/genetics.116.193987

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., Suveges, D., Vrousgou, O., Whetzel, P. L., Amode, R., Guillen, J. A., Riat, H. S., Trevanion, S. J., Hall, P., Junkins, H., … Parkinson, H. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, *47*(D1), D1005–D1012. https://doi.org/10.1093/nar/gky1120

Burdick, J. T., Chen, W.-M., Abecasis, G. R., & Cheung, V. G. (2006). In silico method for inferring genotypes in pedigrees. *Nature Genetics*, *38*(9), 1002–1004. https://doi.org/10.1038/ng1863

Burton, P. R., Hansell, A. L., Fortier, I., Manolio, T. A., Khoury, M. J., Little, J., & Elliott, P. (2009). Size matters: Just how big is BIG? *International Journal of Epidemiology*, *38*(1), 263–273. https://doi.org/10.1093/ije/dyn147

Chatterjee, N., Shi, J., & García-Closas, M. (2016). Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nature Reviews. Genetics*, *17*(7), 392–406. https://doi.org/10.1038/nrg.2016.27

DePristo, M. A., Banks, E., Poplin, R. E., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., & Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–498. https://doi.org/10.1038/ng.806

Diggle, P., Liang, K.-Y., Heagerty, P. J., & Zeger, S. (2002). *Analysis of Longitudinal Data*. OUP Oxford.

Dudbridge, F., & Gusnanto, A. (2008a). Estimation of significance thresholds for genomewide association scans. *Genetic Epidemiology*, *32*(3), 227–234. https://doi.org/10.1002/gepi.20297

Dudbridge, F., & Gusnanto, A. (2008b). Estimation of significance thresholds for genomewide association scans. *Genetic Epidemiology*, *32*(3), 227–234. https://doi.org/10.1002/gepi.20297

Efron, B., Tibshirani, R., Storey, J. D., & Tusher, V. (2001). Empirical Bayes Analysis of a Microarray Experiment. *Journal of the American Statistical Association*, *96*(456), 1151–1160. https://doi.org/10.1198/016214501753382129

Exome Aggregation Consortium, Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., Tukiainen, T., Birnbaum, D. P., Kosmicki, J. A., Duncan, L. E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., … MacArthur, D. G. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *536*(7616), 285–291. https://doi.org/10.1038/nature19057

Fox, E. J., Reid-Bayliss, K. S., Emond, M. J., & Loeb, L. A. (2014). Accuracy of Next Generation Sequencing Platforms. *Next Generation, Sequencing & Applications*, *1*. https://doi.org/10.4172/jngsa.1000106

Fritsche, L. G., Beesley, L. J., VandeHaar, P., Peng, R. B., Salvatore, M., Zawistowski, M., Taliun, S. A. G., Das, S., LeFaive, J., Kaleba, E. O., Klumpner, T. T., Moser, S. E., Blanc, V. M., Brummett, C. M., Kheterpal, S., Abecasis, G. R., Gruber, S. B., & Mukherjee, B. (2019). Exploring various polygenic risk scores for skin cancer in the phenomes of the Michigan genomics initiative and the UK Biobank with a visual catalog: PRSWeb. *PLOS Genetics*, *15*(6), e1008202. https://doi.org/10.1371/journal.pgen.1008202

Fuchsberger, C., Flannick, J., Teslovich, T. M., Mahajan, A., Agarwala, V., Gaulton, K. J., Ma, C., Fontanillas, P., Moutsianas, L., McCarthy, D. J., Rivas, M. A., Perry, J. R. B., Sim, X., Blackwell, T. W., Robertson, N. R., Rayner, N. W., Cingolani, P., Locke, A. E., Tajes, J. F., … McCarthy, M. I. (2016). The genetic architecture of type 2 diabetes. *Nature*, *536*(7614), 41–47. https://doi.org/10.1038/nature18642

Goeman, J. J., & Solari, A. (2014). Multiple hypothesis testing in genomics. *Statistics in Medicine*, *33*(11), 1946–1978. https://doi.org/10.1002/sim.6082

Guan, Y., & Stephens, M. (2011). Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *The Annals of Applied Statistics*, *5*(3), 1780–1815. https://doi.org/10.1214/11-AOAS455

Held, L., & Ott, M. (2018). On P-Values and Bayes Factors. *Annual Review of Statistics and Its Application*, *5*(1), 393–419. https://doi.org/10.1146/annurev-statistics-031017-100307

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

Hoggart, C. J., Whittaker, J. C., De Iorio, M., & Balding, D. J. (2008). Simultaneous Analysis of All SNPs in Genome-Wide and Re-Sequencing Association Studies. *PLoS Genetics*, *4*(7). https://doi.org/10.1371/journal.pgen.1000130

Hout, C. V. V., Tachmazidou, I., Backman, J. D., Hoffman, J. X., Ye, B., Pandey, A. K., Gonzaga-Jauregui, C., Khalid, S., Liu, D., Banerjee, N., Li, A. H., Colm, O., Marcketta, A., Staples, J., Schurmann, C., Hawes, A., Maxwell, E., Barnard, L., Lopez, A., … Center, on behalf of the R. G. (2019). Whole exome sequencing and characterization of coding variation in 49,960 individuals in the UK Biobank. *BioRxiv*, 572347. https://doi.org/10.1101/572347

International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. https://doi.org/10.1038/35057062

Izawa, T., Rohatgi, N., Fukunaga, T., Wang, Q.-T., Silva, M. J., Gardner, M. J., McDaniel, M. L., Abumrad, N. A., Semenkovich, C. F., Teitelbaum, S. L., & Zou, W. (2015). ASXL2 regulates glucose, lipid and skeletal homeostasis. *Cell Reports*, *11*(10), 1625–1637. https://doi.org/10.1016/j.celrep.2015.05.019

Jeong, S. W., Chung, M., Park, S.-J., Cho, S. B., & Hong, K.-W. (2014). Genome-Wide Association Study of Metabolic Syndrome in Koreans. *Genomics & Informatics*, *12*(4), 187–194. https://doi.org/10.5808/GI.2014.12.4.187

Jiang, W., Chen, S.-Y., Wang, H., Li, D.-Z., & Wiens, J. J. (2014). Should genes with missing data be excluded from phylogenetic analyses? *Molecular Phylogenetics and Evolution*, *80*, 308–318. https://doi.org/10.1016/j.ympev.2014.08.006

Johnson, V. E. (2005). Bayes Factors Based on Test Statistics. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *67*(5), 689–701. JSTOR.

Jun, G., Wing, M. K., Abecasis, G. R., & Kang, H. M. (2015). An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. *Genome Research*, *25*(6), 918–925. https://doi.org/10.1101/gr.176552.114

Kathiresan, S., Willer, C. J., Peloso, G. M., Demissie, S., Musunuru, K., Schadt, E. E., Kaplan, L., Bennett, D., Li, Y., Tanaka, T., Voight, B. F., Bonnycastle, L. L., Jackson, A. U., Crawford, G., Surti, A., Guiducci, C., Burtt, N. P., Parish, S., Clarke, R., … Cupples, L. A. (2009). Common variants at 30 loci contribute to polygenic dyslipidemia. *Nature Genetics*, *41*(1), 56–65. https://doi.org/10.1038/ng.291

Lango Allen, H., Estrada, K., Lettre, G., Berndt, S. I., Weedon, M. N., Rivadeneira, F., Willer, C. J., Jackson, A. U., Vedantam, S., Raychaudhuri, S., Ferreira, T., Wood, A. R., Weyant, R. J., Segrè, A. V., Speliotes, E. K., Wheeler, E., Soranzo, N., Park, J.-H., Yang, J., … Hirschhorn, J. N. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, *467*(7317), 832–838. https://doi.org/10.1038/nature09410

Lee, Seunggeun, Emond, M. J., Bamshad, M. J., Barnes, K. C., Rieder, M. J., Nickerson, D. A., Christiani, D. C., Wurfel, M. M., & Lin, X. (2012). Optimal Unified Approach for Rare-Variant Association Testing with Application to Small-Sample Case-Control Whole-Exome Sequencing Studies. *American Journal of Human Genetics*, *91*(2), 224–237. https://doi.org/10.1016/j.ajhg.2012.06.007

Lee, Seunggeun, Teslovich, T. M., Boehnke, M., & Lin, X. (2013). General Framework for Meta-analysis of Rare Variants in Sequencing Association Studies. *American Journal of Human Genetics*, *93*(1), 42–53. https://doi.org/10.1016/j.ajhg.2013.05.010

Lee, Seunggeung, Abecasis, G. R., Boehnke, M., & Lin, X. (2014). Rare-Variant Association Analysis: Study Designs and Statistical Tests. *American Journal of Human Genetics*, *95*(1), 5–23. https://doi.org/10.1016/j.ajhg.2014.06.009

Li, B., & Leal, S. M. (2008). Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data. *American Journal of Human Genetics*, *83*(3), 311–321. https://doi.org/10.1016/j.ajhg.2008.06.024

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, Y., Sidore, C., Kang, H. M., Boehnke, M., & Abecasis, G. R. (2011). Low-coverage sequencing: Implications for design of complex trait association studies. *Genome Research*, *21*(6), 940–951. https://doi.org/10.1101/gr.117259.110

Li, Y., Willer, C. J., Ding, J., Scheet, P., & Abecasis, G. R. (2010). MaCH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology*, *34*(8), 816–834. https://doi.org/10.1002/gepi.20533

Li, Y., Willer, C., Sanna, S., & Abecasis, G. (2009a). Genotype imputation. *Annual Review of Genomics and Human Genetics*, *10*, 387–406. https://doi.org/10.1146/annurev.genom.9.081307.164242

Li, Y., Willer, C., Sanna, S., & Abecasis, G. (2009b). Genotype Imputation. *Annual Review of Genomics and Human Genetics*, *10*(1), 387–406. https://doi.org/10.1146/annurev.genom.9.081307.164242

Lin, D. (2019). A simple and accurate method to determine genomewide significance for association tests in sequencing studies. *Genetic Epidemiology*, *43*(4), 365–372. https://doi.org/10.1002/gepi.22183

Lin, D. Y., & Zeng, D. (2010). Meta-analysis of genome-wide association studies: No efficiency gain in using individual participant data. *Genetic Epidemiology*, *34*(1), 60–66. https://doi.org/10.1002/gepi.20435

Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., Day, F. R., Powell, C., Vedantam, S., Buchkovich, M. L., Yang, J., Croteau-Chonka, D. C., Esko, T., Fall, T., Ferreira, T., Gustafsson, S., Kutalik, Z., Luan, J., Mägi, R., Randall, J. C., … Speliotes, E. K. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature*, *518*(7538), 197–206. PubMed. https://doi.org/10.1038/nature14177

Locke, A. E., Steinberg, K. M., Chiang, C. W., Service, S. K., Havulinna, A. S., Stell, L., Pirinen, M., Abel, H. J., Chiang, C. C., Fulton, R. S., Jackson, A. U., Kang, C. J., Kanchi, K. L., Koboldt, D. C., Larson, D. E., Nelson, J., Nicholas, T. J., Pietilä, A., Ramensky, V., … Freimer, N. B. (2019). Exome sequencing identifies high-impact trait-associated alleles enriched in Finns. *BioRxiv*, 464255. https://doi.org/10.1101/464255

Luo, M., Liu, A., Wang, S., Wang, T., Hu, D., Wu, S., & Peng, D. (2017). ApoCIII enrichment in HDL impairs HDL-mediated cholesterol efflux capacity. *Scientific Reports*, *7*. https://doi.org/10.1038/s41598-017-02601-7

Mägi, R., Horikoshi, M., Sofer, T., Mahajan, A., Kitajima, H., Franceschini, N., McCarthy, M. I., COGENT-Kidney Consortium, T2D-GENES Consortium, & Morris, A. P. (2017). Trans-ethnic meta-regression of genome-wide association studies accounting for ancestry increases power for discovery and improves fine-mapping resolution. *Human Molecular Genetics*, *26*(18), 3639–3650. https://doi.org/10.1093/hmg/ddx280

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., & Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. *Nature Reviews Genetics*, *9*(5), 356–369. https://doi.org/10.1038/nrg2344

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., Kang, H. M., Fuchsberger, C., Danecek, P., Sharp, K., Luo, Y., Sidore, C., Kwong, A., Timpson, N., Koskinen, S., Vrieze, S., Scott, L. J., Zhang, H., Mahajan, A., … Haplotype Reference Consortium. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*, *48*(10), 1279–1283. https://doi.org/10.1038/ng.3643

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297. https://doi.org/10.1101/gr.107524.110

Morris, A. P., Voight, B. F., Teslovich, T. M., Ferreira, T., Segrè, A. V., Steinthorsdottir, V., Strawbridge, R. J., Khan, H., Grallert, H., Mahajan, A., Prokopenko, I., Kang, H. M., Dina, C., Esko, T., Fraser, R. M., Kanoni, S., Kumar, A., Lagou, V., Langenberg, C., … McCarthy, M. I. (2012). Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature Genetics*, *44*(9), 981–990. https://doi.org/10.1038/ng.2383

Moser, G., Lee, S. H., Hayes, B. J., Goddard, M. E., Wray, N. R., & Visscher, P. M. (2015). Simultaneous Discovery, Estimation and Prediction Analysis of Complex Traits Using a Bayesian Mixture Model. *PLoS Genetics*, *11*(4). https://doi.org/10.1371/journal.pgen.1004969

Müller, P., Parmigiani, G., Robert, C., & Rousseau, J. (2004). Optimal Sample Size for Multiple Testing: The Case of Gene Expression Microarrays. *Journal of the American Statistical Association*, *99*(468), 990–1001. JSTOR.

Nielsen, J. B., Thorolfsdottir, R. B., Fritsche, L. G., Zhou, W., Skov, M. W., Graham, S. E., Herron, T. J., McCarthy, S., Schmidt, E. M., Sveinbjornsson, G., Surakka, I., Mathis, M. R., Yamazaki, M., Crawford, R. D., Gabrielsen, M. E., Skogholt, A. H., Holmen, O. L., Lin, M., Wolford, B. N., … Willer, C. J. (2018). Biobank-driven genomic discovery yields new insight into atrial fibrillation biology. *Nature Genetics*, *50*(9), 1234–1239. https://doi.org/10.1038/s41588-018-0171-3

Okada, Y., Momozawa, Y., Sakaue, S., Kanai, M., Ishigaki, K., Akiyama, M., Kishikawa, T., Arai, Y., Sasaki, T., Kosaki, K., Suematsu, M., Matsuda, K., Yamamoto, K., Kubo, M., Hirose, N., & Kamatani, Y. (2018). Deep whole-genome sequencing reveals recent selection signatures linked to evolution and disease risk of Japanese. *Nature Communications*, *9*. https://doi.org/10.1038/s41467-018-03274-0

Paltoo, D. N., Rodriguez, L. L., Feolo, M., Gillanders, E., Ramos, E. M., Rutter, J. L., Sherry, S., Wang, V. O., Bailey, A., Baker, R., Caulder, M., Harris, E. L., Langlais, K., Leeds, H., Luetkemeier, E., Paine, T., Roomian, T., Tryka, K., Patterson, A., … National Institutes of Health Genomic Data Sharing Governance Committees. (2014). Data use under the NIH GWAS data sharing policy and future directions. *Nature Genetics*, *46*(9), 934–938. PubMed. https://doi.org/10.1038/ng.3062

Panagiotou, O. A., Ioannidis, J. P. A., & for the Genome-Wide Significance Project. (2012). What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. *International Journal of Epidemiology*, *41*(1), 273–286. https://doi.org/10.1093/ije/dyr178

Park, J.-H., Wacholder, S., Gail, M. H., Peters, U., Jacobs, K. B., Chanock, S. J., & Chatterjee, N. (2010). Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nature Genetics*, *42*(7), 570–575. https://doi.org/10.1038/ng.610

Pe'er, I., Yelensky, R., Altshuler, D., & Daly, M. J. (2008a). Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genetic Epidemiology*, *32*(4), 381–385. https://doi.org/10.1002/gepi.20303

Pe'er, I., Yelensky, R., Altshuler, D., & Daly, M. J. (2008b). Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genetic Epidemiology*, *32*(4), 381–385. https://doi.org/10.1002/gepi.20303

Pers, T. H., Karjalainen, J. M., Chan, Y., Westra, H.-J., Wood, A. R., Yang, J., Lui, J. C., Vedantam, S., Gustafsson, S., Esko, T., Frayling, T., Speliotes, E. K., Boehnke, M., Raychaudhuri, S., Fehrmann, R. S. N., Hirschhorn, J. N., & Franke, L. (2015). Biological interpretation of genome-wide association studies using predicted gene functions. *Nature Communications*, *6*(1), 5890. https://doi.org/10.1038/ncomms6890

Peterson, C. B., Bogomolov, M., Benjamini, Y., & Sabatti, C. (2016). Many Phenotypes Without Many False Discoveries: Error Controlling Strategies for Multitrait Association Studies. *Genetic Epidemiology*, *40*(1), 45–56. https://doi.org/10.1002/gepi.21942

Pulit, S. L., de With, S. A. J., & de Bakker, P. I. W. (2017). Resetting the bar: Statistical significance in whole-genome sequencing-based association studies of global populations: P ULIT ET AL . *Genetic Epidemiology*, *41*(2), 145–151. https://doi.org/10.1002/gepi.22032

Pulit, S. L., Stoneman, C., Morris, A. P., Wood, A. R., Glastonbury, C. A., Tyrrell, J., Yengo, L., Ferreira, T., Marouli, E., Ji, Y., Yang, J., Jones, S., Beaumont, R., Croteau-Chonka, D. C., Winkler, T. W., Hattersley, A. T., Loos, R. J. F., Hirschhorn, J. N., Visscher, P. M., … Lindgren, C. M. (2019). Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry. *Human Molecular Genetics*, *28*(1), 166–174. https://doi.org/10.1093/hmg/ddy327

Ray, D., & Boehnke, M. (2018). Methods for meta-analysis of multiple traits using GWAS summary statistics. *Genetic Epidemiology*, *42*(2), 134–145. https://doi.org/10.1002/gepi.22105

Risch, N., & Merikangas, K. (1996). The Future of Genetic Studies of Complex Human Diseases. *Science*, *273*(5281), 1516–1517. https://doi.org/10.1126/science.273.5281.1516

Schwartzman, A., & Lin, X. (2011). The effect of correlation in false discovery rate estimation. *Biometrika*, *98*(1), 199–214. https://doi.org/10.1093/biomet/asq075

Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration of ρ Values for Testing Precise Null Hypotheses. *The American Statistician*, *55*(1), 62–71. https://doi.org/10.1198/000313001300339950

Servin, B., & Stephens, M. (2007). Imputation-Based Analysis of Association Studies: Candidate Regions and Quantitative Traits. *PLoS Genetics*, *3*(7), 13.

Shen, J. J., Huang, L., Li, L., Jorgez, C., Matzuk, M. M., & Brown, C. W. (2009). Deficiency of Growth Differentiation Factor 3 Protects against Diet-Induced Obesity by Selectively Acting on White Adipose. *Molecular Endocrinology*, *23*(1), 113–123. https://doi.org/10.1210/me.2007-0322

Siegmund, D. O., Zhang, N. R., & Yakir, B. (2011). False discovery rate for scanning statistics. *Biometrika*, *98*(4), 979–985. https://doi.org/10.1093/biomet/asr057

Speliotes, E. K., Willer, C. J., Berndt, S. I., Monda, K. L., Thorleifsson, G., Jackson, A. U., Lango Allen, H., Lindgren, C. M., Luan, J., Mägi, R., Randall, J. C., Vedantam, S., Winkler, T. W., Qi, L., Workalemahu, T., Heid, I. M., Steinthorsdottir, V., Stringham, H. M., Weedon, M. N., … Loos, R. J. F. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics*, *42*(11), 937–948. https://doi.org/10.1038/ng.686

Stanley, A., Ponde, C. K., Rajani, R. M., & Ashavaid, T. F. (2017). Association between genetic loci linked to HDL-C levels and Indian patients with CAD: A pilot study. *Heart Asia*, *9*(1), 9–13. https://doi.org/10.1136/heartasia-2016-010822

Strunz, T., Lauwen, S., Kiel, C., Hollander, A. den, & Weber, B. H. F. (2020). A transcriptome-wide association study based on 27 tissues identifies 106 genes potentially relevant for disease pathology in age-related macular degeneration. *Scientific Reports*, *10*(1), 1584. https://doi.org/10.1038/s41598-020-58510-9

Sun, L., Craiu, R. V., Paterson, A. D., & Bull, S. B. (2006). Stratified false discovery control for large-scale hypothesis testing with application to genome-wide association studies. *Genetic Epidemiology*, *30*(6), 519–530. https://doi.org/10.1002/gepi.20164

Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., Taliun, S. A. G., Corvelo, A., Gogarten, S. M., Kang, H. M., Pitsillides, A. N., LeFaive, J., Lee, S., Tian, X., Browning, B. L., Das, S., Emde, A.-K., Clarke, W. E., Loesch, D. P., … Abecasis, G.

R. (2019). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *BioRxiv*, 563866. https://doi.org/10.1101/563866

Tang, Y., Ghosal, S., & Roy, A. (2007). Nonparametric Bayesian Estimation of Positive False Discovery Rates. *Biometrics*, *63*(4), 1126–1134. https://doi.org/10.1111/j.1541-0420.2007.00819.x

Tang, Z.-Z., & Lin, D.-Y. (2015). Meta-analysis for Discovering Rare-Variant Associations: Statistical Methods and Software Programs. *American Journal of Human Genetics*, *97*(1), 35–53. https://doi.org/10.1016/j.ajhg.2015.05.001

Teslovich, T. M., Musunuru, K., Smith, A. V., Edmondson, A. C., Stylianou, I. M., Koseki, M., Pirruccello, J. P., Ripatti, S., Chasman, D. I., Willer, C. J., Johansen, C. T., Fouchier, S. W., Isaacs, A., Peloso, G. M., Barbalic, M., Ricketts, S. L., Bis, J. C., Aulchenko, Y. S., Thorleifsson, G., … Kathiresan, S. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, *466*(7307), 707–713. https://doi.org/10.1038/nature09270

The 1000 Genomes Project Consortium, Gibbs, R. A., Boerwinkle, E., Doddapaneni, H., Han, Y., Korchina, V., Kovar, C., Lee, S., Muzny, D., Reid, J. G., Zhu, Y., Wang, J., Chang, Y., Feng, Q., Fang, X., Guo, X., Jian, M., Jiang, H., Jin, X., … Rasheed, A. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

The International HapMap Consortium. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, *449*(7164), 851–861. https://doi.org/10.1038/nature06258

The Wellcome Trust Case Control Consortium, Maller, J. B., McVean, G., Byrnes, J., Vukcevic, D., Palin, K., Su, Z., Howson, J. M. M., Auton, A., Myers, S., Morris, A., Pirinen, M., Brown, M. A., Burton, P. R., Caulfield, M. J., Compston, A., Farrall, M., Hall, A. S., Hattersley, A. T., … Donnelly, P. (2012). Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nature Genetics*, *44*(12), 1294–1301. https://doi.org/10.1038/ng.2435

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., & DePristo, M. A. (2013). From FastQ data to high confidence variant calls: The Genome Analysis Toolkit best practices pipeline. *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]*, *11*(1110), 11.10.1-11.10.33. https://doi.org/10.1002/0471250953.bi1110s43

Wakefield, J. (2007a). A Bayesian Measure of the Probability of False Discovery in Genetic Epidemiology Studies. *American Journal of Human Genetics*, *81*(2), 208–227.

Wakefield, J. (2007b). A Bayesian Measure of the Probability of False Discovery in Genetic Epidemiology Studies. *The American Journal of Human Genetics*, *81*(2), 208–227. https://doi.org/10.1086/519024

Wakefield, J. (2009). Bayes factors for genome-wide association studies: Comparison with P-values. *Genetic Epidemiology*, *33*(1), 79–86. https://doi.org/10.1002/gepi.20359

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA Statement on p-Values: Context, Process, and Purpose. *The American Statistician*, *70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a World Beyond "p < 0.05." *The American Statistician*, *73*(sup1), 1–19. https://doi.org/10.1080/00031305.2019.1583913

Wellcome Trust Case Control Consortium. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, *447*(7145), 661–678. PubMed. https://doi.org/10.1038/nature05911

Wen, X. (2016). Molecular QTL discovery incorporating genomic annotations using Bayesian false discovery rate control. *The Annals of Applied Statistics*, *10*(3), 1619–1638. https://doi.org/10.1214/16-AOAS952

Wen, X. (2017). Robust Bayesian FDR Control Using Bayes Factors, with Applications to Multi-tissue eQTL Discovery. *Statistics in Biosciences*, *9*(1), 28–49. https://doi.org/10.1007/s12561-016-9153-0

Whittemore, A. S. (2007). A Bayesian False Discovery Rate for Multiple Testing. *Journal of Applied Statistics*, *34*(1), 1–9. https://doi.org/10.1080/02664760600994745

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

Willer, C. J., Sanna, S., Jackson, A. U., Scuteri, A., Bonnycastle, L. L., Clarke, R., Heath, S. C., Timpson, N. J., Najjar, S. S., Stringham, H. M., Strait, J., Duren, W. L., Maschio, A., Busonero, F., Mulas, A., Albai, G., Swift, A. J., Morken, M. A., Narisu, N., … Abecasis, G. R. (2008). Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nature Genetics*, *40*(2), 161–169. https://doi.org/10.1038/ng.76

Willer, C. J., Schmidt, E. M., Sengupta, S., Peloso, G. M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M. L., Mora, S., Beckmann, J. S., Bragg-Gresham, J. L., Chang, H.-Y., Demirkan, A., Den Hertog, H. M., Do, R., Donnelly, L. A., Ehret, G. B., Esko, T., … Global Lipids Genetics Consortium. (2013). Discovery and refinement of loci associated with lipid levels. *Nature Genetics*, *45*(11), 1274–1283. https://doi.org/10.1038/ng.2797

Wilson, M. A., Iversen, E. S., Clyde, M. A., Schmidler, S. C., & Schildkraut, J. M. (2010). BAYESIAN MODEL SEARCH AND MULTILEVEL INFERENCE FOR SNP ASSOCIATION STUDIES. *The Annals of Applied Statistics*, *4*(3), 1342–1364.

Winkler, T. W., Day, F. R., Croteau-Chonka, D. C., Wood, A. R., Locke, A. E., Mägi, R., Ferreira, T., Fall, T., Graff, M., Justice, A. E., Luan, J., Gustafsson, S., Randall, J. C., Vedantam, S., Workalemahu, T., Kilpeläinen, T. O., Scherag, A., Esko, T., Kutalik, Z., … Loos, R. J. (2014). Quality control and conduct of genome-wide association meta-analyses. *Nature Protocols*, *9*(5), 1192–1212. https://doi.org/10.1038/nprot.2014.071

Wood, A. R., Esko, T., Yang, J., Vedantam, S., Pers, T. H., Gustafsson, S., Chu, A. Y., Estrada, K., Luan, J., Kutalik, Z., Amin, N., Buchkovich, M. L., Croteau-Chonka, D. C., Day, F. R., Duan, Y., Fall, T., Fehrmann, R., Ferreira, T., Jackson, A. U., … Frayling, T. M. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics*, *46*(11), 1173–1186. PubMed. https://doi.org/10.1038/ng.3097

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., & Lin, X. (2011). Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. *American Journal of Human Genetics*, *89*(1), 82–93. https://doi.org/10.1016/j.ajhg.2011.05.029

Xu, ChangJiang, Ciampi, A., & Greenwood, C. M. T. (2014). Exploring the potential benefits of stratified false discovery rates for region-based testing of association with rare genetic variation. *Frontiers in Genetics*, *5*. https://doi.org/10.3389/fgene.2014.00011

Xu, Chao, Wu, K., Zhang, J.-G., Shen, H., & Deng, H.-W. (2017). Low-, high-coverage, and two-stage DNA sequencing in the design of the genetic association study. *Genetic Epidemiology*, *41*(3), 187–197. https://doi.org/10.1002/gepi.22015

Yang, H., & Wang, K. (2015). Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nature Protocols*, *10*(10), 1556–1566. https://doi.org/10.1038/nprot.2015.105

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

Yengo, L., Sidorenko, J., Kemper, K. E., Zheng, Z., Wood, A. R., Weedon, M. N., Frayling, T. M., Hirschhorn, J., Yang, J., Visscher, P. M., & GIANT Consortium. (2018). Meta-analysis of genome-wide association studies for height and body mass index in ∼700000 individuals of European ancestry. *Human Molecular Genetics*, *27*(20), 3641–3649. https://doi.org/10.1093/hmg/ddy271

Yi, N., & Xu, S. (2008). Bayesian LASSO for Quantitative Trait Loci Mapping. *Genetics*, *179*(2), 1045–1055. https://doi.org/10.1534/genetics.107.085589

Zeggini, E., & Ioannidis, J. P. A. (2009). Meta-analysis in genome-wide association studies. *Pharmacogenomics*, *10*(2), 191–201. https://doi.org/10.2217/14622416.10.2.191

Zhang, Y., Qi, G., Park, J.-H., & Chatterjee, N. (2018). Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. *Nature Genetics*, *50*(9), 1318–1326. https://doi.org/10.1038/s41588-018-0193-x

Zhou, W., Nielsen, J. B., Fritsche, L. G., Dey, R., Gabrielsen, M. E., Wolford, B. N., LeFaive, J., VandeHaar, P., Gagliano, S. A., Gifford, A., Bastarache, L. A., Wei, W.-Q., Denny, J. C., Lin, M., Hveem, K., Kang, H. M., Abecasis, G. R., Willer, C. J., & Lee, S. (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature Genetics*, *50*(9), 1335–1341. https://doi.org/10.1038/s41588-018-0184-y

Zhou, X., Carbonetto, P., & Stephens, M. (2013). Polygenic Modeling with Bayesian Sparse Linear Mixed Models. *PLoS Genetics*, *9*(2). https://doi.org/10.1371/journal.pgen.1003264

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M. R., Powell, J. E., Montgomery, G. W., Goddard, M. E., Wray, N. R., Visscher, P. M., & Yang, J. (2016a). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature Genetics*, *48*(5), 481–487. https://doi.org/10.1038/ng.3538

Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M. R., Powell, J. E., Montgomery, G. W., Goddard, M. E., Wray, N. R., Visscher, P. M., & Yang, J. (2016b). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature Genetics*, *48*(5), 481–487. https://doi.org/10.1038/ng.3538

Zuk, O., Schaffner, S. F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M. J., Neale, B. M., Sunyaev, S. R., & Lander, E. S. (2014). Searching for missing heritability: Designing rare variant association studies. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(4), E455-464. https://doi.org/10.1073/pnas.1322563111